



Universidad Politécnica
de Madrid

**Escuela Técnica Superior de
Ingenieros Informáticos**



Máster en Data Science

Trabajo Fin de Máster

**Modelos de Regresión de Poisson
Aplicados a la Propagación del COVID-19
a través del Tráfico Aéreo**

Autor(a): David Cabornero Pascual

Madrid, Julio, 2022

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

Trabajo Fin de Máster
Máster en Data Science

Título: Modelos de Regresión de Poisson Aplicados a la Propagación del COVID-19 a través del Tráfico Aéreo
Julio, 2022

Autor: David Cabornero Pascual
Tutores: Alfonso Mateos Caballero y Arminda Moreno Díaz
Grupo de Análisis de Decisiones y Estadística
Departamento de Inteligencia Artificial
ETSI Informáticos
Universidad Politécnica de Madrid

Agradecimientos

Con este máster completo mi formación y mi etapa como estudiante. No puedo acabar sin agradecerle todo el esfuerzo que han puesto en mí mis padres, mi abuela. De la misma forma, quiero dar gracias por todo al apoyo a Laura y Juanmi. Sin vosotros no hubiera podido llegar hasta aquí.

Por supuesto, también tengo que agradecer la dedicación que han puesto en mí Ar-minda y Alfonso, sin los cuales este TFM no tendría la calidad con la que lo entrego.

Para finalizar, también se agradece el apoyo proporcionado por CRIDA (Centro de Referencia I+D+i ATM) y los proyectos MadridDataSpace4Pandemics-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).

Resumen

La COVID-19 es una infección vírica detectada en diciembre de 2019 que se convirtió rápidamente en una pandemia mundial. Desde entonces, la necesidad de crear modelos de predicción para predecir epidemias ha ido creciendo, no solo por la necesidad en si sino por la cantidad de datos generados debido a la pandemia.

Dentro de todas las posibilidades que existen para predecir contagios, los modelos de regresión de Poisson son una buena opción, ya que los contagios, al igual que la distribución de Poisson, se pueden definir como un conteo de casos. De esta forma, se evita también que los algoritmos predigan un número negativo de contagios. Por ello, en este trabajo se han elegido estos algoritmos para predecir el número de contagiados en distintos países.

Utilizamos una base de datos con todos los vuelos en 2020, donde se incluyen datos relevantes como países de origen y destino, pasajeros por vuelo o contagios por región. Con todos estos datos, se puede estimar el número de casos entrantes a cada país, que se ha denominado *casos importados*, que se considera uno de los factores más importantes en fases tempranas de la pandemia, donde los contagios venían principalmente de países externos.

De la misma manera, el *riesgo importado* es una variable similar, pero en este caso se utilizan modelos SIR para calcular con más precisión cómo influye la duración de los vuelos en el número de infectados entrantes. La predicción del número de infectados por país mediante modelos de Poisson utilizando estas variables ha sido el objetivo principal de este trabajo.

Para utilizar datos que sean lo más fiables posibles, nuestro estudio se restringe únicamente a Europa. Además, solo se ha trabajado con las fechas comprendidas entre el 15 de febrero y el 15 de marzo de 2020, ya que antes de estas fechas no hay casi contagios detectados en Europa y después los estados empezaron a aplicar severas restricciones al tráfico aéreo, por lo que los contagios dentro del país pasaron a tener mucha más relevancia que los exteriores a partir de esta fecha.

Dado que la COVID-19 tenía un tiempo de incubación promedio de una semana, se van a predecir los positivos de los países tras una semana, sabiendo el riesgo importado y los casos confirmados actuales. De esta forma, para cada día se entrenará un nuevo modelo. El modelo principal es el modelo de Poisson, pero dado el fenómeno de sobredispersión que se experimenta en él también se han utilizado dos variantes: el modelo quasi-Poisson y el de la binomial negativa.

Tras entrenar los modelos, se ha analizado el cumplimiento de las hipótesis básicas, la relevancia de las variables independientes, la sobredispersión, los coeficientes de

los estimadores, los residuos y la bondad de ajuste. Tras todo esto, se ha concluido que teóricamente este modelo es adecuado para el problema que nos atañe (se cumplen las hipótesis), todas las variables son relevantes, la sobredispersión aumenta con el tiempo (por lo que las variantes del modelo de Poisson toman fuerza frente al original), la variable independiente más importante comienza siendo los contagios importados (contagios de fuera) y acaba siendo los contagios confirmados (contagios del país) y que la varianza de los residuos no aumenta con el tiempo.

Una vez que se trata de predecir con el modelo del día anterior el día siguiente, se ve que todos los modelos tienen errores parecidos mostrando resultados muy buenos que mejoran porcentualmente según avanza el tiempo. Con todo esto, se concluye que los modelos de regresión de Poisson con los contagios importados dan lugar a buenos modelos de regresión para predecir los contagios de un país.

Abstract

COVID-19 is a viral infection detected in December 2019 that quickly escalated to a global pandemic. Since then, the necessity to create predictive models to predict epidemics has increased dramatically, due not only to the need itself but to the amount of data generated during the pandemic.

Among the possibilities that exist to predict contagions, Poisson regression models are a good option, since the contagions and the Poisson distribution can be defined as a count of cases. In this way, predictions of a negative number of cases are avoided. Therefore, these algorithms have been chosen in this work in order to predict the number of COVID-19 cases in different countries.

We use a database with all flights in 2020, including relevant data such as country of origin and destination, passengers per flight, and contagions per region. With all these data, we can estimate the number of incoming cases to each country due to air traffic. This number has been referred to as the number of imported cases, which is considered one of the most relevant factors in the early stages of the pandemic, where infections came mainly from external countries.

In the same way, the imported risk is a similar variable, but in this case, SIR models are used to calculate more accurately how the duration of flights affects the number of incoming infectees. The prediction of the number of cases per country through Poisson models using these variables has been the main objective of this project.

In order to warrant that the data are as reliable as possible, our study is restricted to Europe. In addition, we have only considered dates between 15th February 2020 and 15th March 2020, since before these dates there were almost no detectable cases in Europe, and after that, the states started to apply severe restrictions on air traffic, so that contagions within the country became much more relevant than those imported from this date onwards.

Since COVID-19 had an average incubation time of one week, country positives will be predicted after one week, knowing the imported risk and the current confirmed cases. Thus, a new model will be trained for each day. The main model is the Poisson model, but given the overdispersion phenomenon experimented in it, two variants have also been proposed: the quasi-Poisson model and the negative binomial model.

After training the models, the compliance with the basic hypotheses was analyzed, as well as the relevance of the independent variables, the overdispersion, the estimator coefficients, residuals and goodness of fit. After all this, it has been concluded that, theoretically speaking, this model is adequate for our problem (the hypothesis are fulfilled), all the variables are relevant, the overdispersion increases over time (so

the variants of the Poisson model works better compared to the original over time), the most relevant independent variable is initially the imported risk (contagions from abroad), but finally confirmed cases are more relevant (contagions within the country) and the variance of the residuals does not increase over time.

Once we try to predict the next day with the model trained with the previous day, we can observe that every model has similar errors and shows good results that improve in percentage over time. With all these results, it is concluded that Poisson regression models and imported cases form good models for predicting the contagions of a country.

Tabla de contenidos

| | |
|--|-----------|
| 1. Introducción | 1 |
| 1.1. Objetivos | 2 |
| 1.2. Estructura del documento | 2 |
| 2. Estado del Arte | 5 |
| 2.1. Modelo SIR | 5 |
| 2.2. Riesgo Importado | 7 |
| 2.2.1. Cálculo del riesgo importado | 7 |
| 2.3. Modelos de regresión en epidemiología | 8 |
| 2.3.1. Modelo de regresión Poisson | 9 |
| 2.3.2. Modelos con sobredispersión | 10 |
| 2.3.2.1. Modelo quasi-Poisson | 10 |
| 2.3.2.2. Modelo de la binomial negativa | 11 |
| 2.4. Métricas para medir el error en regresión | 11 |
| 3. Desarrollo e implementación | 13 |
| 3.1. Extracción de datos | 13 |
| 3.1.1. Estructura de la base de datos | 13 |
| 3.1.2. Criterios para la extracción de datos | 15 |
| 3.1.3. <i>Set up</i> y formulación de consultas | 16 |
| 3.2. Análisis y predicción | 19 |
| 4. Resultados y discusión | 23 |
| 4.1. Análisis exploratorio | 23 |
| 4.2. Análisis de las predicciones | 26 |
| 4.2.1. Cumplimiento de las hipótesis | 26 |
| 4.2.2. Análisis de las variables a incluir | 28 |
| 4.2.3. Análisis de la sobredispersión | 28 |
| 4.2.4. Análisis de los coeficientes de los modelos | 28 |
| 4.2.5. Análisis de residuos | 29 |
| 4.2.6. Bondad de ajuste | 30 |
| 4.2.7. Métricas de evaluación | 30 |
| 5. Conclusiones | 41 |
| 5.1. Trabajo futuro | 42 |
| Bibliografía | 44 |

Capítulo 1

Introducción

La COVID-19 es un virus que comenzó a detectarse en diciembre de 2019, y en pocos meses se extendió hasta convertirse rápidamente en una pandemia mundial. Desde entonces, el interés por predecir cómo el virus puede extenderse por los países se ha incrementado en todo el mundo. Sin embargo, este es un tema que ya se ha tratado con otras epidemias, como la influenza [12], el ébola [5] o el zika [4]. De la misma forma, ya hay muchos estudios que han tratado de predecir la expansión de la COVID-19 teniendo en cuenta únicamente los contagios [7] o otros factores determinantes, como la tasa de vacunación y el estado meteorológico [1].

Sin embargo, no hay tantos que hayan mostrado interés directo por la influencia que el tráfico aéreo tuvo en los primeros momentos de la pandemia, aunque haya estudios como [15] que demuestran que fue uno de los factores determinantes. La principal razón se debe, posiblemente, a que se necesita una base de datos extensa que dé suficiente información sobre los vuelos y los pasajeros. Sin embargo, nosotros disponemos de una base de datos suficientemente rica para esta tarea. Por ello, este trabajo se centra en crear modelos de regresión para calcular el número de contagiados que hubo en países europeos sabiendo el tráfico aéreo que vino de países de riesgo.

Los modelos de regresión de Poisson ya han sido utilizados en [15] para calcular el número de infectados de países a partir del tráfico aéreo. Estos modelos son ideales para nuestro tipo de problemas, ya que intentamos calcular una variable basada en el conteo, con datos que deben ser un conteo. Sin embargo, este artículo solo utiliza como variable el número de pasajeros que llega a cada país en 2018. Aunque es una buena aproximación de cómo el tráfico aéreo influye en los contagios por COVID-19, no está teniendo en cuenta el origen de los infectados ni la situación de cada país, cosa que nosotros sí podemos hacer con la base de datos de la que disponemos.

Por otro lado, tenemos el concepto *riesgo importado* creado por García Moreno et al. en [14], donde para cada vuelo se tiene en cuenta el número de infectados que se estima que entran en el avión. Además, dentro del avión se simula el modelo epidemiológico SIR (Susceptibles-Infectados-Recuperados) [6, Capítulo 2.1] para tener en cuenta nuevos infectados que hayan podido producirse durante el vuelo. De esta forma, para cada vuelo tenemos el número aproximado de infectados que se han importado, y por tanto el número de infectados importado a cada aeropuerto o país por día. Este dato es mucho más preciso que el anterior, basado únicamente en los

pasajeros entrantes por país. Sin embargo, este dato por sí solo no vale para predecir el número de infectados que habrá en el país; se necesitan los modelos de regresión anteriormente mencionados.

1.1. Objetivos

Los objetivos que se abordan en este trabajo son:

- **Extraer datos** de la base de datos provista con todos los vuelos de 2020. Aunque es muy rica, es tan grande que tratar con ella no es una tarea trivial. Es necesario extraer el riesgo de cada vuelo, así como su correcta agrupación por países para obtener una base de datos de un tamaño razonable.
- Realización de un **análisis exploratorio**. Antes de predecir con nuestros modelos, hay que concluir si las fechas escogidas son adecuadas o si finalmente el tráfico aéreo no es una variable relevante. De la misma manera, puede que no todo el intervalo de fechas sea relevante y haya que podar algunos valores. También se va a realizar un análisis en busca de *outliers* que evite la pérdida de precisión de nuestro modelo.
- Aplicar **modelos de regresión de Poisson** para predecir el número de infectados en países sabiendo el riesgo importado de días anteriores. Como el tráfico aéreo es especialmente relevante en las fases tempranas de la pandemia, se ha decidido analizar la evolución de casos positivos en Europa durante febrero y marzo de 2020.
- Analizar la **calidad de los resultados**. Para ello, se realiza un análisis exhaustivo de los coeficientes de los modelos, de los residuos y del cumplimiento de las hipótesis básicas. Además, mediante las métricas habituales usadas en regresión se evaluará si el modelo es adecuado para el problema en cuestión.

1.2. Estructura del documento

El documento se compone de cuatro partes:

- **Estado del arte.** Repasa los conocimientos básicos para entender el resto del documento. Para ello, explica cómo se elabora el riesgo importado de cada vuelo, cómo funciona el modelo SIR y los modelos de regresión de Poisson que se utilizarán posteriormente.
- **Desarrollo e implementación.** Se detallan los pasos a seguir para llegar a los resultados y hacer el experimento repetible. En esta parte, se muestra el *set up* y los comandos de *Neo4j* utilizados para extraer la información, el análisis exploratorio y los pasos a seguir para analizar correctamente los resultados obtenidos con los modelos de regresión de Poisson.
- **Resultados y discusión.** Explica y analiza los resultados obtenidos con los modelos de regresión y en el análisis exploratorio. Se concluye si el uso del riesgo importado es relevante y la calidad de predicción de los modelos utilizados.
- **Conclusiones.** El documento finaliza mostrando los resultados más importantes y recopila las ideas más relevantes del trabajo. De la misma forma, indica las

Introducción

líneas futuras de investigación que no se han abarcado en este proyecto.

Capítulo 2

Estado del Arte

2.1. Modelo SIR

El modelo básico SIR (Susceptibles-Infectados-Recuperados) es uno de los modelos clásicos para modelizar problemas epidemiológicos. En él, se considera que una cierta enfermedad ha comenzado a proliferar en una zona cerrada y homogénea. Se distinguen tres variables dependientes del tiempo:

- **Susceptibles (S)**. Número de personas que aún no han sufrido la enfermedad.
- **Infectados (I)**. Número de personas que están contagiadas.
- **Recuperados (R)**. Número de personas que ya han pasado la enfermedad y no pueden contagiarse ni contagiar.

Como vemos, no se tienen en cuenta fallecidos en el modelo, ya que a efectos prácticos actúan como recuperados. Sin embargo, existen modelos más sofisticados en los que se pueden tener en cuenta las tasas de natalidad, mortalidad y mortalidad causada por la enfermedad para tener en cuenta que la población no es constante. No obstante, en el modelo clásico SIR la población N no depende del tiempo, y es resultado de la ecuación:

$$N = S(t) + I(t) + R(t). \quad (2.1)$$

Para el caso que nos concierne, este modelo constituye una buena aproximación de lo que deseamos calcular. Sin embargo, existen otros modelos para otros tipos de problemas epidemiológicos. Por ejemplo, el modelo SI es más adecuado para enfermedades en las que los infectados nunca se recuperan, y el modelo SIS sirve para aquellos casos en los que, si bien los infectados se recuperan de la enfermedad, pueden volver a contraerla.

Respecto a cómo quedan definidos los valores S, I, R conforme avanza el tiempo, se

calculan mediante la ecuación:

$$S(t+1) = S(t) - \beta S(t) \frac{I(t)}{N} \quad (2.2)$$

$$I(t+1) = I(t) + \beta S(t) \frac{I(t)}{N} - \alpha I(t) \quad (2.3)$$

$$R(t+1) = R(t) + \alpha I(t), \quad (2.4)$$

donde $\beta > 0$ es la tasa de infección y $\alpha > 0$ es la tasa de recuperación. Como vemos, el número de susceptibles que se infectan tras cada instante es directamente proporcional al número de susceptibles y al número de infectados. De la misma forma, el número de infectados que se recuperan es directamente proporcional al número de infectados. Además, sumando las tres ecuaciones resulta obvio el resultado adelantado en 2.1 en el que se aseguraba que la población permanece constante.

Tras fijar estos parámetros, solo queda escoger los parámetros iniciales, es decir, a tiempo 0. Se suele comenzar con casi toda la población susceptible, una pequeña cantidad de infectados y ningún recuperado.

Si se trabaja con tiempo infinitesimal, nuestro problema se convierte en un sistema de ecuaciones diferenciales, con el que se puede obtener un resultado aproximado de la evolución de la epidemia. No obstante, trabajando con tiempo discreto se pueden obtener simulaciones muy sencillas de estas ecuaciones. Por poner un ejemplo, en la Figura 2.1 se refleja la evolución del modelo SIR para el caso de 100.000 habitantes susceptibles iniciales y un único infectado inicial. Los parámetros α y β son $1/9$ y 0.253 , que como se verá en la Sección 2.2.1, son los valores elegidos para simular un modelo SIR de la pandemia de COVID-19.

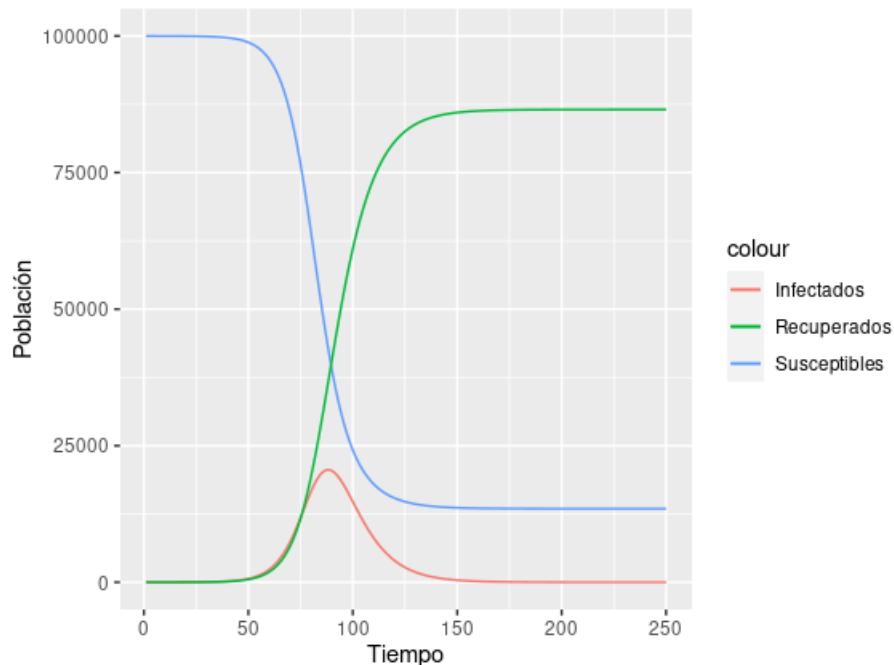


Figura 2.1: Evolución del modelo SIR en 250 unidades de tiempo.

Como vemos, tras un comienzo exponencial del número de infectados y un pico de

algo más de 200.000 infectados (con los parámetros α y β dados), el número de infectados vuelve a descender hasta que la enfermedad desaparece de nuevo. Uno de los puntos más interesantes de este modelo es que no toda la población sufre el virus, sino que aproximadamente una quinta parte seguirá siendo susceptible cuando acabe la pandemia.

2.2. Riesgo Importado

Como ya se ha mencionado, el tráfico aéreo tuvo una gran influencia en las fases tempranas de la pandemia de COVID-19. Prediciendo los aeropuertos con mayor riesgo de propagar el virus, los estados podrían haber tomado medidas cautelares que hubieran mitigado la fase explosiva de la pandemia.

Teniendo una base de datos completa con todos los vuelos a nivel internacional, el número de pasajeros y el número de infectados por región, podemos estimar cuántos contagiados viajan en cada vuelo, y así determinar cuántos infectados se estima que recibirá cada aeropuerto diariamente. Es a esta estimación a la que llamaremos **infectados importados** o **casos importados**. Sin embargo, en vuelos de larga duración se espera que haya un cierto contagio entre pasajeros.

Es aquí donde entra el concepto de **riesgo importado**: un parámetro que estima el número de infectados que desembarcan de cada vuelo. Con este parámetro no solo se puede estimar los infectados entrantes de cada vuelo, sino que además se tiene en cuenta el tiempo que los pasajeros pasaron en el avión para fabricar un modelo SIR que estime mejor el número final de infectados. La suma de todos los vuelos que llegan a un cierto aeropuerto nos dará, por lo tanto, el riesgo importado de dicho aeropuerto. Además, se puede seguir el mismo procedimiento para estimar cuántos infectados entran en una región o un país.

2.2.1. Cálculo del riesgo importado

Para calcular el riesgo importado de un vuelo, primero debemos **estimar el número de pasajeros contagiados** que entran en el avión. Para esto, se siguen los siguientes pasos (para cada paso se adjunta la fuente que justifica el factor que se utiliza):

1. Se utilizan los datos disponibles para obtener el número de contagiados durante los últimos 7 días en la región, lo cual se considera el periodo de incubación promedio [11].
2. Se multiplica dicho valor por 10, ya que al comienzo de una pandemia los datos de infectados no son del todo fiables [10].
3. Dado que se presupone que los infectados en el avión son asintomáticos, presintomáticos o con síntomas leves, el número anterior se multiplica por un factor de 0.75, dado que no son tan contagiosos [3].
4. Se multiplica el resultado por 0.5 para reflejar que, como los pasajeros suelen tener una renta elevada, tienen menos probabilidades de contraer la enfermedad [2].

Con todo esto, finalmente el número de pasajeros contagiados estimado en el vuelo

2.3. Modelos de regresión en epidemiología

es:

$$n \cdot \frac{10 \cdot 0.75 \cdot 0.5 \cdot I}{N} = n \cdot \frac{3.75 \cdot I}{N},$$

donde n es el número de pasajeros, I el de infectados indicados en la región, y N la población de la región.

Tras esto, para tener en cuenta la duración del vuelo, **se aplica un modelo SIR** en el que se comienza con el número ya mencionado de infectados, el resto de pasajeros susceptibles y ningún pasajero recuperado. Para determinar este modelo en el caso de la COVID-19, necesitamos los parámetros α y β :

- La tasa de recuperación α es $\frac{1}{9}$, ya que en promedio un individuo se recupera en 9 días de la COVID-19.
- La tasa de contagio β es calculada teniendo en cuenta múltiples factores, como que los usuarios del avión llevan mascarilla; o si el avión ha sido llenado completamente o solo a mitad de su capacidad, dejando un asiento libre entre pasajeros. Siguiendo esta casuística, la tasa de contagio es:
 - 0.253 si el avión está completo y todos los pasajeros llevan mascarilla, en un avión cuyas filas están separadas en dos partes (por un pasillo central) con tres asientos en cada parte. Para el caso de no disponer de mascarilla u otro tipo de avión esta tasa también se podría recalcular, ya que depende de las distancias entre los asientos y de la fila en la que los pasajeros se sitúen con respecto a la que ocupa el infectado.
 - 0.081 si el avión ha dejado libres los asientos intermedios, con las mismas indicaciones que en el caso anterior.

Con estos datos y las ecuaciones obtenidas en la Sección 2.1, el tiempo se contabiliza en minutos para obtener iterativamente los valores finales del modelo SIR.

2.3. Modelos de regresión en epidemiología

El modelo SIR que hemos visto es especialmente útil si nos encontramos en un entorno cerrado. Sin embargo, este modelo muestra ciertas limitaciones si queremos hallar los infectados dentro de un país, ya que solo podríamos tener en cuenta los infectados del propio país. Esto es un grave problema, ya que uno de los factores más importantes al principio de la pandemia fue el número de casos importados.

De la misma forma, el riesgo importado estima la cantidad de casos que llegan de otros lugares al aeropuerto/región/país de destino. Es un buen indicador de la cantidad de casos que habrá en una o dos semanas, en especial en los primeros momentos de la pandemia. Sin embargo, no podemos predecir directamente el número de infecciones que habrá en un futuro con este dato.

Por estas razones, surge la necesidad de usar modelos alternativos para predecir el número de infectados por país. En concreto, los modelos de regresión de Poisson son ideales para nuestro caso, ya que se utilizan cuando, de alguna forma, nuestra variable de respuesta (en este caso el número de infectados) se puede interpretar como un conteo.

Los tres modelos de regresión que se van a tratar son el modelo **Poisson**, el modelo **quasi-Poisson** y el modelo de la **binomial negativa**.

2.3.1. Modelo de regresión Poisson

Este modelo se basa en que la variable dependiente sea una variable aleatoria de Poisson, por lo que los valores que adopta son enteros no negativos. El valor que debemos estimar, λ , representa el promedio de ocurrencias por unidad. En nuestro caso, representará el promedio de infectados por día en una cierta región (y por lo tanto la predicción final de infectados).

Siendo las variables independientes $\{x_1, x_2, \dots, x_n\}$, el número de infectados esperados se define como:

$$\log \lambda = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n, \quad (2.5)$$

donde los coeficientes β_i se determinan mediante la minimización del error cuadrático (aunque existen más métodos, como máxima verosimilitud).

Sin embargo, en epidemiología hay veces que el número bruto de infectados no tiene tanto sentido como el número de infectados proporcional a la población. En este caso, se suele introducir un **offset** o compensación que hace que las predicciones se calculen en base a la población, aunque por supuesto el número de infectados bruto es información recuperable. Siendo N la población total de la región, en este caso la regresión se calcula como:

$$\log \frac{\lambda}{N} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n,$$

y tras aplicar propiedades básicas de logaritmos obtenemos:

$$\log \lambda = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \log N.$$

Es decir, los coeficientes serán exactamente los mismos, salvo β_0 , que se verá reducido por el nuevo término.

Por otro lado, si queremos hacer inferencia sobre nuestro modelo correctamente, se deben cumplir cuatro hipótesis básicas:

- La variable de respuesta (en este caso, el número de infectados por día en una cierta región) debe seguir una distribución de Poisson. A continuación se muestran algunas distribuciones de Poisson a modo de ejemplo en la Figura 2.2 con distintos λ .
- Las distintas observaciones de la variable respuesta deben ser independientes.
- La media y la varianza de la variable respuesta deben ser iguales.
- Debe existir linealidad entre el logaritmo de la variable respuesta y las variables regresoras.

En el caso de la última hipótesis, muchas veces es salvable gracias a transformaciones de las variables independientes. Lo mismo ocurre con la tercera hipótesis, ya que es muy normal que la varianza esté por encima de la media. A este fenómeno se le conoce como **sobredispersión**, y aunque esto impide que se cumplan las hipótesis del modelo, existen variantes del modelo de Poisson creadas para mitigar este problema.

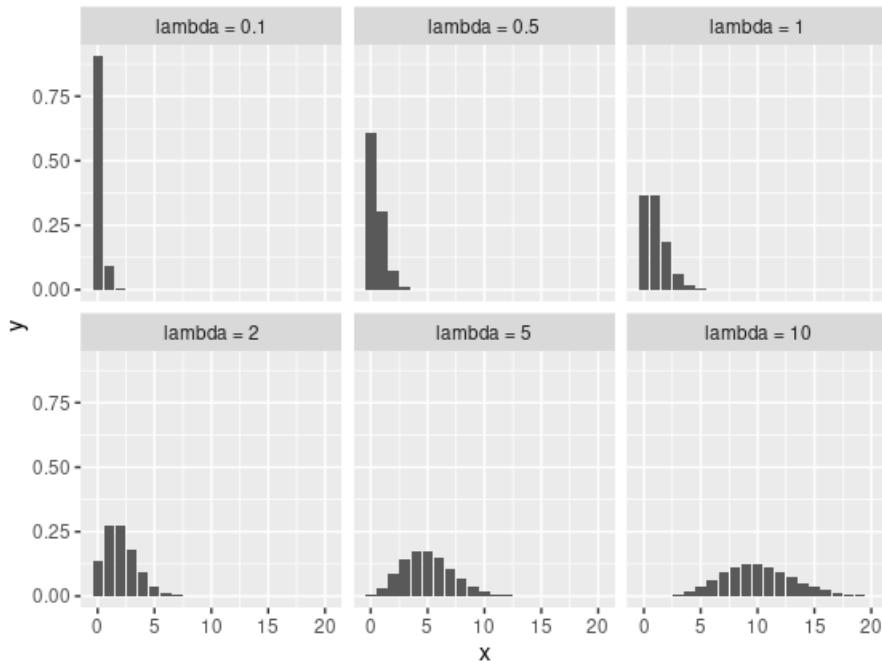


Figura 2.2: Función de densidad de la distribución de Poisson para varios valores λ .

2.3.2. Modelos con sobredispersión

Estos modelos deben usarse cuando la variable respuesta varía por encima de lo que se espera en un modelo de Poisson. Si no se toman medidas, estaremos creando modelos con desviaciones de los errores por debajo de sus valores reales, por lo que los p -valores generados también serán artificialmente más pequeños de lo que realmente son.

Se van a tratar dos modelos que tratan la sobredispersión: el quasi-Poisson y la binomial negativa.

2.3.2.1. Modelo quasi-Poisson

Este modelo es el más intuitivo de los dos. Si los errores calculados son demasiado pequeños, los multiplicamos por un factor en función de la magnitud de la sobredispersión. Para definir este factor, primero tenemos que definir el residuo de Pearson de cada predicción. Siendo O_i el valor real (u observado) y E_i el valor predicho (o esperado), el residuo de Pearson se define como:

$$r_i = \frac{O_i - E_i}{\sqrt{E_i}}. \quad (2.6)$$

Una vez definidos los residuos, el factor se define como la suma de los cuadrados de los residuos dividida entre sus grados de libertad:

$$\phi = \frac{\sum_{i=1}^n r_i^2}{n - p}, \quad (2.7)$$

donde n es el número de muestras y p es el número de parámetros del modelo.

Ahora, podemos multiplicar la varianza de los errores de los residuos por este factor y obtener un nuevo valor, que inflará los p -valores y permitirá que obtengamos intervalos de confianza más realistas.

2.3.2.2. Modelo de la binomial negativa

La binomial negativa es una distribución alternativa a la de Poisson donde en vez de un solo parámetro se determinan dos, dando mayor flexibilidad al modelo. Matemáticamente, su definición se basa en la repetición de experimentos de Bernoulli, es decir, experimentos donde existe una probabilidad p de éxito y una probabilidad $1 - p$ de fracaso. En concreto, la distribución consiste en el número de fracasos que hemos tenido hasta haber conseguido un número $r \in \mathbb{N}$ de aciertos.

Sin embargo, aunque no hay una forma sencilla de relacionar esta definición con el problema que nos atañe, existe una fuerte relación con la distribución de Poisson que nos permite volver a nuestro problema. En concreto, pongamos que nuestra variable dependiente Y sigue una binomial negativa, es decir, $Y \sim BN(r, p)$. En tal caso, esto es equivalente a afirmar que $Y \sim Poisson(\lambda)$, pero λ ahora no es un valor fijado, sino que es otra variable aleatoria con distribución $\lambda \sim Gamma(r, \frac{p}{1-p})$.

En esta distribución, el parámetro de dispersión viene dado por:

$$\text{sobredisp} = \frac{(1-p)^2 r}{p^2}. \quad (2.8)$$

2.4. Métricas para medir el error en regresión

Las métricas en regresión comparan los valores observados con los predichos, de forma que un valor mayor indica una mayor discrepancia entre el modelo y las observaciones. En cuanto a la notación, se va a suponer que hay n observaciones, la i -ésima observación se denota como y_i y su correspondiente predicción se denota como \hat{y}_i .

En este trabajo, hay dos métricas que son de interés:

- RMSE (Raíz del error cuadrático medio). La Ecuación 2.9 define dicha métrica. El cuadrado evita la influencia del signo, mientras que la raíz cuadrada reescala el error para que sea interpretable como un error promediado. El punto fuerte de esta fórmula es la intuición con la que se pueden interpretar los resultados, pero a cambio tenemos que los resultados no serán comparables si cambiamos la escala.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}}. \quad (2.9)$$

Evidentemente, no es lo mismo equivocarse en 100 casos si había que predecir 2 casos o si había que predecir 2.000. Es por esto que este método se dice que no depende de la escala.

- SMAPE (Error porcentual de la media simétrica absoluta). La Ecuación 2.10 define esta métrica. Como indica la definición, en este caso esta métrica indica intuitivamente el error relativo que se ha cometido al hacer la predicción. Existe

2.4. Métricas para medir el error en regresión

también la métrica MAPE, muy similar a esta, pero no está preparada para que los valores observados sean 0, caso que sí se da en este trabajo.

$$SMAPE = \frac{2}{n} \sum_{i=1}^n \frac{|\hat{y}_i - y_i|}{|\hat{y}_i| + |y_i|}. \quad (2.10)$$

Capítulo 3

Desarrollo e implementación

La implementación y desarrollo del trabajo se divide en dos partes: primero, la extracción de datos para conseguir la información necesaria en formato tabular y después un análisis de los datos, que incluye una parte exploratoria y la predicción con modelos de regresión de Poisson.

Nuestra principal suposición consiste en que, al menos al comienzo de la pandemia de la COVID-19, el tráfico aéreo desde países de riesgo fue uno de los factores determinantes para la difusión del virus a nivel internacional. Por ello, se intentará predecir el número de infectados en los países europeos tras una semana de los datos desde el 15 de febrero hasta el 15 de marzo.

En concreto, el trabajo se centra en la predicción de positivos de cada país de Europa basándose en el número de positivos que había una semana antes y el tráfico aéreo desde distintos países. La combinación del tráfico aéreo proveniente de cada país hacia nuestro país destino y el número de infectados en el país origen nos darán la información que buscamos: el número estimado de contagiados que llegaron a un cierto país por vía aérea.

3.1. Extracción de datos

La base de datos provista incluye información sobre todos los vuelos desde enero hasta septiembre de 2020, combinada con información sobre los positivos por COVID-19 de cada región. Sin embargo, para este trabajo solo nos interesan los valores mencionados anteriormente, que son un pequeño subconjunto de toda la información de la que disponemos. En concreto, la base de datos inicial está compuesta por más de 11 GB, mientras que la base de datos que acabamos usando posteriormente no supera los 3 MB.

3.1.1. Estructura de la base de datos

Esta información viene dada en un modelo de base de datos orientada a grafos en formato implementada en lenguaje **Neo4j**. Esto quiere decir que cada **nodo** es una entidad, que bien puede ser un cierto aeropuerto, un cierto vuelo, un cierto país...

Cada uno de estos nodos, a su vez, puede tener propiedades. Por ejemplo, un vuelo tiene, entre otras propiedades, la hora de salida y llegada, el porcentaje de ocupación

o la marca comercial del avión.

Finalmente, los nodos se relacionan entre sí mediante **aristas** dirigidas. Por ejemplo, en nuestra base de datos los países y los aeropuertos se relacionan entre sí mediante una arista hacia los aeropuertos llamada *INFLUENCE_ZONE*, que indica en qué país opera dicho aeropuerto.

Respecto a la estructura en sí de la base de datos, se rige por el siguiente esquema mostrado en la Figura 3.1.

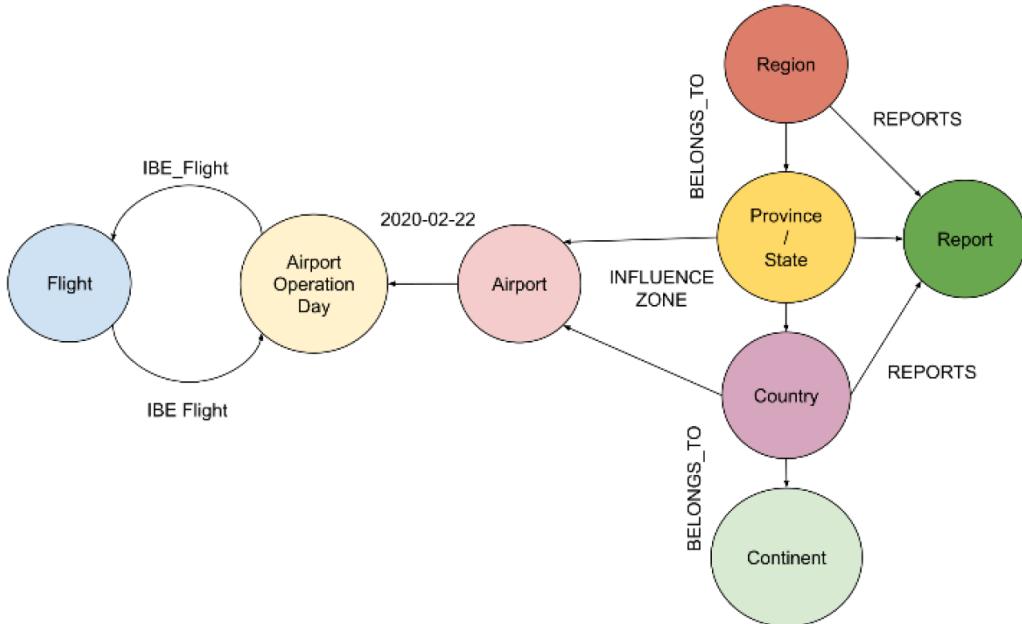


Figura 3.1: Estructura de la base de datos. Fuente: [14].

A continuación se procede a detallar los nodos y aquellas propiedades que resultarán útiles para la extracción de datos:

- *Flight* son los datos correspondientes a cada vuelo. Las propiedades relevantes son la fecha de salida (*dateOfDeparture*), la fecha de llegada (*dateOfArrival*), el riesgo importado del vuelo (*flightImportRisk*), el aforo (*seatsCapacity*) y el porcentaje de ocupación estimado (*occupancyPercentage*). En este caso, no es necesario extraer la duración de los vuelos, ya que el riesgo importado ya ha sido calculado previamente en la base de datos.
- *AirportOperationDay* (AOD a partir de ahora) muestra ciertos datos de cada aeropuerto en cada día. En nuestro caso, ninguna propiedad es relevante. Si la arista va dirigida hacia el nodo *Flight* bajo el nombre *IBE_Flight*, el AOD correspondiente es el origen del vuelo. Si la arista toma el sentido contrario, el AOD es el de destino.
- *Airport* es la entidad correspondiente a cada aeropuerto. De nuevo, no hay ninguna propiedad relevante. Las aristas apuntan a las entidades AOD, siendo el nombre de esta relación la fecha del AOD.
- *Continent*, *Country*, *Province/State*, *Region* corresponden respectivamente a los continentes, países, provincias y regiones donde se encuentran los aeropuertos.

Desarrollo e implementación

Están relacionados con los aeropuertos mediante una arista bajo el nombre *INFLUENCE_ZONE* y que una zona pertenezca a otra se indica mediante la arista *BELONGS_TO*.

No todos los aeropuertos tienen regiones o provincias asociadas, ya que el nivel de especificidad depende del país en cuestión. Por ejemplo, en Estados Unidos y China sí que se ofrece la provincia y región de sus aeropuertos, en España solo la provincia (que es la CC.AA.) y en Afganistán solo el país en sí.

Respecto a las propiedades relevantes en este caso, el nombre de los países viene dado por *countryName*, el de los continentes por *continentName*... Por otro lado, la propiedad *population*, que indica la población de la región en cuestión, también será relevante.

- *Report* muestra los datos de infecciones de COVID-19 en una cierta región y un cierto día provisto por la base de datos *Johns Hopkins* [9]. En Europa, solo se tienen los datos de infectados por países, y solamente se tienen *Reports* más específicos en países de gran tamaño como EEUU o China. En este caso, la relación que les une con la región en cuestión es *REPORTS*, dirigido hacia el informe en sí. Las propiedades relevantes en este caso son *releaseDate*, indicando la fecha del informe, y *confirmed* indicando el número de casos confirmados ese día.

Con objetivo de sintetizar las propiedades relevantes de cada entidad, se adjunta la Tabla 3.1 a modo aclaratorio.

Tabla 3.1: Propiedades relevantes de los nodos del grafo para este problema.

| <i>Flight</i> | <i>AOD</i> | <i>Airport</i> | <i>Country, Province...</i> | <i>Report</i> |
|----------------------------|------------|----------------|-----------------------------|---------------------|
| <i>dateOfDeparture</i> | | | <i>regionName</i> | <i>releasedDate</i> |
| <i>dateOfArrival</i> | | | <i>population</i> | <i>confirmed</i> |
| <i>flightIfFinal</i> | | | | |
| <i>seatsCapacity</i> | | | | |
| <i>occupancyPercentage</i> | | | | |

3.1.2. Criterios para la extracción de datos

No solo la extracción eficiente de datos va a ser relevante, sino también unos criterios correctos que nos ayuden a realizar predicciones precisas.

Para empezar, como ya se ha comentado anteriormente, las **fechas** con las que vamos a trabajar están comprendidas entre febrero y marzo de 2020, ambos incluidos. Esto es debido a que el tráfico aéreo tuvo una influencia especial en los momentos iniciales, ya que es cuando los infectados producidos por movimientos internacionales eran mucho más influyentes que los intranacionales. Además, a finales de marzo empezaron confinamientos en muchas partes de Europa, lo cual empezaría a producir predicciones menos dependientes de nuestras variables regresoras. Posteriormente, en el análisis explicativo se concluirá que solo es necesario utilizar los datos que van desde el 15 de febrero al 15 de marzo, ambos inclusive.

Como ya se ha mencionado en varias ocasiones, solo se van a realizar predicciones sobre **países europeos**. La causa principal es que al resto de países la pandemia llegó después, así que la mayoría de países solo nos aportarían ceros. Además, la fiabilidad

de estos datos es mayor que en muchos otros países cuyo presupuesto en sanidad es más reducido.

Respecto a **China**, aunque no se van a predecir sus estadísticas sobre positivos, sí los usaremos para estimar los infectados que llegan a Europa. Esto se hace porque en este caso no estamos cuestionando la fiabilidad de los datos en China, sino que se excluye el país porque se tomaron medidas drásticas muy pronto y podría resultar un *outlier* para nuestras predicciones.

Respecto a los **vuelos dentro del propio país**, uno podría pensar si son relevantes o no para calcular el número de casos importados al país. En este caso no se han tenido en cuenta, ya que el hecho de que el tráfico aéreo esté íntimamente relacionado con el comienzo de la pandemia se debe esencialmente a que el virus se importó de otros países. Sin embargo, esta es una decisión de diseño y el punto de vista contrario también podría ser perfectamente válido y justificable.

Aunque existan datos de regiones dentro de los países (por ejemplo, en España se tomaron datos a nivel de CC.AA.), en la base de datos provista solo se contemplan *Reports* a nivel nacional en Europa. Por ello, aunque trabajar por regiones nos aportaría una fuente de datos más rica, nos vemos obligados a trabajar **a nivel nacional**.

Por último, nuestra variable dependiente son los infectados detectados **7 días después**, es decir, se predicen los infectados desde el 22 de febrero hasta el 22 de marzo. Este valor ha sido determinado por diversos estudios [11], que situaban el tiempo de incubación entre 4.5 y 7 días.

3.1.3. Set up y formulación de consultas

En esta sección se especifican las instrucciones que se deben seguir para reproducir de nuevo la extracción de datos llevada a cabo.

Para empezar, todas las operaciones han sido llevadas a cabo mediante la interfaz de **Neo4j Desktop**, la cual es especialmente recomendable si no se ha trabajado previamente con estas tecnologías. Los requisitos iniciales para las ejecuciones que siguen son:

- En el archivo de configuración, los atributos *dbms.memory.heap.initial_size* y *dbms.memory.heap.max_size* no deben estar fijados, es decir, hay que borrar o comentar las líneas en las que se les asigna un valor. Debido a los grandes volúmenes con los que se va a trabajar, es mejor flexibilizar estos valores para evitar errores durante las transacciones.
- Se debe instalar el *plugin APOC*, que va a permitirnos importar bases de datos de gran volumen.

Tras todo esto, podemos iniciar nuestra base de datos e importar el archivo mediante el comando:

```
1 CALL apoc.import.graphml("covid19.graphml", {readLabels: true, batchSize:1000})
```

Al fijar *batchSize*, estamos evitando que se almacene todo el archivo en memoria, ya que produciría un error en lectura por superar la memoria RAM disponible. Si fuera necesario, este valor podría reducirse aún más, dependiendo de los requisitos del

Desarrollo e implementación

ordenador en que se vayan a ejecutar las *queries*. Esta operación y las que siguen pueden tardar varios minutos cada una.

Para podar y reducir la base de datos se han llevado a cabo las siguientes instrucciones:

1. La base de datos abarca los vuelos desde enero hasta septiembre de 2020. Para reducir el tamaño, eliminamos todos los vuelos que no pertenezcan a febrero y marzo.¹

```
1 :auto WITH date({year: 2020, month: 2, day: 1}) AS fechaInicio,
2 date({year: 2020, month: 4, day: 1}) AS fechaFin
3 MATCH (f:FLIGHT)
4 WHERE date(f.dateOfArrival) < fechaInicio OR date(f.dateOfArrival) >= fechaFin OR f.
    dateOfArrival IS NULL
5 CALL {
6     WITH f
7     DETACH DELETE f
8 } IN TRANSACTIONS OF 20 ROWS
```

2. Ahora solo se han borrado los vuelos. Sin embargo, muchos *AirportOperationDay* se han quedado sin ningún vuelo asociado. A continuación se eliminan también.

```
1 :auto MATCH (aod:AirportOperationDay) WHERE NOT exists((aod)-[]-(:FLIGHT))
2 CALL {
3     WITH aod
4     DETACH DELETE aod
5 } IN TRANSACTIONS OF 20 ROWS
```

3. Solo estamos interesados en aquellos aeropuertos que sean o bien europeos o bien chinos. El resto deben ser eliminados.

```
1 MATCH (a:Airport)<-[:INFLUENCE_ZONE]-()-[{:BELONGS_TO*0..2}]->(country:Country)-[:.
    BELONGS_TO]->(cont:Continent)
2 WHERE country.name <> "China" AND cont.continentName <> "Europe"
3 CALL {
4     WITH a
5     DETACH DELETE a
6 }
```

4. De nuevo, muchos vuelos y *AirportOperationDay* se han quedado sin aeropuerto asociado. Se eliminan en el siguiente orden:

```
1 :auto MATCH (aod:AirportOperationDay) WHERE NOT exists((aod)<--(:Airport))
2 CALL {
3     WITH aod
4     DETACH DELETE aod
5 } IN TRANSACTIONS OF 20 ROWS
```

```
1 :auto MATCH (f:FLIGHT) WHERE NOT exists((:AirportOperationDay)<--(f))
2 CALL {
3     WITH f
4     DETACH DELETE f
5 } IN TRANSACTIONS OF 20 ROWS
```

¹Si la operación falla porque se supera la memoria que puede utilizar el ordenador, se recomienda reducir el número de transacciones. En cambio, si el PC puede soportar grandes cargas, aumentar el tamaño de las transacciones puede reducir mucho los tiempos de espera.

5. Por último, para facilitar posteriores *queries*, recordemos que el nombre de un país venía dado por el atributo *countryName*. Sin embargo, en las regiones venía dado por *regionName*, en los continentes por *continentName*... Esto puede hacer que tengamos que trabajar con *queries* mucho más complicadas. Por ello, todas estas propiedades pasarán a llamarse genéricamente *name*. Se pone a continuación como ejemplo el caso de los países:

```

1 MATCH (c:Country)
2 WITH collect(c) AS countries
3 CALL apoc.refactor.rename.nodeProperty("countryName", "name", countries)
4 YIELD batches, total, timeTaken, committedOperations
5 RETURN batches, total, timeTaken, committedOperations

```

6. Finalmente, se eliminan aquellos vuelos que se produzcan dentro de un mismo país.

```

1 :auto MATCH (n:FLIGHT)-->(op:AirportOperationDay)<--(:Airport)<-[ :INFLUENCE_ZONE]-() -[ :BELONGS_TO*0..2]->(c),
2 (n)<--(opOri:AirportOperationDay)<--(:Airport)<-[ :INFLUENCE_ZONE]-() -[ :BELONGS_TO
*0..2]->(cOri)
3 WHERE c.name = cOri.name
4 CALL {
5     WITH n
6     DETACH DELETE n
7 } IN TRANSACTIONS OF 20 ROWS

```

Con todas estas modificaciones, la base de datos ya no pesa los 11 GB que tenía inicialmente, sino únicamente 2GB. Aunque sigue siendo un tamaño considerable, ahora sí se pueden ejecutar queries en ordenadores con no demasiados recursos. Sin embargo, si esto no fuera suficiente se puede separar la base de datos en varias partes (por ejemplo, una para febrero y otra para marzo).

Finalmente, se va a ejecutar la consulta que permitirá extraer el CSV con los datos que nos interesan, cuyas columnas son:

- Fecha.
- Nombre del país de destino de los vuelos, casos confirmados en ese país y casos confirmados una semana después.
- Nombre del país de origen de los vuelos, casos confirmados en ese país y población de ese país.
- Número de pasajeros diarios que vuelan desde el país de origen al país de destino.
- Riesgo importado total de los vuelos correspondientes.

Con esto, tenemos $num_paises \cdot (num_paises - 1) \cdot num_dias$ filas en nuestra base de datos, lo cual ya se puede considerar tratable para trabajar con normalidad. La consulta que devuelve este CSV es:

```

1 MATCH (n:FLIGHT)-->(op:AirportOperationDay)<--(:Airport)<-[ :INFLUENCE_ZONE]-() -[ :BELONGS_TO
*0..2]->(c)-[:REPORTS]->(r:Report),
2 (c)-[:BELONGS_TO*1..3]-(cont:Continent),
3 (c)-[:REPORTS]->(rToday:Report),
4 (n)<--(opOri:AirportOperationDay)<--(:Airport)<-[ :INFLUENCE_ZONE]-() -[ :BELONGS_TO*0..2]->(cOri)
    -[:REPORTS]->(rOri:Report)

```

Desarrollo e implementación

```
5 WHERE date(n.dateOfArrival) + Duration({days: 7}) = date(r.releaseDate)
6 AND n.dateOfArrival = rToday.releaseDate
7 AND rOri.releaseDate = n.dateOfArrival
8 AND cont.continentName = "Europe"
9 RETURN
10 c.name AS Zone,
11 n.dateOfArrival AS Date,
12 rToday.confirmed AS ConfirmedCases,
13 r.confirmed AS ConfirmedCasesWeekLater,
14 cOri.name AS Origin,
15 SUM(n.occupancyPercentage*n.seatsCapacity/100) AS NumPassengers,
16 rOri.confirmed AS OriginConfirmed,
17 cOri.population AS OriginPopulation,
18 SUM(n.flightIfinal) AS ImportedRisk
19 ORDER BY c.name, n.dateOfArrival, cOri.name
```

Finalmente, se debe calcular a partir de esta base de datos el número de casos importados estimados en cada país de origen. Para ello, siendo pas_{ijk} el número de pasajeros que van un cierto día k desde el país i al país j , pobl_j la población del país i , import_{ik} el número estimado de casos importados al país i en la fecha k , EU el conjunto de países de Europa y conf_{ik} el número de casos confirmados en el país i en la fecha k , tenemos que:

$$\text{import}_{ik} = \sum_{j \in \text{EU}} \text{pas}_{jik} \frac{\text{conf}_{jk}}{\text{pobl}_j}. \quad (3.1)$$

Con esto, la tabla que obtenemos finalmente tiene como columnas: país, fecha, casos confirmados en la fecha correspondiente y una semana después, población, riesgo importado total (suma de los riesgo importados de cada país origen) y casos estimados importados obtenidos mediante 3.1.

3.2. Análisis y predicción

Tanto el análisis exploratorio como los posteriores modelos de predicción han sido llevados a cabo mediante la herramienta *R*.

Inicialmente, se ha llevado a cabo un análisis exploratorio para determinar que no todo febrero y marzo son de utilidad para llevar a cabo buenas predicciones (la primera quincena de febrero y la última de marzo se eliminaron, de hecho) y para saber si las variables regresoras tienen alguna correlación con los infectados confirmados una semana después. Además, se llevará a cabo una detección de *outliers* para evitar problemas posteriores.

Por otro lado, se han llevado a cabo las predicciones en sí con los tres modelos estadísticos descritos en la Sección 2.3. Para predecir los datos de un día, se ha entrenado un modelo con los datos del día anterior, de forma que para cada día hay tres modelos de predicción nuevos diseñados únicamente para predecir el día siguiente. Esto significa que no solo hay que centrarse en analizar los tres modelos, sino su evolución a lo largo del mes en el que hemos centrado el análisis. Todas las fases de dicho análisis se muestran a continuación:

1. Primero, es necesario comprobar que las **hipótesis teóricas** de la regresión de Poisson se cumplen en cada día. Son cuatro, contrastadas como sigue:

- La variable dependiente debe seguir una distribución de Poisson. Esta hipótesis es más complicada de tratar, ya que se debería estimar el parámetro

λ que mejor se adapte a nuestros datos (se estima mediante la media muestral de los datos) y después evaluar con algún test si los datos siguen una Poisson. Por simplicidad, se adjuntarán las gráficas de la distribución de los datos por fecha y se evaluará si los datos siguen una Poisson o no a partir de ellas.

- La media debe ser igual a la varianza de la variable dependiente. Se adjuntará la evolución del cociente varianza-media, pero se espera que haya un fenómeno de sobredispersión que debe ser analizado en profundidad mediante los modelos quasi-Poisson y binomial negativa.
 - Para comprobar que existe linealidad entre el logaritmo de la variable dependiente y las independientes, se adjuntan gráficas por día en las que se muestra si existe esa linealidad. En caso de no existir, se aplicarán transformaciones a las variables independientes para intentar mejorar el ajuste.
 - La independencia de las muestras debe ser abordado desde el marco teórico. Evidentemente, los datos no han sido recogidos aleatoriamente, ya que no ha habido ningún tipo de muestreo. Al ser todos cercanos en tiempo y lugar, es lógico que cada dato recogido esté estrechamente relacionado con el resto. Sin embargo, aunque exista cierta dependencia no hay mucho que se pueda hacer al respecto en este caso.
2. En segundo lugar, nuestro estudio se centra en la predicción de positivos futuros dados los casos entrantes en el país por vía aérea. No obstante, también se ha querido incluir el número de positivos del día actual como dato relevante. Sin embargo, es necesario saber si este dato es realmente importante para nuestro modelo o no aporta nueva información. Para ello, se llevará a cabo un **test ANOVA** que evalúe si la nueva variable debe permanecer o no.
 3. Una vez determinada la importancia de esta segunda variable, es importante concluir si existe o no **sobredispersión** y cómo abordarla. Para ello, se analizará en profundidad si existe dicho fenómeno y cuál es la evolución del factor de sobredispersión en caso de que exista en aquellos modelos que nos ofrecen herramientas para arreglar este problema (quasi-Poisson y binomial negativa).
 4. El siguiente paso es no tener en cuenta solo el resultado final, sino también entender el modelo en sí. De esta forma, podremos entender qué papel juega cada variable regresora dentro de la predicción. Para ello, se va a realizar un **análisis de coeficientes**, observando así la evolución de los modelos para ver cómo se adaptan a las nuevas situaciones ofrecidas por los datos.
- Por otro lado, también se ofrecen los intervalos de confianza de cada coeficiente. De esta forma, podremos saber cómo varía la precisión de las predicciones en función del tamaño de dichos intervalos. Además, es posible que sea mucho más ilustrativo que los intervalos de confianza de la predicción en sí, ya que al haber solo tres coeficientes y tres modelos vamos a tener 9 intervalos de confianza. En cambio, tenemos más de 20 países y 3 modelos, por lo que la visualización de datos sería más complicada si miramos a fondo las más de 60 líneas de predicción.
5. Además de los propios coeficientes, existen **p-valores asociados a los coeficientes** que nos contrastan la hipótesis nula $H_0 : \beta_i = 0$. Por lo tanto, un *p*-valor

cercano a 0 nos asegura que dicho coeficiente está influyendo en la predicción. Uno alto, en cambio indica que nuestro coeficiente podría ser 0 y que la variable asociada podría no tener ningún peso en la predicción. Todo esto tiene cierta relación con el análisis anterior, ya que si en las gráficas anteriores tenemos al 0 dentro del intervalo de confianza debemos esperar un *p*-valor alto.

6. Después, se incluye un **análisis de residuos**. Los residuos de Pearson deben seguir una **varianza constante** si el modelo es correcto. Los gráficos que se adjuntan pueden ser engañosos incluso para las vistas más entrenadas, por lo que el análisis de sobredispersión puede ser una métrica más analítica. Sin embargo, resultados no simétricos, los *outliers* o una excesiva dispersión en ciertos momentos pueden ser síntomas de problemas que son detectables mediante este análisis.
7. Además, se realizarán algunos test de **bondad de ajuste**, basados esencialmente en test χ^2 sobre los residuos. Se llevarán a cabo dos test:
 - En el primero, se contrasta la hipótesis nula de que un modelo constante, sin ninguna variable independiente, es mejor que el nuestro. Es deseable un *p*-valor bajo.
 - En el segundo, se contrasta la hipótesis nula de que nuestro modelo es mejor que un modelo saturado. Un modelo saturado es aquel en el que hay 0 grados de libertad, es decir, tenemos tantos parámetros como observaciones, y por lo tanto el error en entrenamiento es 0, causando un claro problema de *overfitting*. Es deseable un *p*-valor alto.
8. Para finalizar, se van a implementar varias **métricas** donde se compararán los tres modelos de predicción. Asimismo, se ha trabajado durante todo el trabajo con los infectados estimados en vez de con el riesgo importado, cuando el riesgo importado podría suponer una pequeña mejora en las predicciones. Para ello, se compararán también los modelos cambiando los infectados estimados por el riesgo importado. No es necesario repetir el resto de pasos para el riesgo importado, ya que las predicciones y los modelos son muy similares en ambos casos.

Las métricas a utilizar son dos:

- El RMSE para mostrar en valores absolutos cómo ha evolucionado el error de los modelos.
- El SMAPE para mostrar de forma relativa la evolución del error, ya que al ser mucho mayor el número de infectados en marzo se espera que los errores absolutos sean mucho mayores en marzo.

Para calcular estas métricas, en cada modelo se utiliza como datos de entrenamiento el número de infectados el día anterior (variable dependiente), el número de infectados hace 8 días y los casos importados de hace 8 días (variables independientes). Con esto, el conjunto de test son los infectados del día de referencia (variable dependiente), el número de infectados y casos importados de hace una semana (variables independientes). Con esto, estamos evaluando cómo funciona un modelo que utiliza los datos de hace una semana para conseguir los datos de hoy.

Capítulo 4

Resultados y discusión

4.1. Análisis exploratorio

Inicialmente, se ha comenzado a trabajar con todo febrero y marzo por ser el comienzo de la pandemia en Europa. Sin embargo, es posible que no todos los días sean útiles para nuestro propósito. A principios de febrero no había casi ningún positivo detectado en toda Europa, por lo que es posible que haya que podar el comienzo de la base de datos. Para decidir esto, se puede mirar la evolución de casos positivos por país en la Figura 4.1.

Como se puede ver, hay menos de una decena de casos en todo Europa durante estos días. Los modelos en este caso serían triviales, e incluso la binomial negativa experimenta errores numéricos en el ajuste.

Recordemos que íbamos a usar dos variables para predecir el número de positivos una semana después en cada país: el número de positivos el día actual y los casos importados (sustituible por el riesgo importado). Por lo tanto, es importante que haya algún tipo de relación entre estas variables y la variable dependiente. Para ello, se van a mostrar en una gráfica conjunta, Figura 4.1, que refleje la relación con los casos importados¹.

En este caso, se puede ver cómo a partir del 15 de marzo los países toman medidas restrictivas y empiezan a cerrar fronteras. La estimación de positivos importados, en muchos casos, o bien se frena o bien empieza a bajar, mientras que los positivos del país siguen subiendo cada vez más. Por esta razón, se ha decidido prescindir también de esta segunda quincena de marzo.

Además, se pueden ver relaciones más profundas entre los positivos una semana después y los infectados importados si se utilizan herramientas analíticas. Por ejemplo, en la Figura 4.2 observamos por un lado la evolución del coeficiente de correlación de Pearson entre estas dos variables, y por otro lado observamos esta misma correlación entre el logaritmo de la variable dependiente y los infectados importados. En la primera gráfica se puede observar cómo baja la correlación entre los días de interés respecto a la primera quincena. Al estar todos los datos de contagiados tan cercanos a cero, es posible que la correlación no sea un buen medidor durante esta quincena.

¹Nótese que los positivos una semana después no son los del 15 de febrero al 15 de marzo, sino los del 22 de febrero al 22 de marzo.

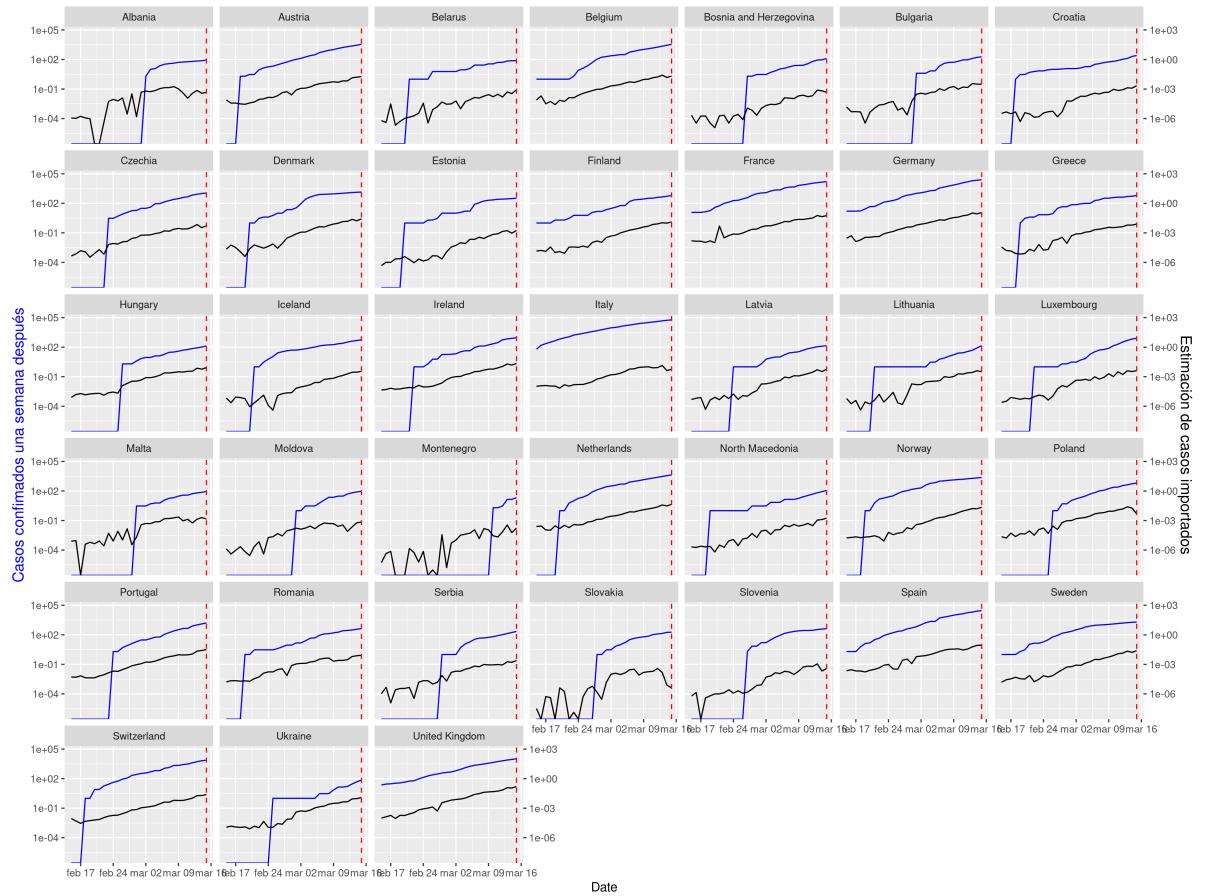


Figura 4.1: Evolución de los positivos una semana después junto con los casos importados por países (escala logarítmica).

Respecto a la segunda gráfica, podemos ver una mayor correlación durante los dos meses. Esto es positivo, y es preferible tener esta correlación alta en la segunda gráfica, ya que nuestro modelo busca linealidad entre el logaritmo de la variable dependiente y las independientes (como ya vimos en la Ecuación 2.5). Si, en cambio, tuviéramos una alta correlación en el primer caso, nuestro modelo debería estar más ligado a la regresión lineal u otros modelos similares.

Otro tema a tratar es si realmente hay una diferencia sustancial entre el riesgo importado y los casos importados. El primer caso es una versión más específica del segundo, ya que en el riesgo importado, además de estimar la densidad de casos en la región de origen de los vuelo, aplica un modelo SIR durante el viaje para tener en cuenta la duración de cada vuelo. Como los vuelos duran unas pocas horas, se espera generalmente que el riesgo importado, sea siempre mayor, pero sin una gran diferencia. En la Figura 4.3 se muestra la distribución de la diferencia entre el riesgo importado y los casos importados.

Como vemos, la diferencia es generalmente ínfima: el 77.6 % de los valores están entre 0 y 0.01 y el 83 % están entre 0 y 0.02. La media de las diferencias es 0.006 y su desviación típica es 0.021. De hecho, también se ha hecho una prueba *t de Student*, donde la hipótesis nula es que la diferencia entre ambas variables es en media 0. El *p*-valor obtenido de esta prueba es 0.9568, indicando que podemos confirmar que estas dos

Resultados y discusión

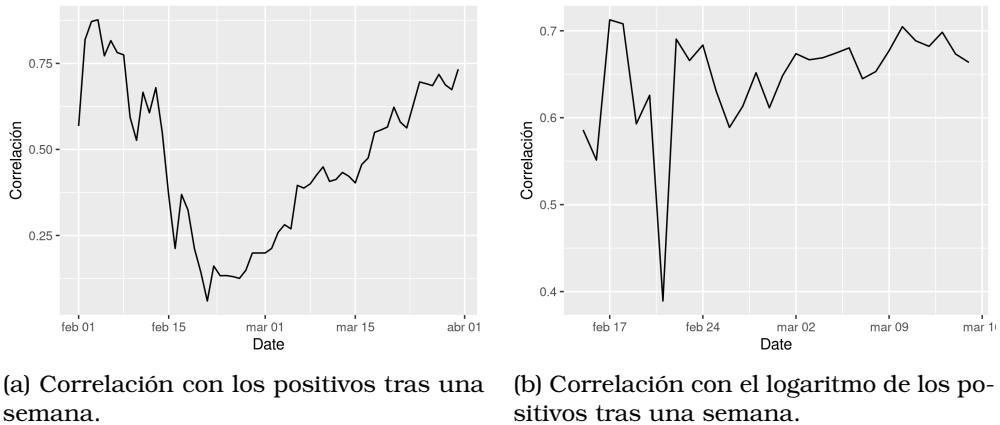


Figura 4.2: Correlación de los casos importados con transformaciones de la variable dependiente.

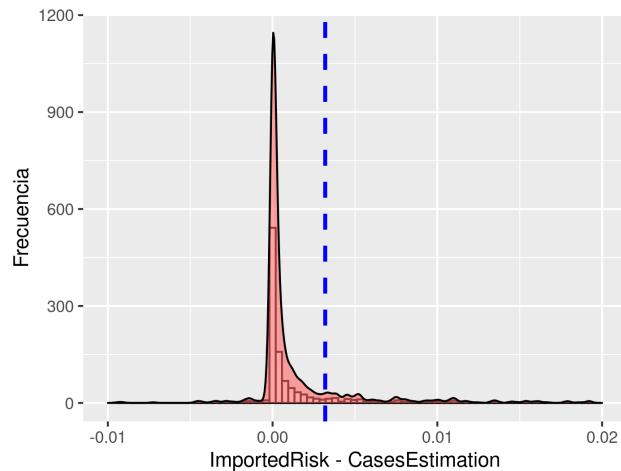


Figura 4.3: Diagrama de frecuencias de la diferencia entre el riesgo importado y la estimación de casos importados. La línea azul representa la media.

variables tienen, la misma media. Todo esto indica que ambas variables son muy parecidas, pero aún así el riesgo importado podría dar a las predicciones una cierta mejora, al ser más específico por tener en cuenta la duración de los vuelos, por lo que se hará un análisis posterior analizando mejoras en las predicciones. Sin embargo, al ser tan similares no se espera una mejora sustancial.

Finalmente, es muy importante **eliminar los outliers**. Como se está intentando predecir el logaritmo de la variable dependiente, el aumento desmesurado de las variables independientes (casos confirmados y infectados importados) en un dato a predecir aumentará el error de la predicción exponencialmente. Se han eliminado, por ello dos casos:

- El caso de Italia, ya que la pandemia comenzó mucho antes y sus datos son siempre *outliers*.
- Como se puede ver en la Figura 4.4, el día 21/02 la estimación de infectados importados se disparó más de 50 veces durante un día, volviendo después a

la normalidad. La predicción, al crecer exponencialmente, daría como resultado un número astronómico de infectados. Por ello, el dato de Francia durante el 21/02 será retirado también.

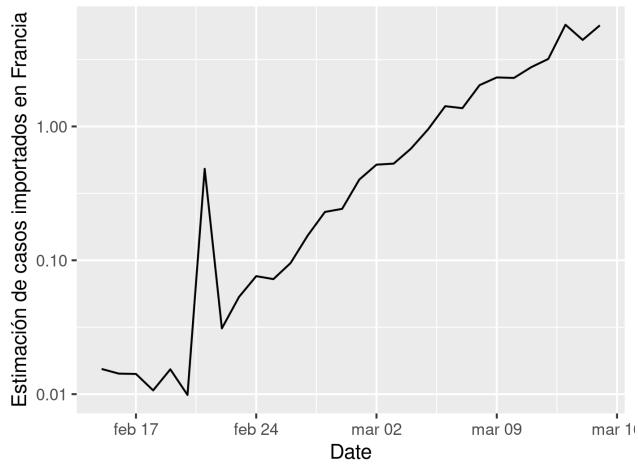


Figura 4.4: Evolución de la estimación de casos importados en Francia (escala logarítmica).

4.2. Análisis de las predicciones

Como ya se ha mencionado, se han utilizado tres modelos de predicción basados en la distribución de Poisson: el modelo de regresión de Poisson, el modelo quasi-Poisson y la binomial negativa. Se adjuntan las predicciones asociadas a cada país en la Figura 4.5.

Como puede verse, todos los modelos muestran buenas predicciones, y en general similares. No obstante, hay algunos países donde las predicciones difieren por mucho del valor real, sobrestimando el resultado real. En concreto, estamos hablando de Hungría, Lituania, Montenegro, Macedonia del Norte y Ucrania. La razón por la que podría estar ocurriendo esto es una incógnita, pero una posibilidad sería que estos países no hicieron todas las pruebas que deberían y la fiabilidad de sus datos fue muy reducida.

4.2.1. Cumplimiento de las hipótesis

Dado que la hipótesis de independencia ya se ha discutido teóricamente, quedan tres a tratar: la variable dependiente debe seguir una distribución de Poisson, la media debe ser igual que la varianza y linealidad tal y como se muestra en la Ecuación 2.5.

Comencemos por ver si la variable dependiente sigue una **distribución de Poisson**. En la Figura 4.6 se pueden ver los histogramas de las frecuencias de casos confirmados por fecha. Como ya se ha mencionado, no se va a comprobar si sigue una Poisson de forma cuantitativa, sino que más bien se va a comprobar que las gráficas tienen la forma que deberían (es decir, si tienen la forma característica de alguna de las gráficas mostradas en 2.2).

Como se puede ver, todas las gráficas se adaptan a una Poisson con un parámetro

Resultados y discusión

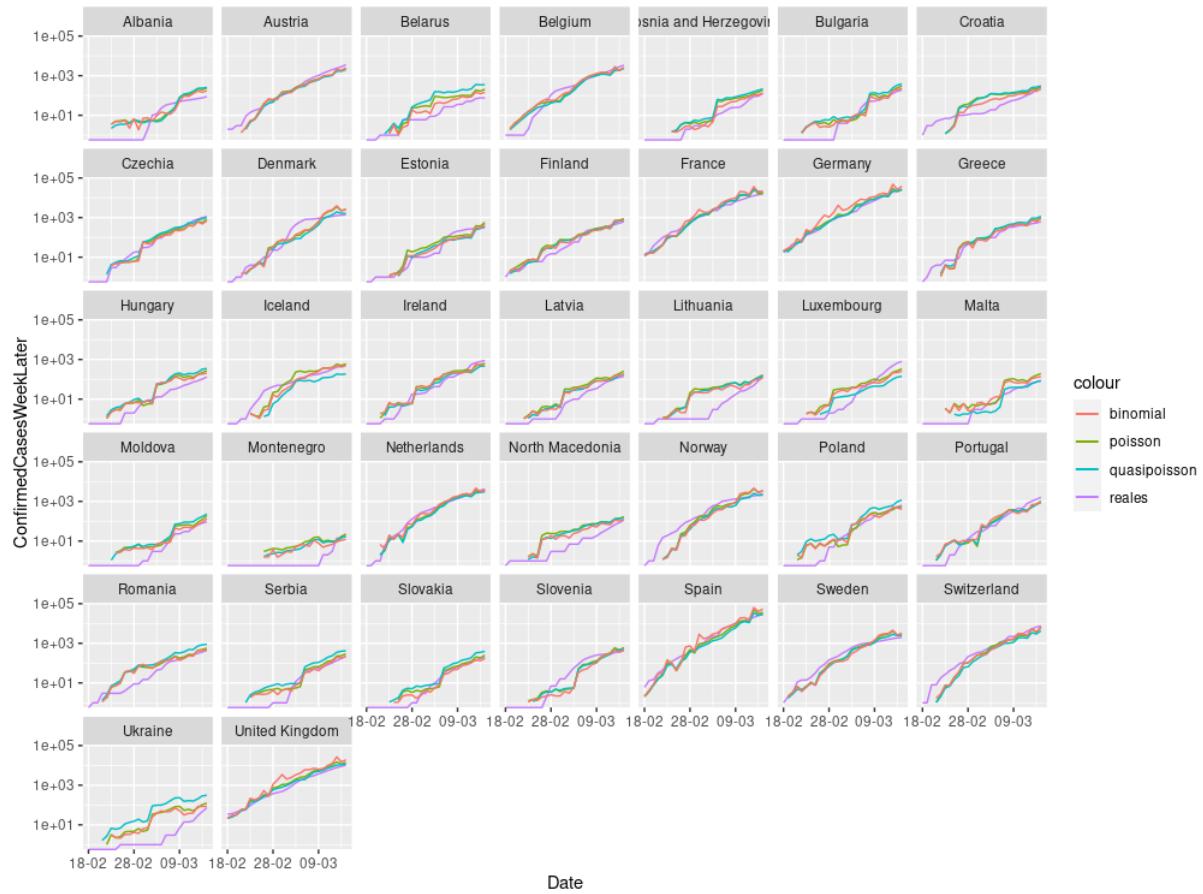


Figura 4.5: Predicciones de los infectados por COVID-19 una semana después según países (escala logarítmica).

$\lambda < 1$, y en general es más bajo conforme nos acercamos a los primeros días, donde solo unos pocos datos eran distintos de 0.

Por otro lado, tenemos que comprobar que las **medias y varianzas** de la variable dependiente sean similares. Para ello, se muestra la evolución del cociente entre la varianza y la media según pasan los días en la Figura 4.7. Como se puede ver, esta hipótesis no solo no se cumple, sino que va a peor conforme avanzan los días (más teniendo en cuenta que la escala es logarítmica). Esto está causando un problema de sobredispersión, que debe ser resuelto mediante modelos quasi-Poisson o de binomial negativa, que serán tratados más adelante.

Por último, se ha tratado la hipótesis de **linealidad** de las variables independientes respecto al logaritmo de la variable dependiente. En la Figura 4.8 se puede ver que no existe una linealidad clara, pero que siguen una función exponencial que podemos tratar de aproximar. En concreto, las funciones que mejor se adaptan son:

- Para la estimación de casos importados, lo mejor es aplicar la raíz cúbica a esta variable.
- Para los casos confirmados, lo mejor es aplicar la raíz quinta a esta variable.

Tras aplicar estas funciones, el resultado queda como se muestra en la Figura 4.9. Se puede ver que realmente se podría aplicar un exponente distinto a cada día, ya

que los casos de febrero no cuadran tanto con estos exponentes como los de marzo. Sin embargo, en favor de un modelo más interpretable y con menos parámetros que estimar, se va a mantener el mismo exponente en todos los casos.

4.2.2. Análisis de las variables a incluir

Como ya se ha mencionado, los casos confirmados son una variable adicional en nuestro estudio, y debe tenerse en consideración si es realmente relevante o no. La Figura 4.10 muestra los resultados del p -valor del test ANOVA, donde un valor menor al 5% (que usaremos como nivel de referencia) significa que tenemos evidencias para confirmar que la variable es relevante.

En general, en todos los casos tenemos que los infectados confirmados son información relevante con un p -valor que es prácticamente 0. Existe un único caso en el que el p -valor no está por debajo del umbral: el 25 de febrero para el modelo quasi-Poisson. Sin embargo, tras eso se vuelve a los p -valores iniciales, por lo que podemos concluir que ese p -valor es un *outlier* dentro de todos los modelos y que la raíz quinta de los casos confirmados sí **es relevante** para nuestras predicciones.

Además, que los p -valores del modelo quasi-Poisson sean más altos no es casualidad. Al inflar la varianza de los estimadores para reducir la sobredispersión, los valores de los test estadísticos han disminuido. Como esta varianza va en el denominador del valor de los test estadísticos, es lógico que los p -valores aumenten. Una vez dicho esto, el hecho de que solo se haya disparado el valor en unos pocos días refuerza la idea de que la nueva variable es relevante.

4.2.3. Análisis de la sobredispersión

La evolución de los parámetros encargados de controlar la sobredispersión en los modelos quasi-Poisson y binomial negativa están respectivamente en la Figura 4.11. En este caso, los valores de quasi-Poisson son mucho más altos, pero no son comparables, ya que cada factor tiene un significado distinto. Sin embargo, un factor mayor que 1 implica en ambas ocasiones un caso de sobredispersión, lo cual entra en consonancia con la Figura 4.7. Inicialmente el factor de sobredispersión es menor que 1, de lo que concluimos que incluso si la varianza es mayor que la media el efecto de sobredispersión no tiene por qué ser mitigado si no es muy pronunciado.

Sin embargo, la clara diferencia es que el factor de dispersión del modelo quasi-Poisson es creciente mientras que el factor de la binomial negativa se estanca en valores cercanos al 10. Esto implica que, en el modelo quasi-Poisson, los efectos de la sobredispersión son cada vez mayores (y por tanto los intervalos del modelo de Poisson son cada vez menos acertados). Sin embargo, el factor de la binomial negativa parece mantenerse constante pese al incremento del cociente varianza-media.

4.2.4. Análisis de los coeficientes de los modelos

En la Figura 4.12 se puede observar la evolución de los coeficientes de cada modelo. Salvando alguna excepción, a nivel cualitativo la evolución de los coeficientes en los tres modelos es muy similar y puede analizarse conjuntamente. Sin embargo, hay un caso que sí que hay que tratar aparte: el intervalo de confianza en el modelo de

Resultados y discusión

Poisson acaba siendo prácticamente 0 en marzo. Esto es debido al problema de sobre-dispersión, que provoca que los intervalos se vuelvan artificialmente más estrechos. En el fondo, existe sobredispersión todo el tiempo, pero como ya se vio en la Figura 4.7 se escala a cocientes varianza-media 100 veces mayores.

Por otro lado, la constante β_0 (llamada *intercept*), evoluciona de valores negativos en febrero a valores positivos a finales de febrero y principios de marzo, para acabar finalmente cerca del 0. Esto marca cómo va evolucionando la pandemia: en febrero, con cero casos importados y cero casos confirmados tenemos $\log \lambda = \beta_0 < 0$ (siendo λ el promedio de la variable dependiente), por lo que se espera que los casos confirmados una semana después sean muy bajos. En cambio, si a la mitad de la evolución empieza a crecer β_0 significa que, con cero casos importados y cero casos confirmados esperamos un gran número de casos confirmados una semana después. Esto muestra una tendencia de cómo la pandemia va creciendo conforme avanza el tiempo.

Respecto a los infectados importados, los coeficientes positivos y altos indican que esta variable tiene una gran relevancia en nuestro modelo y que, como era de esperar, influye positivamente. Su pico se alcanza a finales de febrero y principios de marzo, lo que indica que es aquí cuando más influye el tráfico aéreo en la expansión de la pandemia. Sin embargo, a partir de la segunda semana de marzo los casos confirmados del país empiezan a tener más importancia, indicando que ahora el problema de la pandemia no viene de fuera, sino que está dentro del propio país.

Por otro lado, los intervalos de confianza más anchos de la variable de infectados importados se alcanzan a finales de febrero, indicando que, aunque el valor del parámetro sea alto, el modelo no es capaz de dar el valor exacto, por lo que la predicción también será susceptible de estar en un intervalo más amplio.

Respecto al coeficiente de los casos confirmados, se mantiene más o menos constante durante todo el tiempo, aunque con mayores valores al principio. Esto posiblemente se deba a que, sin ningún tipo de medida sanitaria y con casi todos los países con cero casos, aquellos que tengan infectados dentro experimentarán una subida mucho más explosiva que el resto.

Respecto a los **p-valores** de los coeficientes, la evolución se adjunta en la Figura 4.13. Exceptuando algunos picos, lo que más destaca es que el modelo quasi-Poisson calcule *p*-valores tan altos para los infectados importados. Sin embargo, como ya se ha explicado anteriormente, el hecho de que se inflen las varianzas de los estimadores provoca este efecto.

4.2.5. Análisis de residuos

Se ha mostrado la evolución de los residuos en cada uno de los modelos mediante tres gráficos de dispersión en la Figura 4.14. Por lo general, la varianza de los residuos es constante y su media es cercana a cero en todos los modelos. Sin embargo, solo existen *outliers* positivos, por lo que existe una asimetría con los residuos en este sentido. En este caso, los residuos no son la observación menos la predicción, sino que son los residuos de Pearson, definidos como se mostraba en la Ecuación 2.6. Aún así, que los *outliers* sean siempre positivos significa que, cuando los modelos realizan malas predicciones, es porque se subestimó la cantidad de positivos que iba a haber en una semana.

4.2.6. Bondad de ajuste

Las dos gráficas de bondad de ajuste se adjuntan en la Figura 4.15. En la primera, se utiliza como hipótesis nula el modelo ajustado y como alternativa el modelo saturado, que tiene tantos parámetros como observaciones y que ajusta los datos a la perfección. Como vemos, los primeros días obtenemos p -valores muy altos, lo que indica que nuestro modelo es mejor. Sin embargo, el resto de días se rechaza la hipótesis nula. Este problema se debe generalmente a que hay variables que no se han tenido en cuenta, que los *outliers* están influyendo demasiado o que nuestro modelo no tiene suficiente flexibilidad para abarcar correctamente este problema. Sin embargo, esto no significa directamente que nuestros modelos sean malos, ya que cabe la posibilidad de que alcanzar un modelo mejor que el saturado sea una directriz muy exigente en este caso.

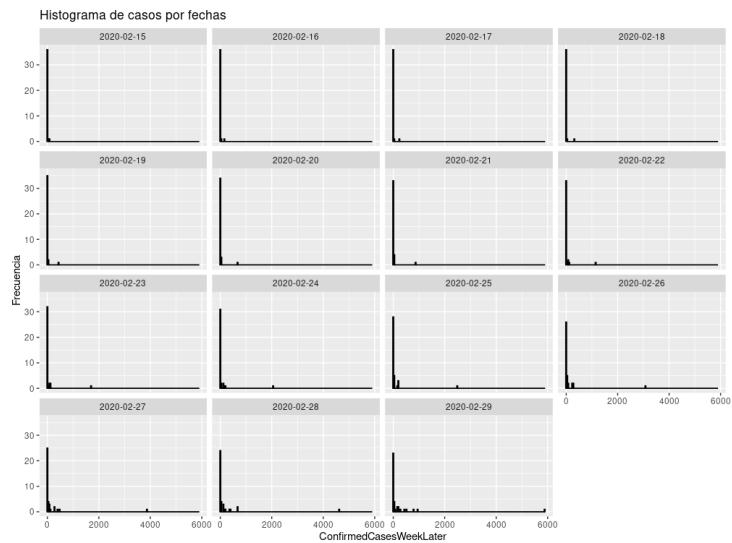
Respecto a la segunda gráfica de bondad de ajuste, la hipótesis nula contrastada contiene al modelo nulo (predicción constante) frente a la hipótesis alternativa en la que se encuentra el modelo ajustado. Aquí, en todo momento los valores están en valores muy cercanos a 0, indicando de esta manera que existe una gran evidencia de que introducir los infectados importados y los infectados del país ayudan a los modelos a predecir mejor los resultados.

4.2.7. Métricas de evaluación

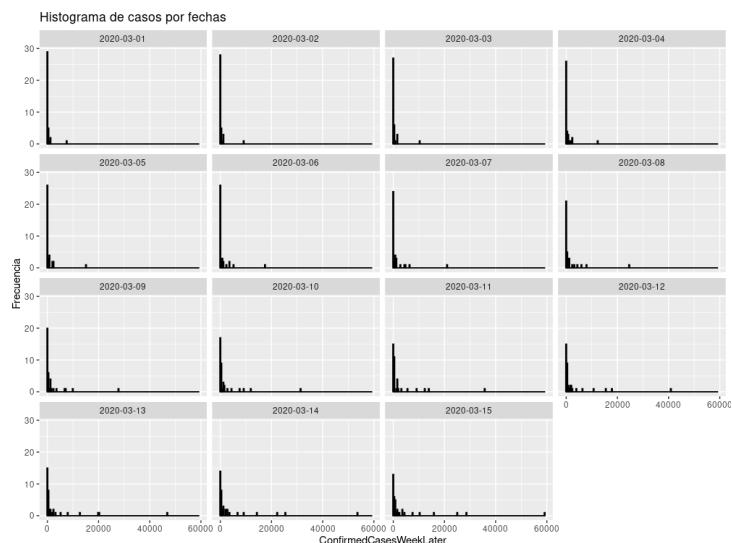
Las gráficas de la evolución del RMSE y el SMAPE se muestran respectivamente en la Figura 4.16. No se muestra el caso del riesgo importado porque ambas gráficas se superponen, es decir, los resultados son prácticamente iguales siempre. En cambio, se muestra la diferencia entre el error cometido usando los infectados importados y el error cometido usando el riesgo importado en la Figura 4.17. Con esta gráfica se puede ver que **el riesgo importado no beneficia a las predicciones**, ya que para ello la diferencia de errores tendría que ser mayoritariamente positiva. En cambio, puede verse que los errores más graves de la predicción se cometen cuando se usa el riesgo importado. Sin embargo, por lo general no parece que ninguna de las dos variables beneficie o perjudique a las predicciones más que en algún caso puntual. Como ya se comentó en la Sección 4.1, ambas variables son prácticamente iguales y no se esperan grandes diferencias a la hora de hacer regresión.

Si miramos el RMSE, comprobaremos que en los tres modelos va aumentando el error con el tiempo, debido al cambio de escala en las predicciones (en febrero, ningún país tiene más de 10 positivos, mientras que en marzo se alcanzan los cientos en muchos países). En general, se puede observar mayor inestabilidad en el modelo binomial negativo, donde además el RMSE es superior casi siempre. El modelo Poisson y quasi-Poisson son algo más parecidos, aunque parece más estable el segundo.

Respecto a los valores relativos que nos ofrece la métrica SMAPE, parece que porcentualmente mejoran las predicciones, por lo que en este sentido nuestro modelo funciona mejor en marzo. Además, el modelo quasi-Poisson pasa de dar las mejores predicciones en marzo a las peores, aunque en este caso la diferencia entre los tres modelos es mucho menor (nunca supera el 0.2 %).



(a) En febrero.



(b) En marzo.

Figura 4.6: Histograma de los casos confirmados una semana después por fecha.

4.2. Análisis de las predicciones

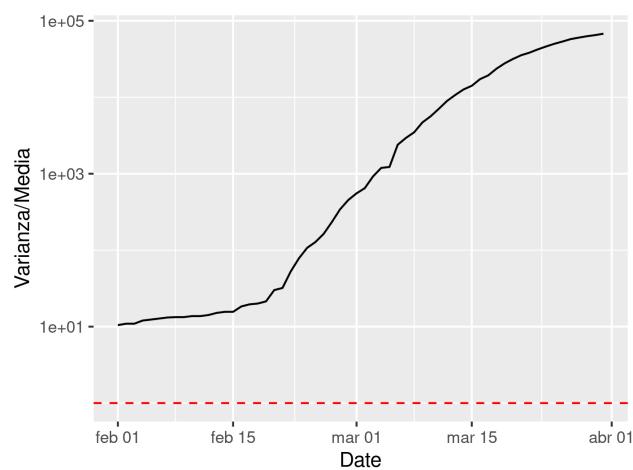
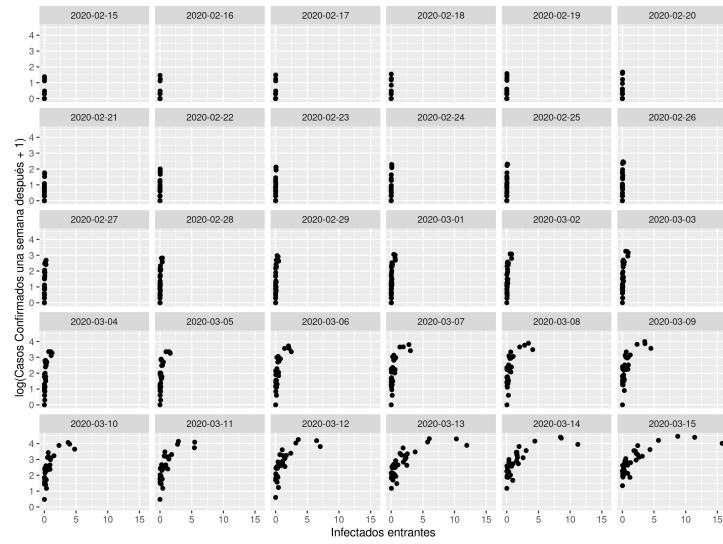
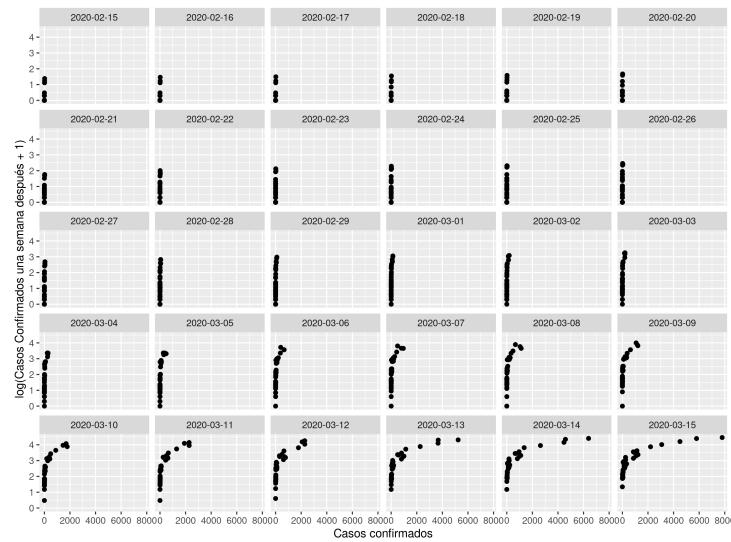


Figura 4.7: Evolución del cociente entre varianza y media de los casos confirmados una semana después (escala logarítmica).

Resultados y discusión



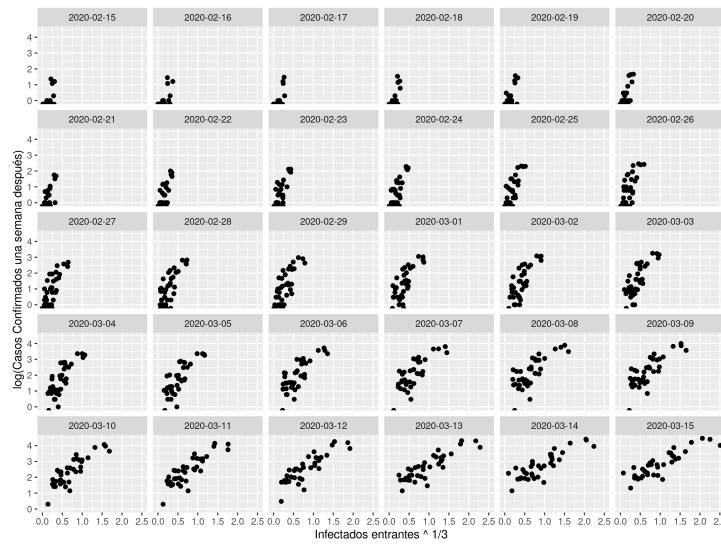
(a) Frente a la estimación de casos importados.



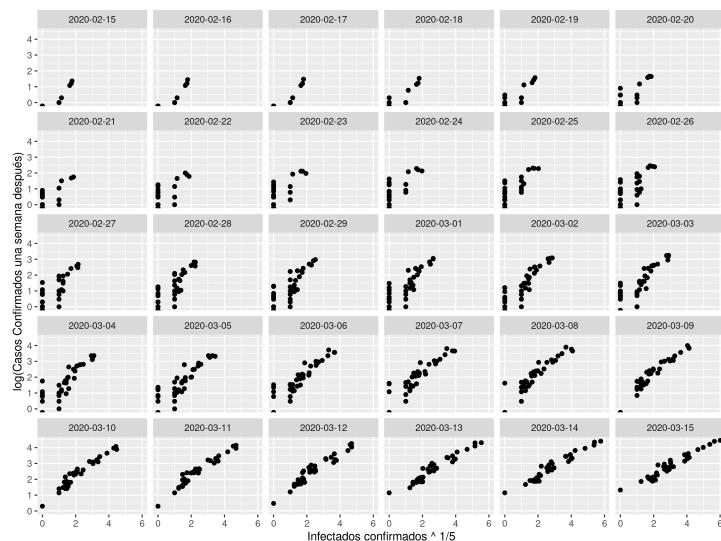
(b) Frente a los casos confirmados.

Figura 4.8: Logaritmo de la variable dependiente frente a las variables independientes.

4.2. Análisis de las predicciones



(a) Frente a la raíz cúbica de la estimación de casos importados.



(b) Frente a la raíz quinta de los casos confirmados.

Figura 4.9: Logaritmo de la variable dependiente frente a las variables independientes transformadas.

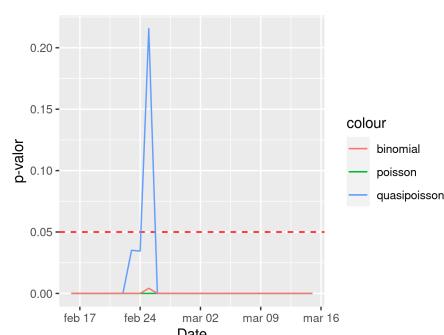


Figura 4.10: Evolución del test ANOVA con los tres modelos.

Resultados y discusión

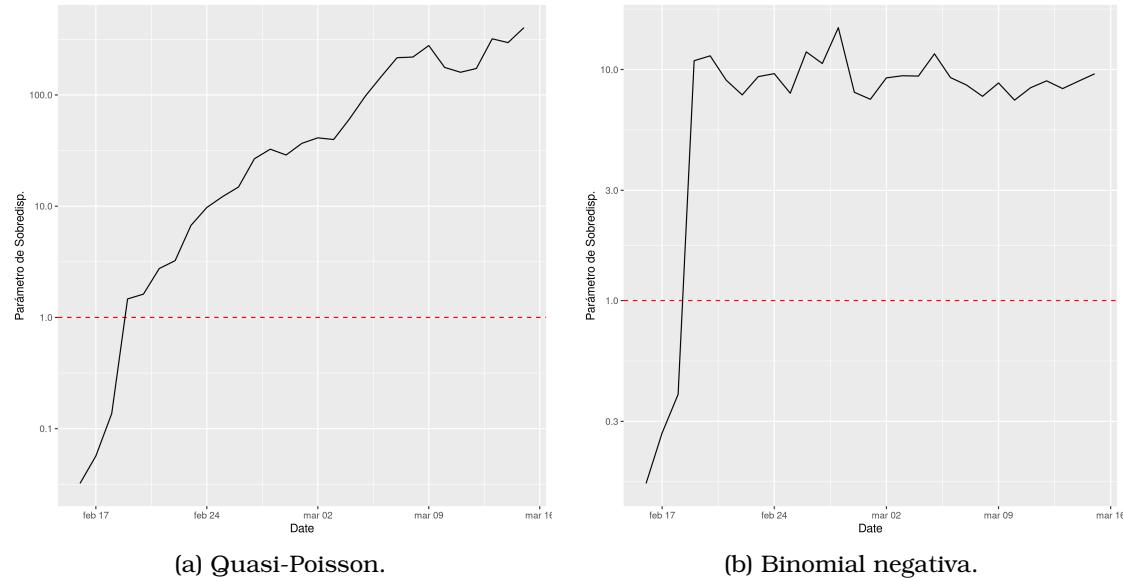


Figura 4.11: Evolución del factor de sobredispersión en los modelos que existe.

4.2. Análisis de las predicciones

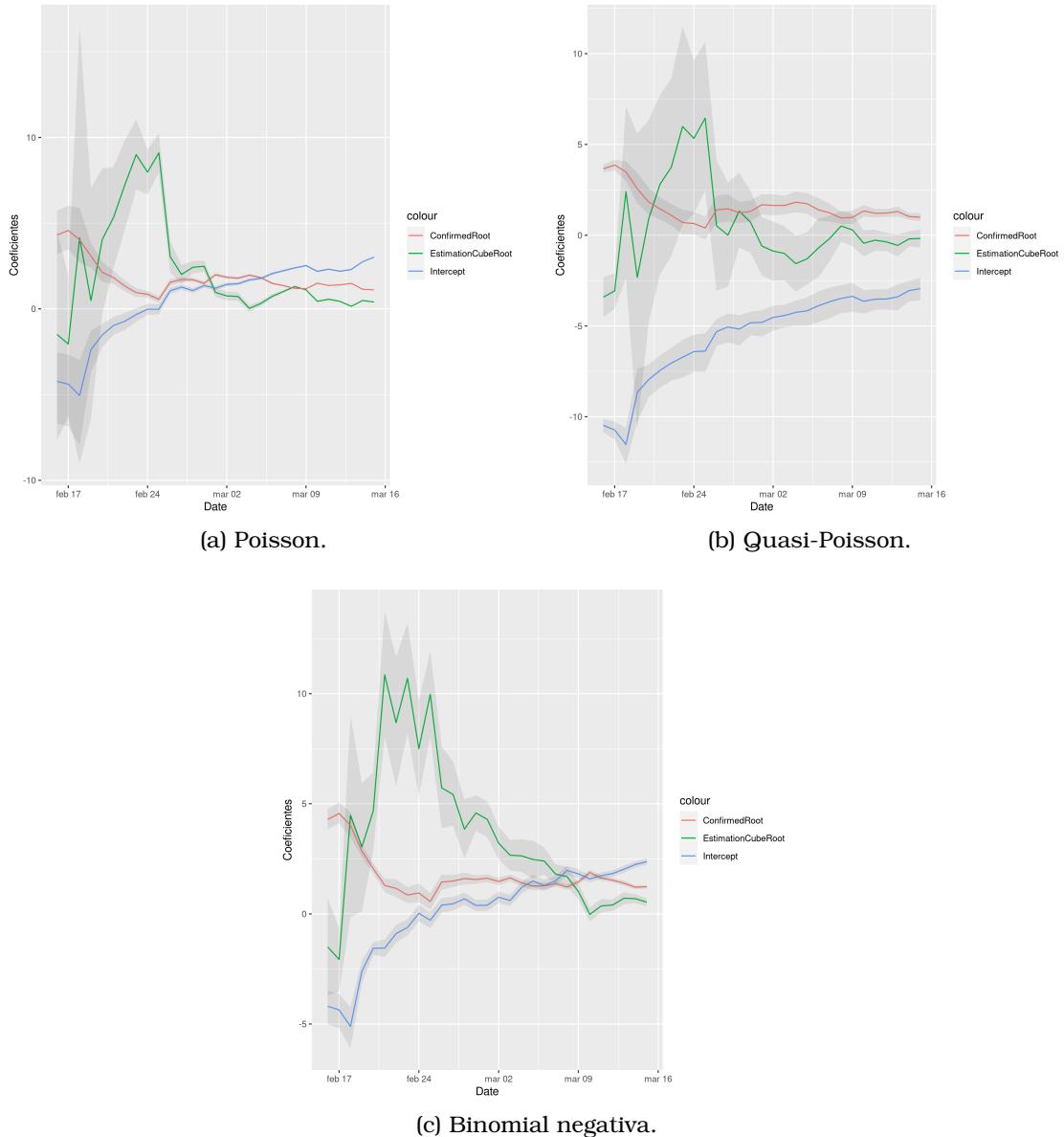


Figura 4.12: Evolución de los coeficientes de los modelos con sus intervalos de confianza. *ConfirmedRoot* corresponde a la raíz quinta de los casos confirmados y *EstimationCubeRoot* a la raíz cúbica de los casos importados.

Resultados y discusión

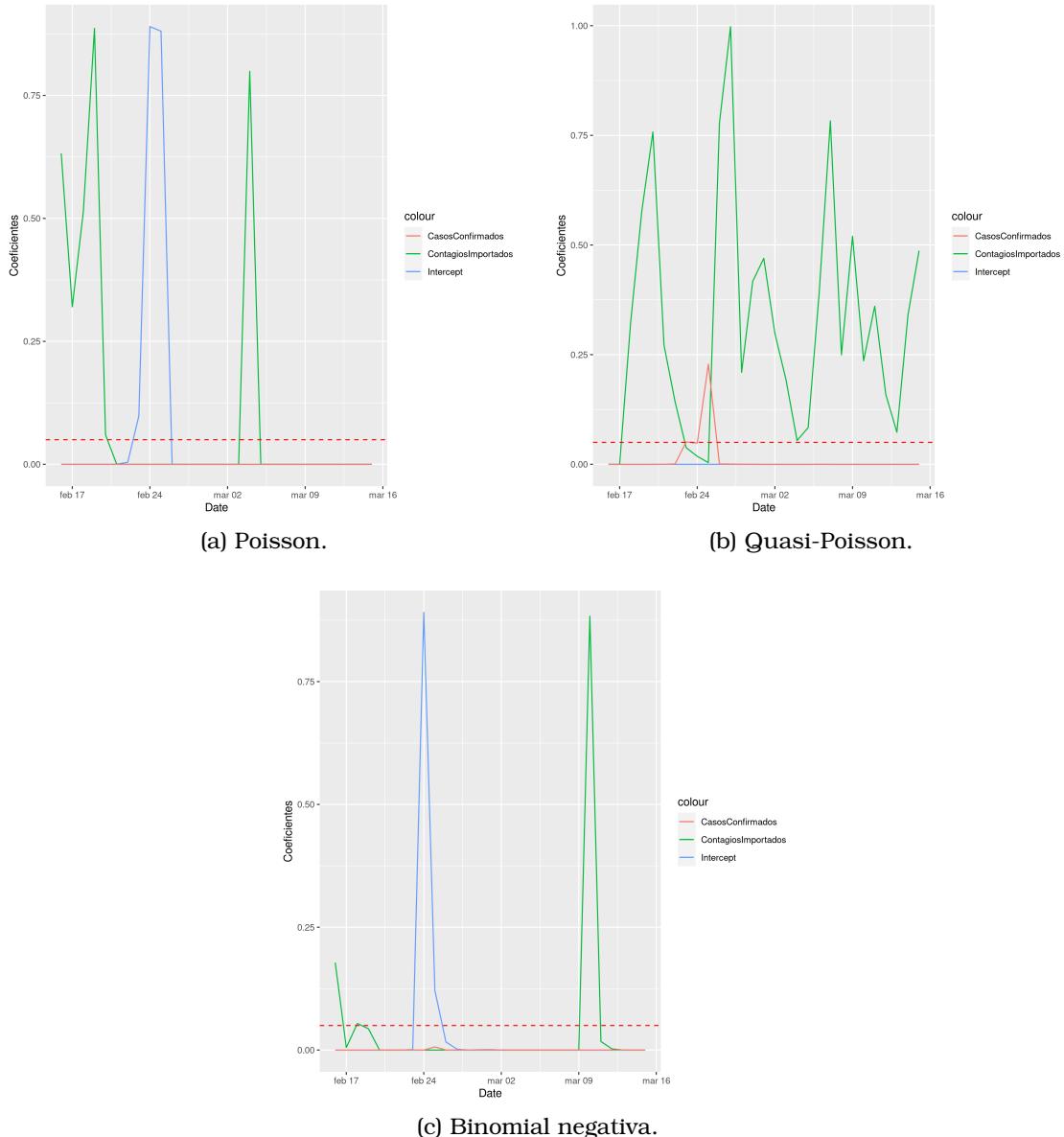


Figura 4.13: Evolución de los p -valores asociados a los coeficientes en los modelos.

4.2. Análisis de las predicciones

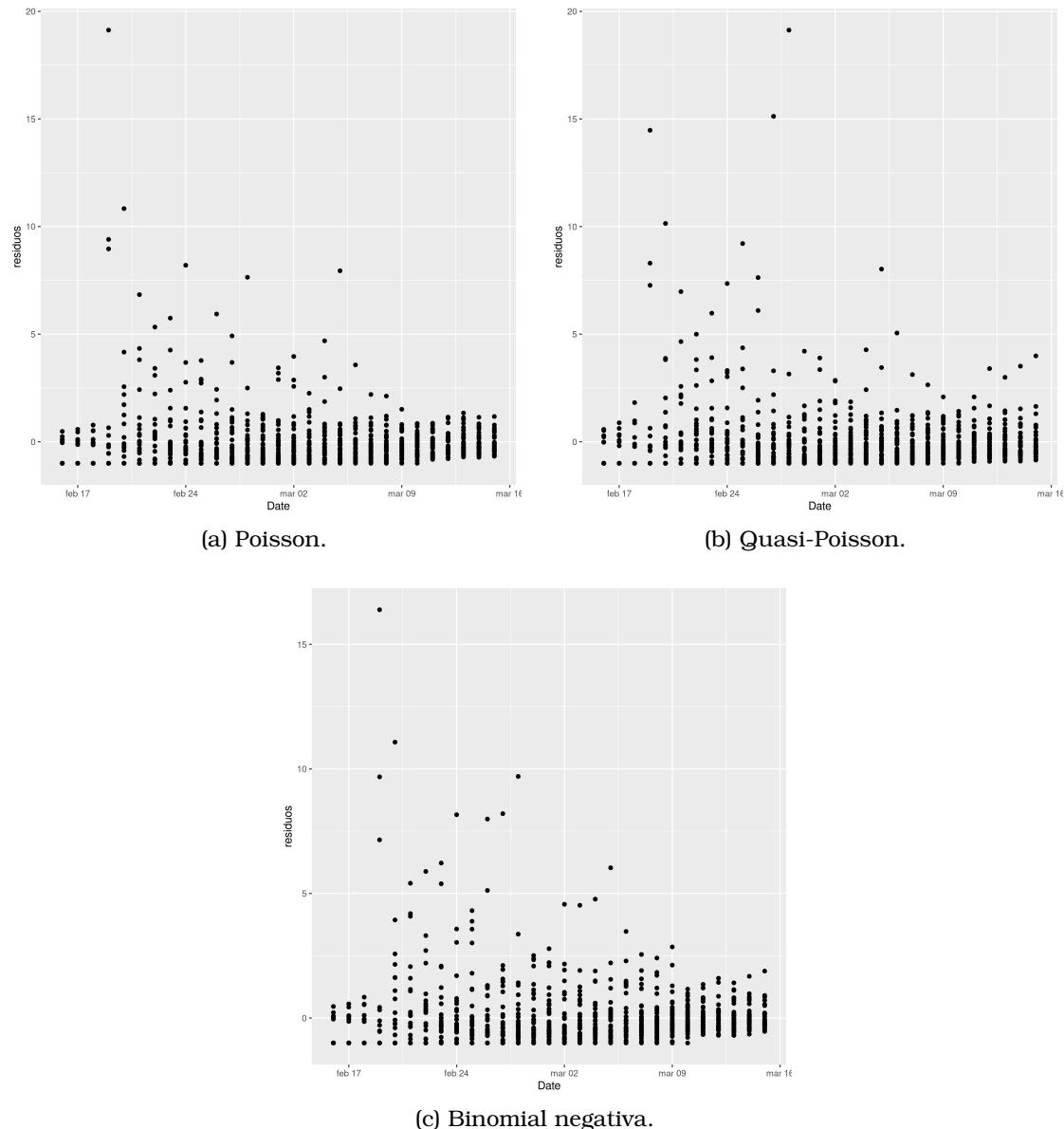


Figura 4.14: Diagrama de dispersión de la evolución de los residuos en los modelos.

Resultados y discusión

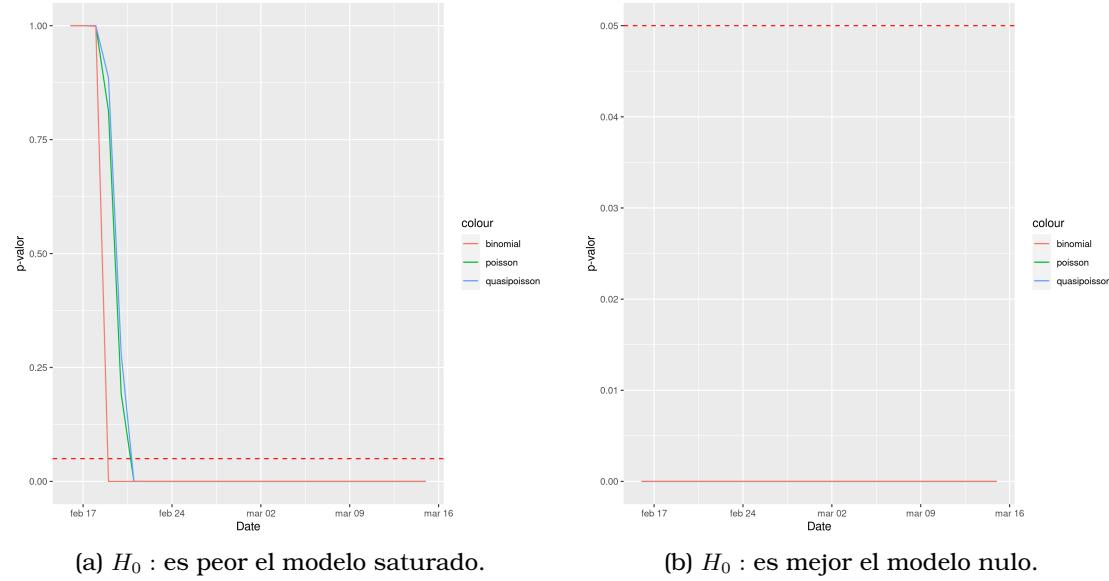


Figura 4.15: Evolución de la bondad de ajuste según diversas hipótesis nulas.

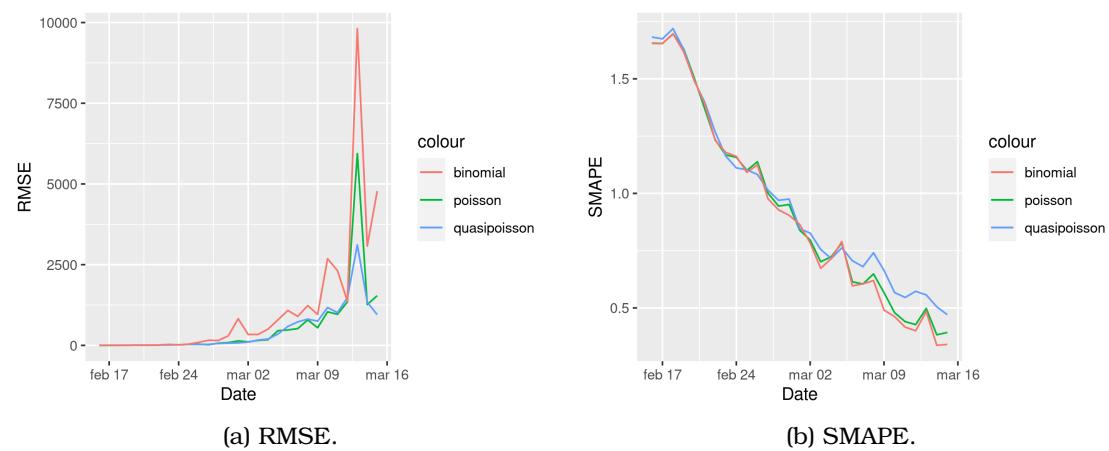
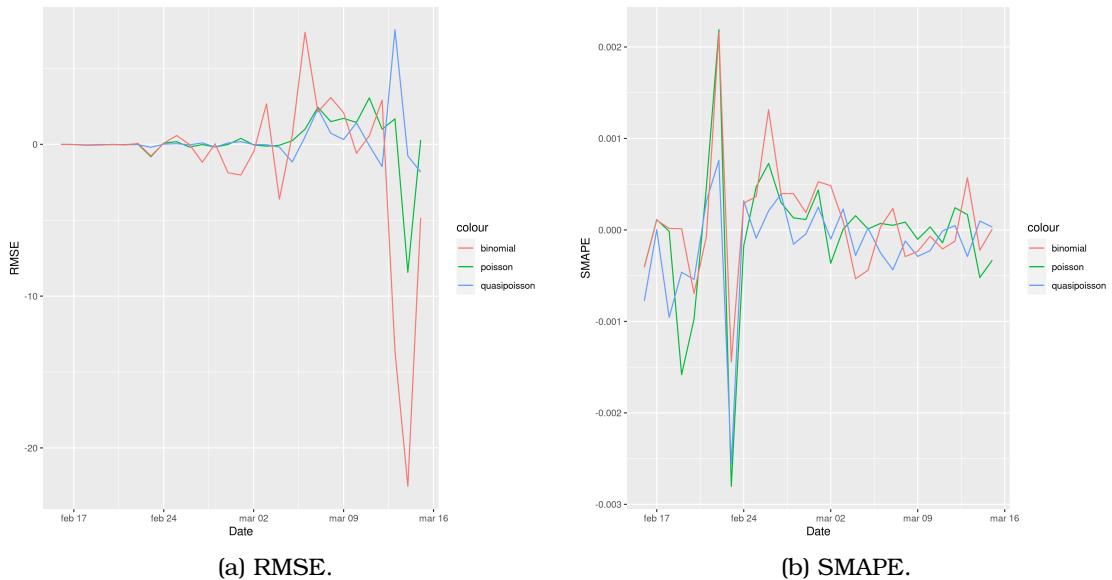


Figura 4.16: Evolución del error de los modelos con la variable de contagios importados según las métricas RMSE y SMAPE.



(a) RMSE.

(b) SMAPE.

Figura 4.17: Evolución de la diferencia entre el error prediciendo con infectados estimados y el error prediciendo con el riesgo importado según las métricas RMSE y SMAPE.

Capítulo 5

Conclusiones

En este trabajo se han utilizado modelos de regresión de Poisson para obtener modelos de predicción de contagios de COVID-19 a partir del tráfico aéreo. Esta clase de modelos son adecuados para esta situación, ya que se predicen conteos y las hipótesis básicas de estos modelos se cumplen en su mayoría.

Para aumentar la fiabilidad de los datos, hemos tomado solo países europeos. Además, el tráfico aéreo solo era relevante al comienzo de la pandemia, por lo que se han predicho los contagios con los datos que comprenden desde el 15 de febrero hasta el 15 de marzo.

Se ha utilizado una base de datos basada en grafos con todos los vuelos del año. Aunque de gran tamaño y por lo tanto complicada de manejar, su flexibilidad y riqueza han permitido obtener una estimación muy precisa sobre el número de infectados que entraban en cada país.

Se han tratado tres modelos de regresión: el de Poisson, el quasi-Poisson y la binomial negativa. El segundo y tercer modelo se pueden considerar generalizaciones del primero, ya que han sido diseñados para ser utilizados cuando falla una de las hipótesis básicas (la media debe ser igual a la varianza) y se genera sobredispersión en la variable dependiente.

La variable dependiente a predecir ha sido el número de casos de COVID-19 por país tras 7 días. La variable regresora principal ha sido la estimación de casos importados sin tener en cuenta el tiempo de vuelo. Alternativamente, se ha usado el riesgo importado, que aplica modelo SIR para tener en cuenta el tiempo de los vuelos y crear una estimación mucho más exacta. Sin embargo, no es suficiente para arrojar mejores resultados, ya que no hay una diferencia sustancial entre ambas variables. Junto a esta variable, se ha añadido también el número de contagios del día actual en dicho país, ya que se ha demostrado que mejora las predicciones.

Aunque el error bruto (RMSE) aumenta conforme avanza el tiempo (y por tanto hay más contagios), la métrica relacionada con el error relativo (SMAPE) ha mostrado que el modelo gana precisión conforme avanza el tiempo. Esto demuestra que el modelo creado va ganando exactitud conforme la infección se extiende. Sin embargo, la influencia del tráfico aéreo se reduce según nos aproximamos a marzo, mientras que los casos confirmados van ganando fuerza como variable explicativa.

Además, aunque la precisión sea muy buena, los intervalos de confianza relacionados

con los coeficientes de los contagios importados son muy anchos. Esto es negativo, ya que significa que el modelo no puede adoptar un coeficiente preciso con suficiente certeza, generando incertidumbre en las predicciones.

Sin embargo, este defecto no tiene por qué significar que el tráfico aéreo no es una buena variable regresora. Podría deberse a la falta de datos, ya que estos modelos solo se entranan con 37 datos (correspondientes a los 37 países europeos que se gestionan). En general, se puede considerar bueno el modelo y capaz de predecir con precisión los contagios tras una semana.

5.1. Trabajo futuro

Las principales áreas por las que se puede continuar trabajando si se parte de este proyecto son:

- Ampliación de nuestra base de datos gracias a la *Johns Hopkins*[9]. Una de las principales limitaciones de nuestro modelo ha sido la pequeña cantidad de datos con la que se ha trabajado al restringirnos a países europeos. Sin embargo, existen datos de regiones europeas dentro de cada país que han sido recopiladas en la base de datos *Johns Hopkins*. La introducción de estos datos permitiría aumentar rápidamente el número de ejemplos por día, dando al modelo mayor riqueza y mejorando posiblemente los resultados.
- Como se ha visto, el riesgo importado no aporta una mejora sustancial a nuestro modelo porque es muy similar a los contagios importados. Queda como línea de investigación ampliar el rango de países (o regiones) que se tratan, de forma que los vuelos aumenten su duración y el modelo SIR sea más relevante.
- Usar modelos con mayor flexibilidad. En concreto, en [8] se usan técnicas de *Deep Learning* para predecir el número de contagios. Este tipo de técnicas permiten añadir una cantidad mucho mayor de variables a nuestro modelo. Por ejemplo, no tendríamos que conformarnos con mostrar únicamente los infectados entrantes a nuestro país, sino que podríamos mostrar los infectados desde cada aeropuerto junto con la distancia a dicho aeropuerto. Sin embargo, esto último es solo una idea, ya que la gama de posibilidades se abre mucho con un algoritmo tan flexible.
- También se han utilizado modelos *ARIMA* y *Seasonal ARIMA* para la predicción de la COVID-19 en [7]. Estos modelos son muy transparentes y se puede ver en profundidad la influencia de distintos factores. En general, sí admiten variables explicativas en general, como es el caso de *ARIMAX*. Podrían ser muy interesantes para analizar los datos con los que hemos trabajado.
- Frente a las hipótesis que hemos tenido que contrastar con nuestro modelo, hay otros en los que estas hipótesis son más flexibles. Por ejemplo, los modelos de regresión con procesos Gaussianos han sido utilizados también [1] para la predicción de la COVID-19 con buenos resultados en India y Brasil. En este caso, se han utilizado datos sobre el índice de vacunación, la edad promedio y las condiciones meteorológicas para explicar el número de contagios. Sin embargo, no se ha utilizado la influencia del tráfico aéreo en fases tempranas de la pandemia, por lo que utilizar este modelo para la predicción de contagios podría ser interesante.

Bibliografía

- [1] Y. Alali, F. Harrou, and Y. Sun. A proficient approach to forecast covid-19 spread via optimized dynamic machine learning models. *Scientific Reports*, 12(1), 2022.
- [2] A. Barnett. Covid-19 risk among airline passengers: Should the middle seat stay empty? 2020.
- [3] A. Bilinski, R. Birger, S. Burn, M. Chitwood, E. Clarke-Deelder, T. Copple, J. Eaton, H. Ehrlich, M. Erlendsdottir, S. Eshghi, and et al. Defining high-value information for covid-19 decision-making. 2020.
- [4] I. I. Bogoch, O. J. Brady, M. U. G. Kraemer, M. German, M. I. Creatore, M. A. Kulkarni, J. S. Brownstein, S. R. Mekaru, S. I. Hay, E. Groot, A. Watts, and K. Khan. Anticipating the international spread of zika virus from brazil. *The Lancet*, 387(10016):335–336, 2016.
- [5] I. I. Bogoch, M. I. Creatore, M. S. Cetron, J. S. Brownstein, N. Pesik, J. Miniota, T. Tam, W. Hu, A. Nicolucci, S. Ahmed, J. W. Yoon, I. Berry, S. I. Hay, A. Aneema, A. J. Tatem, D. MacFadden, M. German, and K. Khan. 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, 2015.
- [6] V. Capasso. *Mathematical structures of Epidemic Systems*. Springer, 2008.
- [7] T. Dehesh, H. Mardani-Fard, and P. Dehesh. Forecasting of covid-19 confirmed cases in different countries with arima models. *medRxiv*, 2020.
- [8] J. Devaraj, R. Madurai Elavarasan, R. Pugazhendhi, G. Shafiullah, S. Ganesan, A. K. Jeysree, I. A. Khan, and E. Hossain. Forecasting of covid-19 cases using deep learning models: Is it reliable and practically significant? *Results in Physics*, 21:103817, 2021.
- [9] E. Dong, H. Du, and L. Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet Infectious Diseases*, 20(5):533–534, 2020.
- [10] F. P. Havers, C. Reed, T. Lim, J. M. Montgomery, J. D. Klena, A. J. Hall, A. M. Fry, D. L. Cannon, C.-F. Chiang, A. Gibbons, I. Krapiunaya, M. Morales-Betouille, K. Roguski, M. A. U. Rasheed, B. Freeman, S. Lester, L. Mills, D. S. Carroll, S. M. Owen, J. A. Johnson, V. Semenova, C. Blackmore, D. Blog, S. J. Chai, A. Dunn, J. Hand, S. Jain, S. Lindquist, R. Lynfield, S. Pritchard, T. Sokol, L. Sosa, G. Turabelidze, S. M. Watkins, J. Wiesman, R. W. Williams, S. Yendell, J. Schiffer, and N. J. Thornburg. Seroprevalence of Antibodies to SARS-CoV-2 in 10 Sites in the United States, March 23-May 12, 2020. *JAMA Internal Medicine*, 180(12):1576–1586, 12 2020.

- [11] W. He, G. Y. Yi, and Y. Zhu. Estimation of the basic reproduction number, average incubation time, asymptomatic infection rate, and case fatality rate for covid-19: Meta-analysis and sensitivity analysis. *Journal of Medical Virology*, 92(11):2543–2550, 2020.
- [12] M. J. Kane, N. Price, M. Scotch, and P. Rabinowitz. Comparison of arima and random forest time series models for prediction of avian influenza h5n1 outbreaks. *BMC Bioinformatics*, 15(1), 2014.
- [13] P. R. Legler and Julie. Beyond multiple linear regression, Jan 2021.
- [14] J. G. Moreno, J. Poveda, Óscar Villasante, P. Sánchez-Escalonilla, A. M. Caballero, E. V. Cestero, and R. Lorenzo-Redondo. On-line platform for the short-term prediction of risk of expansion of epidemics. *Fourteenth USA/Europe Air Traffic Management Research and Development Seminar*, 54, 2021.
- [15] Y. M. Sokadjo and M. N. Atchadé. The influence of passenger air traffic on the spread of covid-19 in the world. *Transportation Research Interdisciplinary Perspectives*, 8, 2020.