

**UNIVERSIDAD COMPLUTENSE DE
MADRID**

FACULTAD DE INFORMÁTICA



TRABAJO DE FIN DE GRADO

Propagación Covid 19: Recolección, Tratamiento y
Análisis de los datos

Covid 19 Spread: Data Collection, Treatment and
Analysis

Dirigido por: Sonia Estévez Martín

Iván Jesús Saugar Peinado

Sergio Sanchez Ortiz

Álvaro San José Guerra

Grado en Ingeniería Informática

Curso académico 2020-21

Convocatoria Septiembre

Agradecimientos

Este Trabajo de Fin de Grado representa el final de una etapa de nuestra vida, es el resultado de mucho trabajo y esfuerzo durante varios años. Es por ello por lo que, en primer lugar, queremos dar las gracias a aquellas personas que siempre han estado y confiado en nosotros, haciendo que nos superásemos día a día, por eso, nuestro agradecimiento va en especial a nuestras familias y amigos más cercanos, que siempre han estado ahí, para motivarnos y darnos ese apoyo incondicional cuando lo hemos necesitado.

En segundo lugar, nos gustaría dar las gracias a todos nuestros compañeros y profesores, que hemos conocido a lo largo de nuestra trayectoria académica, en especial a nuestro pastor Carlos, por todo su apoyo y paciencia a lo largo de la carrera.

Sin olvidarnos de hacer una mención especial y agradecimiento a Sonia Estévez y María Victoria, por el esfuerzo, las ganas y el tiempo que han dedicado a ayudarnos con la realización de este trabajo.

Resumen

Este Trabajo de Fin de Grado es una aportación al estudio analítico sobre la evolución del COVID-19, aportando un repositorio capaz de generar conjuntos de datos de interés. El estudio se centra en América, Asia y Europa, analizando los países más afectados. Mediante un ranking cuantitativo con los datos de los países considerados, se han seleccionado un conjunto de países para realizar un estudio con varios índices para medir la evolución y peligrosidad de la pandemia. Se han utilizado tres índices: índice de transmisión R_0 , incidencia acumulada (a 14 días) e índice de peligrosidad (IP). Además se ha realizado una clasificación mediante el algoritmo K-Means utilizando como referencia los índices citados durante un periodo de 25 días alrededor de la segunda ola. El estudio se ha realizado por regiones a dos niveles: Comunidades Autónomas en España y diversas regiones en el mundo. Se han generado y analizado gráficas a partir de las variables más representativas e índices con el objetivo de observar el impacto causado por el COVID-19 sobre las distintas regiones.

Palabras Clave: *Covid19, Análisis de datos, Clasificación, Índice de Peligrosidad, Factor R_0 , Incidencia Acumulada, Procesamiento de datos*

Abstract

This Final Degree Project is a contribution to the analytical study on the evolution of COVID-19 worldwide, providing a repository capable of generating a database from the data provided by John Hopkins University. The study focuses on America, Asia and Europe, analysing the most affected countries and classifying them. Once classified, graphs were generated and analysed based on the most representative variables in order to observe the impact caused by COVID-19 on the different regions. As a complementary study, an assessment was carried out by looking at the worst affected region of the most affected countries.

Keywords: *Covid19, Data Analysis, Classification, Danger Index, R0 Factor, Accumulated incidence, Data processing*

Índice general

1. Introducción	9
1.1. Contexto	9
1.2. Motivación	10
1.3. Objetivos	11
1.4. Estructura del documento	11
2. Introduction	13
2.1. Context	13
2.2. Motivation	14
2.3. Objectives	15
2.4. Structure of the document	15
3. Estado del Arte	17
3.1. Organización Mundial de la Salud (OMS)	17
3.2. Universidad Johns Hopkins (JHU)	17
3.3. Datos oficiales nacionales	19
4. Selección de datos	24
4.1. Datos oficiales	24
4.2. Tabla Bases de datos por continente	25
4.3. Puntuación	29
4.4. América	30
4.5. Asia	31
4.6. Europa	34
4.7. Países Seleccionados	35
5. Resultados selección de datos	37
5.1. GitHub	37
5.2. Dataset Generado	40
5.3. Análisis de datos oficiales nacionales	42

6. Metodología del análisis	52
6.1. Creación de scripts y datasets	52
6.2. Variables e Indicadores	55
6.3. Método de clustering	57
7. Resultados del Análisis Clúster	61
7.1. Clasificación por regiones de España	61
7.2. Clasificación por regiones de Asia, Europa y América . .	71
8. Contribuciones individuales	80
8.1. Contribución de Iván Jesús Saugar Peinado	80
8.2. Contribución de Sergio Sánchez Ortiz	82
8.3. Contribución de Álvaro San José Guerra	83
9. Conclusiones y Trabajo futuro	85
9.1. Conclusiones	85
9.2. Trabajo Futuro	86
10. Conclusions and Future work	87
10.1. Conclusions	87
10.2. Future Work	88
Bibliografía	88

Índice de figuras

3.1. Dashboard de la OMS	18
3.2. Países más afectados según la OMS	18
3.3. Datasets ofrecidos por la JHU	19
3.4. Dashboard de la JHU	20
3.5. Datos del Ministerio de Sanidad de España	20
3.6. Incidencia Acumulada en España	20
3.7. Datos descargables Costa Rica	21
3.8. Datos ofrecidos por Irak	21
3.9. Datos ofrecidos por Japón	22
3.10. Datos ofrecidos por Mongolia	23
4.1. Bases de datos mundiales	25
4.2. Bases de datos oficiales de América	26
4.3. Bases de datos oficiales de Asia	27
4.4. Bases de datos oficiales de Europa	28
4.5. Puntuación empleada para la clasificación de los países. .	30
4.6. Distribución de puntos respecto al número de infectados.	30
4.7. Clasificación América del norte	30
4.8. Clasificación América Central	31
4.9. Clasificación América del Sur	31
4.10. Clasificación Asia Septentrional	31
4.11. Clasificación Asia Central	32
4.12. Clasificación Asia Occidental	32
4.13. Clasificación Asia Meridional	33
4.14. Clasificación Asia Oriental	33
4.15. Clasificación Sudeste Asiático	33
4.16. Clasificación Europa Meridional	34
4.17. Clasificación Europa Septentrional	34
4.18. Clasificación Europa Central	35
4.19. Clasificación Europa Occidental	35
4.20. Clasificación Europa Oriental	35
5.1. GitHub	38

Propagación Covid 19
Recolección, Tratamiento y Análisis de los datos

5.2. GitHub Deploy	38
5.3. GitHub Counter-Filter, USA-State y World whitout USA	39
5.4. GitHub R	39
5.5. GitHub R FUNTIONS	40
5.6. Clasificación América del norte	42
5.7. Clasificación América Central	43
5.8. Clasificación América del Sur	43
5.9. Clasificación Asia Septentrional	44
5.10. Clasificación Asia Central	44
5.11. Clasificación Asia Occidental	45
5.12. Clasificación Asia Meridional	46
5.13. Clasificación Asia Oriental	46
5.14. Clasificación Sudeste Asiático	47
5.15. Clasificación Europa Meridional	48
5.16. Clasificación Europa Septentrional	49
5.17. Clasificación Europa Central	50
5.18. Clasificación Europa Occidental	50
5.19. Clasificación Europa Oriental	51
6.1. Algoritmo K-Means. Fuente: Towards Data Science	57
6.2. K-Means: Paso de inicialización (Pasos 1,2,3). Fuente: Wikipedia	58
6.3. K-Means: Pasos generales (Pasos 4,5). Fuente: Wikipedia	58
6.4. K-Means: Pasos generales (Paso 6). Fuente: Wikipedia	59
6.5. K-Means: Pasos generales (Paso 7). Fuente: Wikipedia	59
7.1. Gráficas Generales por comunidades España	62
7.2. Análisis cluster por regiones de España en base a R0 y a la segunda ola	64
7.3. Gráficas R0 España	65
7.4. Análisis cluster por regiones de España en base a IP y a la segunda ola	67
7.5. Gráficas Índice Peligrosidad España	68
7.6. Análisis cluster por regiones de España en base a IA y a la segunda ola	70
7.7. Gráfica General Regiones América	72
7.8. Gráfica General Regiones Asia	73
7.9. Gráfica General Regiones Europa	74
7.10. Análisis cluster por regiones de Asia, Europa y América en base a R0 y a la segunda ola	76
7.11. Análisis cluster por regiones de Asia, Europa y América en base a IP y a la segunda ola	77

7.12. Análisis cluster por regiones de Asia, Europa y América
en base a IA y a la segunda ola 78

Capítulo 1

Introducción

1.1. Contexto

A finales del año 2019, se detectó una epidemia de neumonía de origen desconocido que fue declarada pandemia a principios de 2020 [1] a causa del virus SARS-Cov-2.

Los primeros casos se concentraron en uno de los mercados de China y un equipo del Centro para el Control y Prevención de Enfermedades (CDC) fue asignado para investigar los contagios [2]. Tras las investigaciones se descubrieron más casos y recogieron muestras a los pacientes. Una vez identificada como una nueva enfermedad, los expertos se percataron también de la facilidad de transmisión entre la población. El 30 de enero, la OMS declaró la enfermedad como una emergencia de salud pública de interés internacional [3]. El 11 de marzo y con la enfermedad presente en más de cien territorios a nivel mundial la OMS eleva a la categoría de Pandemia la anterior emergencia sanitaria [4]. La evolución del virus fue de tal magnitud y de semejante velocidad, que la mayoría de los gobiernos no fue capaz de reaccionar a tiempo, provocando una situación crítica en la cual se tomaron medidas drásticas para poder frenar la propagación del virus, hasta que se dispusieron medidas eficaces para paliar los efectos. Una de las medidas tomadas fue el confinamiento masivo de los habitantes cuyos efectos paralizaron parcial o totalmente ciertos sectores de la economía [5]. Otro de los grandes afectados fue el sector sanitario, que se vio sobrepasado en la mayoría de los países. Para solventar este problema, muchos gobiernos optaron por la creación de hospitales de campaña para poder abastecer de camas al gran número de enfermos de Covid-19 [6]. Además, se empezó una carrera por el desarrollo de una vacuna [7].

Tras medidas de confinamiento y del uso de mascarillas, se redujeron las curvas de contagios y de muertes de manera significativa comparado

con los inicios de la pandemia, y fue entonces en torno a los meses de verano, cuando se levantaron las medidas de restricción de movilidad y de reunión con personas [8]. Siendo este el origen del segundo repunte de casos en muchos de los países afectados, aunque, en este caso con menor impacto que en los peores momentos de la primera ola. En noviembre de 2020 se empezaron a testar las primeras vacunas [9] contra el virus, cuya efectividad se encontraba en más de un noventa por ciento [10].

El día 11 de noviembre de 2020, la Comisión Europea oficializó el contrato con las compañías farmacéuticas para el desarrollo de estas vacunas y la potencial compra de estas [11]. El 21 de diciembre del mismo año, la Unión Europea autorizó la comercialización de dichas vacunas.

La tasa de vacunación se incrementa día a día [12], en algunos países, como Israel, cuentan con aproximadamente un 60 % de la población vacunada a mayo de 2021, por este motivo, se comenzó a suavizar las restricciones en la mayoría de los países desarrollados [13].

1.2. Motivación

Desde la aparición del virus se ha generado una gran cantidad de datos sobre la pandemia de fuentes muy diversas, que van desde instituciones académicas hasta las propias agencias de salud o ministerios de cada país. Dado la diversidad de las fuentes, no existe homogeneidad en la información facilitada, variando tanto en la forma de presentarla como en la variedad de los datos y la frecuencia en la actualización de estos. Además, los índices utilizados por los distintos países son también muy diversos.

Por una parte, la motivación de este estudio está en analizar índices que puedan informar de la evolución de la pandemia y comparar la situación en diferentes países y regiones. Para este estudio, se han considerado diversas variables, como por ejemplo el número de casos infectados/recuperados/muertos diariamente. Algunos países no publican todas las variables y otros publican información agregada. Aunque en este estudio se han recopilado más variables, las más importantes son las relativas a datos diarios sobre infectados, recuperados y fallecidos. Esto es porque los índices que se van a estudiar para analizar la evolución de la pandemia se calculan a partir de estas variables. Estos índices son tres: El índice de transmisión (R_0), la incidencia acumulada (IA) y el índice de peligrosidad (IP). En los trabajos [14, 15, 16] se definen y se justifican estos índices con detalle.

Por otra parte, la mayoría de los gobiernos e instituciones facilitan datos agregados o visualizaciones mediante aplicaciones de cuadros de mando (dashboard), no existiendo una manera directa de acceso a los

datos crudos a nivel mundial. Por esta razón, se ha decidido aportar la creación de un repositorio abierto con los datos reunidos de forma que, cualquier otro organismo o particular pueda acceder a los mismos y emplearlos, así como varios estudios de la evolución de la pandemia a nivel mundial en base a los datos y variables generadas.

1.3. Objetivos

El trabajo aquí presentado tiene los siguientes objetivos:

1. Recolección y preprocesado de los datos. En este trabajo se ha creado una base de datos fruto de la recolección de datos fiables (procedentes de organismos oficiales) con el mayor número de países posible. Además, se han realizado operaciones de tratamiento de datos (extracción, filtrado, transformación de datos) permitiendo un análisis posterior más adecuado.
2. Clasificación y selección. El siguiente objetivo consiste en la selección de los países más afectados por el Covid-19 de los siguientes continentes: Europa, Asia y América. Mediante una clasificación, dividiendo cada continente en regiones, puntuando los países en función de la calidad de la información aportada y el número de infectados.
3. Análisis. Se han realizado dos análisis de clasificación utilizando el algoritmo K-Means: el primero sobre las comunidades autónomas de España y el segundo sobre las regiones de los países con la puntuación más alta obtenida en la clasificación de cada continente. Estos análisis se centran en índices anteriormente citados: Índice de transmisión (R_0), Índice de peligrosidad (IP) e Incidencia Acumulada (IA) El análisis comparativo se ha realizado sobre una ventana de información correspondiente a la segunda ola de cada región/país.
4. Difusión. Este trabajo incluye repositorio GitHub con todos los procesos desarrollados para facilitar futuros análisis, tanto para organizaciones como para investigadores o estudiantes.

1.4. Estructura del documento

La memoria estará dividida en los siguientes apartados:

1. Introducción: Breve resumen del proyecto, motivaciones y objetivos de este.

2. Estado del arte: Datos proporcionados por instituciones gubernamentales e instituciones académicas.
3. Selección de Datos: Bases de datos que son de interés para los estudios posteriores y clasificación de los países de América, Asia y Europa mediante una puntuación propia..
4. Resultados de la selección de datos: Resultados obtenidos de la selección de datos, tanto el repositorio creado en GitHub como el análisis de la clasificación creada.
5. Metodología: Memoria de la metodología llevada a cabo para la obtención de los objetivos del proyecto.
6. Resultados: Exposición y análisis de los datos obtenidos de los estudios realizados.
7. Contribuciones Individuales: Contiene las contribuciones realizadas por cada miembro del equipo.
8. Conclusiones y trabajo a futuro: Contiene las conclusiones realizadas tras el desarrollo del proyecto y las recomendaciones para el trabajo futuro.

Capítulo 2

Introduction

2.1. Context

At the end of 2019, a pneumonia outbreak of unknown origin was detected and declared a pandemic in early 2020 [1] due to the SARS-Cov-2 virus.

The first cases were concentrated in one of China's markets, and a team from the Center for Disease Control and Prevention was assigned to investigate the contagions [2]. After the research was made more cases were discovered and samples were taken from the patients. Once it was identified as a new disease the experts realized the ease in which the disease spreaded between the population. In the 30 of January of 2020, the World Health Organization declared the disease as an international health emergency [3]. By the 11 of March, the disease was present in more than one hundred territories worldwide the WHO elevated the health emergency to global pandemic [4].

The evolution of the virus was of such big a magnitude and spreaded so fast that the majority of the governments were not able to react on time. This cause a critic situation in which some drastic decisions were taken in order to slow down the propagation of the virus until the needed tools to fight out the efect of it were available. One of the measures taken was a massive lockdown of the inhabitants, this made some sectors of the economy partially or fully paralyzed [5]. The sanitary sector was one of the most affected ones and was surpassed by the situation in most of the countries. To solve out this issue, many governments created field hospitals to supply with beds to the large number of Covid-19 patients[6]. What's more, a race for the development of the vaccine started [7].

After confinement measures and the use of face masks, the infection curve and deaths were reduced significantly compared with the start of the pandemic, it was by the time of summer when the precautions of

travel restrictions were removed [8]. Being this factor, the origin of the second upturn of cases in many of the affected countries, although, this upturn had less impact than the first one. In November 2020 the first vaccines were tested [9], their effectiveness was around 90

The eleventh of November 2020, the European Commission officialized the contract with the pharmaceutical companies for the development and a potential purchase of these vaccines [11]. The 21 of December of that same year, the European Union authorized its commercialization.

The vaccination rate increases day by day [12], in some countries such as Israel, had on May 2021 nearly the 60% of their population vaccinated. Due to this fast rate, the majority of the restrictions were lowered in most of the countries [13].

2.2. Motivation

Since the appearance of the Covid-19, a large amount of data has been generated from very diverse sources, this goes from academic institutions to the health agencies or ministries of each country. Due to the diversity of the sources, there is no homogeneity in the information provided, varying both in the way it is presented and in the variety of data and the frequency with which it is updated. Moreover, all the indicators used by each country are very diverse.

On the one hand, the aim of this study relies on analyzing indicators that can provide information about the evolution of the pandemic and compare this situation between different countries. For this research, a variety of variables have been taken into consideration, for example the daily number of: new infected cases, recovered cases and deaths. Some countries, do not release all these variables and other just present the data aggregated. Even though, in this research more variables have been gathered, the most relevant ones are those that treat the data daily. This is motivated by the fact that the indicators that are going to be studied to analyze the evolution of the pandemic are calculated using this data. These indicators are the following: transmission index (R_0), Danger index (IP) and cumulative incidence (IA). In [14, 15, 16] they are defined and justified with detail.

On the other hand, most governments and institutions provide aggregated data or visualizations through dashboard applications, and there is no direct way to access the raw data at the global level. For this reason, it has been decided to contribute to the creation of an open repository with the collected data so that any other organization or individual can access and use them, as well as several studies of the evolution of the pandemic worldwide based on the data and variables generated.

2.3. Objectives

This project the following objectives:

1. Data collection and preprocessing. In this work, a data base has been created as a result of the collection of reliable (official agency records) with the largest possible number of countries. In addition, data processing operations (data extraction, filtering, data transformation) have been carried out, allowing a more adequate future analysis.
2. Ranking and selection. The following objective is to select the countries most affected by Covid-19 from the following continents: Europe, Asia and America. By means of a classification, dividing each continent into regions, scoring the countries according into regions, scoring the countries according to the quality of the information provided and the number of people infected.
3. Analysis. Two analysis have been carried out using the K-Means algorithm: the first one of all the autonomous communities of Spain and the second one on the regions of the countries with the highest score obtained in the classification of each continent. These analyses are focused on the previously mentioned indexes: Transmission Index (R0), Hazard Index (IP) and Cumulative Incidence (CI). The comparative analysis has been carried out on a window of information corresponding to the second wave of each region/country.
4. Diffusion. This project includes a GitHub repository with all the processes developed in order to facilitate future analysis, both for organizations and for researchers or students.

2.4. Structure of the document

The memory is divided in the following chapters:

1. Introduction: Brief summary of the project, motivations and objectives.
2. State of the art: Data given by the government and academic institutions.
3. Data selection: Databases that are relevant for further studies and classification of countries in the Americas, Asia and Europe using a self-made score.

4. Data selection results: Results obtained in the data selection, a repository created in GitHub and a deeper analysis of the ranking made before.
5. Methodology: Memory of the methodology performed to reach the objectives of the project.
6. Results: Exposure and analysis of the studies made.
7. Individual contributions: Contains the contributions made by every member of the team.
8. Conclusions and future work: Explains the conclusions made after the development of the project and recommendations for the future work.

Capítulo 3

Estado del Arte

Los estudios más seguros y fiables a nivel mundial han sido proporcionados por la Organización Mundial de la Salud (OMS) y la universidad Johns Hopkins, incluso algunos países utilizan los datos recopilados por estos estudios. A continuación, se explica más en detalle estos datos.

3.1. Organización Mundial de la Salud (OMS)

Los estudios realizados por la OMS son bastantes extensos, proporcionando el número de casos de contagios, número de muertes, tipo de transmisión del virus, ya sea mostrando estos datos de forma acumulada de los últimos siete días o solamente las últimas veinticuatro horas. Su página web [17] utiliza estos mismos datos para generar un dashboard, ver Figura 3.1, incorporando un mapa geográfico mundial de calor permitiendo mostrar los países donde se puede encontrar más casos confirmados, al igual que el número de muertes o transmisión del virus. También incorpora histogramas lineales acumulativos semanales y diarios de cada tipo de caso, permitiendo ver de una forma rápida las curvas acumulativas de muertos e infectados, además de un porcentaje de variación semanal y diario. En caso de no seleccionar ningún país, muestra informes globales sobre las distintas regiones de la OMS, en las que se comparan entre dichas regiones. Por último, la Figura 3.2 muestra una clasificación de los países más afectados ya sea, por casos de contagio o número de muertos en formato histograma. Cabe destacar, que la OMS mide la evolución del virus en tiempo real a partir de informes sanitarios.

3.2. Universidad Johns Hopkins (JHU)

La universidad de medicina Johns Hopkins también abarca un estudio a nivel mundial de los casos surgidos por el Covid-19, a diferencia de la

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

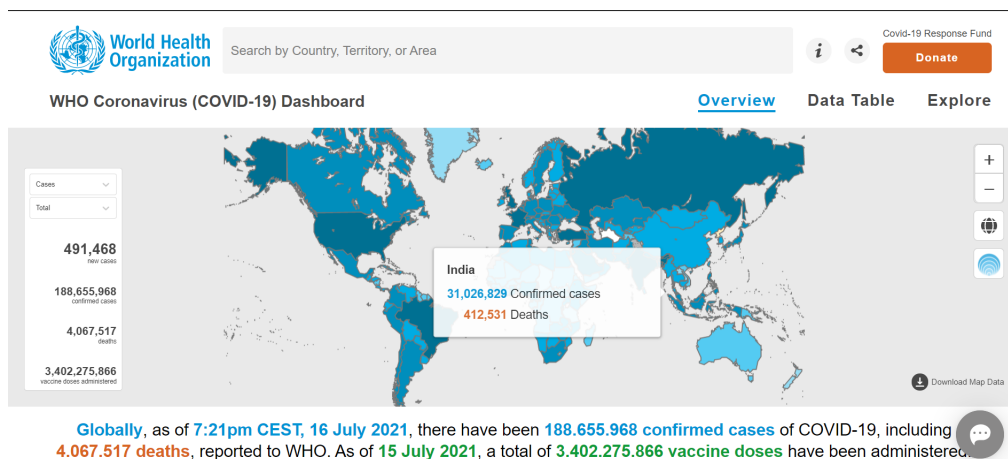


Figura 3.1: Dashboard de la OMS

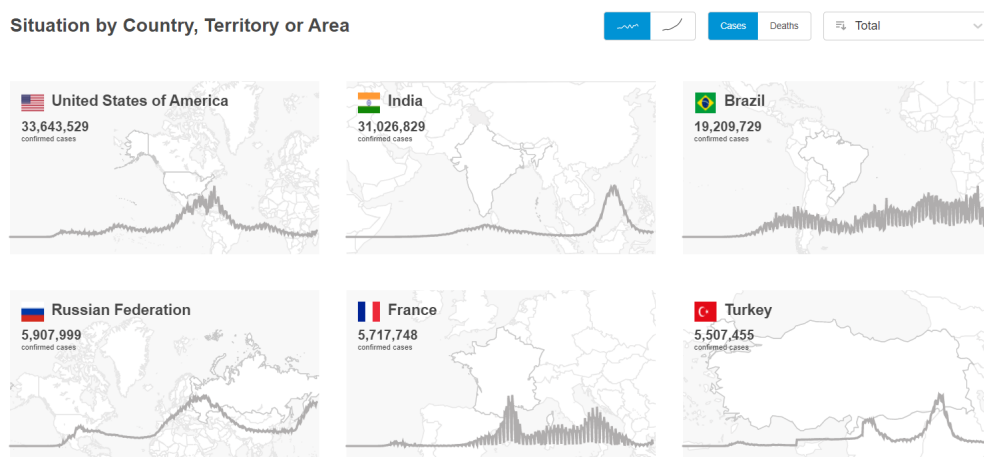


Figura 3.2: Países más afectados según la OMS

OMS, que se actualiza en tiempo real, la JHU CSSE se actualiza en función de repositorios de datos generados por distintas organizaciones Figura 3.3, por ejemplo: OMS, European Centre for Disease Prevention and Control (ECDC), Centers for Disease Control and Prevention (CDC) o BNO News entre otros.

La base de datos generada a partir de los datos recopilados por la Universidad Johns Hopkins utiliza las siguientes variables para todos los países: Province State, Country Region, Last Update, Lat, Long, Confirmed, Deaths, Recovered, Active, Incident Rate y Case Fatality Ratio, exceptuando los Estados Unidos de América, facilitando para este país cuatro variables adicionales: Total Test Results, People Hospitalized, Testing Rate y Hospitalization Rate.

Utilizando las variables anteriormente mencionadas, la Universidad de medicina Johns Hopkins, genera un mapa geográfico mundial Figura 3.4, permitiendo seleccionar un país y mostrando un mapa de calor el cual, dependiendo del número de casos que tenga el área afectada, la densidad del volumen varía. Una ventaja que tiene en comparación con el gráfico de la OMS es que, este distingue entre las diferentes regiones afectadas de algunos países como el caso de España o Alemania, no aportando solo la suma nacional. Por último, indicar que también muestra una clasificación con los datos acumulativos[18].

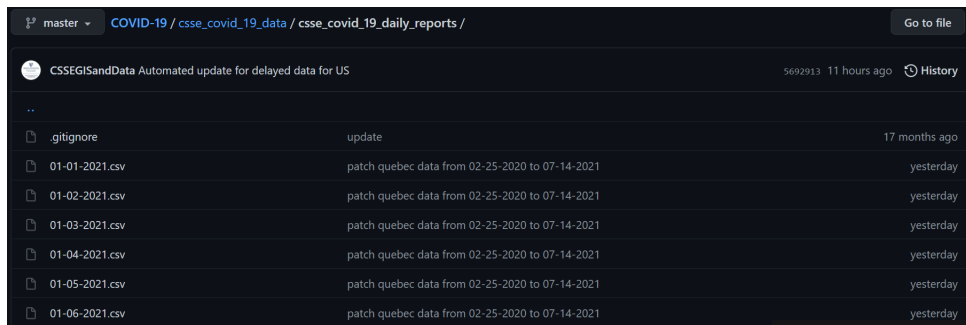


Figura 3.3: Datasets ofrecidos por la JHU

3.3. Datos oficiales nacionales

Aparte de lo expuesto con anterioridad en general, cada país expone de forma oficial sus propios datos, a través de diferentes fuentes, como ministerios, agencias gubernamentales u otras instituciones oficiales. La Figura 3.5 muestra los datos de España el día 14 de mayo de 2021 y la Figura 3.6 muestra la gráfica del índice Incidencia Acumulada en el periodo julio 2020 - mayo 2021.

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

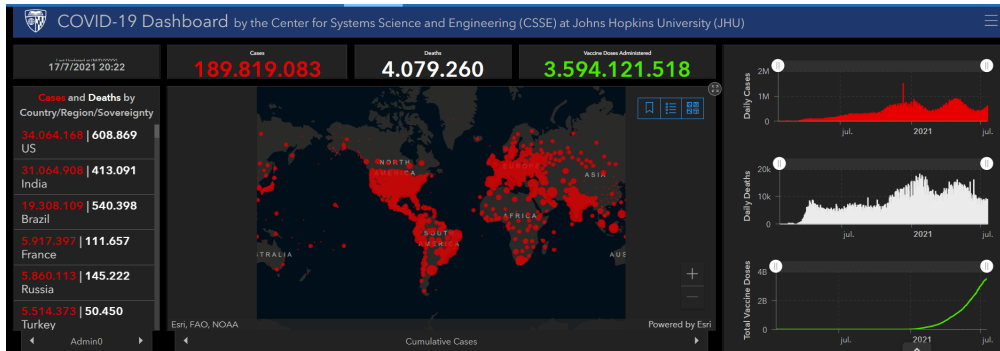


Figura 3.4: Dashboard de la JHU

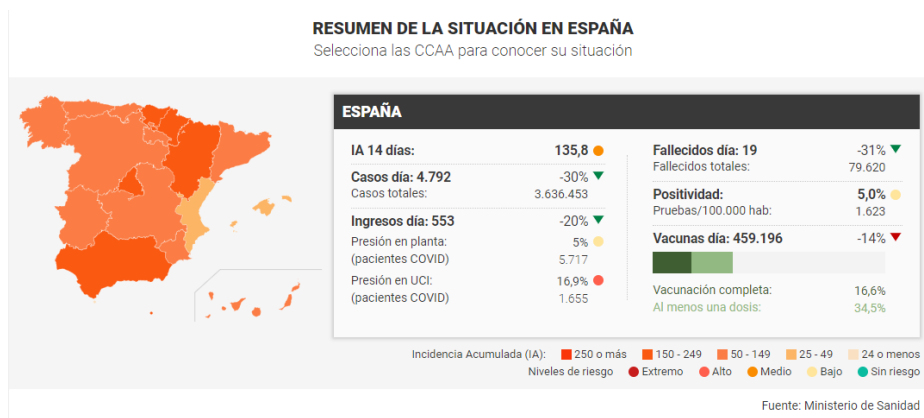


Figura 3.5: Datos del Ministerio de Sanidad de España



Figura 3.6: Incidencia Acumulada en España

Propagación Covid 19 Recolección, Tratamiento y Análisis de los datos

Algunos países además de ofrecer datos regionales y proporcionar un dashboard, dan facilidades para descargar los datos, como es el caso de Costa Rica, Figura 3.7.

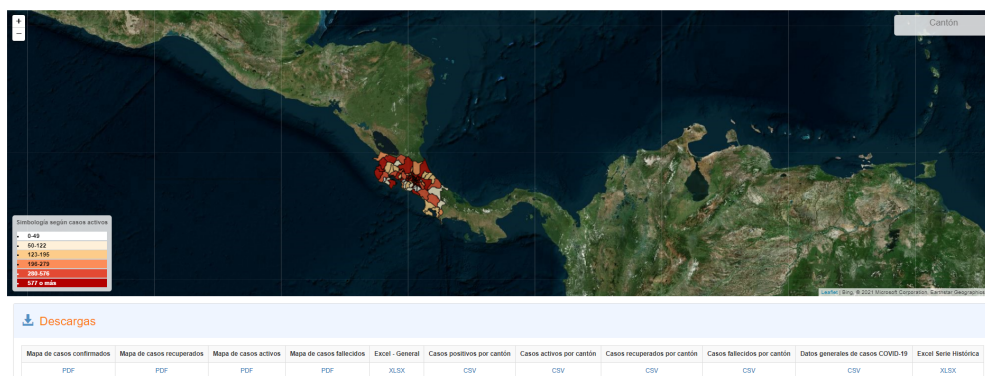


Figura 3.7: Datos descargables Costa Rica

En ocasiones, los países ofrecen multitud de datos y gráficas, mientras que en otros apenas ofrecen información, un ejemplo de ello, sería Japón e Irak, mientras que Japón ofrece un amplio abanico de información, Figura 3.9, Irak apenas aporta 5 variables, Figura 3.8.



Figura 3.8: Datos ofrecidos por Irak

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

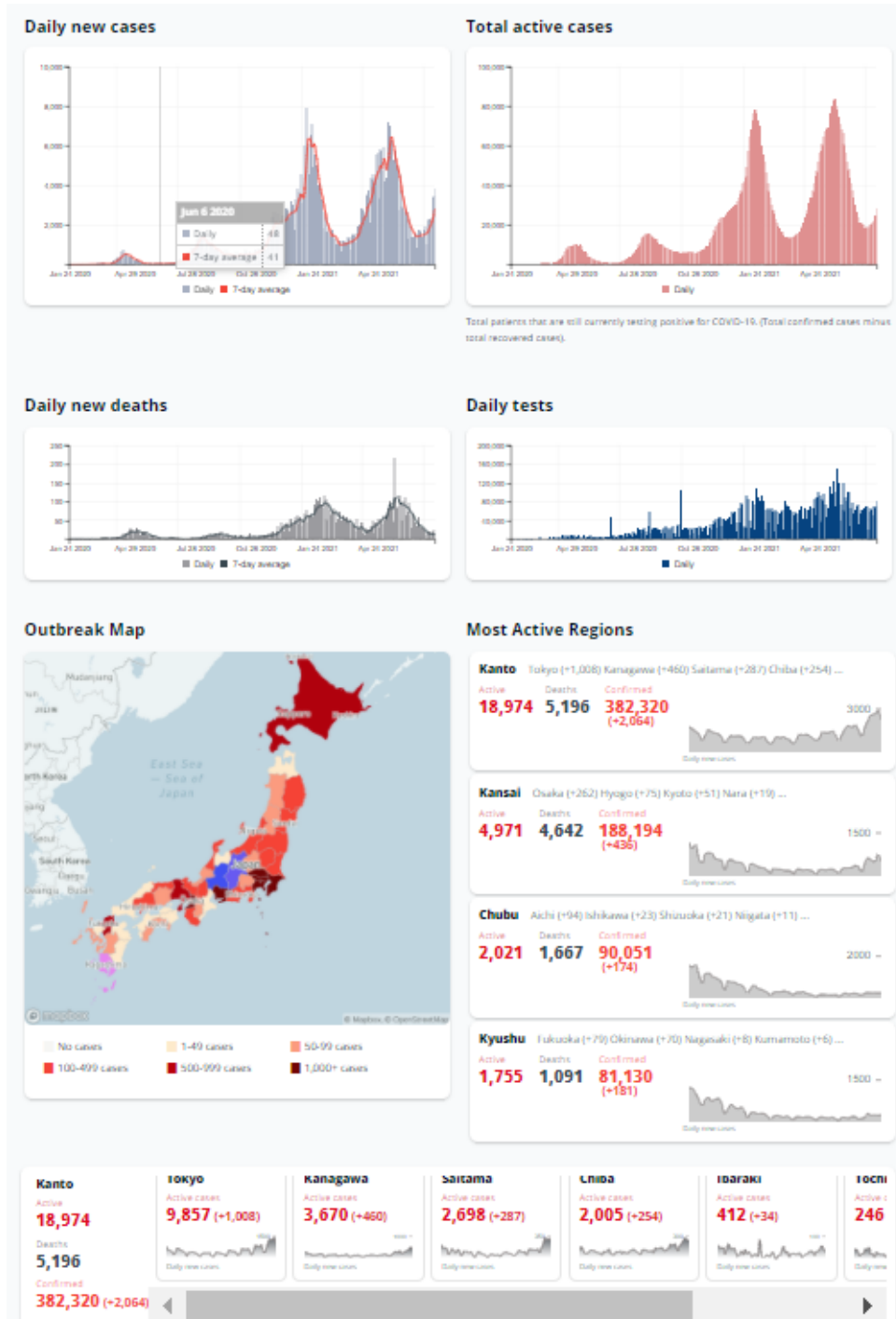


Figura 3.9: Datos ofrecidos por Japón

Propagación Covid 19 Recolección, Tratamiento y Análisis de los datos

Por último, también podemos encontrar casos en los cuales el propio país en vez de recoger y mostrar sus propios datos utiliza organizaciones externas para esta tarea, este es el caso de Mongolia, Figura 3.10, utiliza tanto el dashboard como la base de datos de la JHU.

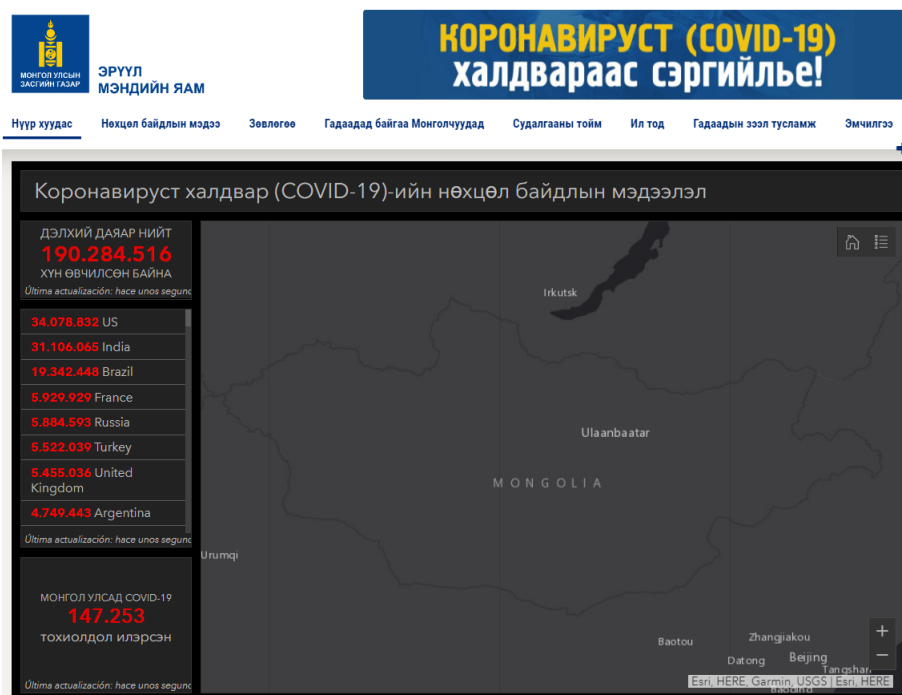


Figura 3.10: Datos ofrecidos por Mongolia

Capítulo 4

Selección de datos

Desde el inicio de la pandemia los gobiernos han ido recopilando gran cantidad de datos relacionados con los casos de coronavirus que se han generado en su territorio, para ello, cada gobierno ha tomado diferentes variables e indicadores de interés, siendo algunas de ellos los siguientes: casos activos, índice reproductivo, tasa de incidencia o número de casos positivos con fuente de infección conocida en los últimos quince días. Estas bases de datos son aportadas generalmente por el gobierno central, el ministerio de sanidad u otras organizaciones públicas como universidades o centros de investigación, siendo de total interés para concienciar a la población o para preparar medidas preventivas entre otras posibilidades.

Es por este motivo, que se ha tomado como referencia estas bases de datos como punto de partida.

Una vez agrupados todos los datos, se ha creado un sistema de puntuación para poder clasificar y puntuar a todos los países, con el objetivo de realizar un estudio con aquellos países que obtengan una mayor cantidad de puntos.

4.1. Datos oficiales

Para facilitar la comprensión de los datos se ha generado una tabla por cada agrupación territorial, tomando como variables de interés:

- Link: URL de acceso a la base de datos.
- Oficial: indica si son datos ofrecidos por el gobierno.
- Datos diarios / acumulados: refleja que tipo de datos se aportan.
- Recuperados: indica si posee datos sobre el número de recuperados.

- Transmisiones OMS: se utiliza la clasificación de transmisión oficial de la OMS.
- CSV: indica si permite descargar la base de datos.
- Colores:
 - Verde: datos recogidos por gobierno o ministerios.
 - Amarillo: datos recogidos por organizaciones oficiales o universidades.
 - Rojo: datos no disponibles.

4.2. Tabla Bases de datos por continente

En esta sección se van a mostrar las tablas obtenidas de todos los países Figura 4.1 pertenecientes a América Figura 4.2, Asia Figura 4.3 y Europa Figura 4.4.

Link Gobierno	Oficial	Datos diarios / Acumulados	Recuperados	Transmisiones OMS	CSV
BASE DATOS OFICIAL OMS					
https://covid19.who.int/	SI	Datos diarios / Acumulados	NO	ALL	SI
https://www.who.int/em	SI	Informes			
BASE DATOS CSSE JOHNS HOPKINS					
https://github.com/CSSE	SI	Datos diarios / Acumulados	SI	NO	SI
BASE DATOS WIKIPEDIA					
https://google.com/covic	SI	Datos Acumulados	SI	NO	NO

Figura 4.1: Bases de datos mundiales

Propagación Covid 19 Recolección, Tratamiento y Análisis de los datos

PAISES	Link Gobierno	Oficial	Datos diarios / Acumulados	Recuperados	Transmisiones OMS	CSV
NORTH AMERICA						
United States of America	https://covid.cdc.gov/cov	SI	Diarios / Acumulados	SI	Community transmission	SI
Canada	https://www.canada.ca/t	SI	Diarios / Acumulados	NO	Community transmission	SI
CENTRAL AMERICA						
Mexico	https://covid19.sinave.go	SI	Diarios / Acumulados	SI	Community transmission	SI
Guatemala	https://www.belize.gov.b	NO	-	-	Community transmission	NO
Belice	https://www.belize.gov.b	NO	-	-	Community transmission	NO
Honduras	https://covid19honduras.	SI	Acumulados	NO	Community transmission	NO
El Salvador	https://covid19.gob.sv/	SI	Diarios / Acumulados	SI	Community transmission	NO
Nicaragua	https://observatorioni.or	NO	Diarios / Acumulados	NO	Community transmission	NO
Costa Rica	https://geovision.uned.ac	SI	Diarios / Acumulados	SI	Community transmission	SI
Panama	http://minsa.gob.pa/corc	SI	Acumulados	NO	Community transmission	NO
Cuba	https://www.presidencia.	SI	Acumulados	NO	Community transmission	NO
SOUTH AMERICA						
Argentina	https://www.argentina.g	SI	Diarios / Acumulados	SI	Community transmission	SI
Bolivia	https://www.boliviasegur	SI	Diarios / Acumulados	SI	Community transmission	NO
Brazil	https://qsprod.saude.gov	SI	Diarios / Acumulados	NO	Community transmission	NO
Chile	https://www.gob.cl/coro	SI	Acumulados	SI	Community transmission	SI
Colombia	https://www.minsalud.g	SI	Diarios / Acumulados	SI	Community transmission	SI
Ecuador	https://www.salud.gob.e	SI	Acumulados	SI	Community transmission	NO
Guyana	https://dpi.gov.gy/	SI	Acumulados	NO	Community transmission	NO
Paraguay	https://www.mspbs.gov.i	SI	Diarios / Acumulados	SI	Community transmission	NO
Perú	https://www.gob.pe/corc	SI	Diarios / Acumulados	SI	Community transmission	NO
Suriname	http://www.gov.sr/	NO	-	-	Clusters of cases	NO
Uruguay	https://www.gub.uy/	SI	Diarios / Acumulados	SI	Community transmission	SI
Venezuela	https://presidenciave.con	SI	Acumulados	SI	Community transmission	NO

Color	Informacion
	Datos recogidos por el gobierno o ministerios.
	Datos recogidos por organizaciones o universidades.
	Datos no disponibles.

Figura 4.2: Bases de datos oficiales de América

Propagación Covid 19 Recolección, Tratamiento y Análisis de los datos

TABLA DE DATOS OFICIALES CONTINENTE ASIA

PAISES	Link Gobierno	Oficial	Datos diarios / Acumulados	Recuperados	Transmisiones OMS	CSV
ASIA SEPTENTRIONAL						
Rusia	https://xn--80aesfpeba	SI	Datos diarios (Mes) / Acumulados	SI	Clusters of cases	NO
ASIA CENTRAL						
Kazajistán	https://www.coronavir	SI	Datos Acumulados	SI	Clusters of cases	NO
Kirguistán	https://www.gov.kg/ky	-	-	-	Clusters of cases	-
	http://www.med.kg/ru/informatsii.html	-	-	-	-	-
Tayikistán	https://covid.tj/	SI	Datos diarios / Acumulados	SI	Clusters of cases	NO
Uzbekistán	https://coronavirus.uz	SI	Datos Acumulados / 24h	SI	Clusters of cases	NO
Turkmenistan	-	-	-	-	No cases	-
ASIA OCCIDENTAL						
Arabia Saudita	https://covid19.moh.gov.sa	SI	Datos diarios / Acumulados	SI	Sporadic Cases	NO
Armenia	https://ncdc.am/coron	SI	Datos diarios / Acumulados	SI	Community transmission	NO
Azerbaiyán	https://its.gov.az/page	SI	Datos Acumulados	SI	Clusters of cases	NO
	https://koronavirusinf	SI	-	-	-	PDF
Baréin	https://www.moh.gov.bh	SI	Datos Acumulados	SI	Clusters of cases	NO
Catar	https://covid19.moh.gov.qa	SI	Datos Acumulados / 24h	SI	Community transmission	NO
Chipre	https://covid19.ucy.ac.cy	SI	Datos Acumulados	SI	Clusters of cases	NO
Emiratos Árabes Unidos	https://covid19.ncema.ae	SI	Datos diarios / Acumulados	SI	Community transmission	NO
Georgia	https://stopcov.ge/en	SI	Datos Acumulados	SI	Community transmission	NO
Irak	https://coronavirus.iq/	SI	Datos Acumulados	SI	Community transmission	NO
Israel	https://datadashboard	SI	Datos Acumulados / 24h	SI	Community transmission	NO
Jordania	https://portal.jordan.gov.jo	-	-	-	Community transmission	-
Kuwait	https://corona.e.gov.kw	SI	Datos diarios / Acumulados	SI	Community transmission	NO
Libano	https://corona.ministry.gov.lb	SI	Datos Acumulados / 15 días	SI	Community transmission	NO
Omán	https://www.oman.or.gov.om	-	-	-	Community transmission	-
Siria	http://www.mol.gov.sy	-	-	-	Community transmission	-
Turquía	https://covid19.tubitak.gov.tr	SI	Datos diarios / Acumulados	SI	Community transmission	SI
	https://corona.cbddo.gov.tr	SI	Datos Acumulados	SI	SI	NO
	https://covid19.saglik.gov.tr	SI	Datos Acumulados	SI	SI	NO
Yemen	https://www.akhbar1.gov.ye	SI	Datos diarios / Acumulados	SI	Community transmission	NO
ASIA MERIDIONAL						
Afganistan	https://moph.gov.af/d	SI	-	-	Clusters of cases	-
Bangladés	https://www.gov.bd/	SI	Datos Acumulados / 24h	SI	Community transmission	NO
	https://corona.gov.bd/	SI	-	-	-	-
Bután	https://www.gov.bt/cc	SI	Datos Acumulados	SI	Sporadic Cases	NO
India	https://www.mygov.in	SI	Datos Acumulados / 24h	SI	Clusters of cases	NO
Irán	https://behdasht.gov.ir	SI	Informes	-	Community transmission	-
Maldivas	https://covid19.health.gov.mv	SI	Datos diarios / Acumulados	SI	Clusters of cases	NO
Nepal	https://covid19.moh.gov.np	SI	Datos Acumulados / 24h	SI	Clusters of cases	NO
	https://covid19.ndrmi.gov.np	SI	-	-	-	-
Pakistán	http://covid.gov.pk/	SI	Datos diarios / Acumulados	SI	Clusters of cases	NO
	http://covid.gov.pk/str	SI	-	-	-	-
Sri Lanka	https://covid19.gov.lk/	SI	Datos Acumulados	SI	Clusters of cases	NO
ASIA ORIENTAL						
China	http://weekly.chinacdc.org/	SI	Datos diarios	SI	Clusters of cases	NO
China	http://www.nhc.gov.cn/	SI	Datos diarios / Acumulados	SI	Clusters of cases	NO
Corea del Norte	-	-	-	-	No cases	-
Corea del sur	http://ncov.mohw.go.kr	SI	Datos Acumulados / Last 7d.	SI	Clusters of cases	NO
Japon	https://www.mhlw.go.jp/e/	SI	Datos diarios / Acumulados	SI	Clusters of cases	SI
	https://www.mhlw.go.jp/	SI	-	-	-	-
	https://www.mhlw.go.jp/e/	SI	Informes	-	-	-
	https://corona.go.jp/e/	SI	Datos Acumulados / semanales	SI	-	NO
Japon	https://covid19japan.cdc.gov	SI	Datos diarios / Acumulados	SI	Clusters of cases	NO
	https://github.com/reustle/covid19japan-data/	SI	-	-	-	SI
Mongolia	https://covid19.moh.gov.mn	SI	Datos Acumulados	S	Sporadic Cases	NO
Taiwan	https://www.cdc.gov/taiwan/	SI	Datos Acumulados	SI	Clusters of cases	NO
	https://sites.google.com/cdc.gov.tw/2019ncov/taiwan?authuser=0	SI	-	-	-	NO
SUDESTE ASIÁTICO						
Myanmar (Birmania)	-	-	-	-	Clusters of cases	-
Brunei	https://www.healthinf.gov.bn	SI	Datos Acumulados	SI	Sporadic Cases	NO
Camboya	http://www.cdcmoh.gov.kh/	SI	Datos diarios / Acumulados	SI	Sporadic Cases	NO
	https://covid19-map.cdcmoh.gov.kh/	SI	-	-	-	-
Filipinas	https://www.gov.ph/	SI	-	-	Community transmission	-
	https://www.facebook.com/OfficialDOHgov/	SI	-	-	-	-
Indonesia	http://covid19.bnpb.go.id	SI	Datos diarios / Acumulados	SI	Community transmission	NO
	https://covid19.kemke.go.id	SI	-	-	-	-
Laos	https://www.covid19.gov.la/	SI	Datos Acumulados	SI	Sporadic Cases	NO
	https://moh.gov.la/	SI	-	-	-	-
Malasai	http://covid-19.moh.gov.my	SI	Datos diarios / Acumulados	SI	Clusters of cases	NO
Singapur	https://www.moh.gov.sg/	SI	Informes	SI	Community transmission	NO
Tailandia	https://ddc.moph.go.th/	SI	Datos diarios / Acumulados	NO	Clusters of cases	NO
Timor Oriental	https://covid19.gov.tl/	SI	-	-	Sporadic Cases	-
Vietnam	https://ncov.moh.gov.vn/	SI	Datos diarios / Acumulados	SI	Clusters of cases	NO

Leyenda:

color	Informacion
	Datos recogidos por gobiernos o ministerios
	Datos recogidos por organizaciones o universidades
	Datos no disponibles

Figura 4.3: Bases de datos oficiales de Asia

Propagación Covid 19 Recolección, Tratamiento y Análisis de los datos

TABLA DE DATOS OFICIALES CONTINENTE EUROPA						
PAISES	Link gobierno	Oficial	Datos diarios/Acumulados	Recuperados	Transmisiones OMS	CSV
EUROPA MERIDIONAL						
Albania	https://shendetesia	SI	Datos diarios / Acumulados	SI	Cluster of cases	NO
Andorra	https://www.govern	SI	Datos semanales / Acumulados	SI	Community transmission	NO
Bosnia y Herzegovina	http://www.fbihvlj	SI	Acumulados	SI	Community transmission	NO
Ciudad del Vaticano	http://www.vatican	-	-	-	Sporadic cases	-
Croacia	https://www.koron	SI	Datos diarios / Acumulados	SI	Community transmission	NO
Eslovenia	https://www.gov.si/	SI	Datos diarios / Acumulados	SI	Cluster of cases	SI
España	https://www.mscbs	SI	Datos diarios / Acumulados	SI	Community transmission	SI
Grecia	https://eody.gov.gr/	SI	Datos diarios / Acumulados	SI	Community transmission	NO
Italia	http://www.salute.g	SI	Datos diarios / Acumulados	SI	Cluster of cases	SI
Macedonia del Norte	https://koronavirus	SI	Datos diarios / Acumulados	SI	Cluster of cases	NO
Malta	https://www.gov.mt	SI	Datos diarios / Acumulados	SI	Cluster of cases	NO
Montenegro	https://www.gov.me	SI	Acumulados	SI	Cluster of cases	NO
Portugal	https://covid19esta	SI	Datos diarios / Acumulados	SI	Cluster of cases	NO
San Marino	-	SI	-	-	Community transmission	-
Serbia	https://www.srbija.j	SI	Datos diarios / Acumulados	NO	Community transmission	NO
EUROPA SEPTENTRIONAL						
Dinamarca	https://www.sst.dk/	SI	Datos diarios / Acumulados	SI	Community transmission	NO
Estonia	https://www.tervise	SI	Datos diarios / Acumulados	SI	Cluster of cases	NO
Finlandia	https://thl.fi/en/web	SI	Datos diarios / Acumulados	NO	Community transmission	NO
Irlanda	https://www.gov.ie/	SI	Datos diarios / Acumulados	SI	Community transmission	NO
Islandia	https://www.covid.i	SI	Datos diarios / Acumulados	SI	Community transmission	NO
Letonia	https://www.rfa.gc	SI	Datos diarios / Acumulados	SI	Community transmission	NO
Noruega	https://www.regjeri	SI	Datos diarios / Acumulados	SI	Cluster of cases	NO
Reino Unido	https://www.gov.uk	SI	Datos diarios / Acumulados	SI	Community transmission	NO
Suecia	https://www.goverr	SI	Datos diarios / Acumulados	SI	Community transmission	NO
EUROPA CENTRAL						
Alemania	https://www.rki.de/	SI	Datos diarios / Acumulados	SI	Community transmission	SI
Austria	https://www.sozialr	SI	Datos diarios / Acumulados	SI	Community transmission	SI
Eslovaquia	https://www.mzv.sk	SI	Datos diarios / Acumulados	SI	Cluster of cases	NO
Hungría	http://abouthungar	SI	Datos diarios / Acumulados	SI	Community transmission	NO
Liechtenstein	https://www.liechte	SI	Datos diarios / Acumulados	SI	Sporadic cases	NO
Lituania	https://koronastop.l	SI	Datos diarios / Acumulados	SI	Community transmission	NO
Luxemburgo	https://covid19.pub	SI	Datos diarios / Acumulados	SI	Community transmission	SI
Polonia	https://www.gov.pl/	SI	Datos diarios / Acumulados	SI	Community transmission	NO
República Checa	https://www.mvcr.c	SI	Datos diarios / Acumulados	SI	Community transmission	NO
Rumania	https://www.gov.ro	SI	Datos diarios	SI	Community transmission	NO
Suiza	https://www.swissir	SI	Datos diarios	SI	Community transmission	NO
EUROPA OCCIDENTAL						
Belgica	https://epistat.wiv-i	SI	Datos diarios / Acumulados	SI	Community transmission	SI
Francia	https://www.gouver	SI	Datos diarios / Acumulados	SI	Community transmission	SI
Monaco	https://en.gouv.mc/	SI	Datos diarios / Acumulados	SI	Sporadic cases	NO
Países Bajos	https://www.rivm.n	SI	Datos diarios / Acumulados	SI	Community transmission	SI
EUROPA ORIENTAL						
Armenia	https://www.gov.am	SI	Acumulados	SI	Community transmission	NO
Azerbaiyan	https://nk.gov.az/	SI	Datos diarios / Acumulados	SI	Cluster of cases	NO
Bielorrusia	https://www.belaru	SI	Acumulados	SI	Community transmission	NO
Bulgaria	https://www.gov.bg	SI	Acumulados	SI	Cluster of cases	NO
Georgia	https://dph.georgia	SI	Datos diarios / Acumulados	SI	Community transmission	NO
Moldavia	https://statistica.go	SI	Datos diarios / Acumulados	SI	Community transmission	NO
Ucrania	https://covid19.gov	SI	Datos diarios / Acumulados	SI	Community transmission	NO

Color	Información
	Datos recogidos por el gobierno o ministerios
	Datos recogidos por organizaciones o universidades
	Datos no disponibles

Figura 4.4: Bases de datos oficiales de Europa

4.3. Puntuación

Los países han sido clasificados atendiendo a un sistema de puntuaciones creado por el equipo, se han tenido en cuenta diferentes variables para destacar unos países frente a otros. Una vez calculadas las puntuaciones se ha recalcado de color verde, los países que han obtenido mayor puntuación y en amarillo los países que les siguen. Con el objetivo de centrarse en realizar un estudio conjunto de los países con mayor calificación obtenida.

Para ello, se han puntuado los siguientes casos:

- Datos oficiales nacionales.
- Universidad Johns Hopkins (UJH).
- Organización Mundial de la Salud (OMS).
- Número de infectados.

4.3.1. Datos oficiales nacionales

Se han puntuado los datos que ofrecen los gobiernos valorando el nivel de detalle que estos ofrecen. Se han ido puntuando de forma creciente los datos según la frecuencia, siendo estos: datos acumulados, datos diarios, división de los datos por regiones, gráficas y si facilitan el acceso a sus datos, es decir, si permiten descargar los datos en CSV Figura 4.5.

4.3.2. Universidad Johns Hopkins

Se ha puntuado dos variables, siendo la primera si el país se encuentra dentro de la base de datos de la UJH, y la segunda, si a la vez está dividida en regiones Figura 4.5.

4.3.3. Organización Mundial de la Salud

Se ha tenido en cuenta la existencia de datos de dicho país dentro de la base de datos de la OMS Figura 4.5.

4.3.4. Número de infectados

Esta variable es la más relevante para nuestro estudio, ya que, nos interesa estudiar las distintas olas sufridas en cada país, es por ello, que se han puntuado con mayor número los países más infectados Figura 4.6.

Propagación Covid 19 Recolección, Tratamiento y Análisis de los datos

GOVERNMENT/MINISTRY OF HEALTH DATA			CSSE JHU	
DATA		POINTS	DATA	POINTS
	Cumulative	1	Country	3
	Daily	2	Region	2
	Regional	3		
	Graphics	4	WHO	
	CSV	5	DATA	POINTS
				3

* = COVID-19 (SARS-CoV-2 Infection), Others
* = CSSE JHU

Ranking 1
Ranking 2

Figura 4.5: Puntuación empleada para la clasificación de los países.

INFECTED AMERICA		INFECTED ASIA	
INFECTED CASES	POINTS	INFECTED CASES	POINTS
< 1.000.000	3	< 100.000	3
1.000.000-3.000.000	5	100.000-300.000	5
> 3.000.000	10	300.000-1.000.000	7
		> 1.000.000	10
		> 10.000.000	12

INFECTED EUROPE	
INFECTED CASES	POINTS
< 1.000.000	3
1.000.000-3.000.000	5
> 3.000.000	10

Figura 4.6: Distribución de puntos respecto al número de infectados.

4.4. América

Este continente está dividido en tres regiones:

4.4.1. América del Norte

En el caso de América del Norte tanto Canadá, como Estados Unidos de América obtienen la máxima puntuación en las secciones de los datos gubernamentales. Diferenciándose en el número de infectados Figura 4.7.

AREA	COUNTRY	GOVERNMENT/MINISTRY OF HEALTH DATA	CSSE JHU	WHO	INFECTED	TOTAL
NORTH AMERICA						
	United States	14*	5	3	10	32
	Canada	14*	5	3	5	27

Figura 4.7: Clasificación América del norte

4.4.2. América Central

En América central destaca México que obtiene una puntuación total de 27 puntos y siendo el país con mayor puntuación de la región Figura

Propagación Covid 19
Recolección, Tratamiento y Análisis de los datos

4.8.

AREA	COUNTRY	GOVERNMENT/MINISTRY OF HEALTH DATA	CSSE JHU	WHO	INFECTED	TOTAL
CENTRAL AMERICA						
	Mexico	14*	5	3	5	27
	Guatemala	0	3	3	3	9
	Belice	0	3	3	3	9
	Honduras	4	3	3	3	13
	El Salvador	6	3	3	3	15
	Nicaragua	10*	3	3	3	19
	Costa Rica	14*	3	3	3	23
	Panama	6	3	3	3	15
	Cuba	1	3	3	3	10

Figura 4.8: Clasificación América Central

4.4.3. América del Sur

En América del Sur destacan tres países, siendo de mayor a menor puntuación: Colombia, Argentina y Brasil Figura 4.9.

AREA	COUNTRY	GOVERNMENT/MINISTRY OF HEALTH DATA	CSSE JHU	WHO	INFECTED	TOTAL
SOUTH AMERICA						
	Argentina	14*	3	3	10	30
	Bolivia	6	3	3	3	15
	Brazil	10*	5	3	10	28
	Chile	14*	5	3	5	27
	Colombia	14*	5	3	10	32
	Ecuador	8*	3	3	3	17
	Guyana	3	3	3	3	12
	Paraguay	10	3	3	3	19
	Perú	10*	5	3	5	23
	Suriname	0	3	3	3	9
	Uruguay	14*	3	3	3	23
	Venezuela	1	3	3	3	10

Figura 4.9: Clasificación América del Sur

4.5. Asia

Este continente se divide en seis regiones:

4.5.1. Asia Septentrional

En esta región tan solo se encuentra Rusia, país que ha obtenido 28 puntos Figura 4.10.

REGION	COUNTRY	GOVERNMENT/MINISTRY OF HEALTH DATA	CSSE JHU	WHO	INFECTED	TOTAL
NORTH ASIA						
	Russia		10	5	3	10
						28

Figura 4.10: Clasificación Asia Septentrional

4.5.2. Asia Central

En Asia central destaca Kazajistán como el país que obtiene la mayor puntuación con 18 puntos, muy cerca de él se encuentra Tayikistán con 16 Figura 4.11.

REGION	COUNTRY	GOVERNMENT/MINISTRY OF HEALTH DATA	CSSE JHU	WHO	INFECTED	TOTAL	
CENTRAL ASIA							
	kazakhstan		5	3	3	7	18
	Kyrgyzstan		0	3	3	3	9
	Tajikistan		7	3	3	3	16
	Uzbekistan		5	3	3	3	14
	Turkmenistan		0	0	0	0	0

Figura 4.11: Clasificación Asia Central

4.5.3. Asia Occidental

En Asia Occidental, Turquía es el país que más puntuación obtiene con 30 puntos, después se encuentra Israel que obtiene 22 Figura 4.12.

REGION	COUNTRY	GOVERNMENT/MINISTRY OF HEALTH DATA	CSSE JHU	WHO	INFECTED	TOTAL	
WESTERN ASIA							
	Saudi Arabia		7*	3	3	7	19
	Armenia		10	3	3	5	21
	Azerbaijan		7	3	3	7	20
	Bahrain		1	3	3	5	12
	Qatar		3	3	3	5	14
	Cyprus		10*	3	3	3	22
	United Arab Emirates		5	3	3	7	18
	Georgia		3	3	3	7	16
	Iraq		1	3	3	10	17
	Israel		7	5	3	7	22
	Jordan		0	3	3	7	13
	Kuwait		7	3	3	5	18
	Lebanon		10	3	3	7	23
	Oman		0	3	3	5	11
	Syria		0	3	3	3	9
	Turkey		14	3	3	10	30
	Yemen		3	3	3	3	12

Figura 4.12: Clasificación Asia Occidental

4.5.4. Asia Meridional

En esta región, destaca India como el país con mayor puntuación, 26 puntos, seguida muy de cerca por Pakistán que obtiene 25 puntos y Nepal con 23 Figura 4.13.

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

REGION	COUNTRY	GOVERNMENT/MINISTRY OF HEALTH DATA	CSSE JHU	WHO	INFECTED	TOTAL
SOUTH ASIA						
	Afghanistan	0	3	3	3	9
	Bangladesh	7	3	3	7	20
	Bhutan	3	3	3	3	12
	India	6	5	3	12	26
	Iran	0	3	3	10	16
	Maldives	3	3	3	3	12
	Nepal	10	3	3	7	23
	Pakistán	10	5	3	7	25
	Sri Lanka	6	3	3	3	15

Figura 4.13: Clasificación Asia Meridional

4.5.5. Asia Oriental

En Asia Oriental destaca Japón, que obtiene 29 puntos siendo uno de los países con mayor puntuación dentro del continente Figura 4.14.

REGION	COUNTRY	GOVERNMENT/MINISTRY OF HEALTH DATA	CSSE JHU	WHO	INFECTED	TOTAL
NORTHEAST ASIA						
	China	5	5	3	5	18
	North Korea	0	0	0	0	0
	South Korea	10	3	3	5	21
	Japan	14	5	3	7	29
	Mongolia	1*	3	3	3	10
	Taiwan	7*	3	0	3	13

Figura 4.14: Clasificación Asia Oriental

4.5.6. Sudeste Asiático

En esta región destaca Indonesia, con 26 puntos, seguido de cerca por Malasia con 23 y Camboya con 22 Figura 4.15.

REGION	COUNTRY	GOVERNMENT/MINISTRY OF HEALTH DATA	CSSE JHU	WHO	INFECTED	TOTAL
SOUTHEAST ASIA						
	Myanmar (Birmania)	0	0	3	3	6
	Brunei	5	3	3	3	14
	Cambodia	10**	3	3	3	22
	Philippines	0	3	3	10	16
	Indonesia	10	3	3	10	26
	Laos	3	3	3	3	12
	Malaysia	10	3	3	7	23
	Singapore	3	3	3	3	12
	Thailand	3	3	3	3	12
	East Timor	0	3	3	3	9
	Vietnam	10	3	3	3	19

Figura 4.15: Clasificación Sudeste Asiático

4.6. Europa

Este continente se divide en cinco regiones:

4.6.1. Europa Meridional

En esta región destacan dos países. Siendo ellas, España e Italia con 32 puntos Figura 4.16.

EUROPE AREA	COUNTRY	GOVERNMENT/MINISTRY OF HEALTH DA	CSSE JHU	WHO	INFECTED	TOTAL
EUROPA MERIDIONAL						
	Albania		3	3	3	12
	Andorra		5	3	3	14
	Bosnia y Herzegovina		8	3	3	17
	Ciudad del Vaticano		0	3	3	9
	Croacia		10	3	3	19
	Eslovenia		13*	3	3	22
	España		14*	5	3	32
	Grecia		0	3	3	9
	Italia		14*	5	3	32
	Macedonia del Norte		3	3	3	12
	Malta		0	3	3	9
	Montenegro		0	3	3	9
	Portugal		10*	3	3	19
	San Marino		0	3	3	9
	Serbia		3	3	3	12

Figura 4.16: Clasificación Europa Meridional

4.6.2. Europa Septentrional

En esta región destacan Reino Unido y Dinamarca, obteniendo 25 y 21 puntos respectivamente Figura 4.17.

EUROPE AREA	COUNTRY	GOVERNMENT/MINISTRY OF HEALTH DA	CSSE JHU	WHO	INFECTED	TOTAL
EUROPA SEPTENTRIONAL						
	Dinamarca		10*	5	3	21
	Estonia		10	3	3	19
	Finlandia		0	3	3	9
	Irlanda		3	3	3	12
	Islandia		3	3	3	12
	Letonia		0	3	3	9
	Noruega		0	3	3	9
	Reino Unido		7*	5	3	25
	Suecia		6	5	3	19

Figura 4.17: Clasificación Europa Septentrional

4.6.3. Europa Central

En Europa Central destaca Alemania con 32 puntos obteniendo la máxima puntuación, seguido por Austria que obtiene 23 Figura 4.18.

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

EUROPE AREA	COUNTRY	GOVERNMENT/MINISTRY OF HEALTH DA	CSSE JHU	WHO	INFECTED	TOTAL
EUROPA CENTRAL						
	Alemania		14*	5	3	32
	Austria		14*	3	3	23
	Eslovaquia		0	3	3	9
	Hungría		3	3	3	12
	Liechtenstein		0	3	3	9
	Lituania		3	3	3	12
	Luxemburgo		14*	3	3	23
	Polonia		0	3	3	11
	Republica Checa		0	3	3	11
	Rumania		0	3	3	11
	Suiza		10	3	3	19

Figura 4.18: Clasificación Europa Central

4.6.4. Europa Occidental

En esta región destaca Francia que obtiene 32 puntos, seguido de cerca por Bélgica y Países Bajos, ambos con 27 Figura 4.19.

EUROPE AREA	COUNTRY	GOVERNMENT/MINISTRY OF HEALTH DA	CSSE JHU	WHO	INFECTED	TOTAL
EUROPA OCCIDENTAL						
	Belgica		14*	5	3	27
	Francia		14*	5	3	32
	Monaco		3	3	3	12
	Países Bajos		14*	5	3	27

Figura 4.19: Clasificación Europa Occidental

4.6.5. Europa Oriental

En esta región destacan Ucrania y Geórgica como los países que han obtenido una mayor puntuación siendo está 16 puntos Figura 4.20.

EUROPE AREA	COUNTRY	GOVERNMENT/MINISTRY OF HEALTH DA	CSSE JHU	WHO	INFECTED	TOTAL
EUROPA ORIENTAL						
	Armenia		0	3	3	9
	Azerbaiyan		3	3	3	12
	Bielorrusia		0	3	3	9
	Bulgaria		1	3	3	10
	Georgia		7	3	3	16
	Moldavia		0	3	3	9
	Ucrania		3*	5	3	16

Figura 4.20: Clasificación Europa Oriental

4.7. Países Seleccionados

A partir de estas puntuaciones, se han obtenido los siguientes países como los más puntuados de cada región:

- América: Estados Unidos, México, Argentina, Brasil y Colombia.

Propagación Covid 19
Recolección, Tratamiento y Análisis de los datos

- Asia: Rusia, Kazajistán, Turquía, India, Japón e Indonesia.
- Europa: España, Italia, Alemania, Francia y Países Bajos.

De los cuales, hemos seleccionado tres países de cada región, para posteriores estudios, en este caso hemos seleccionado los siguientes países: Alemania, Brasil, España, Estado Unidos, India, Italia, Japón, México y Rusia.

Capítulo 5

Resultados selección de datos

Tras la recopilación de datos y la selección de estos mismos, se ha creado un repositorio público en GitHub, donde se encuentra todo lo necesario para poder generar el dataset de este proyecto. A continuación, se explica más en profundidad la estructura del repositorio y el dataset final. Además, de un análisis más en profundidad de las puntuaciones obtenidas por los países tras nuestra clasificación.

5.1. GitHub

El resultado final se puede encontrar en GitHub [19], este repositorio contiene todo lo necesario para generar una base de datos completa, así como, ejemplos e información adicional para ayudar a generarla. Como se puede observar en la Figura 5.1, el repositorio está formado por cinco directorios, para ofrecer una experiencia más intuitiva.

1. Deploy: Para desplegar el entorno, contiene tutoriales y scripts.
2. World whitout USA: Los datos de todo el mundo, agrupados por país y región, a excepción de USA.
3. USA-State: Los datos de USA, agrupados por estados.
4. CountryFilter/RegionOptional: Filtro para generar los datos de un país o la región de un país en concreto.
5. R: Formado por scripts y funciones para calcular el algoritmo K-Means, gráficas y variables adicionales.

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

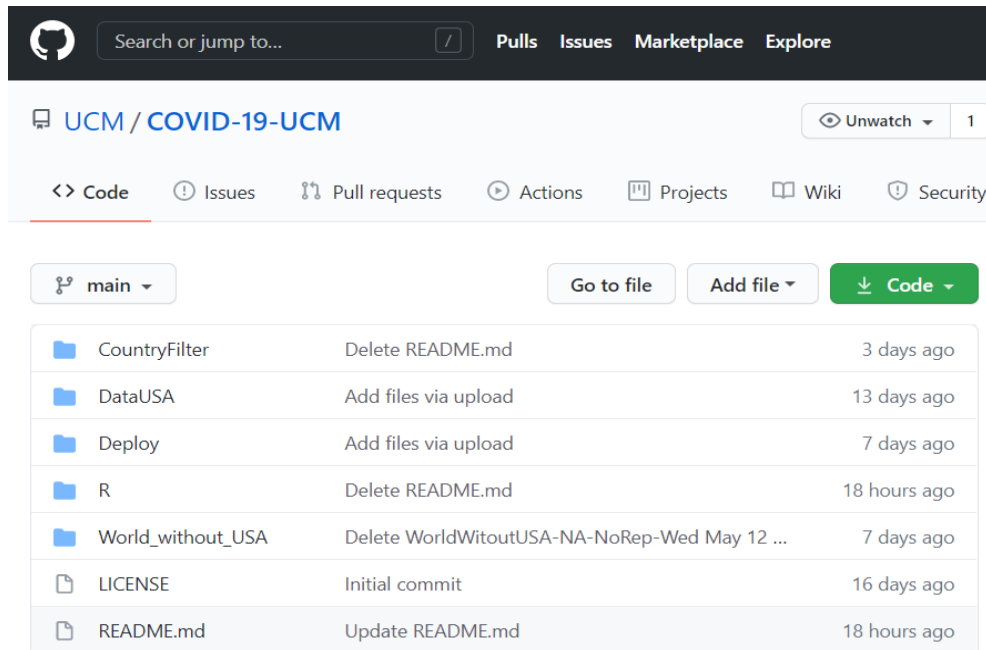


Figura 5.1: GitHub

5.1.1. GitHub Deploy

En esta carpeta encontramos lo necesario para preparar y desplegar el entorno necesario para poder usar el resto de las herramientas facilitadas en el repositorio. Por un lado, tenemos las guías de instalación tanto de Spark y las bibliotecas necesarias en entorno Windows, como la guía para instalar una máquina virtual con un entorno Unix con lo necesario. Aparte se encuentran dos archivos con la secuencia de comandos necesarios para crear los directorios, y descargar en los mismos los archivos del repositorio de la JHU y su corrección de formato en aquellos archivos que se detectaron con errores Figura 5.2.

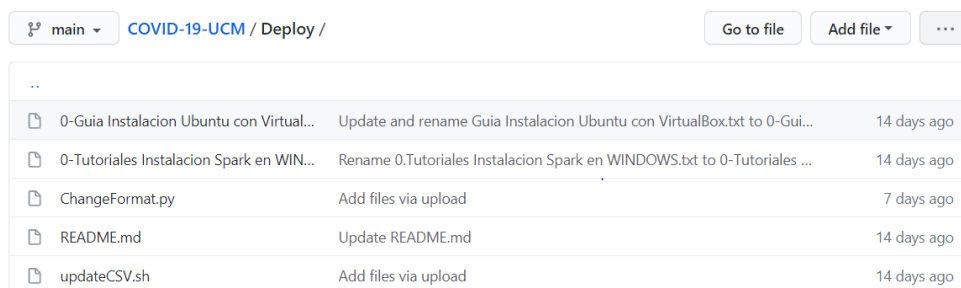


Figura 5.2: GitHub Deploy

5.1.2. GitHub Counter-Filter, USA-State y World whitout USA

En estos tres directorios encontramos una estructura idéntica, con una carpeta EXAMPLES donde se encuentran los CSV generados mediante las diferentes versiones de los scripts, y otra carpeta SCRIPTS que contiene las diferentes iteraciones y aproximaciones que se han ido desarrollando para generar cada uno de los conjuntos de datos. Todos contienen cinco archivos con la secuencia de comandos que generan los datos corregidos, los datos con resultados diarios negativos con y sin fechas repetidas y por último los que generan un resultado donde los valores negativos en las columnas diarias se sustituyen por valores nulos con fechas repetidas y sin estas Figura 5.3.

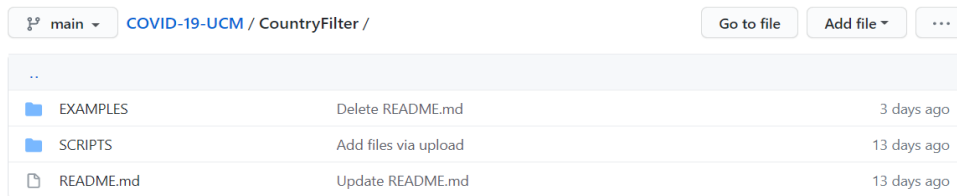


Figura 5.3: GitHub Counter-Filter, USA-State y World whitout USA

5.1.3. R

Como se puede observar en la Figura 5.4, este directorio está a su vez formado por tres subdirectorios. Siguiendo el orden de la imagen, en la carpeta FUNCTIONS, se pueden encontrar diferentes funciones para ampliar la cantidad de variables de un dataset, como calcular la media móvil de las variables daily Figura 5.5. Además de una carpeta, con ejemplos para mostrar los resultados.

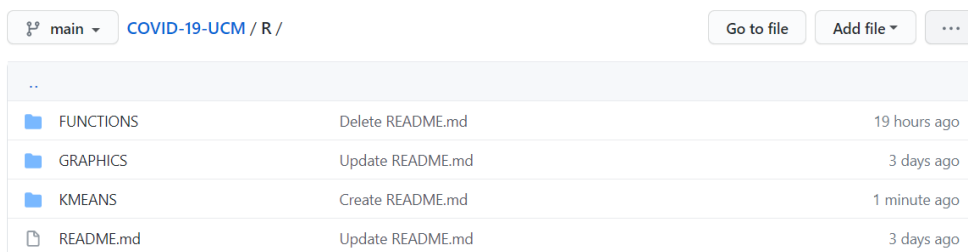


Figura 5.4: GitHub R

En la carpeta GRAPHICS, se encuentran los scripts que generan las gráficas de este proyecto. Este directorio también posee una carpeta con

Propagación Covid 19
Recolección, Tratamiento y Análisis de los datos

R0	Daily_ConfirmedMM	Daily_MuertosMM	Daily_RecuperadosMM	Daily_ActivosMM	IPMM	R0MM
0.63122172	587.57143	28.8571429	484.28571	74.428571	132.1428571	0.7660644
0.74516908	587.00000	30.0000000	516.71429	40.285714	100.2857143	0.7918674
0.78526841	615.28571	29.8571429	529.28571	56.142857	115.8571429	0.8537166
0.95956454	650.71429	30.4285714	534.28571	86.000000	146.8571429	0.9314928
1.47021277	677.57143	31.0000000	548.28571	98.285714	160.2857143	1.0100085
1.20103986	699.85714	30.7142857	551.42857	117.714286	179.1428571	1.0643059
1.09170306	716.57143	31.4285714	568.85714	116.285714	179.1428571	1.1072848
1.08115942	703.14286	31.0000000	571.28571	100.857143	162.8571429	1.0655986
1.05170068	646.00000	30.7142857	575.00000	40.285714	101.7142857	0.9335260
1.02472527	598.14286	30.5714286	551.14286	16.428571	77.5714286	0.8792524

Figura 5.5: GitHub R FUNTIONS

ejemplos, para que el usuario visualice las gráficas que se puede realizar. Como dato adicional, se adjunta un anexo de gráficas, para poder visualizar las gráficas generadas de los países más puntuados en la sección dedicada al estado del arte. Por último, el directorio KMEANS, contiene las tablas generadas tanto de España, como del mundo R0, IP e IA. Además, se ha adjuntado los K-Means generados como ejemplo.

5.2. Dataset Generado

Gracias al GitHub aportado, se puede generar un dataset que contiene las siguientes variables e indicadores, diferenciando con un asterisco, variables solo aportadas en el caso de Estados Unidos.

1. **Country_Region**: Nombre que identifica el país.
2. **Province_State**: Identificador de la provincia o estado del país.
3. **Lat**: Latitud correspondiente a el país o provincia al que pertenecen los datos.
4. **Long_***: Longitud correspondiente a el país o provincia al que pertenecen los datos.
5. **Last_Update1**: Fecha en la que se insertaron los datos, normalmente corresponde a los datos del día anterior (por ejemplo, un valor de 21/02/2021 corresponde a los datos del día 20/02/2021).
6. **Confirmed**: Número de confirmados acumulados hasta ese día.
7. **Deaths**: Número de muertos acumulados hasta ese día.
8. **Recovered**: Número de personas que han superado la enfermedad hasta esa fecha.

9. **Active:** Número de casos activos, que no son ni recuperados ni han fallecido, hasta la fecha.
10. **People _ Tested:** Personas a las que se les ha practicado las pruebas para ver si están infectadas.
11. **People _ Hospitalized:** Personas que están hospitalizadas a causa del virus.
12. **Daily _ Confirmed:** Casos confirmados diarios para ese día.
13. **Daily _ Deaths:** Muertes diarias para ese día.
14. **Daily _ Recovered:** Casos recuperados diarios para ese día.
15. **Daily _ Active:** Casos activos para ese día.
16. **Daily _ Tested:** Personas a las que se les ha realizado una prueba ese día.
17. **Daily _ Hospitalized:** Hospitalizados diarios, solo en EEUU.
18. **IP:** Índice de peligrosidad, se calcula con los datos diarios.
19. **R0:** Número de reproducción o factor de transmisión R0, también se calcula con los datos diarios, describe la intensidad de una enfermedad infecciosa.
20. **Daily _ ConfirmedMM:** Media móvil (a siete días) para los datos diarios de confirmados.
21. **Daily _ MuertosMM:** Media móvil (a siete días) para los datos diarios de muertos.
22. **Daily _ RecuperadosMM:** Media móvil (a siete días) para los datos de recuperados diarios.
23. **Daily _ ActivosMM:** Media móvil (a siete días) para los datos de activos.
24. **IPMM:** Índice de peligrosidad que usa los nuevos datos suavizados con la media móvil.
25. **R0MM:** Número de reproducción o factor de transmisión R0 que utiliza los nuevos datos suavizados con la media móvil.

Cabe destacar, que este dataset ha sido utilizado por otro equipo, en su propio trabajo de fin de grado "Desarrollo de un dashboard para la evolución de la Covid19"[20] en esta misma universidad (Universidad Complutense de Madrid), con el objetivo de generar un dashboard.

5.3. Análisis de datos oficiales nacionales

A continuación, se expone un análisis más detallado y con datos actualizados a mayo de 2021 tras la clasificación realizada anteriormente sobre las regiones de América, Asia y Europa.

5.3.1. América

Es el segundo continente más grande del mundo, tras Asia, este continente tiene una población de 970 millones de habitantes [21], a su vez este continente está dividido en tres regiones distintas, dependiendo de su localización, estas son: América del norte, América Central y América del Sur. En este caso, se ha tomado como referencia de división el geoesquema estándar M49 de la ONU [22], por lo tanto, se ha incluido México en América central.

América del Norte

Formado por Canadá y Estados Unidos, en esta zona de América destaca enormemente Estados Unidos, con más de 30 millones de contagios, y más de 550.000 fallecidos, estos datos son facilitados por el Centro para el Control y Prevención de Enfermedades (CDC) [23].

Por otro lado, está Canadá con más de un millón de casos y más de 20 mil muertos, ha ido aportando tanto datos totales como diarios, además de permitir exportar la información en CSV [24].

Como se puede observar en la Figura 5.6 estos países han tenido datos muy fiables, actualizados diariamente y de fuentes oficiales, pero la cantidad de contagios y muertes de Estados Unidos lo sitúa muy por encima en la clasificación, debido a que aproximadamente el 20 % de los contagios y muertes a nivel mundial a mayo de 2021 han sido en Estados Unidos. [17].

AREA	COUNTRY	GOVERNMENT/MINISTRY OF HEALTH DATA	CSSE JHU	WHO	INFECTED	TOTAL	
NORTH AMERICA							
	United States		14*	5	3	10	32
	Canada		14*	5	3	5	27

Figura 5.6: Clasificación América del norte

América Central

En la Figura 5.7 destaca México, con más de 3 millones de contagios, y acumulando la suma de 70 mil muertos. Además, ha aportado reportes diarios de datos con una información bastante completa sobre la evolución de la pandemia.[25].

Propagación Covid 19
Recolección, Tratamiento y Análisis de los datos

En el resto de Centroamérica no se encuentran países que destaquen por el número de infectados, ya que, ninguno supera el millón de contagios. Además, en esta zona, pocos países aportan datos oficiales sobre la pandemia y, los que aportan, en general han ofrecido muy poca información. En contraparte, se han encontrado países que ofrecen gran cantidad de datos, es el caso por ejemplo de Nicaragua, el cual aporta tanto datos como gráficas, y también es el caso de Costa Rica, que además permite descargar los datos en formato CSV [26].

AREA	COUNTRY	GOVERNMENT/MINISTRY OF HEALTH DATA	CSSE JHU	WHO	INFECTED	TOTAL
CENTRAL AMERICA						
	Mexico		14*	5	3	5
	Guatemala		0	3	3	3
	Belice		0	3	3	3
	Honduras		4	3	3	3
	El Salvador		6	3	3	3
	Nicaragua		10*	3	3	3
	Costa Rica		14*	3	3	3
	Panama		6	3	3	3
	Cuba		1	3	3	3

Figura 5.7: Clasificación América Central

América del Sur

En esta Figura 5.8 destacan tres países por su elevado número de confirmados, con más de 3 millones de contagios en caso de Argentina [27] y Colombia [28], y superando los 15 millones en caso de Brasil [29]. La causa de esta gran diferencia en el número de contagios recae sobre la variante brasileña P1 del Covid-19, la cual es mucho más contagiosa que la cepa original [30]. Tanta es la diferencia en esta región, que ningún otro país supera los tres millones de contagios, y solo otros dos, superan el millón a mayo de 2021.

AREA	COUNTRY	GOVERNMENT/MINISTRY OF HEALTH DATA	CSSE JHU	WHO	INFECTED	TOTAL
SOUTH AMERICA						
	Argentina		14*	3	3	10
	Bolivia		6	3	3	3
	Brazil		10*	5	3	10
	Chile		14*	5	3	5
	Colombia		14*	5	3	10
	Ecuador		8*	3	3	3
	Guyana		3	3	3	3
	Paraguay		10	3	3	3
	Perú		10*	5	3	5
	Suriname		0	3	3	3
	Uruguay		14*	3	3	3
	Venezuela		1	3	3	3

Figura 5.8: Clasificación América del Sur

5.3.2. Asia

Este continente está representado como el más grande del mundo, es por ello, que la Organización de las Naciones Unidas (ONU) [31], divide el continente en seis regiones según la situación geográfica, por lo cual, fue bastante útil particionar toda la información de Asia en estas seis regiones (Asia Septentrional, Asia Central, Asia Occidental, Asia Meridional, Asia Oriental y Sudeste Asiático).

Asia Septentrional

Como se observa en la Figura 5.9, el gobierno de Rusia [32] ha mostrado gran cantidad de datos, permitiendo visualizar en histogramas tanto los datos acumulativos como los diarios, en concreto muestra los datos de: infectados, recuperados, muertos e identificados. Además de incluir, un mapa donde se puede ver los datos acumulados de cada región.

REGION	COUNTRY	GOVERNMENT/MINISTRY OF HEALTH DATA	CSSE JHU	WHO	INFECTED	TOTAL	
NORTH ASIA	Russia		10	5	3	10	28

Figura 5.9: Clasificación Asia Septentrional

Asia Central

En esta región (Figura 5.10) hay dos países que marcan la diferencia, siendo Kazajistán el país mejor puntuado por el número de infectados, teniendo cerca de 400 mil casos [33]. Tayikistán como se muestra en la clasificación 5.10, es el país con mayor puntuación en datos ofrecidos por el gobierno [34], pero en contra, no presenta el mayor número de infectados, con solo 17 mil casos a mayo del 2021, por esto se encuentra como segundo en la clasificación.

REGION	COUNTRY	GOVERNMENT/MINISTRY OF HEALTH DATA	CSSE JHU	WHO	INFECTED	TOTAL	
CENTRAL ASIA	kazakhstan		5	3	3	7	18
	Kyrgyzstan		0	3	3	3	9
	Tajikistan		7	3	3	3	16
	Uzbekistan		5	3	3	3	14
	Turkmenistan		0	0	0	0	0

Figura 5.10: Clasificación Asia Central

Asia Occidental

Esta región de Asia es la que está formada por más países, siendo Irak y Turquía los países con mayor número de contagios como se puede

Propagación Covid 19 Recolección, Tratamiento y Análisis de los datos

observar en la Figura 5.11. En el caso de Irak, el gobierno [35] apenas transmite información. En cambio, si tomamos Israel como referencia, que ha tenido un número de infectados menor con un total de 839.000 infectados a mayo del 2021, el gobierno de Israel [36] ha transmitido información de todo tipo, desde datos globales, hasta la situación general del país, indicando en todo momento los nuevos casos que se generaban, los hospitalizados, el porcentaje de pruebas realizadas que han dado positivo y negativo, y el número de vacunas suministradas.

En el caso de Turquía, es el país con mayor número de contagiados de esta región a mayo del 2021, llegando a superar los 5 millones, esto es debido a que a partir de diciembre del 2020 hubo un repunte de casos que repercutió negativamente y antes de estabilizar la situación, se generó una segunda ola de infectados, teniendo como consecuencia el mayor número de contagios en un día, en concreto el 12 de abril del 2021, con más de 400.000 casos [37].

REGION	COUNTRY	GOVERNMENT/MINISTRY OF HEALTH DATA	CSSE JHU	WHO	INFECTED	TOTAL
WESTERN ASIA						
	Saudi Arabia	7*	3	3	7	19
	Armenia	10	3	3	5	21
	Azerbaijan	7	3	3	7	20
	Bahrain	1	3	3	5	12
	Qatar	3	3	3	5	14
	Cyprus	10*	3	3	3	22
	United Arab Emirates	5	3	3	7	18
	Georgia	3	3	3	7	16
	Iraq	1	3	3	10	17
	Israel	7	5	3	7	22
	Jordan	0	3	3	7	13
	Kuwait	7	3	3	5	18
	Lebanon	10	3	3	7	23
	Oman	0	3	3	5	11
	Syria	0	3	3	3	9
	Turkey	14	3	3	10	30
	Yemen	3	3	3	3	12

Figura 5.11: Clasificación Asia Occidental

Asia Meridional

Es la región más afectada de Asia, habiendo tenido los países de esta región un número de infectados muy altos, en concreto la India, la cual ha tenido más de 24 millones de infectados acumulados a mayo del 2021 [38]. Además, el día 3 de mayo es el día con el mayor número de casos de infectados confirmados en un día, en concreto 2.738.957 infectados nuevos. Al tener un número tan alto de infectados le hace estar en lo más alto de la clasificación, siguiéndole Pakistán, el cual, su gobierno [39] como se puede observar en la Figura 5.12 ha mostrado en todo momento la actualidad, incorporando gráficas y dividiendo los datos por regiones, aunque no llegue al millón, Pakistán es el único país de la región que ha

Propagación Covid 19 Recolección, Tratamiento y Análisis de los datos

tenido tres olas claramente pronunciadas a mayo de 2021.

REGION	COUNTRY	GOVERNMENT/MINISTRY OF HEALTH DATA	CSSE JHU	WHO	INFECTED	TOTAL
SOUTH ASIA						
	Afghanistan	0	3	3	3	9
	Bangladesh	7	3	3	7	20
	Bhutan	3	3	3	3	12
	India	6	5	3	12	26
	Iran	0	3	3	10	16
	Maldives	3	3	3	3	12
	Nepal	10	3	3	7	23
	Pakistan	10	5	3	7	25
	Sri Lanka	6	3	3	3	15

Figura 5.12: Clasificación Asia Meridional

Asia Oriental

Japón es el país con más contagiados de la región Oriental, superando los 660 mil infectados, además de que es el único país de Asia Oriental, que llega a obtener la puntuación más alta en la puntuación de los datos ofrecidos por el gobierno Figura 5.13. Esto es debido a que el gobierno de Japón [40] generó varios tipos de informes sobre el Covid-19 desde datos acumulados, incluyendo los hospitalizados hasta informes de mapas de calor, dividiendo el país en ciudades y comparándolas entre sí. La gran parte de informes generados se centran en mostrar los datos con diferentes tipos de histogramas, en concreto sobre los últimos siete días, comparándolo con una gráfica lineal acumulativa.

REGION	COUNTRY	GOVERNMENT/MINISTRY OF HEALTH DATA	CSSE JHU	WHO	INFECTED	TOTAL
NORTHEAST ASIA						
	China	5	5	3	5	18
	North Korea	0	0	0	0	0
	South Korea	10	3	3	5	21
	Japan	14	5	3	7	29
	Mongolia	1*	3	3	3	10
	Taiwan	7*	3	0	3	13

Figura 5.13: Clasificación Asia Oriental

Sudeste Asiático

En esta región, se puede observar que la mayoría de los casos han sido por transmisión comunitaria, resulta curioso si observamos la Figura 5.14, Filipinas tiene un número de infectados bastante alto a mayo del 2021, el gobierno no ha mostrado datos, en cambio Indonesia, para ser el país de la región con más infectados a mayo del 2021, un total de más de un millón 1.700.000 casos, el gobierno quiso mostrar todos los datos relevantes incluyendo gráficas e incluso un mapa de calor para ver donde se ha producido el mayor número de infecciones [41]. En cambio, Malasia

Propagación Covid 19
Recolección, Tratamiento y Análisis de los datos

no teniendo un número tan alto de infectados, en mayo de 2021, se situó por debajo de los 500 mil infectados, pero al tener un país con alrededor 27 millones de habitantes, le hace destacar como un país donde los picos de infectados han sido altos, e incluso a mayo de 2021 todavía no ha controlado la última ola generada. [42].

REGION	COUNTRY	GOVERNMENT/MINISTRY OF HEALTH DATA	CSSE JHU	WHO	INFECTED	TOTAL
SOUTHEAST ASIA						
	Myanmar (Birmania)	0	0	3	3	6
	Brunei	5	3	3	3	14
	Cambodia	10**	3	3	3	22
	Philippines	0	3	3	10	16
	Indonesia	10	3	3	10	26
	Laos	3	3	3	3	12
	Malaysia	10	3	3	7	23
	Singapore	3	3	3	3	12
	Thailand	3	3	3	3	12
	East Timor	0	3	3	3	9
	Vietnam	10	3	3	3	19

Figura 5.14: Clasificación Sudeste Asiático

5.3.3. Europa

Forma parte del supercontinente Euroasiático, en términos de superficie es el segundo continente más pequeño tan solo por delante de Oceanía, está formado por cincuenta países, para realizar un análisis en profundidad lo hemos dividido en cinco regiones por la situación geográfica distintas: Europa Meridional, Europa Septentrional, Europa Central, Europa Occidental y Europa Oriental.

Europa Meridional

Es una de las regiones que más se han visto afectadas dentro de Europa y está formada por la mayor cantidad de países. La mayoría de los casos han sido por transmisión comunitaria. Esta región cuenta con dos de los países que más casos tienen en el continente, siendo estos España e Italia, superando ambos los 3.500.000 casos acumulados. Los gobiernos de ambos países proporcionan una gran cantidad de datos detallados y divididos por regiones. Algo que tienen en común los dos países más afectados de la región, es que fueron los primeros que presentaron casos del virus en Europa [43], siendo el primero de estos en Italia, Figura 5.15.

EUROPE AREA	COUNTRY	GOVERNMENT/MINISTRY OF HEALTH DA	CSSE JHU	WHO	INFECTED	TOTAL
EUROPA MERIDIONAL						
	Albania		3	3	3	12
	Andorra		5	3	3	14
	Bosnia y Herzegovina		8	3	3	17
	Ciudad del Vaticano		0	3	3	9
	Croacia		10	3	3	19
	Eslovenia		13*	3	3	22
	España		14*	5	3	32
	Grecia		0	3	3	9
	Italia		14*	5	3	32
	Macedonia del Norte		3	3	3	12
	Malta		0	3	3	9
	Montenegro		0	3	3	9
	Portugal		10*	3	3	19
	San Marino		0	3	3	9
	Serbia		3	3	3	12

Figura 5.15: Clasificación Europa Meridional

Europa Septentrional

En esta región destaca uno de los países en los que más impacto ha tenido el Covid-19, a mayo de 2021 presentando más de cuatro millones de casos y más de 100 mil fallecimientos, de hecho, en este país ha aparecido una nueva variante del virus [44]. A parte de Reino Unido, Dinamarca es uno de los países que ofrecen mayor cantidad de datos aportando tanto datos diarios como datos acumulados, además de aportar gráficas, también destaca por dar datos del número de hospitalizados, aunque los casos confirmados del país apenas llegan a los 300 mil. El resto de los países de esta región coinciden en que no destacan por el número de casos que tienen, además de esto, no aportan apenas datos, Figura 5.16.

EUROPE AREA	COUNTRY	GOVERNMENT/MINISTRY OF HEALTH DA	CSSE JHU	WHO	INFECTED	TOTAL	
EUROPA SEPTENTRIONAL							
	Dinamarca		10*	5	3	3	21
	Estonia		10	3	3	3	19
	Finlandia		0	3	3	3	9
	Irlanda		3	3	3	3	12
	Islandia		3	3	3	3	12
	Letonia		0	3	3	3	9
	Noruega		0	3	3	3	9
	Reino Unido		7*	5	3	10	25
	Suecia		6	5	3	5	19

Figura 5.16: Clasificación Europa Septentrional

Europa Central

En esta Región, destaca Alemania frente al resto de los países que la componen, presenta más de 3 millones y medio de casos a mediados de mayo de 2021 y con más de 80 mil fallecimientos. El país alemán presenta gran cantidad de datos e informes diarios [45] además de disponer de su propio dashboard donde aportan datos diarios y acumulados divididos por regiones.

Austria es otro país de esta región que presenta los datos de manera similar a Alemania, aunque se ha visto menos afectado por el virus que su país vecino, cabe destacar que pese a tener más de 700 mil casos confirmados apenas tiene poco más de 10 mil fallecidos [46], una tasa muy baja comparada con el resto de los países europeos y muy similar a los datos que presenta Suiza.

El resto de los países dentro de esta región destacan por la falta de información que ofrecen sus respectivos gobiernos, exceptuando a Luxemburgo cuyo gobierno aporta una gran cantidad de datos y gráficos relevantes, aunque apenas presenta casos en comparación al resto de países, Figura 5.17.

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

EUROPE AREA	COUNTRY	GOVERNMENT/MINISTRY OF HEALTH DA	CSSE JHU	WHO	INFECTED	TOTAL
EUROPA CENTRAL						
	Alemania		14*	5	3	10
	Austria		14*	3	3	3
	Eslovaquia		0	3	3	3
	Hungria		3	3	3	3
	Liechtenstein		0	3	3	3
	Lituania		3	3	3	3
	Luxemburgo		14*	3	3	3
	Polonia		0	3	3	5
	Republica Checa		0	3	3	5
	Rumania		0	3	3	5
	Suiza		10	3	3	3

Figura 5.17: Clasificación Europa Central

Europa Occidental

Formado por Bélgica, Francia, Países Bajos y Mónaco, Figura 5.18.

Francia es el país que más destaca dentro de esta región puesto que es la región de Europa que más casos presenta a 2021, presenta más de 5.800.000 casos confirmados y más de 50.0000 fallecimientos. El gobierno francés presenta reportes diarios y además proporciona una plataforma donde se puede buscar la información por región [47].

Países Bajos de forma similar a Francia aporta una gran cantidad de datos de manera diaria, proporciona resúmenes semanales y dispone de gráficos y mapas para representar los datos[48],de manera similar trata los datos Bélgica que aporta una gran cantidad de gráficas y datos tanto diarios como acumulados, además de proporcionar medias de los datos de la última semana y dos semanas [49].

EUROPE AREA	COUNTRY	GOVERNMENT/MINISTRY OF HEALTH DA	CSSE JHU	WHO	INFECTED	TOTAL
EUROPA OCCIDENTAL						
	Belgica		14*	5	3	5
	Francia		14*	5	3	10
	Monaco		3	3	3	3
	Países Bajos		14*	5	3	5

Figura 5.18: Clasificación Europa Occidental

Europa Oriental

Todos los países de esta región Figura 5.19 comparten que apenas aportan datos oficiales y además de esto ninguno destaca por tener una gran cantidad de infectados, a excepción de Ucrania. Ucrania a mayo de 2021 presenta más de 2.000.000 casos confirmados y roza los 50.000 fallecimientos, el gobierno ucraniano aporta datos, pero no aporta gráficas ni ningún tipo de dashboard [50].

EUROPE AREA	COUNTRY	GOVERNMENT/MINISTRY OF HEALTH DA	CSSE JHU	WHO	INFECTED	TOTAL
EUROPA ORIENTAL						0
	Armenia		0	3	3	9
	Azerbaijan		3	3	3	12
	Bielorrusia		0	3	3	9
	Bulgaria		1	3	3	10
	Georgia		7	3	3	16
	Moldavia		0	3	3	9
	Ucrania		3*	5	3	16

Figura 5.19: Clasificación Europa Oriental

Capítulo 6

Metodología del análisis

En este capítulo, se va a explicar la metodología del proyecto incluyendo la creación de los scripts para la obtención de los datasets, las ecuaciones de los indicadores utilizados y el método de clustering. Todo ello para conseguir el objetivo de comparar el comportamiento de la segunda ola en diferentes regiones.

Dependiendo de cada país la pandemia empezó en un periodo diferente y se aplicaron políticas diferentes, por lo tanto, hemos decidido clasificarlos en un tramo que sea de igual longitud para todos los países. Los tramos elegidos para poder realizar de forma normalizada esta comparación se sitúan en el centro de la segunda ola y se añaden 14 días tanto a la izquierda como a la derecha de la segunda ola. Se ha escogido la segunda puesto que la estructura de datos que proveían los distintos países era más adecuada ya que estaba dividida por regiones haciendo que los datos estuviesen más depurados en comparación con la primera ola.

6.1. Creación de scripts y datasets

Se realizó un estudio global de los diferentes repositorios existentes, buscando la mejor opción para generar la base de datos. Este estudio abarcó tanto las bases de datos globales como las nacionales que se publicaban por agencias o instituciones oficiales, una vez obtenida la información necesaria, se clasificó los países teniendo en cuenta tanto los datos que aportan, como el número de infectados de cada uno. Generando un ranking dividido por regiones continentales.

Tras haber analizado las posibles bases de datos, se tomó la decisión de seleccionar una base de datos global, en concreto la base de datos de la Johns Hopkins University Center for Systems Science and Engineering, ya que, contiene reportes de todos los gobiernos, proporciona las variables

que se requerían y además se actualizaba de forma diaria.

Una vez terminado el análisis, se procedió a la preparación del entorno y la obtención de los datos para su posterior recopilación, dando como resultado el desarrollo e implementación de un script para descargar y preparar los directorios donde se colocarían los CSV con los datos a procesar. El script *updateCSV.sh* da soporte para la actualización con los nuevos datos disponibles en el repositorio de la JHU, para añadirlos a nuestro conjunto de datos, este script se puede encontrar en el repositorio Github [19].

Tras estudiar en detenimiento la información disponible en la JHU, se observó que los datos ofrecidos se trataban de datos acumulados, siendo de mayor interés para el proyecto los datos diarios, que serán los que se usen más tarde para los cálculos de las tasas y los diferentes índices. Para realizar el procesado de los diferentes datos diarios, primero los agrupamos por país y luego, de existir, por región. Aunque, en un principio se pensó en usar los datos de latitud y longitud, había disparidad en los valores de una misma región, así que se decidió prescindir de esas columnas. Una vez agrupados se calculó los valores diarios mediante el uso de una función, permitiendo restar los valores que representa un día con los datos acumulados del día anterior.

Tras una primera iteración, se observó que había incongruencias en los datos facilitados, existiendo casos en que los datos acumulados de un día eran menores a los del día anterior, dando como resultado el cálculo de datos diarios negativos. Si bien se consideró que no era extraño para los casos activos, que pueden pasar a ser recuperados o muertos, eran cifras extrañas para el resto de los datos diarios. Tras una deliberación y varias iteraciones con soluciones y aproximaciones diferentes, al final, se llegó al consenso de no modificar los datos acumulados, que en un principio se consideró corregir por datos acumulados del día anterior dando como resultado cero en el cálculo de los diarios, y poner a nulos todos los datos diarios menores que cero para indicar posibles datos erróneos o fuera de rango.

Una vez solucionado esto, se detectó que había un rango de fechas donde el formato de las tablas de origen era significativamente diferente al resto, por lo que se producían conflictos a la hora de hacer la unión entre todos los CSV. Primero se procedió a acotar las fechas en las que el formato era dispar para hacer un primer proceso, recogido en el archivo *changeFormat.py*. Tras renombrar las diferentes columnas y unificar los nombres de varios países que eran incluidos de diferente forma, por ejemplo, Corea del Sur que aparecía también como Corea del Sur, se volvió a procesar todos los datos para su análisis. En esta iteración se encontró otro problema de homogeneidad, esta vez en el campo que indicaba la

fecha a la que pertenecían los datos. Teniendo en cuenta esto, se añadieron los mecanismos necesarios para detectar el formato de la columna y transformarlo en el tipo adecuado para poder trabajar con ello.

Tras recoger diferentes evidencias y considerar los datos como aptos para los siguientes cálculos, por ejemplo, el índice de peligrosidad, y teniendo en cuenta que desde un principio se orientó el proyecto de forma que, los datos se clasifican por diferentes continentes, se procedió a generar el código y los filtros necesarios para separar los resultados. Como desenlace se generó el código de tres scripts diferentes:

1. *DataUSA.py*: Con él se generan los datos pertenecientes a EE. UU.
2. *WorldWithoutUSA.py*: Se generan los datos del resto del mundo.
3. *CountryFilter.py*: Se generan los datos del país indicado en los argumentos, dando la opción al usuario de filtrar también por una región o estado particular del mismo.

Tras ejecutar estos scripts y recibir como resultado un conjunto de datos con todas las variables `Daily_Confirmed`, `Daily_Death`, `Daily_Recovered` y el indicar IP, se ejecutó la función en R, para calcular la intensidad de una enfermedad infecciosa.

Para facilitar la iteración entre los diferentes procesos y componentes del grupo, se creó un repositorio en GitHub [19] donde se disponía del código que se desarrollaba para agilizar los cambios y llevar un histórico de los mismos.

Una vez obtenido el conjunto de datos, se detectaron problemas a la hora de realizar el estudio de las olas de los países más impactados, debido a la inclusión de datos nulos generando picos bruscos en las curvas de los datos, los cuales distorsionaban la visualización de estos. Por este motivo se llegó al consenso de la necesaria utilización de aplicar una media móvil de siete días en los datos diarios, con un doble objetivo, en primer lugar, solventar el problema de los datos nulos al rellenarse estos con la media móvil. Y, en segundo lugar, solucionando los picos que se generaban anteriormente con las variables diarias, generando visualizaciones más suaves y facilitando el estudio posterior de las olas en las diferentes regiones. Tras esto, se decidió añadir las variables `R0` e `IP`, pero esta vez calculadas con los datos suavizados con la media móvil anterior.

Una vez obtenidas las variables `R0` e `IP`, se realizó un proceso de selección de aquellos países que más afectados se han visto por el virus. Para dicha selección se realizó un análisis de los datos aportados por todos los países de cada continente, dependiendo de los datos que aportase cada país estos obtienen una puntuación, una vez calificados todos los países, aquellos tres de cada continente que han obtenido la mayor puntuación

son los seleccionados para pasar a la próxima fase del estudio. Cabe remarcar que un factor relevante dentro de nuestro estudio es que los países dividan los datos por regiones. Una vez seleccionados se escogieron tres de estos cinco países de cada continente y de cada país se seleccionaron las tres regiones que más se han visto afectadas por el virus.

En este caso, se tomó la segunda oleada de la pandemia como referencia, ya que, la primera en muchos casos no está dividida por regiones. Una vez encontrados los picos, se han creado tres tablas donde se almacenan todos los datos correspondientes a R_0 , IP e IA en las fechas de 12 días antes del pico y 12 días después del mismo.

6.2. Variables e Indicadores

Con estos scripts y funciones en R, se incorpora de forma automática las siguientes variables e indicadores, cabe destacar que los indicadores se han calculado tanto como con los datos sin tratar como con las variables calculadas con las medias móviles y que estos datos son ofrecidos de forma acumulada, por tanto, las ecuaciones se corresponden con los datos diarios teniendo la variable i el valor 1 para el primer día que se hubiera recogido datos.

1. Daily_Confirmed: Casos confirmados diarios para ese día y para ese país o región, se corresponde con las anotaciones que dicho país tiene para esa región mediante cualquier tipo de test.

$$C_i = CA_i - CA_{i-1}, \forall (i > 1)$$

2. Daily_Deaths: Muertes diarias para ese día y para ese país o región.

$$F_i = FA_i - FA_{i-1}, \forall (i > 1)$$

3. Daily_Recovered: Casos recuperados diarios para ese día y para ese país o región, que se corresponde con las anotaciones de cada región sobre las personas que han superado el virus.

$$R_i = RA_i - RA_{i-1}, \forall (i > 1)$$

4. R0: Número de reproducción o factor de transmisión R0, también se puede estimar con los datos diarios, describe la capacidad de transmisión que tiene en un momento dado una enfermedad infecciosa. Un valor inferior a 1 indica que no hay peligro.

$$R_0^i = \frac{C_i}{C_{i-14}}, \forall i \geq 15$$

5. IA: Incidencia acumulada, esta variable refleja la expansión de una enfermedad en un tiempo y en una población determinada, según la OMS una IA inferior a 150 es un nivel medio bajo de peligro, entre 150-250 un nivel alto y con más de 250, el riesgo sería extremo.

$$Inc^i = \left(\sum_{j=1}^{14} C_{i-j} \right) * 100000/P, \forall i \geq 15$$

6. IP: Índice de peligrosidad, se calcula con los datos diarios utilizando las tres variables C, F y R, representando así el equilibrio entre ellas.

$$IP^i = C_i + F_i - R_i, \forall i \geq 1$$

7. Media móvil de siete días:

$$\forall i \geq 4, x_i = \frac{1}{7} \sum_{k=i-3}^{i+3} x_k$$

8. R0MM: Número de reproducción o factor de transmisión R0, que utiliza los nuevos datos suavizados con la media móvil.

$$R_0^i = \frac{mm(C_i)}{mm(C_{i-14})}, \forall i \geq 15$$

9. IPMM: Índice de peligrosidad, que usa los nuevos datos suavizados con la media móvil.

$$IP^i = mm(C_i) + mm(F_i) - mm(R_i) , \forall i \geq 1$$

6.3. Método de clustering

El algoritmo de K-Means [51] es un algoritmo de machine learning, que consiste en la agrupación de un conjunto de datos en diferentes grupos o clústeres. Estas agrupaciones se realizan por la similitud que presentan las componentes de cada grupo, creando de esta manera una agrupación en los datos por componentes de características similares, Figura 6.1.

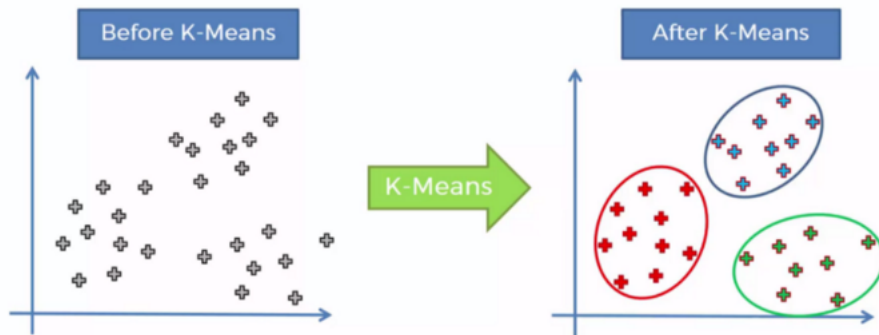


Figura 6.1: Algoritmo K-Means. Fuente: Towards Data Science

Tomando un grupo de observaciones $(x_1, x_2 \dots x_N)$ donde cada observación es un vector, el algoritmo de K-means construye una agrupación de todas las observaciones con el objetivo de minimizar la diferencia entre los elementos de cada conjunto y maximizar la diferencia entre los elementos de conjuntos distintos.

Cálculo de centroides

Para poder calcular de manera precisa el número de centroides óptimo para la agrupación final, siendo un centroide la observación central dentro de cada grupo generado, el algoritmo itera alternando entre dos pasos distintos.

- Primer paso: Escogemos un valor para K que se considere más adecuado.

- Segundo paso: Seleccionar aleatoriamente K observaciones y asignar a cada una de ellas una clase.
- Tercer paso: Asignar como centroide de cada clase la observación que se ha asignado en el segundo paso, Figura 6.2.

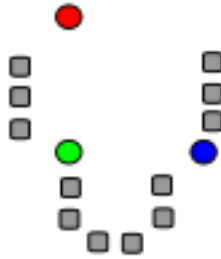


Figura 6.2: K-Means: Paso de inicialización (Pasos 1,2,3). Fuente: Wikipedia

- Cuarto paso: Calculamos la distancia de las observaciones a los todos los centroides.
- Quinto paso: Asignar a cada observación la misma clase del centroide de más cercano calculado en el paso anterior, si se asigna el mismo centroide que en la iteración anterior se para el proceso, Figura 6.3.

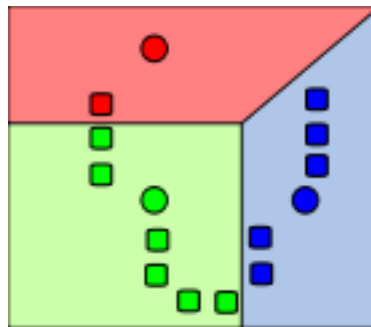


Figura 6.3: K-Means: Pasos generales (Pasos 4,5). Fuente: Wikipedia

- Sexto paso: Recalcular los centroides de cada clase, si al recalcular los centroides no hay cambios se para el proceso, Figura 6.4.

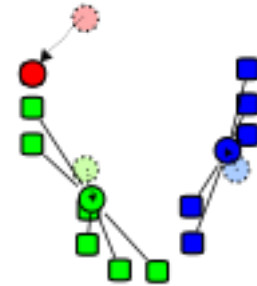


Figura 6.4: K-Means: Pasos generales (Paso 6). Fuente: Wikipedia

- Séptimo paso: Volver al cuarto paso de nuevo, si se han realizado ya demasiados pasos (un número determinado) se para el proceso, Figura 6.5.



Figura 6.5: K-Means: Pasos generales (Paso 7). Fuente: Wikipedia

Función K-Means en R

El problema que presenta el algoritmo de K-means en R es, que el número de clústeres identificados es un parámetro de entrada, por eso tenemos que calcular cual es el número de K clústeres óptimos, mediante el estudio de unos ratios que devuelve la función de K-means. Estos ratios son los siguientes:

- **betweenSS**: Representa la media de las distancias entre los centros de los conjuntos. Interesa que este ratio sea lo más alto posible, ya que, esto implicaría que los diferentes grupos son los más heterogéneos posibles.
- **withinSS**: Representa la suma de cuadrados dentro del clúster. Dando como resultado un vector con un número para cada conjunto. Interesa que este ratio sea lo más bajo posible para cada clúster, ya que, indicaría que hay mayor homogeneidad dentro de los clústeres.

Una vez identificados estos ratios, se llevó a cabo el proceso para identificar qué valor de K era el más óptimo para el proyecto. A continuación, se ejecutó el algoritmo dado a K valores entre $1 \dots N$, anotando en cada ejecución los ratios devueltos por **betweenSS** y **withinSS**, con el objetivo de seleccionar el k que cumpla, que el valor devuelto por **betweenSS** sea el más elevado posible y los valores devueltos en el vector de **withinSS** sean los más bajos posibles.

Capítulo 7

Resultados del Análisis Clúster

En este capítulo se van a mostrar los resultados de la aplicación del algoritmo K-Means por regiones en España y en las regiones más afectadas en América, Asia y Europa. Todo ello aplicado para el periodo de días 14 días antes y 14 días después tomando como punto central de referencia el pico de casos dentro de la segunda ola para cada país. Todas las regiones son estudiadas en base al rango de días mencionado anteriormente, aunque estos días puedan corresponder a fechas diferentes ya que el pico de la segunda ola no ocurrió de manera simultánea en todas las regiones.

7.1. Clasificación por regiones de España

Se ha utilizado el método de clasificación machine learning K-Means para agrupar las distintas regiones a partir de las distintas variables disponibles. Tras la realización de las tres tablas R0, IP e IA calculadas anteriormente, hemos extraído una serie de conclusiones de cada tabla. El número de clústeres elegidos para realizar el K-Means es seis, ya que, es el número de clúster que maximiza la distancia entre grupos y minimiza la distancia intragrupal. En la Figura: 7.1 se puede observar la relación entre las variables de confirmados, recuperados y muertes en todas las regiones de España.

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos



Figura 7.1: Gráficas Generales por comunidades España

7.1.1. Factor de transmisión R0

Como se puede observar en la Figura: 7.2. El primer clúster que se puede observar en color negro contiene las comunidades de Navarra, Extremadura, Castilla y León, Castilla - La Mancha, Baleares, Asturias y Andalucía, estas regiones comienzan en el pico de la ola con un valor de R0 en torno al 1.5 y a los doce días del pico, dicho valor desciende considerablemente en torno a un 55 %, a excepción de Baleares que mantiene una tendencia bajista durante todo el periodo. El clúster rojo compuesto por Murcia, Melilla y Aragón presenta un valor medio igual a 2.4 en el pico de la ola, en la segunda mitad del periodo dicho valor desciende un 49 %, hasta alcanzar valores similares a los iniciales. El Clúster verde formado por Ceuta y Catalonia, posee inicialmente un valor en torno a 2, y tiende a descender hasta alcanzar valores incluso un 74 % inferiores. El clúster azul marino, contiene las provincias de País Vasco, Cantabria y Comunidad Valenciana, estas tienen en común que el valor de R0 apenas varía durante todo el periodo. El clúster en color celeste está formado únicamente por las islas Canarias, puesto que no se asemeja su R0 al resto de regiones de la Península Ibérica. El último clúster, en color rosa y formado por Madrid, La Rioja y Galicia, tiene en común que su R0, en el pico oscila en torno a uno, con una media inicial de 1.37 y en la última parte del periodo su valor decremента alrededor de un 50 % alcanzando el 0.68 y manteniendo siempre un valor reducido. La Figura: 7.3 muestra todos los datos de R0 dentro de cada región.

Propagación Covid 19
Recolección, Tratamiento y Análisis de los datos

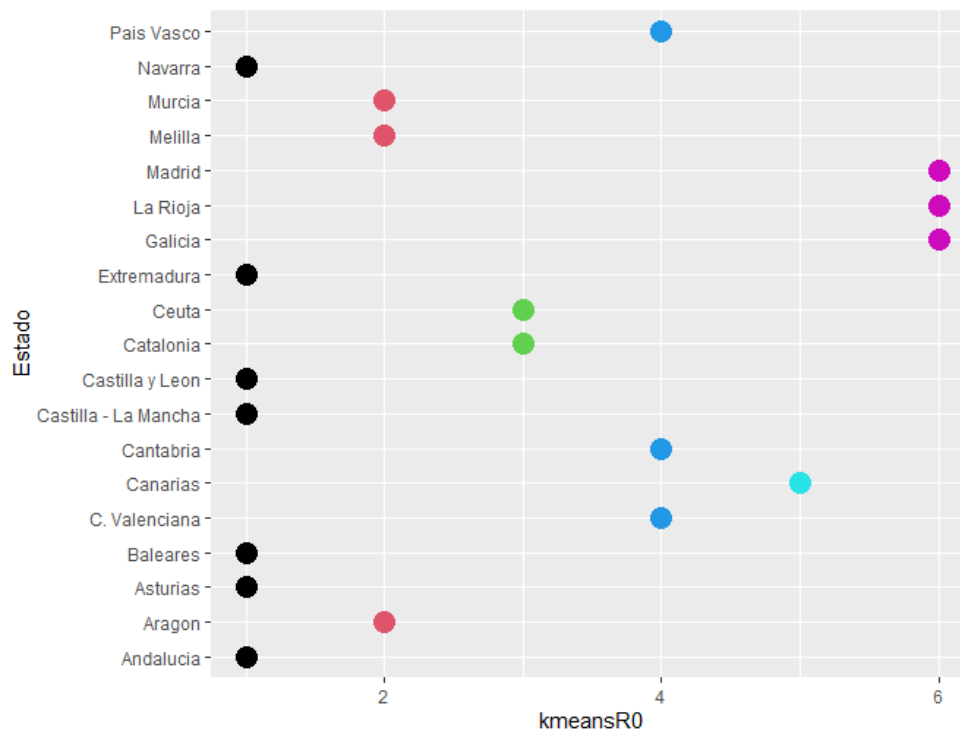
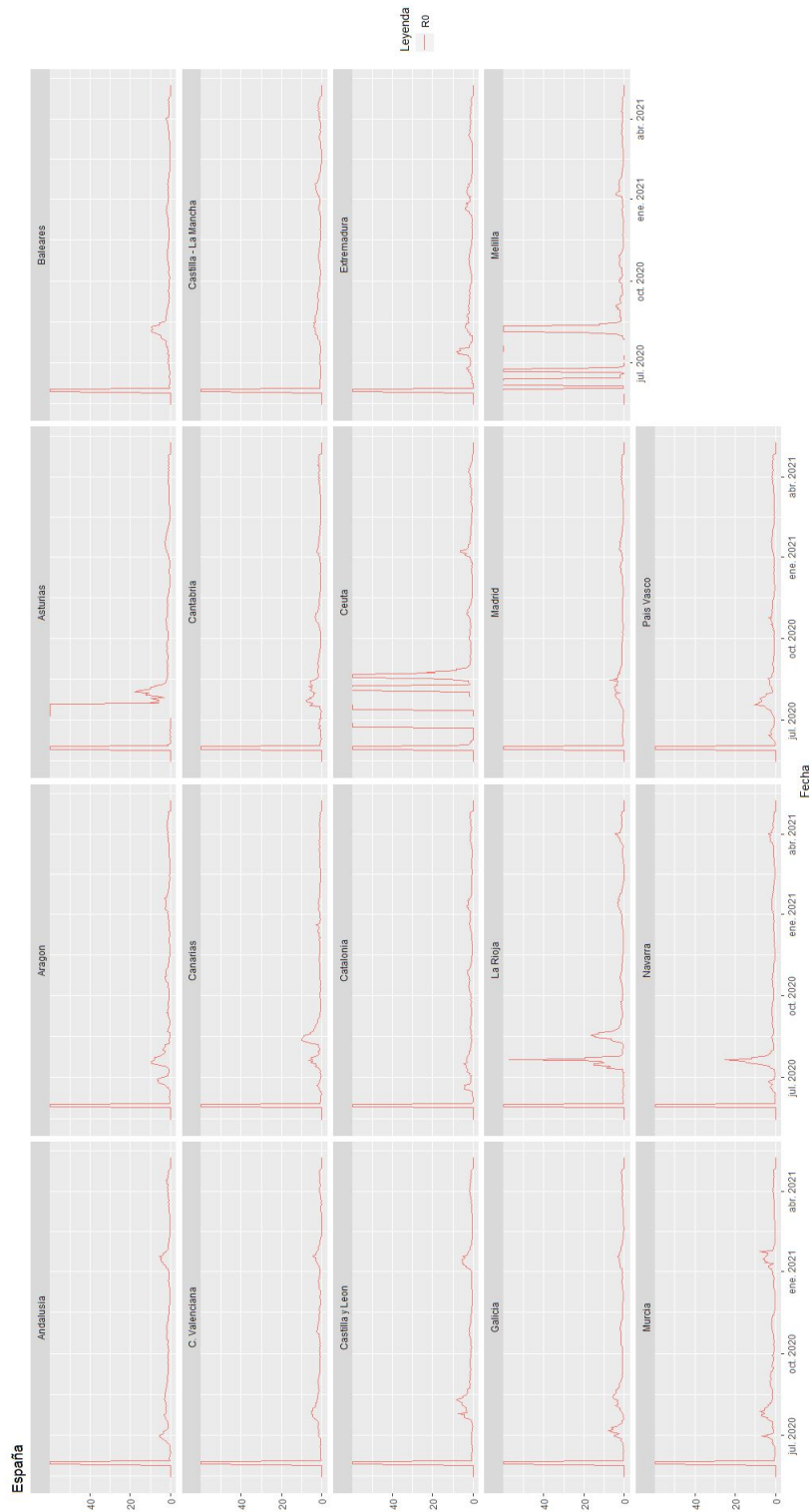


Figura 7.2: Análisis cluster por regiones de España en base a R0 y a la segunda ola

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos



7.1.2. Índice de peligrosidad (IP)

Como se puede observar en la Figura: 7.4. El clúster negro, formado por Navarra, Galicia y Baleares, comparte un valor medio inicial igual a 519.42, que apenas varía a lo largo del estudio, mientras que, en el clúster en color rojo, este rango oscila entre 300 y 500, con valor medio inicial de 362, de 518 a mitad del periodo, y de 292.28 al final, lo que supone una caída del 40 % con respecto al pico de la ola, conteniendo este clúster a Extremadura y Asturias. Con respecto al clúster en color verde, formado por País Vasco, Murcia, Castilla - La Mancha, Comunidad Valenciana y Aragón, se puede observar cómo el valor medio inicial es de 658.65 y los primeros días antes del pico de la ola toman valores entre 500 y 1300, dichos valores escalan hasta 1023.73 de media en el pico de la ola. El clúster en azul marino presenta valores muy bajos durante el periodo de estudio, no superando en ningún caso el valor de 200 y estando formado por Melilla, La Rioja, Ceuta, Cantabria y Canarias. El quinto clúster, en color celeste, contiene únicamente a Madrid y a Castilla y León, este es un caso interesante, ya que, ambas regiones comienzan el periodo con datos muy dispares, pero a lo largo de los días, tienden a aproximarse llegando incluso a cruzarse. Por último, el clúster rosa pertenece a Catalonia y Andalucía donde el IP alcanza valores más altos, empezando con un valor medio igual a 2686, llegando a aumentar hasta un 39 % tomando valores de 4398.85 en el centro del estudio, y disminuyendo en la segunda mitad para acabar por encima de 3593.78. La Figura: 7.5 muestra todos los datos de IP dentro de cada región.

Propagación Covid 19
Recolección, Tratamiento y Análisis de los datos

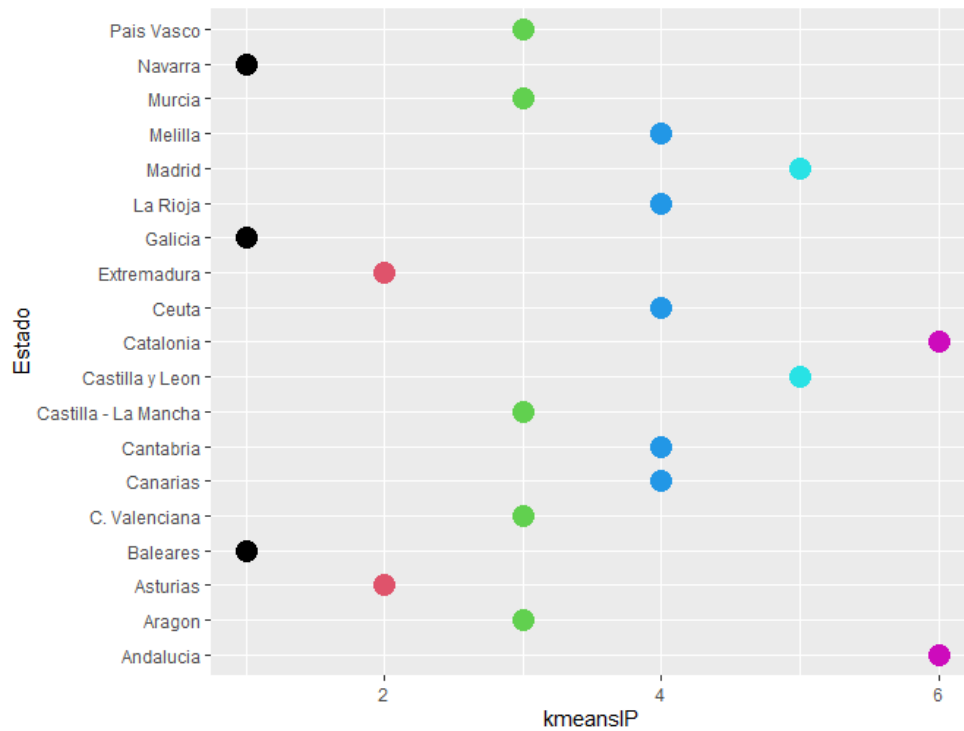


Figura 7.4: Análisis cluster por regiones de España en base a IP y a la segunda ola

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

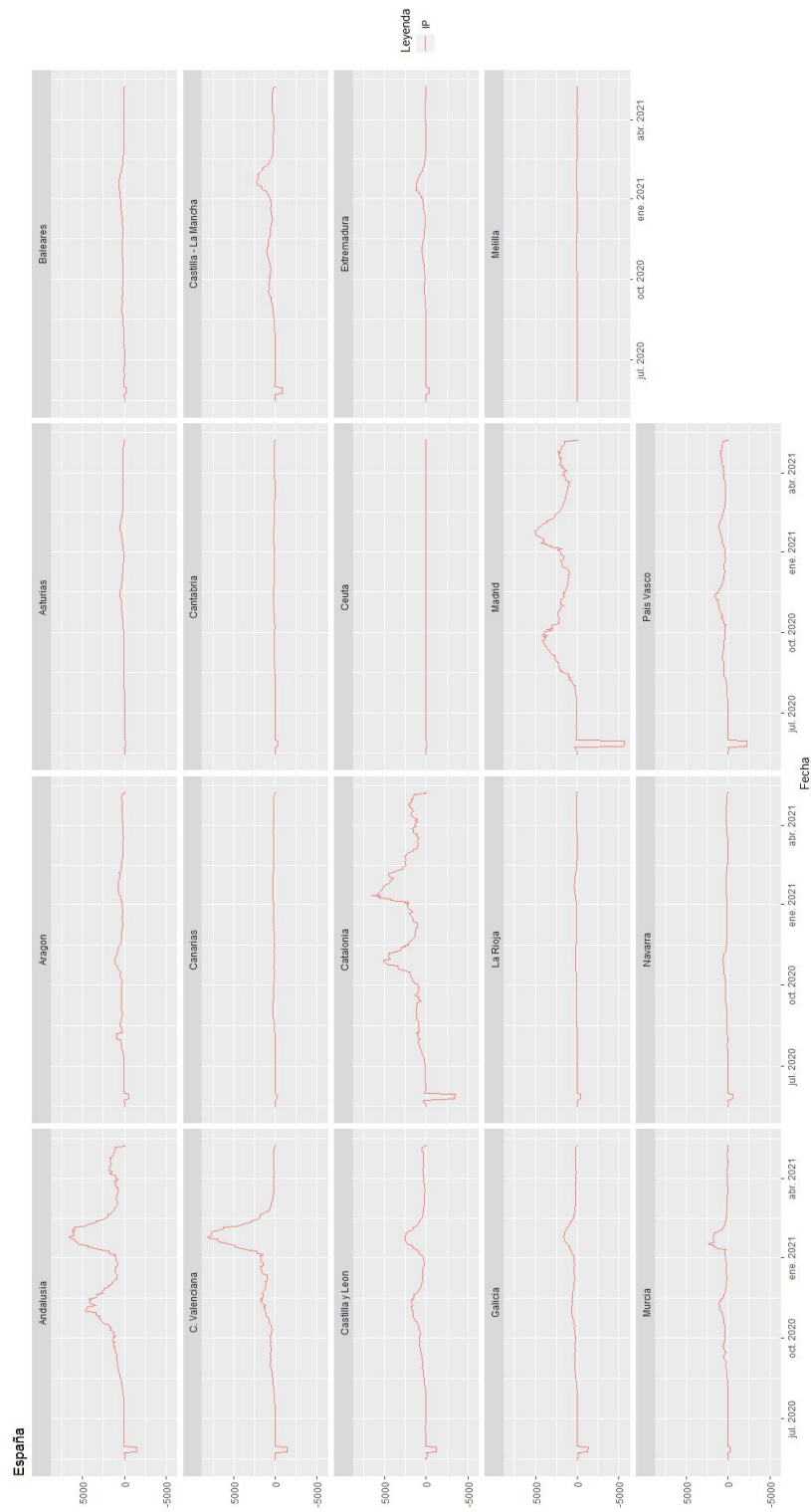


Figura 7.5: Gráficas Índice Peligrosidad España

7.1.3. Incidencia acumulada (IA)

Como se puede observar en la Figura: 7.6. El primer clúster, en color negro, está formado por Madrid, Galicia, Cantabria y Comunidad Valenciana, presenta una incidencia acumulada media de 379.59 durante el pico de la ola, al final del periodo estos valores aumentan levemente un 3% hasta los 390.47. Para el clúster de color rojo, el cual contiene a Extremadura, Castilla - La Mancha, Baleares, Asturias y Andalucía, esta variable toma cifras medias de 376.2 al inicio del periodo. las cuales aumentan hasta los 600.8 a mitad del estudio, manteniéndose en ese rango durante el resto del periodo. El clúster de color verde está formado únicamente por las islas canarias, ya que, toma valores muy inferiores al del resto de regiones, por esa razón se encuentra aislado. En el clúster de color azul marino, las regiones de País Vasco, Murcia, Catalonia y Aragón evolucionan de manera muy similar, encontrándose al empezar con una incidencia acumulada en torno a los 347.14 de valor medio, pero este valor va aumentando gradualmente, tomando en el centro del estudio, valores en torno a los 638.2, y finalizando en 905.26. El clúster celeste, agrupa las regiones de Navarra y Melilla, con valores iniciales que rondan los 581 de valor medio y que tienen una tendencia alcista durante todo el periodo, finalizando con valores superiores a los 1300. Las regiones de La Rioja, Ceuta y Castilla y León forman parte del clúster rosa, comienzan con un valor medio en torno a 514.63, pero escala rápidamente hasta un 43% en el centro del estudio a valores en torno a 891.3, manteniendo estos valores en el resto del periodo.

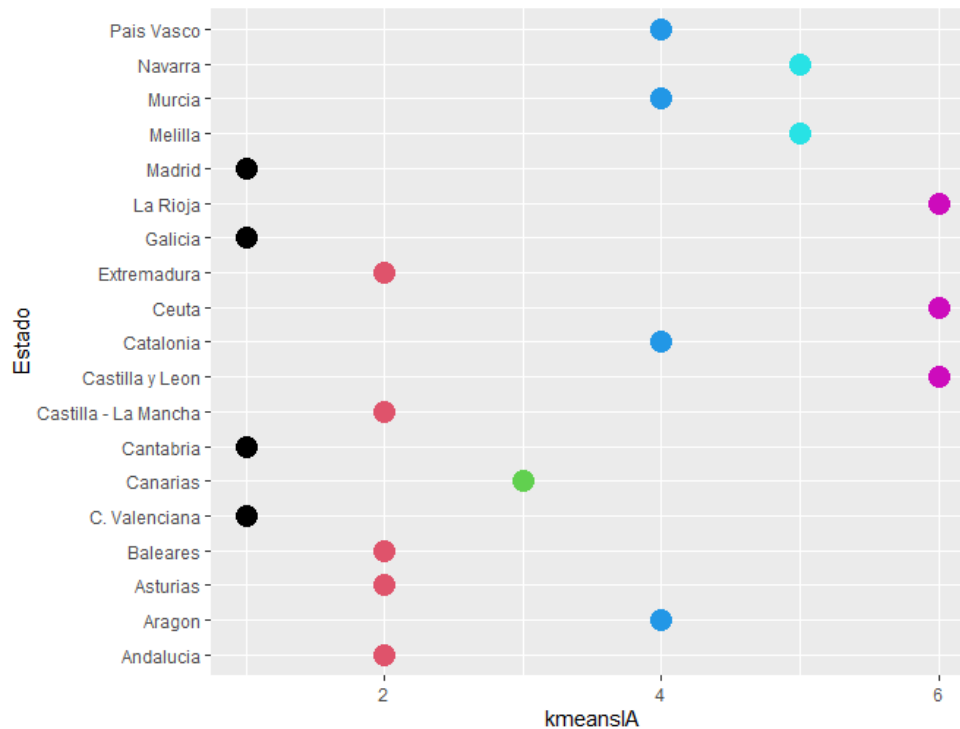


Figura 7.6: Análisis cluster por regiones de España en base a IA y a la segunda ola

7.1.4. Análisis K-Means

Tomando como referencia los tres análisis, se puede observar que hay una serie de regiones que comparten clúster, siendo estas Asturias con Extremadura y Aragón con Murcia. Por otra parte, destaca Canarias, la cual siempre se encuentra aislada a excepción de la variable IP, debido a que sus valores son muy inferiores a los del resto. Por otro lado, el resto de las regiones no coinciden entre ellas, a excepción de dos pares de regiones, siendo una Ceuta y La Rioja y la otra, la Comunidad Valenciana junto al País Vasco, que comparten clúster R0 e IP, R0 e IA respectivamente.

Las siguientes regiones han tenido un comportamiento similar en los peores momentos de la segunda ola en todas o alguna de las variables.

- Asturias y Extremadura.
- Aragón y Murcia.
- Ceuta y La Rioja (IP,IA).
- Comunidad Valenciana y País Vasco (R0, IP).

Es muy interesante a la hora de analizar las regiones, ver como Canarias al tener un impacto mucho menor al del resto de comunidades peninsulares, llega incluso a estar aislado en un clúster.

7.2. Clasificación por regiones de Asia, Europa y América

Aplicando el mismo método que con las regiones de España, se ha repetido el proceso con las regiones seleccionadas de los continentes estudiados anteriormente, con el objetivo de buscar correlación entre regiones de diferentes países y continentes. El número de clústeres elegidos para realizar el K-Means es seis, ya que, es el número de clúster que maximiza la distancia entre grupos y minimiza la distancia intragrupal. En las Figuras: 7.7, 7.8 y 7.9, se puede ver la relación de las variables de casos confirmados, recuperados y muertes en las regiones escogidas de cada continente.

Propagación Covid 19 Recolección, Tratamiento y Análisis de los datos

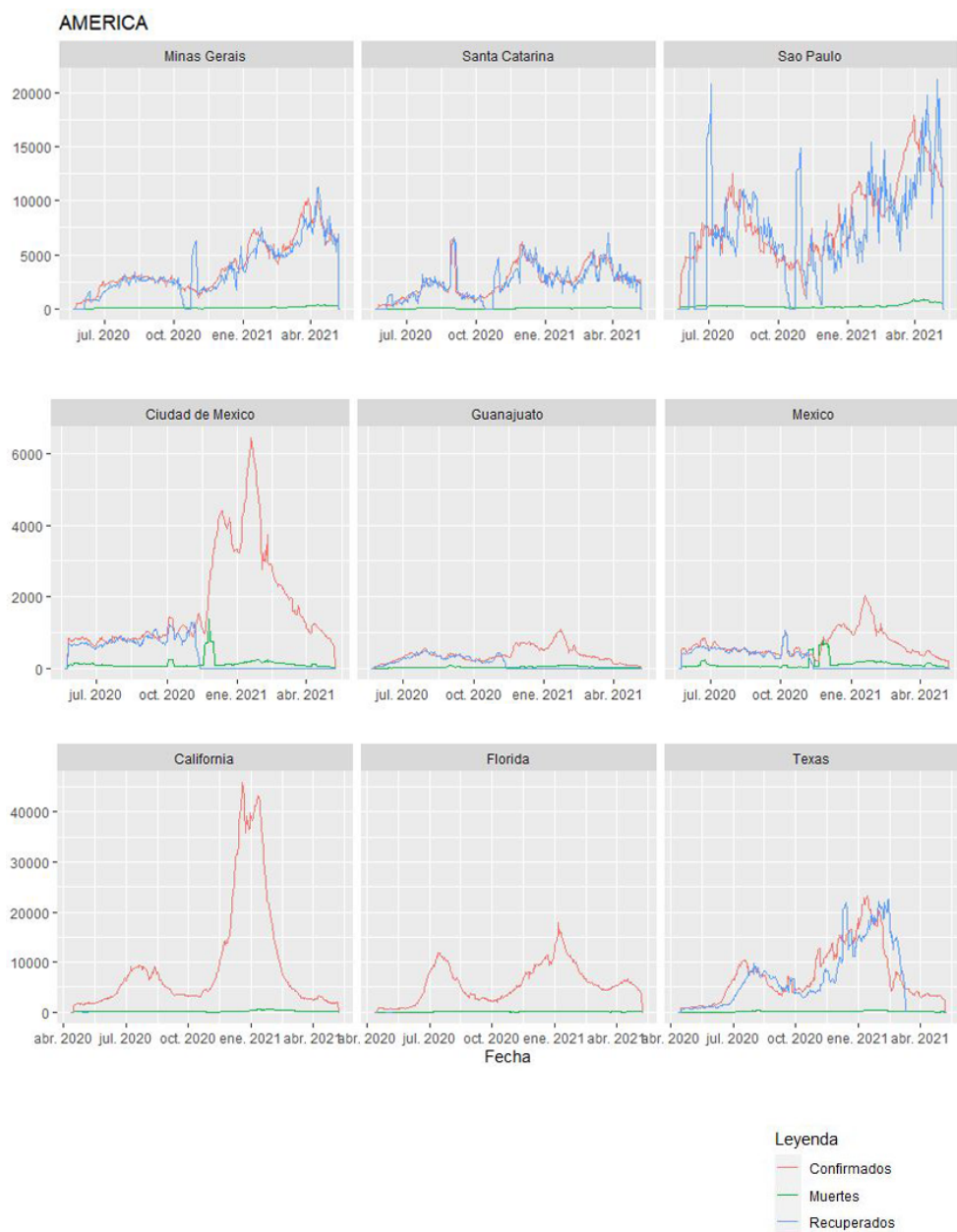


Figura 7.7: Gráfica General Regiones América

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

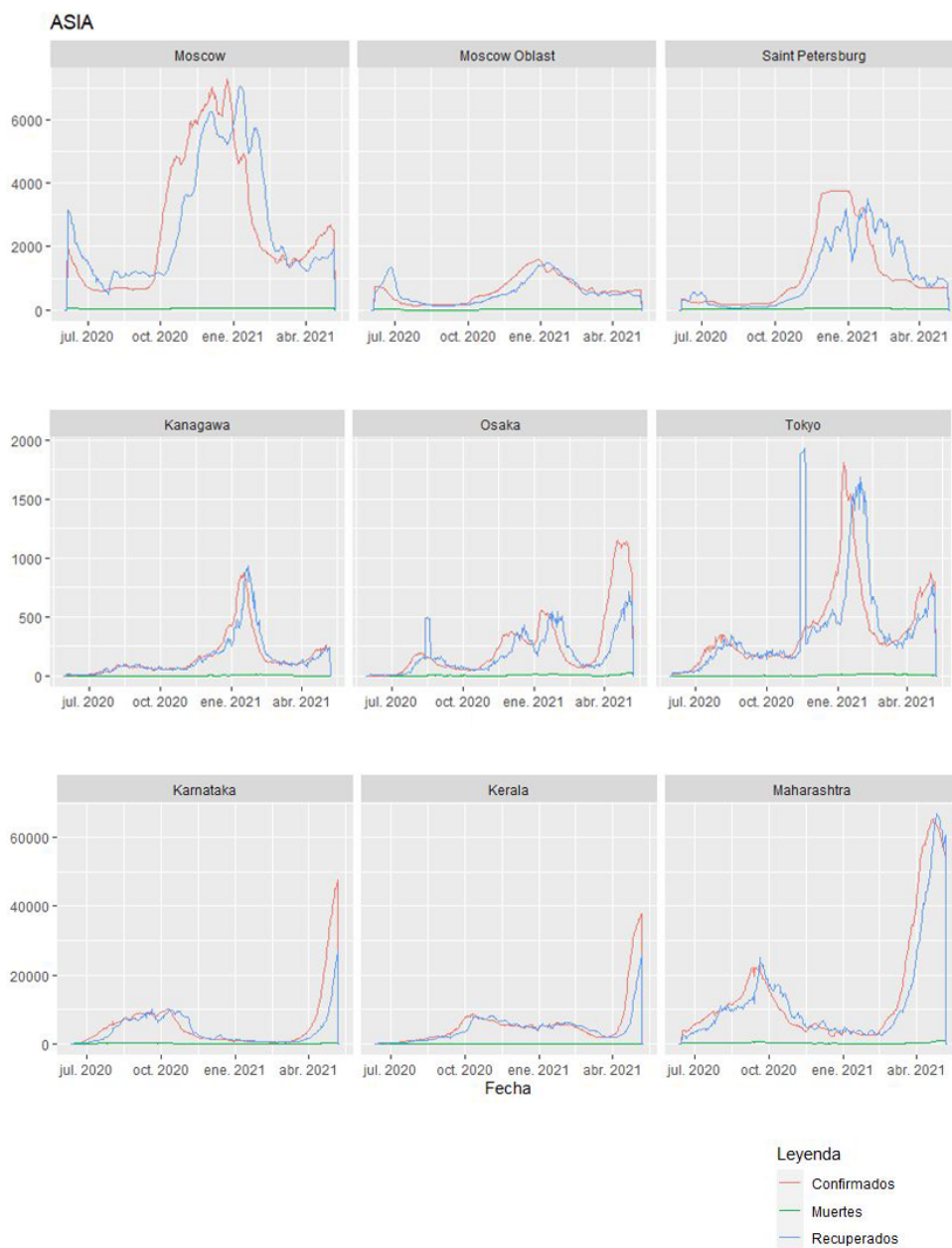


Figura 7.8: Gráfica General Regiones Asia

Propagación Covid 19 Recolección, Tratamiento y Análisis de los datos

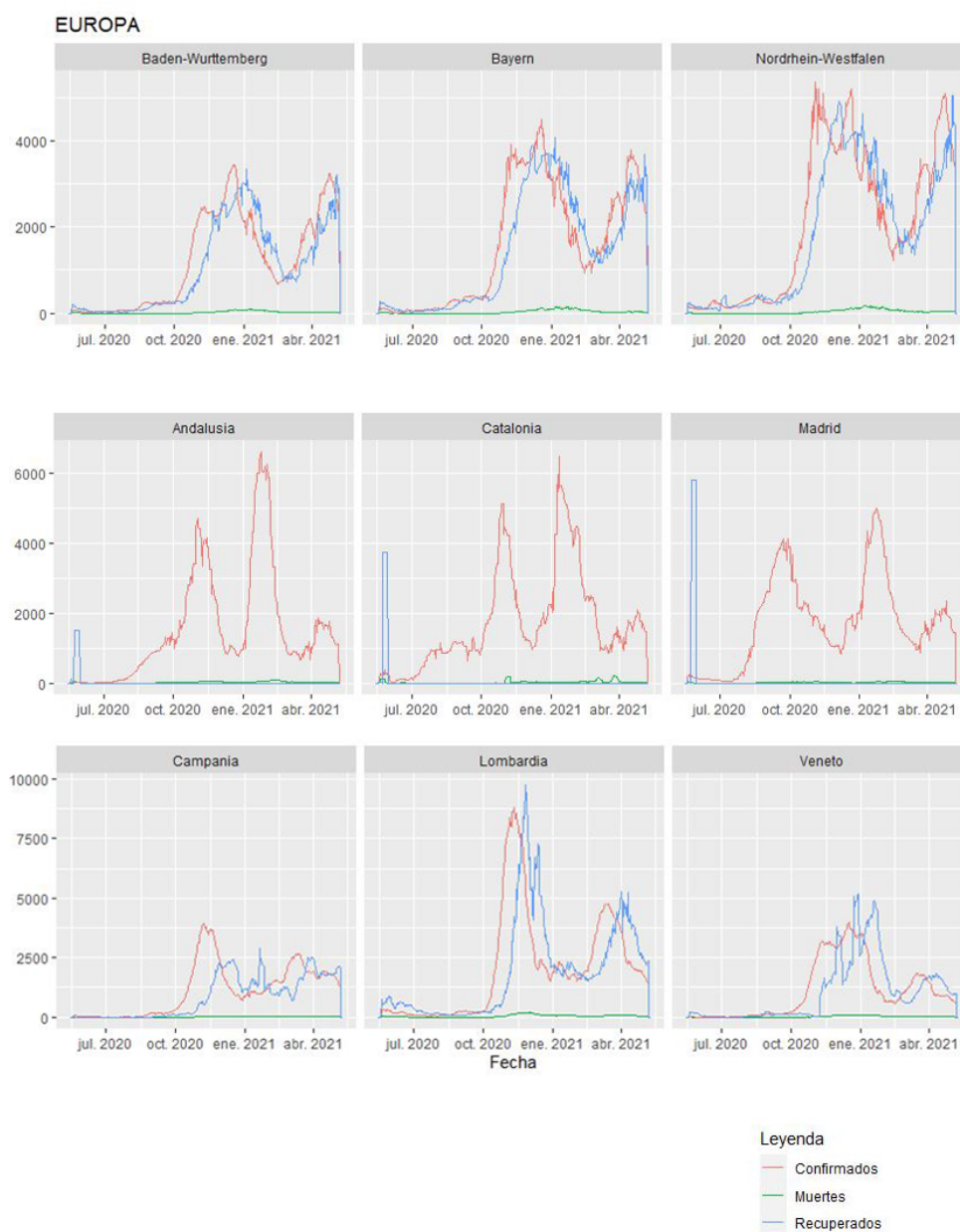


Figura 7.9: Gráfica General Regiones Europa

7.2.1. Factor de transmisión R0

Como se puede observar en la Figura: 7.10. En el clúster de color negro, podemos encontrar las regiones de Santa Catarina, Moscow, Minas Gerais, Madrid y Bayern, al comienzo del periodo presentan valores en torno a 0.9, una vez alcanzado el pico de la ola, estos valores se incrementan de forma sensible hasta un 5 %, volviendo a disminuir al finalizar la ola. El clúster de color rojo contiene a Tokyo, Osaka y México, estas regiones tienen en común, que, una vez llegada la cima de la ola, todas sus regiones adoptan un valor en torno a 2, en la segunda parte del periodo mantiene una tendencia bajista hasta alcanzar valores cercanos a 1. El clúster verde, está formado por el mayor número de regiones, estas son Veneto, Texas, Sao Paulo, Saint Petersburg, Nordrhein-Westfalen, Moscow Oblast, Karnataka y Baden-Wuttemberg, que comienzan el periodo de estudio en valores cercanos a 1, manteniendo valores similares hasta el pico de la ola y no es hasta el final del periodo, cuando todos ellos descienden progresivamente hasta a 0.9. En el clúster de color azul, podemos encontrar las regiones de Kerala, Kanagawa, Catalonia, California y Andalucía, las cuales no comparten una similitud en los datos, sino que se asemejan en la evolución durante el periodo, llegando a perder alrededor de 1.2 si comparamos su valor inicial y final. El clúster en color celeste está formado por Maharashtra, Guanajuato, Florida y Ciudad de México, estas regiones comienzan su periodo en valores similares a los del clúster verde, diferenciándose en que el crecimiento y decrecimiento de este clúster es más agresivo, alcanzando cotas más altas en el pico en torno a 1.5 y más bajas al finalizar el periodo, descendiendo aproximadamente más de un 40 %. Por último, el clúster rosa está únicamente compuesto por Lombardía y Campania, comienza el rango de estudio con valores muy superiores al del resto, en torno a 3.5, este valor presenta una tendencia bajista acentuada a lo largo de todo el periodo, finalizando en valores cercanos a 0.9.

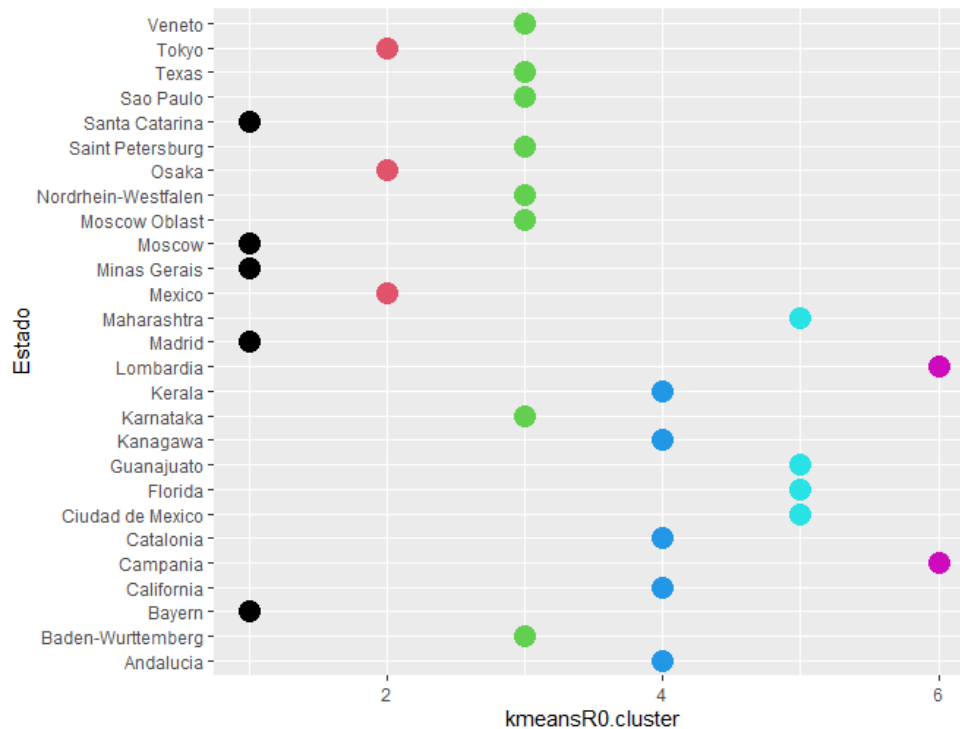


Figura 7.10: Análisis cluster por regiones de Asia, Europa y América en base a R0 y a la segunda ola

7.2.2. Índice de peligrosidad (IP)

Como se puede observar en la Figura: 7.11. El primer clúster, en color negro, está formado por las regiones de Texas, Ciudad de México, Cataluña y Andalucía, que comienzan el estudio con un índice de peligrosidad entre 2000 y 4000, alcanzando un aumento de hasta el 90 % en el centro del estudio, pasada la ola estos valores vuelven a un estado similar al inicial. El clúster rojo, contiene las regiones de Saint Petersburg, México, Madrid y Campania, este clúster mantiene en todo momento valores muy similares, fluctuando desde el comienzo hasta el final del estudio, dicha fluctuación esta acotada entre 1000 y 3000, siendo 3000 un caso único. El clúster verde está compuesto por Véneto, Moscow y Karnataka, comparte en el pico de la ola valores cercanos a 2000, sufriendo un descenso drástico y alcanzando valores inferiores a -1000. El clúster en color azul está formado por regiones como Sao Paulo y Kerala, cuyos valores iniciales son elevados, tomando valores entre 3500 y 6000, estos valores descienden de forma drástica durante todo el periodo, llegando a terminar en valores cercanos a 350. El clúster celeste está formado por Maharashtra y Lombardia, a diferencia del anterior, los valores aumentan en la primera mitad

del periodo, para descender de una forma violenta alcanzando valores inferiores a 500 e incluso negativos al finalizar el periodo. El clúster rosa, está formado por la mayoría de las regiones, como Tokyo, Santa Catarina, Osaka, Nordrhein-Westfalen, Moscow Oblast, Minas Gerais, Kanagawa, Guanajuato, Florida, California, Bayern y Baden-Wuttemberg ya que, sus valores fluctúan entre -1000 y 1000 de manera continua en todo el periodo, sin destacar ninguna tendencia creciente ni decreciente. Cabe resaltar que en este grupo, California y Florida no aportan datos sobre los recuperados, por tanto, tienen valores nulos.

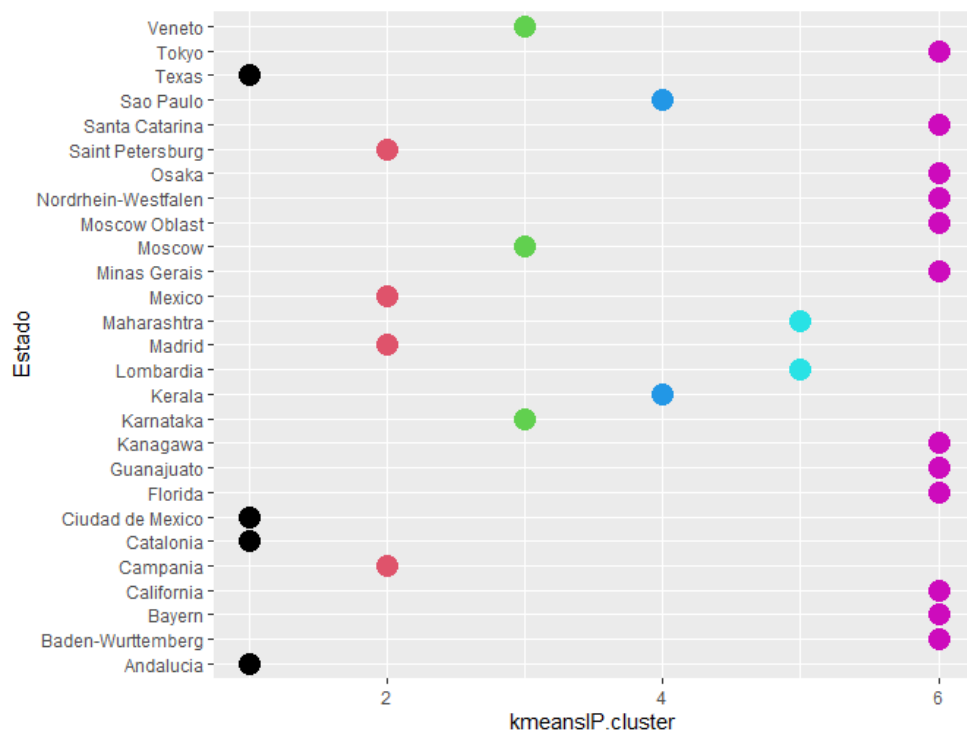


Figura 7.11: Análisis cluster por regiones de Asia, Europa y América en base a IP y a la segunda ola

7.2.3. Incidencia acumulada (IA)

Como se puede observar en la Figura: 7.12. El primer clúster, de color negro, está compuesto por Osaka, Maharashtra, Karnataka y Guanajuato, estas regiones conservan una tendencia alcista durante todo el periodo, concretamente alcanzando un aumento del 30% en la mitad del periodo, el cual, continúo aumentando un 12% adicional hasta finalizar el periodo. El clúster rojo está formado por Véneto, Texas, Saint Petersburg, Lombardía, Florida y California, en estas regiones la incidencia

acumulada aumenta hasta llegar al pico de la ola obteniendo valores parejos, y manteniendo estos valores estables hasta el final del periodo. En el clúster verde, se encuentran Santa Catarina, Moscow, Ciudad de México, Catalonia, Campania y Andalucía, la media de los valores presenta, una tendencia alcista a lo largo del periodo. El clúster azul está comprendido por Nordrhein-Westfalen, Moscow Oblast, Minas Gerais, Madrid, Kerala, Bayern y Baden-Wuttemberg, en estas regiones la media de los valores a lo largo del periodo es muy similar, teniendo una cima, en el centro del periodo un 3% superior a los extremos. El clúster celeste, contiene las regiones con valores más bajos de incidencia acumulada como Tokyo, México y Kanagawa, siendo la media del comienzo del periodo 72 y aumentando esta hasta 146 al final del periodo. Por último, el clúster rosa formado únicamente por Sao Paulo se encuentra totalmente aislado, a causa de su alta tasa de incidencia acumulada.

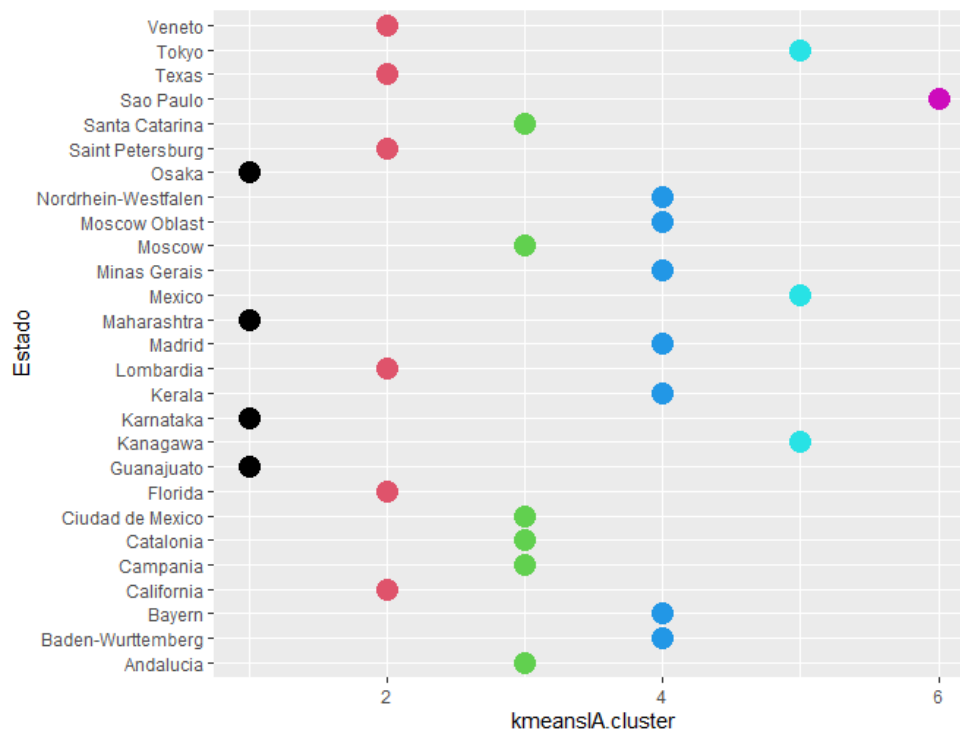


Figura 7.12: Análisis cluster por regiones de Asia, Europa y América en base a IA y a la segunda ola

7.2.4. Análisis K-Means

Tras realizar los tres análisis, encontramos una serie de regiones que comparten clúster sin importar la variable de estudio, es decir, atendiendo

a la información que nos aportan las variables estudiadas se puede concluir que han tenido un comportamiento similar en los peores momentos de la segunda ola.

- Minas Gerais (Brasil) y Bayern (Alemania).
- Nordrhein-Westfalen (Alemania), Moscow Oblast (Rusia) y Baden-Wurttemberg (Alemania).
- Cataluña y Andalucía (España).

Otra serie de regiones comparten clúster en al menos dos variables, siendo esto bastante natural en las regiones pertenecientes al mismo país, pero, por otro lado, es bastante llamativo como regiones de países lejanos como puede ser Italia y EE. UU, tienen tanta similitud.

- Santa Catarina (Brasil), Minas Gerais (Brasil), Bayern (Alemania) (R0, IP).
- Minas Gerais (Brasil), Bayern (Alemania), Moscow Oblast (Rusia) (IP IA).
- Tokio y Osaka (Japón) (R0 IP).
- Tokio (Japón) y México (México) (R0 IA).
- Tokio (Japón) y Kanagawa (Japón) (IP IA).
- Véneto (Italia) y Texas (Estados Unidos) (R0 IA).
- Véneto (Italia) y Karnataka (India) (R0 IP).
- Cataluña (España), Andalucía (España) y Ciudad de México (México) (IP IA).
- Maharashtra (India) y Guanajuato (México) (R0 IA).
- Guanajuato (México) y Florida (Estados Unidos) (R0 IP).

Capítulo 8

Contribuciones individuales

A lo largo del desarrollo del proyecto todos los integrantes del equipo han participado en cada una de las partes y fases, aun así, cada uno ha participado más en unas fases que en otras, y para que el trabajo individual quede contrastado, cada uno de los integrantes ha compartido su trabajo en este apartado.

8.1. Contribución de Iván Jesús Saugar Peinado

Tras las primeras reuniones y concretar más los diferentes objetivos del proyecto se procedió a repartir las diferentes tareas que se iban a realizar a lo largo del mismo. Por un lado, se me otorgo la tarea de revisar, corregir y estructurar la memoria del trabajo. Por otro, también se me encargo una primera recopilación de datos de los diferentes países pertenecientes al continente que tenía asignado, Asia en mi caso, (posteriormente esta información se utilizaría para la realización de un análisis). Tras buscar las estadísticas en diferentes fuentes y de cada país por separado, reunimos la información del resto de continentes y notamos que esta era muy poco homogénea, tanto en su formato, que iban desde CSV a simples hojas de documento, como en los datos que se aportaban, algunos países presentaban más datos que otros. Tras esto y encontrar una fuente de información donde se presentaban los datos de todos los países juntos, se decidió usar este nuevo repositorio como fuente principal de estadísticas a tratar. Dado mi conocimiento en Big Data y tratamiento de datos, me encargue de realizar primero el montaje de un entorno para llevar a cabo las tareas y luego de programar varios scripts que se iban a encargar de la recogida, filtrado y tratamiento de los datos que íbamos a necesitar. En primer lugar, y dado la versatilidad y el cono-

cimiento previo, decidí trabajar sobre una máquina virtual con sistema operativo Ubuntu, en el cual instalaría todas las herramientas necesarias para realizar las tareas (biblioteca Python, Scala, Spark). Dado que los usuarios no tienen por qué usar una plataforma Unix también desarrolle un par de guías para crear un entorno en Windows. Tras montar toda la infraestructura procedí a una primera fase, que consistió en un pequeño Script que únicamente filtraba y recogía los datos del dataset original, que era facilitado por la JHU, con los valores para cada día de los diferentes países. Posteriormente, procedimos a probarlo con los candidatos que se habían elegido anteriormente y compararlos con los datos que ya teníamos para ver si la información se ajustaba a lo esperado. Dada la potencia que nos proporcionaba Spark y a que la JHU subía datos a diario de todos los países, se decidió a crear un script más que recopilara los datos, por un lado, de EE. UU., que venía por separado y por otro del resto del mundo. También se observó que los datos que nos mostraban eran acumulados, habiéndose decidido que era más interesante presentarlos de forma diaria, así que, mi nueva tarea sería agregar nuevas columnas con los cálculos diarios de las estadísticas que nos interesaban. Como resultado, realice una primera versión de los scripts para los datos de EE. UU. por un lado y los del resto por otro, así como una segunda versión del ya realizado que filtraba por un país nada más. Tras diversas reuniones con el otro equipo que iba a usar los datos generados por nuestros scripts, se detectaron diversos fallos que tuve que ir subsanando, como la existencia de datos diarios negativos por incongruencias de los datos ofrecidos por la JHU, errores de formatos en los primeros conjuntos de datos del principio y que cambiaron a posteriori o datos replicados. Según avanzo el proyecto, se fueron añadiendo otras variables nuevas para los diferentes cálculos con R con el objetivo de realizar estudios aplicando el algoritmo K-Means, en el cual ayude. Tras comprobar que los scripts ya procesaban, limpiaban y presentaban los datos de forma correcta, se me encargo la tarea de documentar diferentes guías de uso, así como de generar nuevos scripts preparados para el despliegue y actualización de los datos de origen. Como resultado, se entregó un conjunto de scripts que permiten desde desplegar todo lo necesario en el equipo, como las diferentes soluciones encontradas a los problemas que fueron surgiendo durante el desarrollo de estos, como, por ejemplo, corregir datos erróneos usando los datos del día anterior o ponerlos a vacío únicamente y eliminar o no los datos de fechas repetidas, así como usar diferentes separadores para los datos (coma o punto y coma). Al darnos cuenta de que esta información podía ser de gran utilidad para investigaciones, creé un GitHub público, con el objetivo de facilitar todos los procesos desarrollados a lo largo del proyecto, con sus respectivas guías y ejemplos, con el objetivo

de ayudar a investigaciones o estudios personales además de dar pie a crear una comunidad para añadir nuevas variables o mejoras funcionales en los scripts.

8.2. Contribución de Sergio Sánchez Ortiz

En primer lugar, realicé un estudio exhaustivo sobre las bases de datos existentes a nivel mundial. Centrándome principalmente en las bases de datos que aportaban los gobiernos de América, analizando la calidad de los datos, la regularidad con la cual los actualizaban, y la facilidad que dieran los gobiernos para descargar estos datos, preferiblemente en CSV. Tras esto, se generó una tabla, con variables parametrizadas para medir y puntuar la calidad de la información de cada país, que más tarde se usaría para escoger los países sobre los cuales realizar el estudio. Una vez realizado el estudio, y utilizando RStudio generé una primera versión de la base de datos, esta se rellenó con los datos de los países que los aportaban y en caso de que los gobiernos no facilitaran su descarga, los datos de ese país eran obtenidos de la base de datos de la WHO. Esta base de datos tenía varios problemas. El más importante consistía en la falta de variables esenciales, como los recuperados, además al utilizar datos de diferentes fuentes estos tenían variables distintas, ya que algunos datos aportaban más variables que otros. Esto derivó en la decisión de utilizar únicamente la base de datos de la Universidad Johns Hopkins, puesto que solucionaba los problemas anteriormente expuestos. Para utilizar esta base de datos nos vimos en la obligación de utilizar Spark, para facilitar la recopilación de los datos ya que estos se dividían en distintos CSV. Aunque no realicé el desarrollo principal de los scripts generados, si trabajé con ellos, solucionando diversos errores como nombres erróneos de países. Una vez obtenida la base de datos, queríamos introducir nuevas variables, esto lo realicé generando distintas funciones en RStudio, cada una de ellas orientada a generar una de las nuevas variables requeridas, entre ellas el R0, las medias móviles, el IPMM y R0MM. En ese momento, cuando quisimos representar los datos en gráficas observamos que en varios días la variable Daily_Death era 0, en muchas ocasiones a causa de la falta de actualización de datos de los gobiernos. Por ejemplo, en España, los fines de semana no había contagios. Para poner solución a esto, generé una nueva función, la cual esta vez realizaba un reparto ponderado de los datos en caso de que hubiera días con cero nuevos contagios alrededor de días con contagios. Tras todo lo anterior, tocaba pasar a la generación de las tablas K-Means para estudiar las diferentes regiones, por tanto, desarrollé un script el cual, a partir del nombre, la fecha del pico y la población introducía en las tablas K-Means R0, K-Means IP y K-Means

IA una nueva fila con los valores de estas variables y en un espacio de tiempo de veinticinco días con la fecha del pico en el centro. Una vez generadas las tablas, utilice la función K-Means de RStudio para realizar el estudio de con la finalidad de realizar su estudio, buscando el número de clústeres que mejor se adaptaba a las variables de este. Por último, redacté todas las partes de la memoria las cuales incluyen a América, o hablan sobre el trabajo que realicé, y participé junto a mis compañeros en el análisis de los K-Means, cuya resolución pueden consultar tanto en resultados como en conclusiones.

8.3. Contribución de Álvaro San José Guerra

Una vez establecidos el tema principal del proyecto, los objetivos y la motivación acordamos en dividirnos el trabajo de investigación y de recolección de datos por continentes. En mi caso me tocó realizar el estudio de Europa.

En primer lugar, realicé un estudio exhaustivo de todos los datos que aportan todos los gobiernos de países de Europa, analizando la cantidad de datos que aportaban, la regularidad con la que estos aportaban datos, haciendo hincapié en comprobar si los países aportaban datos diarios y si los aportaban divididos por regiones ya que esto era de gran interés para nuestro proyecto. También valoramos altamente si los países aportaban la posibilidad de descargar sus datos en formato CSV ya que es de mucha ayuda a la hora de analizar sus datos detenidamente usando R.

Una vez analizados todos los países guardé toda la información recopilada en una tabla y tras generar en equipo un sistema de puntuaciones para valorar los datos aportados por cada gobierno use el sistema diseñado para puntuar a todos los países de Europa, con el objetivo final de seleccionar aquellos países dentro del continente que mayor puntuación obtengan para utilizarlos en el análisis haciendo uso del algoritmo de agrupación K-means.

Aunque no realice el desarrollo principal ayudé a la hora de la creación del script de R para la generación de las tablas del K-means y al estudio de las variables devueltas por la función de K-means (`betweenss` y `withinss`) para obtener el número óptimo de clústeres para nuestro estudio.

Una vez calculadas todas las tablas los diferentes K-means, me encargue de la creación de un primer análisis de ambos estudios tanto el de las regiones de España como el de las regiones de los 3 continentes. Utilizando R cree diferentes tablas para cada grupo con el objetivo de poder visualizar los datos numéricos de forma que pudiésemos valorar y

comparar los datos de forma más sencilla. Este análisis se dividió dentro de cada estudio en 3 diferentes, cada uno atendiendo al indicador estudiado, estos indicadores son R_0 , que representa el ritmo de reproducción o el factor de transmisión R_0 , IP , que es el índice de peligrosidad e IA , que representa la incidencia acumulada. Tras realizar los diferentes análisis en base a los indicadores realicé un análisis en conjunto enfocado a tratar aquellos casos en los que los países se encuentran de manera repetida en los mismos conjuntos, dando a entender que el efecto que ha tenido la pandemia en aquellos que compartan grupos en todos o la mayoría de los indicadores es muy similar.

Además de esto me encargué de realizar un estudio de las posibilidades que ofrece la librería de R `ggplot2`. Esta librería lo que ofrece son diferentes formas de representar datos en gráficas dependiendo de los tipos de datos que sean y del tipo de gráfica que se desee crear, puedes crear desde histogramas hasta gráficas lineales pasando por gráficas de puntos. Tras realizar el estudio de esta librería y sus usos, fui el encargado de realizar todas las gráficas del proyecto que representan datos obtenidos por nosotros tanto gráficas de casos confirmados, fallecimientos o recuperados entre otras como las gráficas de los diferentes K-Means.

Por último, también me he encargado de toda la documentación del proyecto y de añadir las correspondientes citas donde fuesen requeridas.

Capítulo 9

Conclusiones y Trabajo futuro

9.1. Conclusiones

Se ha generado un análisis del impacto del coronavirus en los diferentes países de cada uno de los continentes estudiados (América, Asia y Europa), para ello se han utilizado las bases de datos tanto oficiales de cada país, como generales de la WHO o la JHU, centrándonos en la cantidad de datos distintos aportados. Tras hacer los primeros estudios y buscar la mejor relación de datos que podíamos obtener con los diferentes datasets, estos se utilizaron para generar nuestro propio dataset, al cual añadimos datos diarios de todas las variables, medias móviles a 7 días de los datos e indicadores como el índice de peligrosidad o el factor de transmisión R_0 , los cuales se calcularon tanto con los datos sin tratar como con las medias móviles, y por último se realizaron dos estudios con el algoritmo de Machine learning k-Means, el primero enfocado a las zonas con más contagios de cada país, seleccionadas con el análisis previo, y el segundo, con las diferentes comunidades de España.

A lo largo del proyecto se ha utilizado R, Panda, Bash o Git y librerías de tratado como dplyr, ggplot2, tidyverse y quantreg para desarrollar la base de datos, utilizando los cursos ofrecidos por la Universidad Complutense de Madrid, impartidos por nuestras tutoras para aprender a utilizar la herramienta R, o aplicando los conocimientos de asignaturas como Cloud y Big data o Sistemas Operativos enfocadas a analizar grandes cantidades de datos, o para automatizar el tratamiento de datos respectivamente. Por último, se ha utilizado Latex para la realización de este documento. No solamente se ha aprendido a utilizar estas herramientas, sino que además se ha adquirido conocimiento acerca del cálculo de diferentes indicadores, y su importancia a nivel muestral.

Una habilidad adquirida también a lo largo del desarrollo de este trabajo consiste en la aplicación de varias técnicas analíticas, a lo largo de

proyecto se han desarrollado varios análisis y estudios que han favorecido el desarrollo de esta habilidad, ya sea, a la hora de analizar bases de datos fiables, la calidad de los propios datos o la varianza de los valores tanto a lo largo del tiempo como en los diferentes países. Por otro lado, es importante que la base de datos creada contenga la mayor cantidad de información posible, tanto en forma de diversos indicadores como en forma de exposición de estos, ya que, esta base de datos será utilizada por usuarios en busca de muy diversos tipos de información, por tanto, la disponibilidad de esos datos ha sido uno de los objetivos principales.

Por último, concluir que es necesario el aporte de todos los datos por parte de los países, para poder extraer o definir indicadores adecuados que nos permitan tener información de valor, analizar similitudes en ventanas temporales, comparar las olas de los países en determinados periodos, o también para comparar países que aplican políticas preventivas y paliativas similares, y por supuesto, para predecir y prevenir situaciones basadas en conocimientos similares.

9.2. Trabajo Futuro

Tras finalizar el proyecto, el equipo está satisfecho con los resultados obtenidos, a pesar de ser completamente funcional somos conscientes de que existe margen de mejora, el gran alcance de la base de datos realizada proporciona diversos caminos de mejora, con la obtención de nuevos indicadores o scripts de tratamiento de datos.

El repositorio GitHub es un buen punto de partida para futuras investigaciones, pudiendo evolucionar de forma no lucrativa con el apoyo de personas ajenas al proyecto inicial, ya sea, aportando mejoras funcionales con nuevos métodos o scripts para la generación y visualización de la base de datos, o incluyendo indicadores que no han sido contemplados en este proyecto, las siguientes versiones estarán enfocadas a futuras necesidades.

Capítulo 10

Conclusions and Future work

10.1. Conclusions

An analysis of the impact of the coronavirus in the different countries of each of the continents studied (America, Asia and Europe) has been generated using the official databases of each country, as well as the general databases of the WHO and the JHU, focusing on the amount of different data provided. After making the first studies and looking for the best data relationship that we could obtain with the different datasets, these were used to generate our own dataset, to which we added daily data of all the variables, 7-day moving averages of the data and indicators such as the danger index or the R_0 , which were calculated with both the raw data and the moving averages, and finally, two studies were carried out with the Machine learning k-Means algorithm, the first focused on the areas with the most contagions in each country, selected with the previous analysis, and the second, with the different communities in Spain. Throughout the project we have used R, Panda, Bash or Git and treaty libraries such as dplyr, ggplot2, tidyverse and quantreg to develop the database, using the courses offered by the Complutense University of Madrid, taught by our tutors to learn how to use the R tool, or applying the knowledge of subjects such as Cloud and Big data or Operating Systems focused on analysing large amounts of data, or to automate data processing respectively. Finally, Latex has been used to produce this document. Not only have we learned to use these tools, but we have also acquired knowledge about the calculation of different indicators, and their importance at the sample level. A skill also acquired throughout the development of this work consists of the application of various analytical techniques. Throughout the spreading Covid 19 Data Collection, Processing and Analysis project, various analyses and studies have been carried out that have favoured the development of this skill,

whether it be when analysing reliable databases, the quality of the data itself or the variance of the values both over time and in the different countries. On the other hand, it is important that the database created contains as much information as possible, both in the form of various indicators and in the form of their exposition, as this database will be used by users in search of very different types of information, therefore, the availability of such data has been one of the main objectives. Finally, to conclude that it is necessary to provide all the data from the countries in order to extract or define appropriate indicators that allow us to have valuable information, to analyse similarities in time windows, to compare the waves of countries in certain periods, or also to compare countries that apply similar preventive and palliative policies, and of course, to predict and prevent situations based on similar knowledge.

10.2. Future Work

At the end of the project, the team is satisfied with the results obtained. obtained, although it is fully functional, we are aware that there is room for improvement. that there is room for improvement, the large scope of the database database provides several ways of improvement, with new indicators or data processing scripts. new indicators or data processing scripts. The GitHub repository is a good starting point for future research, and can evolve in a non-profit way with the support of people outside the from people outside the initial project, either by contributing functional improvements, new methods or scripts for with new methods or scripts for database generation and visualisation, or by including the database, or indicators that have not been contemplated in this project. The following versions will be focused on future needs.

Bibliografía

- [1] UNESCO. La campaña "La Nueva Normalidad" de la UNESCO. 6 2020. <https://es.unesco.org/campaign/nextnormal>.
- [2] Organización Mundial de la Salud. Neumonía de causa desconocida – China .1 2020. <http://www.who.int/csr/don/05-january-2020-pneumonia-of-unkown-cause-china/es/>.
- [3] Organización Mundial de la Salud. Declaración sobre la segunda reunión del Comité de Emergencias del Reglamento Sanitario Internacional (2005) acerca del brote del nuevo coronavirus (2019-nCoV), 1 2020.
- [4] Organización Mundial de la Salud. Alocución de apertura del Director General de la OMS en la rueda de prensa sobre la COVID-19, 3 2020.
- [5] Álvaro Sánchez. Coronavirus: Cuando el confinamiento golpeó la economía | Economía | EL PAÍS. 12 2020. <https://elpais.com/economia/2020-12-30/cuando-el-confinamiento-golpeo-la-economia.html>.
- [6] BBC News Mundo. Coronavirus: así son los hospitales de campaña en el mundo para tratar a los pacientes con covid-19. 3 2021. <https://www.bbc.com/mundo/noticias-internacional-52098615>.
- [7] Tung Thanh Le, Zacharias Andreadakis, Arun Kumar, Raúl Gómez Román, Stig Tollefsen, Melanie Saville, and Stephen Mayhew. The COVID-19 vaccine development landscape, 5 2020.
- [8] La Nueva España. Madrid prevé mantener algunas restricciones por el Covid "todo el verano". 6 2020. <https://www.lne.es/espana/2020/06/13/madrid-preve-mantener-restricciones-covid-14514238.html>.
- [9] BBC News Mundo. Vacuna contra el coronavirus: en qué etapa está el resto de vacunas contra el coronavirus además de la de Pfizer. 11 2020. <https://www.bbc.com/mundo/noticias-54879489>.

- [10] BBC News Mundo. Vacuna contra el coronavirus: Pfizer asegura que la suya es eficaz en más de un 90 % mientras los expertos piden cautela. 11 2020. <https://www.bbc.com/mundo/noticias-54873572>.
- [11] Comisión Europea. European Commission-Press release Coronavirus: Commission approves contract with BioNTech-Pfizer alliance to ensure access to a potential vaccine. Technical report, 11 2020. https://ec.europa.eu/commission/presscorner/detail/en/ip_20_2081.
- [12] Mallorca Diario. La vacunación mundial contra el coronavirus en tiempo real: 44,11 dosis por cada cien habitantes en España. 5 2021. <https://www.mallorcadiarario.com/mapa-vacunacion-coronavirus-en-tiempo-real>.
- [13] La Prensa. Así te contamos el avance del coronavirus a nivel mundial este lunes , 3 2020.
- [14] Yury A. Jimenez, Victoria Lopez, and Walter Chanava. Identification of outliers through classification based on evolution indices and their application to covid-19. *CAEPIA-TAMIDA-2021, X Symposium of Theory and Applications of Data Mining-Proceedings of the XIX Conference of the Spanish Association for Artificial Intelligence*, 2021.
- [15] V. Lopez and M. Čukić. The comparison of trends in spain and the nederland: a dynamical compartment model of the transmission of coronavirus. 2020.
- [16] V. Lopez and M. Čukić. A dynamical model of sars-cov-2 based on people flow networks. *Safety Science*, 134:105034, 2021.
- [17] Organización Mundial de la Salud. WHO Coronavirus (COVID-19) Dashboard.2021. <https://covid19.who.int/>.
- [18] COVID-19 Map - Johns Hopkins Coronavirus Resource Center. 2021. <https://coronavirus.jhu.edu/map.html>.
- [19] Repositorio de GitHub. 2021. <https://github.com/isaugar/COVID-19-UCM>.
- [20] Adrián García Sánchez, Harold Luis Pascua Cajucom, and Dale Francis Valencia Calicdan. Development of a dashboard for the evolution of covid-19, <https://eprints.ucm.es/id/eprint/67033/>, <https://sites.google.com/ucm.es/coda>. *eprints Universidad Complutense de Madrid*, 2021.

- [21] Fondo Monetario Internacional. 4 2021. <https://www.imf.org/external/datamapper/LP@WEO/OEMDC/ADVEC/WEOWORLD>.
- [22] División de Estadísticas de Naciones Unidas. 2021. <https://unstats.un.org/unsd/methodology/m49/>.
- [23] CDC COVID Data Tracker. 2021. <https://covid.cdc.gov/covid-data-tracker/index.html#county-view>.
- [24] Gobierno de Canadá. 2021. <https://www.canada.ca/en.html>.
- [25] Gobierno de México. 2021. <https://datos.covid-19.conacyt.mx/#DownZCSV>.
- [26] Ministerio de Salud de Costa Rica. 2021. <https://geovision.uned.ac.cr/oges/>.
- [27] Datos Abiertos del Ministerio de Salud. Covid-19. Casos registrados en la República Argentina. 2021. <http://datos.salud.gob.ar/dataset/covid-19-casos-registrados-en-la-republica-argentina>.
- [28] Ministerio de Salud de Colombia. 2021. https://www.minsalud.gov.co/salud/publica/PET/Paginas/Covid-19_copia.aspx.
- [29] Ministerio de Salud de Brasil. 2021. https://susanalitico.saude.gov.br/extensions/covid-19_html/covid-19_html.html.
- [30] Brief report: New Variant Strain of SARS-CoV-2 Identified in Travelers from Brazil. 2021. <https://doi.org/10.1101/2020.12.31.425021>.
- [31] Wikipedia. Asia. 2021. <https://es.wikipedia.org/wiki/Asia>.
- [32] Gobierno de Rusia. 2021. <https://xn--80aesfpebagmflc0a.xn--p1ai/information/>.
- [33] Gobierno de Kazajistán. 2021. <https://www.coronavirus2020.kz/kz>.
- [34] Gobierno de Tayikistán. 2021. <https://covid.tj/>.
- [35] Gobierno de Irak. 2021. <https://coronavirus.iq/>.
- [36] Gobierno de Israel. 2021. https://datadashboard.health.gov.il/COVID-19/general?utm_source=go.gov.il&utm_medium=referral.

- [37] Turkey: WHO Coronavirus Disease (COVID-19) Dashboard With Vaccination Data. 2021. <https://covid19.who.int/region/euro/country/tr>.
- [38] Organización Mundial de la Salud. India: WHO Coronavirus Disease (COVID-19) Dashboard With Vaccination Data. 2021. <https://covid19.who.int/region/searo/country/in>.
- [39] Gobierno de Pakistan. 2021. <https://covid.gov.pk/>.
- [40] Gobierno de Japón. 2021. https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000164708_00001.html.
- [41] Gobierno de Indonesia. 2021. <http://covid19.bnpb.go.id/>.
- [42] Gobierno de Malasia. 2021. <http://covid-19.moh.gov.my/>.
- [43] Manuel Tori. Coronavirus Italia: Italia, el foco de la pandemia en Europa, un año después del primer caso | Público, 2 2021.
- [44] James Gallagher. Coronavirus en Reino Unido: la variante del virus que pone en riesgo los planes de reapertura del país. 5 2021. <https://www.bbc.com/mundo/noticias-57131055>.
- [45] Instituto Robert Koch de Alemania. 2021 https://www.rki.de/DE/Home/homepage_node.html.
- [46] Austria: WHO Coronavirus Disease (COVID-19) Dashboard With Vaccination Data. 2021. <https://covid19.who.int/region/euro/country/at>.
- [47] Gobierno de Francia. 2021. <https://www.gouvernement.fr/info-coronavirus/carte-et-donnees>.
- [48] Ministerio de Salud de Holanda. 2021. <https://www.rivm.nl/en/coronavirus-covid-19>.
- [49] Epistat – COVID-19 Belgian Dashboard. 2021. <https://epistat.wiv-isp.be/covid/covid-19.html>.
- [50] Gobierno de Ucrania. 2021. <https://covid19.gov.ua/en/>.
- [51] John A Hartigan. Clustering algorithms. 1937. <https://archive.org/details/clusteringalgori0000hart/page/n5/mode/2up>.

Anexo Gráficas

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

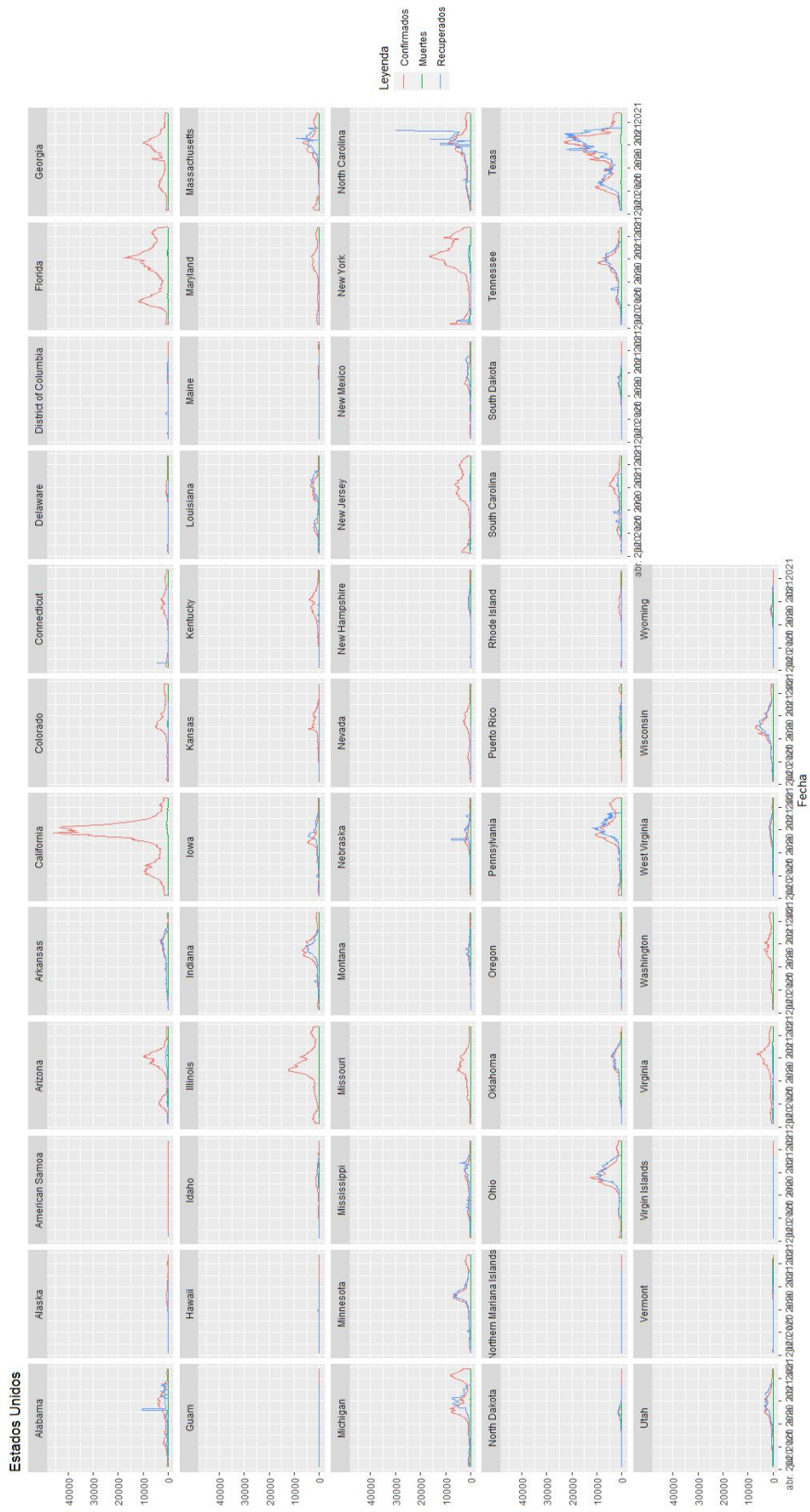


Figura 1: Curva contagios EEUU

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos



Figura 2: Curva contagios México

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

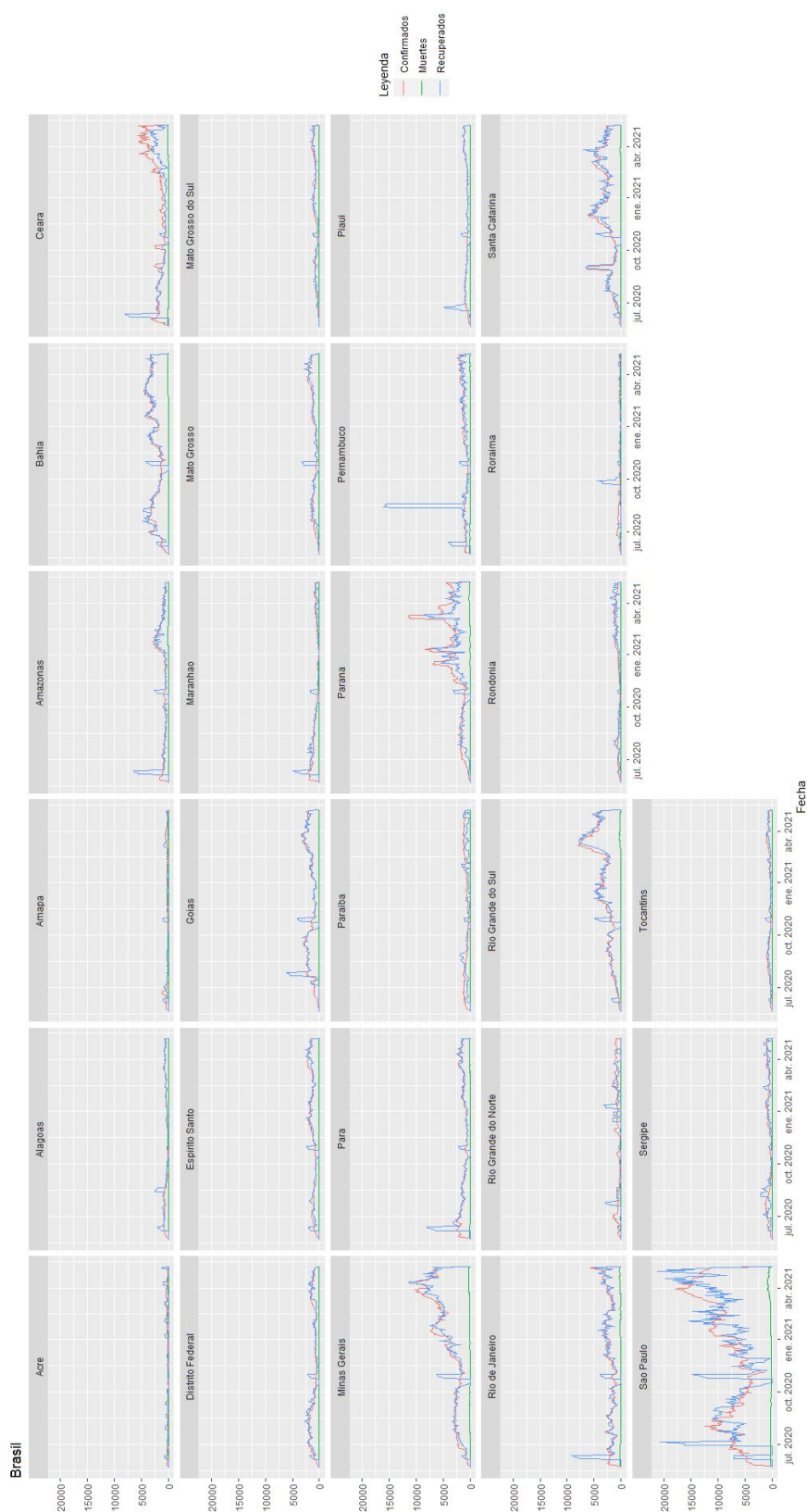


Figura 3: Curva contagios Argentina

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

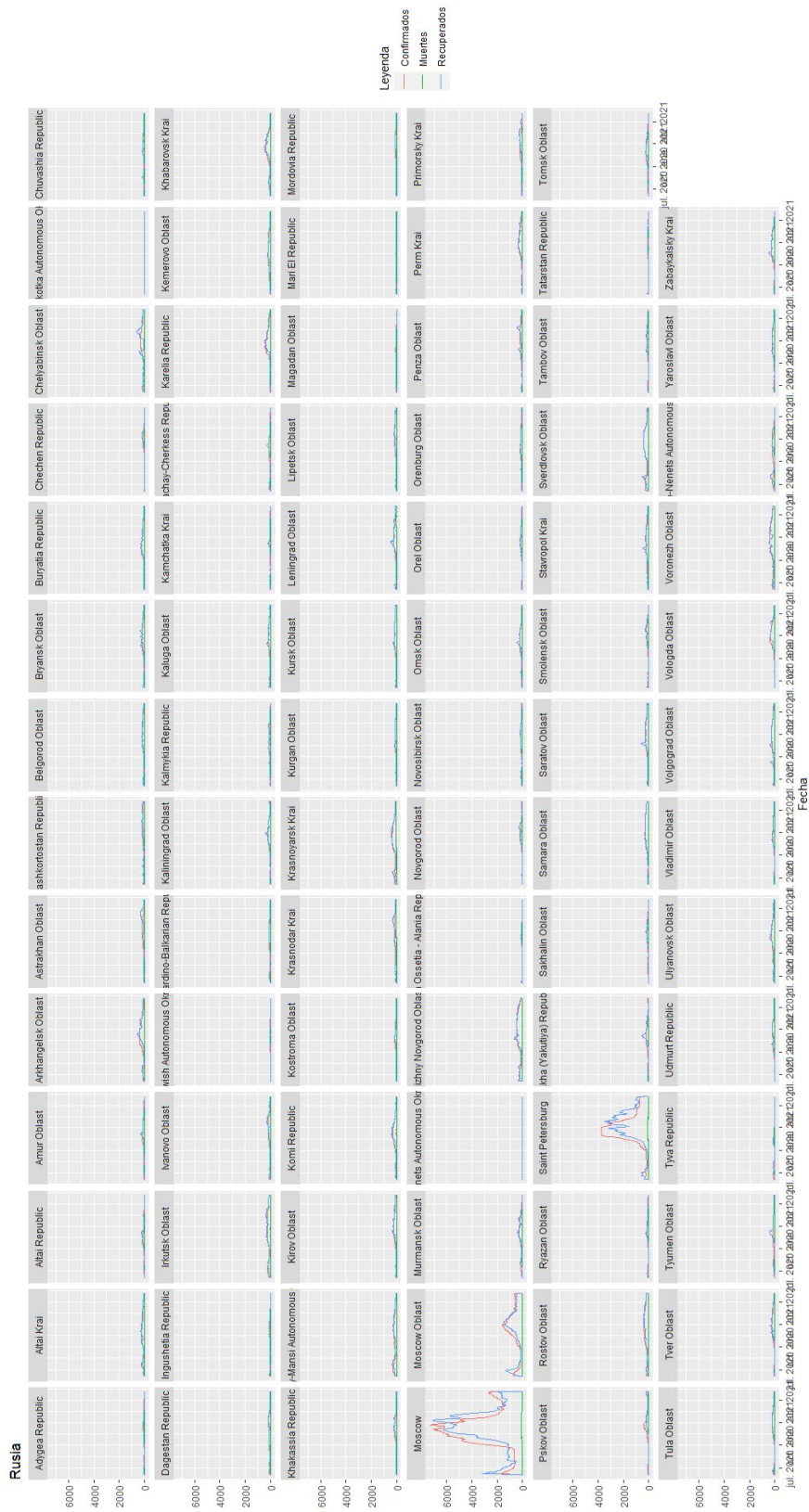


Figura 4: Curva contagios Rusia

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

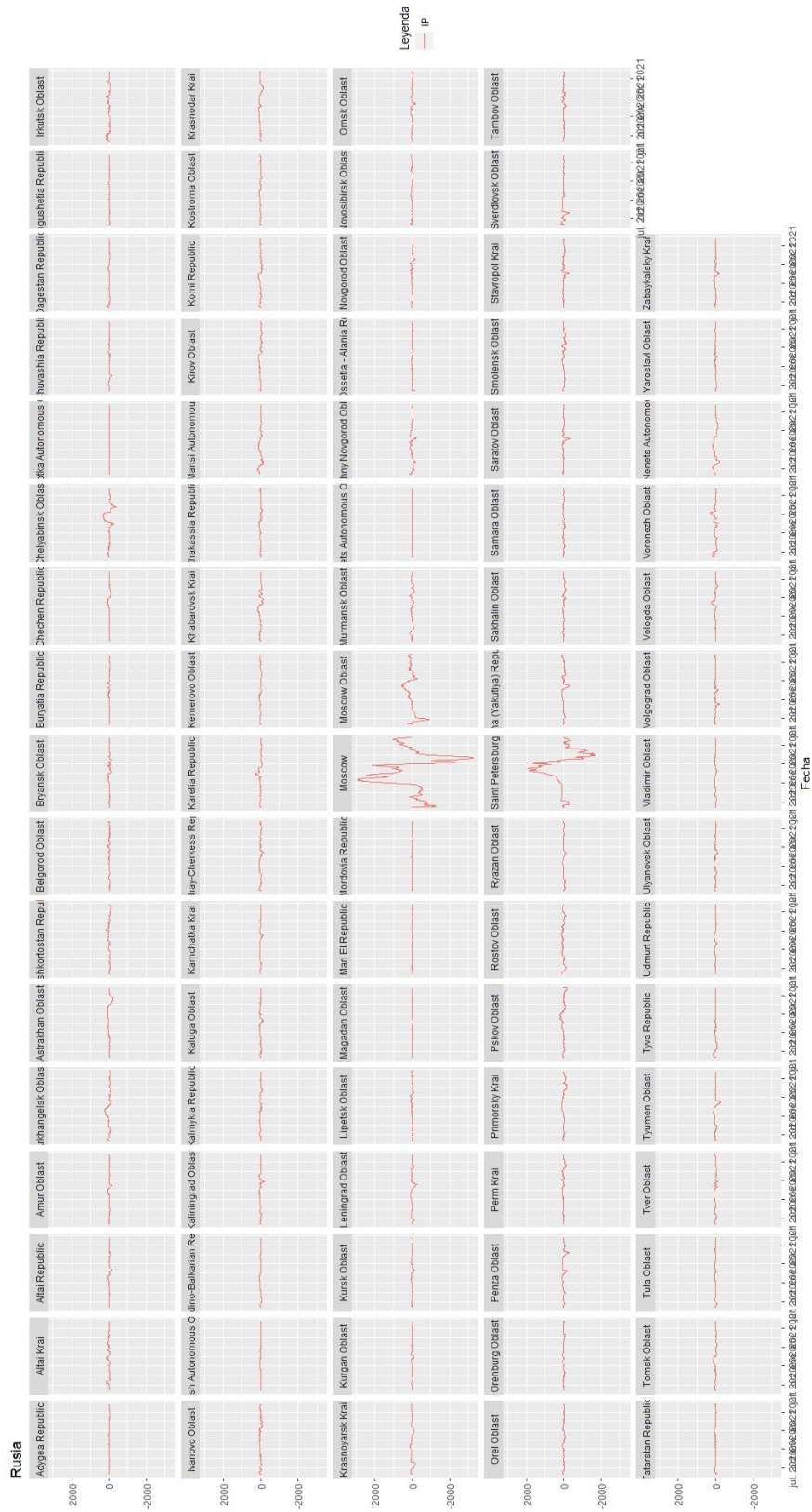


Figura 5: Curva IP Rusia

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

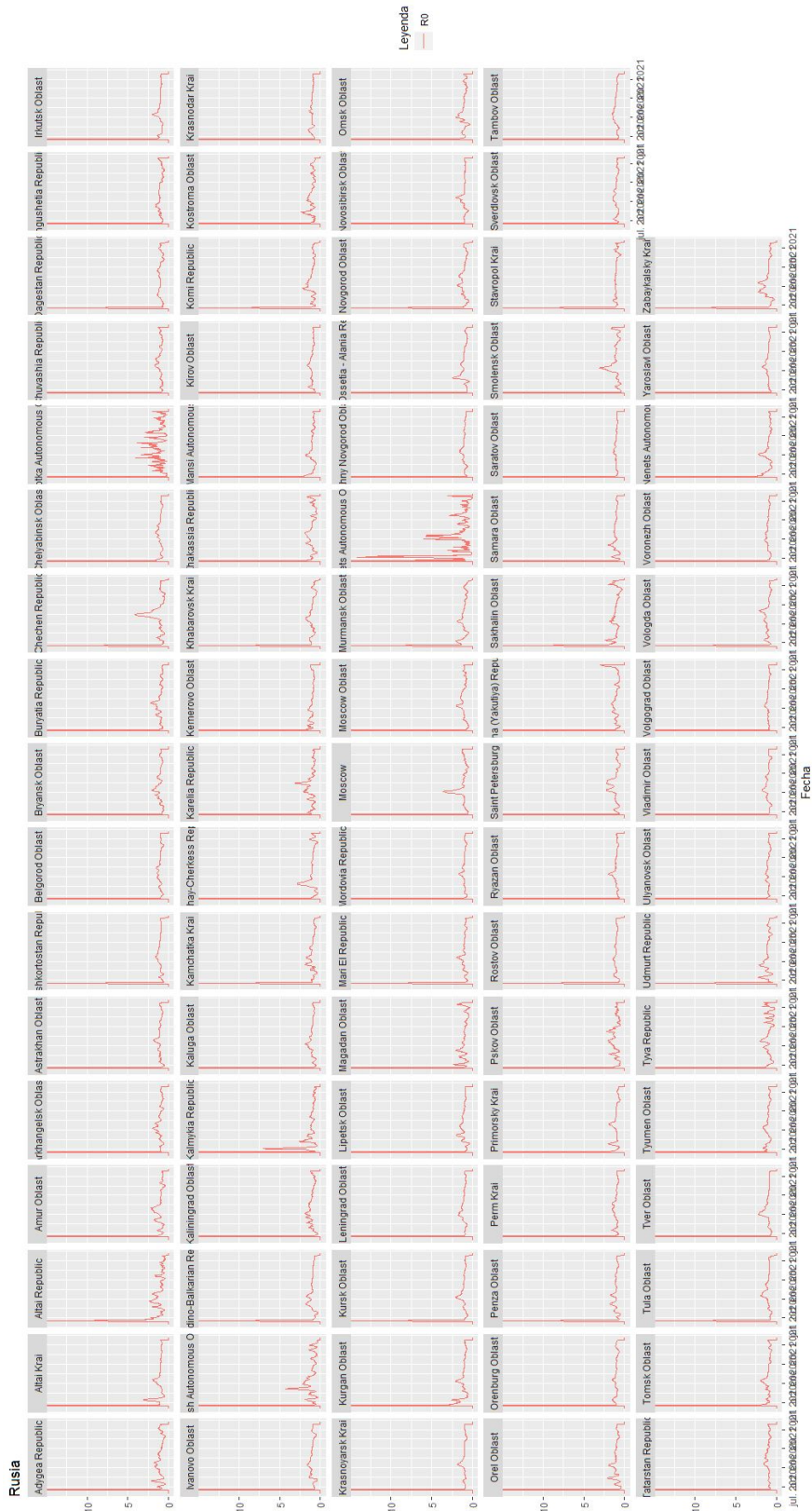


Figura 6: Curva R0 Rusia

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

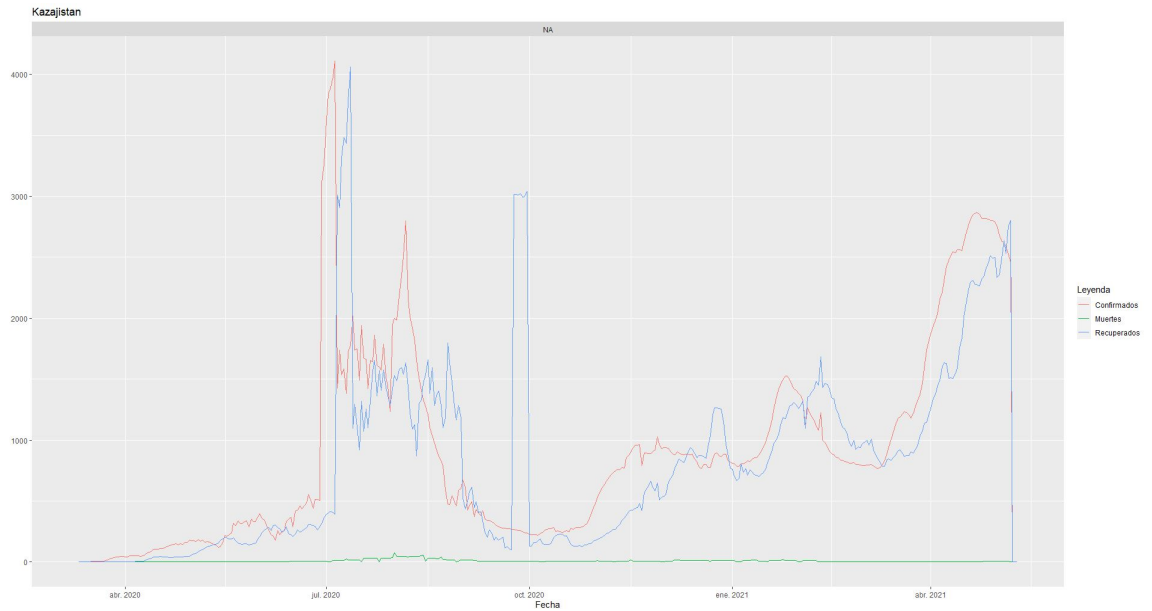


Figura 7: Curva contagios Kazajistán

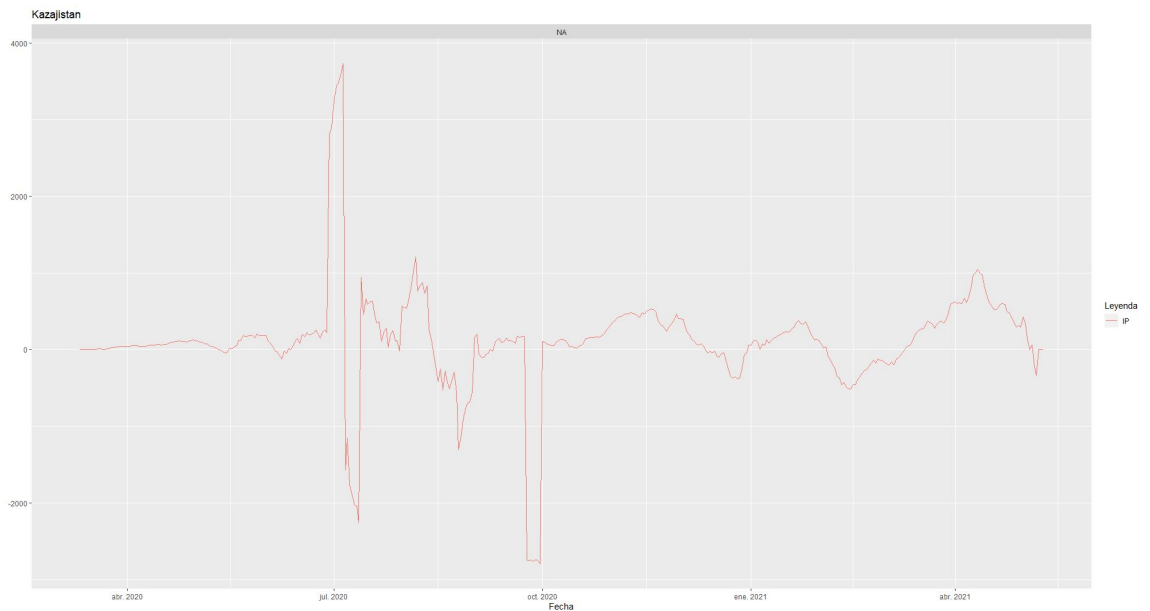


Figura 8: Curva IP Kazajistán

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

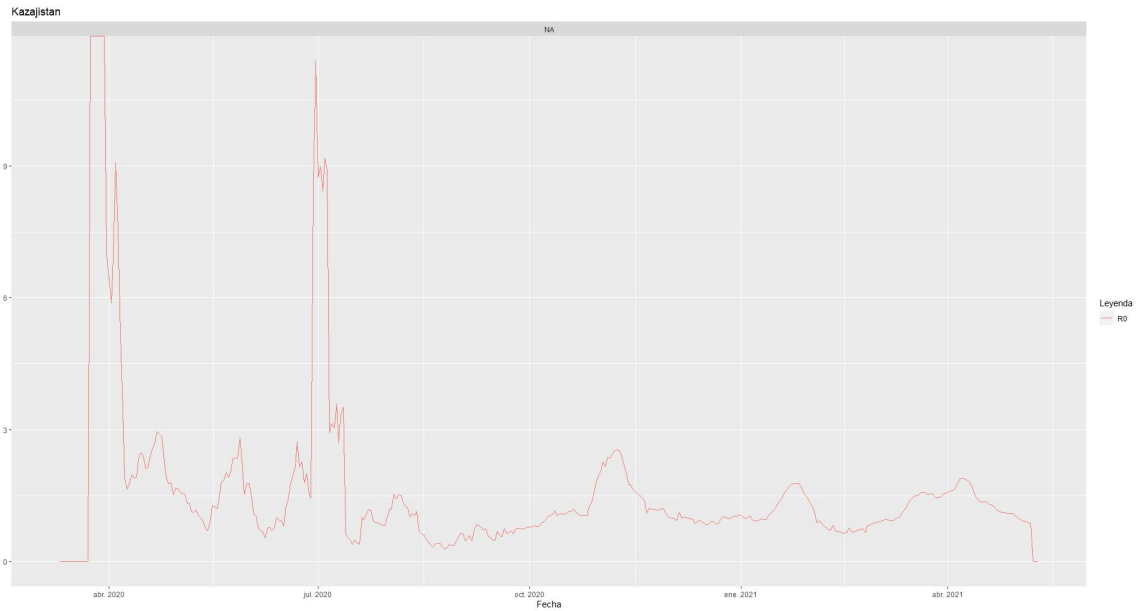


Figura 9: Curva R0 Kazajistán

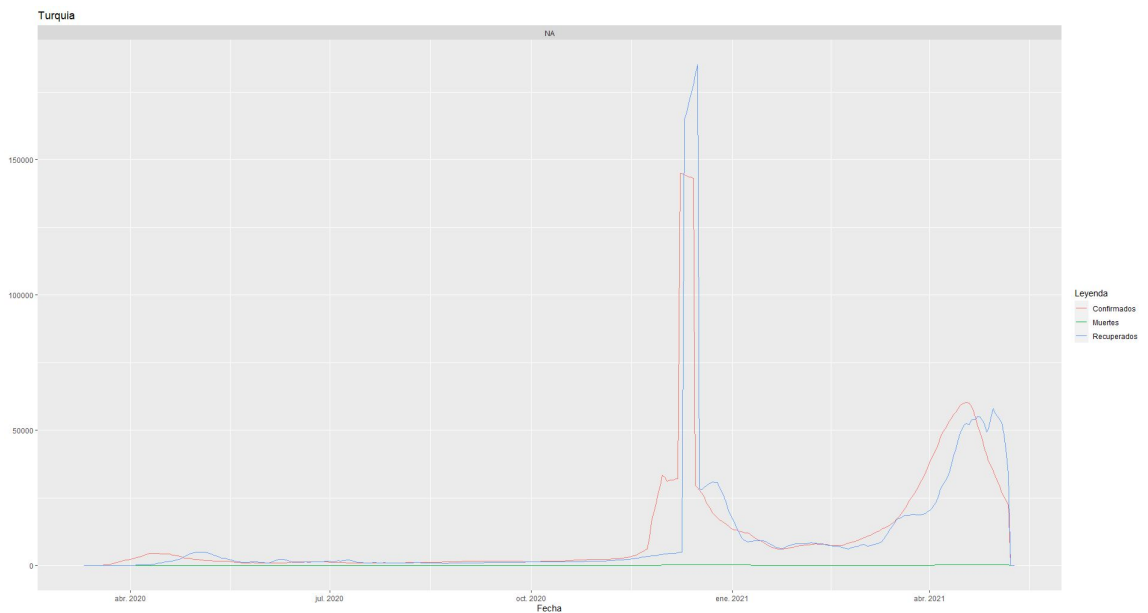


Figura 10: Curva contagios Turquía

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

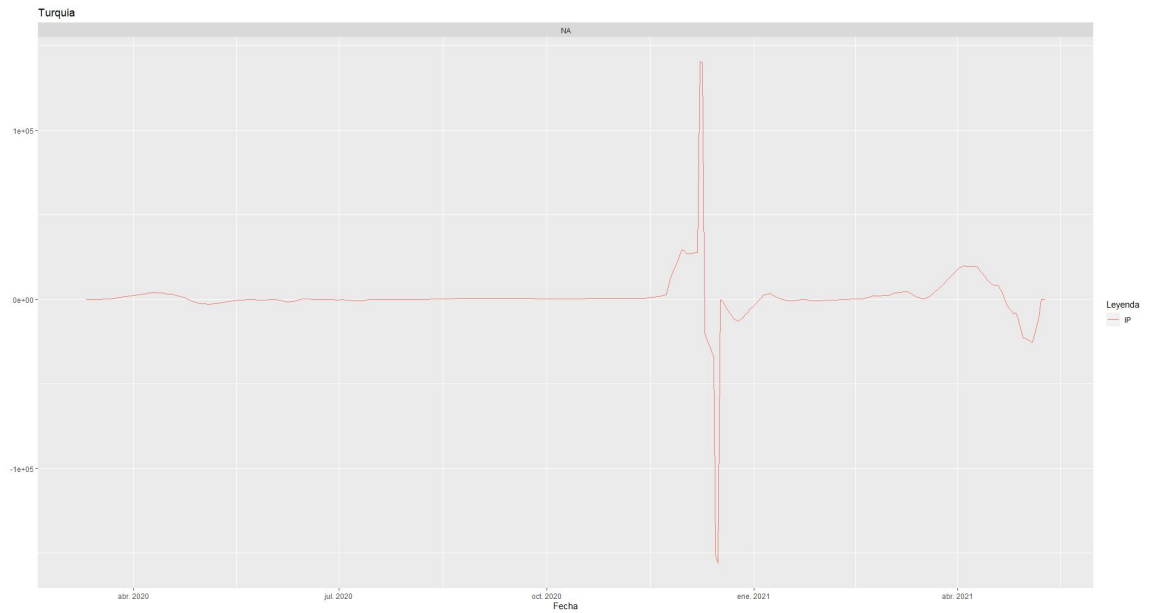


Figura 11: Curva IP Turquía

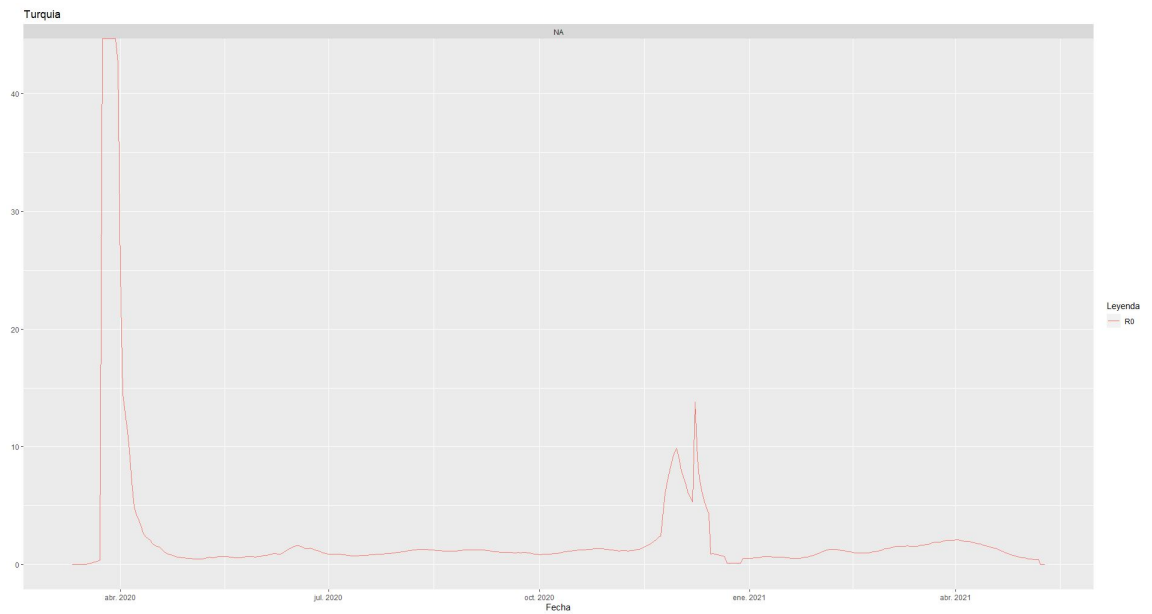


Figura 12: Curva R0 Turquía

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos



Figura 13: Curva contagios India
103

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

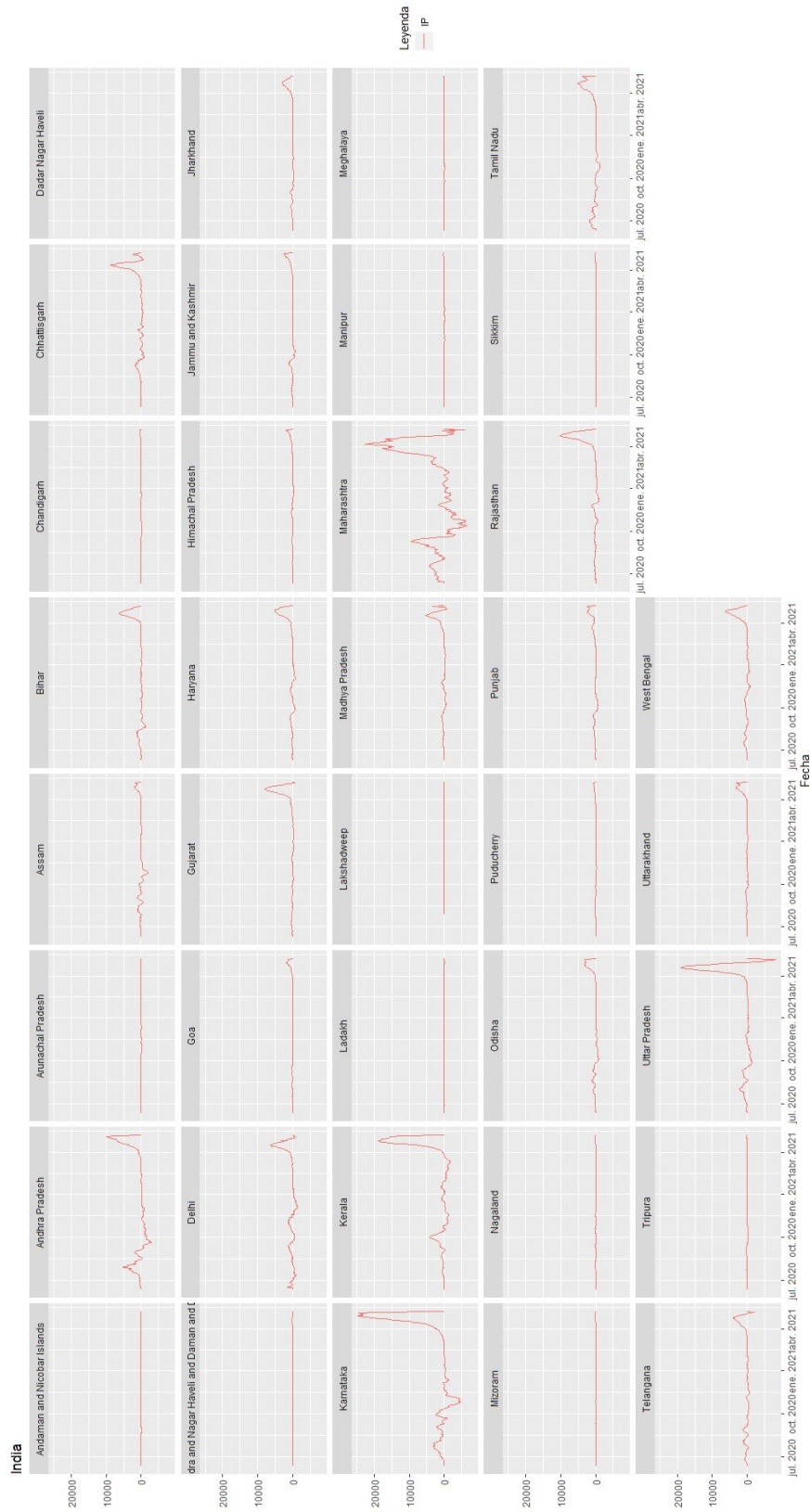


Figura 14: Curva IP India
104

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

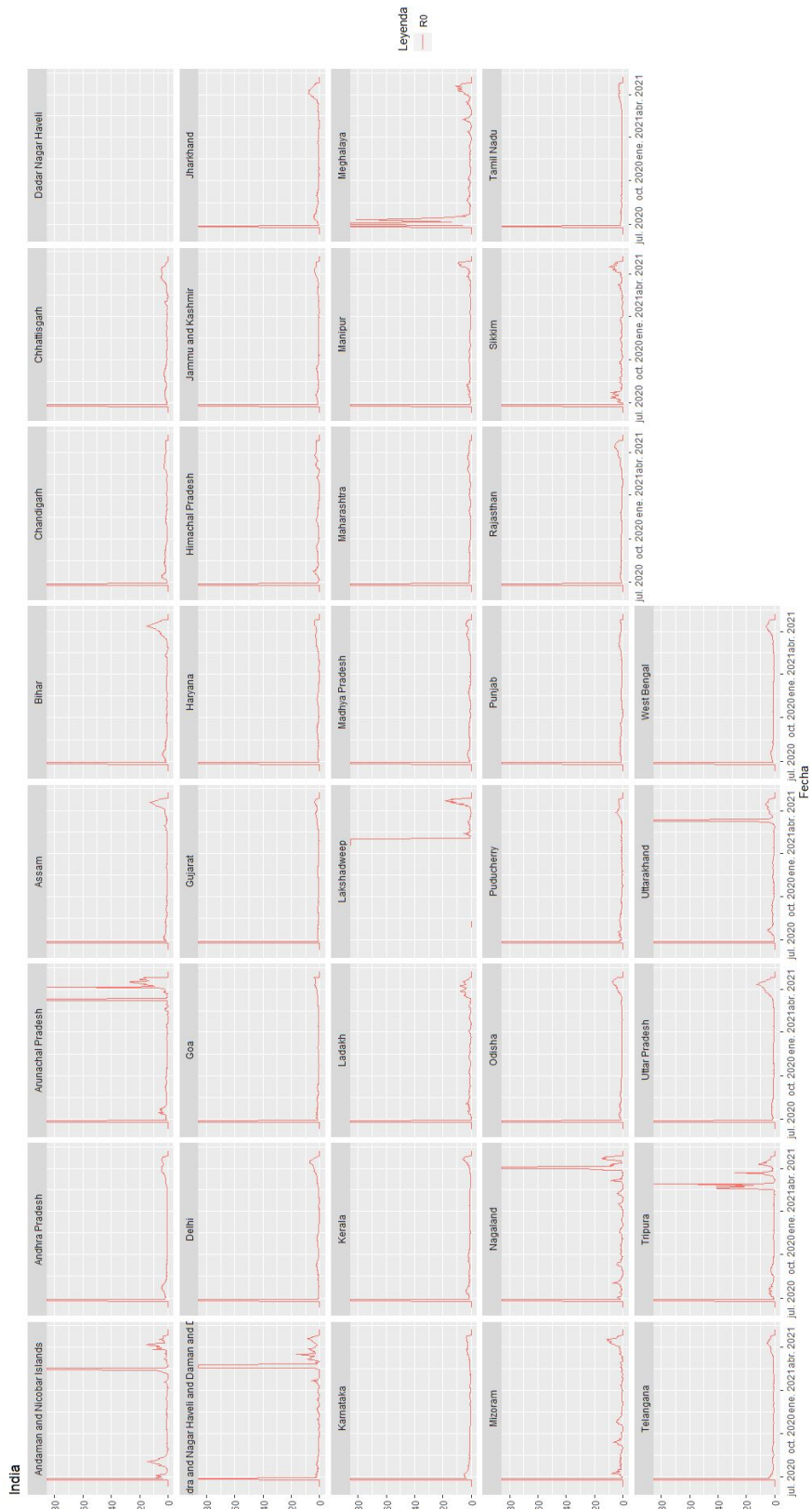


Figura 15: Curva R0 India
105

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

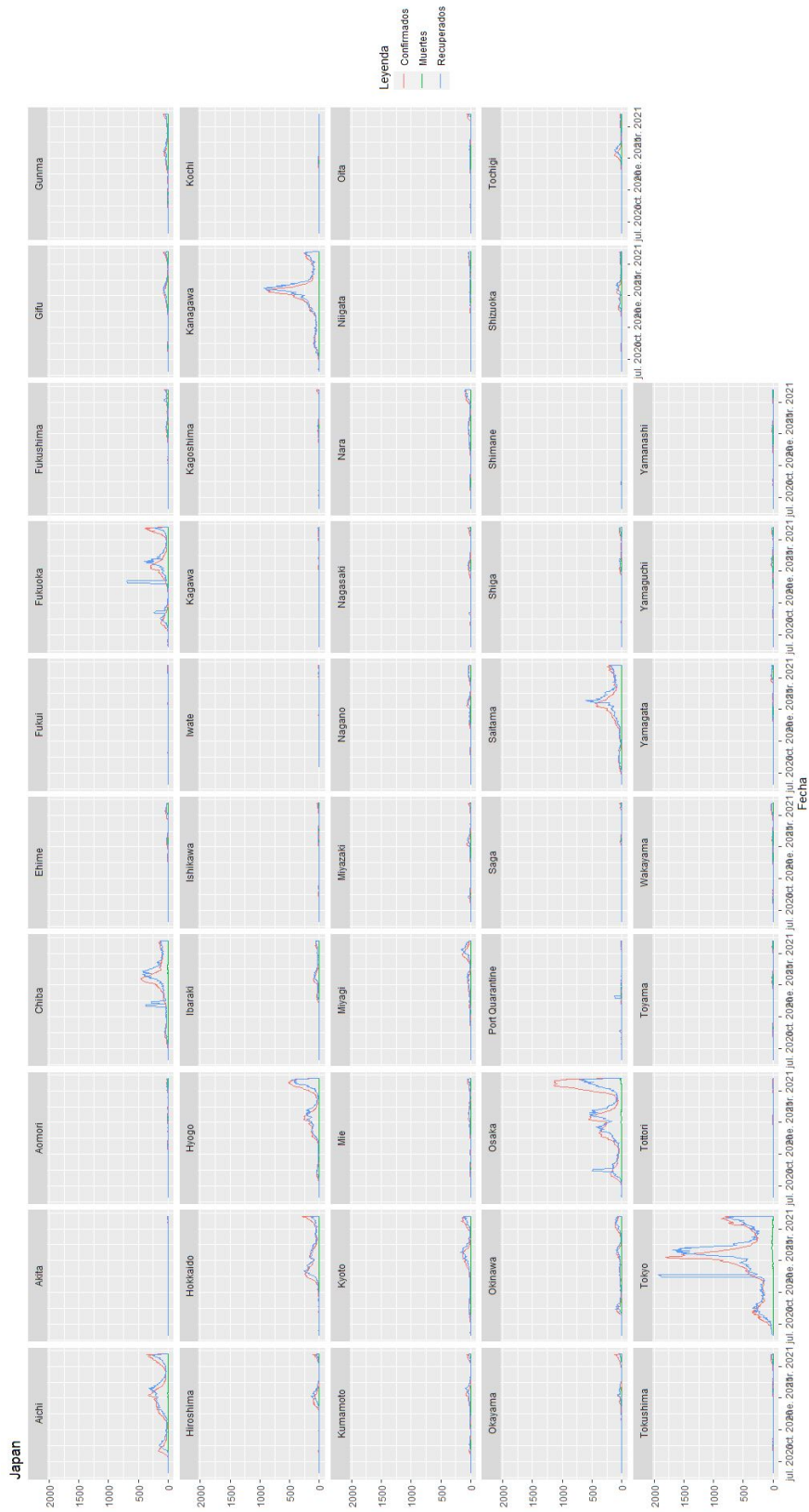


Figura 16: Curva contagios Japón

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

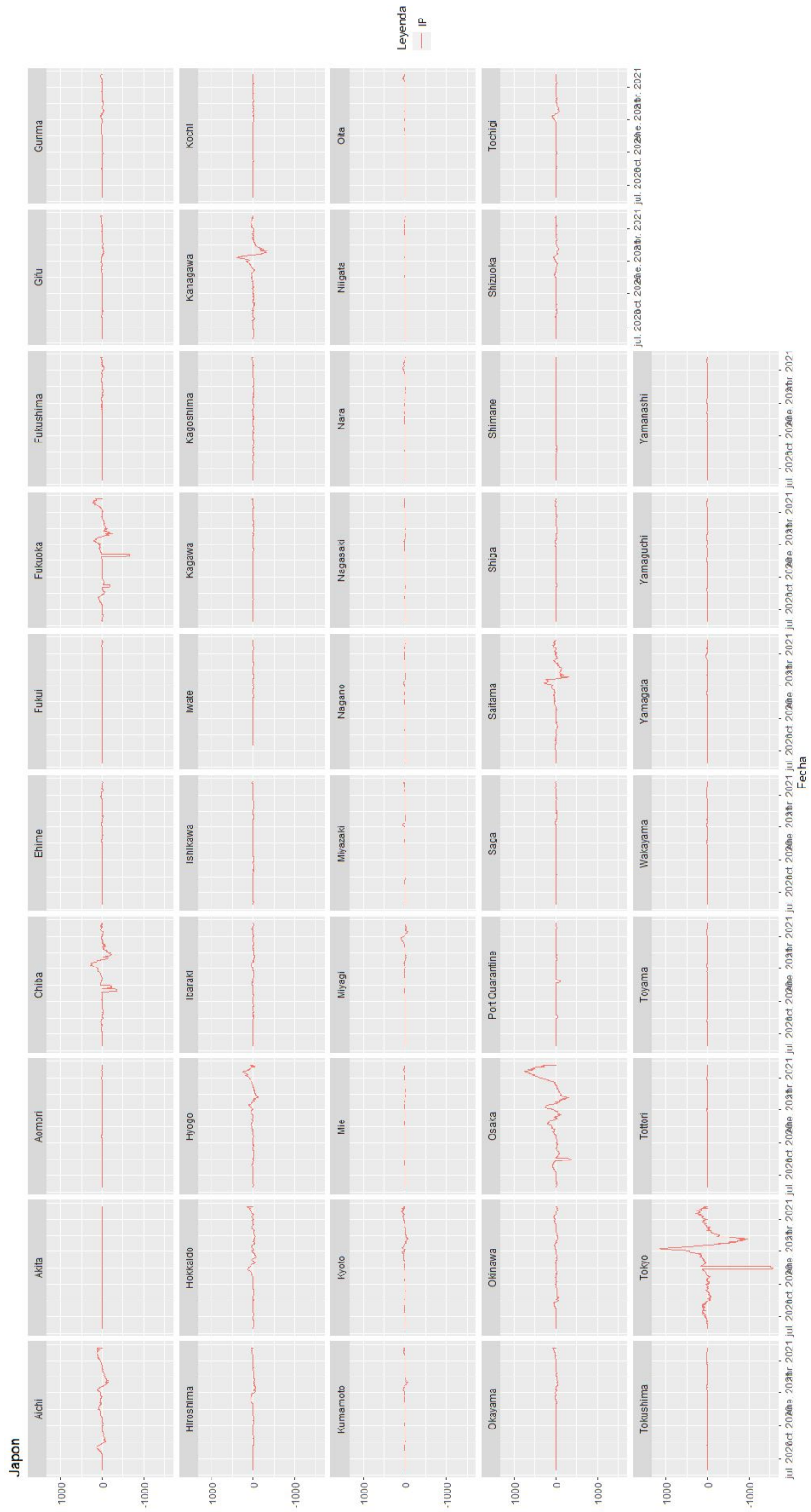


Figura 17: Curva IP Japón
107

Propagación Covid 19 Recolección, Tratamiento y Análisis de los datos

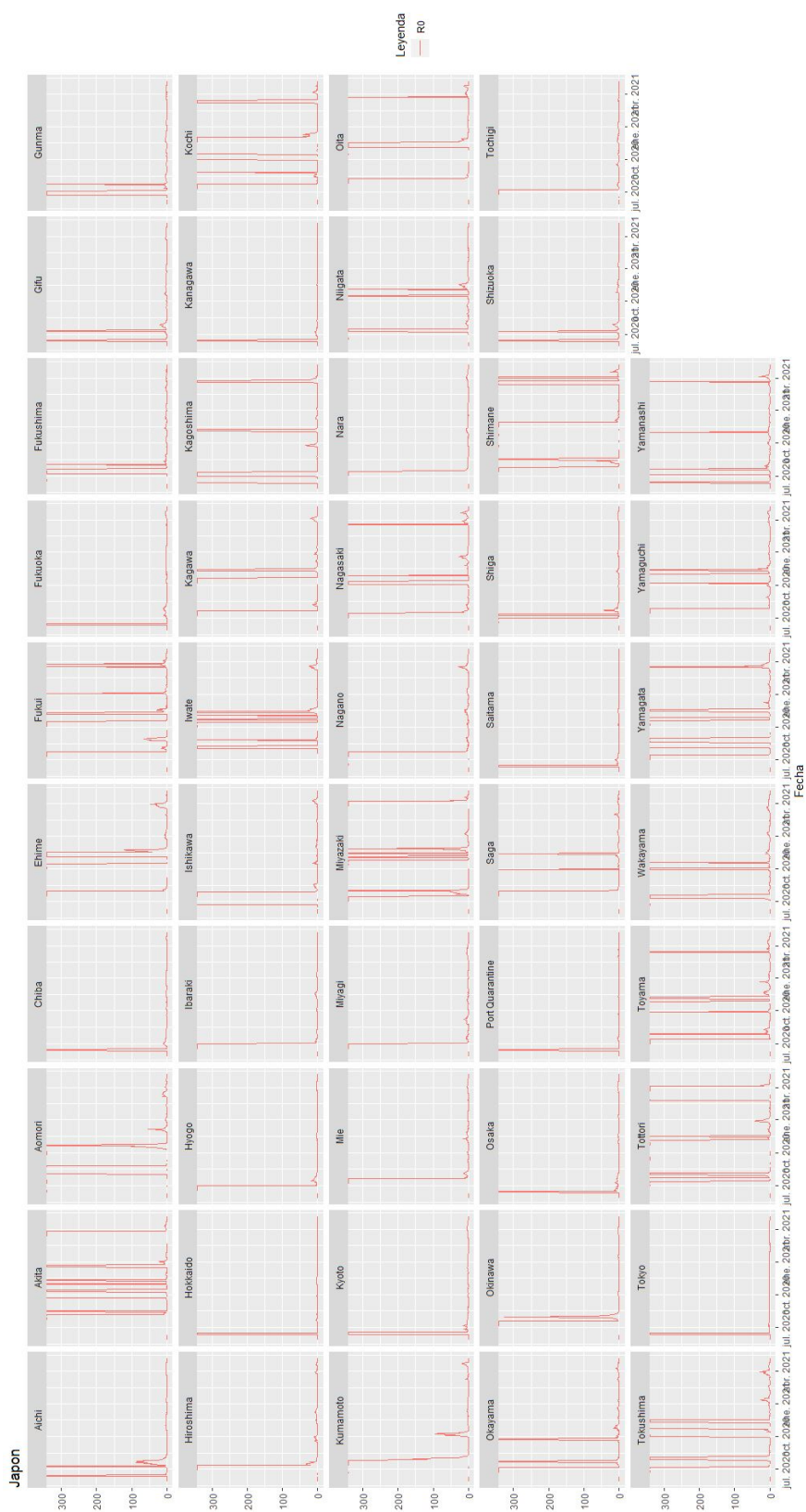


Figura 18: Curva R0 Japón

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

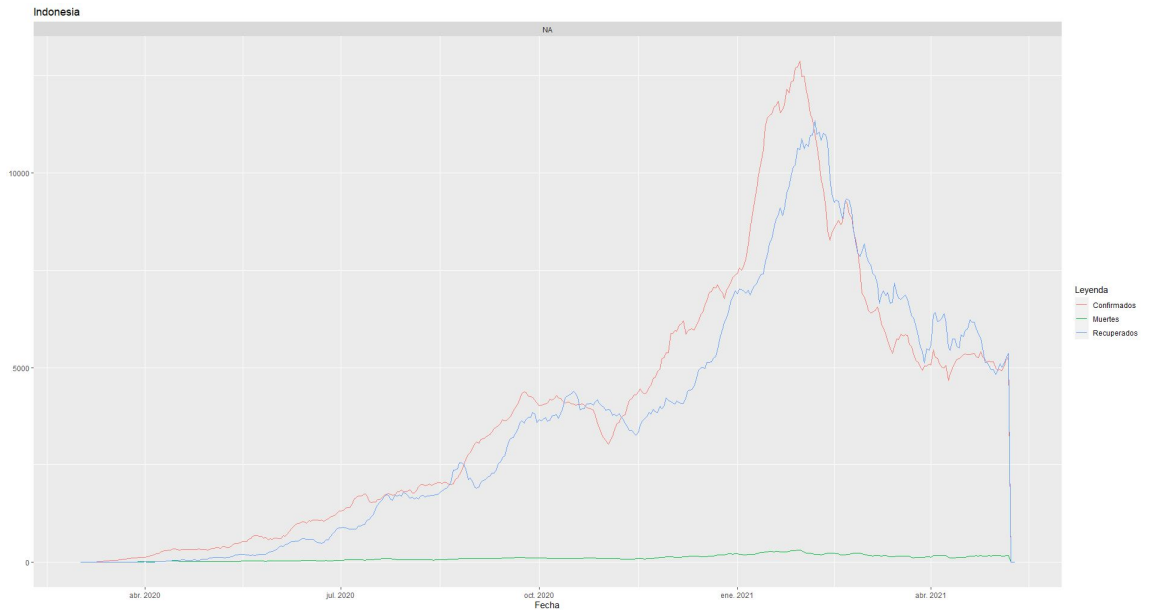


Figura 19: Curva contagios Indonesia

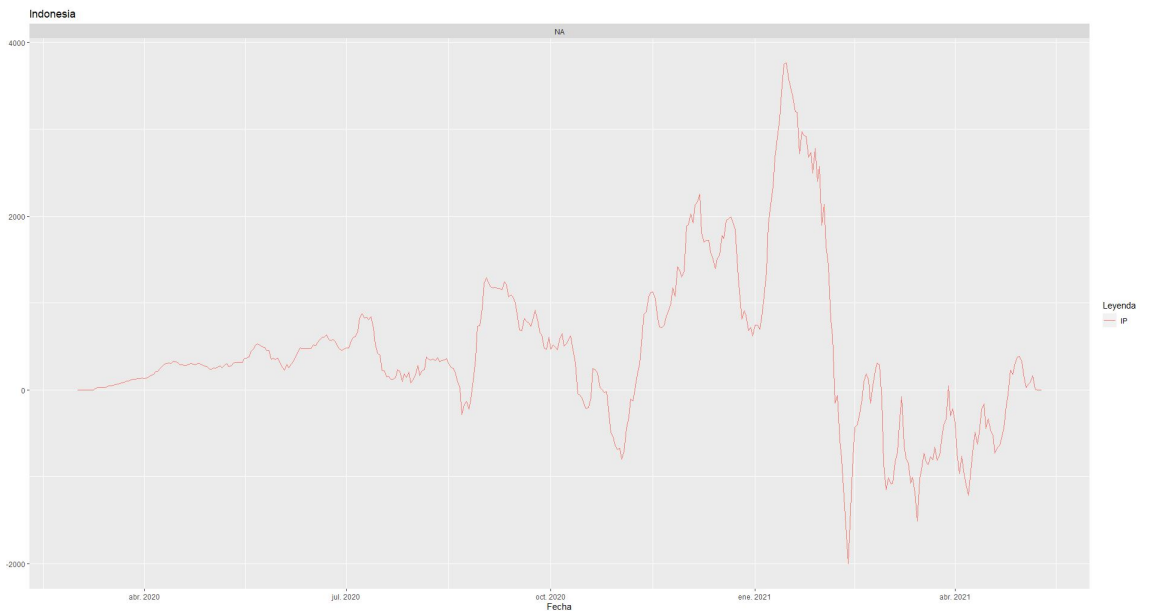


Figura 20: Curva IP Indonesia

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos



Figura 22: Curva contagios España
111

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

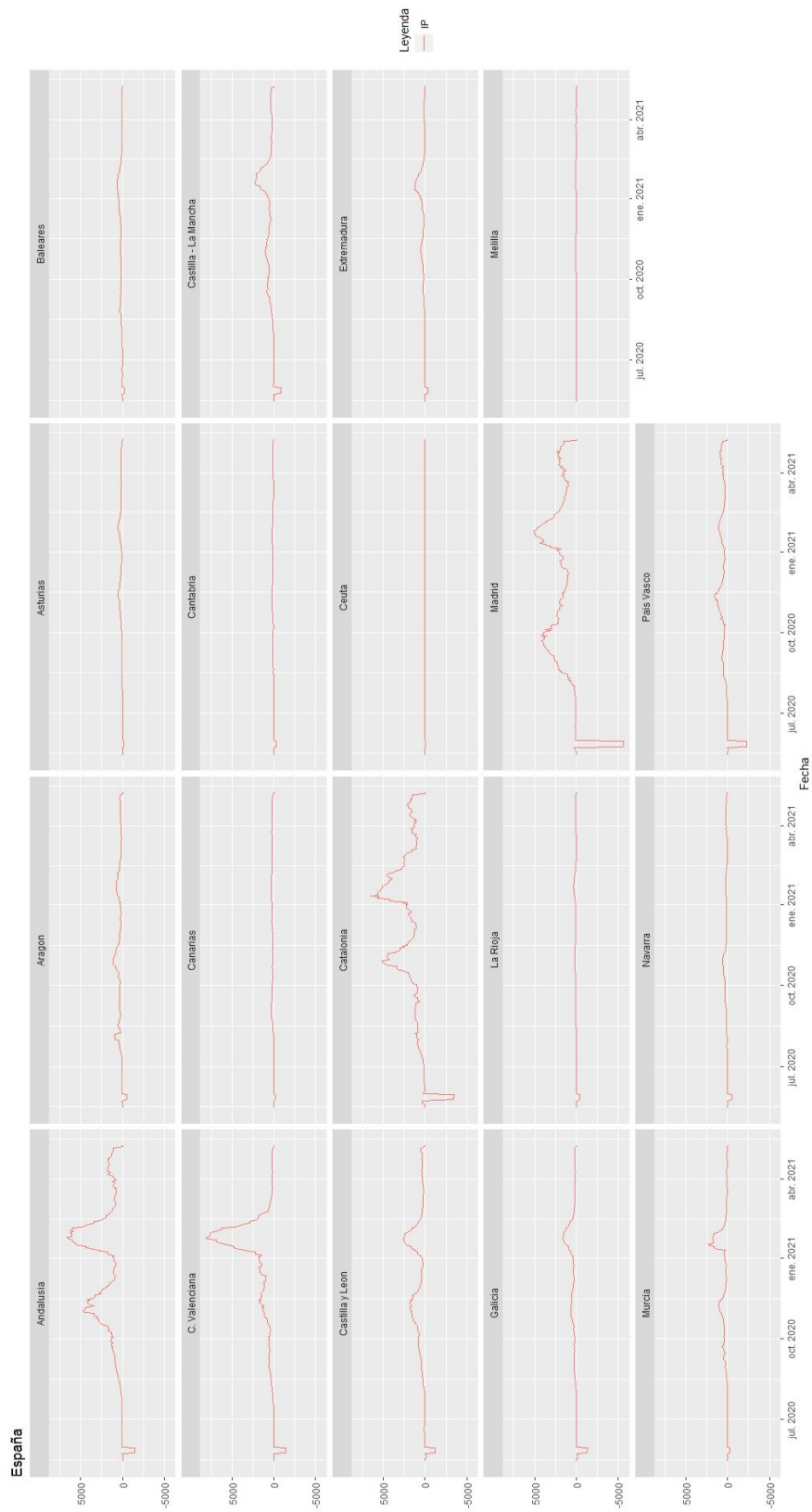
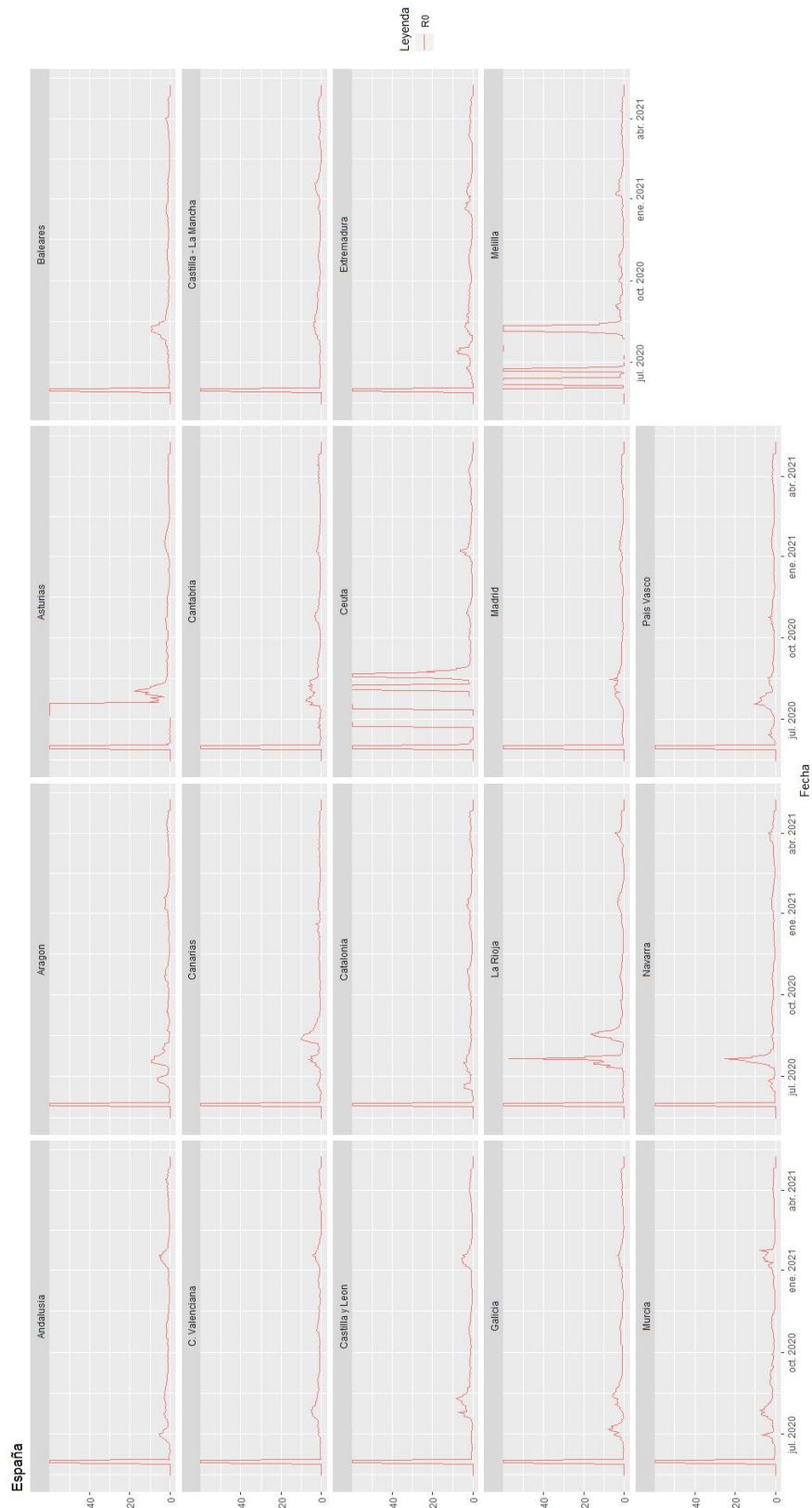


Figura 23: Curva R0 España

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos



Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

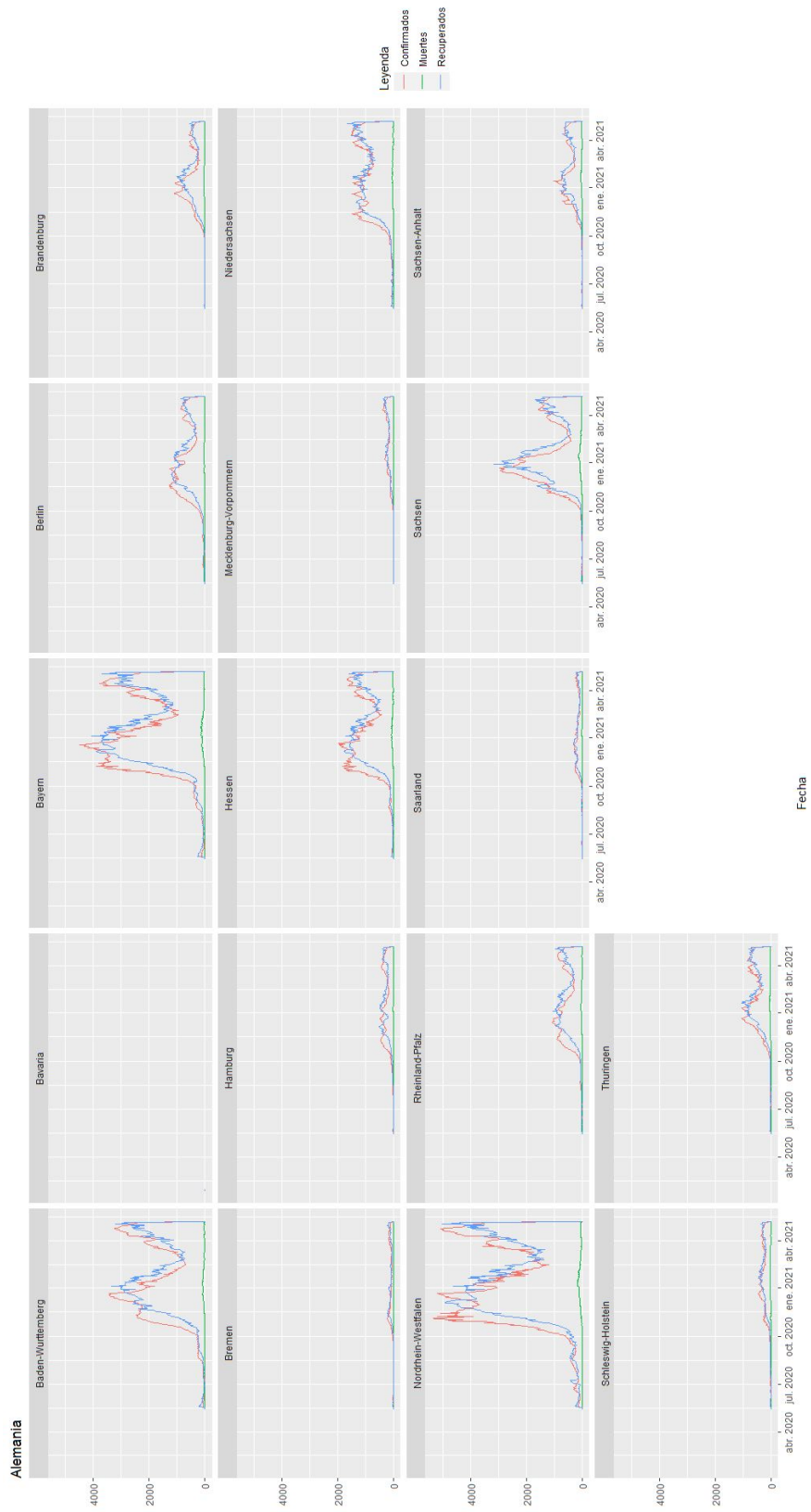


Figura 25: Curva contagios Alemania

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

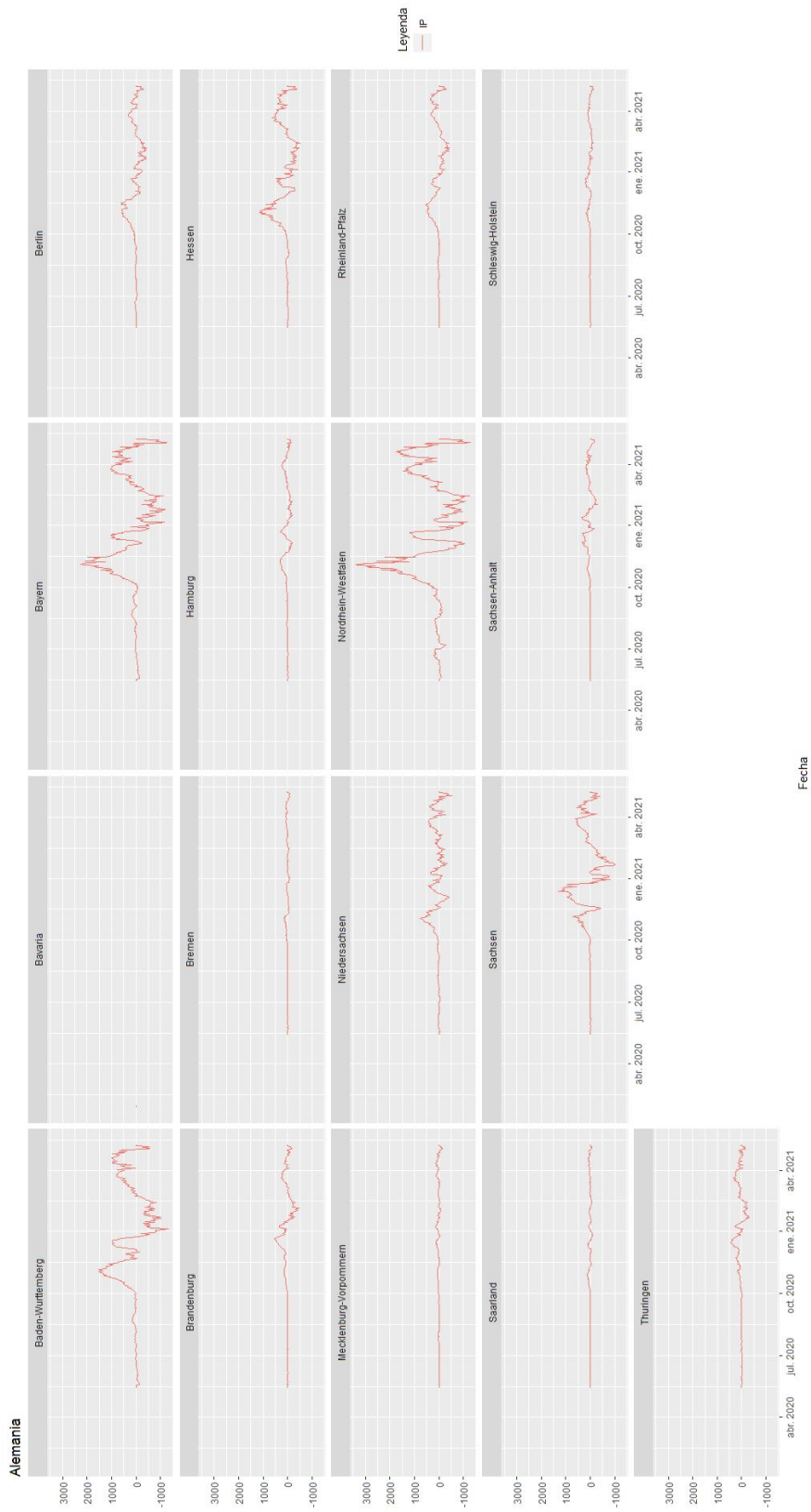


Figura 26: Curva IP Alemania
115

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

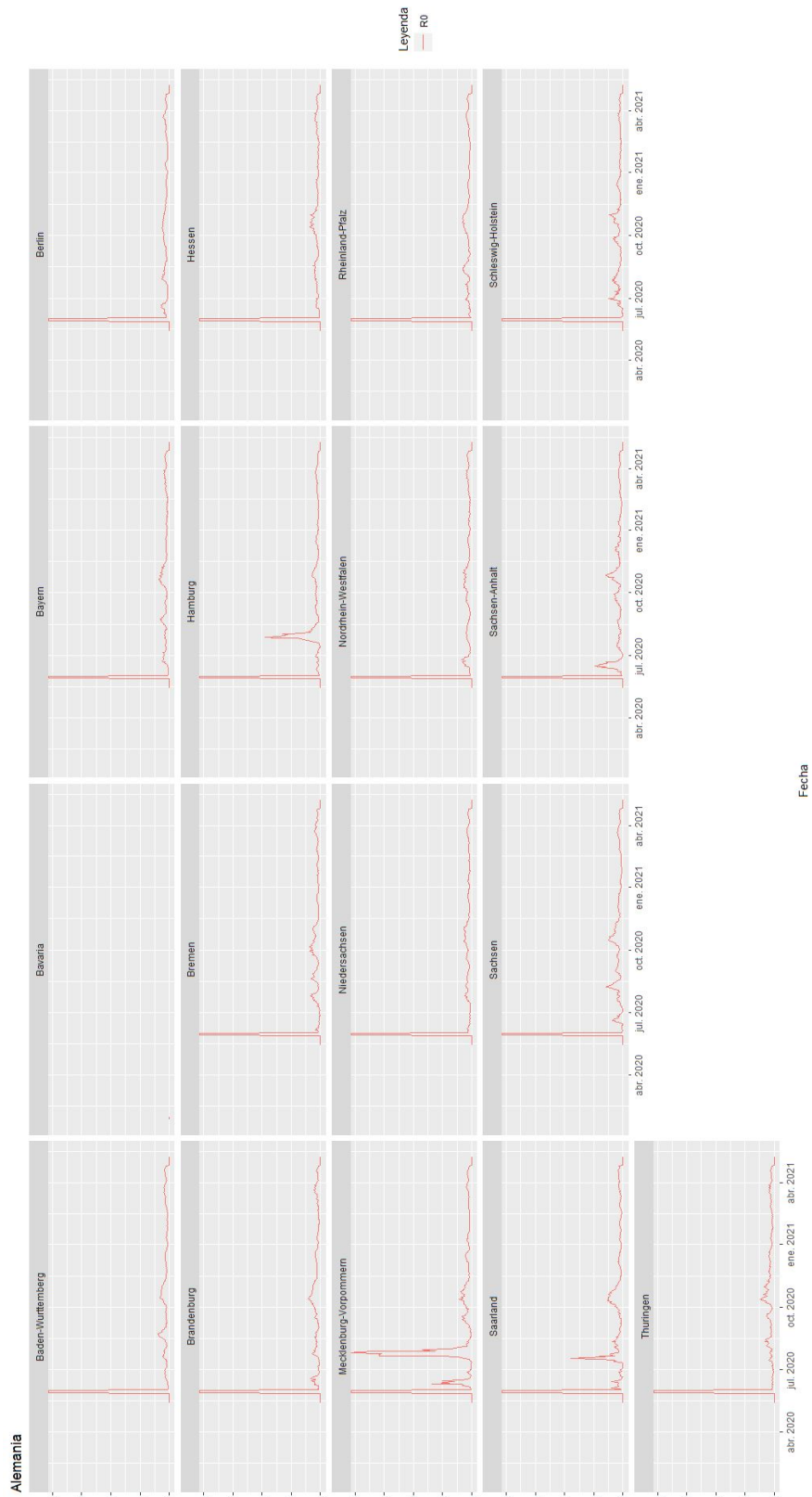


Figura 27: Curva R0 Alemania

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

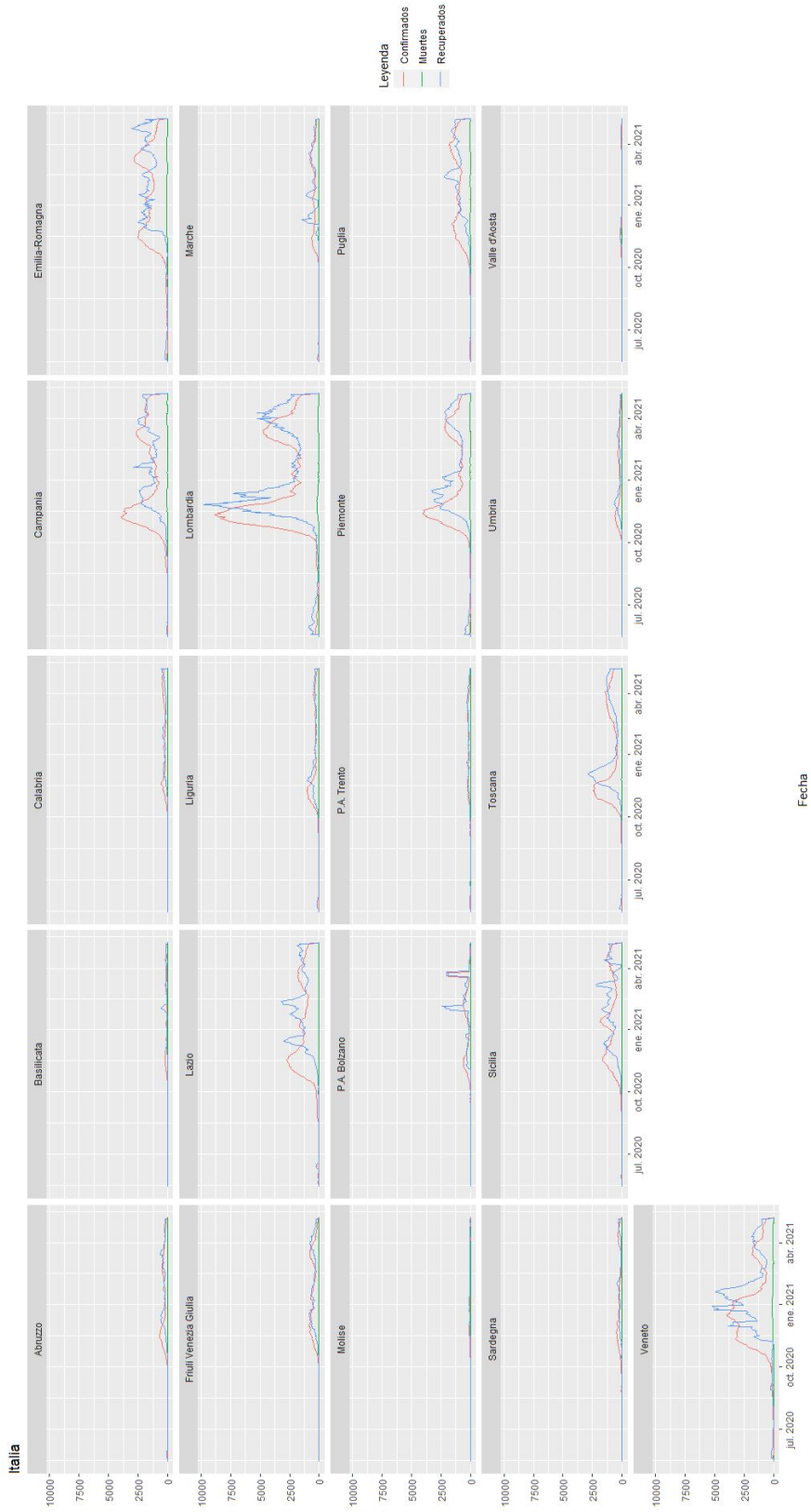


Figura 28: Curva contagios Italia

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

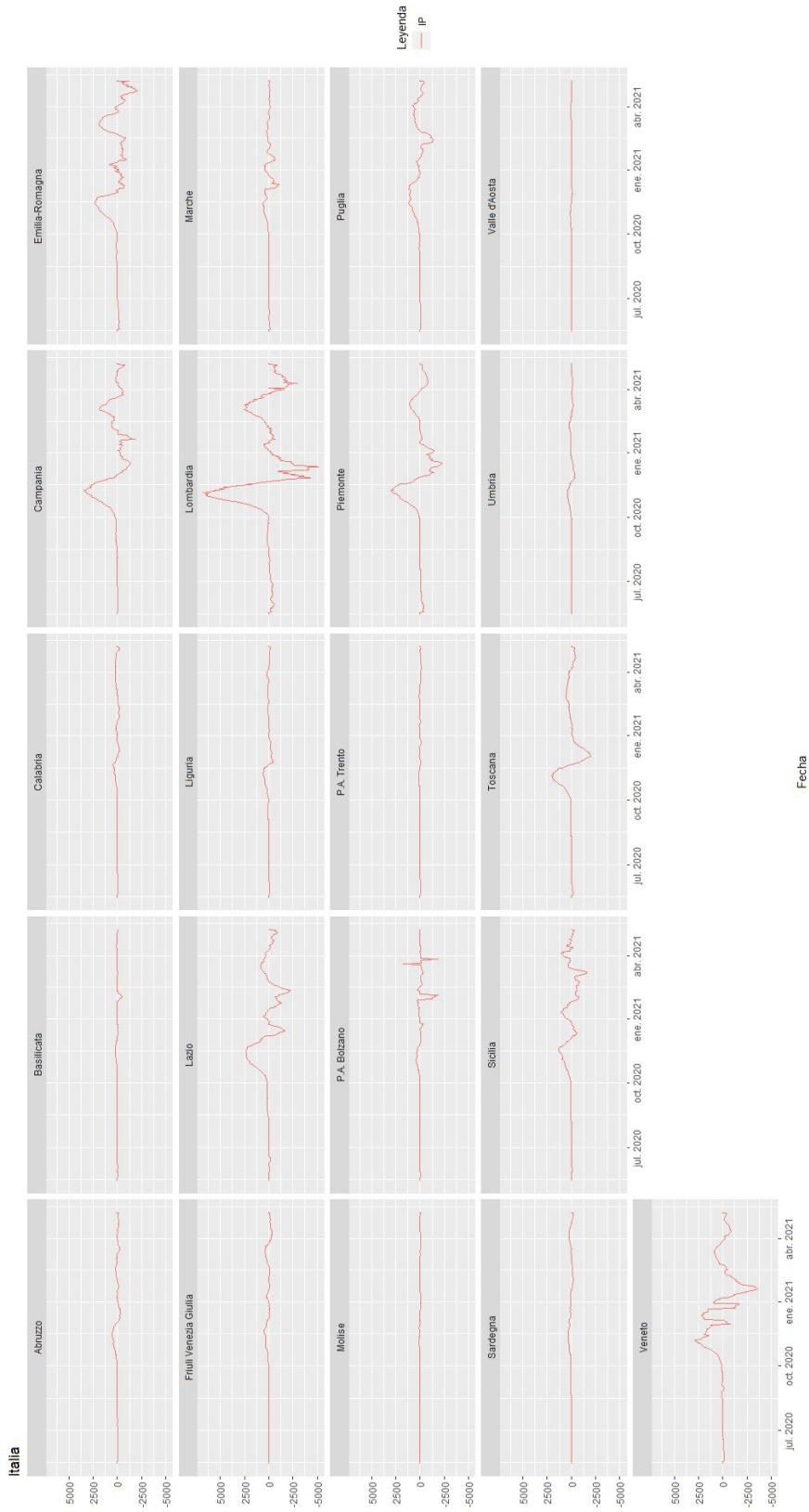


Figura 29: Curva IP Italia
118

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos



Figura 30: Curva R0 Italia

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

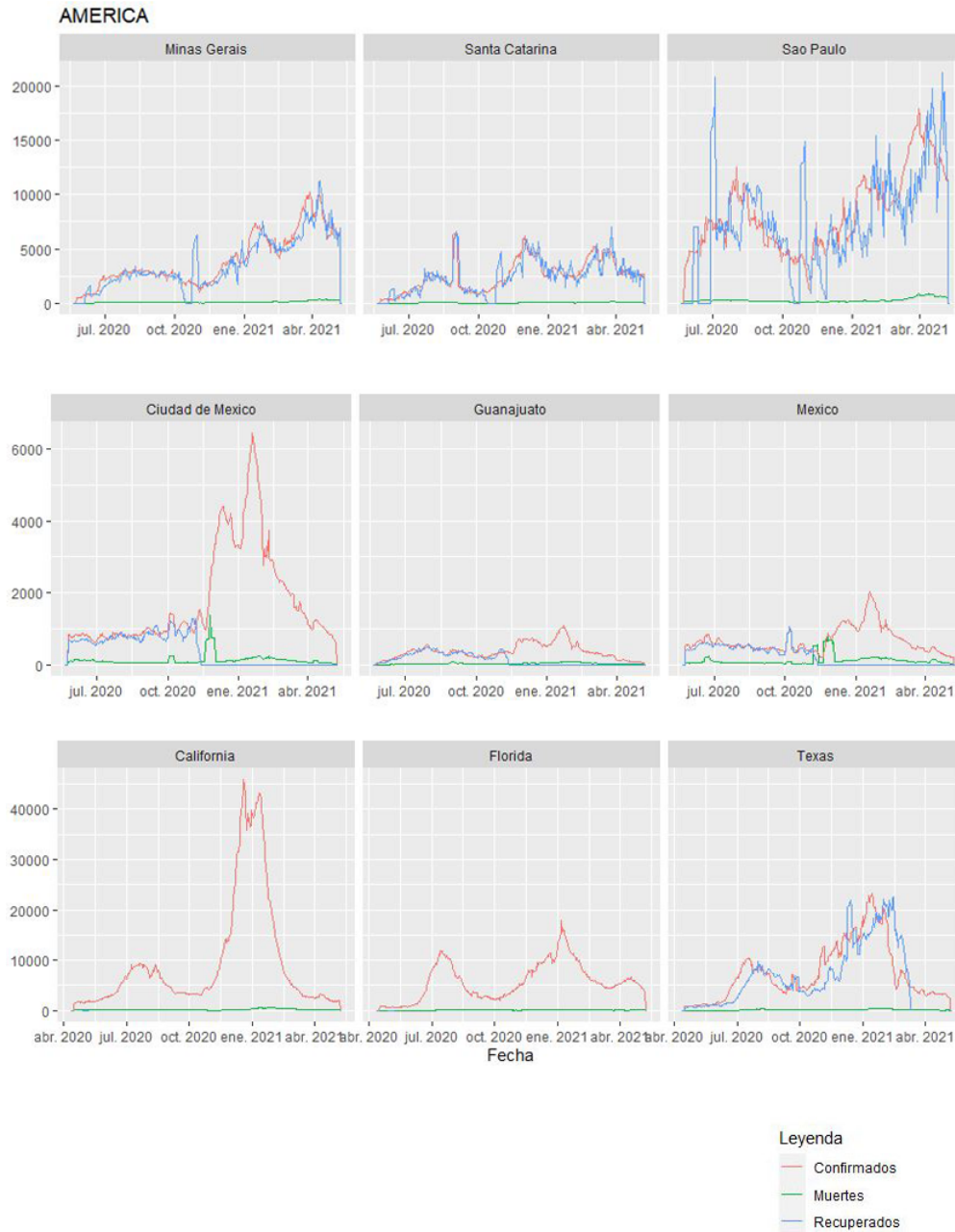


Figura 31: Gráficas regiones América

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

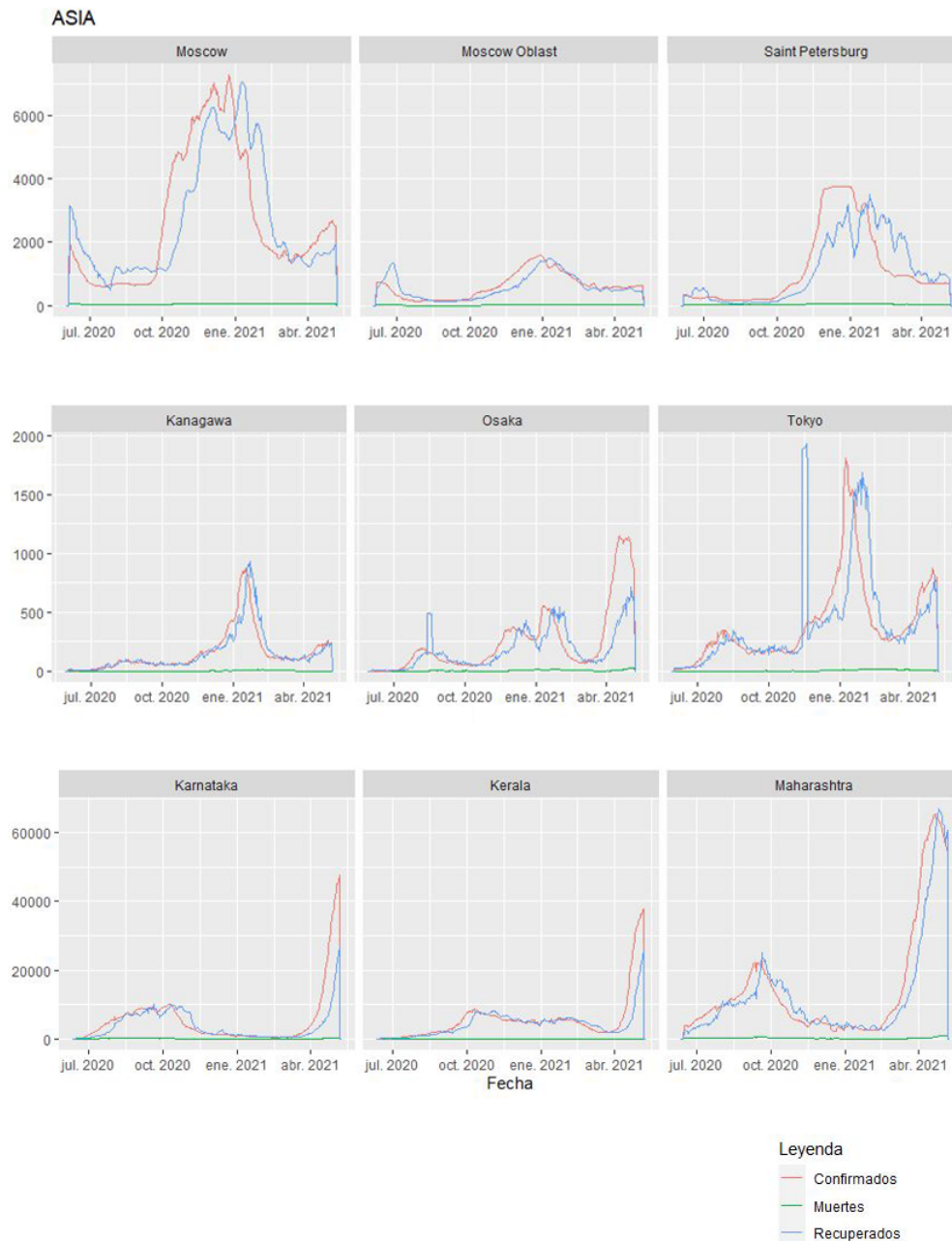


Figura 32: Gráficas regiones Asia

Propagación Covid 19

Recolección, Tratamiento y Análisis de los datos

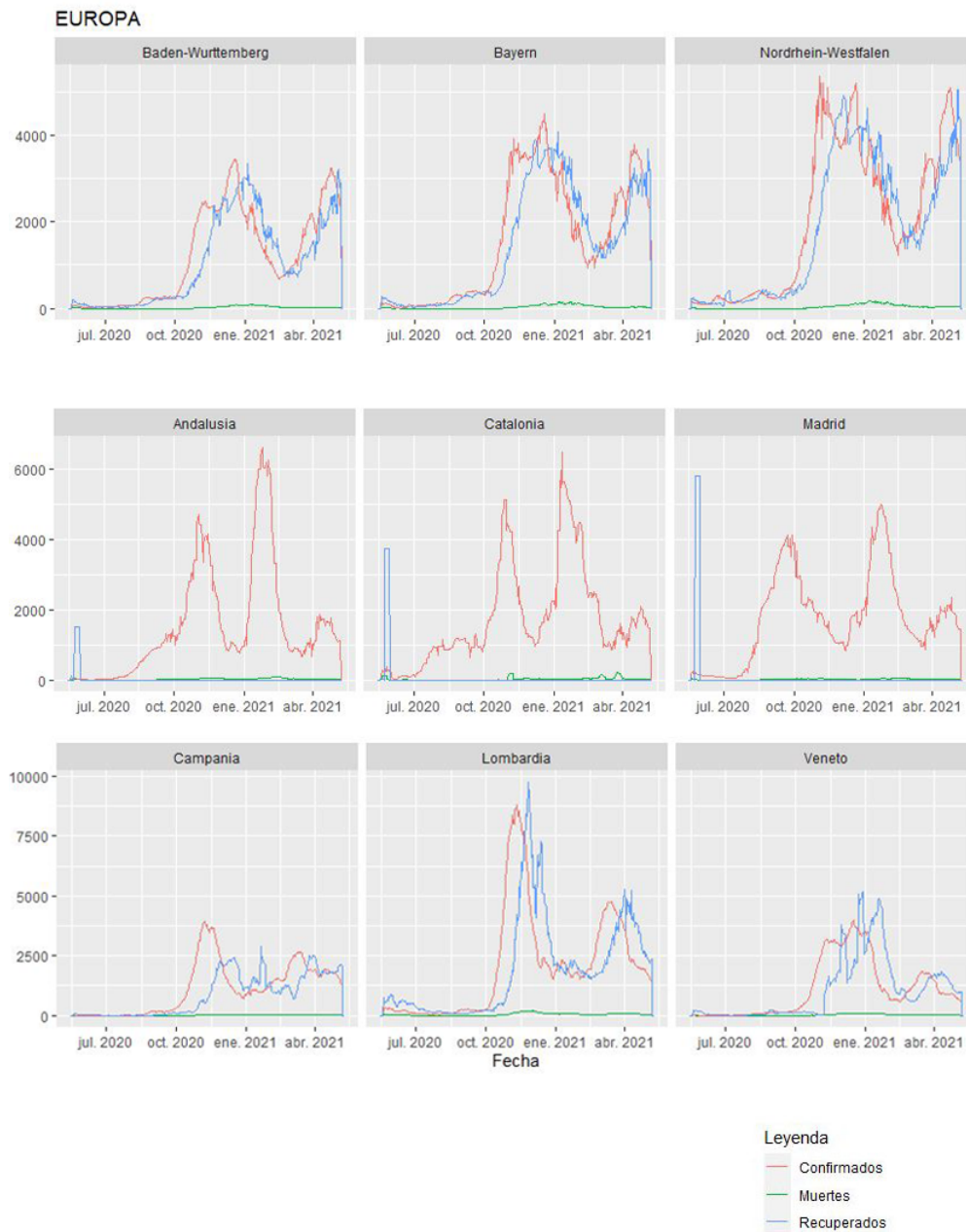


Figura 33: Gráficas regiones Europa