Universidad Politécnica
de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Máster Universitario en Inteligencia Artificial

Trabajo Fin de Máster

Análisis de la Relación de la Propagación del COVID-19 con la Movilidad de Origen y Destino Aeropuertos.

Autor(a): Javier Carbonell García
Tutor(a): Antonio Jiménez Martín y Alfonso Mateos Caballero

Madrid, Julio 2022

Este Trabajo Fin de Máster se ha depositado en la ETSI Informáticos de la Universidad Politécnica de Madrid para su defensa.

Trabajo Fin de Máster
Máster Universitario en Inteligencia Artificial

Título: Análisis de la Relación de la Propagación del COVID-19 con la Movilidad de Origen y Destino Aeropuertos.

Julio 2022

Autor(a): Javier Carbonell García
Tutor(a): Antonio Jiménez Martín y Alfonso Mateos Caballero
Inteligencia Artificial
ETSI Informáticos
Universidad Politécnica de Madrid

Resumen

En 2020 el virus COVID-19 se reconoció como pandemia mundial ante su gran expansión en todo el mundo. Europa se convierte en uno de los epicentros mundiales y sufre las consecuencias de ello. A su vez, España se convierte en uno de los focos principales dentro de Europa lo que provoca una reacción por parte del gobierno. Esta reacción es en forma de estrictas restricciones como la reducción de la movilidad o el confinamiento domiciliario.

La importancia de tener mecanismos para el análisis de la propagación a la hora de tomar medidas de restricción es muy grande. Tener estos mecanismos ayudan a tomar restricciones eficaces estableciendo periodos de tiempo, escala e incluso el tipo de medida acorde con las características de ella.

Este Trabajo de Fin de Máster tiene varios objetivos principales. El primero es la recopilación de fuentes de información sobre la movilidad internacional por carretera, ferrovial y marítima con origen España, para que proyectos futuros puedan dar uso de ellas. El segundo es un análisis de la movilidad por carretera con origen y destino los aeropuertos. Mediante este análisis se pretende mejorar la metodología SIR para la estimación del riesgo exportado de cada aeropuerto español y evaluar de forma detallada el impacto de la movilidad aérea a España a nivel de municipios.

Todos los datos utilizados en este proyecto han sido proporcionados por instituciones públicas, tanto por el gobierno de España como por los gobiernos autonómicos. Por tanto, todos los datos utilizados están respaldados por estas instituciones.

A partir de estos datos se ha elaborado una base de datos orientada a grafos que facilita la consulta de información sobre la movilidad e incidencia en un periodo de tiempo determinado. Esta base de datos también permite calcular estimaciones como el riesgo exportado de cada aeropuerto. La base de datos se ha estructurado para poder extenderse de forma sencilla para su uso en futuros estudios.

Por último se han utilizado los datos de movilidad, incidencia y riesgo importado, para elaborar modelos de Poisson, Quassipoisson y binomial negativo para estudiar su impacto en la propagación del virus en la comunidad de Madrid

Abstract

In 2020, the COVID-19 virus was recognized as a global pandemic due to its worldwide spread. Europe becomes one of the world epicenters and suffers the consequences of this. In turn, Spain becomes one of the main focal points within Europe, which provokes a reaction from the government. This reaction is in the form of strict restrictions such as reduced mobility or home confinement.

The importance of having mechanisms for propagation analysis when taking restriction measures is very significant. Having these mechanisms help to take effective restrictions by establishing time periods, scale and even the type of measure according to the characteristics of the measure.

This Master's thesis has several main objectives. The first is the compilation of sources of information on international mobility by road, rail and sea to Spain, so that future projects can make use of them. The second is an analysis of road mobility to and from airports. This analysis is intended to improve the SIR methodology for estimating the exported risk of each Spanish airport and to evaluate in detail the impact of air mobility to Spain at the level of municipalities.

All the data used in this project have been provided by public institutions, both by the Spanish and regional governments. Therefore, all the data used are supported by these institutions.

From these data, a graph-oriented database has been developed to facilitate the consultation of information on mobility and incidence in a given period of time. This database also allows the calculation of estimates such as the exported risk of each airport. The database has been structured so that it can be easily extended for use in future studies.

Finally, the mobility, incidence and imported risk data have been used to develop Poisson, Quassipoisson and negative binomial models to study their impact on the spread of the virus in the community of Madrid.

Agradecimientos

En primer lugar a mi familia, pareja y amigos por darme fuerzas en la distancia cuando más lo necesitaba.

A mis tutores, Antonio y Alfonso, por ayudarme a mejorar cada día.

Se agradece el apoyo proporcionado por CRIDA (Centro de Referencia I+D+i ATM) y los proyectos MadidDataSpace4Pandemics-CM (financiado por la Consejería de Educación, Universidades, Ciencia y Portavocía de la Comunidad de Madrid y la Unión Europea-FEDER como parte de la respuesta de la Unión a la pandemia de COVID-19) y PID202-122209OB-C31 (financiado por el Ministerio de Ciencia e Innovación).

Tabla de contenidos

1. Introducción **1**
1.1. El transporte y la propagación de enfermedades 2
1.2. Estudios internacionales para la propagación del COVID 4
1.3. Estudios nacionales para la propagación del COVID 5

2. Desarrollo **7**
2.1. Primera temática: Búsqueda de fuentes sobre movilidad internacional hacia España 7
2.1.1. Objetivos 7
2.1.2. Fuente datos 8
2.1.2.1. Datos de movilidad 8
2.1.2.2. Frontour (INEM) 9
2.1.2.3. Facebook API 11
2.1.2.4. MITMA 11
2.1.2.5. INE 13
2.1.2.6. Lifesight 13
2.1.2.7. DataCore 14
2.2. Segunda temática: 15
2.2.1. Objetivos 15
2.2.2. Búsqueda e identificación de fuentes de información 15
2.2.2.1. Datos de movilidad y estructura geográfica 16
2.2.2.2. Datos de Incidencia y estructura geográfica 19
2.2.2.3. Riesgo Importado 26
2.2.3. Tratamiento de datos 27
2.2.3.1. Preprocesamiento Movilidad 28
2.2.3.2. Preprocesamiento Estructura geográfica 29
2.2.3.3. Preprocesamiento Incidencia 30
2.2.3.4. RiesgoImportado 33
2.2.4. Elaboración de la base de datos 34
2.2.4.1. Neo4j y bases de datos orientados a grafos 34
2.2.4.2. Estructura 35
2.2.4.3. Estadísticas 39
2.2.5. Modelado 44
2.2.5.1. Primera parte: Mejora detallada del cálculo del riesgo exportado en Aeropuertos españoles 44
2.2.5.2. Segunda parte: Modelo probabilístico del impacto de la movilidad con origen Aeropuertos en la incidencia COVID en Madrid 46

3. Resultados **51**
3.1. Estimación Riesgo Exportado 51
3.2. Resultados modelo probabilístico 54

4. Conclusiones **57**

Bibliografía **61**

Capítulo 1

Introducción

A finales de 2019 un nuevo virus respiratorio agudo aparece en la ciudad de Wuhan, China, llamado SARS-COV-2 (más tarde llamado COVID-19). La propagación de este virus a nivel nacional e internacional fue tan rápida que en pocos meses la gran mayoría de países sufrían en mayor o menor medida las consecuencias de este virus. El 30 de enero de 2020 la Organización Mundial de la Salud la declaró una emergencia de salud pública de importancia internacional y la reconoció como una pandemia el 11 de marzo de 2020, pandemia del COVID-19.

Este virus se transmite generalmente a través de pequeñas gotas de saliva que se emiten al hablar, estornudar, toser o respirar. Por esto, la importancia de tomar medidas en el aforo,uso de mascarillas, movilidad y distancia de seguridad cobraron una gran importancia para frenar esta propagación.

Europa ,en poco tiempo, se volvió uno de los grandes focos mundiales de propagación del virus COVID-19. Por lo que los gobiernos de los distintos países tuvieron que tomar medidas muy drásticas y rígidas para frenar esta expansión. Entre estas medidas se pueden encontrar periodos extensos de confinamiento domiciliario, gran reducción en la movilidad de todos los medios medios de transporte, cierre de centros públicos, comercios. etc. Dependiendo del país estas restricciones fueron a mayor o menor escala. España ,dentro de la gran incidencia general europea, se convirtió en gran velocidad en uno de los focos europeos en la propagación del virus. Ante este situación extrema el gobierno de España decidió tomar muchas de estas restricciones a gran escala.

La decisión de tomar estas restricciones repercute a la situación económica y sanitaria de cualquier país. Por lo que la toma de ellas deben ser estudiadas con el máximo detalle posible. La mayoría de gobiernos tuvieron que decirlas en poco tiempo por la situación en la que se encontraban pero, actualmente, en 2022 es posible echar la vista atrás y estudiarlas para tener una base en casos futuros.

Los principales estudios para mejorar este tipo de decisiones pasan desde estudios clínicos o biológicos ,donde se estudian las características del propio virus y su tratamiento, hasta los estudios epidemiológicos, donde se estudia la transmisión del virus en la sociedad.

Gracias a los estudios epidemiológicos es posible recopilar una gran información sobre la propagación del virus que ayudarían en gran medida a la toma de restricciones.

Como se transmite el virus en lugares cerrado y abiertos, como afecta la densidad de población, la movilidad, el nivel económico o la situación del año pueden ser estudios muy interesantes o necesarios a la hora de la toma de decisiones.

Este proyecto se centrará en estudiar la relación de la movilidad en España con la propagación del virus. Esta temática se ha estudiado en diversos proyectos y abarca una amplia variedad de temas diferentes. Estas diferencias pueden ser en el tipo de transporte, en el tipo de movilidad (ocio o laboral), escala del viaje (internacional o nacional), origen de los viajes etc.

En este proyecto concretamente se estudiarán dos aspectos distintos de la movilidad. El primero es un análisis de fuentes de información sobre la movilidad no aérea internacional con destino España. El objetivo principal de esta era encontrar datos fiables para analizar la relación entre esta movilidad y la incidencia española. Sin la disposición de estos, este tema del proyecto se centro en analizar los posibles datos encontrados para ser usados en futuros proyectos.El segundo aspecto o tema a estudiar en este proyecto es el análisis de la movilidad por carretera con origen y destino aeropuertos. La idea principal se centra en la importancia que tienen estos puntos en la propagación del virus. Por tanto, se intenta detallar la llegada de personas a un aeropuerto origen, para detallar el riesgo exportado, y como se expanden desde un aeropuerto destino, detallando su municipio de destino.

De esta forma se podría dividir está temática en dos ramas: Mejora del riesgo exportado en Aeropuertos y Análisis de la movilidad con origen aeropuertos y la propagación del virus a nivel municipal.

Por otro lado, en esta sección también se mostrará una revisión bibliográfica de artículos relacionados con la temática de este proyecto. Es decir, se verán varios artículos relacionados con la propagación del COVID-19 y la relación que esta tiene con la movilidad en varios tipos de transporte. Para empezar, se describirán varios estudios donde se analiza este tipo de relación con otras enfermedades diferentes al COVID-19. Se han incluido estos estudios realizados antes de la pandemia del COVID-19 porque mantienen unos objetivos y metodologías similares a los de este proyecto. Luego, se mostrarán estudios más específicos en los que se analiza la propagación del COVID-19 en distintos países del mundo, como China o países europeos. Con esto se muestra con detalle como se ha estudiado este virus en el pasado. Por último, se describirán estudios realizados en España cuyos objetivos son muy similares a los estudiados en este proyecto. En estos últimos se mostrarán los detalles a tener en cuenta para el estudio del virus en el territorio español.

1.1. El transporte y la propagación de enfermedades

La influencia de la movilidad por diferentes medios de transporte es un tema muy estudiado desde antes de que el virus COVID-19 apareciera. En este ámbito vemos estudios más generales intentando modelar la propagación de un virus u otra enfermedad, como en [Tatem *et al.*, 2006] o en [Colizza *et al.*, 2006], o modelos de enfermedades concretas como [MacFadden *et al.*, 2015], [Grais*et al.*, 2003], [Bogoch*et al.*, 2014], [Bogoch*et al.*, 2016] o [Tian *et al.*, 2017].

Dentro de los estudios más generales, se encuentra [Tatem *et al.*, 2006], donde se hizo una recopilación de información de grandes propagaciones históricas de enfer-

medades a lo largo de la historia y la influencia que tuvo el transporte en ellas. Se describieron casos como el cólera, la influenza, la malaria o el dengue. A parte de la recopilación histórica, se describe la importancia de crear modelos a la hora de tomar medidas para la propagación de ellas, aportando posibles tipos de modelos propios. Otro de los estudios más generales es [Colizza *et al.*, 2006], donde se estudia también el transporte, concretamente el aéreo, en la propagación de virus. En este artículo se propone un *framework* de previsiones de pandemias mundiales que considera el tráfico aéreo mundial y el censo poblacional de cada país. Gracias a un gran *dataset*, en el artículo se describe una base de datos que recoge la información del 99 % del movimiento aéreo realizado en el año 2002. Por otro lado utiliza la metodología SIR para el cálculo del riesgo de infección de cada vuelo. Con esto, se propone un modelo de previsión de pandemias.

Dentro de los estudios de la propagación de enfermedades concretas es posible encontrar varios estudios sobre la influenza, ébola, zika o denge entre otros. El primero en destacar es [MacFadden *et al.*, 2015], donde se presenta un estudio en el que se simula como se hubiera propagado la pandemia del virus de influenza de 1968 en el año 2000. El objetivo de esta simulación es estudiar como el aumento en el tráfico aéreo afectaría a la propagación de tal virus y demostrar la importancia de tomar decisiones de propagación comunes en situaciones de pandemia. Otro artículo específico es [Grais*et al.*, 2003] donde se estudia un caso concreto de propagación de virus a través del transporte aéreo. Este caso es el de India y el virus *New Delhi Metallo-beta-lactamase 1*, este virus pareció tener origen en este país para luego propagarse a otros países. Este estudio se centra en encontrar una relación entre la movilidad con India y los principales países contagiados. Para este modelo se utiliza una regresión logística multivariante para predecir el número de infectados en cada país y con varias variables predictoras como la población, el número de pasajeros o el nivel económico. Los resultados obtenidos en el estudio confirman la influencia clara de esta movilidad en la propagación del virus. Otros dos artículos con una metodología parecida son [Bogoch*et al.*, 2014] y [Bogoch*et al.*, 2016]. En [Bogoch*et al.*, 2014], ante la amenaza de una posible pandemia de ébola en 2013 se realizó un estudio de como la movilidad por aire con los 3 países más afectados influían en su propagación en el resto del mundo. Por tanto, se recopiló información sobre los vuelos internacionales con origen y destino Guinea, Liberia o Sierra Leona. Para calcular el riesgo potencial de cada vuelo se utilizaron los datos de casos activos (Confirmado, probable y sospechoso), la población estimada de cada país y el número de pasajeros que realizaron un viaje a las zonas de contagio. Por otro lado, en [Bogoch*et al.*, 2016] se realiza un estudio similar al anterior pero para el virus del zika transportado por mosquitos. El objetivo principal de este estudio era obtener las zonas de riesgo de este mosquito en Brasil y otras zonas de Latinoamérica. Para ello se estudió la movilidad de los principales aeropuertos de Brasil, concretamente los vuelos realizados a las zonas de Asia y África afectadas con este virus. Por último,en [Tian *et al.*, 2017] se estudia la propagación del virus denge en Asia. El objetivo del estudio era ver la influencia de los distintos medios de transporte en su propagación. Los resultados fueron que el medio de transporte que más afecto a la propagación en Asia fue el aéreo y que su principal zona de riesgo u origen de la propagación fue Tailandia.

En general se observa como en un gran número de estudios sobre la propagación de enfermedades se coloca el foco de atención en la movilidad, concretamente en la aérea. Por tanto, con otro virus como el SARS-CoV-2 no iba a ser una excepción.

1.2. Estudios internacionales para la propagación del COVID

Una vez el COVID-19 apareció en el escenario internacional surgieron un gran número de estudios buscando el origen de esta propagación a escala mundial. Uno de los orígenes que la gran mayoría propuso fue la movilidad internacional y las ineficaces o tardías medidas impuestas por los distintos países. Por tanto, los primeros estudios de esta temática están relacionados con la movilidad internacional en los primeros meses de pandemia. Los primeros estudios que analizaron la movilidad internacional en los primeros momentos de la pandemia se centraban en la movilidad con origen China, ya que era considerada epicentro de la enfermedad. Uno de los primeros estudios que analizaron esta movilidad fue [Pullano *et al.*, 2020]. Este artículo se realizó en enero de 2020, cuando en Europa solo existían 3 casos detectados en Francia y Alemania. El objetivo principal de este fue la estimación del riesgo de que cada aeropuerto europeo llegase un infectado con origen Chino, es decir, el riesgo importado. Los resultados obtenidos revelaban que los países con mayor riesgo estaban acorde a su población, es decir, Alemania, Francia, Inglaterra, España e Italia encabezaban este podio.

En [Nakamura *et al.*, 2020] se propone un cálculo de riesgo, tanto importado como exportado, de cada uno de los principales aeropuertos mundiales. Para el cálculo del riesgo exportado se utilizan datos estimados de población y número de infectados. En cambio, para el cálculo del riesgo importado es necesario recopilar información sobre los vuelos con destino el aeropuerto propuesto y calcular el riesgo exportado de cada aeropuerto origen, es decir, a partir del número de casos y su población. La principal diferencia de este estudio es el no utilizar como epicentro único del virus China si no utilizar el riesgo exportado de cada aeropuerto, apareciendo así varios epicentros. Los resultados mostraron como países como Estados Unidos o Europa ya mostraban niveles altos de riesgo en momentos en el que apenas solo se detectaban casos en China.

En [Sokadjo *et al.*, 2020] se propone varios modelos probabilísticos que contrasten la hipótesis de la influencia que tiene la movilidad aérea en la propagación del COVID-19 en cada país. La variable de interés utilizada es el número de casos detectados en un país, es decir una variable de conteo. Por esto, utiliza los modelos de *Poisson, quasi-Poisson*, *Negativebinomial* , *zero-inflated* y *Hurdle*. Como se ha dicho, el número de casos detectados sería la variable de interés mientras que en las variables predictoras solo se encuentra el número de pasajeros que viaja a un país en una determinada fecha. En este artículo no se trabaja con el riesgo importado si no que se utiliza el número de pasajeros sin ningún cálculo añadido. Una vez realizados los modelos se contrastan utilizando los datos de incidencia de su respectivo país en un periodo de tiempo. Los resultados obtenidos en este estudio revelaban que cuando el trafico aéreo aumentaba en uno su número de pasajeros, el número de casos aumentaba en uno.

En [Lau *et al.*, 2020] se presentan dos estudios diferentes relacionados con la expansión del virus en China. En este se estudia, por un lado, la relación entre las movilidad internacional con China y la incidencia en el país de destino. Y, por otro lado, la movilidad aérea nacional en China. Los datos utilizados para este estudio son estimaciones realizadas con la movilidad y rutas de años anteriores. Los resultados fueron distintos en la movilidad internacional que en la nacional. Para la movilidad

internacional se encontró una gran correlación entre el número de casos de un país y el número de pasajeros procedentes de China. Pero para la movilidad aérea nacional se encontró una menor correlación. Como conclusión, proponen que el motivo de esta menor correlación es que la expansión del COVID-19 nacional en China era ya muy grande y, por tanto, otros medios de transportes más comunes como por carretera o ferroviario tuvieron más influencia en esta propagación. En [Kraemer *et al.*, 2020], se realiza un estudio de la movilidad nacional en China estableciendo dos espacios temporales, antes de cortar la movilidad con Wuhan y después. El objetivo principal del estudio es demostrar si la medida de cortar la movilidad con esta región frenó el contagio en el resto del país. Los resultados de este fueron que existía una gran correlación entre la movilidad con esta región y la incidencia en el resto de China antes de la medida comentada. Luego, en las primeras semanas, se demostró un descenso muy grande en el crecimiento de la incidencia del país. Al pasar estas semanas, el crecimiento de las demás regiones volvió a subir sin tener relación este con la movilidad con Wuhan. Estos resultados dan a entender como este tipo de medidas reducen considerablemente el aumento de contagios pero de forma temporal.

1.3. Estudios nacionales para la propagación del COVID

En [Orea *et al.*, 2020] se estudia la eficacia de las medidas tomadas en España para frenar la propagación en el país. Los resultados obtenidos demuestran como el inicio y la intensidad de la propagación en un región están muy relacionadas con la movilidad internacional realizada con países focos como Italia. Esta movilidad afectó a mayor nivel en las primeras semanas de pandemia, siendo una de las principales causas. Más tarde esta influencia se va reduciendo mientras que aumenta la de otros medios de transporte nacionales. Con esto, se demuestra como la movilidad interprovincial tenía una gran relación con la incidencia provincial en momentos más avanzado de la pandemia, especialmente en las zonas cercanas a los focos locales como Madrid y Barcelona. Estas zonas fueron las más beneficiadas con el estado de alarma español.

En [Mazzoli *et al.*, 2020] se muestra un estudio distinto para la movilidad dentro de España. Este estudio se centra en analizar los movimientos con los epicentros de propagación local en España, es decir los puntos con incidencia más alta. Por tanto, los movimientos con más atención serían ciudades como Madrid, Barcelona o Valencia. Aunque este estudio se ha realizado en los primeros momentos de la pandemia, el virus ya estaba presente en los puntos mencionados. En los resultados obtenidos se puede observar una gran correlación entre la movilidad con estos puntos de riesgo y la incidencia.

Por último,en [García-Abadillo, 2021] se evalúa el impacto de la movilidad por carretera en la propagación del COVID-19 en las primeras semanas de pandemia. Para esto se utilizan datos de ubicación de terminales móviles obteniendo así la movilidad nacional. Junto a estos datos se obtuvieron datos de incidencia de distintas zonas de España. Por otro lado, también se obtiene unos estimadores de riesgos de cada uno de estos territorios españoles. Estos datos de movilidad e incidencia en España se obtuvieron de organismos públicos del gobierno de España.

Capítulo 2

Desarrollo

En esta sección se describirá con detalle como se ha desarrollado este proyecto. En un inicio se ha explicado como la estructura del proyecto tiene varias ramas diferenciadas. Estás ramas representan las dos temáticas diferentes que se han estudiado en el proyecto. En un primer momento se planteó el estudio de la movilidad internacional hacia España en todos los medios de transporte excepto el aéreo. Este estudio quedó incompleto al no encontrar ninguna fuente de información con los datos necesarios para realizarse por lo que el proyecto se centro en una segunda temática. Esta segunda temática era el estudio de la movilidad por carretera entre localidades y aeropuertos. A su vez esta temática se dividió en dos objetivos principales: La mejora del riesgo exportado de cada aeropuerto español y el análisis de la movilidad con origen aeropuertos y la propagación del virus en el destino.

2.1. Primera temática: Búsqueda de fuentes sobre movilidad internacional hacia España

Como se ha comentado este proyecto se puede dividir en dos etapas. En esta sección se explicará la primera. Esta primera etapa tenía como objetivo analizar el tráfico por carretera, marítima y ferroviaria internacional hacia España, dándole más importancia a este primer medio de transporte.

2.1.1. Objetivos

Por tanto, el objetivo de esta primera temática es calcular el riesgo importado para cada uno de los medios de transportes mencionados de forma diaria y crear un modelo probabilístico que estudie la influencia de esta en los destinos del viaje.

Los pasos que se plantearon para esta temática fueron los siguientes:

- Revisión Bibliográfica: Estudiar proyectos similares ya realizados.
- Búsqueda e identificación de fuentes de información: La búsqueda de datos contrastados que respalden nuestros resultados
- Tratamiento de datos: Extracción y corrección de los datos necesarios a partir de los datos encontrados.

- Modelado: Elaboración del propio modelo probabilístico.
- Análisis de resultados: Comprensión de los resultados obtenidos con el modelo y conclusión.

Desgraciadamente el segundo paso de este proyecto no se pudo completar. Durante este, se encontraron numerosas bases de datos pero no se ajustaban a las necesidades de este estudio. Por tanto, tras analizarlas a fondo se decidió modificar la temática del proyecto.

Aun así las fuentes de información encontradas pueden ser de gran interés ya que se pueden ajustar a proyectos similares. Por tanto, se describirán estas fuentes de información.

2.1.2. Fuente datos

2.1.2.1. Datos de movilidad

La búsqueda de fuentes de datos sobre la movilidad internacional ha sido una tarea complicada de realizar. Esta dificultad estaba causada por los pocos registros existentes de los trayectos fronterizos en Europa al existir libre circulación por el continente. Para este proyecto en específico era necesaria una base de datos con una información concreta esencial. Estos datos debían cumplir las siguientes características:

- Deben tener información de la movilidad internacional por carretera, marítima y ferrovial. Esta información puede estar en una misma base de datos o en varias distintas.
- La información debe estar organizada de forma diaria para el análisis de esta.
- La información debe de ser, temporalmente, de antes de los eventos sucedidos con la aparición del virus SARS-CoV-2. Por esto, el momento perfecto para el uso de estos datos es el año 2019.
- Debe tener información sobre el país de origen de los visitantes.

Con estos objetivos establecidos se comenzó la búsqueda de información encontrando una serie de posibilidades muy distintas. Varias de las fuentes encontradas no cumplen con todos los requisitos establecidos anteriormente pero pueden ser útil para proyectos similares a este con prerrequisitos parecidos.

Una vez realizada la búsqueda de datos sobre movilidad se han encontrado diferentes metodologías de obtención de estos datos. Respecto a los datos de movilidad en los principales medios de transporte como son avión, ferrocarril y marítimo es posible obtener datos de las principales compañías de cada sector para su uso estadístico o utilizar estudios ya realizados por el Instituto Nacional de Estadística. Por otro lado, para los datos de movilidad por carretera es necesario una metodología más compleja para su obtención. Estas metodologías las podemos agrupar en los siguientes tipos:

- *INE*: Estudios estadísticos realizados por el Instituto Nacional de Estadística a través de distintas vías.
- *Geolocalización por redes sociales*: En esta metodología se recoge información de las cuentas personales de una o más redes sociales.

- *Geolocalización por terminales móviles*: Utiliza la geolocalización proporcionada por las compañías telefónicas para construir la base de datos de movilidad.

A continuación se describirán las fuentes de datos encontradas sobre la movilidad ferrovial, marítima y por carretera en España. En esta descripción se intentará mostrar sus principales características y las metodologías utilizadas con intención de facilitar su uso en el futuro.

2.1.2.2. Frontour (INEM)

Frontour o Estadística de movimientos turísticos [Frontur] en frontera es una base de datos elaborada por el Instituto Nacional de Estadística para el control del turismo internacional hacia España. En esta base de datos se encuentra información desde el año 2015 hasta la actualidad mostrando esta información de forma mensual. Los datos utilizados para la elaboración de estas estadísticas fueron obtenidos de la Encuesta de Movimientos Turísticos en Frontera y Gasto Turístico encuesta cuyo objetivo era proporcionar estimaciones mensuales o anuales de los visitantes no residentes en España en un momento dado y sus principales características.

Los objetivos principales de este estudio serían: Medir el número de no residentes en España que llegan a nuestro país cada mes. Estos se clasifican por su vías de acceso (Carretera, Marítima, Aeropuerto y Ferrovial), por ser turista o excursionista y otros detalles de su viaje como el destino, tipo de alojamiento, país de residencia, motivo del viaje o la forma de organización.

La población objeto de este estudio son las personas no residentes en España que entran o salen de él, habiendo realizado pernoctación o que simplemente estén en el país de tránsito. Por tanto, para este proyecto ha sido necesario estudiar el tráfico en las principales fronteras de España, prestando especial atención en los extranjeros. De esta forma se recopiló información detallada sobre su visita turística. Una visita turística se considera cualquier desplazamiento fuera del entorno de residencia de la persona y sin relación laboral. Si esta visita implica al menos una pernotación se considera viaje turístico. En cambio si no implica ninguna pernoctación se considera excursión.

Para realizar la estimación a partir de las encuestas realizadas es necesario algún tipo de registro del número de personas que entran al país por cada vía de transporte. Estos registros han sido proporcionados por entidades distintas dependiendo de su vía de acceso. Por carretera la Dirección General de Tráfico proporciona los registros obtenidos por cámara en los principales puntos fronterizos del país. Por avión, AECFA y AENA han proporcionado registro sobre la totalidad de los vuelos a territorio español y unas estimaciones del número y nacionalidad de viajeros. Por vía ferroviaria, Renfe ha proporcionado registros de la totalidad de trenes con destino español junto a una estimación de sus pasajeros. Por último, por vía marítima Puertos del Estado ha proporcionado un registro de los barcos y pasajeros embarcados y desembarcados en los principales puertos de España.

Esta base de datos tiene una serie de atributos que permite su clasificación por tablas. Los atributos se muestran en la Tabla 2.1

Cuadro 2.1: Atributos Frontour

VARIABLE	DESCRIPCIÓN
Vía de acceso	Representa la forma de acceso al país. Este atributo puede tener los valores puerto, tren, avión y carretera.
Comunidad autónoma de acceso	Como su nombre indica este atributo representa la comunidad destino del viaje y tiene como valores las 17 comunidades españolas.
País de residencia	Representa el país de origen del viajero en cuestión. En este atributo podemos encontrar como valores a los principales países de cada continente representando a los países restantes como resto de y su respectivo continente.
Motivo del viaje	Este atributo intenta diferenciar los viaje turísticos y los de negocio.
Tipo de Alojamiento	Este atributo representa la forma en la que el o los viajeros se han hospedado en el país. No se profundizará mucho en los valores de este atributo pero alguno de sus valores son hotel, vivienda en alquiler o vivienda en propiedad.
Forma de organización del viaje	Este atributo intenta diferenciar los viajes según si se ha organizado en forma de paquete turístico o sin él.
Duración del viaje	Este atributo representa el número de noches que tiene de duración un viaje dado. Los valores se representan mediante atributos siendo 1 noche, de 2 a 3 noches, de 4 a 7 noches, de 8 a 15 noches o más de 15.
Periodo	Representa temporalmente la información teniendo valores mensuales desde 2015 hasta 2021, ambos inclusive.

La unidad de medida de estas estadísticas son el número de viajeros o número de viajes. Respecto a la confidencialidad del estudio, el Instituto Nacional de Estadística asegura el cumplimento de la *Ley 12/1989, de 9 de mayo, de la Función Estadística Pública y del Reglamento (CE) nº 223/2009 del Parlamento Europeo y del Consejo, de 11 de marzo de 2009*, adoptando las medidas lógicas, físicas y administrativas necesarias para que la protección de los datos confidenciales sea efectiva, desde la recogida de datos hasta su publicación.

La garantía de calidad de este estudio se basa en el uso de El Código de Buenas prácticas de las Estadísticas Europeas.

Este estudio ha proporcionado información estadística muy útil tanto para este como para todo tipo de estudios relacionados con la movilidad extranjera y el turismo en España, tanto en época de cuarentena como en años anteriores.

Esta base de datos es muy completa cumpliendo la mayoría de los prerequisitos establecidos y añadiendo nueva información muy útil. El único prerrequisito que no cumpliría es su división temporal mensual ya que se necesitaría una diaria.

2.1.2.3. Facebook API

En este apartado se explicará como crear una base de datos producida por *Facebook Marketing API* y *Ads Manager interface*. En el artículo [Gendronneau *et al.*, 2021] se utiliza estas herramientas para desarrollar un estudio sobre la inmigración entre países de la Unión Europea. Esta dos herramientas permiten registrar los usuarios que encajen en unas determinadas características introducidas previamente y están a disposición de cualquier usuario registrado. Estas características o atributos que deben ser establecidos previamente y están relacionados con la demografía, localización, intereses y comportamientos. Concretamente en este artículo se centran en los atributos relacionados con la localización, como en este proyecto. En este artículo mencionado se utilizan los atributos de localización para calcular el número de personas que son de origen extranjero pero actualmente están viviendo en España. Además de las dos herramientas se utiliza un software para estimar los valores llamado *PySocialWatcher* disponible en *GitHub* [PySocialWatcher].

Por contraparte para el objetivo de este proyecto se podría generar una *query* que recopilara la información de los perfiles con residencia en el extranjero y que actualmente estén en España. Esto se podría conseguir mediante el atributo "geolocations"que representa la localización actual en ese momento. Con esto podríamos generar las estadísticas tanto mensuales como diarias de las personas extranjeras que viajan a España. Los principales problemas de esta base de datos serían la no clasificación de viajes por su vía de entrada y la imposibilidad de recopilar información del pasado lo que imposibilita la obtención de datos de antes de la pandemia.

2.1.2.4. MITMA

En 2020, el Ministerio de transporte, movilidad y agenda urbana realizó un estudio de la movilidad de los ciudadanos españoles en el interior del país [OpenData Movilidad]. El objetivo de este era el estudio de esta movilidad durante los primeros meses de pandemia. La gran diferencia de este informe fue la novedosa metodología con la que obtuvieron toda la información necesaria. Esta metodología consistía en obtener datos de los terminales móviles proporcionados por las principales compañías telefónicas del país. También fue el primer contacto con esta metodología para su uso en posteriores estudios. El objetivo principal de este proyecto es el estudio de la movilidad española durante el Estado de Alarma, es decir, cuantificar el número de personas en movimiento y hacia donde se dirigen.

Esta metodología fue todo un reto y a la vez una gran oportunidad de información más exacta sobre movilidad. El principal reto fue analizar esta gran cantidad de datos para obtener la información necesaria para el estudio. Como recompensa, se obtuvieron una información más exacta capaz de proporcionar datos sobre la movilidad interterritorial de los españoles.

Este *dataset* tiene un nivel de detalle muy alto. Este, recoge información sobre los viajes realizado por horas y por longitud del trayecto. Al recoger la información a partir de datos móviles pueden tener un error de unos metros en ciudad hasta varios kilómetros en zonas rurales.

Temporalmente este estudio se divide de dos formas distintas. La primera recoge información diaria desde Febrero de 2020 hasta Mayo de 2021. Y la segunda resume la información primera por meses.

Geográficamente el estudio divide el territorio español en una serie de distritos divididos por su población etiquetándolos con un id. La relación de estos id con sus coordenadas geográficas está contenida en un archivo *shapelife*. Aún teniendo las coordenadas geográficas es complicado relacionar el distrito con sus respectivos municipios de forma automática.

El estudio se ha estructurado en dos matrices distintas. La primera matriz centrada en la información de cada viaje. Y la segunda matriz centrada en los viajes por persona. Concretamente la matriz 1 o matriz de viajes contiene el número de viajes y de viajeros para cada día y cada combinación de origen y destino, actividad realizada, residencia, hora y distancia. Con más detalle vemos sus atributos en la Tabla 2.2.

Cuadro 2.2: Variables Matriz 1 MITMA

VARIABLE	DESCRIPCIÓN
fecha	Día en el que se realizó el viaje. FORMATO='AAAA-MM-DD'
origen	Id del distrito origen del viaje.
destino	Id del distrito destino del viaje.
actividad_origen	Actividad realizada en el distrito origen.
actividad_destino	Actividad realizada en el distrito destino.
residencia	Lugar de residencia del viajero.
edad	Edad del viajero.
periodo	Periodo horario en el que se realizo el viaje.
distancia	Distancia recorrida (Por rangos).
viajes	Número de viajes.

En la segunda matriz o matriz viajes por persona se reduce el número de variables y muestra el número de personas que han hecho viajes en cada fecha y combinación de distrito y número de viajes. Concretamente sus variables se muestran en la Tabla 2.3.

Cuadro 2.3: Variables Matriz 2 MITMA

VARIABLE	DESCRIPCIÓN
fecha	Fecha en la que se realizaron los viajes. FORMATO='AAAA-MM-DD'
distrito	Id del distrito destino.
numero_viajes	Número de viajes que ha realizado el viajero.
personas	Personas que han realizado un viaje de tal caracteristicas

Un aspecto importante que mencionar de este estudio es su exactitud. La unidad mas pequeña en la que se puede medir la posición de un terminal es el área de

movilidad. Por tanto, es imposible conocer con más detalle la posición dentro de este área. También destacar que la mayoría de movimientos se suelen realizar dentro de su respectiva área de residencia. Por otro lado, hay que tener cuidado con los datos obtenidos de movilidad con áreas colindantes al área de residencia ya que pueden haber sido provocadas por errores en la ubicación de los terminales móviles.

Por último, respecto a la confidencialidad de los datos, las operadoras móviles han proporcionados los datos de forma anónima de forma que desde MITMA es imposible rastrear cualquier movimiento individual. Las medidas tomadas están acorde a la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales.

Respecto al estudio de este proyecto, esta base de datos nos proporcionaría toda la información necesaria si en ella se usaran los números de teléfonos con prefijo extranjero. Desgraciadamente para el estudio se rechazó todo número sin prefijo español por lo que no es posible su uso en este proyecto concreto.

2.1.2.5. INE

Este *dataset* es muy interesante. Proviene de un estudio realizado por el Instituto Nacional de Estadística. Utiliza una metodología similar a la anteriormente mencionada, por datos móviles. A diferencia del anterior proporciona unos datos mejor estructurados, explicados y sencillos de usar, además de una documentación muy clara. En general, tenía como objetivo el estudio de la movilidad dentro de España, como en MITMA. Por todo esto, esta base de datos se utilizaría en la segunda temática del proyecto y en su sección será descrita al detalle.

2.1.2.6. Lifesight

En este apartado se describirán algunas bases de datos ofrecidas por la empresa Lifesight en la plataforma Datarade. Estas bases de datos son las primeras bases de datos no públicas que se decriben. El precio de estas varía dependiendo de la base de datos y se especificarán a continuación. Estas bases de datos utilizan los datos recibidos por los terminales móviles para localizar en vivo su posición. También son base de datos formadas por dispositivos de varios países entre ellos España y gran mayoría de los países europeos.

La primera que describiremos será la llamada *Lifesight Mobility/ Raw Location Data | Global mobile location data* [Lifesight Mobility]. Esta base de datos cuenta con un volumen de 295 millones de usuarios activos y una media de 1.92 billones de registros al día. El precio de esta base es de 4.500$ al mes. Como se ha explicado anteriormente, esta base de datos esta compuesta por la geolocalización de dispositivos móviles en vivo. En ella se encuentran una serie de atributos que nos proporcionan información especifica de cada registro. Los principales atributos se muestran en 2.4:

Cuadro 2.4: Atributos Lifesight Mobility

VARIABLE	DESCRIPCIÓN
Maid	Id identificativo del dispositivo.
Latitude	Latitud de la localización del dispositivo.
Longitude	Longitud de la localización del dispositivo.
Timestamp	Instante temporal del evento. FORMATO = 'dd/MM/YYYY hh:mm:ss'.
Country	País de origen del dispositivo.
State hasc	Estado, provincia o región de origen.
City hasc	Ciudad de origen.
Vertical accuracy	Precisión vertical.
Ha accuracy	Precisión horizontal.

Con toda esta información se es capaz de analizar el número de viajantes internacionales que cruzan la fronteras de España e incluso se podrían extrapolar estos datos a cualquier país de la Unión Europea. Además, gracias al seguimiento temporal dado se puede calcular las estadísticas de forma diaria.

La segunda base de datos que se describirá es *Lifesight Human Mobility data by time of day, aggregated to Geohash/Quadkey/hex* [Lifesight Human Mobility]. Esta base de datos es similar a la anteriormente explicada, pero sus registros se centran en el lugar visitado en lugar del visitante. También está compuesta de los datos de países de todo el mundo incluyendo España y Europa. En este caso también se encuentran atributos útiles para el proyecto. Estos atributos se muestran en 2.5:

Cuadro 2.5: Atributos Lifesight Human Mobility data

VARIABLE	DESCRIPCIÓN
Location Id	Id identificativo del lugar visitado.
Location Name	Dirección del lugar visitado.
Location Day of the week	Día en el que fue visitado.
Number of visitors	Número de visitantes.
Visitor Home	Lugar de origen del visitante.

En este caso los atributos no proporcionan directamente los datos necesarios para este proyecto pero sería posible realizar estimaciones para obtenerlos como resultado.

2.1.2.7. DataCore

Esta base de datos la ofrece la empresa Data-Core en la plataforma Datarade [Location-based data]. Esta base de datos está formada por la información recopilada de las principales redes sociales de la actualidad como son Facebook, Instagram, TikTok y Twitter y contiene datos globales donde se incluye España y Europa. Tanto el precio como los detalles metodológicos de la base se realizan de forma personalizada con la empresa, seleccionando la información necesarias para el estudio correspondiente.

Con esto finaliza la primera temática del proyecto al no poder utilizar estos datos. .

2.2. Segunda temática:

En está sección se describirá la segunda temática del proyecto, la movilidad por carretera con los aeropuertos.

2.2.1. Objetivos

En esta temática se pueden observar dos principales objetivos. El primero de ellos es la mejora de la metodología SIR para la propagación del Coronavirus en España, concretamente detallando la estimación de personas con riesgo infectado que llegan a cada aeropuerto del país antes de efectuar el trayecto. El segundo objetivo de esta temática es el análisis de un modelo probabilístico que contraste la relación entre la movilidad con origen los aeropuertos de España y la incidencia en los municipios de destino.

Para alcanzar estos objetivos principales se han estructurado una serie de objetivos secundarios necesarios. Estos objetivos mencionados son:

- Revisión Bibliográfica.
- Búsqueda e identificación de fuentes de información: Para alcanzar los objetivos principales es necesario que el estudio esté respaldado de una serie de datos veraces, contrastados y documentados. Para mejorar la reproducción, extensión y uso de estos modelos se han buscado datos ofrecidos por instituciones públicas del estado.
- Tratamiento de datos: El uso de datos puros,sin tratar, es una tarea muy complicada o, en la mayoría de casos, imposible. Por tanto, la limpieza, estructuración y corrección de lo datos es una tarea fundamental. En este paso se pretende unificar datos, formatearlos, eliminar ceros, añadir variables etc.
- Elaboración de la base de datos: Una vez obtenido unos datos estructurados, limpios y con la información necesaria es momento de introducirlos en la base de datos. En este paso se pretende estructurar la base de datos de forma intuitiva y fácil de usar.
- Modelado: En este paso se utilizarán los datos contenidos en la base de datos creada para poder elaborar los modelos propuestos.
- Análisis de resultados: En este paso se interpretarán los resultados obtenidos en los modelos.

2.2.2. Búsqueda e identificación de fuentes de información

Aunque en este proyecto se pueden observar dos objetivos principales, para ambos es necesario una información similar. Tanto para detallar la movilidad hacia los aeropuertos, con su respectivo riesgo exportado, como para evaluar la relación entre movilidad con origen aeropuertos y la incidencia en los destinos es necesario encontrar la siguiente información:

- Movilidad por carretera en España: Un pilar fundamental para desarrollar los objetivos del proyecto es encontrar información detallada sobre la movilidad de

personas en España. Esta información debe ser lo más específica, tanto temporal como geográfica, para poder realizar el modelaje con el máximo detalle posible.

- Datos de incidencia en España: Un segundo pilar en este proyecto es la información sobre la propagación en el país. Esta información se debe ajustar lo máximo posible a las dimensiones de los datos encontrados sobre movilidad. Dentro de estos datos se debe encontrar alguna métrica que cuantifique el número de infectados por unidad de tiempo.
- Datos geográficos de España: Ya que ambos objetivos analizan el flujo de personas y la propagación en el interior del país es necesario mantener un buena estructura geográfica para intentar analizar con el máximo de detalle la información encontrada sobre movilidad e incidencia. Dentro de esta, es necesario encontrar información sobre la división de España a nivel de comunidad autónoma, provincias, municipios o la situación de los principales aeropuertos. Para explicar la búsqueda e identificación de la información geográfica necesaria para este proyecto la agruparemos junto a los datos de movilidad e incidencia. Se ha estructurado de esta manera porque a lo largo del proyecto el nivel de detalle geográfico se ha ajustado a los de estos dos tipos de datos.
- Riesgo Importado a España: Para realizar el segundo objetivo mencionado, modelado probabilístico sobre la relación entre la movilidad con origen aeropuertos y la propagación del virus, es necesario tener datos sobre los vuelos con destino España y la respectiva incidencia en el país de origen.

2.2.2.1. Datos de movilidad y estructura geográfica

El objetivo principal de la búsqueda de datos de movilidad era encontrar un *dataset* que especificase el origen y destino geográfico del flujo de personas en el interior de España, número de personas que realiza estos desplazamientos y la fecha o franja temporal en la que se realizan.

Respecto a la estructura geográfica era necesario tener la información más detallada posible de la división de territorios del país para así utilizar los datos de movilidad de la forma más cómoda y eficaz posible.

Para obtener toda esta información se encontró un estudio realizado por el Instituto Nacional de Estadística [INE Movilidad] durante los años 2019, 2020 y 2021. Aunque el objetivo principal de este proyecto varió durante su desarrollo, en todas sus etapas se estudió la movilidad en España a través de la telefonía móvil. El análisis estadístico de la información espacio-temporal de los teléfonos móviles intenta sustituir la metodología tradicional obtenida a partir de censos de Población y Vivienda. El estudio en total se puede dividir en cuatro etapas con leves diferencias que se explicarán más adelante.

Las fuentes de estos datos han sido las tres principales operadoras de telefonías móviles de España. En el estudio se especifica como estas operadoras abarcan la gran mayoría de cuota de mercado. A continuación se muestra la Tabla 2.6 con los datos oficiales publicados por la Comisión Nacional de los Mercados y la Competencia (julio de 2019).

Cuadro 2.6: Cuota de Mercado por Operador

Operador	Terminales	Cuota de Mercado
Movistar	16.270.132	30,3 %
Orange	13.705.972	25,5 %
Vodafone	12.293.129	22,9 %
Grupo Masmóvil	7.018.220	13,1 %
Operadores móviles virtuales	4.471.248	8,3 %
Total	53.758.801	100 %

Como se puede observar las tres principales compañías abarcan en total un 78,6 % de los terminales móviles de España. También se especifica que, además de abarcar este gran porcentaje de mercado, las compañías se distribuyen de forma homogénea por todo el territorio español.

Respecto a los ámbitos de la investigación, se especifica que el ámbito poblacional está constituido por todo teléfono con numeración nacional excluyendo de esta manera todo número extranjero o con *roaming*. El ámbito geográfico es toda España dividiéndola en 3.214 distritos preestablecidos. Estos distritos son denominados áreas de movilidad INE y cumplen una serie de características. Si un municipio supera los 5.000 empadronados y no supera los 50.000 será catalogado como una área. Si no se supera este mínimo se agrupará con otros municipios de características similares para que en su totalidad cumplan este requisito. Por contra parte, si un municipio supera el máximo se dividirá por barrios o distritos. Por otro lado, durante todo el proyecto se ha establecido un umbral mínimo con el que decidir si mostrar o no un flujo origen-destino. Es decir, si el flujo de personas origen-destino en un intervalo de tiempo es menos que el umbral dado no se utilizará en el estudio. Esta decisión, se ha tomado con el objetivo de garantizar la privacidad de estas personas.

Aparte de toda la información respecto al flujo, que es el principal objetivo de este estudio, se añadió mucha información relacionada con las áreas de movilidad.

Como se ha comentado anteriormente, este proyecto se puede dividir en distintas partes o matrices en las que se analizan los datos de forma distinta dependiendo de su objetivo. En lugar de explicar con detalle cada una de las matrices del proyecto solo se explicará la única que se ha utilizado. La matriz utilizada es llamada de Población Cotidiana. En esta matriz para cada área de movilidad se proporcionan los siguientes datos:

- Población residente obtenida del Padrón y actualizada el 1 Enero de 2020.
- Población residente encontrada durante el día en otra área.
- Población no residente que se encuentra en ese área.
- Población total.
- Diferencia entre la población que entra y que sale.
- Flujo que tengo origen o destino en el área considerada.
- Flujo entra las principales ciudades.

Una vez explicado con detalle las características principales del estudio encontrado se

identificarán los archivos que proporciona este estudio y los atributos que contienen estos datos. Los archivos o conjunto de archivos proporcionados serían de tres tipos:

- Descripción de las áreas de movilidad y su población: Se especifica con detalle todas las características de las áreas de movilidad. Conjunto de archivos de tablas diferenciadas por el tramo temporal en el que se usaron. Estos archivos si han sido usados en el estudio y se explicarán sus variables a detalle.
- Límites geográficos de las áreas de movilidad y provincias: Conjunto de archivos en formato *shape* para detallar las coordenadas geográficas de cada área de movilidad. El uso de estos datos no ha sido necesario para el proyecto.
- Información detallada: Conjunto de archivos en los que se muestra la movilidad. El conjunto esta dividido por el nivel territorial de la movilidad y temporalmente por días. Este conjunto de archivos si ha sido utilizado y ha sido de gran importancia.

Una vez se ha comentado qué archivos se han usado, se mostrarán las Tablas 2.7 y 2.8 las variables encontradas en cada uno de ellos.

Cuadro 2.7: Atributos datos Distritos Movilidad

VARIABLE	DESCRIPCIÓN
CUMUN	Código (INE) de Municipio
CPRO	Código (INE) de Provincia
NPRO	Nombre de la Provincia
NMUN	Nombre del Municipio
AREA GEO	Tipo de área geográfica
ID AREA GEO	Identificador del área geográfica
POB AREA GEO	Población empadronada a 1 de enero de 2019 en el área geográfica
COD LITERAL SCD	Nombre del SCD
ID GRUPO	Identificador (INE) del área de movilidad
POB GRUPO	Población empadronada a 1 de enero de 2019 en el área de movilidad
ID COMPLETO GRUPO	Identificador y Nombre completo del área de movilidad
LITERAL GRUPO	Nombre que se le ha dado al área de movilidad

Cuadro 2.8: Atributos de Movilidad

VARIABLE	DESCRIPCIÓN
CELDA ORIGEN	Id área origen
NOMBRE CELDA ORIGEN	Nombre área origen
CELDA DESTINO	Id área destino
NOMBRE CELDA DESTINO	Nombre área destino
FLUJO	Número de personas

Para resumir, con todos estos datos que ha proporcionado este estudio ya es posible extraer mucha de la información necesaria para el proyecto. En concreto, la información extraída es la división del territorio en distritos (con sus respectivos datos asociados) y la movilidad en España entre distritos.

Por último, con estos datos se puede dividir el territorio español en distritos y conocer su respectiva movilidad pero no es posible conocer la movilidad hacia o desde aeropuertos. Por tanto, se decidió que cada aeropuerto iba a estar representado por el distrito al que pertenece. Con esta idea fue necesario encontrar todos los aeropuertos de España y su localización. Esta información fue proporcionada por Aena TAL, empresa española encargada a la gestión de los aeropuertos españoles. Con esto encontramos todos los aeropuertos y helipuertos tanto de pasajeros como de carga. Se decidió introducir todos en nuestra base de datos. En la Figura 2.1 se muestran todos los aeropuertos españoles, siendo los gestionados por AENA los introducidos en el proyecto.

Figura 2.1: Mapa aeropuertos España

2.2.2.2. Datos de Incidencia y estructura geográfica

El objetivo principal de esta búsqueda es encontrar un *dataset* de la incidencia en España al nivel de detalle geográfico y temporal más pequeño posible. Por tanto, se intentó encontrar la incidencia diaria a nivel de municipio en toda España. La problemática de esta búsqueda fue no encontrar una fuente unificada de todos los municipios del país. Es decir, cada comunidad autónoma proporcionaba sus respectivos informes y datos de incidencia por separado. Esto provocó una gran diferencia en estructuras y variables entre los datos encontrados. Para algunas comunidades autónomas ha sido imposible encontrar los datos necesarios. El no encontrar datos de incidencia a nivel municipal en ciertas comunidades autónomas provocó tener que asociarle a municipios incidencias a un nivel más alto como a nivel de área sanitaria o incluso nivel provincial. Por tanto, también se buscó datos a nivel provincial para asegurar que todo municipio tenga asociado como mínimo una incidencia.

La importancia de decidir un periodo temporal para el estudio es una tarea muy

importante. Se tiene que analizar las posibles diferencias que aparecerán dependiendo del periodo escogido. También se tiene que tener en cuenta la disponibilidad de datos en ese periodo. Por tanto para el primer objetivo, la mejora de la metodología SIR, interesa la mayor disponibilidad de los datos y se eligieron los meses de Junio y Julio de 2021 ya que un gran número de comunidades autónomas tenían datos disponibles. Para el segundo objetivo, analizar la relación entre la movilidad desde aeropuertos y la incidencia en el destino, interesa analizar el inicio de la pandemia pero es necesario que los datos de incidencia no estén vacíos, cosa bastantes común en las primeras fechas. Por tanto, se decidió utilizar los meses de Marzo y Abril de 2020 ya que, aunque cubre la época de confinamiento en España, sigue habiendo varios vuelos en Madrid, lugar donde se sitúa el estudio.

Provincias Con la problemática definida se empezará a explicar una a una todas las fuentes de información. La primera fuente encontrada fue un estudio del Instituto Nacional de Epidemiología que proporciona datos a nivel provincial sobre incidencia, hospitalizaciones y defunciones. El objetivo principal de este estudio es el agrupamiento de datos relacionados con el COVID-19 de cada una de las comunidades autónomas formando así un único *dataset*. La información se obtiene a partir de la declaración de los casos de COVID-19 a la Red Nacional de Vigilancia Epidemiológica (RENAVE) a través de la plataforma informática vía Web *SiViES* (Sistema de Vigilancia de España) que gestiona el Centro Nacional de Epidemiología (CNE). En la Tabla 2.9 se mostrarán las variables del *dataset* que proporciona este estudio.

Cuadro 2.9: Atributos Incidencia Provincias

VARIABLE	DESCRIPCIÓN
provincia iso	Código ISO de la provincia. NC= no consta.
fecha	Fecha de diagnóstico. Formato = AAAA-MM-DD.
sexo	Sexo. H=hombre, M= mujer, NC= no consta.
grupo edad	Grupo de edad al que pertenece.
num casos	Número de casos notificados.
num hosp	Número de casos hospitalizados
num uci	Número de casos ingresados en UCI.
num def	Número de defunciones.

Como se comentó, el objetivo de esta búsqueda es encontrar una información más detallada de la incidencia y esta se encuentra dividida por comunidades autónomas. Por tanto, primero se mostrará la Tabla 2.10 con todas las comunidades autónomas de España, si ha sido posible utilizar datos de ellas y alguna observación necesaria. Tras esto se detallarán cada comunidad de forma más detallada. Por otro lado, también fue necesario buscar información sobre la población de cada una de las provincias.

Cuadro 2.10: Comunidades Autónomas utilizadas

Comunidad Autónoma	Utilizado	Observaciones	Fuente
Andalucía	SI	Nivel de detalle: Área Sanitaria	[Datos Andalucía]
Aragón	SI	Nivel de detalle: Comarca	[Datos Aragón]
Asturias	SI	Nivel de detalle: Área Sanitaria	[Datos Asturias]
Baleares	SI	-	[Datos Baleares]
Comunidad Valenciana	SI	-	[Datos C.Valenciana]
Canarias	SI	Nivel de detalle: Islas	[Datos Canarias]
Cantabria	NO	Enlace con más detalle caído, no es posible descargar.	[Datos Cantabria]
Castilla y León	SI	Nivel de detalle: Agrupación de municipios	[Datos Castilla y León]
Castilla La Mancha	NO	No existe histórico, solo datos de los últimos 14 días y de personas +60 años.	[Datos Castilla y la Mancha]
Cataluña	SI	-	[Datos Cataluña]
Ceuta	NO	Se decide utilizar directamente datos provinciales.	
Extremadura	NO	Enlace a datos libres caído, ninguna información detallada.	[Datos Extremadura]
Galicia	SI	Nivel de detalle: Área Sanitaria	[Datos Galicia]
La Rioja	NO	No encontrado datos de la fecha necesaria.	[Datos La Rioja]
Madrid	SI	-	[Datos Madrid]
Melilla	NO	Se decide utilizar directamente datos provinciales.	
Murcia	NO	Datos vacíos.	[Datos Murcia]
Navarra	NO	Se decide no utilizar ya que tiene un 90 % de municipios con valores en 0.	[Datos Navarra]
País Vasco	NO	Formato JSON: Error al transformar.	[Datos País Vasco]

Andalucía La primera comunidad autónoma que se comentará es Andalucía. Los datos relacionados a esta comunidad autónoma fueron encontrados en unos estudios elaborados por la Consejería de Salud y Familias [Datos Andalucía] para dar seguimiento a la COVID-19 en Andalucía. Desde el comienzo de la pandemia se ha venido publicando la información diaria sobre número y tasas de casos confirmados, hospitalizados, curados y fallecidos desagregadas por grupos de edad, sexo y territorio. Aquí se encuentran una gran variedad de datos disponibles relacionados con el COVID-19 divididos por territorios o fechas. En concreto los datos encontrados a nivel municipal solo mostraban datos de los últimos 14 días por lo que fue rechazada. Por esto se decidió elegir los datos con nivel de detalle mas pequeño, en este caso fue nivel distrito Sanitario. Un distrito sanitario o Áreas de Gestión Sanitaria,son modelo de organización que gestiona de forma unitaria los niveles de atención primaria y hospitalaria, en una demarcación territorial específica. Es decir, cada distrito sanitario representará un conjunto de municipios. A continuación se mostrará la Tabla 2.11 con los atributos de los datos obtenidos.

Cuadro 2.11: Atributos Andalucía

VARIABLE	DESCRIPCIÓN
Territorio	Nombre de Área Sanitaria
Fecha diagnóstico	Fecha de diagnóstico. Formato = DD/MM/AAAA.
Número de Casos	Nº personas infectadas.
Número de Defunciones	Nº personas fallecidas.
Número de Hospitalizaciones	Nº personas hospitalizadas.

Ya que el nivel geográfico en estos datos son las áreas sanitarias es necesario buscar información sobre ellas y los municipios que lo forman. En la propia consejería encontramos la información buscada, identificación y municipios que forman cada distrito sanitario.

Aragón La siguiente comunidad autónoma es Aragón. Sus datos relacionados con el covid son proporcionados por el Gobierno de Aragón [Datos Aragón]. En este estudio se muestran los datos diarios de número de casos totales, provinciales, comarcales y por distritos sanitarios, número de pruebas realizadas y ocupación hospitalaria. Como se observa, el nivel de detalle es la comarca por lo que será necesario buscar información relacionada a ellas. Concretamente, los archivos encontrados están separados por día y están formadas por tablas independientes clasificadas por género y edad y distribuciones por provincia, Sector, Comarca y zona básica de salud. La tabla utilizada es la distribución por comarca y sus atributos se muestran en la Tabla 2.12.

Cuadro 2.12: Atributos Aragón

VARIABLE	DESCRIPCIÓN
Zona Básica de Salud	Nombre de la comarca
Casos	Casos diarios.

Como en el caso anterior es necesario identificar y conocer los municipios de cada comarca. Esta información la proporciona el gobierno de Aragón.

Asturias La siguiente comunidad autónoma será Asturias. Los datos de esta comunidad fueron encontrados en un estudio realizado por el Observatorio de salud en Asturias [Datos Asturias]. Las fuentes de este estudio son vigilancia epidemiológica, SESPA, descargas de laboratorios y GO.DATA. En este estudio se proporciona una información muy variada y detallada. Se disponen de datos a nivel provincial, distrito sanitario y de concejos de forma diaria. Concretamente se escogen los datos a nivel de distrito sanitario, muy detallados y completos. Estos datos poseen 72 atributos clasificados en variables de incidencia, de mortalidad, relativas a mayores de 65 años, relativas a hospitales, relativas a pruebas y de rastreo. Como son un gran número de variables se van a explicar las de interés, variables de incidencia. Estos atributos se muestran la Tabla 2.13.

Cuadro 2.13: Atributos Asturias

VARIABLE	DESCRIPCIÓN
fecha	Fecha diagnóstico. Formato= YYYY-MM-DD
casos diarios	Casos diarios.
crec casos intradia	% de crecimiento de casos respecto al día anterior.
tend casos	Media de casos.
casos acum	Casos acumulados hasta la fecha.
porc positivos pobl obs	% de población como positiva.
casos 7D	Casos acumulados últimos 7 días.
casos 14D	Casos acumulados últimos 14 días.
IA7	Incidencia acumulada por 100.000 habitantes en los últimos 7 días.
crec IA7	% de crecimiento de la variable IA7.
ratio IA7	Ratio de crecimiento de la incidencia.
porc positivos pobl obs 7D	% población que ha dado positiva en los últimos 7 días.

Como en varias comunidades también fue necesaria la información sobre las áreas sanitarias.

Islas Baleares La siguiente comunidad autónoma ha mencionar es las Islas Baleares. Como con el resto de las comunidades estos datos los ofrece el Gobierno de las Islas Baleares [Datos Baleares]. Esta organización ofrece mucha información sobre el COVID-19, uno de los estudios es Datos COVID-19 Casos Confirmados Illes Balears (Por Municipios). En este conjunto de datos se ofrece los casos diarios de cada municipio desde el 9 de Febrero de 2020 hasta la actualidad. En la Tabla 2.14 se mostrarán los atributos de estos datos:

Cuadro 2.14: Atributos Islas Baleares

VARIABLE	DESCRIPCIÓN
Data diagnòsic	Fecha de diagnóstico. Formato=DD/MM/AAAA
Codi illa	Código de la isla a la que pertenece el municipio.
Illa	Nombre de la isla a la que pertenece el municipio.
Codi municipi	Código del municipio.
Municipi	Nombre del municipio.
Sexe	Sexo. home = Hombre, dona = Mujer.
Gran franja d'edat	Franja de edad.
Nombre casos	Número de casos diarios.

Comunidad Valenciana La siguiente comunidad autónoma en la lista es la Comunidad Valenciana. Los datos relacionados a esta comunidad los pone a disposición la *Conselleria de participación, transparencia, cooperación y calidad democràtica* [Datos C.Valenciana]. La fuente de estos datos son de la *Conselleria de Sanitat Universal i Salut Pública*. Esta información esta dividida en un archivo cada tres días desde el 31 de enero de 2020. Por otro lado, el nivel de detalle geográfico de estos datos es por municipio. Para detallar más esta información se mostrarán todos sus atributos en la Tabla 2.15.

Cuadro 2.15: Atributos C.Valenciana

VARIABLE	DESCRIPCIÓN
Casos PCR	Número de casos de COVID-19 acumulados.
Casos PCR + 14 dies	Número de casos de COVID-19 acumulados en los últimos 14 días.
CodMunicipio	Código del municipio.
Defuncions	Número de defunciones por COVID-19.
Incidència acumulada PCR+	Incidencia acumulada de casos PCR positivos por 100.000 habitantes.
Incidència acumulada PCR+14	Incidencia acumulada de los últimos 14 días de casos PCR positivos por 100.000 habitantes.
Municipi	Nombre del municipio.
Taxa de defunció	Tasa de mortalidad por 100.000 habitantes.

Canarias Para los datos relacionados con el COVID-19 el gobierno de Canarias ofrece una serie de estudios diferentes. Uno de ellos es el Datos epidemológicos COVID-19 [Datos Canarias]. Esto es un conjunto de datos que ofrece la incidencia diaria por municipio desde 1 de Enero de 2021 hasta la actualidad. En la Tabla 2.16 se verán los atributos de estos datos.

Cuadro 2.16: Atributos Canarias

VARIABLE	DESCRIPCIÓN
fecha_datos	Fecha de actualización. FORMATO = MM/DD/AAAA
isla	Isla a la que pertenece el municipio.
municipio	Nombre del municipio.
sexo	Sexo
grupo edad	Franja de edad a la que pertenecen.
fecha caso	Fecha de diagnóstico. FORMATO = MM/DD/AAAA
fecha fallecido	Fecha de defuncion. FORMATO = MM/DD/AAAA
fecha curado	Fecha de curación. FORMATO = MM/DD/AAAA
estado caso	Estado del infectado.

Castilla y León La información de contagios de la comunidad de Castilla y León la pone a disposición la dirección general de planificación sanitaria investigación e innovación [Datos Castilla y León]. Se ofrecen datos por área sanitaria desde el 30 de Febrero de 2020 hasta la actualidad. Estos datos tienen 33 atributos, muchos de ellos vacíos o a 0. Por tanto, en la Tabla 2.17 se mostrarán únicamente los más necesarios.

Cuadro 2.17: Atributos Castilla y León

VARIABLE	DESCRIPCIÓN
FECHA	Fecha de diagnóstico. FORMATO = AAAA-MM-DD.
GERENCIA	Código de la gerencia.
NOMBREGERENCIA	Nombre de la gerencia.
CS	Código del centro sanitario.
CENTRO	Nombre del centro sanitario.
PCR POSITIVOS	Número de casos diarios.
X GEO, Y GEO	Coordenadas geográficas.
PROVINCIA	Nombre de la provincia.
MUNICIPIO	Nombre de los municipios que pertenecen al área.

En este caso no es necesario buscar más información sobre los centros sanitario ya que el propio *dataset* los tiene.

Cataluña En el caso de la comunidad autónoma de Cataluña los datos los proporciona el Departamento de Salud de Cataluña [Datos Cataluña]. En este *dataset* se muestran la incidencia diaria de municipios desde el 1 de Marzo de 2020 hasta la actualidad. En la Tabla 2.18 se muestran sus atributos.

Cuadro 2.18: Atributos Cataluña

VARIABLE	DESCRIPCIÓN
TipusCasData	Fecha de diagnóstico. FORMATO = DD/MM/AAAA.
ComarcaCodi	Código de la comarca.
ComarcaDescripcio	Nombre de la comarca.
MunicipiCodi	Código del municipio.
MunicipiDescripcio	Nombre del municipio.
DistricteCodi	Código del distrito.
DistricteDescripcio	Nombre del distrito.
SexeCodi	Código sexo. 1=Mujer, 0=Hombre.
SexeDescripcio	Sexo. Dona = Mujer, Home = Hombre
TipusCasDescripcio	Prueba realizada.
NumCasos	Casos

Galicia Para Galicia los datos los proporciona la *Xunta de Galicia* [Datos Galicia]. En este caso no es posible encontrar un histórico a nivel de municipios ya que solo se dispone de los últimos 14 días. Por tanto, se escoge un *dataset* a nivel de Distrito Sanitario con casos diarios desde el 6 de Marzo de 2020 hasta la actualidad. Los atributos de este *dataset* se muestran en la Tabla 2.19.

Cuadro 2.19: Atributos Galicia

VARIABLE	DESCRIPCIÓN
Fecha	Fecha de diagnóstico. FORMATO =AAAA-MM-DD.
Area Sanitaria	Nombre del área sanitaria.
Personas Infectadas	Número de personas infectadas.

En este caso, se vuelve a necesitar información extra sobre los distritos sanitarios y sus respectivos municipios.

Madrid Por último, se hablará de la comunidad de Madrid. Los datos encontrados los proporciona la Consejería de Sanidad de la Comunidad de Madrid y la fuente es Red de Vigilancia Epidemiológica de la Comunidad de Madrid [Datos Madrid]. En estos datos se muestra la incidencia acumulados en los últimos 14 días de cada municipio y distrito cada semana desde el 2 de Julio de 2020 hasta el 29 de Marzo de 2022. Por otro lado, se debían encontrar datos también de las primeras semanas de pandemia. Para este caso la información encontrada sigue siendo a nivel municipal y ofreciendo la incidencia acumulada en los últimos 14 días pero esta vez de forma diaria.

Ambos *datasets* tienen los mismos atributos y se muestrán en la Tabla 2.20.

Cuadro 2.20: Atributos Madrid

VARIABLE	DESCRIPCIÓN
municipio distrito	Nombre del municipio/distrito
fecha informe	Fecha de diagnóstico. FORMATO = AAAA/MM/DD hh:mm:ss
Casos confirmados últimos 14días	casos acumulados en los últimos 14 días.
tasa incidencia acumulada últimos 14días	Incidencia de casos por 100.000 habitantes
casos confirmados totales	Casos acumulados totales.
tasa incidencia acumulada total	Incidencia casos totales por 100.000 habitantes

Al trabajar con los datos se observó que el municipio de Madrid a su vez están dividido en los distritos de la ciudad. Sin embargo, los distritos de movilidad están divididos por los barrios de la mismas. Por tanto, se tuvo que buscar información sobre los distritos de Madrid y los barrios que lo forman.

2.2.2.3. Riesgo Importado

Para el segundo objetivo de este proyecto es necesario trabajar con el riesgo importado que llega a los aeropuertos de España, concretamente al de Madrid. Se eligió este aeropuerto por ser el aeropuerto con más número de vuelos y, por tanto, es posible mostrar mejor como estos afectan a la incidencia de la comunidad. Por tanto el objetivo es buscar el riesgo importando diario en el aeropuerto de Barajas en el mes de Marzo. Para encontrar datos relacionados se utilizó la base de datos ya creada por C.R.I.D.A. Esta es una base de datos orientada a grafos que contiene la información de los vuelos realizados hacia Europa. Por tanto, en un principio es posible obtener la información necesaria en esta base de datos.

La estructura de esta base de datos es bastante sencilla, a continuación se describirá la información que contiene y se mostrará en la Figura 2.21 sus componentes. En resumen, en la base de datos aparecen los aeropuertos a nivel mundial con la información de cada vuelo con origen o destino este. A su vez, cada aeropuerto tiene asociado una provincia o estado con sus respectivas regiones y un país con su respectivo continente. Por último, todas estas unidades geográficas tienen recogidas su incidencia ordenada temporalmente.

Cuadro 2.21: Nodos C.R.I.D.A

Nodo	Descripción
Flight	Vuelo con origen y destino ciertos aeropuertos y realizados cierto día.
Airport Operation Day	Entidad encargada de agrupar los vuelos realizados en un día y con origen o destino un aeropuerto.
Airport	Aeropuerto perteneciente a una región y a un país y con una serie de vuelos totales diarios.
Province/State	Provincia o Estado formada por una serie de regiones y perteneciente a un país. Tienen cierta incidencia a lo largo del tiempo.
Country	País formado por provincias y perteneciente a un continente. Tienen cierta incidencia a lo largo del tiempo.
Region	Región perteneciente a un país. Tienen cierta incidencia a lo largo del tiempo.
Continent	Continente al que pertenecen un conjunto de países.
Report	Información sobre la incidencia de países, provincias o regiones.

Por último, en la Figura 2.2 se mostrará el esquema completo de la base de datos de CRIDA obtenido de un propio artículo de la empresa.

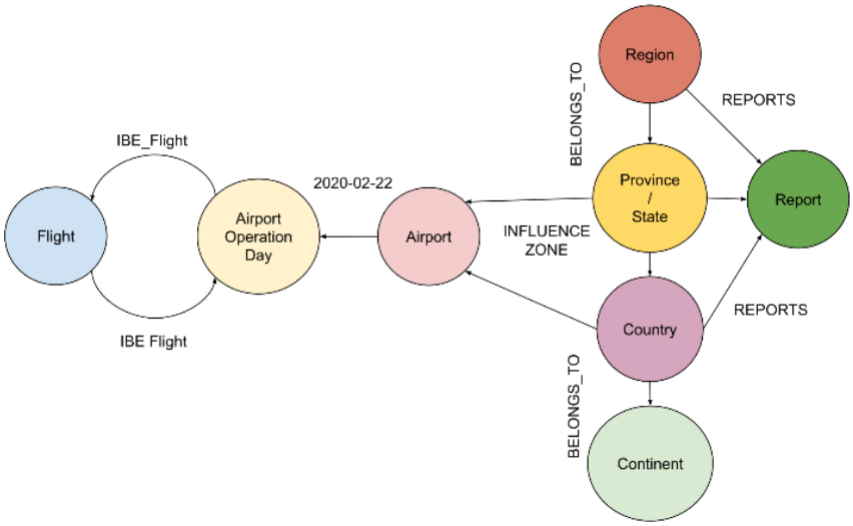

Figura 2.2: Esquema base de datos CRIDA

2.2.3. Tratamiento de datos

En esta sección se explicará el preprocesamiento de los datos antes de ser introducidos en la base de datos. Esta tarea se complica al observar el gran número de archivos de origen distinto, con distintos atributos, formatos etc. Para resumir todos los archivos con los que se comienza a realizar el preprocesamiento se mostrará la Tabla 2.22 con todos ellos.

Cuadro 2.22: Archivos preprocesamiento

Comunidad Autónoma	Archivos
Distritos	Distritos.csv
Movilidad	conjunto{Movilidad+Dia.csv}
Provincias	IncidenciaProvincias.csv
Andalucía	IncidenciaAndalucia.csv DistritosSanitariosAndalucia.csv
Aragón	IncidenciaAragon.csv ComarcasAragón.csv
Asturias	IncidenciaAsturias.csv AreasSanitariasAsturias.csv
Baleares	IncidenciaBaleares.csv
Comunidad Valenciana	conjunto{IncidenciaValenciana+Dia.csv}
Canarias	IncidenciaCanarias.csv
Castilla y León	IncidenciaCYL.csv
Cataluña	IncidenciaCataluña.csv
Galicia	IncidenciaGalicia.csv AreasSanitariasGalicia.csv
Madrid	IncidenciaMadrid1.csv IncidenciaMadrid2.csv
RiesgoImportado	RiesgoImportado.csv

Para mantener la misma estructura se dividirá el tratamiento de los datos en tres secciones: Movilidad, estructura geográfica e incidencia.

2.2.3.1. Preprocesamiento Movilidad

La información esencial que se busca obtener de este conjunto de tablas son dos archivos con la información organizada por fechas y que detalle el flujo de movilidad entre los distritos origen y destino. Se recuerda que especificamos dos archivos uno para cada periodo temporal que será usado en cada objetivo. Para todo este preprocesamiento se utilizó *PYTHON 3.9.7* y demás librerías, Pandas principalmente. Para conseguir esta información se realizaron una serie de pasos que se detallarán a continuación:

- Creación de tabla de aeropuertos: Con todos los aeropuertos de España identificados, se creó una tabla y se enlazó cada aeropuerto (variable AEROPUERTO) manualmente al distrito/área al que pertenece añadiendo las variables: DISTRITO Y DISTRITO ID. Estás variables representan el nombre e id del distrito al que pertenece el aeropuerto y fueron obtenidas del archivo *Distritos.csv*.
- Unificación de archivos: Unificar, por un lado, los 8 archivos y los 39 archivos de movilidad de cada espacio temporal en dos únicos archivos. Al ir añadiendo un archivo se añade la variable FECHA de su día correspondiente con formato AAAA-MM-DD.
- Eliminación de variables: Trabajando con estos dos único archivos se intentó reducir al máximo las variables eliminando: Comunidad Autónoma de residencia, Comunidad Autónoma de destino, Provincia de residencia,Provincia de destino,

Nombre área de residencia y Nombre área de destino. El resultado son dos tabla con 4 variables: Origen, Destino, Flujo y Fecha. Siendo Origen y Destino los id de sus respectivos distritos.

- Filtrado: En este último paso se intentó reducir el tamaño de los datos eliminando información innecesaria. Para ello, en la primera tabla se eliminó todo dato cuyo destino no fuese un aeropuerto. Y por otro lado, en la segunda tabla se eliminó todo dato que no tuviese como origen un aeropuerto. Toda la información sobre los aeropuertos fue sacada de la tabla *Aeropuertos.csv*.
- Diferenciación de archivos: Aunque es posible unificar ambos archivos en uno solo se decidió dividirlo en dos por control de errores y visibilidad de los datos. Al archivo con la movilidad destino aeropuertos se le llamó *Movilidad.csv* mientras que el de origen aeropuertos se le llamó *Expansión.csv* .

2.2.3.2. Preprocesamiento Estructura geográfica

El objetivo principal de este procesamiento es obtener una serie de tablas diferenciadas con toda la información necesaria de cada división geográfica y su relación entre ellas. Lo primero a destacar es que se decide unificar toda división geográfica que sea un conjunto de municipios para tratarlos de una misma forma y se les denominará como distrito sanitario aunque realmente no todas lo sean. Con esto nuestro resultado esperado son 4 tablas: Distritos, Municipios, Distritos Sanitarios y Provincias.

Distritos El resultado requerido para de este procesamiento era obtener una única tabla identificando los distritos y condensando la información de cada una. Toda la información se obtuvo a partir del archivo *distritos.csv*. Los pasos a seguir fueron los siguientes:

- Eliminación de variables: En este paso se eliminaron varias variables redundantes y relacionadas con los municipios (se utilizarán en otra tabla). Se eliminaron todas las variables excepto: NPRO,ID GRUPO, POB GRUPO y LITERAL GRUPO. Estas variables pasaron a llamarse respectivamente: Provincia, DistritoID, Población y Distrito.
- Variable DistritoMunicipal: Para añadir la información de los distritos de Madrid se utilizó la información obtenida sobre ellos y se identificó cada barrio con su respectivo distrito en la nueva variable DistritoMunicipal.
- Variable PoblaciónDistritoMunicipal: La población de estos distritos Municipales era necesaria y teniendo la población de los barrios (distritos de movilidad) se calculó la población de cada uno de ellos en la variable PoblacionDistritoMunicipal.

Con todo esto obtenemos una única tabla con 6 variables: Provincia, DistritoId, Distrito, Población, DistritoMunicipal y PoblaciónMunicipal, siendo estas dos últimas solo para los barrios de Madrid.

Municipios Respecto a los datos de municipios se podrían haber introducidos en la propia tabla Distritos.csv ya que en ella se unificaban la información de ambas unidades geográficas pero para facilitar la comprensión y uso de ellas se decidió dividirla en dos. Por tanto, el objetivo principal de este tratamiento es la creación de

una única tabla con la información necesaria de los municipios. Los pasos de este preprocesamiento fueron los siguiente:

- Eliminación de variables: En este paso se volvieron a eliminar las variables innecesarias para mostrar la información sobre municipios. En este caso se eliminaron todas las variables excepto: NPROV,NMUN, POB AREA GEO y ID GRUPO. Respectivamente estas variables pasaron a llamarse: ProvinciaId, Municipio, Población y DistritoId. La variable ProvinciaId se modificó y pasó de ser el literal de la provincia a su código ISO.
- Formateo de Municipio: Al no tener una etiqueta o id común para cada municipio fue necesario utilizar el literal de cada uno de ellos como id. Con esto aparecieron numerosos problemas al encontrar en los datos de incidencia modificaciones del nombre de los municipios como: Literales sin acentos, separación de los artículos de forma distinta, literales en mayúsculas o algunos problemas con los idiomas. Por esto se decidió parsear todos los literales de la misma forma. El formato de un literal seria el siguiente: *'Artículo(OPT) + Nombre/ Artículo2Idioma(OPT) + Nombre2Idioma(OPT)'* con todas las primeras letras en mayúsculas y sin acentos.

Con todo esto se obtuvo una tabla con las siguientes variables: Municipio, DistritoId, ProvinciaId y Poblacion.

Provincias El objetivo de este preprocesamiento es obtener una única tabla que identifique cada provincia de España y su respectiva población. Por tanto, se creó una tabla con tan solo tres variables: ProvinciaId, Provincia y Población. La primera variable es el código ISO de cada provincia, la segunda el nombre y la tercera la población. El código de Navarra 'NA' se modificó al identificarlo python como un dato vacío.

Distritos Sanitarios El objetivo principal sería tener una única tabla con la información necesaria de los distritos sanitarios, comarcas y áreas sanitarias. Para cada uno se calcularon las poblaciones de cada una de las unidades geográficas sumando las poblaciones de cada uno de los municipios que pertenecen a ella. Por tanto, el resultado es una tabla con dos variables DistritoSanitario y Poblacion.

2.2.3.3. Preprocesamiento Incidencia

El preprocesamiento de la incidencia ha sido una tarea bastante tediosa ya que se tienen un gran número de archivos con atributos y formatos muy distintos. La forma en la que se ha decidido estructurar estos datos es en tres archivos: IncidenciaMunicipios, IncidenciaDistritosSanitarios e IncidenciaProvincias.

IncidenciaMunicipios Como se ha mencionado se intenta unificar todos los datos de incidencia que tengan nivel de detalle municipios. Estos son los de las provincias de: Baleares,Comunidad Valenciana,Canarias,Cataluña y Madrid. A continuación se intentarán resumir los pasos que se han seguido para conseguir nuestro objetivo:

- Unificación[U]: Este caso es solo opcional si existe un archivo por día registrado. En este caso es necesario unirlos en un único archivo.

- Formateo de fechas [F.F]: Tiene un objetivo parecido al anterior, todas las fechas de este estudio deben tener el mismo formato: 'AAAA-MM-DD'.

- Filtrado [F]: Este es el caso contrario al anterior y por tanto también es opcional. Se ha realizado en el caso de que exista un único archivo con un periodo temporal más grande de lo deseado. En este caso se eliminará todo valor que no entre en esta franja deseada.

- Eliminación de Variables [E.V]: Este paso simplemente elimina esas variables que no aportan información relevante para el proyecto.Y respectivamente se les cambian el nombre a: Municipio, Fecha, Casos, Casos7 y Casos14. Dependiendo de la información que se tenga varias de estas tres últimas pueden estar vacías.

- GroupingBy[G]: Este paso es también opcional y ocurre en las tablas que dividen a la población en franjas de edad o sexo: El objetivo es unificarlas todas en una.

- Formateo de municipios [F.M]: Este paso es igual al explicado en el procesamientos de municipios. Se intenta transforma cualquier nombre de municipio en un estándar para su uso como id.

- Estimación de fechas [E.F]: Este paso es el más complejo de todos y en general su objetivo es estimar los datos para todos los días y calcular los datos que están vacíos. Esta fase se divide a su vez en dos fases. La primera fase solo se daría en el caso de que la tabla a estimar no tenga datos de forma diaria y se añadirían los datos con las fechas restantes estimando los casos con una media ponderada según la distancia a cada fecha anterior y posterior. En la segunda fase se trata de rellenar los datos vacíos dentro de las tres variables: Casos, Casos7, Casos14. Dependiendo del campo que ya tenga relleno se estimará de una manera o de otra. Si Casos esta completo se contarán los 7 o 14 últimos días. Si son otras variables se estimarán dividiéndolas entre el número de días contados.

- Por último, cuando ya tienes cada archivo limpio y estructurado se unifican en una única tabla, IncidenciaMunicipios (Es posible hacer este paso antes).

Una vez se han descrito los pasos se especificará en la Tabla 2.23 que comunidades autónomas han necesitado estos pasos. En el caso de la etapa [E.V] se muestran las variables que no se han eliminado y que se le cambiaron el nombre a Municipio, Fecha, Casos, Casos7, Casos14 respectivamente. Si en algún caso la columna está vacía se representa con "NaN".

Cuadro 2.23: Tareas IncidenciaMunicipal

Comunidad Autónoma	U	F	E.V	G	F.M	F.F	E.F
Islas Baleares	No	Si	(Municipi, Data diagnosic, Nombre casos, NaN, NaN)	(Edad,Sexo)	Si	Si	Si
Comunidad Valenciana	Si	No	(Municipi, Fecha, Casos PCR, NaN, Casos PCR+14 dics)	No	Si	No	Si
Islas Canarias	No	Si	(municipio, fecha datos, casos, NaN, NaN)	(Edad,Sexo, Estado caso)	Si	Si	Si
Cataluña	No	Si	(MunicipiDescripcio, TipusCasData, NumCasos, NaN, NaN)	(Edad,Sexo)	Si	Si	Si
Madrid	No	Si	(municipio distrito, fecha informe, NaN,NaN, casos confirmados ultimos 14dias)	No	Si	Si	Si

IncidenciaDistritosSanitarios El objetivo principal de este preprocesamiento es obtener un archivo único con la información de cada municipio, su respectivo distrito en un día determinado. Esta vez, no se explicarán una por una las fases ya que se realizan todas las nombradas anteriormente. Solamente se explicará una nueva tarea que únicamente se realiza para los distritos sanitarios.

Esta tarea es denominada Enlace Municipio-Distrito[E.M-D]. Consiste realmente en dividir cada dato de incidencia de la tabla en tantos como municipios este formado, identificándolos. Se realiza de esta manera para facilitar en un futuro su introducción en la base de datos.

De igual manera que en el apartado anterior se mostrarán en la Tabla 2.24 las comunidades autónomas a nivel distrito sanitario y que tareas se les ha tenido que aplicar. De igual forma en los datos del apartado E.V los atributos mostrados son los que se han mantenido y se les ha renombrado de la siguiente forma: Municipio, Fecha, DistritoSanitario, Casos, Casos7 y Casos14.

Cuadro 2.24: Tareas IncidenciaDistritoSanitario

Comunidad Autónoma	U	F	E.V	E.M-D	G	F.M	F.F	E.F
Andalucía	No	Si	(Municipio, Territorio, Fecha diagnóstico, Numero de Casos, NaN, NaN)	Si	No	Si	Si	Si
Aragón	Si	No	(Municipio, Fecha, Zona Básica de Salud, Casos, NaN, NaN)	Si	No	Si	No	Si
Asturias	No	Si	(Municipio, fecha, casos diarios, casos 7D, casos 14D)	Si	No	Si	No	Si
Castilla y León	No	Si	(Municipio, FECHA, NOMBREGERENCIA, PCR POSITIVOS, NaN, NaN)	Si	No	Si	No	Si
Galicia	No	Si	(Municipio, Fecha, Personas Infectadas, NaN, NaN)	Si	No	Si	No	Si

IncidenciaProvincia En el caso de la *IncidenciaProvincial.csv* se intentará describirlo de forma breve ya que es tan solo un archivo. Primero se eliminaron todas las variables excepto provincia iso, fecha, num casos y fueron renombradas como ProvinciaId, Fecha, Casos. Como en el resto de archivos se añadieron las columnas Casos7 y Casos14 y se estimaron de la misma forma. Con estos dos pasos ya se alcanzó la tabla deseada.

2.2.3.4. RiesgoImportado

Ya con la posibilidad de obtener todos los datos necesarios de la base de datos de CRIDA el objetivo principal es elaborar una *query* para obtener justo la información necesaria para el proyecto. El modelado que se va a realizar afectaría solo a la ciudad de Madrid en la franja temporal del mes de Marzo. Para esto se debe elaborar una *query* que recoja la información de los vuelos de cada día para el Aeropuerto de Madrid. A partir de los vuelos totales por día se calcularía el riesgo importado diario. Con esto ya se obtendrían la información del riesgo importado en el aeropuerto de Barajas en el intervalo de tiempo deseado. Finalmente el resultado obtenido es una tabla con dos parámetros fecha y riesgo importado. La *query* en concreto es la siguiente:

```
1 MATCH (f:FLIGHT)-->(n:AirportOperationDay)<--(a:Airport)
2 WHERE (a.airportId = 'LEMD')
3 RETURN f.dateOfArrival AS Fecha,
```

```
4  a.airportName AS Aeropuerto,
5   SUM(f.flightIfinal) AS RiesgoImportado
```

2.2.4. Elaboración de la base de datos

2.2.4.1. Neo4j y bases de datos orientados a grafos

Bases de datos orientados a grafos(BDOG) En esta sección se describirá con todo detalle la base de datos que se ha elaborado para el proyecto. Desde un primer momento se planeó realizar una base de datos orientada a grafos (BDOG). Las principales características de este tipo de bases de datos es que almacenan la información en vértices y aristas, cumpliendo con la teoría de grafos. Estos grafos se utilizan para representar interacciones complejas entre datos. Como cualquier grafo esta compuesto por los nodos o vértices que representarían las entidades de la base de datos con sus respectivos atributos y las relaciones o aristas que representarían las propias conexiones o asociaciones entre entidades.

Estas bases de datos son muy útiles para trabajar con datos con muchas interconexiones. Ejemplos de casos útiles para utilizar las BDOG son representaciones de rutas, redes sociales, redes de sistemas, estructuras de empresas etc.

Las principales ventajas que pueden tener este tipo de bases de datos están relacionadas con su rendimiento, flexibilidad y visibilidad. A diferencia de las bases de datos relacionales, las BDOG tienen guardadas las propias relaciones en la base de datos por lo que no se tienen que calcular en la consulta. Esto hace que la velocidad de búsqueda sea muy rápida. Por otro lado, las BDOG tienen una estructura de datos muy flexible. Es posible tener nodos de un mismo tipo con un número de atributos distinto. Además las propias relaciones pueden tener sus propios atributos y pueden ser con dirección o sin ellas.

Por último, las BDOG permiten una visualización constante de la estructura de los datos, si la herramienta lo permite. Esto facilita el trabajo a la hora de la creación, mantenimiento y uso de la base de datos. En la Figura 2.3 se puede observar un pequeño esquema de control de cuentas bancarias.

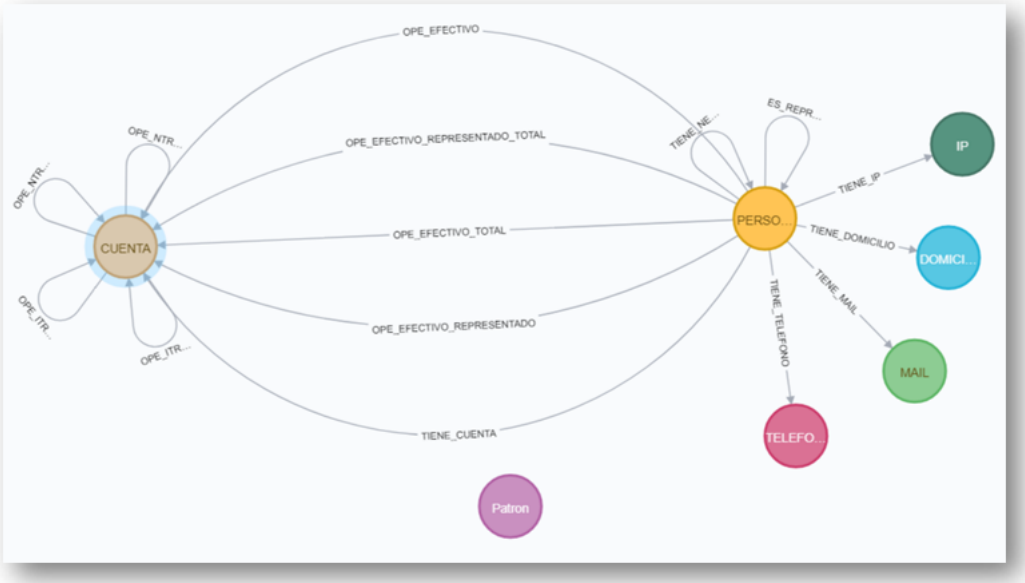

Figura 2.3: DBOG

Neo4j La herramienta que se usó para la creación y gestión de esta base de datos orientada a grafos es Neo4j. Neo4j es un software libre de BDOGs implementada en Java. Neo4j almacena datos estructurados en grafos en lugar de en tablas, es decir, la información se almacena de forma relacionada formando un grafo dirigido entre los nodos y las relaciones entre ellos. Neo4j permite acceder a sus datos de diversas formas y usando distintos lenguajes de consulta. Destacan aquí Cypher, un lenguaje que permite consultar y manipular grafos, y Gremlin, un lenguaje que permite gestionar grafos. Cypher es un lenguaje declarativo, inspirado en SQL, que permite manipular datos en Neo4j, mientras que Gremlin es un lenguaje de flujo de datos funcional que permite a los usuarios expresar recorridos o consultas de gráficos de propiedades complejas de una manera concisa.

2.2.4.2. Estructura

En esta sección se hablará sobre la estructura de la base de datos que se ha implementado. El objetivo de principal de esta base de datos es mantener una estructura fácil de comprender y que aporte toda la información necesaria de forma sencilla. Desde un inicio todo el preprocesamiento de los datos se ideó para facilitar la creación de esta base de datos aunque la información este separada en varias tablas.

Como se ha comentado en la sección anterior esta base de datos es una base de datos orientadas a grafos por lo que la principal tarea al idear una base de datos de este tipo es como transformar el problema en un conjunto de entidades o nodos y un conjunto de relaciones entre nodos.

A continuación se resumirá un poco el problema para entender la elección de entidades y relaciones. Se ha dividido el territorio Español en un gran conjunto de distritos. Estos distritos pueden estar formado por uno o más municipios. A su vez cada municipio formará parte de una provincia y opcionalmente a un área sanitaria. Por otro

lado, se tiene la incidencia a todos los niveles geográficos de forma diaria. A parte, situamos una serie de aeropuertos de España, estos aeropuertos tienen una movilidad diaria con cada distrito. Por último, desde el exterior llegan a los aeropuertos un riesgo exterior diario.

Con este resumen se pueden establecer los nodos,que se muestran en la Tabla 2.25.

Cuadro 2.25: Nodos

Nodos	Descripción
Distrito	Entidad geográfica que realiza la movilidad con los aeropuertos, esta formado por uno o mas municipios.
Municipio	Entidad geográfica. Pertenece a una provincia y puede pertenecer a un distrito sanitario. Puede tener incidencia diaria.
Distrito Sanitario	Entidad geográfica. Esta formado por municipios. Puede tener incidencia diaria.
Provincia	Entidad geográfica. Esta formado por municipios. Puede tener incidencia diaria.
Aeropuerto	Entidad geográfica. Realiza la movilidad con los distritos.
RiesgoExterior	Entidad geográfica. De forma diaria tiene una estimación de personas que llegan infectadas a cada aeropuerto.

Por otro lado, las relaciones entre las entidades se muestran en la Tabla 2.26.

Cuadro 2.26: Aristas

Aristas	Relación
Movilidad	Distrito-Aeropuerto
Expansion	Aeropuerto-Distrito
FormadoPor	Distrito-Municipio
IncidenciaMunicipio	Municipio-Distrito
IncidenciaDistritoMunicipal	Municipio-Distrito
IncidenciaDistritoSanitario	DistritoSanitario- Municipio
IncidenciaProvincia	Provincia-Municipio
RiesgoImportado	RiesgoExterior-Aeropuerto

Con esto ya se han identificado tanto los nodos, como sus respectivas relaciones pero aún queda por especificar los atributos de cada uno de ellos. En la Tabla 2.27 se muestrán estos atributos.

Cuadro 2.27: Atributos

Aristas	Atributos
Aeropuerto	distritoId,distrito,aeropuertoId
Distrito	distritoId,distrito,poblacio,provincia
Municipio	municipioId,provinciaId,poblacion
DistritoSanitario	distritoSanitarioId, poblacion
Provincia	provinciaId, provincia, poblacion
Movilidad	fecha,flujo
Expansion	fecha, flujo
FormadoPor	-
IncidenciaMunicipio	fecha,casos,casos7,casos14
IncidenciaDistritoMunicipal	fecha,casos,casos7,casos14
IncidenciaDistritoSanitario	fecha,casos,casos7,casos14
IncidenciaProvincia	fecha,casos,casos7,casos14
RiesgoImportado	fecha,probCont

Con todo esto y se han definido todas las partes de la base de datos quedando el esquema mostrado en la Figura 2.4

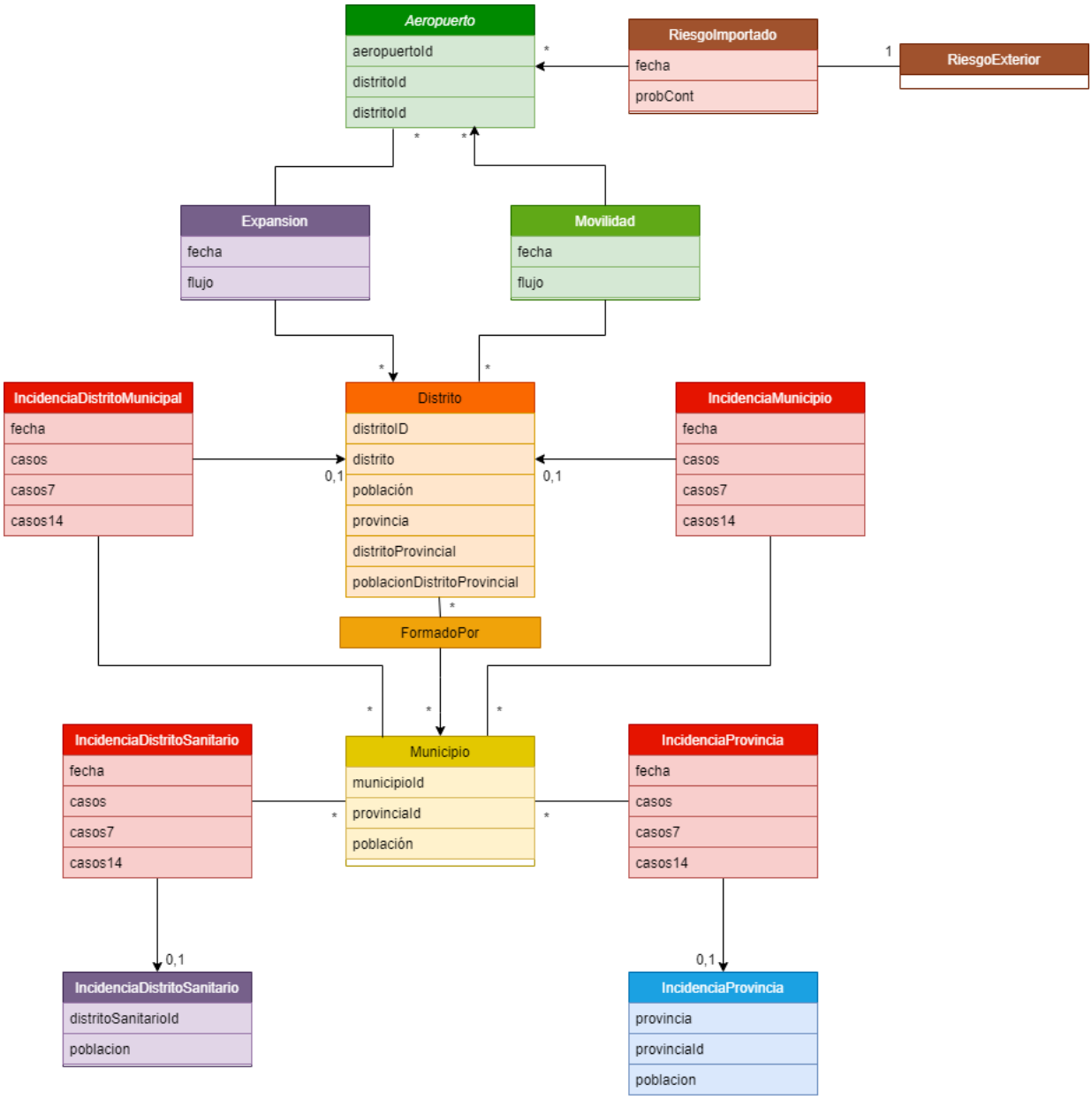

Figura 2.4: UML Base de Datos

Se tiene que especificar que en la base de datos existen unas clases más de las que vienen explicadas anteriormente. Esto es así porque algunas clases han sido duplicadas para el mejor tratamiento de ellas en los dos diferentes objetivos. Este es el caso de las relaciones de incidencias. Para toda relación IncidenciaX existe una IncidenciaXExp con iguales atributos y estructura. En un futuro se unirían ambos tipos de relaciones ya que representan lo mismo pero en periodos de tiempo distinto. Por esto no han sido descritas como una clase aparte.

2.2.4.3. Estadísticas

En esta sección se ha decidido que, una vez construida la base de datos, se muestren varías estadísticas que intenten evaluar el trabajo realizado.

Error Cometido Lo principal que se ha querido observar es el número de municipios que no se han podido enlazar correctamente a causa del identificador. Con identificador me refiero al nombre del municipio, su id. Como se comentó en la sección anterior se ha intentado formatear este nombre para que en todas las tablas tomen el mismo valor. Aún así hubo muchos casos en los que fue imposible formatear. Por tanto se van a mostrar cuantos municipios han sido formateados correctamente y cuales no. Primero se comenzará mostrando las provincias que tienen un nivel de detalle en la incidencia municipal en la Figura 2.5. Para luego mostrar las provincias con un nivel de distrito Sanitario en la Figura 2.6. Por último, no será necesario mostrar las provincias con un nivel provincial ya que se han conseguido introducir todos los municipios a este nivel.

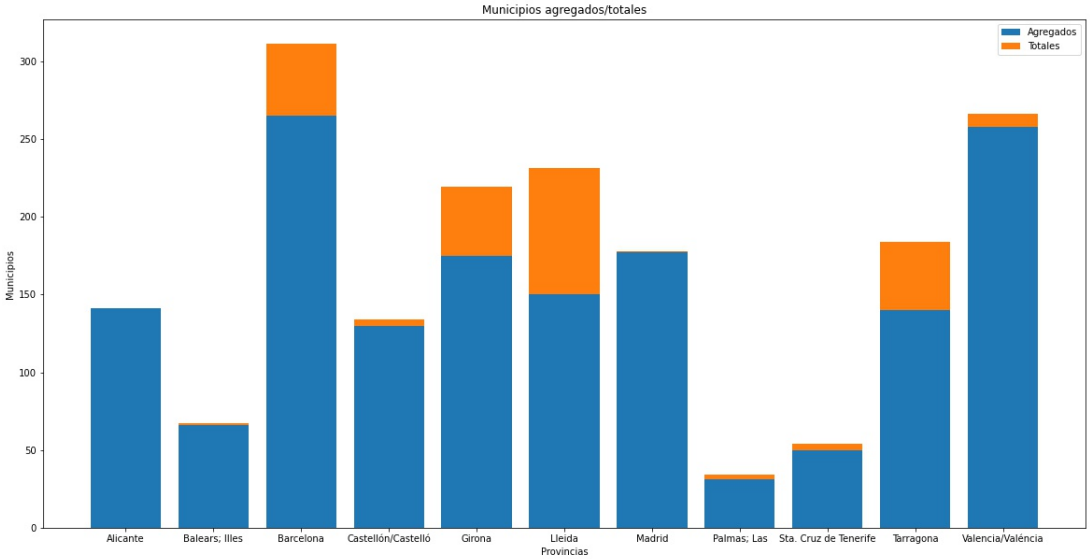

Figura 2.5: Municipios agregados Incidencia Municipal

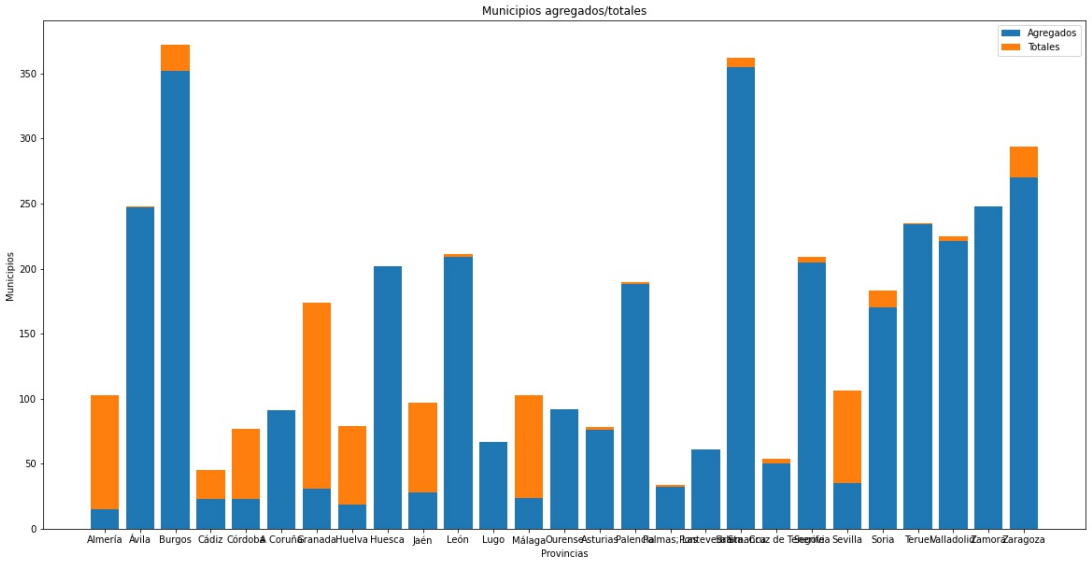

Figura 2.6: Municipios agregados Incidencia Distrito Sanitario

Principalmente se observa como la comunidad autónoma menos representada es la de Andalucía. El problema se identificó en los datos sobre los distritos sanitarios. Estos datos identificaban solo las principales localidades de cada distrito de Andalucía y por tanto no identificaban a un gran número de municipios. Más tarde se encontró más información sobre estos distritos pero estaba muy dispersa y de difícil uso. Por el poco tiempo de proyecto, se rechazó la mejora de está comunidad.

Falta de Información En este apartado se mostrarán los porcentajes de municipios que tienen cada nivel de precisión en la incidencia. Es decir, el porcentaje que tiene incidencia nivel municipal, nivel distrito sanitario y nivel provincial. Como es lógico, se intenta tener el máximo porcentaje para los niveles de detalle más alto. En las Figuras 2.7 y 2.8 se muestran el porcentaje de municipios para cada nivel de detalle. Como estos porcentajes, tal vez, no representen la población que está representada en cada nivel de detalle se mostrará la Figura 2.9 y 2.10 con el porcentaje de población con cada nivel de detalle. Se observan grandes diferencias, ya las comunidades que se han podido introducir en la base de datos son las mas pobladas de España, por tanto los municipios con nivel provincial suelen ser menos poblados. Esto es una buena señal ya que es de interés tener a la mayor parte de la población representada con el mayor detalle posible.

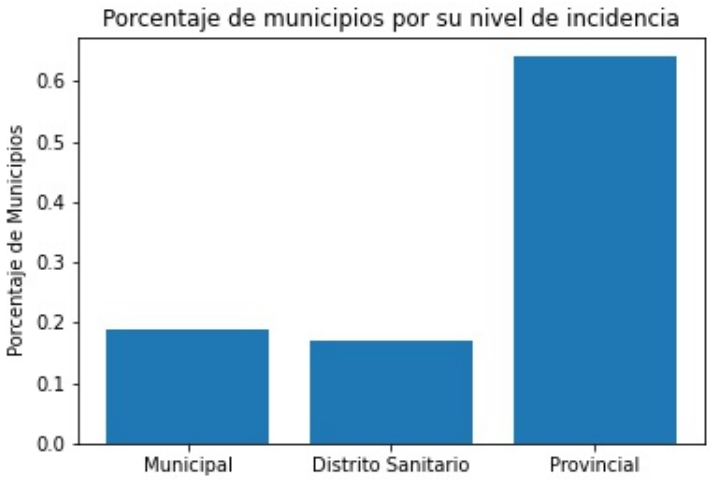

Figura 2.7: Porcentajes Municipios Incidencia

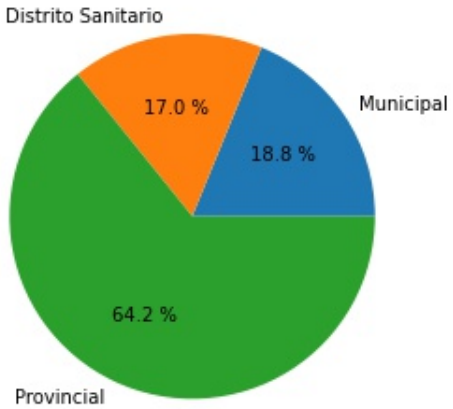

Figura 2.8: Porcentajes Municipios Incidencia

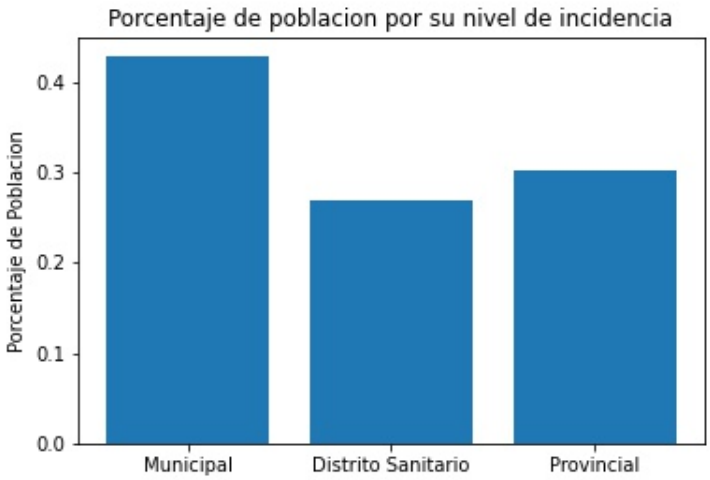

Figura 2.9: Porcentajes Población Incidencia

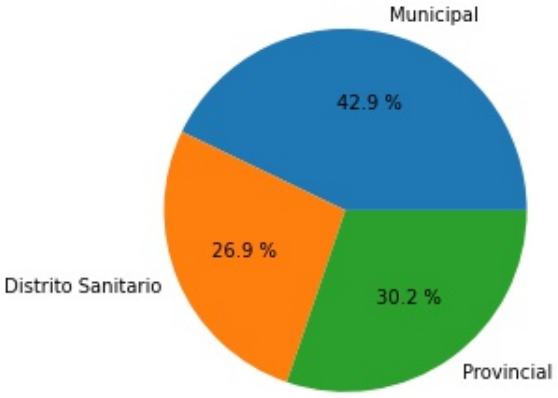

Figura 2.10: Porcentajes Población Incidencia

Movilidad En este apartado se pretenden mostrar como ha sido la movilidad hacia los aeropuertos desde el 15 de Junio de 2021 hasta el 31 de Julio de 2021. Estos datos se muestran en la Figura 2.11. Al ser verano, hay que tener en cuenta el turismo de está época y por eso se pueden observar que los aeropuertos de ciudades costeras tienen una mayor movilidad total.

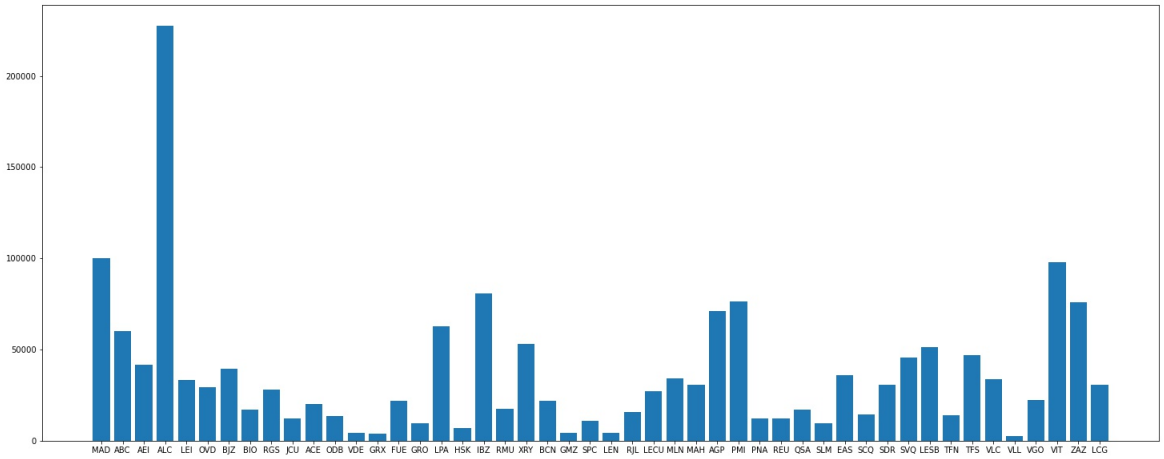

Figura 2.11: Movilidad hacia aeropuertos

Incidencia Con la Figura 2.12 de calor se pretende mostrar como ha sido la incidencia en las distintas provincias de España. Los colores rojos representan un menor nivel de incidencia mientras que los niveles azules muestran niveles más altos. Como era de esperar las grandes capitales muestran unos niveles muy altos como ha ocurrido a lo largo de la pandemia.

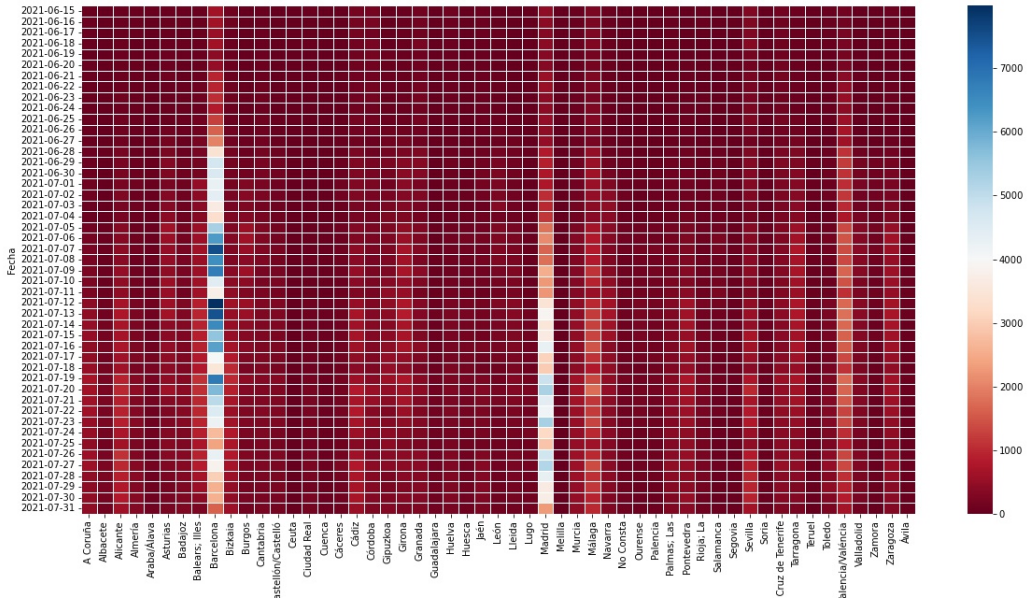

Figura 2.12: Incidencia Diaria

2.2.5. Modelado

Una vez se ha elaborado la base de datos, ya es posible extraer de ella todos los datos necesarios para realizar los modelos propuestos en el proyecto. Como se ha ido explicando a lo largo de este estudio se tenían dos principales objetivos. El primero de ellos es la mejora del cálculo del riesgo exportado en cada aeropuerto de España. Y el segundo de ellos realizar un modelo probabilístico que explique la relación entre el número de personas que llegan a un aeropuerto en vuelos y la incidencia en el lugar que se desplaza dentro de España.

2.2.5.1. Primera parte: Mejora detallada del cálculo del riesgo exportado en Aeropuertos españoles

En esta primera parte se explicará todo lo relacionado con el cálculo del riesgo exportado y su respectiva mejora. El riesgo exportado como definición es una estimación del número de personas contagiadas que llegan a un aeropuerto origen en un día determinado. En la mayoría de estudios esta estimación se realiza mediante la metodología SIR. Esta metodología divide a la población en tres grupos:

- Población Susceptible (S): Personas sin inmunidad ante el virus y que, por tanto, pueden ser infectadas.
- Población Infectada (I): Personas que en el momento dado están infectadas por el virus.
- Población Recuperada (R): Personas con inmunidad ante el virus y que, por tanto, no pueden ser contagiadas.

Con esta división de la población se necesitan ecuaciones que describan como evoluciona cada variable a lo largo del tiempo. Es decir, ecuaciones que muestren el número de recuperados, infectado y susceptibles a lo largo del tiempo. Por tanto, el modelo SIR se define con las siguientes ecuaciones diferenciales:

$$\frac{dS}{dt} = -\beta SI; \frac{dI}{dt} = +\beta SI - \gamma I; \frac{dR}{dt} = \gamma I \tag{2.1}$$

Siendo β la tasa de transmisión y γ la tasa de recuperación.

Cuando se utiliza esta metodología para el cálculo del riesgo exportado en la variable I, población Infectada, se calcula mediante una estimación de la incidencia en el lugar de origen del aeropuerto, siendo ciudad, estado o provincia.

La propuesta de este proyecto es modificar el cálculo de este riesgo con una estimación más detallada. Como se ha visto anteriormente en el estudio se tiene información sobre: aeropuertos de España, movilidad de personas con destino un aeropuerto y origen los distintos distritos de España y por último se tiene la información sobre la incidencia (a distintos niveles de detalles) de todos los municipios españoles. Por tanto, la idea sería calcular por cada aeropuerto de España y de forma diaria su respectiva movilidad. Esta movilidad dividirla por su origen y estimar por separado los posibles contagiados a partir de sus datos de movilidad.

De esta forma, se obtiene una estimación más precisa del número de contagiados que llegan a los aeropuertos españoles. Los aeropuertos que se verán más beneficiados

serán los aeropuertos con mayor tráfico aéreo como Barajas (Madrid), El Prat (Barcelona), Costa del Sol (Málaga) o Miguel Hernández (Alicante), donde se supone que llegan más personas procedentes de distintas provincias de España.

Para el cálculo de esta estimación se barajó varias tecnologías pero se decidió expandir las funcionalidades de la Base de Datos Neo4j mediante un *plugin*. Este *plugin* contendría una serie de funcionalidades que facilitarían el cálculo del riesgo importado por un aeropuerto. De esta forma se facilita la recreación y uso de estas estimaciones.

Para detallar el cálculo de este dato se mostrará la Figura 2.13. En esta figura se puede observar como en una fecha determinada un aeropuerto tiene tres enlaces de movilidad con tres distritos distintos. Cada distrito es de un tipo: un único municipio, un conjunto de municipios y un distrito municipal. Cada uno de los distritos obtendrá su incidencia a partir de el los municipios que lo forman.Hay que recordar que para este objetivo se tienen datos absolutos es decir número de casos detectados. En el caso de un único municipio o el de un distrito municipal solo se calculará el número de contagiados por habitante utilizando la población. En cambio, para el conjunto de municipios se tiene que calcular a partir de los municipios que lo forman, es decir cada incidencia por separado pasarla a casos por habitantes y luego calcular su porcentaje de población el distrito que forma. Por otro lado, en municipios vemos tres niveles de incidencia distintas: provincial, distrito sanitario y municipal. En cada caso respectivamente se elige el nivel de detalle más alto. Con esto ya se tendría el número de casos por habitante en cada municipio y su porcentaje de población en el caso de distritos de varios municipios. Con esto ya es posible calcular el número de infectados estimados en cada flujo de movilidad. Al sumarlos estas estimaciones se obtiene el número de infectados estimado que llegan al cada día.

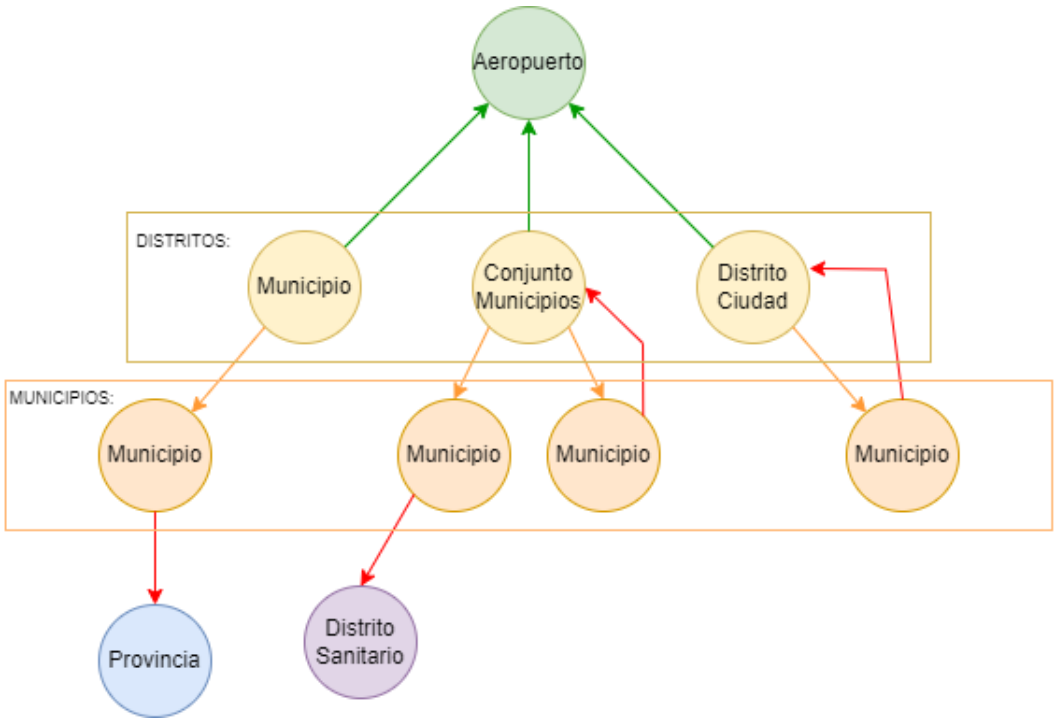

Figura 2.13: Caso de Uso Nº Infectados

Ya explicado el cálculo de la estimación, se mostrará la *query* necesaria para extraerlas de la base de datos:

```
MATCH (a:Aeropuerto)<-[mo:Movilidad]-(d:Distrito)
MATCH (d)-[f:FormadoPor]->(m:Municipio)
OPTIONAL  MATCH (m)-[iMun:IncidenciaMunicipio]->(d) WHERE mo.fecha=iMun.fecha
OPTIONAL MATCH (m)-[iDs:IncidenciaDistritoSanitario]->(ds:DistritoSanitario) WHERE mo.fecha=
    iDs.fecha
OPTIONAL MATCH (m)-[iProv:IncidenciaProvincia]->(p:Provincia) WHERE mo.fecha=iProv.fecha
OPTIONAL MATCH (m)-[iDsM:IncidenciaDistritoMunicipal]->(d) WHERE mo.fecha=iDsM.fecha
WITH a,d,mo,m,p,ds,iMun,iProv,iDs,iDsM
RETURN a.aeropuertoId,mo.fecha,sum(example.casosDiariosEstimados(m.poblacion,COALESCE(p.
    poblacion,0),COALESCE(ds.poblacion,0),d.poblacion,COALESCE(toInteger(d.
    poblacionDistritoMunicipal),0),COALESCE(iMun.casos14,-1.0),COALESCE(iProv.casos14,-1.0),
    COALESCE(iDs.casos14,-1.0),COALESCE(iDsM.casos14,-1.0),mo.flujo)) AS incidencia
```

2.2.5.2. Segunda parte: Modelo probabilístico del impacto de la movilidad con origen Aeropuertos en la incidencia COVID en Madrid

En esta sección se explicará el proceso de creación del modelo de este proyecto. El objetivo principal de este es elaborar un modelo que use predictores para aproximar la variable objetivo mostrando si existe una relación significativa entre ellas. La variable de interés en el proyecto es una variable de conteo, la incidencia en cada municipio de Madrid denominado como Inc. Al tener una variable de conteo como variable objetivo es posible realizar varios tipos de estimaciones como el modelo de Poisson, el modelo quasi-Poisson o el modelo Binomial Negativo.

Por tanto, las variables del modelo se definirán como:

- Variables predictoras: Incidencia del municipio en la fecha que se realiza una movilidad,llamada Inc, y el número de personas infectadas que realizan esta movilidad con destino el municipio,llamada Casos.

- Variable objetivo: Incidencia del municipio 14 días después de haberse realizado la movilidad.

Estas variables se obtienen a partir de la base de datos elaborada. Las variables Inc e Inc14 se obtienen fácilmente a partir de la base de datos. En cambio, para la variable Casos se tienen que agrupar toda la movilidad de un municipio de forma diaria, multiplicarla por el número de personas infectadas que llegan al aeropuerto y dividiendo por el total de personas que realizan la movilidad en ese día. Así se estima el número de personas infectadas que realizan un viaje determinado.

Antes de realizar el modelo se mostrarán los diagramas de dispersión de los datos utilizados.En una primera Figura 2.14 se mostrarán la variable objetivo en el eje y y la variable Casos en la X. En la segunda Figura 2.15 se mostrarán la variable objetivo en el eje y y la variable Inc en la X.

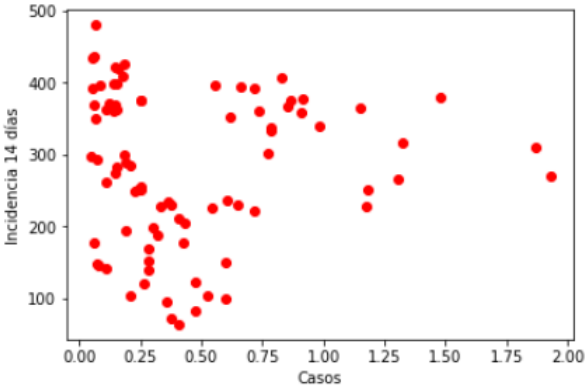

Figura 2.14: Diagrama de dispersión Incidencia 14 días- Casos

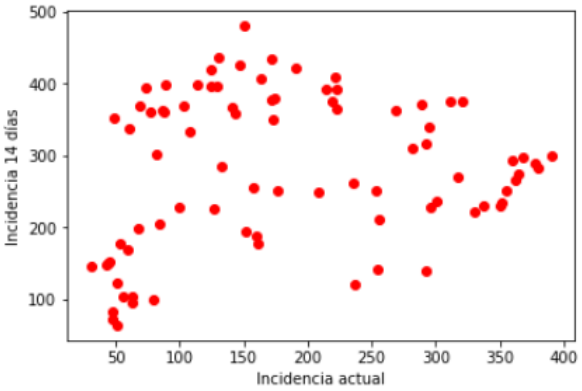

Figura 2.15: Diagrama de dispersión Incidencia 14 días- Incidencia Actual

Se observa que los datos no siguen ningún patrón claro, especialmente en el diagrama de Casos. Esto puede indicar poca relación entre las variables. Se observa también una dispersión excesiva en las variables por lo que es posible que se transformen estas variables.

Para seguir comprobando si las variables tienen relación entre ellas se calculará el coeficiente de correlación de *Pearson* para ver la relación entre las variables predictoras y la variable objetivo. Con la Tabla 2.28 se observa la correlación entre todas las variables del modelo. El coeficiente de correlación de *Pearson* nos indica la relación entre dos variables. Si el coeficiente es igual o está cerca a los valores 1 o -1 querrá decir que una variables es producto de una transformación lineal de la otra. Si el coeficiente es 0 o cercano a cero querrá decir que no existe o no tiene mucha relación lineal aunque si es posible que tenga relación de otro tipo.

Cuadro 2.28: Coeficiente de correlación de Pearson

	Inc14 (Objetivo)	Casos	Inc
Inc14 (Objetivo)	1	0.031457	0.171829
Casos	0.031457	1	0.180638
Inc	0.171829	0.180638	1

Con esta tabla se puede observar que en un principio es posible que la variable Casos, la movilidad, no tenga mucha relación con la variable objetivo, la incidencia.

Para analizar con aún más detalle la correlación entre las variables se dividirá el *dataset* por días analizados y se obtendrán de nuevo los coeficientes de correlación de *Pearson* para cada día y variable. En la Figura 2.16 se muestra está evolución en la correlación respecto a los casos. Por otro lado, en la Figura 2.17 se muestra la de la incidencia en el momento dado.

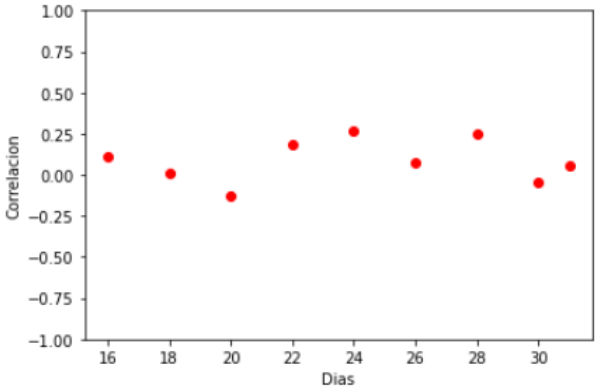

Figura 2.16: Evolución Correlación Inc14-Casos

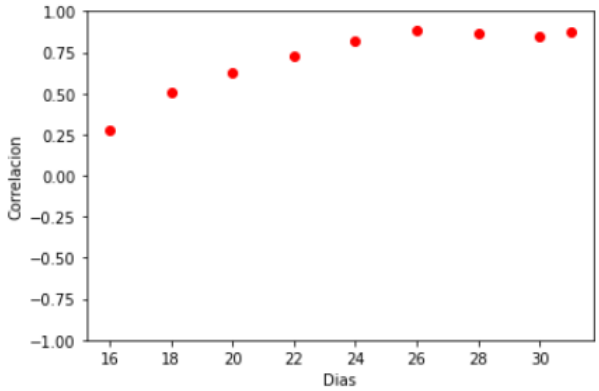

Figura 2.17: Evolución Correlación Inc14-Inc

Como se puede observar la Figura 2.16 la correlación va oscilando levemente pero se suele mantener en el mismo rango de valores.En cambio, 2.17 parece que en los primeros instantes tiene niveles de correlación más bajos pero se va estabilizando a niveles bastante altos de correlación, cosa que tiene mucho sentido.

En este punto se comenzó a realizar los distintos modelos. Lo primero a comentar es que si se hubiese tenido una cantidad más alta de datos se hubieran realizado varios modelos dividiéndolos por tramos temporales para así encontrar el modelo más ajustado. Al tener pocos datos se utilizarán todos los datos disponibles en cada modelo. Lo segundo ha comentar será la división del *dataset*. Para realizar los modelos se dividirá el *dataset* en un 80 % para el conjunto de entrenamiento y un 20 % para

el conjunto de pruebas. Con esto, se realizara el modelo con el conjunto de entrenamiento obteniendo así los p-valores y con el conjunto de pruebas se obtendrán el error cuadrático medio (RMSE) y el error absoluto medio porcentual (MAPE). Por último, se ha comentado la posible dispersión excesiva en las variables del problema, más clara en las variables de incidencia. Por tanto, se propondrán una serie de transformaciones para reducir esta dispersión. Otra transformación que se realiza es la unidad de medida de la variable Casos en nivel de incidencia por 100.000 habitantes para que todas las variables del problema tengan la misma unidad de medida.

En el conjunto de Figuras 2.18 y 2.19 se muestra como varía la dispersión de las variables ante las transformaciones propuestas.

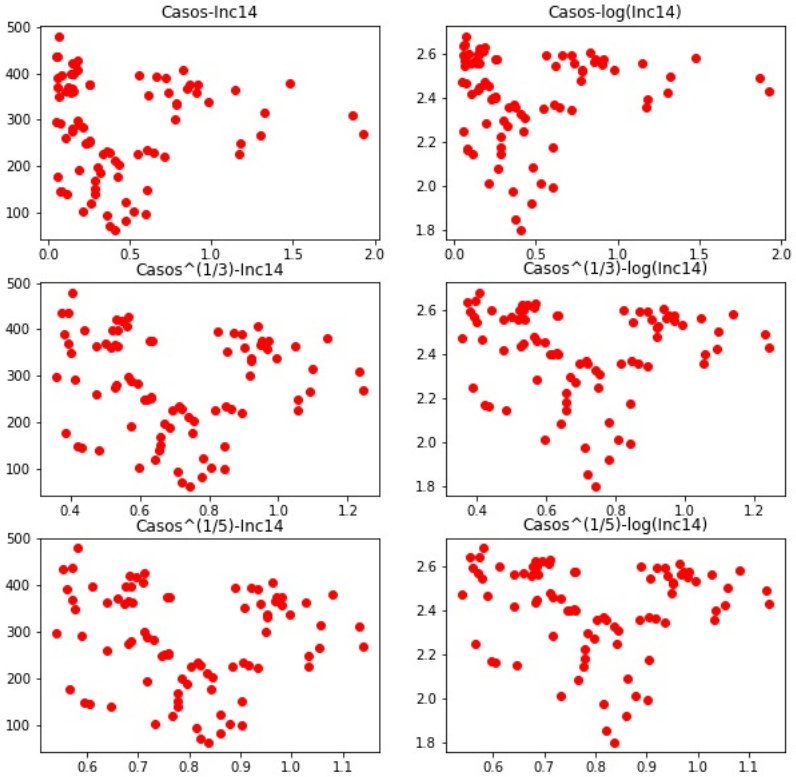

Figura 2.18: Dispersión variable caso transformada

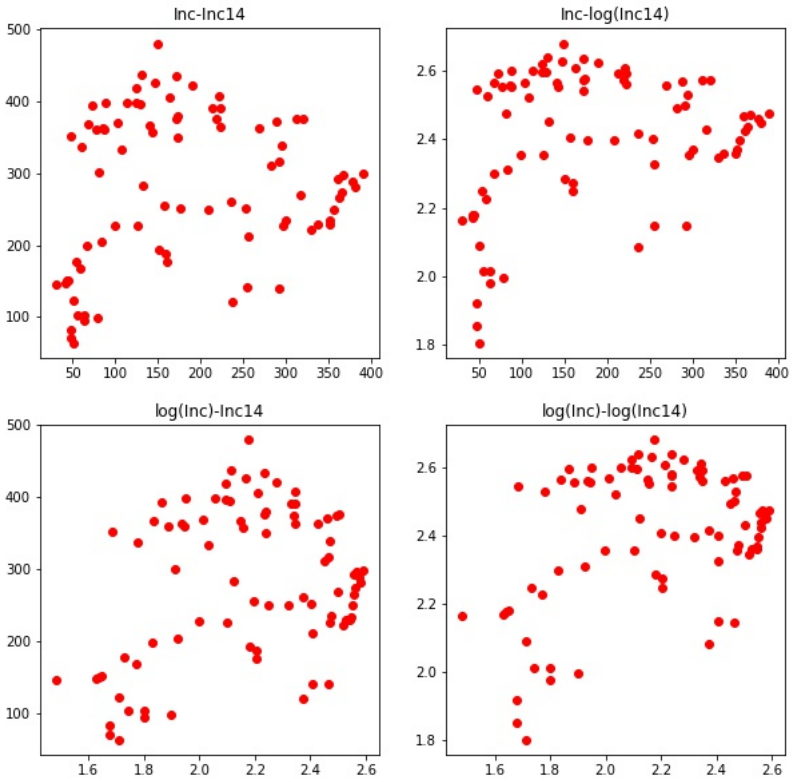

Figura 2.19: Dispersión variable caso transformada

Para decidir que transformaciones usar también se realizo la correlación con las variables transformadas obteniendo la tabla 2.29.

Cuadro 2.29: Correlación de Pearson variables transformadas

Variable	Correlación Inc14	Correlación Log(Inc14)
Inc	0.171829	0.272336
Inc14	1.000000	0.971402
log(Inc)	0.314052	0.406112
log(Inc14)	0.971402	1.000000
Casos	-0.218072	-0.249744
Casos^1/3	-0.218072	-0.249744
Casos^1/5	-0.218072	-0.249744

Al ver la correlación junto con los diagramas de dispersión se puede observar como la correlación mejor con la transformación de logaritmo en las variables de Incidencia, por lo que los modelos tendrán esta transformación. Por otro lado, las transformaciones en la variable Casos,tanto en unidad de medida como las raíces, son un poco más irregulares pero cambia el signo de la correlación por lo que se probarán varias de las transformaciones en los modelos.

Capítulo 3

Resultados

En esta sección se explicarán los resultados obtenidos para los dos objetivos propuestos. Como en el resto del proyecto se dividirán los resultados en dos, uno por cada objetivo.

3.1. Estimación Riesgo Exportado

Los primeros resultados que se mostrarán será la estimación del número de infectados estimados por aeropuerto. Estos resultados se muestran en la Figura 3.1. En esta gráfica se pueden observar como evoluciona esta estimación para cada uno de los aeropuertos españoles durante el mes de Julio. En ella se pueden observar como los aeropuertos de Madrid y Barcelona han pasado a un segundo puesto, especialmente Barcelona. En cambio los aeropuertos de zonas costeras como Alicante, Mallorca, Málaga, Valencia o Ibiza se encuentran liderando los infectados estimados. Estos resultados pueden estar relacionados a la estación temporal escogida. Los datos recogidos son en pleno verano y en un momento en el que el turismo exterior estaba reducido. Por tanto, los principales destinos turísticos veraniegos como las islas y zonas costeras tienen un mayor nivel de contagios estimados. En la segunda Tabla 3.1 se reduce el número de aeropuertos para mostrar los principales aeropuertos de España o los de más alto nivel de infectados.

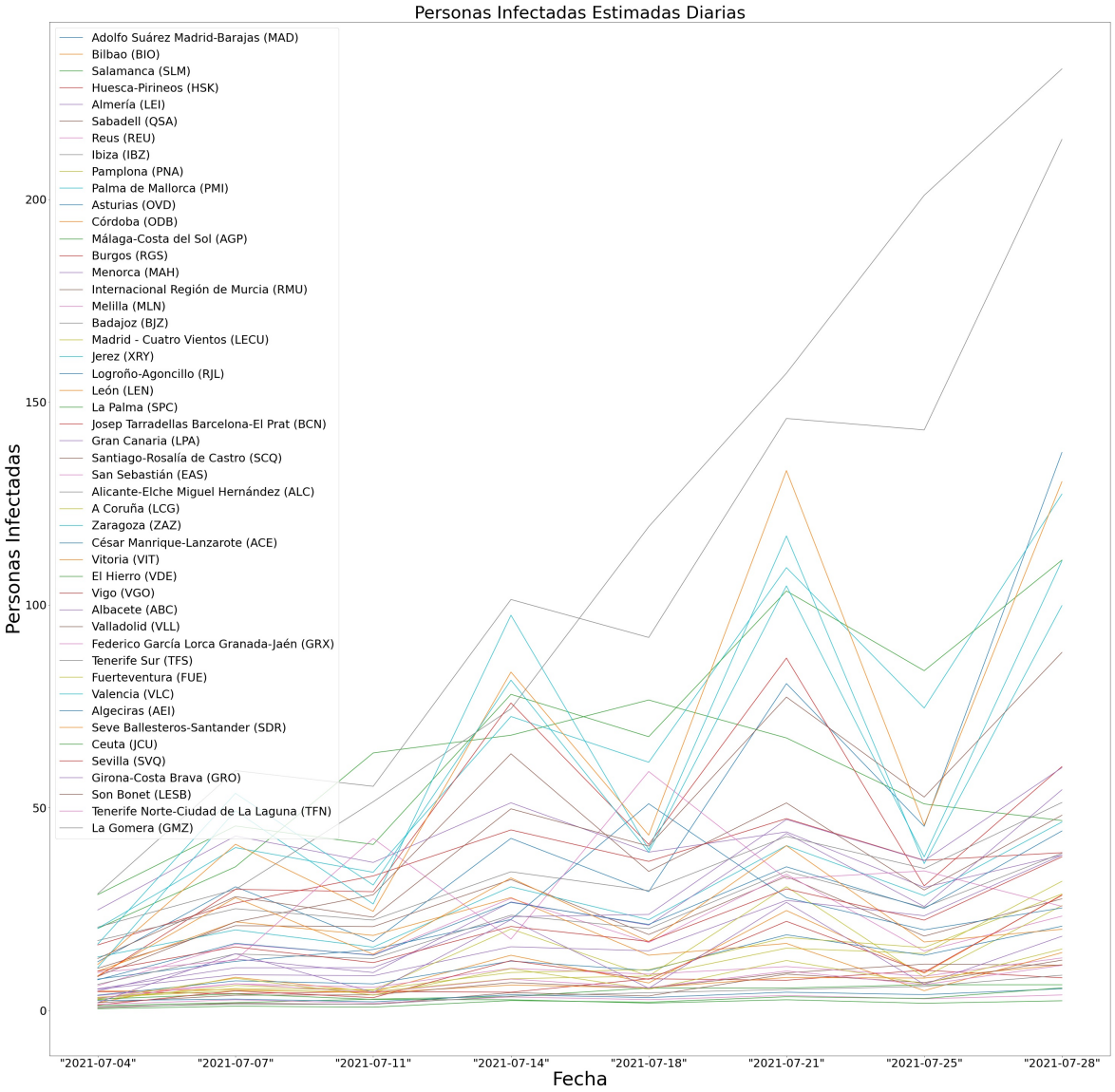

Figura 3.1: Estimación infectados diarios por aeropuerto

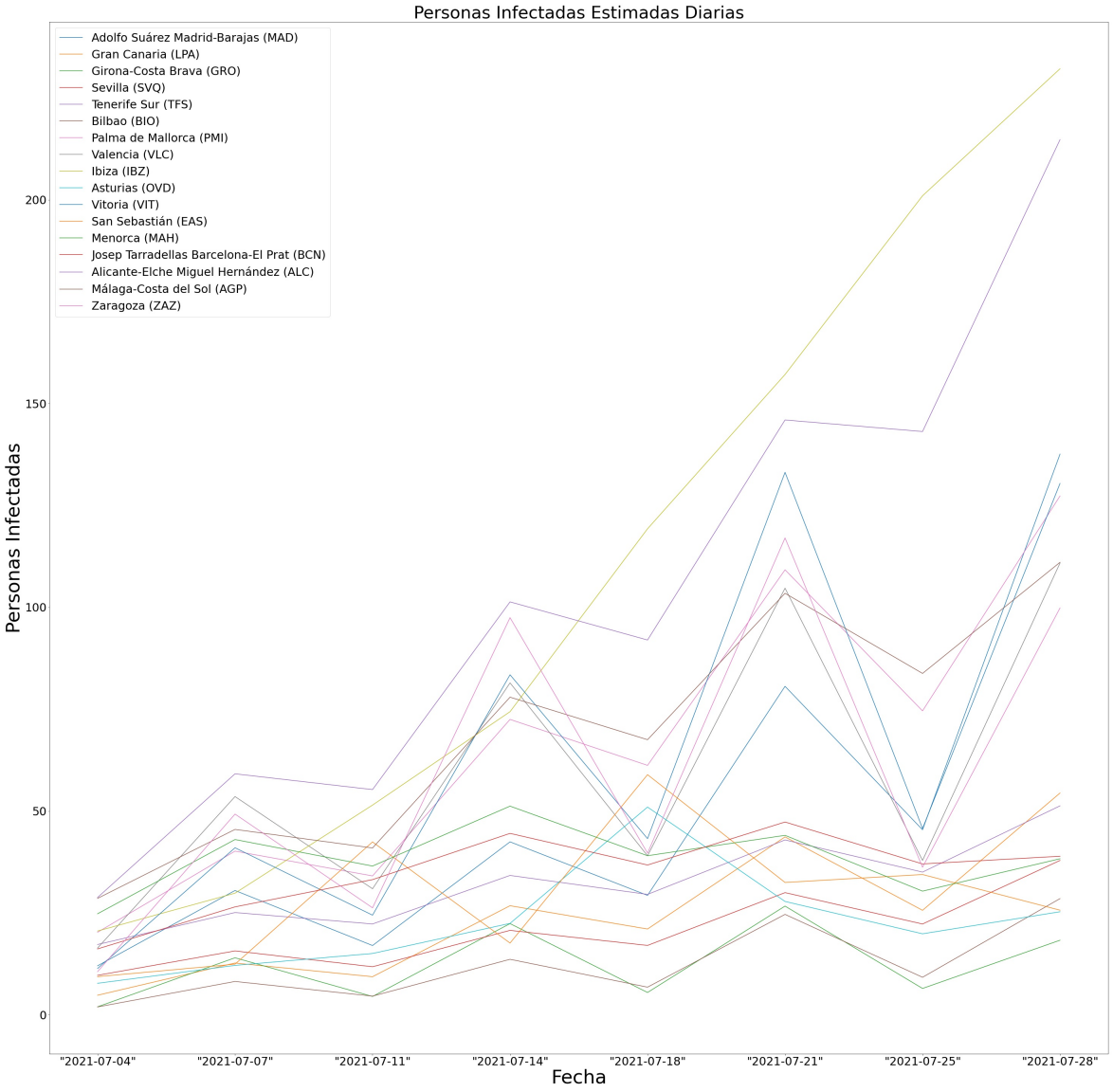

Figura 3.2: Estimación infectados diarios en los principales aeropuertos

En un primer momento se pretendía evaluar esta metodología más especifica de estimación de infectados. Esta idea se tuvo que rechazar por su dificultad. Esta dificultad se encontraba en que las personas que llegan a un aeropuerto se dispersan a lo largo del mundo. Por tanto comprobar si esta estimación está más ajustada a la realidad era complejo. Una posibilidad hubiese sido desarrollar dos modelos probabilísticos

que analizarán la movilidad desde España hasta el resto de países. En uno de los modelos se usaría la metodología estándar de estimación de infectados, mientras que en el segundo se utilizaría la metodología propuesta. Si los resultados del modelo con la metodología propuesta se ajustan más que en la estándar podría indicarnos que la metodología propuesta se ajusta más a la realidad.

3.2. Resultados modelo probabilístico

Los segundos resultados que se mostrarán son los de los modelos probabilísticos elaborados. Como se explicó en la sección anterior se realizaron una serie de modelos distintos para el mismo problema. A continuación se mostrarán las estadísticas obtenidas de cada uno de los modelos. Primero en la Tabla 3.1 se muestran los p-valores de todos los modelos propuestos. Por tanto, en esta tabla se diferencian por un lado las transformaciones realizadas y por otro se muestran los tres tipos de modelos realizados (Poisson, Quasipoisson y Binomial negativo). Lo primero que se puede observar en esta tabla es que el modelo de Quassipoisson suele ser el más ajustado de los tres. Esto es lógico al tener media y varianza distintas y por tanto ajustará mejor que Poisson. Por otro lado, vemos como es común en la gran relación entre la variable objetivo e Inc, esta relación se reduce al transformar únicamente una de las dos con logaritmo. También se observa como los modelos en los que las 3 variables se miden con la misma unidad de medida se ajustan mejor. Por tanto, si se tiene que elegir algún modelo por su *p-valor* se elegirían los tres modelos en los que se aplica logaritmo a Inc14 e Inc. De entre esos tres parece que con la variable Casos sin transformar se obtienen datos un poco mas ajustados.

Cuadro 3.1: Modelos y sus p-valores

Transformación	Poisson	Quasi	Binomial Negativo
log(Inc14) - log(Inc) + Casos	Int:0.287 Casos:0.836 Inc:0.652	Int:<2e-16 *** Casos:0.033072 * Inc:0.000223 ***	Int:0.216 Casos:0.795 Inc:0.643
log(Inc14) - Inc + Casos	Int:3.57e-06*** Casos:0.912 Inc:0.835	Int:<2e-16*** Casos:0.0273* Inc:0.0155*	Int:1.97e-08*** Casos:0.778 Inc:0.756
log(Inc14) - log(Inc) + (Casos)^1/3	Int: 0.285 Casos: 0.846 Inc: 0.684	Int:<2e-16*** Casos:0.033072* Inc:0.000223***	Int:0.216 Casos:0.795 Inc:0.643
log(Inc14) - log(Inc) + (Casos)^1/5	Int: 0.273 Casos: 0.660 Inc: 0.738	Int: <2e-16*** Casos:0.033072* Inc:0.000223***	Int: 0.216 Casos:0.795 Inc:0.643

Por otro lado, es importante ver las estadísticas del error con el conjunto de prueba. Estas estadísticas pueden ayudar a elegir uno de los tres modelos propuestos o incluso rechazarlos. Por tanto, las estadísticas calculadas son el error cuadrático medio (RMSE) y el error absoluto medio porcentual (MAPE). En la Tabla 3.2 se muestran estos errores divididos por transformación y tipo de modelo. Esta tabla ayudaría a decidirse entre uno de los modelos elegidos. Por tanto, si se tiene que elegir uno de los dos modelos por su error cometido se elegiría el modelo con Casos transformado con raíz cinco. Al tener valores de error muy similares se decide escoger el modelo con los p-valores mas bajos, es decir, el modelo con transformaciones logarítmicas en las variables Inc e Inc14 y la variable Casos sin transformar.

Cuadro 3.2: RMSE y MAPE de modelos

Transformación	Poisson	Quasi	Binomial Negativo
log(Inc14) - log(Inc) + (Casos)	RMSE:0.1869598 MAPE:6.56284	RMSE: 0.1804826 MAPE: 6.695522	RMSE:0.1941348 MAPE:7.034969
log(Inc14) - Inc + Casos	RMSE:0.1625685 MAPE:6.023837	RMSE:0.2175678 MAPE:8.27408	RMSE:0.2307951 MAPE:8.442235
log(Inc14) - log(Inc) + (Casos)^1/3	RMSE:0.2198517 MAPE:8.387005	RMSE:0.1936731 MAPE:7.186118	RMSE:0.1797804 MAPE:6.743414
log(Inc14) - log(Inc) + (Casos)^1/5	RMSE:0.2146994 MAPE:8.249373	RMSE:0.1388859 MAPE:4.872107	RMSE:0.1959825 MAPE:7.187679

Por tanto, con este modelo elegido se puede confirmar la existencia de una relación de la movilidad desde Madrid en los municipios. Esta relación es de muy poca influencia pero confirma la existencia de esta relación. En cambio se puede observar como la variable Inc es decir la incidencia en el momento de la movilidad, cosa evidente.

Para terminar, este modelo se realizó con datos desde el 15 de Marzo de 2020 hasta el 1 de Abril de ese mismo año, es decir, en cuarentena. Por tanto, esta poca relación entre variables podría estar causada por varios motivos. Uno de ellos es las restricciones de movilidad, es decir, el número de vuelos y personas que llegan al país es muy reducida. El otro motivo podría se la diferencia de incidencia entre España y el resto de países Europeos. Es decir, en este momento España lideraba junto a Reino Unido el número de contagios y su incidencia, entonces si una persona llega a España desde Europa su probabilidad de estar infectado es menor que la de un propio ciudadano español. Estas dos causas podrían justificar esta poca relación entre la movilidad desde aeropuertos y el número de infectados.

Capítulo 4

Conclusiones

En este proyecto, por tanto, se han cumplido varios de los objetivos propuestos. En una primera etapa se realizó una búsqueda extensa de fuentes de información sobre la movilidad internacional a España, encontrando numerosas fuentes de datos. Lamentablemente ninguna de estas cumplieron los prerrequisitos necesarios para continuar con esta primera etapa, analizar la influencia de la movilidad internacional por Carretera, Ferrovial y Marítima en la incidencia COVID-19 española. Aún así se describieron las fuentes de información más interesantes y mejor estructuradas para poder ser utilizadas en proyectos futuros o similares.

Tras esto, se comenzó la segunda temática del proyecto, análisis de la movilidad con origen y destino aeropuertos. Este análisis se dividiría en dos estudios distintos. El primer estudio sería la mejora de la metodología SIR para la estimación del riesgo exportado. Y la segunda sería estudiar el impacto de la movilidad con origen el aeropuerto de Barajas, Madrid, en la incidencia COVID-19 municipal de la comunidad de Madrid.

Para ambos objetivos es necesario tener una extensa y detallada base de datos donde obtener la información necesaria para elaborar estos estudios. Esta base de datos debería tener la información de la movilidad con origen o destino aeropuertos españoles, por otro lado, debería tener la información lo más detallada posible de la incidencia COVID-19 en el municipio origen o destino y, por último, es necesario tener el riesgo importado de los aeropuertos españoles, concretamente Barajas. Para elaborar, por tanto, esta base de datos fue necesario encontrar fuentes de información fiables. Todas las fuentes de información utilizadas en este proyecto provienen de fuentes públicas del gobierno de España o comunidades autónomas, destacando la fuente de movilidad TAL, la fuente del riesgo importado proporcionado por CRIDA y los datos de incidencia aportados por las distintas organizaciones autonómicas. Estos datos obtenidos fueron preprocesados para obtener una serie de tablas con toda y cada una de la información necesaria para elaborar la base de datos. Especialmente, la incidencia de los municipios de España fue la que más tiempo de preprocesamiento necesitó. Esto se debió a que cada comunidadad autonóma aporta sus respectivos datos con estructuras distintas, periodos temporales distintos, niveles de detalle distinto etc. Por tanto, con este procesamiento se consiguió unificar y estructurar todos los datos facilitando la creación, uso y extensión de la base de datos.

La base de datos creada es de tipo orientada a grafos, es decir, la información que se

ha obtenido se ve representada mediante nodos y aristas. Para elaborarla se utilizó la tecnología Neo4j.

Una vez realizada la base de datos fue momento de realizar los dos estudios comentados. El primero fue la mejora de la metodología SIR para la estimación del riesgo exportado de los aeropuertos españoles. Esta mejora se centra en aumentar el nivel de detalle en la estimación de número de infectados que llegan al aeropuerto origen. En la metodología clásica se estima este valor mediante un único dato de incidencia en el territorio donde se sitúa el aeropuerto. Por tanto, la metodología propuesta consiste en estimar este valor dividiendo las personas que llegan a cada aeropuerto por su origen municipal. Con esta división se calcularía por separado el riesgo de infección con su respectiva incidencia municipal. De esta forma se obtendría una posible estimación más detallada de los infectados que llegan a cada aeropuerto.Se realizó una estimación del riesgo exportado en los aeropuertos españoles en el mes de Julio de 2020. Los resultados obtenidos reflejaban un mayor riesgo en los principales aeropuertos costeros españoles como Alicante, Málaga, o varios aeropuertos de las islas. Esto puede ser debido al momento temporal elegido, verano.

Un objetivo claro de este estudio es la validación de esta metodología. El problema de esta validación es que es necesaria la información de tráfico aéreo de cada aeropuerto español. Ante esta dificultad se dejó esta validación como posible trabajo futuro. Se propone una posible validación. Esta consistiría en realizar dos modelos probabilísticos distintos en los que se analizaría el impacto de la movilidad con origen España en los distintos destinos. Un primer modelo utilizaría la metodología de estimación de riesgo exportado clásica. El otro modelo utilizaría la metodología propuesta. De esta forma es posible comparar ambos resultados y validar la mejora o no en esta estimación de infectados.

Con este primer estudio finalizado, se comenzó a realizar el segundo. Este segundo estudio consistía en analizar el impacto de la movilidad con origen aeropuerto, concretamente Madrid, en la incidencia municipal de la comunidad autónoma. El periodo temporal utilizado es de los primeros mese de pandemia, concretamente, desde el 15 de Marzo al 1 de Abril de 2021. Recordar que en este momento la restricción del confinamiento ya estaba en vigor. Para realizar este modelado se ideó utilizar varios tipos de modelo de variables de conteo distintos, como son Poisson, Quasipoisson y Binomial Negativo. Antes de realizar estos modelos fue necesario probar varias transformaciones ya que las variables de esta tenia una gran dispersión. Los resultado obtenidos demostraban la existencia de una relación entre ambas variables pero de poca fuerza. Esto puede ser debido a la diferencia de niveles de incidencia entre España y el resto de Europa y las medidas de restricciones que estaban presente en este momento. Un posible trabajo futuro podría ser estudiar este mismo problema en momentos temporales distintos como podía ser antes de las restricciones del país, es decir, primeros momentos de pandemia o estudiar este mismo problema cuando España no era foco de infección en Europa. Con estos trabajos se podría estudiar como varía el nivel del impacto en la incidencia a lo largo de la pandemia. Demostrando, o no, la hipótesis de la mayor relación en las primeras semanas de pandemia en España y la menor relación cuanto más avanzaba.

Bibliografía

[Tatem *et al.*, 2006] Tatem, A. J., Rogers, D. J., Hay, S. I. (2006). *Global transport networks and infectious disease spread. Advances in parasitology*, 62, 293-343.

[Colizza *et al.*, 2006] Colizza, V., Barrat, A., Barthélemy, M., Vespignani, A. (2006). *The role of the airline transportation network in the prediction and predictability of global epidemics*. Proceedings of the National Academy of Sciences, 103(7), 2015-2020.

[MacFadden *et al.*, 2015] MacFadden, D. R., Bogoch, I. I., Brownstein, J. S., Daneman, N., Fisman, D., German, M., Khan, K. (2015). *A passage from India: association between air traffic and reported cases of New Delhi Metallo-beta-lactamase 1 from 2007 to 2012.* Travel medicine and infectious disease, 13(4), 295-299.

[Grais*et al.*, 2003] Grais, R. F., Hugh Ellis, J., Glass, G. E. (2003). *Assessing the impact of airline travel on the geographic spread of pandemic influenza*. European journal of epidemiology, 18(11), 1065-1072.

[Bogoch*et al.*, 2014] Bogoch, I. I., Creatore, M. I., Cetron, M. S., Brownstein, J. S., Pesik, N., Miniota, J., ... Khan, K. (2015). *Assessment of the potential for international dissemination of Ebola virus via commercial air travel during the 2014 west African outbreak.* The Lancet, 385(9962), 29-35.

[Bogoch*et al.*, 2016] Bogoch, I. I., Brady, O. J., Kraemer, M. U., German, M., Creatore, M. I., Kulkarni, M. A., ... Khan, K. (2016). *Anticipating the international spread of Zika virus from Brazil.* The Lancet, 387(10016), 335-336.

[Tian *et al.*, 2017] Tian, H., Sun, Z., Faria, N. R., Yang, J., Cazelles, B., Huang, S., ... Xu, B. (2017). *Increasing airline travel may facilitate co-circulation of multiple dengue virus serotypes in Asia.* PLoS neglected tropical diseases, 11(8), e0005694.

[Nakamura *et al.*, 2020] Nakamura, H., Managi, S. (2020). Airport risk of importation and exportation of the COVID-19 pandemic. Transport policy, 96, 40-47.

[Pullano *et al.*, 2020] Pullano, G., Pinotti, F., Valdano, E., Boëlle, P. Y., Poletto, C., Colizza, V. (2020). Novel coronavirus (2019-nCoV) early-stage importation risk to Europe, January 2020. Eurosurveillance, 25(4), 2000057.

[Sokadjo *et al.*, 2020] Sokadjo, Y. M., Atchadé, M. N. (2020). The influence of passenger air traffic on the spread of COVID-19 in the world. Transportation Research Interdisciplinary Perspectives, 8, 100213.

[Lau *et al.*, 2020] Lau, H., Khosrawipour, V., Kocbach, P., Mikolajczyk, A., Ichii, H., Zacharski, M., ... Khosrawipour, T. (2020). The association between internatio-

nal and domestic air traffic and the coronavirus (COVID-19) outbreak. Journal of Microbiology, Immunology and Infection, 53(3), 467-472.

[Kraemer *et al.*, 2020] Kraemer, M. U., Yang, C. H., Gutierrez, B., Wu, C. H., Klein, B., Pigott, D. M., ... Scarpino, S. V. (2020). The effect of human mobility and control measures on the COVID-19 epidemic in China. Science, 368(6490), 493-497.

[Orea *et al.*, 2020] Orea, L., Álvarez, I. C. (2020). How effective has the Spanish lockdown been to battle COVID-19? A spatial analysis of the coronavirus propagation across provinces. Documento de trabajo, 3(2020), 1-27.

[Mazzoli *et al.*, 2020] Mazzoli, M., Mateo, D., Hernando, A., Meloni, S., Ramasco, J. J. (2020). Effects of mobility and multi-seeding on the propagation of the COVID-19 in Spain. MedRxiv.

[García-Abadillo, 2021] García-Abadillo Velasco, J. (2021). Estudio del impacto de la movilidad interprovincial en la propagación del virus en España durante el período de confinamiento (Doctoral dissertation, ETSI Informatica).

[Gendronneau *et al.*, 2021] Gendronneau, C., Wiśniowski, A., Yildiz, D., Zagheni, E., Fiorio, L., Hsiao, Y., ... Hoorens, S. (2019). Measuring labour mobility and migration using big data. European Commission, Brussels.

[Datos Andalucía] *Datos Coronavirus Andalucía*

[Datos Aragón] *Datos Coronavirus Aragón*

[Datos Asturias] *Datos Coronavirus Asturias*

[Datos Baleares] *Datos Coronavirus Islas Baleares*

[Datos C.Valenciana] *Datos Coronavirus Comunidad Valenciana*

[Datos Canarias] *Datos Coronavirus Canarias*

[Datos Cantabria] *Datos Coronavirus Cantabria*

[Datos Castilla y León] *Datos Coronavirus Castilla y León*

[Datos Castilla y la Mancha] *Datos Coronavirus Castilla la Mancha*

[Datos Cataluña] *Datos Coronavirus Cataluña*

[Datos Extremadura] *Datos Coronavirus Extremadura*

[Datos Galicia] *Datos Coronavirus Galicia*

[Datos La Rioja] *Datos Coronavirus La Rioja*

[Datos Madrid] *Datos Coronavirus Madrid*

[Datos Murcia] *Datos Coronavirus Murcia*

[Datos Navarra] *Datos Coronavirus Navarra*

[Datos País Vasco] *Datos Coronavirus País Vasco*

[Frontur] *Estadística de movimientos turísticos en frontera. Frontur.*

[OpenData Movilidad] *Open Data Movilidad*

[INE Movilidad] *Estudios de movilidad de la población a partir de la telefonía móvil 2020-2021*

[Lifesight Mobility] *Lifesight Mobility/ Raw Location Data | Global mobile location data*

[Lifesight Human Mobility] *Lifesight Human Mobility data by time of day, aggregated to Geohash/Quadkey/hex*

[Location-based data] *Location-based data ,DATACORE*

[PySocialWatcher] *PySocialWatcher.*

[Base de Datos Creada] Base de Datos Creada