TESIS
MAESTRÍA EN CIENCIAS FÍSICAS

MODELOS EPIDEMIOLÓGICOS DE COVID-19: SIMULACIONES COMPUTACIONALES, ESTADÍSTICA BAYESIANA Y TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

Orientación: Sistemas Complejos

Lic. Denise S. Cammarota

Dra. Karina F. Laneri
Directora

Lic. Denise S. Cammarota
Maestranda

Instituto Balseiro
Comisión Nacional de Energía Atómica
Universidad Nacional de Cuyo
S.C. de Bariloche, Febrero de 2022

Inventario 24520
08/03/2022

A mis xadres

Índice de símbolos

AIC: Criterio de Información de Akaike
BIC: Criterio de Información de Bayes
CABA: Ciudad Autónoma de Buenos Aires
COVID-19: Coronavirus Disease 2019
OMS: Organización Mundial de la Salud
RN: Red Neuronal
SARS-CoV-2: Severe Acute Respiratory Syndrome Coronavirus 2
SIR: Susceptible-Infectado-Recuperado (Susceptible-Infected-Recovered)
UCI: Unidad de Cuidados Intensivos

Índice de contenidos

Índice de figuras

Índice de tablas

Resumen

En este trabajo, se estudiaron características de la propagación y predicción de casos del virus SARS-COV-2 en la República Argentina, utilizando datos de casos confirmados en el país durante el año 2020 provistos por el Ministerio de Salud. Para ello, se introdujeron herramientas previamente desarrolladas en la bibliografía, como el número reproductivo empírico y los diagramas de riesgo. En particular, se observó y analizó la tendencia del número reproductivo empírico a tomar valores en torno a la unidad. Por otra parte, se optimizó el cálculo de un número reproductivo empírico a partir del diseño y entrenamiento de redes neuronales artificiales.

Adicionalmente, se estudió la posibilidad de reproducir características de la propagación del SARS-COV-2 utilizando modelos del tipo susceptible-infectado-recuperado (SIR) en su versión de campo medio. Se propusieron nuevos modelos modificados para contemplar la reacción social a la cantidad de casos. Esto permitió entender cualitativamente la evolución temporal de casos en algunas localidades en estudio, aunque no totalmente.

Por ello, se estudió la propagación interprovincial de la enfermedad introduciendo el uso de las correlaciones con lag y se estudiaron varias propiedades de estas cantidades. Entre ellas, se encontró que las provincias como Buenos Aires y CABA, que lideran la dinámica epidémica, se caracterizan por tener correlaciones bajas y lags absolutos grandes. Adicionalmente, las provincias del Noroeste también se destacan como jurisdicciones con tendencia a lideran el brote epidémico. Finalmente, para explicar estas observaciones, se proponen diferentes modelos tipo SIR metapoblacionales de subpoblaciones conectadas por una matriz de conectividad. Se exploran estos modelos para asociarlos a las observaciones reales y se realizan para todos ellos los ajustes correspondientes. El modelo que mejor se ajusta a los datos corresponde a un modelo con tasas de contagio diferentes para cada una de las provincias, conectadas entre ellas por una matriz dependiente tanto de sus poblaciones conjuntas como de la distancia entre ellas.

Palabras clave: COVID-19, EPIDEMIOLOGÍA, MODELO METAPOBLACIONAL, DIAGRAMA DE RIESGO, NÚMERO REPRODUCTIVO, CORRELACIONES CON LAG

Abstract

In this work we studied characteristics of the propagation and forecasting of SARS-COV-2 cases in Argentina, using national data from confirmed cases during 2020 as provided per the Ministry of Health. For that purpose, we introduced tools which were previously developed in the literature, such as the empirical reproductive number and risk diagrams. In particular, we observed and analyzed the tendency of the reproductive number to take values in the vicinity of the unit. Furthermore, we propose a way to optimize the calculus of the empirical reproductive number by means of designing and training artificial neural networks.

Aditionally, we studied the possibility of reproducing characteristics of SARS-COV-2 propagation using mean field susceptible-infected-removed (SIR) models. New models were proposed to take into account social reaction to the amount of confirmed cases. These models allowed us to cualitatively understand the temporal evolution of cases in certain regions, though not completely.

Therefore, we began a new analysis that takes into account the spatial dimension of the problem at hand. To do that, we introduced the concept of lagged correlations and we observed several interesting properties of these quantities. Among them, we found that provinces leading the epidemiological dynamics have lower correlations and higher absolute lags, such as Buenos Aires and CABA. Aditionally, provinces in the North-west region also show a tendency to lead the epidemic outbreak. Finally, to explain these observations, several metapopulation SIR models are proposed. These consist of provinces being connected by a connectivity matrix. We explored all presented models, as well as fitting them to real-world data. The best fit ends up corresponding to a model with inhomogeneus infection rates and with a connectivity matrix dependent on the provinces' joint population as well as on distance between them.

Keywords: COVID-19, EPIDEMIOLOGY, METAPOPULATION, RISK DIAGRAM, REPRODUCTIVE NUMBER, LAGGED CORRELATIONS

Capítulo 1

Introducción

La COVID-19 (Coronavirus Disease 2019) es una enfermedad causada por el virus SARS-CoV-2 (Severe Acure Respiratory Syndrome Coronavirus 2). Este pertenece a la familia de los *coronavirus*, que son conocidos por provocar infecciones respiratorias en animales y en humanos. Además del SARS-CoV-2, se pueden citar como ejemplos de coronavirus que infectan humanos el SARS-CoV (Severe Acute Respiratory Syndrome Coronavirus) y el MERS-CoV (Middle East Respiratory Syndrome Coronavirus), los cuales provocaron epidemias en los años 2002 y 2012 respectivamente. Estos tres virus son de origen zoonótico, y son capaces de provocar enfermedades severas e incluso fatales en los seres humanos [1] [3].

Los síntomas más comunes de la COVID-19 son fiebre, tos seca y fatiga. Otros síntomas posibles incluyen pérdida de los sentidos del olfato y del gusto, congestión nasal, conjuntivitis y dolor de garganta. En casos severos de la enfermedad puede manifestarse falta de aire, pérdida del apetito y dolor en el tórax.

Según la OMS (Organizacion Mundial de la Salud), el virus se propaga de persona a persona a través de pequeñas gotas de la nariz o de la boca, que son emitidas por un individuo infectado (por ejemplo, al hablar, estornudar, toser o exhalar) y otra persona puede inhalar. Adicionalmente, estas partículas pueden caer en superficies u objetos cercanos. Entonces, otras personas contraen la enfermedad tocando estas superficies y luego llevándose las manos a los ojos, nariz o boca [4].

Un sujeto expuesto al virus comienza a manifestar síntomas en promedio de 5-6 días luego de la exposición, aunque puede variar en un rango de 1-14 días. De acuerdo con las evidencias recientes, todos los grupos etarios son susceptibles a contraer la enfermedad. Sin embargo, las manifestaciones clínicas difieren con la edad y el estado de salud de las personas. La mayoría de las personas menores de 50 años se recuperan tras esta etapa inicial. En las personas mayores de 60 años o con comorbilidades, se suelen observar manifestaciones severas como neumonía o falta de aire alrededor de 8 días luego de la exposición al virus. Durante esta etapa, puede ser necesaria la hospitalización de

la persona en la UCI (Unidad de Cuidados Intensivos) y la utilización de ventilación mecánica. Un pequeño porcentaje de pacientes, usualmente mayores de 68 años o con comorbilidades, desarrollan un estado crítico que puede llevarlos a la muerte en un rango de 12-20 días después del primer contacto con la enfermedad. Este estado se caracteriza por síntomas como fallo multiorgánico o SDRA (Síndrome de Dificultad Respiratoria Aguda). Los diferentes síntomas de la COVID-19 y su relación con la edad de los pacientes se ilustra en la Figura 1.1 adaptada de [1].

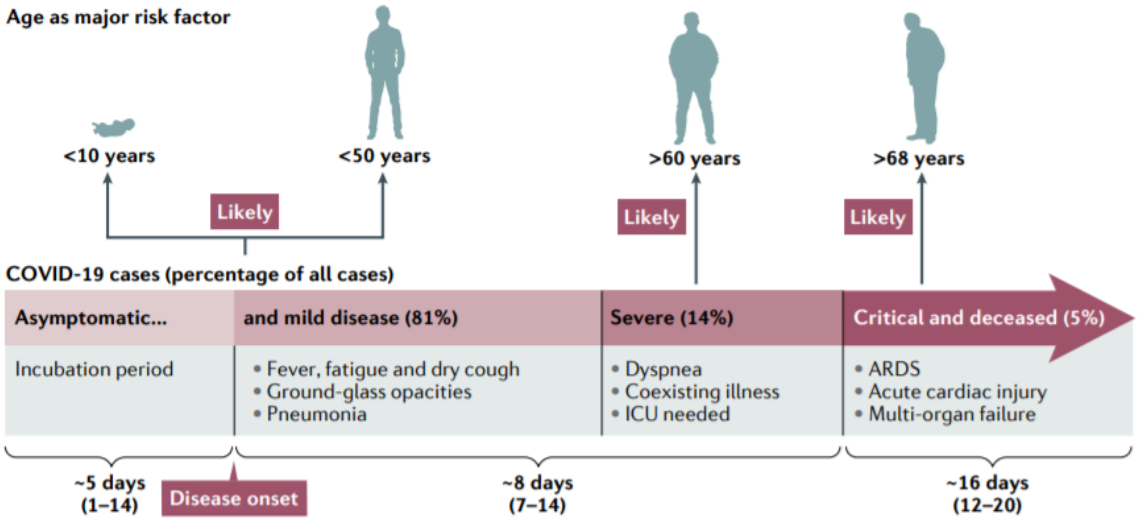


Figura 1.1: Diferentes síntomas de la COVID-19, en función del tiempo desde la exposición a la enfermedad y la edad de los pacientes. Adaptado de [1].

1.1. Breve línea temporal de la propagación.

Los primeros casos de COVID-19 fueron reportados hacia finales de diciembre de 2019 por varios centros de atención médica de la ciudad de Wuhan, provincia de Hubei, China. Estos pacientes presentaban síntomas compatibles con una neumonía cuya causa era entonces desconocida e inicialmente muchos de ellos estaban conectados al mercado *Huanan Seafood Market* en Wuhan. De acuerdo con estudios retrospectivos, los primeros casos podrían remontarse a alrededor del 8-12 de diciembre del 2019.

El día 31 de diciembre de 2019, las autoridades sanitarias de Wuhan reportaron este brote a la OMS. A comienzos del año 2020, el *Huanan Seafood Market* fue cerrado para su sanitización, dada la continuada relación entre nuevos casos de neumonía y este establecimiento. Pronto, nuevos casos que no se conectaban con el lugar aportaron evidencia de una transmisión persona a persona.

El 8 de enero de 2020, se anunció que un nuevo tipo de *coronavirus* era la causa de la nueva enfermedad. Inicialmente, fue llamado 2019-nCov (2019 novel Coronavirus), pero posteriormente OMS cambió su nombre a SARS-CoV-2. Hacia fines de enero de 2020, se detectaron casos de COVID-19 en las 34 provincias de la República Popular

China, tras el aislamiento de la ciudad de Wuhan. Adicionalmente, se reportaron casos importados en países como Tailandia y los Estados Unidos de América. El día 30 de enero la OMS declaró el estado de emergencia sanitaria en todo el mundo. Los casos continuaron en ascenso en China hasta mediados de Febrero, y provocaron estrictas medidas sanitarias de las autoridades chinas, como el aislamiento. Luego, se observó una clara disminución de los casos en ese país, indicando que el brote había sido controlado.

Sin embargo, la propagación internacional se aceleró, resultando en un ascenso rápido de los casos en Europa y en los Estados Unidos, a pesar de las medidas tomadas por las autoridades sanitarias de todo el mundo. El día 11 de Marzo la OMS declaró a la COVID-19 una pandemia mundial y el 13 del mismo mes declaró a Europa como el nuevo epicentro de la pandemia. Hacia el mes de julio, la OMS volvió a redefinir a América, y específicamente a Latinoamérica, como nuevo epicentro de los contagios. Una línea temporal de los acontecimientos se presenta en la Figura 1.2.

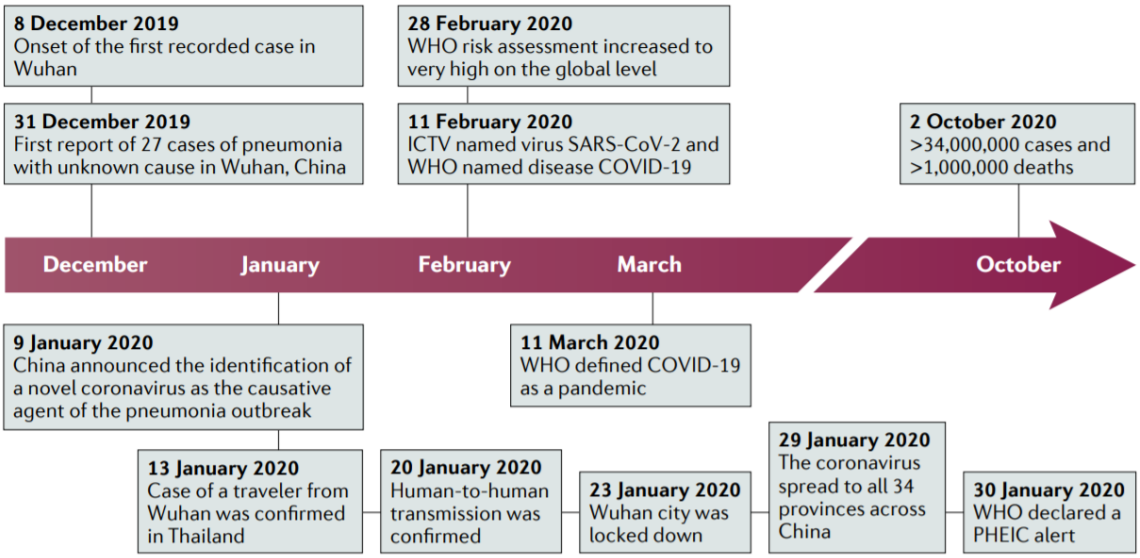
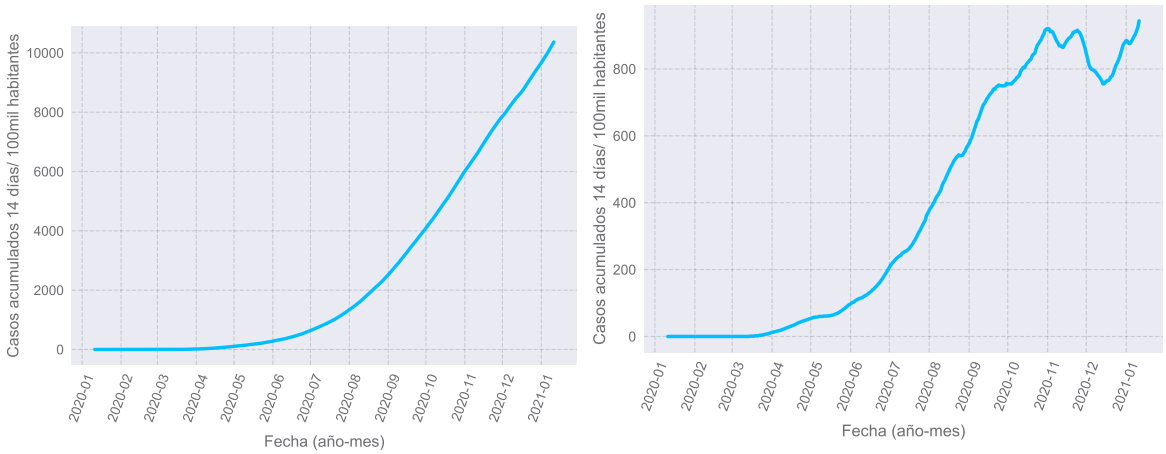


Figura 1.2: Linea de tiempo mundial del origen y de la propagación de la COVID-19, adaptada de [1].

El primer caso registrado en Argentina fue reportado el día 3 de marzo de 2020 como un caso importado de Italia. Le siguieron más casos similares de viajeros que retornaban al país desde diferentes regiones de Europa, y de contactos estrechos con estas personas. La primera mención oficial de la existencia de transmisión comunitaria en nuestro país se hizo el 1 de abril de 2020. Como medida preventiva, se estableció el ASPO (Aislamiento Social Preventivo y Obligatorio) que rigió en la mayor parte del territorio nacional desde el 20 de marzo del 2020 hasta el 8 de noviembre del mismo año.

En Argentina, hacia el día 11 de enero de 2021, se han registrado cerca de 1.7 millones de casos y alrededor de 45 mil fallecimientos, en todas las provincias del país.



(a) Evolución temporal de los casos nacionalmente acumulados cada 100 mil habitantes.

(b) Evolución temporal de los casos nacionalmente acumulados cada 100 mil habitantes en una ventana de 14 días.

Figura 1.3: Evolución temporal de los casos acumulados y de los casos acumulados en 14 días, a nivel nacional.

En la Figura 1.3 se presentan la cantidad de casos nacionalmente acumulados cada 100 mil habitantes en total, desde el comienzo de la pandemia y en una ventana de 14 días.

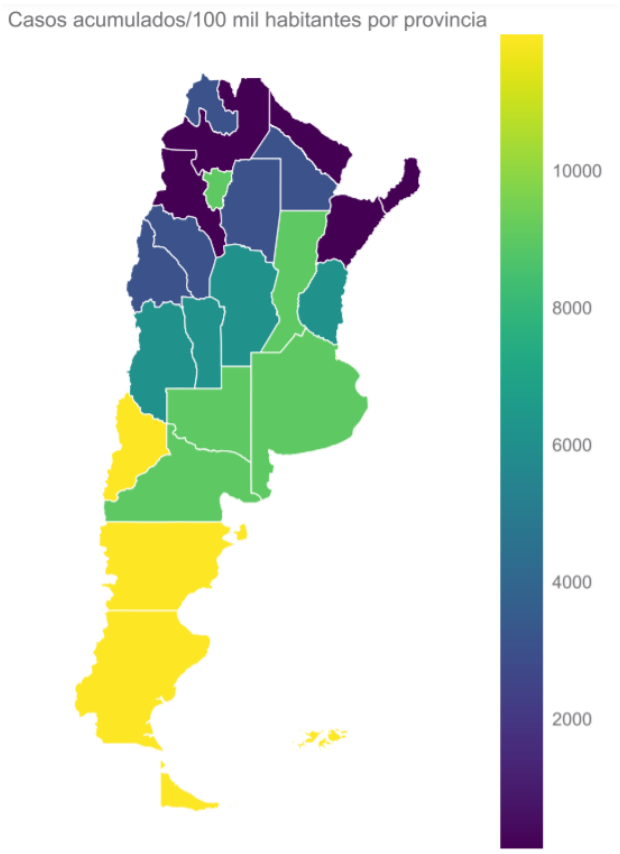


Figura 1.4: Casos confirmados acumulados cada 100 mil habitantes hasta el día 11 de enero de 2021, para cada provincia de Argentina.

Adicionalmente, en la Figura 1.4 se tiene un mapa de Argentina indicando la cantidad

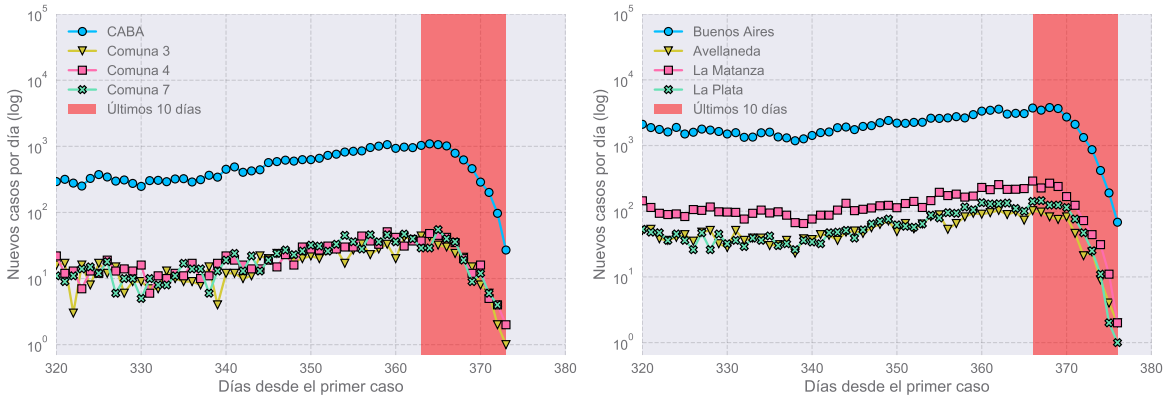
de casos confirmados cada 100 mil habitantes por provincias. Se observa que todas las provincias han tenido personas infectadas, siendo algunas más afectadas que otras.

A nivel mundial, al mes de enero de 2021, la enfermedad se ha propagado a más de 200 países y territorios de los 6 continentes. De acuerdo con la OMS, se han registrado alrededor 88 millones de contagios y 1.9 millones de muertes.

1.2. Datos oficiales de la República Argentina.

Los datos de test positivos en función del tiempo que se utilizan para la realización de esta tesis son calculados a partir de los datos públicos del Ministerio de Salud de la República Argentina. Estos son actualizados todos los días alrededor de las 20:00 horas, hora local en el sitio oficial del ministerio [5]. Estos incluyen todas las determinaciones realizadas en el país, junto con una fecha oficial de apertura del caso y la fecha de inicio de síntomas. En caso de que esta última se encuentre disponible, se la utiliza como el día a partir de la cual una persona está infectada y puede contagiar a otros. En caso contrario, se elije la fecha aleatoriamente en un rango de 8 días centrado en la fecha de apertura. Esta decisión se baso en la estrategia usada en [6] ante la falta de un documento oficial que explicite el criterio para asignar una fecha de apertura.

Otra particularidad de los datos de Argentina es el retraso de origen administrativo en la carga de datos en varios municipios, que puede llegar a ser de alrededor de 10 días. Por ejemplo, en varias localidades de la provincia de Buenos Aires ó en la Ciudad Autónoma de Buenos Aires (CABA). Esto genera que se observe una falsa disminución de casos teniendo en cuenta los días más recientes. Esto se ilustra en la Figura 1.5, donde se presentan los casos confirmados diariamiente en función del tiempo para diferentes municipios de CABA y de la Provincia de Buenos Aires respectivamente.



(a) CABA y tres de sus comunas. **(b)** Buenos Aires y tres de sus municipios.

Figura 1.5: Nuevos casos reportados por día en CABA y algunas de sus comunas, y en la Provincia de Buenos Aires y algunos de sus municipios. En rojo, los últimos 10 días de datos. Se observa una caída aparente de casos en los últimos datos.

Capítulo 2

Modelo semiempírico

2.1. Número reproductivo empírico

En el modelado matemático de enfermedades infecciosas, el número reproductivo empírico a un tiempo determinado R_t mide la cantidad de casos secundarios de enfermedad por cada caso primario. Es decir, es la cantidad de personas contagiadas por una persona infecciosa durante su período activo de contagio. De esta manera, da una idea de la velocidad de propagación de una epidemia en función del tiempo. Con frecuencia se hace referencia al número reproductivo básico R_0, que es el número reproductivo a $t = 0$, en una población susceptible en equilibrio [2].

Para la epidemia de la COVID-19, es útil definir un número reproductivo empírico según la ecuación [6]:

$$R_t = \frac{N_{t-1} + N_t + N_{t+1}}{N_{t-6} + N_{t-5} + N_{t-4}}, \tag{2.1}$$

donde N_t es el número de nuevos casos reportados correspondientes al día t luego del comienzo de síntomas del primer positivo detectado. El promediado de varios días antes y después del número t permite disminuir las posibles fluctuaciones en N_t. Esta definición fue realizada en la bibliografía para el estudio de la primera ola de la COVID-19, y no se ha analizado en el presente trabajo su utilidad posterior (por ejemplo, aplicada a nuevas variantes del virus).

Heurísticamente, la anterior definición puede entenderse de la siguiente manera: el desplazamiento atrás en el tiempo alrededor de una ventana de 5 días tiene en cuenta que las personas infectadas comienzan a contagiar a otros luego de un determinado período de tiempo tras su primera exposición al virus. Adicionalmente, se supone que estos casos tienen impacto sobre los nuevos casos reportados en los días $t-1, t, t+1$. Es decir, estos últimos son casos secundarios de los casos a tiempos $t-6, t-5, t-4$.

Matemáticamente, esta expresión puede obtenerse de considerar que los casos a un día t son contagiados por los casos al día $t-5$. Este retardo de 5 días corresponde a

aproximadamente el día de inicio de síntomas después de la infección, de acuerdo con la OMS. Es decir, consideramos que un infectado contagia a otras personas el día de inicio de síntomas. A su vez, estas desarrollan síntomas 5 días después. Para calcular el número reproductivo R_t, consideraremos que este se mantiene estable en los días $t-6, t-5, t-4$, de manera que los casos secundarios de $N_{t-6}, N_{t-5}, N_{t-4}$ pueden calcularse como:

$$N_{t-6} = R_t N_{t-1} \tag{2.2a}$$

$$N_{t-5} = R_t N_t \tag{2.2b}$$

$$N_{t-4} = R_t N_{t+1}. \tag{2.2c}$$

Sumando estas ecuaciones, puede despejarse la expresión utilizada para calcular el número reproductivo empírico, presentada previamente en la ecuación 4.2.

Cabe aclarar que esta definición de R_t está hecha en base a los casos positivos reportados, que es diferente que el número de nuevos casos totales. Así, en principio, permitiría estimar únicamente la cantidad de nuevos tests positivos a un determinado día en el futuro. Sin embargo, si se asume que el número total de personas infectadas es proporcional a N_t, entonces R_t sirve para el conjunto de todos los casos. De esta manera, se puede esperar que si $R_t > 1$ la epidemia continuará, mientras que si $R_t < 1$ la epidemia se extinguirá.

En la práctica, el número R_t presenta fuertes fluctuaciones. Por ello, se define R_t^7 como el promedio en una ventana de 7 días de esta cantidad, es decir que:

$$R_t^7 = \sum_{i=-3}^{3} \frac{R_{t+i}}{7} = \frac{R_{t-3} + R_{t-2} + R_{t-1} + R_t + R_{t+1} + R_{t+2} + R_{t+3}}{7}. \tag{2.3}$$

2.1.1. Efectos de borde

Se observa que las definiciones de R_t y de R_t^7 presentan problemas cuando se consideran días cercanos al presente. Esto se debe a que para calcular las cantidades mencionadas se necesitarían los valores de N_t para días $t \geq hoy$, donde hoy indica el día del último reporte de casos. Con el objetivo de solucionar este problema, se calculan promedios de N_t a 7 días y se completa la serie extrapolando linealmente. Este proceso se realiza hasta obtener estimaciones de N_t para $t \geq hoy - 4$. Así, esta estrategia permite subsanar los efectos de borde. Sin embargo, es esperable que los valores de N_t y sus cantidades derivadas sufran reajustes a medida que se actualizan los datos, por lo menos hasta que $t < hoy - 4$.

2.2. Número de casos activos

El número R_t^7 permite estimar el número de casos secundarios por cada infectado reportado. Así, si se conoce el número total de infectados en un día determinado, se pueden estimar los contagiados a un tiempo después multiplicándolo por R_t^7. Sin embargo, es imposible conocer con exactitud el número total de personas infecciosas a un tiempo determinado.

Una estimación posible para el número de nuevos casos reportados cada 100 mil habitantes es la incidencia acumulada cada 100 mil habitantes en los últimos 14 días antes de t, denotada por A_t^{14}. Ésta se define en la ecuación 2.4, y para calcularla se debe sumar el número de casos reportados en los 14 días anteriores al día que se considera. Se consideran 14 días como un número adecuado, dado que éste es el período medio de recuperación de un infectado.

$$A_t^{14} = \frac{100000}{N_{pop}} \sum_{i=t-13}^{t} N_i = N_{t-13} + N_{t-12} + ... + N_{t-1} + N_t. \tag{2.4}$$

donde N_{pop} es el número de habitantes de la población en estudio.

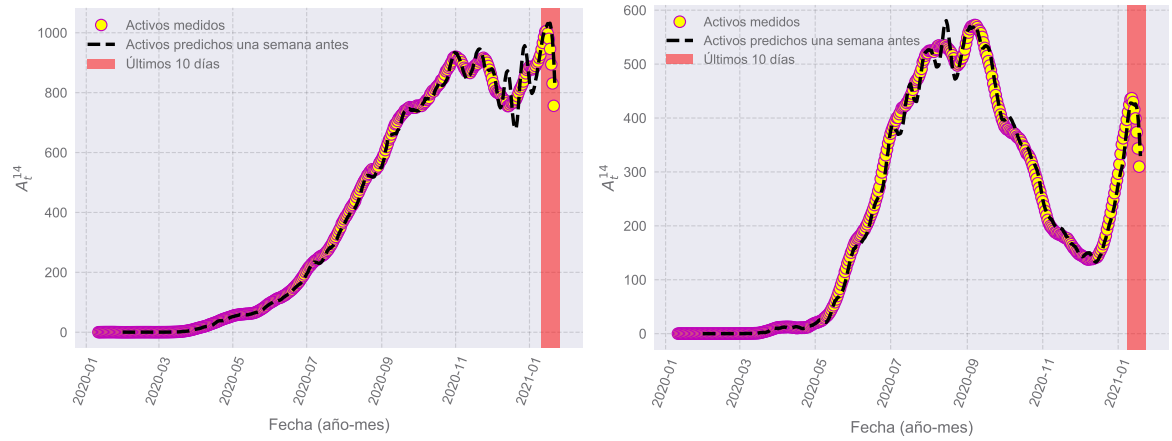
2.3. Predicción de casos

A partir de los cálculos de $\mathcal{R}_t^7$ y A_t^{14} para un día t pueden estimarse la cantidad de casos reportados en el futuro cercano, simplemente multiplicando ambas cantidades. En trabajos previos, se ha correspondido [6] este producto a los nuevos casos del día $t + 7$. Es decir, que si se considera que:

$$\frac{100000}{N_{pop}} N_t = A_t^{14} \simeq R_{t-7}^7 A_{t-7}^{14}, \tag{2.5}$$

el acuerdo entre predicción y los reportes de casos reales parece ser correcto empíricamente. La cantidad a la derecha de la igualdad $R_t^7 A_t^{14}$ recibe en la bibliografía el nombre de índice de crecimiento potencial, y se denota con la sigla P_t [6] [7].

Finalmente, dos ejemplos que ilustran la predicción de casos se presentan en la Figura 2.2. Allí, se grafican los test positivos cada 100 mil habitantes hasta el 11 de enero de 2021 y la predicción efectuada una semana antes para Argentina en total y para CABA. En ambos casos, se señalan los últimos 10 días dado que los retrasos en la carga de datos pueden limitar la capacidad predictiva de las herramientas presentadas.



(a) Argentina. **(b)** CABA.

Figura 2.1: Evolución temporal de casos acumulados A_t^{14} para la Argentina y para CABA. En línea punteada, la predicción para esta cantidad realizada una semana anterior. Se observa un buen acuerdo entre ambas cantidades.

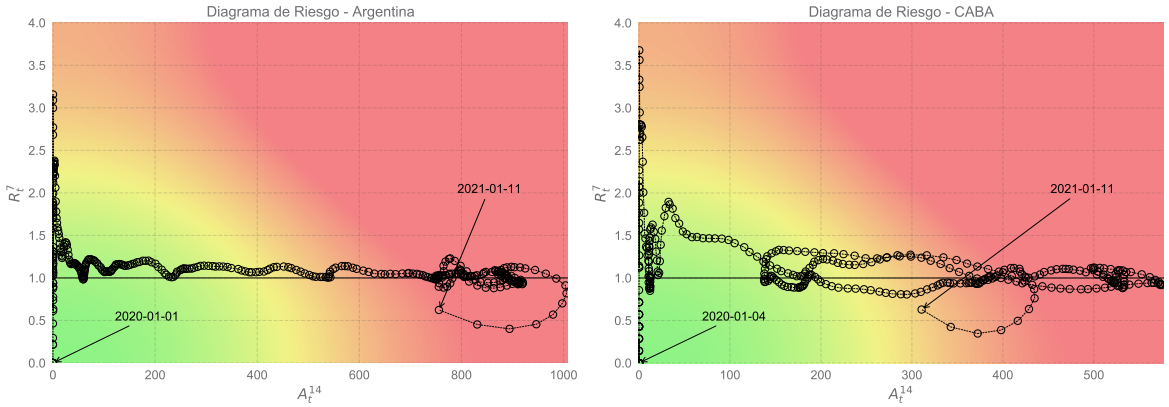
2.4. Diagramas de Riesgo

El riesgo sanitario en los próximos días puede caracterizarse a partir del índice P_t, que da una estimación de la cantidad esperable de tests con resultado positivo al tiempo t cada 100 mil habitantes.

Por un lado, si DTL es la cantidad de tests realizados cada 100 mil habitantes, entonces la situación será de riesgo si $P_t >> DTL$. Esto contempla un riesgo, dado que la cantidad de casos sobrepasaría ampliamente la cantidad de tests. Lo mismo sucederá si la cantidad de personas que requieran atención médica en el futuro sobrepasa la capacidad del sistema sanitario, que es limitada. Para estimar esta cantidad, se supone que f es la fracción de los casos positivos que desarrollan síntomas graves y requieren atención en la UCI, y que C es el número de cuidados disponibles. De esta manera, se puede decir que hay riesgo cuando $P \cdot f > C$. En el caso de Argentina, se puede estimar empíricamente que $f \approx 0{,}035$, sin tener en cuenta rangos etarios. Este número se obtiene realizando el cociente entre los casos que requieren atención hospitalaria y el total de los casos reportados. Esto es posible el Ministerio de Salud indica en su base de datos aquellos casos que requieren atención hospitalaria por su severidad.

Una representación grafica posible del riesgo de la situación sanitaria esta dada por los diagramas de riesgo. En estos, se grafica en el eje horizontal la incidencia A_t^{14} y en el eje vertical el número reproductivo R_t^7. Adicionalmente, para representar $P_t = R_t^7 A_t^{14}$, se utiliza una escala de colores, donde el rojo indica situaciones de mayor riesgo y el verde, de menos. Ejemplos de diagramas de riesgo se pueden visualizar en la Figura 2.2a, donde se muestran el diagrama de riesgo nacional y para CABA. Se observa que la carga incompleta de los datos en los últimos días crea una aparente evolución reciente

hacia situaciones de menor riesgo.



(a) Argentina. **(b)** CABA.

Figura 2.2: Diagramas de riesgo para la Argentina y para CABA. Las flechas indican las fechas de inicio y fin del registro de casos para la realización de estos diagramas.

2.5. Conclusiones

En este capítulo, se han introducido herramientas existentes que permiten generar predicciones aceptables para los nuevos casos reportados de la COVID-19 en el futuro cercano. Adicionalmente, se presentó el concepto de los diagramas de riesgo, que dan una idea cualitativa y cuantitativa del riesgo del colapso del sistema sanitario de una población. En próximos capítulos, nos dedicaremos a estudiar el error de las estimaciones resultantes, y posteriormente a intentar reducirlo. Finalmente, se han introducido y discutido algunas particularidades de los reportes de casos positivos de la Argentina que son importantes a la hora de evaluar la calidad de la predicción realizada.

Capítulo 3

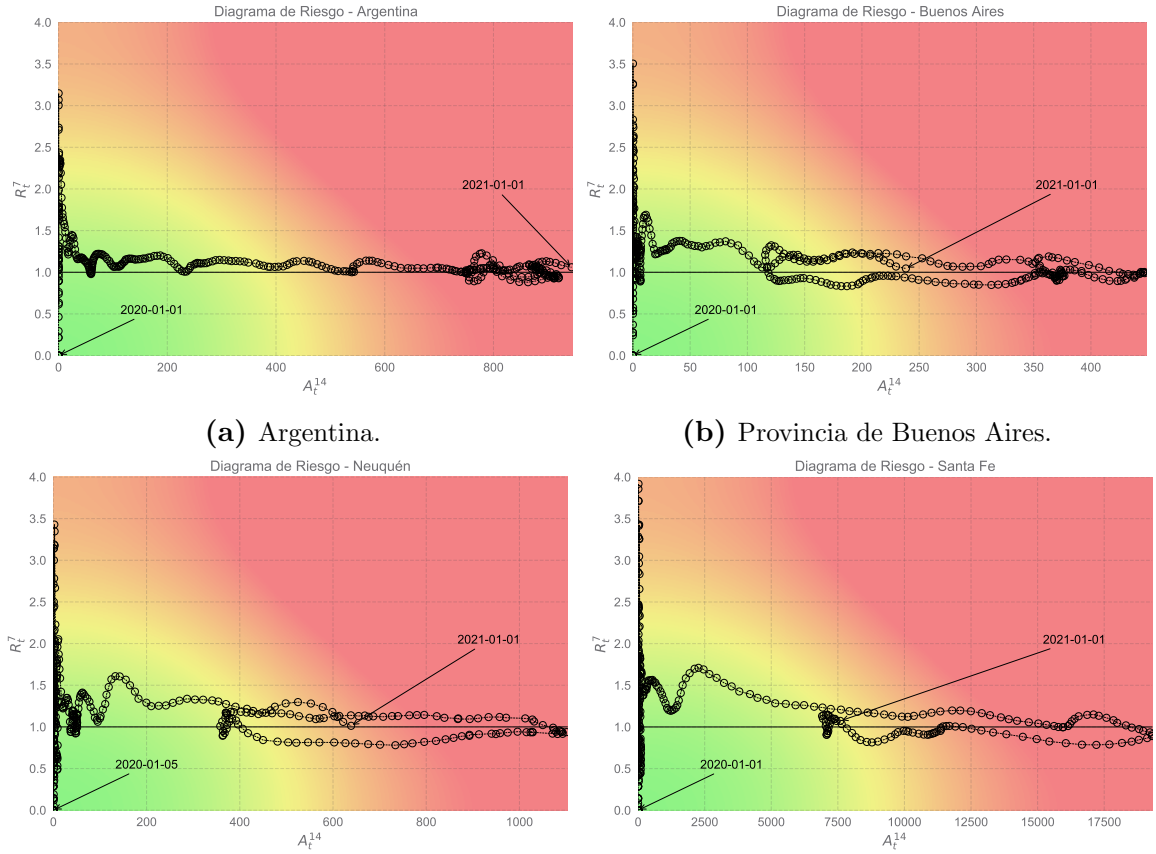
Análisis de las predicciones

En el capítulo 2 presentamos las herramientas básicas desarrolladas para la predicción de casos confirmados de la COVID-19. En este capítulo, dedicaremos la sección 3.1 a hacer algunas observaciones particulares sobre los diagramas de riesgo. Por otra parte, en la sección 3.2 utilizaremos diferentes estrategias para el análisis del error predictivo. Finalmente, en la sección 4.1 estudiaremos la existencia de una cantidad de días de delay óptima como posibilidad para mejorar las predicciones, sin modificar otros aspectos de los métodos del capítulo anterior.

3.1. Análisis de los diagramas de riesgo y del número reproductivo $\mathcal{R}_t^7$.

En la sección 2.4 se introdujo el concepto de diagrama de riesgo, útil para evaluar la evolución temporal del riesgo de colapso del sistema sanitario de una población. En la Figura 3.1 se presentan diagramas de riesgo para la Argentina y algunas de sus provincias. En particular, la Provincia de Buenos Aires, la Provincia de Neuquén y la Provincia de Santa Fe. A diferencia de los ejemplos de la Figura 2.2, en esta ocasión se sustraen los 10 últimos días de datos para no tener en cuenta la carga incompleta de los mismos. En lo que continúa de esta tesis, a menos que se indique lo contrario, se consideran entonces datos hasta el 1 de enero de 2021 inclusive. Esto permite circunscribir el problema, eliminando variables como el cambio en las políticas sanitarias, la vacunación y las nuevas variantes del virus SARS-CoV-2.

En estos diagramas de riesgo se evoluciona temporalmente de una situación de menor a una de mayor riesgo, como es de esperarse. Lógicamente, cuando inicia la epidemia, tanto R_t^7 como A_t^{14} son cercanos a cero. Por otra parte, hasta el último día de datos utilizados se distingue que todas las localidades permanecen en una región de riesgo medio o alto. En cambio, si la epidemia hubiese entrado en remisión, se distinguiría un acercamiento hacia el origen.



(a) Argentina. **(b)** Provincia de Buenos Aires.

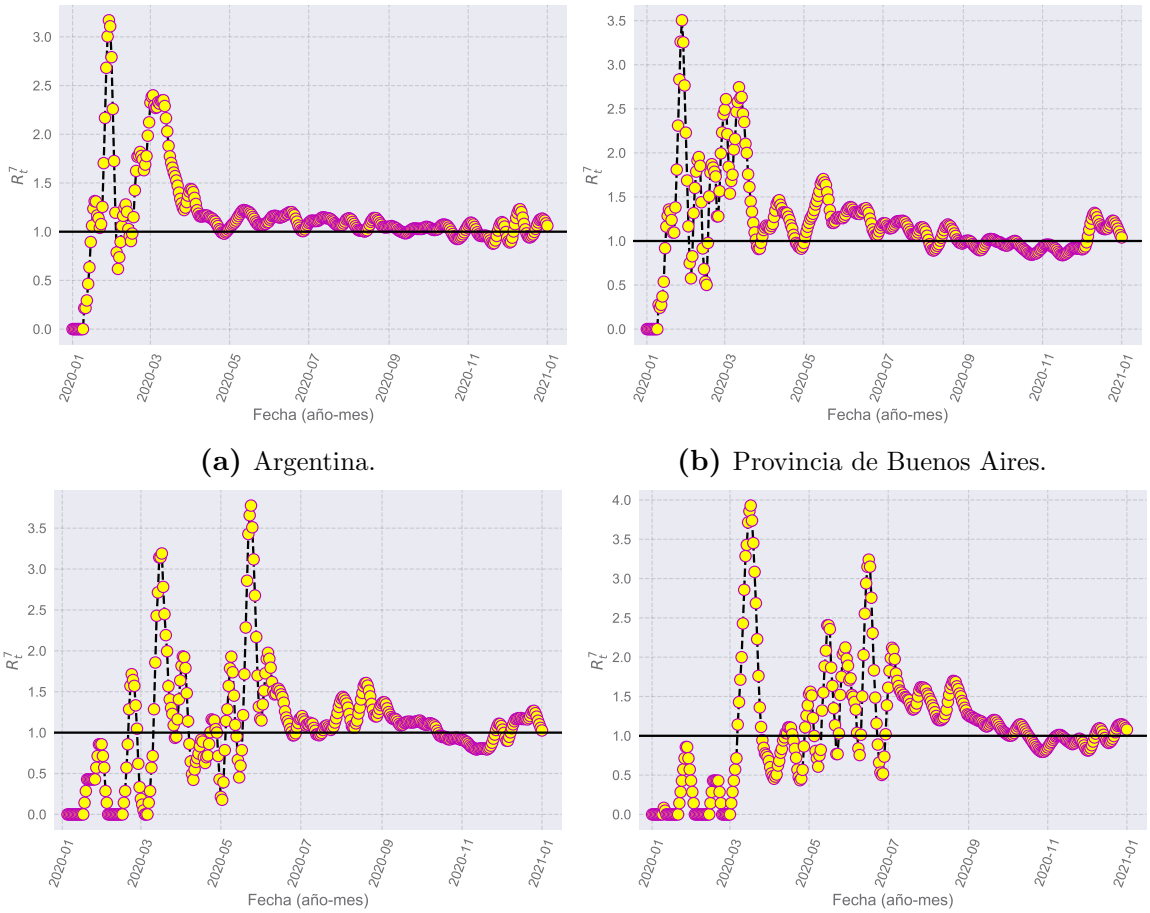
(c) Provincia de Neuquén. **(d)** Provincia de Santa Fe.

Figura 3.1: Diagramas de riesgo para Argentina y las provincias de Buenos Aires, Neuquén y Santa Fe. Las flechas indican el inicio y el último registro de casos en la región correspondiente.

Una observación importante se puede hacer sobre cómo es la evolución temporal de estos diagramas. En los cuatro casos mostrados en la Figura 3.1, el número R_t^7 es inicialmente muy elevado, llegando a valores cercanos a $3 - 3{,}5$, pero luego se torna cercano a la unidad en mayor ó menor medida. Esto se repite en muchas provincias y localidades argentinas, no únicamente en las aquí presentadas.

Esta observación nace de la simple inspección de los diagramas de riesgo, y no parece depender de los valores que toma la variable A_t^{14}. Es útil graficar la evolución temporal de este número reproductivo. Esto se hace en la Figura 3.2 para las poblaciones mencionadas anteriormente. Es claro de aquí que la evolución inicial no es equivalente en todos los casos. Por ejemplo, para Argentina y la Provincia de Buenos Aires, el período inicial de valores de R_t^7 es más corto, y luego rápidamente $R_t^7 \to 1$. Lo mismo no sucede para las provincias de Neuquén y de Santa Fe. Sin embargo, en todos los casos, eventualmente el número reproductivo se acerca a la unidad en forma oscilatoria.

No obstante, no todos los diagramas de riesgo son análogos a los presentados en 3.1. Particularmente, se distingue entre ellos el caso de la ciudad de San Carlos de Bariloche, en la Provincia de Río Negro. En la Figura 3.3a se puede observar su diagrama de



(a) Argentina.

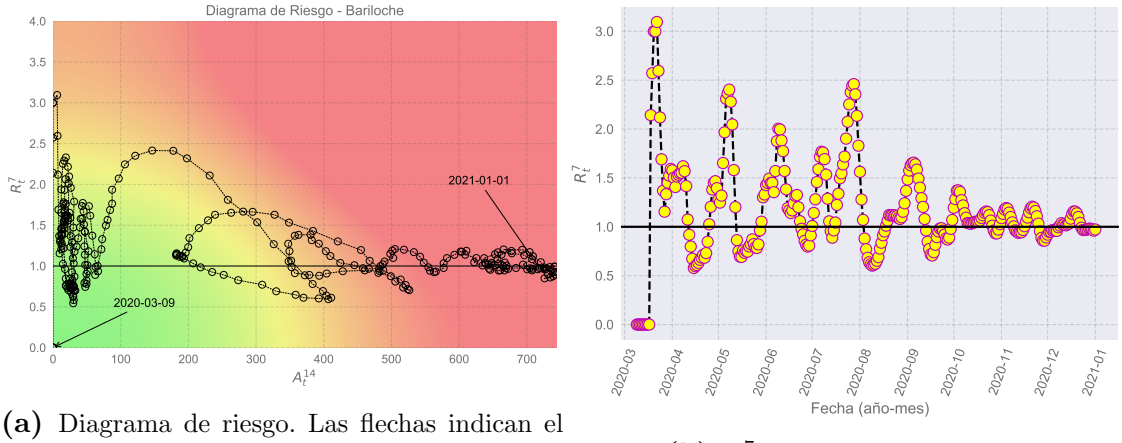
(b) Provincia de Buenos Aires.

(c) Provincia de Neuquén.

(d) Provincia de Santa Fe.

Figura 3.2: R_t^7 para Argentina y las provincias de Buenos Aires, Neuquén y Santa Fe en función del tiempo.

riesgo, que presenta una dinámica con ciclos en el plano $R_t^7 - A_t^{14}$. Esto se refleja en las oscilaciones de R_t^7 en la Figura 3.3b.



(a) Diagrama de riesgo. Las flechas indican el comienzo y el final de los datos analizados.

(b) $\mathcal{R}_t^7$ en función del tiempo.

Figura 3.3: (a) Diagrama de riesgo y (b) Evolución de R_t^7 para la ciudad de San Carlos de Bariloche.

Allí, se tiene que R_t^7 toma el mismo valor por lo menos dos veces. Esto es necesario, pero claramente no suficiente para constituir una de las oscilaciones descriptas. Este tipo de fenómeno se observo únicamente en el caso de la ciudad de San Carlos de Bariloche, y puede deberse a una dinámica espacial particular en este caso. Principalmente, podría asociarse a una primera etapa de brotes espacialmente localizados, antes de que la enfermedad se expandiera a toda la ciudad.

3.2. Análisis del error en las predicciones.

En esta sección, procederemos al análisis del error predictivo. Primero, hacemos una breve introducción sobre la teoría de los fenómenos y exponentes críticos. Luego, procederemos a analizar el error en búsqueda de algo similar a un exponente crítico. Para ello, se desarrollan dos propuestas en las secciones 3.2.3 y 3.2.3.

3.2.1. Modelo de Ising. Fenómenos y exponentes críticos.

El modelo de Ising [8] estudia el comportamiento de una red d-dimensional de N espines σ_i que pueden tomar valores discretos $+1$ o -1, donde i indexa los sitios de la red. Los espines interactúan entre sí a primeros vecinos con una interacción isotrópica. Adicionalmente, se incorpora el efecto de un campo magnético externo H, que se considera constante e uniforme en la red. El Hamiltoniano del modelo de Ising está dado por:

$$H = -J \sum_{<i,j>} \sigma_i \sigma_j - \mu H \sum_i \sigma_i, \tag{3.1}$$

donde $< i, j >$ indica una suma a primeros vecinos, J es una constante que representa la interacción entre espines y μ es el momento magnético neto asociado a un espín.

Este sistema tiene un punto crítico cuando $H \to 0$ y $T \to T_c$, es decir, cuando el campo magnético externo es nulo y la temperatura se acerca a una temperatura T_c, llamada la temperatura crítica. La magnetización neta del sistema M es tal que si $H \to 0$ tiende a un valor limite M_0, que es $M_0 = 0$ para $T \geq T_c$ y $M_0 \neq 0$ si $T < T_c$. En ausencia de campo magnético externo, Si $T >\geq T_c$ la magnetización tiende a anularse, orientándose los espines de la red aleatoriamente. En cambio, para $T < T_c$ los espines tienden a alinearse espontáneamente en una misma dirección, dando lugar a una magnetización neta no nula.

En la vecindad de los puntos críticos, varias cantidades físicas presentan singularidades. Estas se expresan en términos de leyes de potencias caracterizadas por constantes llamadas exponentes críticos [8][9]. En el caso del modelo de Ising, la manera que $M_0 \to 0$ a medida que $T \to T_c$ desde $T < T_c$ define el exponente crítico β a partir de:

$$M_0 \sim (T - T_c)^{\beta}. \tag{3.2}$$

Otros dos exponentes críticos γ y γ' pueden definirse por la manera en la que la susceptibilidad a campos bajos (es decir, cuando $H \to 0$) χ_0 diverge a medida que $T \to T_c$:

$$\chi_0 \sim \left(\frac{\partial M}{\partial H}\right)_{T,H\to 0} \sim \begin{cases} (T - T_c)^{\gamma}, & \text{si } T \gtrsim T_c \\ (T_c - T)^{\gamma'}, & \text{si } T \lesssim T_c \end{cases} \tag{3.3}$$

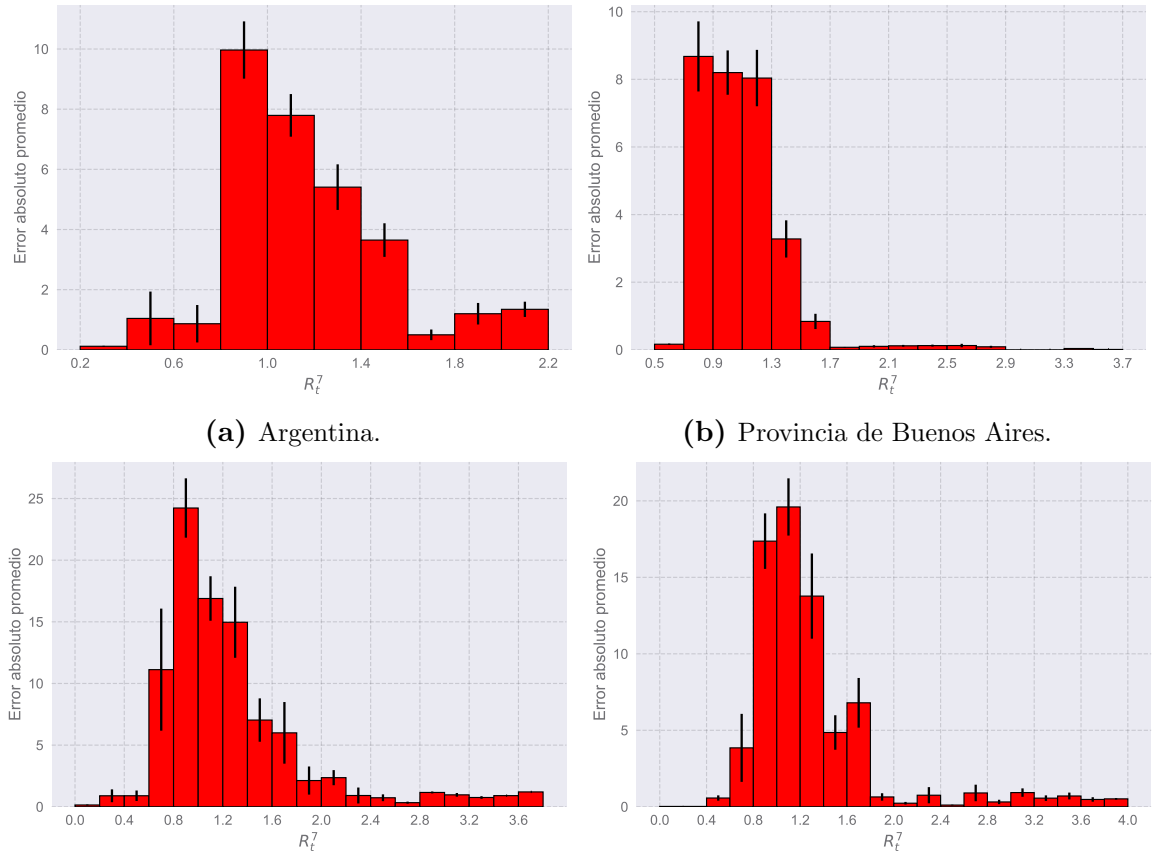
Adicionalmente, pueden estudiarse otras cantidades que divergen cerca del punto crítico, como la longitud de correlación en este caso.

En todos los casos, los exponentes críticos son sumamente importantes por varias razones. De su definición, observamos que si se grafica la cantidad física en estudio en función de $T - T_c$ en escala log-log estos comportamientos son teóricamente fáciles de encontrar por simple inspección para valores de T lo suficientemente cercanos al valor critico T_c. La determinación de los exponentes críticos también es, en teoría, igual de simple. Por otra parte, los exponentes críticos son utilizados con frecuencia para establecer clases de universalidad en diferentes áreas de la física.

3.2.2. Estudio de los histograma convencionales de error predictivo.

Con el objetivo estudiar el error, se realizaron histogramas del error absoluto en función de los valores de R_t^7 correspondientes, para diferentes localidades. Para ello, dada una población, se calculó para cada día t la predicción según la ecuación 2.5 y luego el error absoluto. Dado que correspondía con un día t, se lo asoció inmediatamente con el valor de R_t^7 en ese día. Luego, se dividieron los valores de R_t^7 en bines de 0,2 y se le asignó a cada uno de ellos el promedio del error absoluto para observaciones pertenecientes a cada bin. Los resultados de este procedimiento para Argentina y las provincias de Buenos Aires, Neuquén y de Santa Fe se presentan en la Figura 3.4.

Allí, se observa que el error es máximo para valores de R_t^7 cercanos a la unidad. Esto se repite en gran parte de las localidades analizadas. Sin embargo, se observa que no todos los histogramas son completamente iguales. Por ejemplo, existen localidades como la Provincia de Neuquén que presentan otros máximos locales del error absoluto promedio, para valores mayores de R_t^7, como se observa en la Figura 3.4c. De todas maneras, persiste la existencia de un máximo absoluto para valores de $R_t^7 \simeq 1$. Esto podría ser similar a lo que sucede con cantidades que divergen como la susceptibilidad χ_0 cerca del punto critico en el modelo de Ising. En casos en los que el sistema no se encuentra en el límite termodinámico, la susceptibilidad no diverge, pero sí presenta



(a) Argentina.

(b) Provincia de Buenos Aires.

(c) Provincia de Neuquén.

(d) Provincia de Santa Fe.

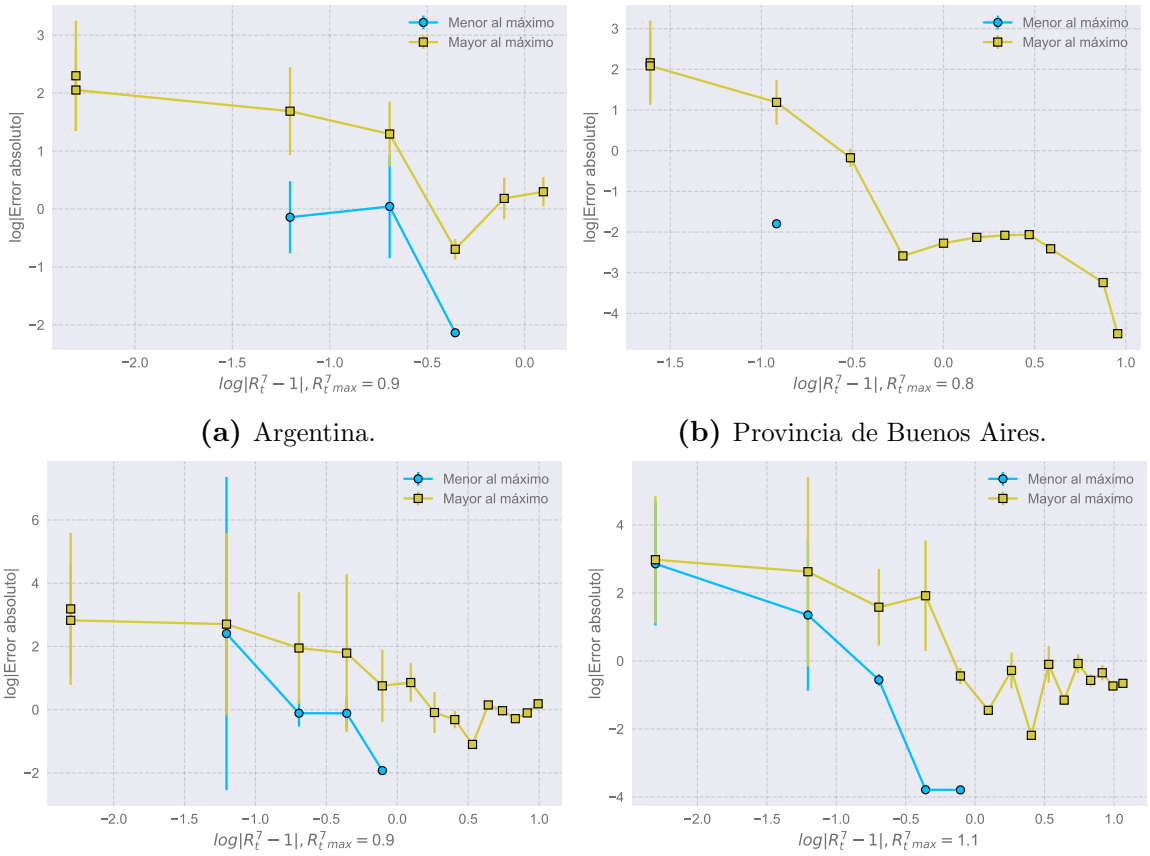
Figura 3.4: Histogramas convencionales de error absoluto promedio en función de R_t^7 para Argentina y las provincias de Buenos Aires, Neuquén y de Santa Fe.

un máximo, como lo hace el error absoluto promedio cerca de $R_t^7 \simeq 1$. Esto motiva la pregunta de si el acercamiento de R_t^7 a la unidad presenta características de fenómeno crítico. Sin embargo, cabe destacar que tendría una diferencia fundamental con el caso del modelo de Ising, que consiste en que R_t^7 no es un parámetro de control como la temperatura. En cambio, el sistema sería tal que R_t^7 se acerca automáticamente en el largo plazo a la unidad, que sería su valor crítico.

Para contemplar la posibilidad de un fenómeno crítico, se graficó en escala log-log el error absoluto promedio en función de la distancia del valor central de cada bin a $R_t^7 = 1$, el cual podría funcionar como un parámetro crítico. Como ya se observó, cerca de este valor el error absoluto promedio es máximo. Adicionalmente, como se discutió en la sección 3.1, es el valor al que se acerca el número reproductivo a largo plazo.

En la Figura 3.5 se presentan los gráficos mencionados para las poblaciones de Argentina y las provincias de Buenos Aires, Neuquén y Santa Fe. En ellos, se indica el valor central del bin tal que el error es máximo en cada caso, al que se denomina $R_{t\,max}^7$. Adicionalmente, se separan gráficamente los datos según si el valor de R_t^7 es mayor o menor a $R_{t\,max}^7$.

De la inspección de la figura anterior, no se puede concluir la existencia o no de



(a) Argentina. **(b)** Provincia de Buenos Aires.

(c) Provincia de Neuquén. **(d)** Provincia de Santa Fe.

Figura 3.5: Error absoluto promedio en función del valor R_t^7 central de cada bin de la Figura 3.4, en escala log-log para Argentina y las provincias de Buenos Aires, Neuquén y de Santa Fe.

un comportamiento similar al esperado de los fenómenos críticos descriptos en 3.2.1. Si bien la tendencia es similar en los casos presentados, no es claro que se corresponda con una ley de potencias lo suficientemente cerca del punto crítico.

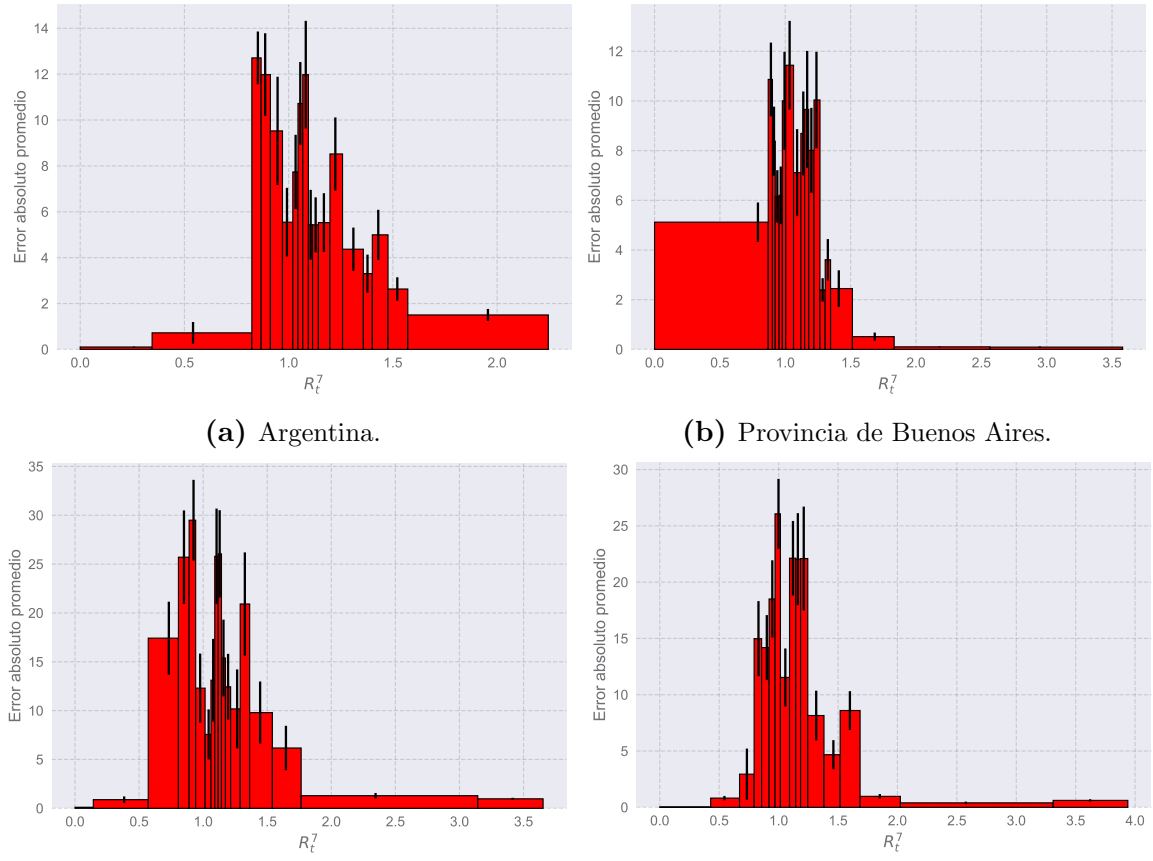
3.2.3. Diseño y estudio de nuevos histogramas de error predictivo.

En el caso de los histogramas anteriores, se organizan las observaciones según su valor de R_t^7 en bines. Así, resulta que hay bines con muchas observaciones, como sucede en torno a la unidad, que tienen del orden de las decenas o centenas. En cambio, para los valores R_t^7 cercanos a los valores máximos o mínimos alcanzados en cada localidad, puede no haber ni siquiera media decena de observaciones. Esto puede provocar que la estadística en estas regiones no sea buena. Para solucionar este problema, se buscó crear histogramas que tuviesen el mismo número de observaciones en cada bin, en contraposición con el caso de los histogramas convencionales.

Para ello, se ordenaron los datos de cada localidad de menor a mayor teniendo en cuenta su R_t^7 y se los dividió en grupos de 20 registros, habiendo asignado previamente

la predicción y el error correspondientes. Así, se calculó para cada grupo de 20 datos el debido error absoluto promedio y la desviación estándar del promedio. De esta manera, se obtienen bines con igual cantidad de datos y con valores de R_t^7 similares entre sí. Los resultados de este procedimiento para Argentina y las provincias de Buenos Aires, Neuquén y de Santa Fe se presentan en la Figura 3.6.

De estos histogramas modificados, se observa claramente que los valores cercanos a la unidad son más frecuentes que los otros, simplemente notando que los bines en esta región de los gráficos son mucho más angostos. En cambio, hacia los extremos del histograma, se tienen que tener en cuenta rangos más grandes de R_t^7 para contar 20 datos y, por lo tanto, los bines se ven más anchos. Por otra parte, estos histogramas no presentan la misma estrutura que los de la Figura 3.4, en los que se observa un máximo absoluto y a lo sumo algún máximo local que toma un valor mucho menor. Por el contrario, en varios de los casos, se observan dos máximos de valores muy similares. Finalmente, la desviación del promedio resulta ser mucho mayor en general.



(a) Argentina.

(b) Provincia de Buenos Aires.

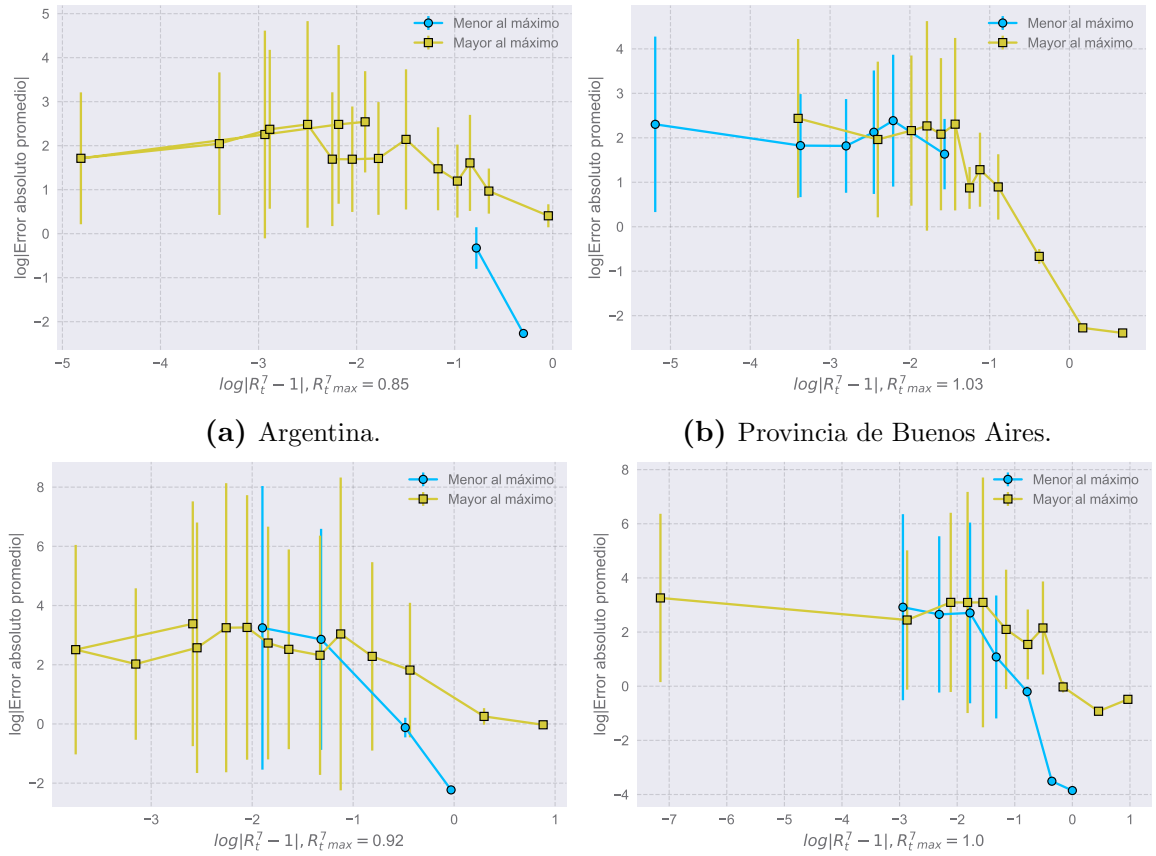
(c) Provincia de Neuquén.

(d) Provincia de Santa Fe.

Figura 3.6: Histogramas modificados de error absoluto promedio en función de R_t^7 para Argentina y las provincias de Buenos Aires, Neuquén y de Santa Fe. Cada bin corresponde a un número constante de 20 observaciones.

Análogamente al caso presentado en la sección anterior, se busca algún tipo de comportamiento asociado a los fenómenos críticos en torno a $R_t^7 = 1$. Para ello, se

repite el procedimiento realizado para obtener de los histogramas convencionales los gráficos de la Figura 3.5, pero a partir de los nuevos histogramas modificados. Los resultados se presentan en la Figura 3.7. Para cada una de las poblaciones en estudio, se indica el valor de $R^7_{t\,max}$ y se distinguen los datos con valor de R^7_t superior e inferior. Al igual que en el caso anterior, se tiene que los valores de $R^7_{t\,max}$ se encuentran cerca de la unidad.



(a) Argentina.

(b) Provincia de Buenos Aires.

(c) Provincia de Neuquén.

(d) Provincia de Santa Fe.

Figura 3.7: Error absoluto promedio en función del valor R^7_t central de cada bin de la Figura 3.6, en escala log-log para Argentina y las provincias de Buenos Aires, Neuquén y de Santa Fe. Cada punto de los gráficos corresponde a 20 observaciones.

Al igual de lo que sucede en los gráficos log-log anteriores, no se distingue claramente la posibilidad de una ley de potencias similar a lo que ocurre en los fenómenos críticos. Sin embargo, tampoco puede descartarse categóricamente. En cambio, a diferencia de lo observado en los gráficos anteriores, se tiene que cerca del valor máximo hay una clara tendencia del error al estancarse, en todos los casos. Cabe destacar nuevamente que los errores, como se observa también en los histogramas de la Figura 3.6, son mucho mayores que para el análisis realizado en la sección .

3.3. Conclusiones.

En este capítulo nos hemos dedicado a observar y estudiar algunas propiedades de los diagramas de riesgo y del número reproductivo R_t^7. En particular, se destaca la tendencia de que $R_t^7 \to 1$ al largo plazo en varias de las poblaciones en estudio.

Luego, se introdujeron conceptos básicos de la teoría de fenómenos críticos, y estudiamos posteriormente la existencia de un fenómeno de tal naturaleza en la relación entre el error absoluto promedio y $R_t^7 - 1$. Sin embargo, no se pudo concluir o descartar la presencia de un fenómeno de este tipo a partir de ninguno de los dos enfoques utilizados: ni de los histogramas tradicionales ni de los nuevos histogramas con número constante de observaciones. Por un lado, esto puede deberse a que el comportamiento tipo ley de potencias se da típicamente muy cerca del punto crítico. En este caso, podría suceder tan cerca del punto crítico de manera que no se observe con la estadística disponible. Por otro lado, en el caso particular de los gráficos log-log de la Figura 3.7, se puede distinguir una tendencia en el comportamiento. Sin embargo, el error de los datos es muy grande.

A futuro se podría estudiar la posibilidad de añadir datos de otros países o regiones para contribuir a la estadística y así avanzar más en el estudio de la existencia de un fenómeno crítico.

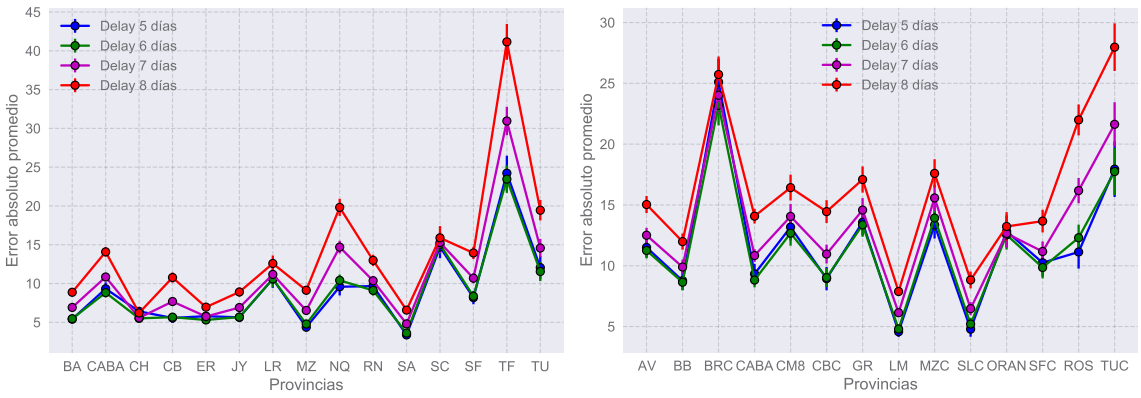
Capítulo 4

Optimización de las predicciones

4.1. Análisis del delay temporal óptimo

Si se recuerdan los conceptos tratados previamente, especialmente en el capítulo 2, se tiene que para predecir los casos a un día t se utilizan los datos a un día $t-r$, con r un retraso de 7 días. De esta manera, se pueden ajustar los casos y predecirlos a 7 días en el futuro razonablemente. Sin embargo, no necesariamente $r = 7$ días es un retraso óptimo con estas herramientas.

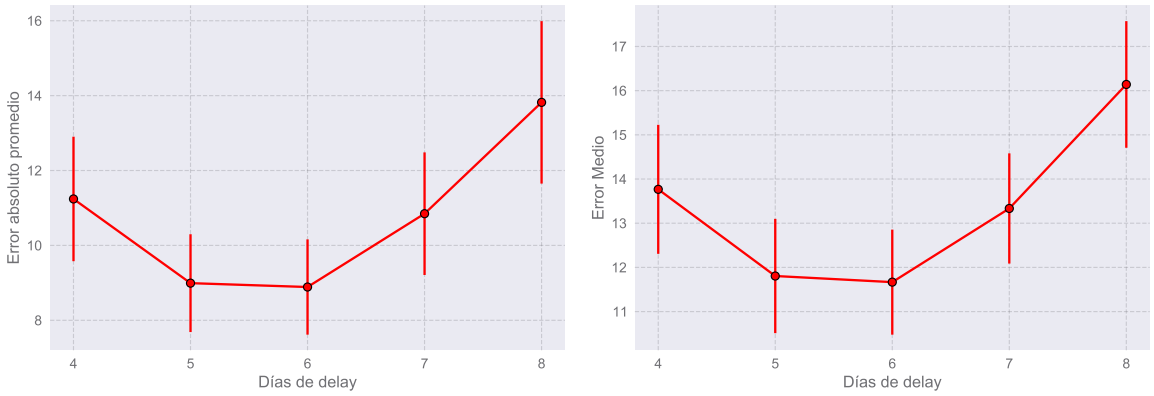
Para dedicarse a hallar un retraso r óptimo, se seleccionaron localidades y provincias de la Argentina, que se detallan en el Apéndice A junto con sus abreviaturas correspondientes. Para cada una de estas poblaciones, se calculó el error absoluto promedio de la predicción para diferentes valores de r. Estos resultados se muestran en la Figura 4.1 con su error correspondiente para un rango de $5-8$ días. Se observa que existen diferencias al utilizar diferentes valores de r, y que las tendencias son similares para la mayoría de las localidades. Esto implica la existencia de un tiempo de retraso óptimo r que es efectivamente un tiempo característico asociado a la enfermedad en cuestión.



(a) Provincias. **(b)** Localidades.

Figura 4.1: Estudio del error absoluto en función de r para diferentes localidades y provincias.

Por otra parte, resulta útil promediar el error absoluto promedio para todas las provincias y localidades, lo cual se hace en la Figura 4.2. Allí, observamos que para el total de las localidades y de las provincias los resultados son similares. En primer lugar, se tiene que no hay una relación monótona entre r y el error promedio. Por otra parte, se observa que $r = 7$ no es el valor óptimo de r. Finalmente, se tiene que el error absoluto promedio es mínimo para $r = 6$, y que para $r = 5$ toma un valor similar al mínimo.



(a) Promedio en provincias. **(b)** Promedio en localidades.

Figura 4.2: Estudio del error absoluto en función de r, teniendo en cuenta todas las provincias y localidades a la vez.

Cabe aclararse que usamos un rango entre $5-8$ días que contiene los valores mínimos para los errores. Adicionalmente, se probaron otros rangos, los cuales mostraron errores significativamente mayores, tanto para $r < 5$ como para $r > 8$.

4.2. Propuestas de modificación del número reproductivo empírico

Hasta ahora, hemos estudiado cambios el valor de r como manera de mejorar las predicciones realizadas, sin modificar de alguna otra manera las herramientas del capítulo 2. A continuación, analizaremos la posibilidad de obtener mejores predicciones cambiando la manera de calcular el número reproductivo empírico R_t. Este fue introducido por primera vez en la ecuación 4.2, donde se lo define como:

$$R_t = \frac{N_{t-1} + N_t + N_{t+1}}{N_{t-6} + N_{t-5} + N_{t-4}},$$

4.2.1. Modificación con 3 pesos

Una primer modificación propuesta para R_t se presenta en la ecuación 4.1. Consiste en agregar pesos a, b y c a los casos primarios de los días $t-6$, $t-5$ y $t-4$ respectiva-

mente. Dado que tienen influencia en las personas que son infectadas en el corto plazo por estos casos, tiene sentido que los coeficientes deban ser positivos o cero. Por otra parte, que a, b y c no sean iguales podría dar cuenta de que las personas en distintas etapas de la enfermedad contagian a otros con diferentes probabilidades. Esto podría deberse a factores biológicos o a factores sociales, como el aislamiento tras el comienzo de síntomas.

$$R_t = \frac{N_{t-1} + N_t + N_{t+1}}{a_3 N_{t-6} + b_3 N_{t-5} + c_3 N_{t-4}}. \tag{4.1}$$

4.2.2. Modificación con 5 pesos

Por otra parte, la ecuación 4.2 presenta otra idea para el cálculo de R_7 utilizando 5 pesos a_5,b_5,c_5,d_5 y e_5 para los infectados en los días $t-6$,$t-5$,$t-4$, $t-3$ y $t-2$. Esta idea está motivada por el razonamiento presentado en la sección 2.1. A diferencia de en ocasiones anteriores, consideramos alguna incidencia de los casos $t-3$ y $t-2$, que corresponden a infectados asintomáticos al momento que contagian a los casos secundarios.

$$R_t = \frac{N_{t-1} + N_t + N_{t+1}}{a_5 N_{t-6} + b_5 N_{t-5} + c_5 N_{t-4} + d_5 N_{t-3} + e_5 N_{t-2}}. \tag{4.2}$$

Cabe aclarar que se han intentado optimizar diversas fórmulas para el cálculo del número reproductivo empíricos. Entre ellas, muchas que ampliaban o modificaban el rango de casos secundarios considerados. Sin embargo, ninguna ha resultado más adecuada que las analizadadas en este capítulo.

4.3. Optimización con redes neuronales

Para encontrar pesos óptimos para el cálculo del número reproductivo empírico, se utilizaron redes neuronales artificiales (RN), una técnica de *machine learning* frecuentemente utilizada en problemas de regresión, procesamiento natural del lenguaje y visión por computadora, entre otros.

Las RN utilizadas en la presente tesis se implementaron utilizando la librería Keras-gpu versión 2.3.1 [10]. Esta librería es de código abierto y permite la construcción y el entrenamiento de RN artificiales. Está escrita en Python y funciona como una interfaz de usuario que permite construir modelos de *machine learning*, basándose en la librería Tensorflow [11]. Esta última es otra librería de Python para programación diferencial, que permite realizar operaciones con tensores, calcular gradientes, entre otras tareas. En particular, se trabajó con la versión 2.0.0 de tensorFlow-GPU. Finalmente, se utilizó una GPU *NVIDIA GeForce GTX 650* para hacer uso de la funcionalidad de estas

librerías, que permiten un rápido entrenamiento de las RN, comparado con lo que pasaría de usarse una CPU.

4.3.1. Redes neuronales utilizadas.

Un esquema de las RN utilizadas se presenta en la Figura 4.3. Se observa que se componen de diversas capas, con y sin pesos. El diseño de las RN fue hecho para que realizaren las operaciones descriptas en el capítulo 2 para las obtener las predicciones, incorporando adicionalmente la posibilidad cálculo de R_t con las propuestas de la sección 4.2. Para su entrenamiento, se usan los datos totales de caso en Argentina, sin distinguir por provincias o localidades, desde el primer caso reportado hasta el día 1 de enero de 2021, como en el resto de esta tesis.

En primer lugar, las redes cuentan con tres capas de entrada.

Una de ellas corresponde a los datos para el denominador del cálculo de R_t para todos los valores posibles de t, es decir, el período temporal en estudio. En el caso que se consideren un R_t que requiere tres coeficientes, se ingresan los datos N_{t-6}, N_{t-5} y N_{t-4}. Para el caso de los cinco coeficientes, se adjuntan también los datos N_{t-3} y N_{t-2}. Así, esta entrada termina siendo una matriz con la cantidad de días estudiados como filas y 3 o 5 columnas, según el caso.

La segunda entrada a la red es similar a la primera, pero contiene los datos necesarios para el numerador del cálculo de R_t. Estos son los valores de N_{t-1}, N_t y N_{t+1} para todo día t. De esta manera, es representada por una matriz con la misma cantidad de filas que en el caso anterior, y 3 columnas.

Finalmente, se ingresan a la red los valores de A_t^{14} para días t validos, es decir, tales que $t > 14$ días desde el inicio de la epidemia.

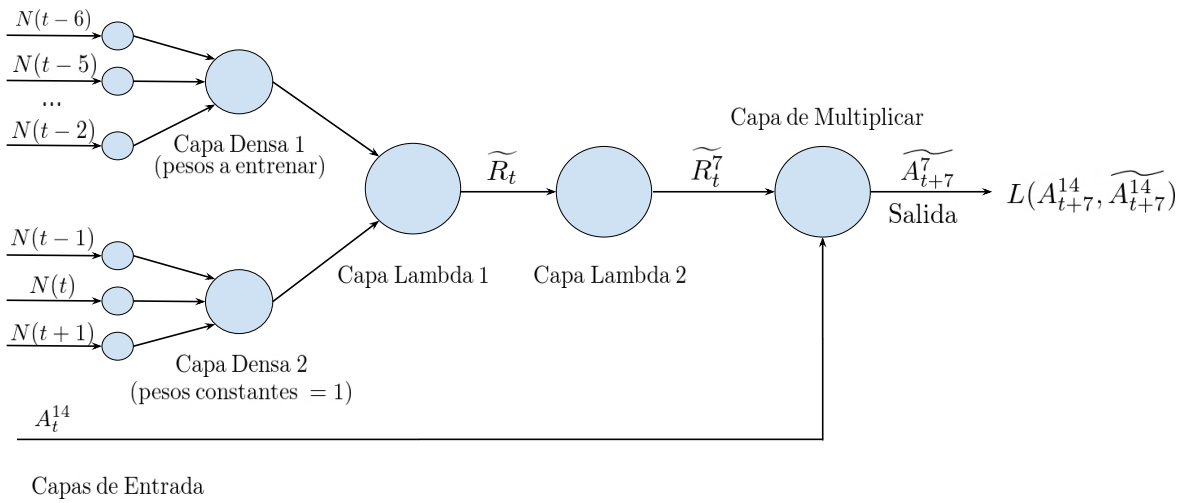


Figura 4.3: Esquema de las redes neuronales utilizadas para la optimización de los pesos de los cálculos de R_t.

Luego de las entradas, se tienen dos capas densas [12]. Una de ellas está conectada a la primera entrada de la red y tiene asignados 3 o 5 pesos a entrenar, según corresponda. Estos pesos están limitados a ser positivos y se inicializan para imitar el cálculo de R_t sin optimizar. De esta manera, la salida de esta capa corresponde al denominador de R_t. La segunda capa densa esta conectada a la segunda entrada, y tiene asignados 3 pesos que no se entrenan, sino que son constantes iguales a 1. La salida de la misma es el numerador de R_t. Siendo las salidas de las capas densas el denominador y el numerador de R_t, estos resultados van a otra capa que realiza la división como corresponda y limita el valor del número reproductivo estimado $\widetilde{R_t}$ para que sea menor a 4. Esta capa recibe el nombre de capa Lambda 1, y es sucedida por la capa Lambda 2. Esta realiza el promedio necesario para calcular un posible valor optimizado de R_t^7 al que llamamos $\widetilde{R_t^7}$ multiplicando la salida de Lambda 1 por una matriz adecuada. Las capa de tipo Lambda son provistas por Keras para crear capas que realicen operaciones deseadas por el usuario y que no estén disponibles en otras capas de la librería [13].

Finalmente, la capa de Multiplicar es la última [14]. Recibe $\widetilde{R_t^7}$ y la tercera entrada a la red, es decir, los valores de A_t^{14}. Entonces, hace la multiplicación valor a valor de los mismos para cada t. En consecuencia, se tiene como resultado un vector de los valores A_{t+7}^{14}. Este vector constituye la salida de la red y, a partir de el, se calcula la función de costo $L(A_t^{14}, \widetilde{A_t^{14}})$ correspondiente. Como función de costo, se ha elegido el error absoluto promedio entre los valores reales A_t^{14} y el resultado de la red $\widetilde{A_t^{14}}$. A través del cálculo de gradientes de la función de costo, se modifican los pesos entrenables de la red, es decir, los pesos de la capa Densa 1 utilizando el optimizador Adam [15]. Se utiliza este optimizador con los parámetros por defecto provistos por Keras, excepto por el *learning rate* o tasa de aprendizaje, que toma un valor $lr = 5e-3$. El número de épocas fue de 3000 para la RN con 3 pesos y 4000 para la RN con 5 pesos. Para calcular los pesos y su error la red se entrena 10 veces obteniéndose en cada ocasión resultados diferentes dada la aleatoriedad en as librerías Keras, Numpy [16] y TensorFlow. Así, cada peso se calcula como el promedio y se le asigna como error la desviación estándar del promedio.

4.3.2. Resultados

Resultados con 3 pesos.

Primero, se analizó la posibilidad de mejorar las predicciones a partir de encontrar pesos óptimos para calcular R_t^7 según la ecuación 4.1. En la Figura 4.4 se presenta un ejemplo de la evolución de la función de costo a lo largo de las 3000 épocas de entrenamiento, tras entrenar la RN descripta anteriormente.Se observa como, para los casos confirmados totales de Argentina, el error absoluto promedio total va disminuyendo hasta un mínimo. Esto quiere decir que se encuentran pesos aproximados que

minimizan el error para los datos alimentados a la RN. Como se describió previamente, se crea y entrena una RN con el diseño adecuado 10 veces.

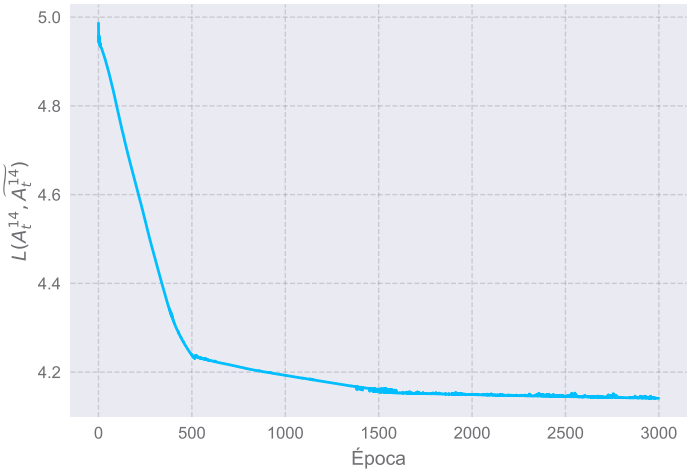


Figura 4.4: Función de costo $L(A_t^{14}, \widetilde{A_t^{14}})$ en función del número de épocas al entrenar una vez una RN de 3 pesos.

Luego, se obtienen los pesos a_3, b_3, c_3 realizando un promedio. En este caso, se obtienen valores $a_3 = 2{,}80(3)$, $b_3 = 0{,}16(3)$ y $c_3 = 0{,}05(1)$. Se observa que el valor de $a_3 > b_3 > c_3$, indicando que la influencia sobre los casos en los días $t-1, t, t+1$ de los infectados en $t-6$ es mayor que en $t-5$ y esta, a su vez, es mayor que para $t-4$.

Con estos pesos, podemos calcular R_t^7 según la ecuación 4.1. Estos pesos fueron optimizados para predecir los casos de todo el país, sin distinguir entre localidades. Para analizar si funcionan para predecir los casos en diferentes regiones, se computa la diferencia entre el error absoluto promedio con el cálculo con los pesos optimizados de R_t^7 y el calculo tradicional, separando los datos por localidades. De esta manera, si el valor de la diferencia es negativo, significa que la predicción original es menos exacta que la realizada a partir de los pesos de la RN. En cambio, si la diferencia es positiva, la predicción original resulta ser mejor.

Una lista con diferentes poblaciones y la diferencia del error absoluto promedio de las predicciones se presenta en la Tabla 4.1. Las abreviaturas de las localidades son las mismas que las utilizadas previamente en este capítulo, y se detallan en el apéndice A. De esta tabla, se desprende que la mayoría de las diferencias resultan ser negativas. Esto quiere decir que las predicciones originales de casos, como propuestas en el capítulo 2, pueden mejorarse utilizando los tres pesos obtenidos a partir del entrenamiento de RN.

Resultados con 5 pesos.

A continuación, se analizó hallar los 5 pesos óptimos para el calculo de R_t^7 según la ecuación 4.2. Para ello, se entrenaron las RN correspondientes durante 4000 épocas. Un ejemplo de la evolución de la función de costo a lo largo de las épocas se presenta

Localidad	Diferencia entre predicciones
Argentina	-0.68 (8)
BRC	5.9 (7)
BA	-0.74 (8)
CABA	-0.9 (2)
CH	0.6 (2)
CB	-1.3 (2)
ER	-0.1 (2)
JY	-1.0 (4)
LR	0.7 (7)
MZ	-1.7 (2)
NQ	-3.9 (7)
RN	0.1 (1)
SA	-1.2 (2)
SC	0.9 (4)
SF	-2.5 (2)
TF	-5 (2)
TU	-1.6 (3)

Tabla 4.1: Diferencia entre el error absoluto promedio de la predicción original y de la predicción con los pesos resultantes de la RN, para R_t^7 calculado según la ecuación 4.1.

en la Figura 4.5. Al igual que en el caso anterior, se observa que para los datos de Argentina se llega a un mínimo de la función de perdida a partir de la modificación de los pesos iniciales.

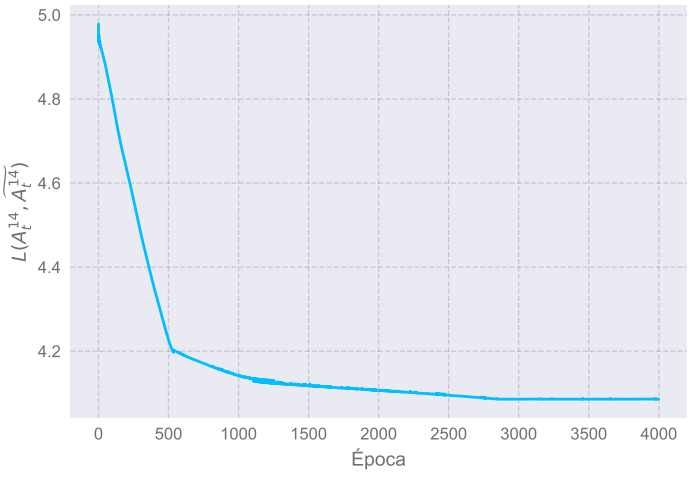


Figura 4.5: Función de costo $L(A_t^{14}, \widetilde{A_t^{14}})$ en función del número de épocas al entrenar una vez una RN de 5 pesos.

En este caso, se obtienen los pesos $a_5 = 2{,}894(5)$, $b_5 = 0$, $c_5 = 0$, $d_5 = 4 \cdot 10^{-5}(2)$ y $e_5 = 1{,}09 \cdot 10^{-1}(5)$. Estos tienen una relación menos evidente entre ellos que los pesos anteriores, y su posible interpretación a nivel biológico no es evidente. Por un lado, a_5 toma un valor similar al caso anterior, pero b_5 y c_5 son nulos. En cambio, el valor del

peso e_5, que corresponde al día $t-2$, tiene un valor no nulo. No resulta evidente una interpretación para estos resultados tales que los infectados al día $t-6$ tienen gran influencia sobre los casos secundarios, los correspondientes a los días $t-5, t-4, t-3$ no la tienen, pero los casos en $t-2$ sí.

Para estos pesos, en la Tabla 4.2 se presentan las diferencias entre los errores absolutos promedios de las predicciones originales y la predicción con los pesos optimizados. En líneas generales, se observa que hay bastantes localidades con diferencias negativas, lo cual indica una mejora respecto de las predicciones originales. Sin embargo, si se comparan con los valores de la Tabla 4.1, se observa que los resultados a partir de la utilización de 3 pesos resultan ser óptimos.

Localidad	Diferencia entre predicciones
Argentina	-0.66 (9)
BRC	7 (1)
BA	-0.69 (7)
CABA	-0.8 (2)
CH	0.8 (1)
CB	-1.3 (2)
ER	0.0 (1)
JY	-0.9 (3)
LR	1.1 (5)
MZ	-1.7 (2)
NQ	-4.2 (8)
RN	0.3 (2)
SA	-1.3 (2)
SC	1.3 (3)
SF	-2.7 (2)
TF	-5 (2)
TU	-1.5 (4)

Tabla 4.2: Diferencia entre el error absoluto promedio de la predicción original y de la predicción con los pesos resultantes de la RN, para R_t^7 calculado según la ecuación 4.2.

4.4. Conclusiones

En este capítulo, examinamos posibilidades para optimizar las herramientas presentadas en el capítulo 2 para predecir el número de casos confirmados en función del tiempo.

Primero, se buscó numéricamente un retraso r óptimo a introducir en la ecuación

2.5 en lugar de la ventana de 7 días utilizada en la bibliografía. Se halló que el error en las predicciones es mínimo para $r = 6$. Adicionalmente, se destaca que el hecho de que exista un r óptimo indica la existencia de un tiempo característico de la evolución de la enfermedad.

Por otra parte, se propuso modificar la manera de calcular el número R_t^7 con dos ecuaciones: una con tres pesos y otra con cinco pesos que cuantifican la influencia de los casos primarios $t-6, t-5, t-4, t-3, t-2$ sobre los casos secundarios $t-1, t, t+1$. Para hallar pesos óptimos que minimicen el error de las predicciones, se diseñaron y entrenaron RN utilizando la librería Keras de Python. Se encontraron tres y cinco pesos óptimos correspondientes a las ecuaciones 4.1 y 4.2 respectivamente. Entre ambas estrategias, la que tiene tres pesos resulta ser la óptima. Por una parte, es la que tiene mejores resultados numéricos. Adicionalmente, este modelo tiene una menor cantidad de parámetros a ajustar, y sus valores tienen una interpretación más clara.

Capítulo 5

Modelo SIR y SIR modificado

El objetivo de este capítulo es introducir y estudiar modelos matemáticos para el modelado de enfermedades infecciosas. En la sección 5.1 se presenta el conocido modelo SIR, uno de los modelos más simples para modelar la propagación de enfermedades infecciosas. En la sección 5.4 se presenta una modificación del mismo. En el caso de ambos modelos, se utilizan las herramientas del capítulo 2 para analizar sus resultados y se los contrasta con propiedades de poblaciones reales.

5.1. Modelo SIR sin demografía clásico

Los modelos más simples para modelar enfermedades agudas (es decir, que causan infección por un periodo de tiempo corto sucedido por inmunidad) son los modelos del tipo SIR. Este formalismo consiste en clasificar a los individuos de una población en estudio en tres grupos: susceptibles (S), infectados (I) y recuperados (R). En el caso más simple, si se ignoran todos los aspectos demográficos y se considera inmunidad de por vida, se pueden dar dos transiciones definidas entre grupos: $S \to I$ y $I \to R$. La primera corresponde a un individuo susceptible que se transforma en infectado por algún mecanismo de transmisión de la enfermedad. En cambio, la segunda transición corresponde a una persona infectada que, tras un período infeccioso determinado se recupera y se torna inmune. Este último podría ser variable, pero usualmente se realiza la simplificación de considerarlo constante.

En la Figura 5.1 se presenta un diagrama de flujo que representa conceptualmente un modelo de este tipo. Allí, las flechas sólidas negras representan los movimientos posibles entre clases, mientras que la flecha gris rayada representa que la tasa a la que se infectan individuos susceptibles puede depender de la enfermedad infecciosa.

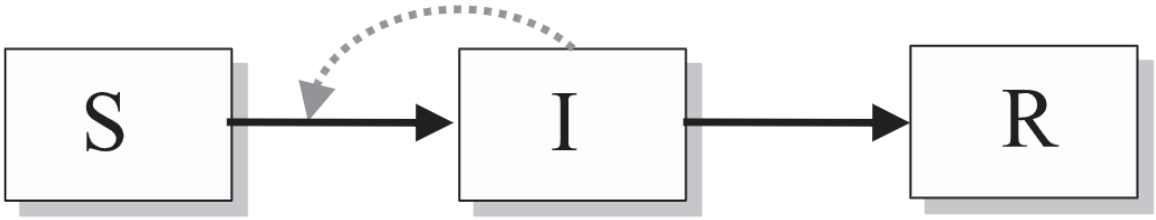


Figura 5.1: Diagrama de flujo representativo de un modelo SIR clásico [2].

Considérese entonces una población de N individuos, de los cuales X son susceptibles, Y están infectados y Z se han recuperado, tal que $X + Y + Z = N$. Alternativamente, se pueden definir las variables: $S = X/N$, $I = Y/N$ y $R = Z/N$, de manera que $S + I + R = 1$. Utilizando estas cantidades, el modelo previamente introducido se describe matemáticamente con el sistema dinámico [2]:

$$\frac{dS}{dt} = -\beta SI \tag{5.1a}$$

$$\frac{dI}{dt} = \beta SI - \gamma I \tag{5.1b}$$

$$\frac{dR}{dt} = \gamma I, \tag{5.1c}$$

donde γ es la inversa del período infeccioso, es decir, es la tasa de recuperación. Por otra parte, β es la transmisión o tasa de contagio que multiplica a las proporciones de susceptibles y de infectados, ya que las infecciones se provocan por contacto entre estas dos poblaciones. Estas ecuaciones tienen condiciones iniciales $s(0) \geq 0$, $i(0) \geq 0$ y $r = 1 - s(0) - i(0)$. A pesar de ser uno de los modelos más simples para tratar enfermedades infecciosas, no es resoluble analíticamente.

Si bien este modelo es demasiado simple para aplicar directamente a nuestro problema, podemos utilizarlo como punto de partida para entender cómo funcionan los modelos de enfermedades infecciosas en campo medio. Adicionalmente, sirve para ilustrar varios principios epidemiológicos cualitativamente, como el fenómeno de umbral o de *threshold*. Este consiste en que es necesaria una mínima fracción de susceptibles inicial $s(0)$ para que una enfermedad infecciosa se propague. Esto se puede ver fácilmente reescribiendo la ecuación 5.6b como:

$$\frac{dI}{dt} = I(\beta S - \gamma), \tag{5.2}$$

de donde se observa que para que $\frac{dI}{dt}(0) > 0$ debe suceder que $S(0) > \frac{\gamma}{\beta}$. De lo contrario, $\frac{dI}{dt}(0) < 0$ y la epidemia finaliza. La cantidad $\frac{\gamma}{\beta}$ suele llamarse la tasa relativa de disipación y el fenómeno de umbral puede interpretarse como la necesidad de que $\frac{\gamma}{\beta}$ sea

lo suficientemente pequeña para permitir la propagación de la enfermedad. A partir esta tasa se define el número reproductivo básico R_0 como su inversa. Este número fue inicialmente introducido y definido en el capítulo 2. El fenómeno del umbral se puede interpretar en base a esta cantidad como: si inicialmente tenemos una población totalmente suceptible, la enfermedad se propagará si y solo si $R_0 > 1$.

5.2. Análisis del modelo SIR

Se procedió a analizar el comportamiento de las cantidades R_t^7 y de los diagramas de riesgo que surgen del modelo SIR tradicional. Para ello, se resolvieron las ecuaciones 5.1 utilizando el método de Euler para parámetros $N = 10000$, $\beta = 0{,}15$ y $gamma = 0{,}075$. Los valores de β y γ son tales que $R_0 = 2$, cercano al número reproductivo aceptado para la COVID-19 en la bibliografía. El valor de γ es $1/14$, donde 14 es el número aceptado de días hasta la recuperación de un infectado. En cuanto a las condiciones iniciales, se eligió $I(0) = 1/N$ y $R(0) = 0$, siendo entonces $S(0) = 1 - I(0) - R(0)$. Se realizaron 250 pasos temporales del algoritmo, separados por un $dt = 1$. Los códigos para la resolución de las ecuaciones fueron de elaboración propia, utilizando el lenguaje de programación Python. La evolución temporal de las cantidades S, I, R se presenta en la Figura 5.2, y es la esperable para un modelo SIR tradicional.

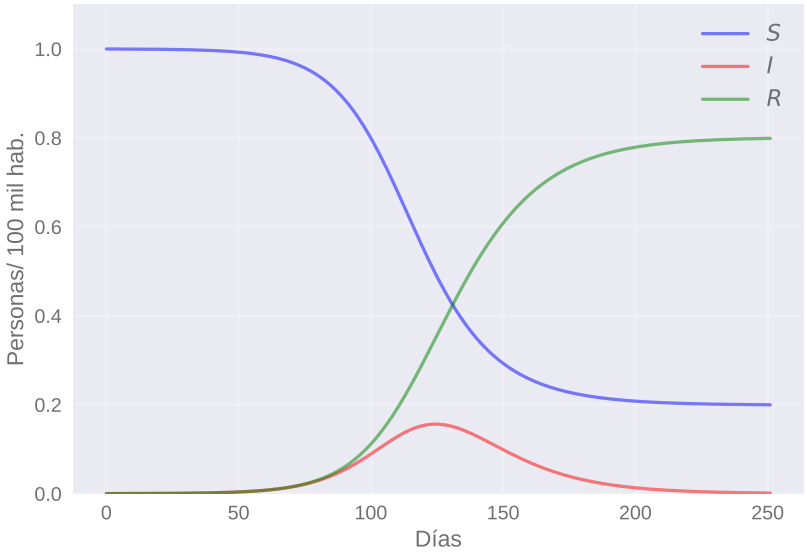
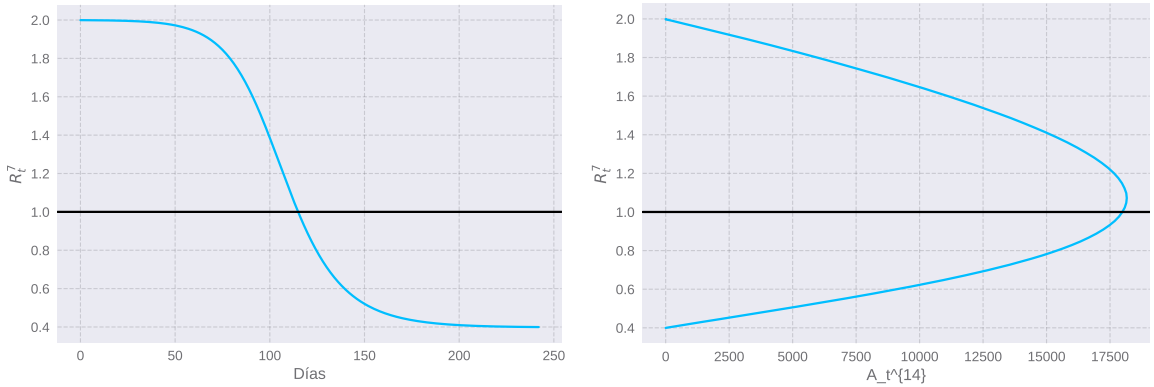


Figura 5.2: Evolución temporal de las cantidades S,I,R en el modelo SIR tradicional.

Por otra parte, en la Figura 5.3 se presenta la evolución temporal de R_t^7 y en la Figura 5.3b se grafica el diagrama de riesgo correspondiente a este modelo. De la primera figura, se observa que el comportamiento de R_t^7 no es similar al real, es decir, este número reproductivo no se acerca de forma oscilatoria a la unidad. En cambio, si bien parece estancarse cerca de un valor por un determinado periodo de tiempo,

el acercamiento no es oscilatorio. Estas diferencias también se ven reflejadas en el diagrama de riesgo correspondiente.



(a) Evolución temporal de R_t^7 para el modelo SIR tradicional.

(b) Diagrama de riesgo para el modelo SIR tradicional.

Figura 5.3: Evolución temporal de las cantidades S,I,R para el modelo SIR tradicional

5.3. Primer modelo SIR sin demografía modificado

Hasta ahora, no hemos podido establecer similitudes significativas entre lo que ocurre en ciertas poblaciones y los resultados del modelo SIR. Para intentar reproducir la dinámica en ciudades reales, como las estudiadas previamente en esta tesis, se propone una modificación del modelo SIR. La misma está descripta por las ecuaciones:

$$\frac{dS}{dt} = -\tilde{\beta} S I \tag{5.3a}$$

$$\frac{dI}{dt} = \tilde{\beta} S I - \gamma I \tag{5.3b}$$

$$\frac{dR}{dt} = \gamma I, \tag{5.3c}$$

con $\tilde{\beta}$ que reemplaza la constante β del modelo SIR tradicional por:

$$\tilde{\beta} = \beta_0 - \alpha \frac{dI_d}{dt}, \tag{5.4}$$

donde $I_d = I(t-d)$ y tal que β_0, α y d son mayores o iguales a cero. Si reemplazamos la ecuación 5.6b en la ecuación 5.7, puede despejarse $\tilde{\beta}$ como:

$$\tilde{\beta} = \frac{\beta_0 + \alpha\, \gamma\, I_d}{1 + \alpha\, I_d\, S_d} \tag{5.5}$$

donde $S_d = S(t-d)$.

Se observa que si $\alpha = d = 0$, se recupera el modelo SIR anteriormente descripto. En caso contrario, el término responsable de la transmisión $\tilde{\beta}$ no es constante, sino que

cambia en función del tiempo según $\frac{dI_d}{dt}$, la derivada del número de infectados a un día $t-d$. La relación entre $\tilde{\beta}$ y $\frac{dI_d}{dt}$ está mediada por la constante α. Dado que $\alpha \geq 0$, si el número de infectados baja a un día $t-d$ la transmisión aumenta y viceversa. Esto intenta modelar una modificación del comportamiento de la población según el aumento o disminución de los infectados. En ocasiones se observa que cuando se percibe un aumento pronunciado de casos la gente reacciona aislándose o tomando medidas de protección, mientras que si los casos disminuyen las personas tienden a relajar los cuidados tomados. Esta reacción puede ser instantánea si $d = 0$, o retardada si $d > 0$. Otros tipos de modelos han sido estudiados en la bibliografía para tener en cuenta este tipo de efectos de reacción social que pueden aumentar o disminuir la tasa de contagio [17].

5.3.1. Análisis del primer modelo SIR modificado

Caso $d = 0$

En primer lugar, se analizo el modelo SIR modificado con reacción instantánea, es decir, para $d = 0$. Se utilizaron como parámetros $N = 10000$, $\beta_0 = 0{,}15$ y $\gamma = 0{,}075$. Estos son análogos a los empleados en el estudio del modelo SIR tradicionales. Adicionalmente, se tomo un valor de $\alpha = 0{,}08\,N$. Las condiciones iniciales son idénticas a las empleadas anteriormente, es decir, $I(0) = 1/N$ y $R(0) = 0$. O sea, corresponde a introducir una persona enferma en una población donde no hay personas inmunes. Para la resolución, se utiliza nuevamente el método de Euler con $dt = 1$. En este caso, se itera por 800 pasos del algoritmo. Los resultados obtenidos de S, I, R en función del tiempo se presentan en la Figura 5.4.

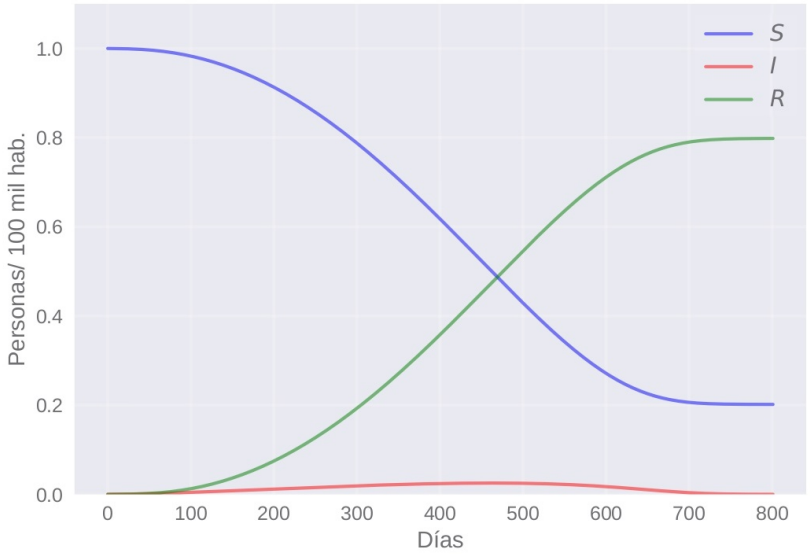
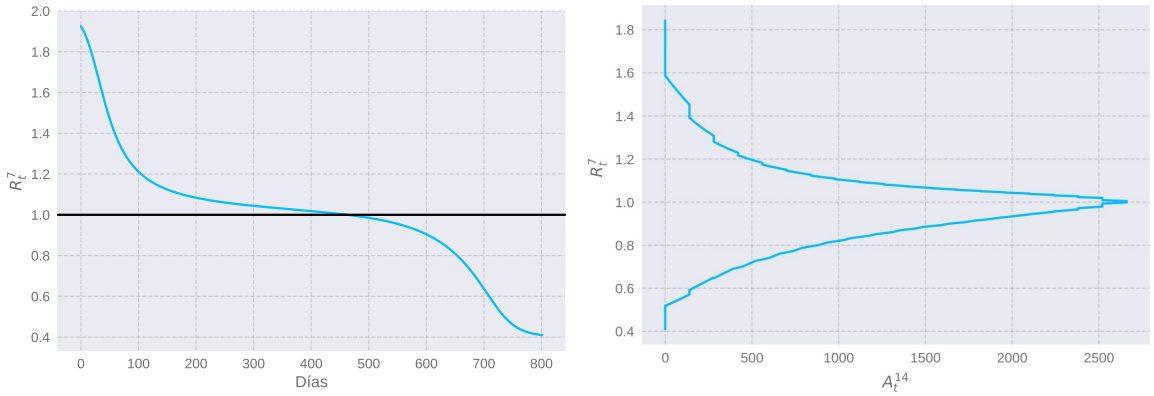


Figura 5.4: Evolución temporal de las cantidades S,I,R en el modelo SIR modificado con $d = 0$.

De estos resultados es posible calcular R_t^7, el cual se grafica en la Figura 5.4. A diferencia de lo que sucede en el caso anterior, se observa que R_t^7 no se acerca y permanece cerca de la unidad como sí se tiene de las observaciones reales. Si bien al principio toma valores grandes, mayores a 2, luego pasa por la unidad antes de anularse y que la epidemia termine. Por otra parte, en la Figura 5.7b se presenta el diagrama de riesgo correspondiente a este modelo. Nuevamente, es claro que este diagrama no se asemeja particularmente a los diagramas de riesgo de ciertas localidades reales.



(a) Evolución temporal de R_t^7 para el modelo SIR modificado con $d = 0$.

(b) Diagrama de riesgo para el modelo SIR modificado con $d = 0$.

Figura 5.5: Evolución temporal de las cantidades S,I,R para el modelo SIR modificado con $d = 0$.

Caso $d \neq 0$

A continuación, se procedió a introducir un retraso $d = 2$ al modelo SIR modificado, como se indica en el sistema de ecuaciones 5.4. En cuanto al resto de los parámetros utilizados, son los mismos que en la sección anterior 5.3.1. Los resultados de S, I, R en función del tiempo se presentan en la Figura 5.6.

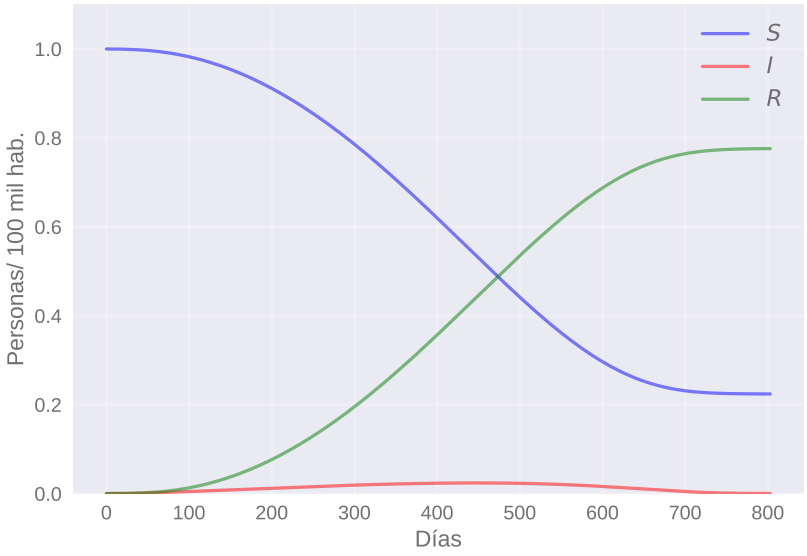
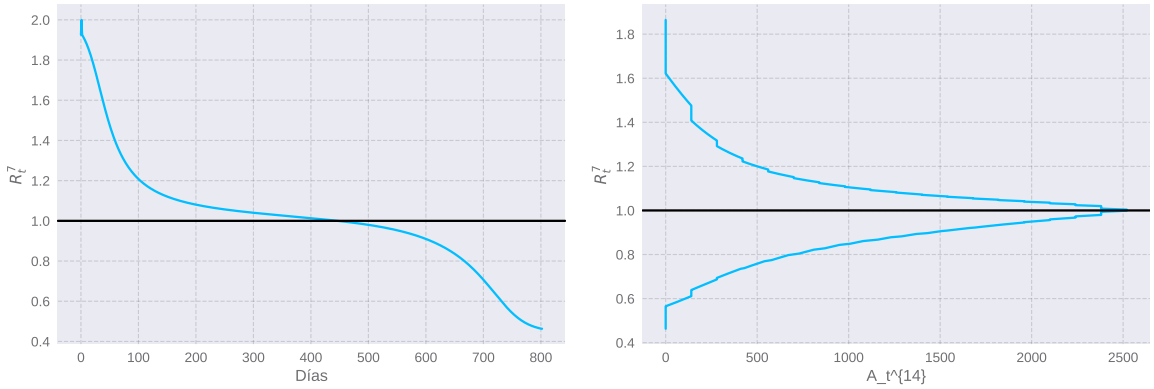


Figura 5.6: Evolución temporal de las cantidades S,I,R en el modelo SIR modificado con $d = 2$.

Adicionalmente, en la Figura 5.7a, se grafica la evolución temporal del número reproductivo R_t^7. Se observa que su comportamiento en función del tiempo es análogo al de la Figura 5.5a, correspondiente al modelo con reacción instantánea. Lo mismo sucede con el diagrama de riesgo correspondiente, el cual se visualiza en la Figura 5.7b. Este es similar cualitativamente al de la Figura 5.7b, confeccionado para $d = 0$.



(a) Evolución temporal de R_t^7 para el modelo SIR modificado con $d = 2$.

(b) Diagrama de riesgo para el modelo SIR modificado con $d = 2$.

Figura 5.7: Evolución temporal de las cantidades S,I,R para el modelo SIR modificado con $d = 2$.

Cabe aclarar que, cambiando el valor de d alrededor de la decena, no se observan grandes diferencias en el comportamiento cualitativo de R_t^7 ni en los diagramas de riesgo. En cambio, para valores mucho mayores se tiene que estos comienzan a parecerse a lo que sucede en ausencia de una reacción. Es decir, en ninguno de los casos es posible reproducir una dinámica como la que se observa en ciudades reales. Lo mismo sucede cambiando el resto de los parámetros como, por ejemplo, modificando el valor de α.

5.4. Segundo modelo SIR sin demografía modificado

Tras no hallar un comportamiento diferente con el primer modelo SIR modificado, proponemos estudiar un nuevo modelo tal que la tasa de contagio se relacione directamente con la cantidad de infectados en vez de con la derivada de los mismos. El nuevo modelo propuesto corresponde al sistema de ecuaciones:

$$\frac{dS}{dt} = -\tilde{\tilde{\beta}}SI \tag{5.6a}$$

$$\frac{dI}{dt} = \tilde{\tilde{\beta}}SI - \gamma I \tag{5.6b}$$

$$\frac{dR}{dt} = \gamma I, \tag{5.6c}$$

con una tasa de contagio $\tilde{\tilde{\beta}}$ que reemplaza a la tradicional contante por la expresión:

$$\tilde{\beta} = \beta_0 - \alpha\left(\frac{I_d}{I_0}\right)^2 \tag{5.7}$$

donde I_0, β_0, d y α son mayores o iguales a cero. El nuevo parámetro I_0 provee una escala con la cual comparar la cantidad de casos I_d. Al igual que en el caso anterior, el modelo SIR tradicional se recupera en el caso que $\alpha = 0$. En caso contrario, se observa que la tasa de contagio disminuye con un aumento de la cantidad de casos, y viceversa. Esta reacción puede nuevamente ser instantánea o retardada, dependiendo de si $d = 0$ o $d > 0$. A continuación, analizamos ambos casos, utilizando los mismos parámetros y las mismas condiciones iniciales que en el estudio del primer modelo modificado. Para la resolución de las ecuaciones, usamos el método de Euler con $dt = 1$.

5.4.1. Análisis del segundo modelo SIR modificado

Caso $d = 0$

En primer lugar, estudiamos el segundo modelo SIR modificado sin delay, es decir, con $d = 0$. Habiendo especificado previamente los parámetros y las condiciones iniciales correspondientes, presentamos el gráfico de las series temporales S, I, R en la Figura 5.8. Se integraron las ecuaciones desde $t = 0$ hasta $t = 900$, intervalo que contempla el desarrollo total del brote epidémico. Se observan que las curvas obtenidas son similares a las de otros modelos, es decir, existe un brote desde las condiciones iniciales tal que la cantidad de casos tiene un máximos y luego decrece hasta llegar a ser nulo.

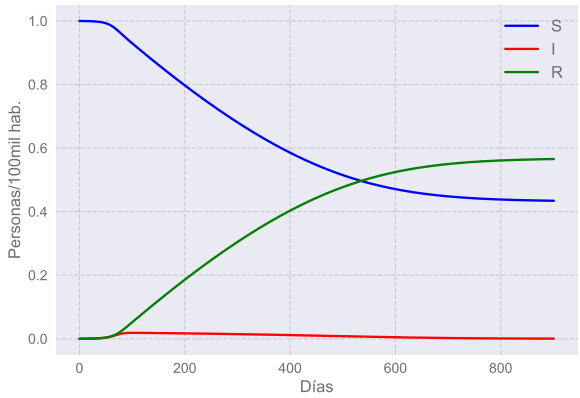
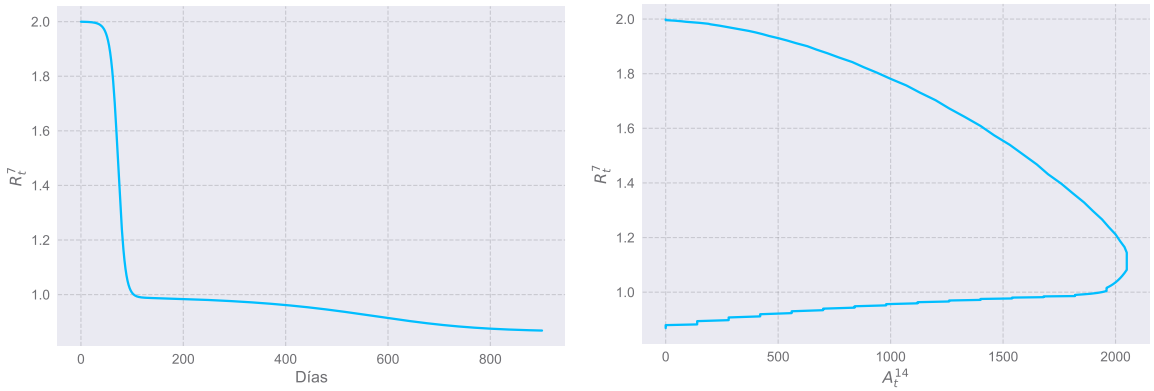


Figura 5.8: Evolución temporal de las cantidades S,I,R en el modelo SIR modificado con $d = 2$.

Adicionalmente, presentamos los gráficos de la serie temporal del número reproductivo R_t^7 y del diagrama de riesgo en la Figura 5.9. En este caso, se tiene que el número reproductivo no se acerca a la unidad oscilando, aunque parece permanecer cerca por un período de tiempo. De todas maneras, ni esta serie temporal ni el diagrama de riesgo parecen aproximarse cualitativamente a lo observado en la realidad.



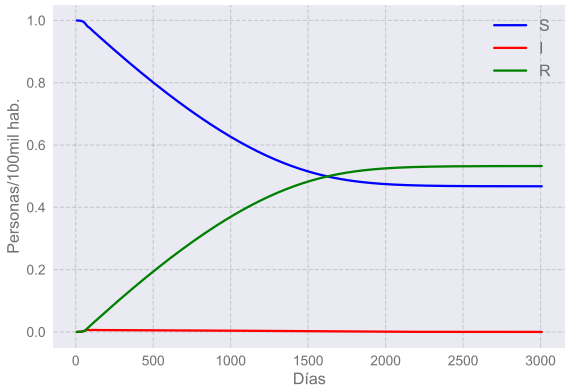
(a) Evolución temporal de R_t^7 para el modelo SIR modificado con $d = 2$.

(b) Diagrama de riesgo para el modelo SIR modificado con $d = 2$.

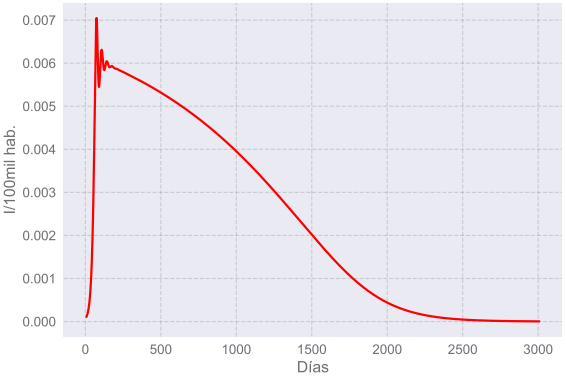
Figura 5.9: Evolución temporal de las cantidades S,I,R para el modelo SIR modificado con $d = 2$.

Caso $d \neq 0$

Por otra parte, realizamos simulaciones para un tiempo de retraso de $d = 7$ días, con los mismos parámetros y condiciones iniciales del caso anterior. Las series temporales S, I, R se presentan en la Figura 5.10a, donde parecen ser cualitativamente similares a las estudiadas anteriormente con el modelo tradicional y el primer modelo modificado. Sin embargo, la serie de infectados tiene unas oscilaciones al comienzo del brote, producto de la reacción social. Para observarlas correctamente, graficamos estos datos separados en la Figura 5.10b. Allí, se visualiza un período de crecimiento vertiginoso de los casos, precedido por oscilaciones y un decremento suave hasta el fin del brote.

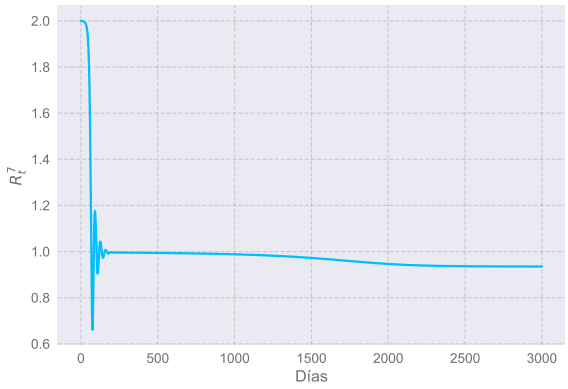


(a) Evolución temporal de R_t^7 para el modelo SIR modificado con $d = 2$.

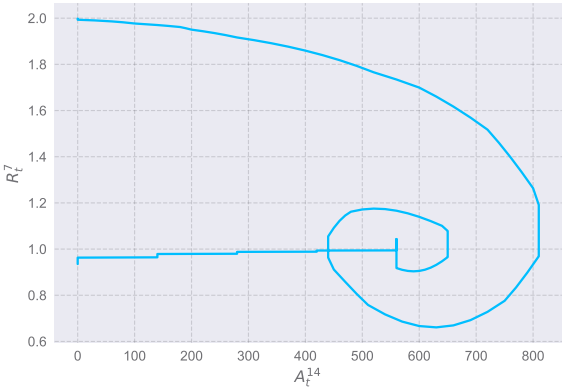


(b) Diagrama de riesgo para el modelo SIR modificado con $d = 2$.

Figura 5.10: Evolución temporal de las cantidades S,I,R para el modelo SIR modificado con $d = 2$.



(a) Evolución temporal de R_t^7 para el modelo SIR modificado con $d = 2$.



(b) Diagrama de riesgo para el modelo SIR modificado con $d = 2$.

Figura 5.11: Evolución temporal de las cantidades S,I,R para el modelo SIR modificado con $d = 2$.

De esta manera, resulta interesante analizar qué sucede con el número reproductivo y el diagrama de riesgo, lo cuales se grafican en la Figura 5.11. Por un lado, se observa que el número reproductivo tiene oscilaciones mientras se acerca a la unidad, una cualidad que se observa en la realidad, al igual que permanecer un tiempo considerable en su vecindad. Esto nos recuerda a lo que sucede en localidades como BA o CABA. Sin embargo, cabe destacar que las oscilaciones en el número de infectados se encuentran directamente relacionadas con oscilaciones en el número reproductivo, algo que no es observado en la realidad. Para los datos reales, la curva de infectados puede no presentar oscilaciones, aunque tengamos oscilaciones en el número reproductivo calculado. Por otra parte, observando el diagrama de riesgo de la Figura 5.11b es claro este no guarda semejanza con los de las localidades mencionadas.

Finalmente, cabe aclarar que se probaron diferentes tipos de modelos de reacción social. Especialmente, modelos similares al segundo, modificando la forma en la que la

reacción social depende del número de infectados. En todos los casos, no fue posible obtener resultados cualitativamente similares a los datos reales. Especialmente, cabe destacar que, si se obtienen oscilaciones, estas son oscilaciones sincronizadas de la cantidad de infectados con el número reproductivo.

5.5. Conclusiones

En este capítulo hemos introducido tres modelos del tipo SIR sin demografía para intentar reproducir el comportamiento de las poblaciones reales previamente analizadas. Sus ecuaciones se resolvieron utilizando el método de Euler, y utilizando parámetros y condiciones iniciales determinadas. Una vez obtenidos S, I, R en función del tiempo, fue posible calcular la evolución de R_t^7 y confeccionar los diagramas de riesgo correspondientes. El objetivo fue intentar reproducir características cualitativas de la propagación de la COVID-19 observadas en poblaciones reales del país.

En primer lugar, se trabajó con el modelo SIR tradicional, uno de los modelos matemáticos mas simples para enfermedades infecciosas. En este caso, no fue posible reproducir un comportamiento similar a lo que sucede en la realidad.

Luego, se examinaron dos modelos SIR modificados, que tienen en cuenta una posible reacción de una población ante el aumento o descenso de los casos. Se consideró que esta reacción podía ser instantánea o con un retraso d. Primero, se modificó la transmisibilidad β por una $\tilde{\beta}$ que dependiese de la derivada de la cantidad de infectados en función del tiempo. En este caso, no fue posible observar características presentes en poblaciones reales. En segundo lugar, se introdujo un modelo tal que la tasa de contagio dependiese directamente de la cantidad de infectados al cuadrado. En este caso, se destaca la tendencia de R_t^7 de acercarse a la unidad oscilando, así como unas oscilaciones en el número de infectados. Sin embargo, estas no están presentes en los datos oficiales y los diagramas de riesgo obtenidos tampoco se asemejan a los calculados de los mismos.

Capítulo 6

Análisis de la propagación espacial

En los capítulos anteriores, se han presentado diversas herramientas para el análisis de brotes de la COVID-19. Estas son útiles para entender el fenómeno de la propagación en una localidad, una provincia o un país. Sin embargo, no tienen en cuenta la conexión entre diferentes poblaciones afectadas por la epidemia. El objetivo de este capítulo es investigar maneras de cuantificar la propagación de la onda epidémica desde ciertas poblaciones fuentes a otras. En este caso, las poblaciones consideradas son las diferentes provincias de la República Argentina.

En la sección 6.1 se hace una breve introducción sobre la propagación inicial de la COVID-19 en nuestro país. En la sección 6.2 se presentan herramientas clásicas para el análisis de una onda epidémica, como las correlaciones y los lags. En la sección 6.3 se explica cómo se calculan estas cantidades computacionalmente. El resto de las secciones se dedican a diferentes análisis de estos indicadores para obtener información relevante a la propagación de la COVID-19 entre las diferentes provincias.

Todas las abreviaciones de provincias utilizadas en este capítulo pueden observarse en el Apéndice A.

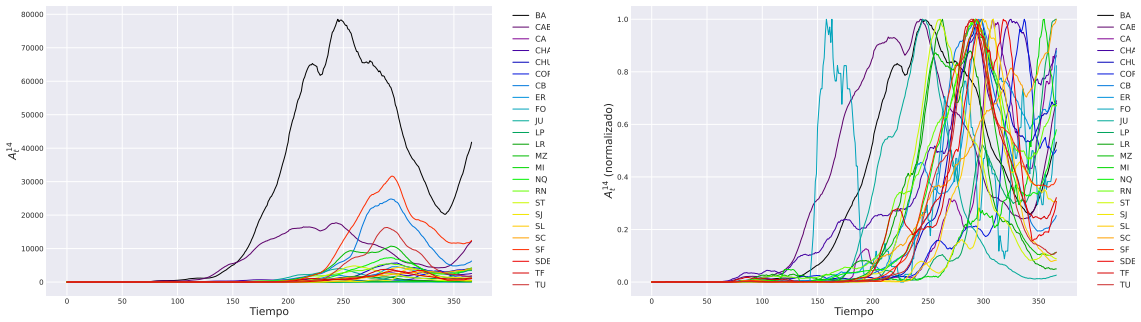
6.1. Propagación de la COVID-19 en Argentina

El virus de la COVID-19 fue originalmente identificado en la ciudad Wuhan, provincia de Hubei, República Popular China, en diciembre del año 2019. Rápidamente, comenzó su propagación por el mundo. En cuanto a Argentina, el 22 de enero de 2020 el Ministerio de Salud emitió un alerta e inició la vigilancia epidemiológica de la enfermedad. Unos días después, el día 3 de marzo de 2020, se detectó el primer caso confirmado en el territorio nacional. Este correspondió a un caso importado proveniente de Europa y perteneció a CABA.

De esta manera, el comienzo de la epidemia en Argentina se debió al ingreso de personas provenientes de países con transmisión activa del virus SARS-COV-2. Inicial-

mente, más de un 50 % de los casos fueron clasificados como importados. Sin embargo, gradualmente fueron incrementándose los casos de contactos estrechos (también llamados conglomerados de casos) y casos de transmisión comunitaria. Para el día 17 de julio de 2020, solo el 0.9 % de los casos eran importados. En cambio, la transmisión comunitaria era responsable por el 51.2 % de los casos y los conglomerados abarcaban al 32.2 % de los mismos [18].

De esta manera, la epidemia se expandió a todo el territorio nacional en el año 2020, siendo algunas provincias más afectadas que otras. Además de la Figura 1.4 que muestra los casos acumulados cada 100mil habitantes de cada provincia, es interesante observar las series temporales correspondientes. Estas proveen información sobre el lugar y fecha de inicio de los brotes de la COVID-19 en nuestro país, así como sobre las provincias que fueron más afectadas. Las mismas pueden observarse en la Figura 6.1, donde se presentan con y sin una normalización a la unidad. Cabe recordar que estas series temporales corresponden hasta el día 1 de enero de 2021, como en el resto de este trabajo. A pesar de existir más datos, no se incluyen con el objetivo de evitar variables como la vacunación que puedan complejizar los modelos matemáticos utilizados.



(a) Series temporales sin normalización. **(b)** Series temporales normalizadas a la unidad.

Figura 6.1: Series temporales de casos, con y sin una normalización a la unidad, para las 24 jurisdicciones argentinas desde el inicio de la pandemia hasta el inicio del 2022.

De estas Figuras, se observa claramente que hay provincias que antecedieron a otras en el surgimiento de un brote. Es decir, se tiene que provincias como CABA y BA tuvieron un fuerte incremento de los casos antes que ocurriera lo mismo con el resto del país. De la Figura 6.1b, cabe destacar que la provincia de Formosa parecer tener un fuerte máximo de casos antes que el resto del país. Sin embargo, esto es contrario a lo que se ha reportado en datos oficiales, según los cuales Formosa no fue una de las provincias con más casos al comienzo de la epidemia. Adicionalmente, diversas fuentes cuestionan el manejo provincial de la epidemia, así como la fiabilidad de los datos reportados [19]. En las secciones siguientes, se harán comentarios adicionales a la hora de estudiar las series temporales en detalle. A continuación, presentaremos las

herramientas necesarias para este análisis.

6.2. Definición de correlaciones y lags entre provincias

Una de las características más importantes en la dinámica espacial de epidemias es la heterogeneidad entre poblaciones. Esta puede referirse tanto a las diferencias espaciales entre poblaciones, como así a las diferencias que puedan emerger a partir de las dinámicas epidemiológicas.

De acuerdo con la bibliografía [2], una medida para cuantificar la heterogeneidad observada es calcular correlaciones entre poblaciones. Estas permiten determinar si las epidemias correspondientes están sincronizadas o no. Si definimos I_i como la proporción de infectados de la población i en función del tiempo, la correlación ρ_{ij} entre la población i y la j se calcula como:

$$\rho_{ij} = \frac{\sum_t (I_i(t) - \overline{I_i})(I_j(t) - \overline{I_j})}{\sqrt{var(I_i)var(I_j)}}, \tag{6.1}$$

donde $\overline{I_i}$ es el promedio y $varI_i$ es la varianza de la serie temporal I_i. De esta ecuación, se observa que si $I_i \propto I_j$, la correlación llega la unidad, que es su valor máximo. En cambio, si ambas series son independientes, la correlación tiene valor nulo. Finalmente, si están fuera de fase, el valor de la correlación es negativa. Esta medida es útil si la heterogeneidad surge de procesos epidémicos que evolucionan de manera estocástica.

En cambio, la heterogeneidad también puede ser causada por ondas epidémicas que viajan de una población a otra. Considerando que los primeros casos de la COVID-19 llegan y se propagan en ciertas provincias antes de llegar a otras, esto resulta más adecuado para el problema en estudio. De esta manera, es más correcto utilizar las correlaciones con lag, calculadas como:

$$\rho_{ij}^{\tau_{ij}} = \frac{\sum_t (I_i(t + \tau_{ij}) - \overline{I_i})(I_j(t + \tau_{ij}) - \overline{I_j})}{\sqrt{var(I_i)var(I_j)}} \tag{6.2}$$

donde τ indica el lag temporal, y $\rho_{ij}^{\tau_{ij}}$ la correlación correspondiente a este lag. Así, puede calcularse un valor de correlación para cada lag posible. Para uno de estos lag, la correlación será la máxima posible. Como convención, adoptaremos que el lag τ_{ij} es negativo si la provincia i antecede a la j, y positivo en el caso contrario. De esta manera si τ_{ij} es negativo, entonces τ_{ji} es positivo. Adicionalmente, es evidente que debe cumplirse que $\tau_{ji} = -\tau ij$.

Cabe aclarar que se adoptan esos valores como los valores adecuados de correlación y lag, y nos referiremos a ellos con estos nombres en el resto de este trabajo.

6.3. Cálculo computacional de correlaciones y de lags entre provincias

Para el cálculo de las correlaciones, se investigaron dos programas posibles en Python: uno que se ejecutaba en CPU y uno ejecutado en GPU. La CPU utilizada tenía un procesador Intel Core i5 9300-H @ 2.40GHz mientras que la GPU correspondió al modelo NVIDIA GeForce GTX 1650. Ambas pertenecieron a una computadora personal. Para el programa en CPU, se utilizó la librería numpy de Python. En cambio, para el GPU, se usó su análogo en paralelo, es decir, la librería cupy.

Como resultado, se obtuvo que la implementación en serie tiene un tiempo de ejecución de $(0{,}09 \pm 0{,}01)$s mientras que el mismo asciende a $(3{,}5 \pm 0{,}6)$s al realizarse en paralelo. Esto se debe a que las series temporales sobre las cuales se calculan las correlaciones tienen alrededor de 360 elementos. Por lo tanto, son muy pequeñas para que la implementación en paralelo provea una aceleración significativa, como puede hacerlo en otros problemas.

Para comprobar esto, generamos vectores aleatorios de una cantidad variable de elementos entre $10 - 10^4$ y calculamos el tiempo que llevaría el cálculo de correlaciones con lag. El resultado puede observarse en la Figura 6.2. De ésta, es claro que la GPU comienza a tener una ventaja en tiempo para vectores de alrededor de 4000 elementos.

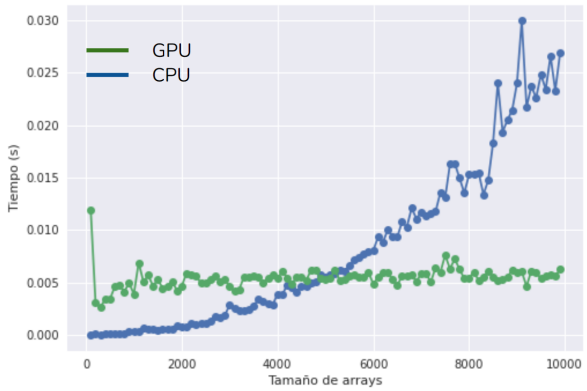


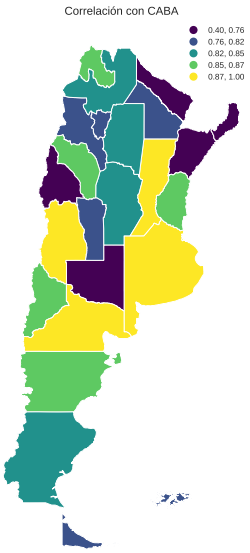
Figura 6.2: Tiempo de cálculo de las correlaciones y de los lags en función del tamaño de las series temporales, utilizando CPU y GPU.

6.4. Representaciones visuales de correlaciones y lags

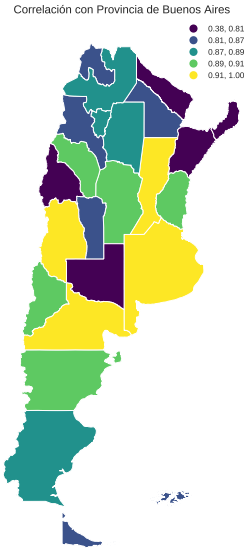
En esta sección, nos dedicaremos a analizar los valores de correlación y de lag obtenidos, utilizando diferentes representaciones visuales. Estos primeros análisis permitieron tener un entendimiento de las características más importantes de estas cantidades, además de proporcionar posibles nuevas líneas de trabajo, al señalar sus limitaciones.

6.4.1. Mapas de correlaciones y de lags con CABA y Provincia de Buenos Aires

El primer tipo de representación visual utilizada fueron los mapas de correlaciones y de lags de una provincia con relación al resto. Elegida una provincia, se tienen 24 valores de correlación y de lags, correspondientes su relación con ella misma y con las otras 23 provincias. En particular, se optó por ver los mapas de correlación y de lags de CABA y BA con relación al resto del país.



(a) Mapa de correlaciones con respecto a CABA.

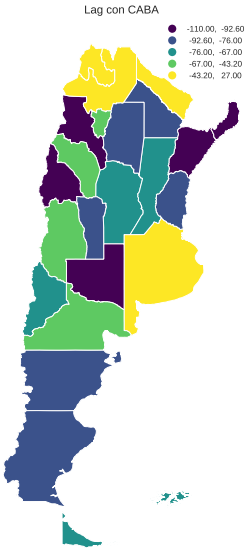


(b) Mapa de correlaciones con respecto a BA.

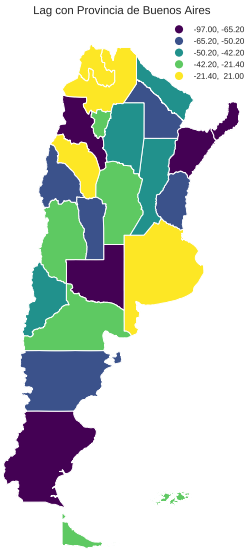
Figura 6.3: Mapas que indican con un código de color las correlaciones de CABA y BA con el resto de las provincias del país.

En la Figura 6.3 se tienen los mapas de correlación de ambas jurisdicciones en

relación al resto del país. De ambas figuras, observamos que las correlaciones de estas provincias con el resto del país toman valores en un rango amplio de $0{,}4 - 1{,}0$. Sin embargo, se tiene que mantienen una correlación por encima de 0,8 con la mayoría de las provincias. Adicionalmente, los lags toman valores en rangos similares para una misma provincia, al analizar su conexión con CABA y con BA.



(a) Mapa de lags con respecto a CABA.



(b) Mapa de correlaciones con respecto a BA.

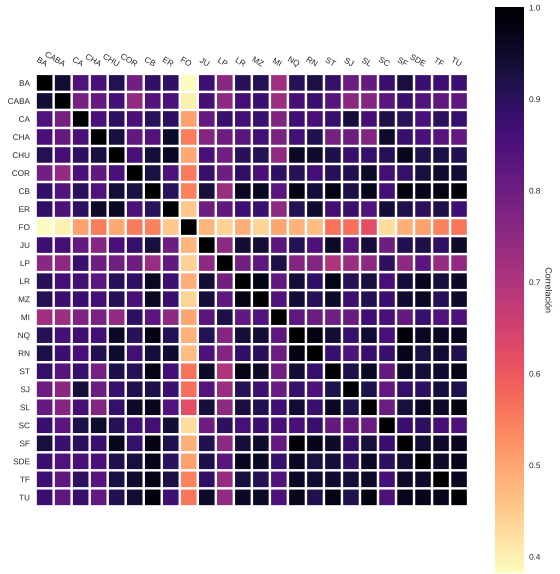
Figura 6.4: Mapas que indican con un código de color los lags de CABA y BA con respecto al resto de las provincias del país.

Por otro lado, en la Figura 6.4 se tienen los mapas de lags correspondientes a CABA y a BA. De estos, se observa que la mayoría de los lags son negativos, lo cual nos indica que CABA y BA se adelantan al resto de las provincias del país en cuestión de casos. El rango de valores tomados por el lag es bastante amplio, siendo de $-100 - 30$ días.

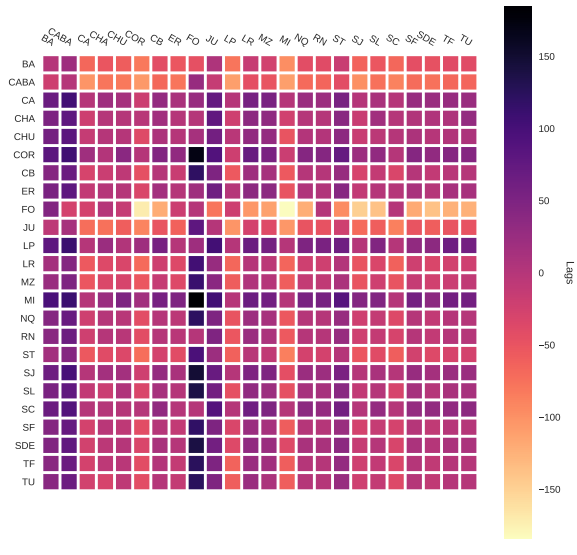
Cabe remarcar que parecen existir varios valores positivos de lags, es decir, provincias que adelantarían a CABA y BA. Esto se puede atribuir en un primer momento a la provincia de Formosa, como hemos destacado previamente al describir las series temporales. Sin embargo, también es interesante que ambos gráficos indican la presencia de otros lags negativos o de pequeño valor, que podrían indicar a las provincias que inmediatamente siguieron a CABA y BA. Incluso, provincias que podrían anteceder a estas jurisdicciones. Estas corresponden a las provincias del Noroeste argentino, como Salta y Jujuy.

6.4.2. Matrices de correlaciones y de lags

Una manera más general de representar visualmente todos los valores de correlaciones y de lags entre las provincias es graficarlas como matrices con un código de colores. Las dos matrices tienen de dimensiones de $24x24$, como corresponde con la cantidad de provincias. Esto nos permite visualizar todos los valores de las cantidades en estudio en apenas dos gráficos: uno para las correlaciones y uno para los lags. Ambos se presentan en la Figura 6.5.



(a) Matriz de correlaciones entre provincias.



(b) Matriz de lags entre provincias.

Figura 6.5: Matrices de correlaciones y de lags entre provincias, donde las escalas de color indican sus valores.

La Figura 6.5a muestra los valores de correlación entre todas las provincias del país. Primero, como es esperable, se observa que las correlaciones de cada provincia con sí misma toman el valor de la unidad. Luego, confirmamos que todas las correlaciones se

encuentran en un rango de valores de $0{,}4 - 1$. Sin embargo, la mayoría de las correlaciones toman valores más bien cercanos a la unidad. En cambio, la provincia de Formosa tiene los valores más bajo de correlación con el resto del país de todas las provincias argentinas. Esto sale a la luz claramente en el gráfico, ya que Formosa corresponde a una línea horizontal y a una línea vertical de colores más claros, mientras que el resto del gráfico tiene una apariencia más homogénea. Por otra parte, la Figura 6.5b es la representación correspondiente de la matriz de lags. Su escala, como es esperable, tiene una apariencia simétrica. Esto se debe a que, si el lag τ_{ij} tiene un valor, entonces $\tau_{ji} = -\tau ij$, como hemos aclarado con anterioridad.

Si bien esta representación sirve para visualizar todos los valores calculados para el lag, es difícil observar qué provincias anteceden y cuáles suceden a otras. Por ello, se procedió a realizar otro gráfico con un nuevo código de colores, dividiendo a los lags en positivos, negativos y nulos. Este gráfico corresponde a la Figura 6.6.

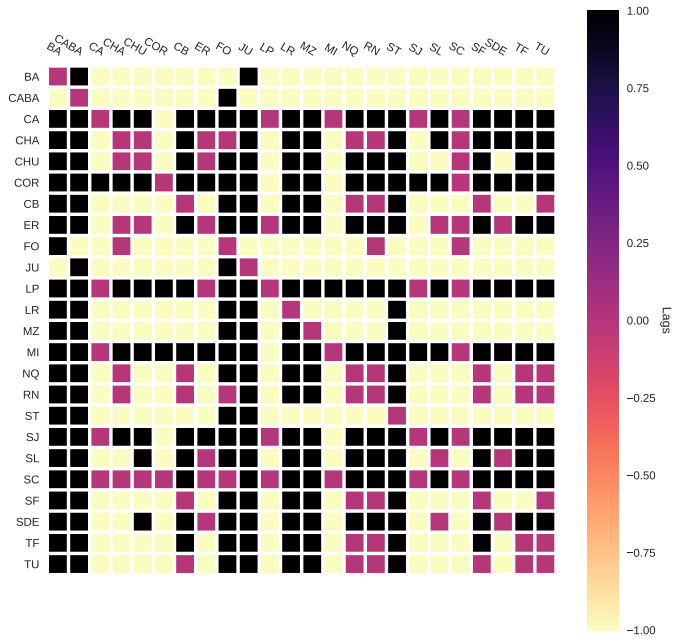


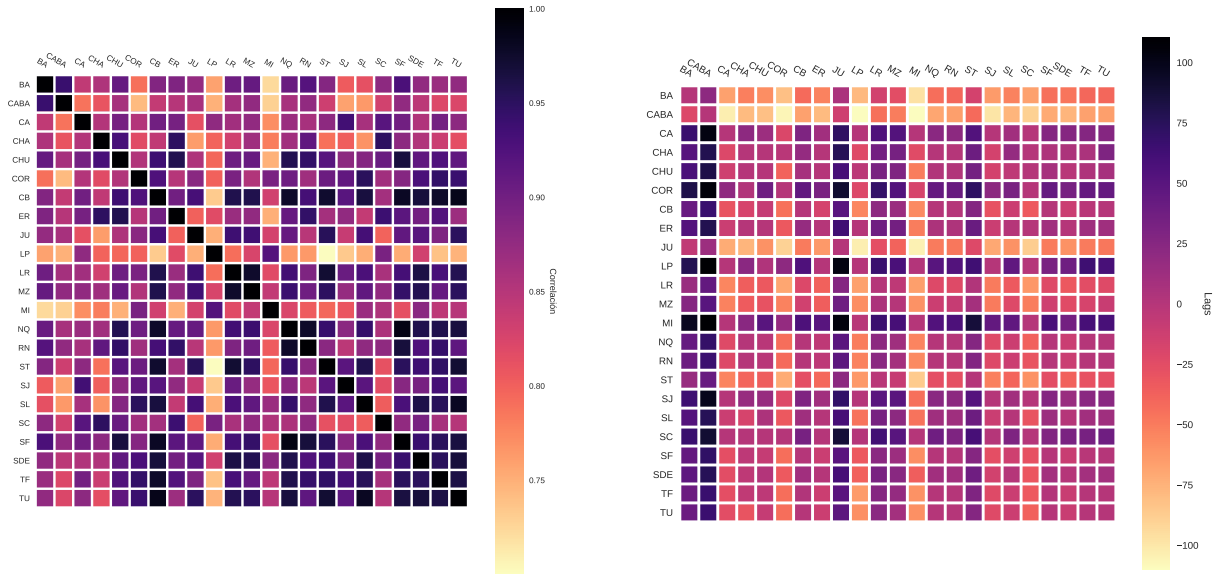
Figura 6.6: Matriz de lags entre provincias, donde el color indica únicamente si el lag en cuestión es positivo, negativo o nulo.

De éste se desprenden varias observaciones interesantes. Por un lado, se tiene que los lags de una provincia con si misma τ_{ii} son nulos, como es esperable. En cuanto a CABA, esta jurisdicción solo parece ser antecedida por Formosa. Por otra parte, se observa que los lags de BA con el resto del país son negativos, con excepción de dos provincias: CABA y Jujuy. Si bien la relación con CABA puede entenderse intuitivamente, el

resultado con respecto a Jujuy no es esperable a priori. Así, nuevamente, se observan particularidades en las cantidades calculadas de provincias del Noroeste.

Matrices de correlaciones y de lags sin Formosa

Como hemos observado de la Figura 6.5a, los datos de provincia de Formosa proporcionan una escala en un rango muy amplio a la hora de realizar el gráfico. Esto no permite observar con claridad las correlaciones entre el resto de las provincias. Por ello, realizamos nuevamente los gráficos de correlaciones y de lags, sin la provincia de Formosa. Estos se presentan en la Figura 6.7. De ambos gráficos, se observa que las escalas se modifican significativamente al quitar esta provincia, que parece tener valores inusuales comparada con el resto de las provincias. Esto es consistente con las observaciones de secciones previas.



(a) Matriz de correlaciones entre provincias, excluyendo Formosa.

(b) Matriz de lags entre provincias, excluyendo Formosa.

Figura 6.7: Matrices de correlaciones y de lags entre provincias, excluyendo la provincia de Formosa. Las escalas de color indican los valores de estas cantidades.

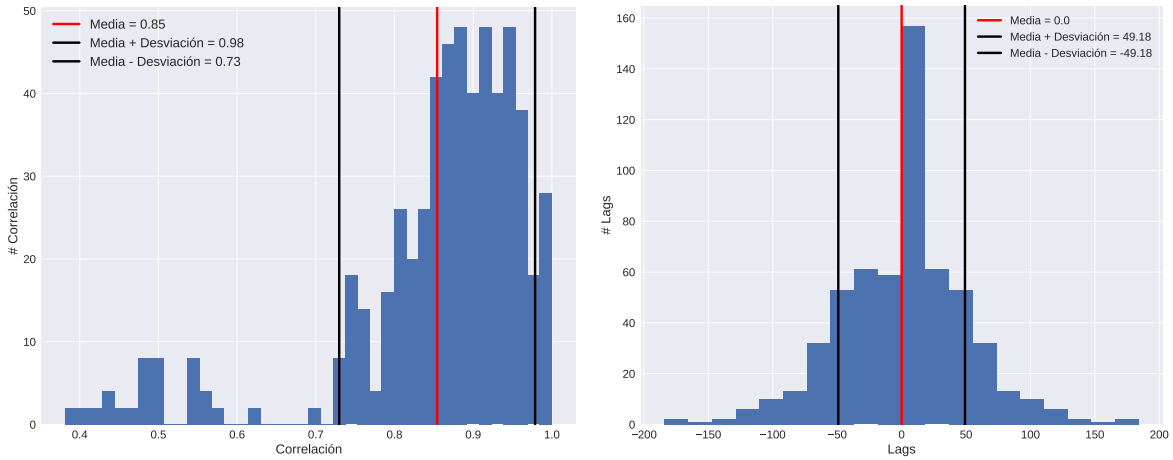
Hasta ahora, hemos investigado dos posibles visualizaciones de los resultados de correlación y de lag, al igual que hemos hecho un breve análisis. En particular, los gráficos de las matrices aportan elementos útiles. Sin embargo, presentan muchos datos y se torna difícil sacar muchas conclusiones a simple vista.

6.4.3. Redes de conexiones entre provincias

Una nueva manera posible de visualizar estos datos es intentar establecer redes de conexión entre provincias a partir de estos indicadores. Por ejemplo, conectando ciertas provincias si el valor de la correlación entre ellas es mayor a un cierto valor de umbral. Así, esto podría decirnos cuáles son las conexiones entre provincias que tienen un mayor peso. Estas podrían ser exclusivas a la dinámica epidémica, o reflejar una red de conexiones intrínseca al país. En ambos casos, resulta interesante estudiar la posibilidad de construir estas redes. Para ello, se comenzará por el estudio de las distribuciones de correlaciones y de lags.

Distribuciones de correlación y de lag

Las distribuciones correspondientes a los valores de correlaciones y de lags se presentan en la Figura 6.8. Por un lado, la distribución de las correlaciones se visualiza en la Figura 6.8a. Allí, se identifican los valores adjudicados a la provincia de Formosa, desplazados hacia la izquierda del gráfico, lejos de donde se acumulan el resto de las correlaciones. El promedio de las correlaciones es 0,85 y la desviación estándar es de 0,13, como se indica con líneas en dicha figura. Por otro lado, la distribución de los lags corresponde a la Figura 6.8b. Esta tiene una apariencia simétrica y su valor promedio es nulo, como es esperable de los lags. Adicionalmente, la desviación estándar vale alrededor de 49,18 días.



(a) Distribución de correlaciones. **(b)** Distribucion de lags.

Figura 6.8: Distribuciones de correlaciones y de lags teniendo en cuenta todas las provincias. Las líneas negras indican el valor promedio correspondiente en cada caso. Por otra parte, las líneas rojas señalan desplazamientos de éste valor teniendo en cuenta las desviaciones estándar.

Redes dadas por la correlación

Tras realizado el estudio de las distribuciones de las correlaciones y los lags, intentaremos establecer redes de conexión entre provincias. Así, nos dedicaremos a esta

tarea, usando inicialmente las correlaciones como indicador de un vínculo.

En un primer momento, sin tener en cuenta el estudio de las correlaciones, se puede pensar en establecer una conexión entre la provincia i y la provincia j si la correlación $\rho_{ij} \neq 0$. La red correspondiente se grafica en la Figura 6.9, donde el código de colores indica el valor de la correlación. Ésta estructura no resulta ser muy útil, debido a que ninguna correlación ρ_{ij} tiene un valor nulo, como hemos identificado antes. Así, se tiene una red poco informativa, donde todas las provincias están conectadas entre sí.

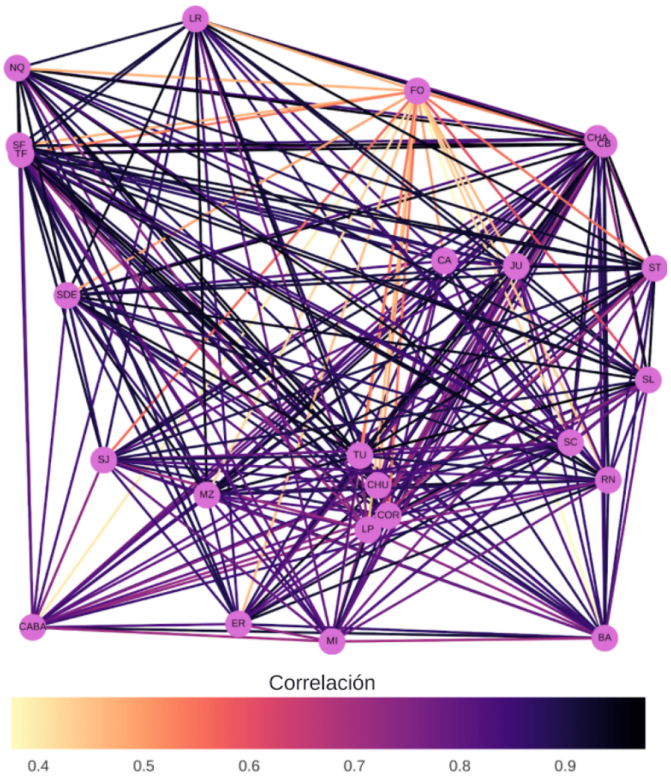


Figura 6.9: Red de conexión entre provincias dada por las correlaciones. Los colores de los enlaces que conectan dos provincias codifican el valor de la correlación entre cada par de ellas.

Por eso, a continuación se probó filtrar conexiones con un valor umbral de correlación. Se exploraron varios valores en un rango de $0{,}7 - 0{,}9$, todos con resultados cualitativamente similares. En la Figura 6.10, se presenta la red resultante de utilizar 0,85 como umbral. Nuevamente, se utiliza una escala de colores para codificar el valor de las correlaciones, modificada por causa del filtro. Esta red presenta menos enlaces que la anterior. Mientras que la primera cuenta con 276 enlaces entre provincias, esta solo captura 182 de ellos. Si nuestra línea de pensamiento anterior es correcta, estos son los más relevantes.

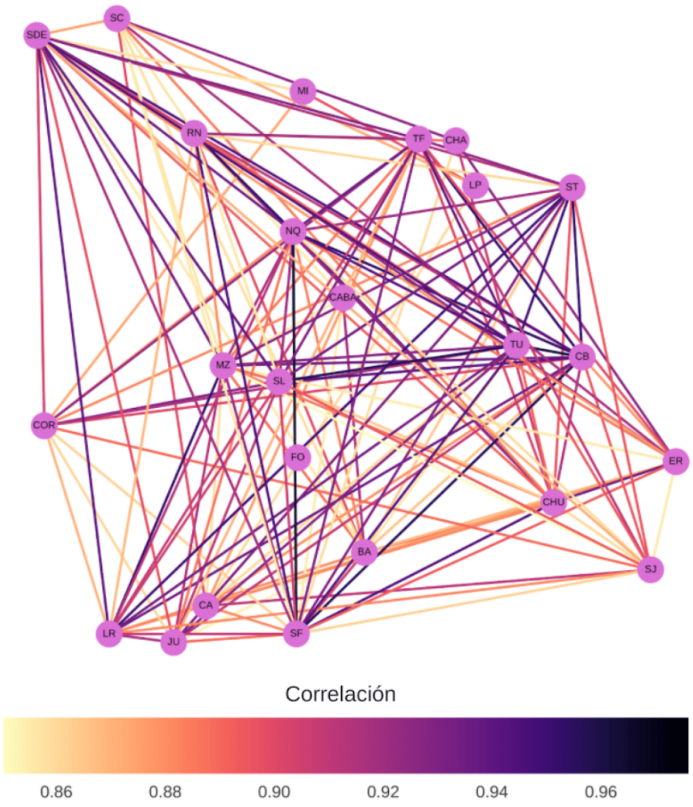


Figura 6.10: Red de conexión entre provincias dada por las correlaciones, filtrando correlaciones que son menores al valor umbral de 0,85. Los colores de los enlaces que conectan dos provincias codifican el valor de la correlación entre cada par de ellas.

Para investigar cuales son las provincias mejores conectadas según este criterio, se graficó el número de conexiones para cada una de ellas. Los resultados se presentan en la Figura 6.11. Este gráfico muestra que la menor cantidad de enlaces la tiene la provincia de Formosa, como es esperable, ya que sus correlaciones con el resto del país son muy pequeñas. Por otra parte, se observa que la siguen provincias como La Pampa o Misiones. Esto puede entenderse pues fueron provincias con pocos casos. En el caso de La Pampa, tuvo medidas muy fuertes en cuanto al aislamiento y a la entrada de personas provenientes de otros lugares. Algo similar pudiera haber ocurrido para Misiones que, adicionalmente, tiene una distancia muy grande con varias jurisdicciones argentinas. Sobre todo, si se considera que el transporte y la movilidad de las personas se vieron disminuidos a causa de la pandemia.

Sin embargo, tanto CABA como BA no se encuentran entre las jurisdicciones mejor conectadas con el resto del país. Por un lado, esto es contrario a lo que se espera intuitivamente, dado que una gran parte de la población del país se concentra en esta región. Adicionalmente, intuitivamente parecen tener mayor movilidad general en relación al resto del país. Por ejemplo, a lo largo de la pandemia, viajeros infectados de CABA o BA llegaron a diversas localidades, dando comienzo a un brote.

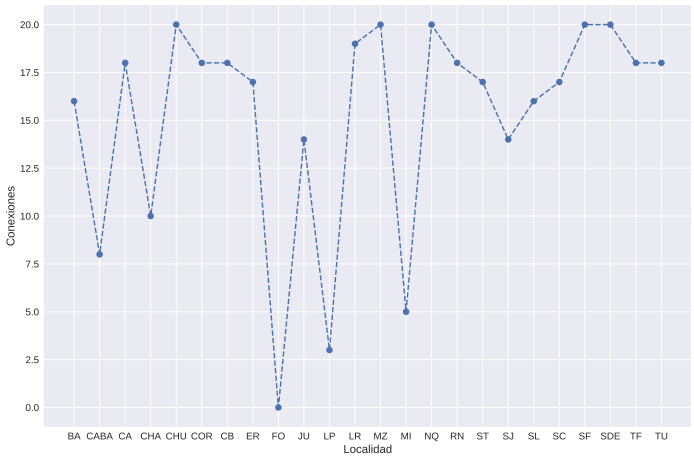


Figura 6.11: Conexiones que restan para cada provincia luego del filtrado de correlaciones según un valor umbral. Se destaca que tanto CABA como BA están conectadas con pocas provincias, en comparación con provincias como Santa Fe o Tucumán.

En cambio, otras provincias más aisladas o con menor población, como Tierra del Fuego, parecen ser más relevantes en la propagación de la enfermedad. De esta manera, este filtro de provincias para conformar una red de conectividades no tiene el resultado esperado.

Como se ha dicho, se han utilizado varios valores umbrales para asegurarse de que el fenómeno no fuera producto de escoger un umbral inadecuado. Este tipo de patrón se repite para todos los valores umbral utilizados que eliminen una cantidad considerable de enlaces. Finalmente, también se ha utilizado el filtro por valores de lag con valores en un rango de $50 - 100$. A diferencia del caso de las correlaciones, se han eliminado los valores de lags mayores al umbral. Es decir, priorizando las conexiones a través de las cuales tarden menos en llegar casos de una jurisdicción a otra. Sin embargo, se observan distribuciones de conexiones por provincia que son similares a las obtenidas en el análisis de las correlaciones. O sea, tanto CABA como BA forman una región conectada débilmente con el resto de Argentina.

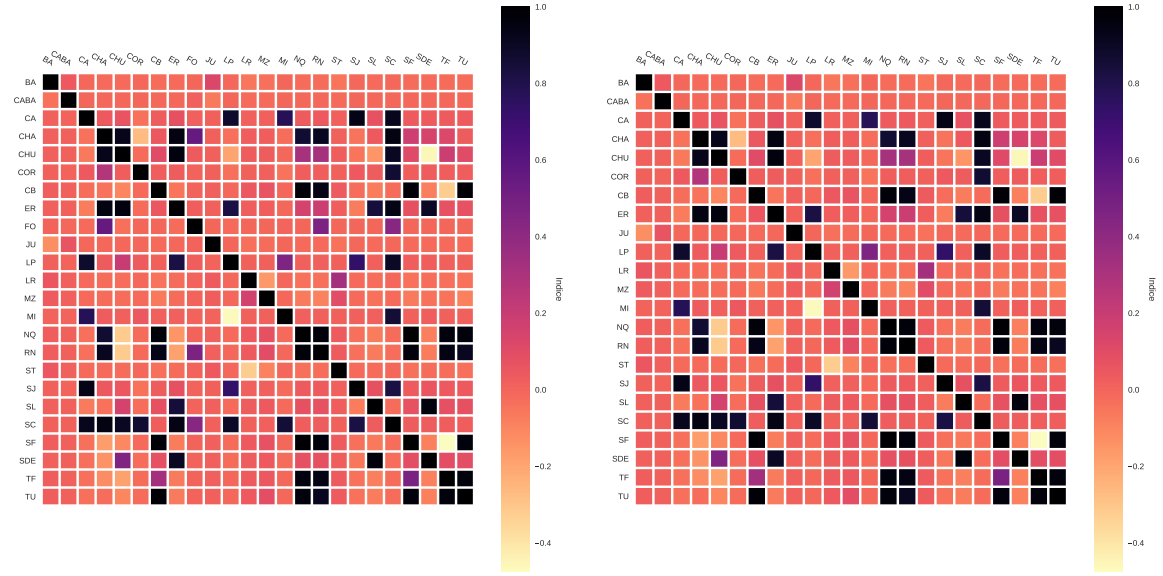
6.5. Propuesta de índice de conectividad

Como manera de poder cuantificar la conectividad entre provincias, también se investigó la idea de proponer una cantidad que tuviera en cuenta las contribuciones de ambos indicadores, es decir, la correlación y el lag. Así, surgió la posibilidad de utilizar un índice de conectividad I_{ij} entre la provincia i y la provincia j, calculado según la ecuación:

$$\mathcal{I}_{\rangle|} = \frac{\rho_{ij}}{\mid \tau_{ij} \mid +1}. \tag{6.3}$$

Esta fórmula refleja claramente la idea intuitiva de que la conectividad es mayor

cuanto mayor es la correlación y menor cuanto mayor es el lag. En el denominador, se suma la unidad, dado que la correlación puede ser nula. En particular, por ejemplo, el lag calculado para una provincia con relación a sí misma. Así, se calculó el índice de conectividad formulado, con y sin la provincia de Formosa. Estos resultados se presentan en la Figura 6.12, en forma de matrices.



(a) Matriz de índices de conectividad entre provincias

(b) Matriz de índices de conectividad entre provincias, excluyendo Formosa.

Figura 6.12: Matrices de índices de conectividad entre provincias, teniendo en cuenta y quitando los datos de Formosa. Las escalas de color indican los valores de esta cantidad.

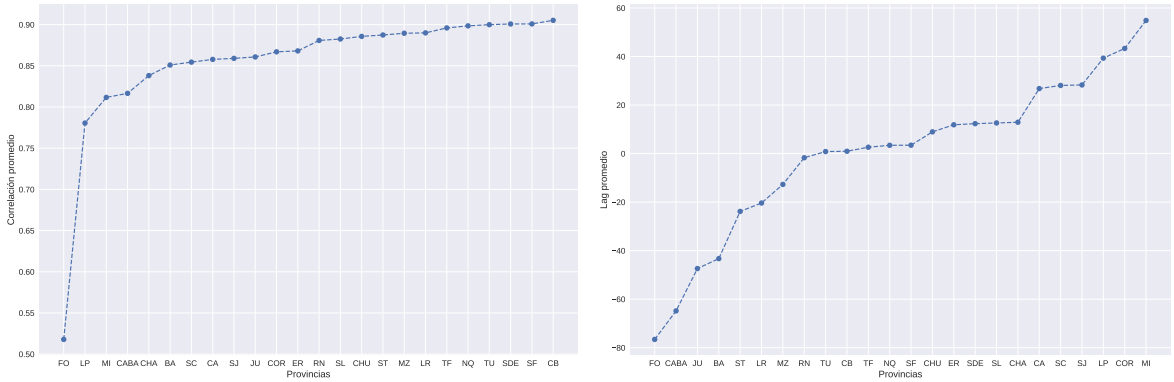
Existen varias objeciones importantes al utilizar este índice. Por un lado, no se observan grandes diferencias de escala al quitar Formosa, como sí sucede en los casos de las matrices de correlación y de lag. En el caso de la Figura 6.12b, la provincia no parece ser particular, como si lo indican las cantidades que constituyen el lag por si solas. Por otra parte, la matriz tiene una apariencia mayormente homogénea. Con lo cual, filtrar valores de índice no nos aportaría mucha información. Se intentó construir redes análogas a las desarrolladas con la correlación y con el lag individualmente. Sin embargo, los resultados tuvieron cualitativamente el mismo comportamiento.

6.6. Correlaciones y lags promedio para las provincias

Hasta ahora, hemos analizado parcialmente las correlaciones y los lags de distintas maneras. El último objetivo formulado fue establecer redes de conexiones entre provincias, que destaquen los enlaces más relevantes para la dinámica epidémica. Sin embargo, todas las ideas intuitivas no han sido adecuadas para lograr este objetivo.

Principalmente, porque provincias como BA y CABA, que tienen una gran importancia en la propagación, parecen tener correlaciones más bajas y lags más altos con respecto al resto del país.

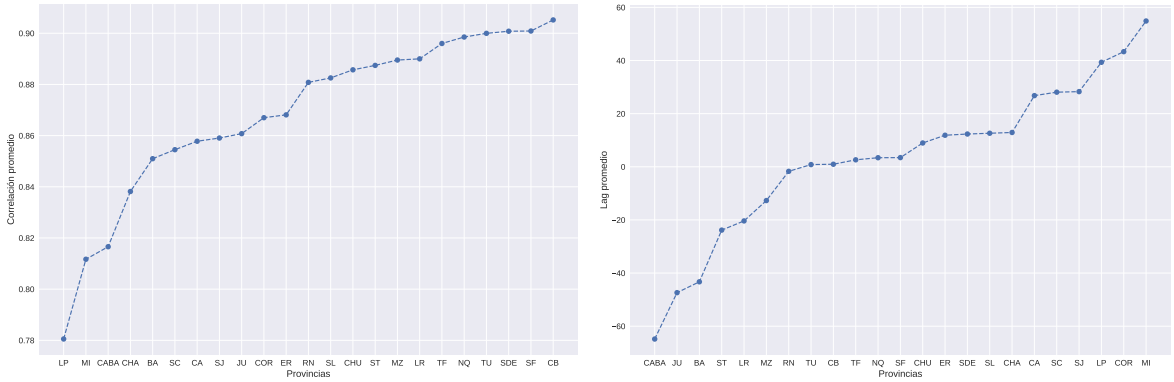
Para estudiar como son las correlaciones y los lags de todas las provincias, se propuso estudiar estas cantidades haciendo un promedio sobre todas las jurisdicciones. Es decir, dada una provincia, la correlación promedio corresponde al promedio de las correlaciones que la involucran. La definición del lag medio es análoga. Esto nos permite saber qué provincias tienen lags y correlaciones mayores o menores, en promedio. Ambas cantidades se presentan en la Figura 6.13. En ambos casos, tanto la correlación como el lag se encuentran ordenados de menor a mayor. De la manera que se calcula el lag promedio, las provincias que tienen mayor tendencia a anteceder a las otras poseen un lag promedio negativo.



(a) Correlaciones promedio para cada provincia. **(b)** Lags promedio para cada provincia.

Figura 6.13: Correlaciones y lags promedio para cada provincia, incluyendo todas las provincias del país.

Por un lado, se tiene que la provincia de Formosa tiene la correlación promedio y el lag promedio mínimos. Esto es claramente esperable de las discusiones anteriores. Dado que la escala de correlaciones se ve fuertemente afectada por esta provincia, realizamos estos gráficos sin ella. Estos se encuentran en la Figura 6.14.



(a) Correlaciones promedio para cada provincia, excluyendo Formosa.

(b) Lags promedio para cada provincia, excluyendo Formosa.

Figura 6.14: Correlaciones y lags promedio para cada provincia, excluyendo la provincia de Formosa.

De la Figura 6.14b, se observa que las cuatro primeras provincias que lideran temporalmente el brote son CABA, Jujuy, BA y Salta. Los valores de los lags resultan ser negativos y muy elevados en valor absoluto. Por un lado, es esperable intuitivamente que CABA y BA sean dos de las provincias que anteceden al resto del país. Por otro lado, hemos señalado con anterioridad varias características de los lags de provincias del Noroeste, como Salta y Jujuy. Una hipótesis es que esto se debe a que los primeros casos importados hayan ingresado simultáneamente tanto en CABA y BA como en Salta y Jujuy. Sin embargo, no se encontraron datos oficiales que den cuenta del ingreso de casos por pasos fronterizos en el Noroeste. Es más, considerando que estos casos podrían haber ingresado en una etapa con cierre de fronteras, deberían haberlo hecho por pasos fronterizos informales. Esto dificultaría aún más tener datos confiables sobre la posible existencia de este fenómeno.

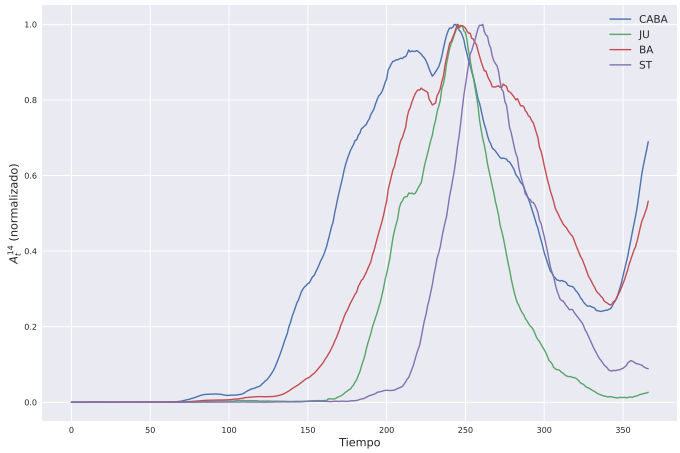


Figura 6.15: Series temporales de casos para las cuatro provincias con lags más pequeños: CABA, Jujuy, BA y Salta.

Las series temporales de estas provincias, normalizadas a la unidad, se presentan en

la Figura 6.15. Allí, se observa claramente el orden en el cual ellas se anteceden entre sí, con diferencias de algunos días.

Por otra parte, de la Figura 6.14a, se tiene que estas cuatro provincias no tienen correlaciones muy elevadas. En particular, como hemos señalado anteriormente, este gráfico confirma que tanto CABA como BA tienen correlaciones medias que se encuentran entre las de menor valor. Es decir, las provincias que anteceden al resto del país no solo no tienen lags pequeños en valor absoluto, sino que tampoco tienen correlaciones elevadas en comparación con el resto de las jurisdicciones. Esto es lo que nos ha impedido formar las redes de conectividades utilizando las correlaciones y los lags. Para confirmar este fenómeno, se realizaron gráficos de lag y lag en valor absoluto en función de la correlación. Estos datos se presentan a continuación, en la Figura 6.16b. Como puede observarse, sobre todo de la Figura 6.16b parece existir una relación inversamente proporcional entre el valor absoluto del lag y la correlación.

(a) Lag en función de la correlación correspondiente para todas las provincias.

(b) Valores absolutos de lag en función de la correlación correspondiente para todas las provincias.

Figura 6.16: Lag y lag absoluto en función de la correlación correspondiente para todas las provincias del país.

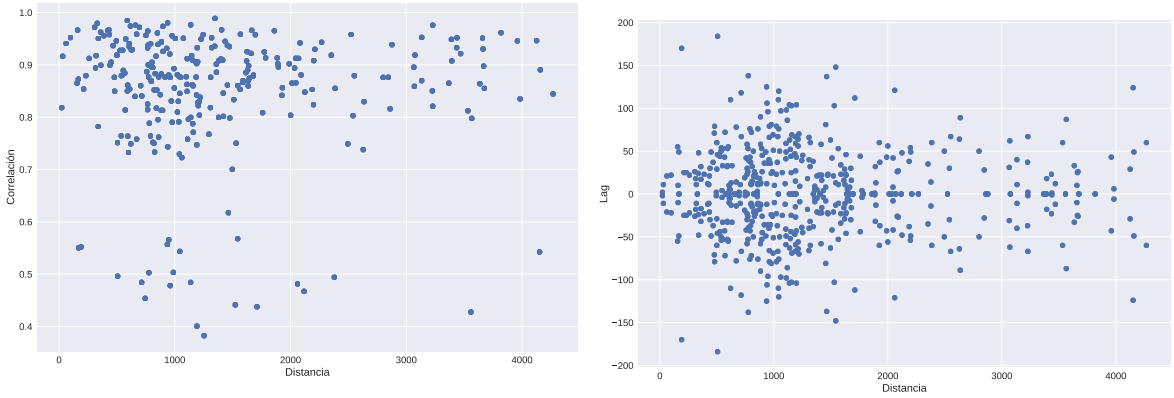
Como consecuencia, debemos observar que puede tomar mucho tiempo que las provincias que lideran la dinámica influencien al resto. En cambio, el resto siguen una dinámica más parecida entre sí. Si no se observa un incremento de casos entre una de ellas y sus jurisdicciones cercanas, esto no significa que esto no vaya a acontecer. Adicionalmente, observar esto nos permite entender mejor por qué las redes de correlación y de lag no son consistentes con las expectativas intuitivas.

6.7. Relación entre correlaciones, lags y distancias

Un problema interesante, que no hemos abordado hasta ahora, es establecer una relación entre las correlaciones o los lags y las distancias entre provincias. Modelos

epidemiológicos tipo SIR de dos poblaciones en la bibliografía [2] muestran una relación clara entre las cantidades calculadas y la distancia entre dos poblaciones. Como es de esperar intuitivamente, estos modelos indican que la correlación disminuye y el lag aumenta con un incremento de la distancia entre localidades.

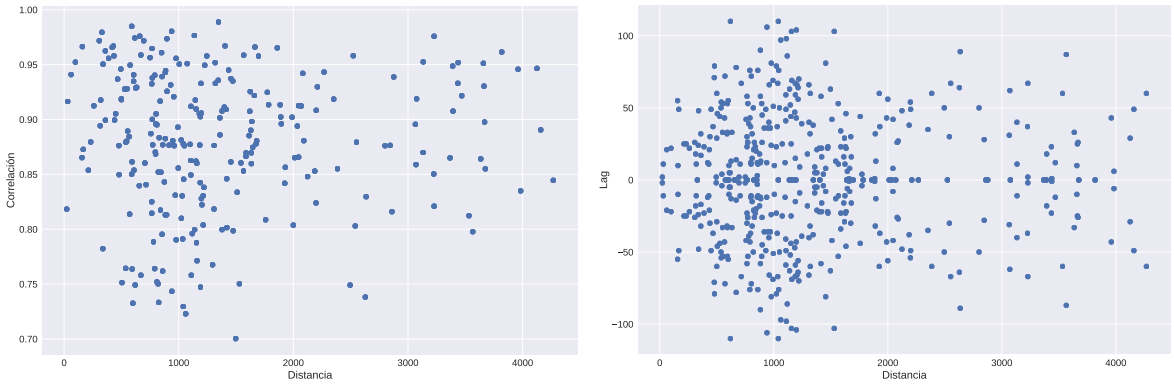
Para observar si este fenómeno se replica en el caso de las correlaciones y lags calculados entre provincias, se toman los valores de distancias entre capitales de provincias por vía terrestre del Instituto Geográfico Nacional (IGN) [20]. Cabe destacar que esta vía es la responsable de la mayor parte de la movilidad en el período analizado, al no ser tan usuales otros medios como los vuelos. De esta manera, es posible elaborar los primeros resultados de correlaciones y lags en función de las distancias entre jurisdicciones, que se presentan en la Figura 6.17. Adicionalmente, se presentan los resultados sin tener en cuenta la provincia de Formosa, en la Figura 6.18.



(a) Correlaciones en función de distancias entre provincias.

(b) Lags en función de distancias entre provincias.

Figura 6.17: Correlaciones y lags entre provincias en función a las distancias correspondientes. Se destaca la falta de una relación clara entre las cantidades estudiadas y las distancias interprovinciales.



(a) Correlaciones en función de distancias entre provincias, excluyendo Formosa.

(b) Lags en función de distancias entre provincias, excluyendo Formosa.

Figura 6.18: Correlaciones y lags entre provincias en función a las distancias correspondientes, sin considerar Formosa. Se destaca nuevamente la falta de una relación clara entre las cantidades estudiadas y las distancias interprovinciales.

En ambos casos, se observa que no hay una relación clara entre cantidades como la correlación o el lag y la distancia entre provincias. Si bien no tener en cuenta los datos formoseños cambia las escalas y modifica la apariencia de los gráficos, la relación entre estas cantidades continua sin ser evidente.

6.8. Conclusiones

En este capítulo, nos hemos dedicado a introducir y a analizar dos herramientas para dar cuenta de una onda epidémica: las correlaciones y los lags entre provincias. Para ello, hemos utilizado diversas representaciones visuales como la confección de mapas y de gráficos de matrices. Adicionalmente, se han intentado construir redes de conexiones entre jurisdicciones de diferentes maneras. Primero, a través del filtrado de conexiones por valores de correlaciones y de lags. Luego, a través de la definición de un índice.

Estas propuestas han resultado infructíferas para resaltar los enlaces más relevantes entre provincias. Uno de sus principales problemas es que daban poca importancia a jurisdicciones centrales, como CABA y BA. Sin embargo, han permitido observar dos características interesantes. Por un lado, la particularidad de la provincia de Formosa al compararse con el resto del país, la cual puede tener raíces en el manejo provincial de la situación epidémica. Por otro lado, el hecho de que provincias como CABA y BA tienen correlaciones bajas y lags grandes en valor absoluto. Esto resulta estar en contradicción con las expectativas intuitivas. En el próximo capítulo, formularemos hipótesis sobre la naturaleza matemática de este fenómeno introduciendo nuevos modelos epidemiológicos. Sin embargo, hasta el momento, no hemos proporcionado explicaciones sobre esta contradicción.

Luego, pasamos a centrarnos en otro tipo de análisis, basado en calcular correlaciones y lags promedios. Este nos ha permitido identificar más claramente el fenómeno de correlaciones bajas para provincias que lideran la onda epidémica. Es decir, aquellas provincias que tienen un lag negativo muy grande en valor absoluto. Entre ellas, se destacan las jurisdicciones de CABA, BA, Jujuy y Salta. Finalmente, también ha sido posible identificar que no existen relaciones claras entre las correlaciones o lags y las distancias entre provincias. Tampoco hemos formulado hipótesis al respecto en este capítulo. En cambio, nos ocuparemos de esto también en el próximo capítulo.

Estas últimas características resultan ser muy interesantes y constituyen las conclusiones más importantes de esta parte del trabajo. Sin embargo, por ahora no hemos aventurado posibles explicaciones claras para ninguna de ellas. Para ello, en el próximo capítulo, nos dedicaremos a proponer modelos que permitan entender los fenómenos subrayados.

Capítulo 7

Modelos metapoblaciones de propagación espacial

En el capítulo anterior, nos hemos dedicado a introducir los conceptos de correlaciones y de lags, al igual que a analizarlos de diversas maneras. El objetivo principal de este capítulo era estudiar aspectos de una onda epidémica que se propagó de ciertas provincias hacia el resto del país, luego de la introducción del virus. Del análisis de las cantidades calculadas, se desprenden diversas observaciones no esperadas y que no se explican intuitivamente. En particular, hemos destacado dos de ellas. Por un lado, el hecho de que provincias que lideran temporalmente la dinámica epidémica presentan correlaciones de valor bajo. Por otro lado, la relación no evidente entre correlación y distancia; además de entre lag y distancia. A diferencia de lo que se puede observar para modelos de dos poblaciones en la bibliografía, la correlación no disminuye y el lag no aumenta con un incremento de la distancia interprovincial.

En este capítulo, nos dedicaremos a proponer modelos SIR de tipo metapoblacional con el objetivo de entender mejor los aspectos no comprendidos que hemos señalado. Para ello, en la sección 7.1 introducimos la idea de los modelos metapoblacionales. A continuación, en la sección 7.2, se detallan los modelos propuestos para explicar la dinámica epidemiológica argentina. La aplicación de los modelos abstractos al país se explica en la Figura 7.3. Adicionalmente, demostraremos cualitativamente cómo pueden explicar los fenómenos observados que nos interesan en la exploración preliminar de la sección 7.4. Hechas estas consideraciones, se procede a realizar y comparar los ajustes de los diferentes modelos presentados a los datos. Finalmente, procederemos a analizar los parámetros del ajuste óptimo en la sección 7.5.

7.1. Modelos Metapoblacionales

Los modelos metapoblacionales constituyen un tipo de modelos particular utilizado en disciplinas como la epidemiología, la biología y la ecología para abordar problemas de dinámica espacial. En particular, varios tipos de modelos metapoblacionales han sido propuestos para estudiar la propagación de la COVID-19 en diferentes regiones del mundo [21] [22] [23] [24].

Los modelos metapoblacionales consisten en dos premisas fundamentales. Por un lado, se supone que una población puede dividirse en distintas subpoblaciones con dinámicas propias. Por otro lado, se supone que las subpoblaciones pueden interactuar unas con otras, modificando la dinámica de cada una de ellas [2] [25].

Como ejemplo, supongamos que tenemos una serie de subpoblaciones. Entonces, consideremos que cada subpoblación i tiene densidades S_i,I_i y R_i de suceptibles, infectados y recuperados. Si suponemos que cada población está regida por una dinámica SIR simple como la presentada en el capítulo 4, entonces las ecuaciones para las cantidades de cada población son:

$$\frac{dS_i}{dt} = -\beta S_i I_i \tag{7.1a}$$

$$\frac{dI_i}{dt} = \beta S_i I_i - \gamma I_i \tag{7.1b}$$

$$\frac{dR_i}{dt} = \gamma I_i, \tag{7.1c}$$

donde hemos asumido que todas las poblaciones tienen la misma tasa de transmisibilidad β y la misma tasa de recuperación γ. Al igual que para el modelo SIR de campo medio, ha de cumplirse que $S_i + I_i + R_i = 1$. Como podemos observar, en este caso, las ecuaciones de diferentes localidades no están acopladas. Así, estas ecuaciones solo representan a una serie de subpoblaciones no interactuantes entre sí. Cada una de ellas seguirá una dinámica de tipo SIR dependiente de sus condiciones iniciales.

7.2. Modelos Metapoblacionales propuestos

7.2.1. Formulación de los modelos

Por otra parte, diversos modelos tienen en cuenta el movimiento de personas entre subpoblaciones al modelar las dinámicas epidemiológicas de cada una de ellas. Con este objetivo, consideraremos algunos modelos posibles que consisten en incorporar interacción entre diferentes poblaciones a través de matrices A_{ij}, B_{ij}, C_{ij} que dan cuenta de la movilidad de susceptibles, infectados y recuperados entre las poblaciones i y j.

De esta manera, las ecuaciones resultantes para cada localidad i están dadas por:

$$\frac{dS_i}{dt} = -\beta S_i I_i + \sum_{j=1}^{n} A_{ij} S_j \tag{7.2a}$$

$$\frac{dI_i}{dt} = \beta S_i I_i - \gamma I_i + \sum_{j=1}^{n} B_{ij} I_j \tag{7.2b}$$

$$\frac{dR_i}{dt} = \gamma I_i + \sum_{j=1}^{n} C_{ij} R_j. \tag{7.2c}$$

Adicionalmente, para que este modelo resulte ser más general, se puede suponer que las tasas de contagio son diferentes para cada una de las poblaciones. De esta manera, si consideramos que esta cantidad toma un valor β_i para la de índice i, entonces las ecuaciones anteriores se convierten en:

$$\frac{dS_i}{dt} = -\beta_i S_i I_i + \sum_{j=1}^{n} A_{ij} S_j \tag{7.3a}$$

$$\frac{dI_i}{dt} = \beta_i S_i I_i - \gamma I_i + \sum_{j=1}^{n} B_{ij} I_j \tag{7.3b}$$

$$\frac{dR_i}{dt} = \gamma I_i + \sum_{j=1}^{n} C_{ij} R_j. \tag{7.3c}$$

En el resto del presente trabajo, utilizaremos números de suceptibles X_i, infectados Y_i y de recuperados Z_i en vez de densidades. Para obtener las ecuaciones correspondientes, debemos efectuar el cambio de variables:

$$\frac{dS_i}{dt} = -\beta_i S_i I_i + \sum_{j=1}^{n} A_{ij} S_j \tag{7.4a}$$

$$\frac{dI_i}{dt} = \beta_i S_i I_i - \gamma I_i + \sum_{j=1}^{n} B_{ij} I_j \tag{7.4b}$$

$$\frac{dR_i}{dt} = \gamma I_i + \sum_{j=1}^{n} C_{ij} R_{j}., \tag{7.4c}$$

donde $N_i = X_i + Y_i + Z_i$ es la población de la población i. Tras este cambio, obtenemos

que las ecuaciones de 7.4 se transforman en el sistema:

$$\frac{dX_i}{dt} = -\frac{\beta X_i Y_i}{N_i} + \sum_{j=1}^{n} A_{ij} X_j \frac{N_i}{N_j} \tag{7.5a}$$

$$\frac{dY_i}{dt} = \frac{\beta X_i Y_i}{N_i} - \gamma Y_i + \sum_{j=1}^{n} B_{ij} Y_j \frac{N_i}{N_j} \tag{7.5b}$$

$$\frac{dZ_i}{dt} = \gamma Y_i + \sum_{j=1}^{n} C_{ij} Z_j \frac{N_i}{N_j}, \tag{7.5c}$$

Adicionalmente, consideraremos que la población en cada parche permanece fija, aunque permitiendo movimientos de personas. Matemáticamente, esta condición se expresa según la ecuación:

$$X_i + Y_i + Z_i = \sum_{j=1}^{n} \Big[A_{ij} X_j + B_{ij} Y_j + C_{ij} Z_j \Big] \frac{N_i}{N_j} = N_i. \tag{7.6}$$

Por otro lado, una simplificación posible de este modelo es asumir que las matrices de contacto son iguales para las tres categorías. Es decir, suponer que $A_{ij} = B_{ij} = C_{ij}$. Llamaremos a la matriz A_{ij} la matriz de conectividad. En ese caso, la ecuación 7.6 se transforma en:

$$\sum_{j=1}^{n} A_{ij} = 0, \forall i = 1, ..., n. \tag{7.7}$$

Finalmente, cabe destacar que esta expresión también es equivalente a pedir que:

$$A_{ii} = \sum_{j \neq i}^{n} A_{ij} = 0, \forall i = 1, ..., n, \tag{7.8}$$

lo cual nos da una manera de calcular los elementos diagonales de la matriz de conectividad.

7.2.2. Propuestas de matrices de contacto

Hasta ahora, hemos planteado diferentes tipos ecuaciones generales para un sistema de n subpoblaciones con N_i habitantes. Sin embargo, no se han propuesto posibles formas para la matriz de conectividad A_{ij} que conecta a las diferentes subpoblaciones, tanto si la tasa de contagio es homogéana como si no. Esto es lo que nos dedicaremos a realizar en esta sección. Para ello, es necesario realizar unas definiciones previas.

Por un lado, dado que tanto la distancia parece ser una variable importante para la conectividad, definiremos d_{ij} como la distancia entre las subpoblaciones i y j. Además,

recordamos también que N_i es el número de habitantes de la población i. Finalmente, denominaremos $N_T = \sum_{i=1}^{n} N_i$ al número de habitantes de la población completa. Con estas definiciones, se proponen cuatro formas posibles para la matriz A_{ij}:

- **Matriz neutral** Se propone en este caso una matriz que tenga elementos A_{ij} constantes para $i \neq j$.

- **Matriz dependiente de la población** Un segundo caso posible es una matriz que esté dada por las poblaciones de manera que:

$$A_{ij} \propto \frac{N_i + N_j}{N_T}, \tag{7.9}$$

  tal que la movilidad es mayor cuanto mayor sea la población en total de ambas subpoblaciones.

- **Matriz dependiente de la distancia** El tercer caso se formula solamente teniendo en cuenta las distancias y consiste en proponer que la matriz es:

$$A_{ij} \propto \frac{1}{d_{ij}}, \tag{7.10}$$

  de forma que la tasa de migración es inversamente proporcional a la distancia entre subpoblaciones.

- **Matriz dependiente de la población y de la distancia** El cuarto caso incorpora directamente los dos anteriores, proponiendo que:

$$A_{ij} \propto \frac{N_i + N_j}{N_T} \frac{1}{d_{ij}}. \tag{7.11}$$

7.3. Idea de aplicación a provincias argentinas

Para la aplicación de estos modelos al caso real, consideraremos que las provincias constituyen las subpoblaciones definidas de una población más grande. Esta correspondería, entonces, a la población total del país.

Para realizar simulaciones, se necesitó averiguar los datos oficiales tanto de distancias entre provincias como de número de habitantes de cada jurisdicción. Como hemos explicado en el capítulo anterior, se considera que las distancias son las distancias por vía terrestre entre capitales de provincias. Estas son reportadas por el IGN [20]. Por otra parte, los números de habitantes utilizados se reportan en la página de internet del Instituto Nacional de Estadística y Censos (INDEC) [26].

Adicionalmente, se toman como condiciones iniciales los números correspondientes a la incidencia acumulada A_{14}^t para cada provincia al día $t = 70$ desde el inicio de la

epidemia. Estos valores se especifican en el Apéndice A.

Finalmente, a menos de que se especifique, se utilizan en todas las simulaciones los parámetros $\gamma = 1/14$ y $beta = 2/14$ de manera que, si se considerase el modelo SIR de campo medio en cada localidad, el número reproductivo fuera $R_0 = 2$. Los valores de la constante de proporcionalidad de los modelos utilizados han sido arbitrarios y se especifican en cada caso.

7.4. Exploración preliminar de los modelos

En esta sección, nos dedicaremos a explorar varias propiedades de los modelos propuestos con anterioridad. Para ello, se utilizarán los datos oficiales y observaciones de incidencia iniciales presentados en la sección anterior. En particular, haremos algunas simulaciones para estudiar características de las correlaciones y de los lags, al igual que la falta de relación entre estas cantidades con las distancias interprovinciales.

7.4.1. Valores de correlación y de lag para provincias fuentes

Para estudiar los valores de correlación y de lag, realizamos una simulación con el modelo neutral para la matriz de conectividad, utilizando los números de habitantes y las condiciones iniciales especificadas con anterioridad. Para los elementos no diagonales de la matriz, utilizamos un valor arbitrario de 0,001. En cambio, los valores diagonales se calculan de la ecuación 7.8. Las series temporales normalizadas correspondientes a esta simulación se presentan en la Figura 7.1. Por un lado, se tiene que estas son curvas que llegan a un máximo desde su condición inicial, y luego decaen. Es decir, la epidemia termina por extinguirse con el tiempo.

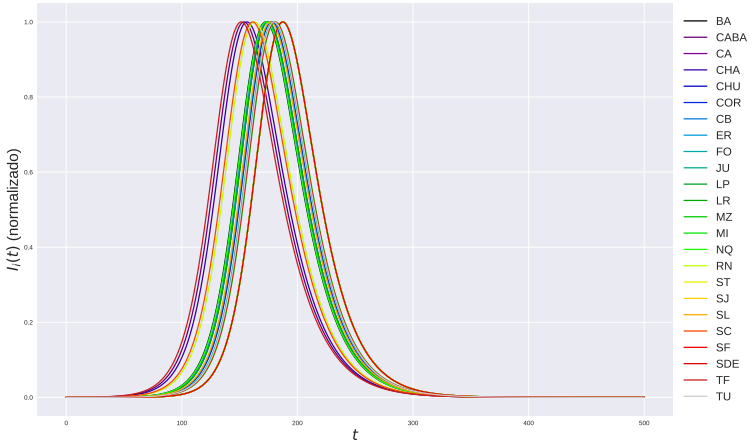
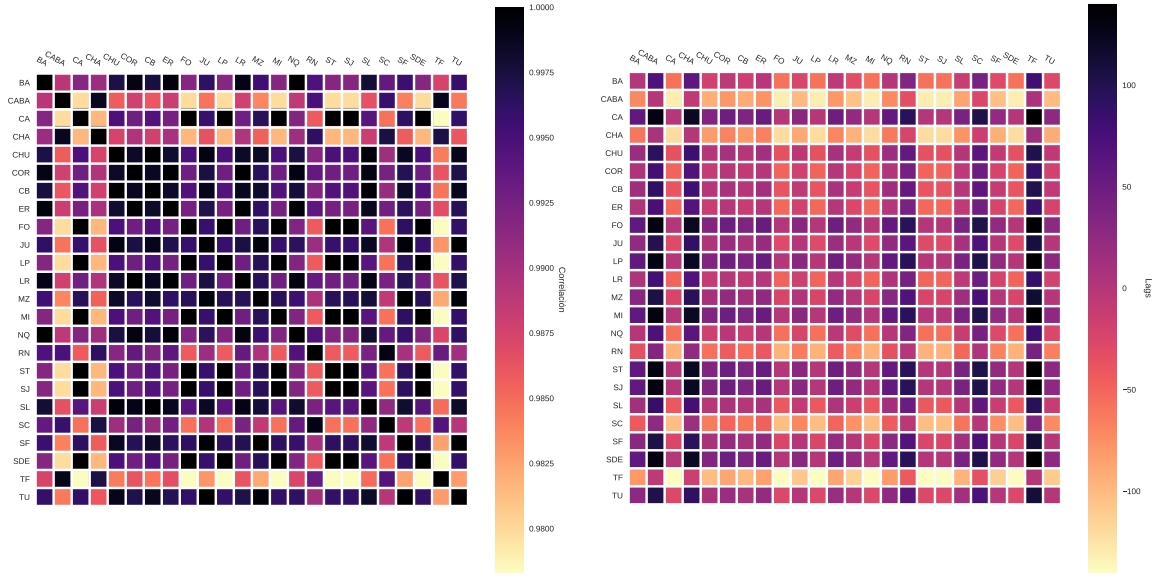


Figura 7.1: Series temporales obtenidas para el modelo neutral, con condiciones iniciales y poblaciones reales.

Adicionalmente, es interesante destacar que parecen existir ciertos grupos de curvas. O sea, grupos de provincias que tienen una dinámica epidemiológica similar entre sí y distinta a la de otras localidades. Dado que las series temporales que provienen de los datos oficiales son más ruidosas, no hemos observado este tipo de fenómeno a simple vista anteriormente.

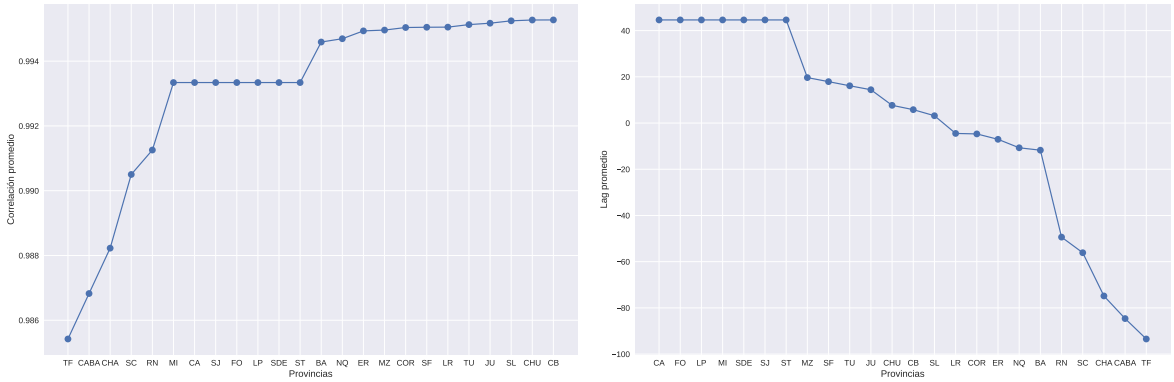
Por otra parte, en la Figura 7.2 se presentan las matrices de correlación y de lag obtenidas de las simulaciones anteriores. En estas matrices, los grupos de provincias se asocian a conglomerados de correlaciones muy altas y lags muy bajos entre sí. En cambio, tienen lags mayores y correlaciones menores con el resto del país en general. Por ejemplo, en este caso, esto es lo que se observa con el grupo de provincias entre Chubut y Entre Ríos. Observando la Figura 7.1, es claro que CABA antecede a estas jurisdicciones, que tienen un pico de casos casi simultáneamente. De esta manera, CABA termina teniendo un lag considerable con estas provincias. Al mismo tiempo, como las series temporales de las mismas son tan parecidas, estas tienen un lag en valor absoluto bajo y una correlación mucho más grande entre sí que con CABA. Sin embargo, CABA puede considerarse como una jurisdicción fuente de la infección, ya que tiene un mayor número de infectados como condición inicial.



(a) Matriz de correlaciones entre provincias. **(b)** Matriz de lags entre provincias.

Figura 7.2: Matrices de correlaciones y de lags entre provincias para el modelo SIR metapoblacional con matriz de conectividad neutral. Tanto las condiciones iniciales como los números de habitantes son tomados de las datos oficiales reales.

Este fenómeno se ve reflejado sobre cantidades como la correlación promedio y el lag promedio. Para estas simulaciones, estas cantidades se encuentran graficadas en la Figura 7.3. Allí, se observa que CABA resulta tener una correlación promedio que se encuentra entre las más bajas del país. Algo análogo sucede con el lag promedio, que resulta ser de los más negativos del país, indicando que es una provincia que



(a) Correlaciones promedio para cada provincia según el modelo neutral.

(b) Lags promedio para cada provincia según el modelo neutral.

Figura 7.3: Correlaciones y lags promedio para cada provincia argentina, de acuerdo con el modelo neutral.

antecede temporalmente el brote en otras localidades. Esto es algo similar a lo que hemos destacado de las observaciones oficiales en el capítulo 6 para varias provincias fuentes de la epidemia.

Simulaciones con condiciones iniciales arbitrarias

Para visualizar más claramente este fenómeno, podemos recurrir a un caso más extremo en el cual dos provincias actuen como fuentes y el resto comiencen con una cantidad nula de infectados. Para ello, realizamos simulaciones asignando las condiciones iniciales realistas a CABA y BA, y considerando que no hay infectados en el resto de las provincias. Las series temporales resultantes se presentan en la Figura 7.4. En este caso límite, el resto de las jurisdicciones tienen exactamente la misma dinámica, de tal manera que no pueden distinguirse las diferentes curvas cuando se observan normalizadas.

Esto tiene su claro correlato en las matrices de correlación y de lag de la Figura 7.5. Entre las provincias que no son CABA ni BA, las correlaciones valen la unidad y los lags son nulos. Los valores son únicamente son diferentes en relación a CABA y BA. De esta manera, es lógico que las correlaciones promedio sean las menores y los lags sean tanto negativos como muy grandes en valor absoluto. Esto no quiere decir que BA o CABA no estén bien conectadas con el resto del país, como habríamos supuesto en el capítulo 6, sino todo lo contrario. Incluso, hemos encontrado un mecanismo para explicar por qué los valores de correlación y de lag tienen esa forma para provincias que son fuentes de la infección, cosa ya que habíamos señalado en dicho capítulo.

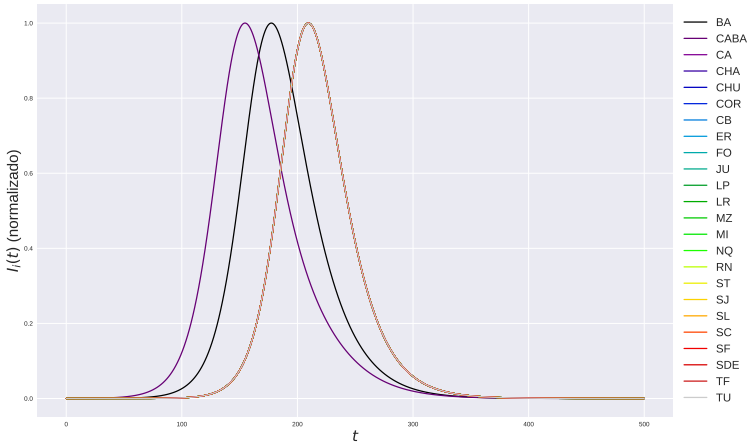
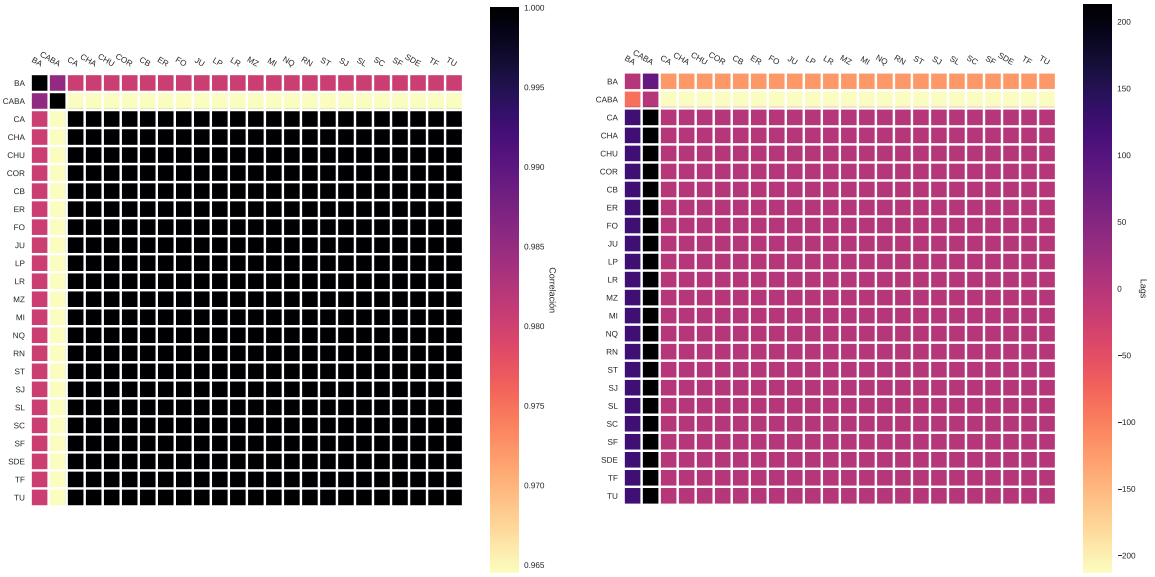


Figura 7.4: Series temporales obtenidas para el modelo neutral, con poblaciones reales y condiciones iniciales arbitrarias.



(a) Matriz de correlaciones entre provincias.

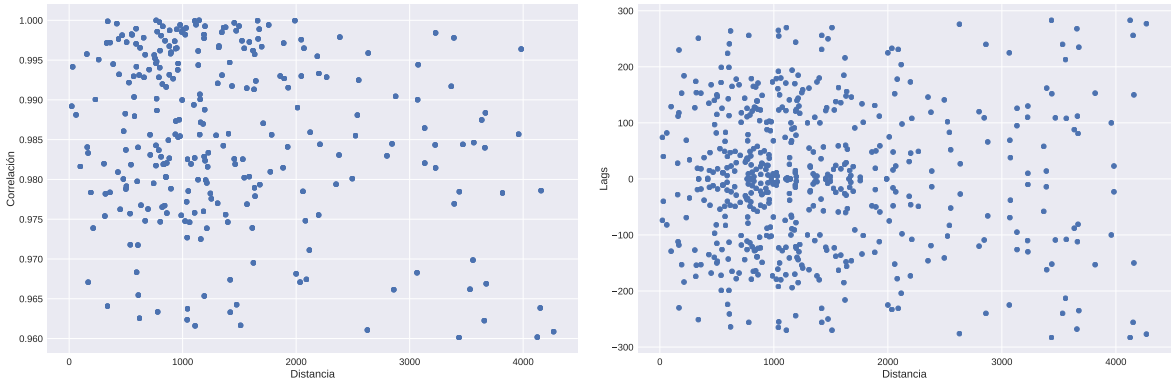
(b) Matriz de lags entre provincias.

Figura 7.5: Matrices de correlaciones y de lags entre provincias para el modelo SIR metapoblacional con matriz de conectividad neutral. Las poblaciones fueron tomadas de datos reales, mientras que las condiciones iniciales son arbitrarias.

Cabe destacar que se han realizado simulaciones análogas, tanto con condiciones iniciales realistas como arbitrarias, para el resto de los modelos presentados. En todos los casos, este mismo fenómeno se reproduce cualitativamente de la misma forma.

7.4.2. Relación entre correlación y lag con la distancia interprovincial

Por otra parte, todavía nos queda explicar el hecho de que no parece existir una relación aparente entre correlación o lag y la distancia interprovincial. Esto se ha observado claramente en la sección 6.7 del capítulo 6. Para verificar si podía reproducirse cualitativamente este tipo de comportamiento de nuestros modelos, realizamos simulaciones para las dos matrices de conectividad propuestas que tienen una dependencia con las distancias entre provincias. En ambos casos, utilizamos las condiciones iniciales realistas, al igual que las distancias y números de habitantes especificados anteriormente. Se utiliza como factor de proporcionalidad un valor arbitrario 0,01 en todas las simulaciones que siguen.



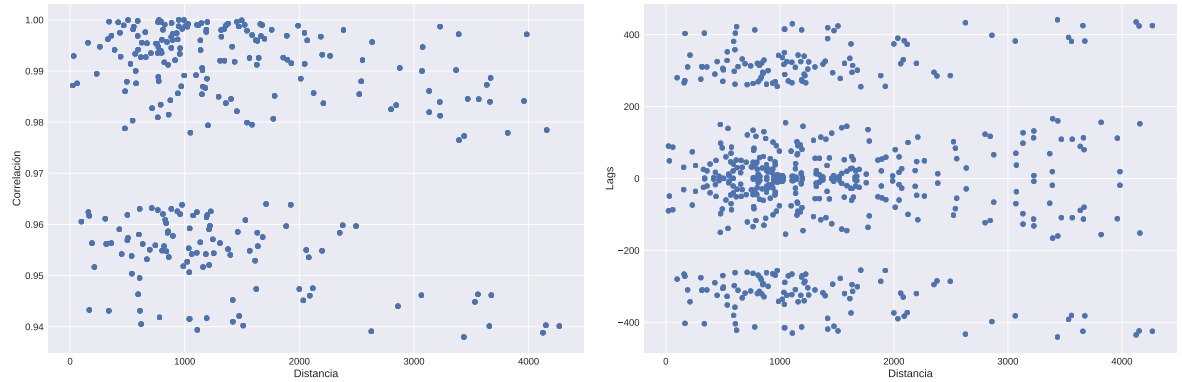
(a) Correlaciones en función de distancias entre provincias para el modelo dependiente de las poblaciones y de las distancias.

(b) Lag en función de distancias entre provincias para el modelo dependiente de las poblaciones y de las distancias.

Figura 7.6: Correlaciones y lags entre provincias en función a las distancias correspondientes, simuladas a partir del modelo dependiente del número de habitantes y de las distancias. Se destaca la falta de una relación clara entre las cantidades estudiadas y las distancias interprovinciales.

Primero, se realizaron simulaciones para el modelo para el cual la matriz de conectividad es inversamente proporcional a las distancias, sin depender del número de habitantes. Los gráficos resultantes de correlación y de lag en función de las distancias se presentan en la Figura 7.6. En este caso, vuelve a observarse la simetría del lag, como es de esperarse. Sin embargo, no se observa una dependencia clara de ninguna de las dos cantidades con la distancia.

Por otra parte, se observa algo similar para el modelo de matriz de conectividad dependiente de la distancia y la población de las diferentes jurisdicciones. Esto puede observarse en la Figura 7.7 donde nuevamente tenemos graficadas las correlaciones y los lags en función de la distancia entre provincias. Si bien no parece tener la misma forma que los gráficos anteriores, siguen observandose nubes de puntos donde la correlación no disminuye con la distancia ni el valor absoluto del lag aumenta con la misma.



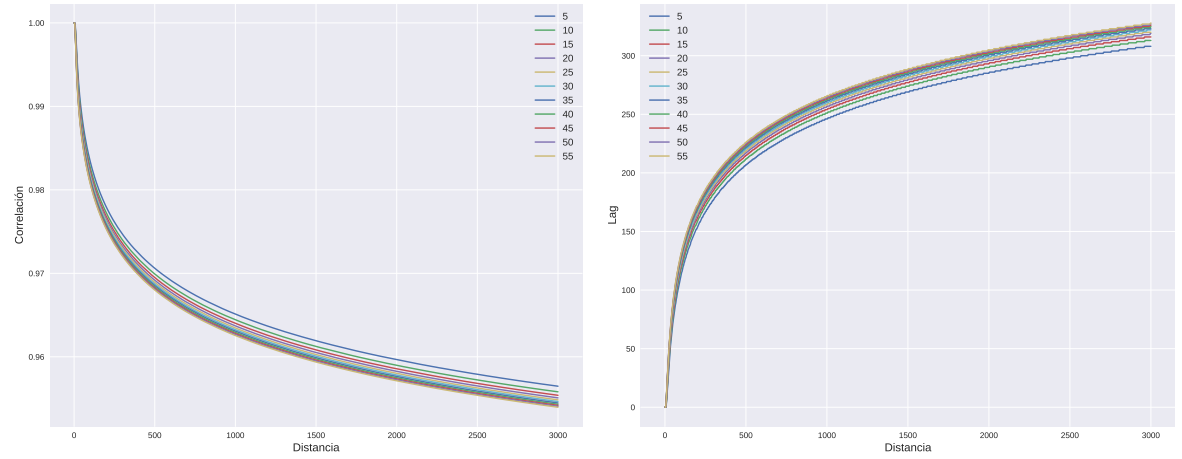
(a) Correlaciones en función de distancias entre provincias para el modelo dependiente de las poblaciones y de las distancias.

(b) Lags en función de distancias entre provincias para el modelo dependiente de las poblaciones y de las distancias.

Figura 7.7: Correlaciones y lags entre provincias en función a las distancias correspondientes, simuladas a partir del modelo dependiente del número de habitantes y de las distancias. Se destaca la falta de una relación clara entre las cantidades estudiadas y las distancias interprovinciales.

Simulaciones de los modelos con dos y tres subpoblaciones

A diferencia de lo que ocurre con las simulaciones anteriores, la distancia parece jugar un papel extremadamente importante en los valores del lag y de la correlación en modelos que se estudian para dos localidades en la bibliografia.



(a) Correlación entre dos subpoblaciones en función de la distancia entre ellas.

(b) Lag entre dos subpoblaciones en función de la distancia entre ellas.

Figura 7.8: Correlaciones y lags entre dos subpoblaciones en función de la distancia entre ellas, simulando un modelo donde la matriz de conectividad depende de la distancia. Se utilizan diferentes condiciones iniciales para una de las localidades, mientras que la otra permanece fija.

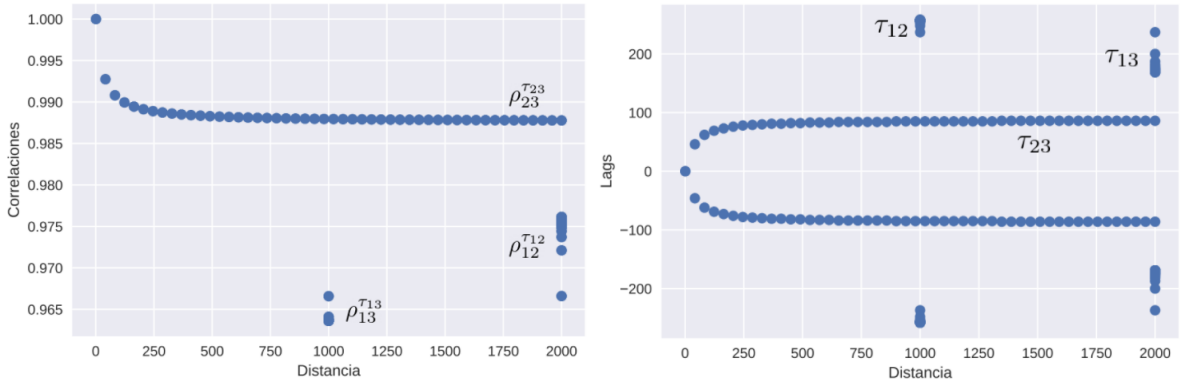
Para ver el comportamiento de nuestros modelos en este caso, realizamos simulaciones con dos subpoblaciones de 10^5 habitantes a una distancia variable. Consideraremos que una de ellas no tiene infectados inicialmente, mientras que la otra tiene una condición inicial que varía en un rango de $5-55$ infectados. Como ambas poblaciones tienen

la misma cantidad de habitantes, la matriz de conectividad termina dependiendo únicamente de la distancia. Se utiliza la misma constante de proporcionalidad que en las simulaciones inmediatamente anteriores. De esta manera, es posible obtener gráficos de correlación y de lag en función de la distancia entre localidades. Estos se corresponden a los gráficos de la Figura 7.8.

Allí, se observa que con dos localidades, el modelo de matriz de conectividad dependiente de la distancia (o de la población y de la distancia simultáneamente) genera correlaciones y lags que se comportan como es esperable intuitivamente. Ambas cantidades parecen aumentar o disminuir, acercándose a un valor asintótico máximo o mínimo, según el caso. Adicionalmente, el cambio de las condiciones iniciales no parece influir en este fenómeno cualitativamente.

Sin embargo, en la realidad se intenta modelar un sistema con muchas otras subpoblaciones en interacción mutua. Con el objetivo de intentar observar que sucede en ese caso, se llevaron a cabo simulaciones con un sistema de tres poblaciones distintas de 10^5 habitantes, a las que llamaremos subpoblaciones 1,2 y 3. Consideramos un esquema tal que la distancia entre las poblaciones 1 y 2, así como la distancia entre la 1 y la 3 son dadas. Las distancias son entonces $d_{12} = 2000$ y $d_{13} = 1000$. En cambio, la distancia d_{23} varía en un rango de $1 - 2000$. Finalmente, asumimos que las cantidades iniciales de infectados son $Y_1 = 0$, $Y_2 = 10$ y $Y_3 = 50$. Se probaron diferentes valores de distancias y condiciones iniciales, observándose en todos los casos resultados análogos.

Para los valores específicados, se grafican las correlaciones y los lags en función de las distancias simuladas en el gráfico 7.9. En cuanto a las correlaciones de la Figura 7.9a, se observa que la correlación entre las subpoblaciones 2 y 3 presenta una curva similar a la que observamos anteriormente. Es decir, disminuye con la distancia hasta alcanzar un valor asintótico. En cambio, las correlaciones ρ_{12} y ρ_{13} cambian su valor dependiendo de d_{23}, sin importar que la distancias d_{12} y d_{13} no se modifiquen. Con el lag, se observa un fenómeno similar.



(a) Correlaciones entre tres subpoblaciones en función de la distancia entre ellas.

(b) Lag entre tres subpoblaciones en función de la distancia entre ellas.

Figura 7.9: Correlaciones y lags entre tres subpoblaciones en función de la distancia entre ellas, simulando un modelo donde la matriz de conectividad depende de la distancia. La distancia entre dos de estas poblaciones es variable. Las condiciones iniciales son fijas.

Por un lado, de este análisis se desprende que en nuestros modelos tanto el número de poblaciones, las condiciones iniciales y las distancias tienen una influencia en la relación de la correlación o el lag con la distancia. Si bien para un par de poblaciones, parece ser correcto que su correlación disminuye con la distancia, esto no puede corroborarse exactamente en la realidad. Esto se debe a que dos poblaciones se encuentran a una distancia fija, que no puede ser cambiada. Así, si queremos comparar lo que sucede con otra distancia, tendremos que tomar otras dos poblaciones, que tienen distintos parámetros y se relacionan de manera diferente. En conclusión, dadas estas diferencias y dada la complejidad de tener varias subpoblaciones, los gráficos pueden dar la idea de que no hay una influencia de la distancia sobre el lag o la correlación, incluso si éste no es el caso.

7.5. Ajustes de los modelos propuestos

Tras hacer una exploración preliminar de los modelos y observar varias características interesantes, el próximo paso es hacer el ajuste de los mismos a los datos reales. Para ello, se intentaron utilizar diversos métodos y librerías de Python. Finalmente, se decidió por utilizar la librería LMFIT [27]. Primero, se investigó la posibilidad de hacer ajustes con tasas de contagios homogéneas. Luego, se introdujo una tasa de contagio diferente β_i para cada provincia.

7.5.1. Ajuste con β constante

Primero, se realizaron los ajustes con una tasa β homogénea. Se hicieron ajustes para los modelos presentados: el modelo neutral, el modelo donde la conectividad depende de la distancia, el modelo en el cual es dependiente de la población y el modelo

que tiene en cuenta ambas contribuciones. Adicionalmente, se tiene en cuenta el modelo SIR sin conectividad entre provincias como un control.

Para comparar la calidad de estos ajustes, se utilizaron tres indicadores proveídos por el paquete LMFIT: el chi cuadrado χ^2, el criterio de información de Akaike (AIC, por sus siglas en inglés) y el criterio de información de Bayes (BIC, por sus siglas en inglés). Cabe aclarar que cuanto más bajos resultan ser estos índices, el ajuste resulta ser más adecuado. Adicionalmente, otros dos parámetros son interesantes: $1/\gamma$ y β/γ. El primero nos da una idea del período de recuperación. En cambio, la segunda cantidad está relacionada con el número reproductivo estimado para la enfermedad.

Los resultados correspondientes se presentan en la tabla 7.1. De la misma, se desprenden muchas observaciones interesantes. Por un lado, se observan que los ajustes para el modelo neutral y el modelo dependiente únicamente de la distancia resultan tener una calidad inferior al control. Esto no sucede para los modelos que incorporan la población. Entre estos dos últimos modelos, el que propone una matriz de conectividad que depende de las poblaciones y las distancias parece ser el que más se ajusta a las observaciones de la realidad.

Modelo	χ^2	AIC	BIC	$1/\gamma$	$\beta/gamma$
Neutral	4.0675e+11	112688.464	112708.692	16.5	1.4564
Distancia	5.9167e+18	215999.626	216019.853	2.59	1.1170
Población	5.8385e+10	100529.070	100549.298	2.52	1.1165
Distancia + Población	4.2820e+10	98586.9646	98607.1923	2.64	1.116214
Control (sin conectividad)	5.8528e+10	100542.417	100555.902	2.51	1.116624

Tabla 7.1: Valores de χ^2,AIC,BIC,$1/\gamma$ y β/γ obtenidos del ajuste de los modelos presentados, asumiendo una tasa de contagio homogénea

En cuanto a los valores de β/γ, se tiene que son cercanos a la unidad en todos los casos. Esto es consistente con observaciones anteriores, especialmente presentadas en el capítulo 3. En cuanto a los valores de $1/\gamma$, tenemos que tienen un valor entre 2 y 3 días. Esto sucede en todos los casos, excepto para el modelo neutral, el cual posee indicadores de un ajuste de baja calidad. En un principio, estos valores pueden parecer erróneos, dado que suelen indicarse entre 10 y 14 días como periodo de recuperación de la COVID-19. Sin embargo, se ha reportado reiteradas veces en la bibliografía que una persona pierde la posibilidad de contagiar a otros alrededor de 3 días después del inicio de síntomas [28][29][30] dado el decremento de la carga viral con el tiempo. Adicionalmente, en nuestro modelo el período de recuperación $1/\gamma$ tiene la única función de señalar cuando una persona infectada puede contagiar a otras. Si pierde esta capacidad, no tiene sentido para nuestro modelo que una persona siga considerándose como infectada.

7.5.2. Ajuste con β_i para cada provincia

A continuación, se realizaron los ajustes de los modelos que permiten un β_i para cada provincia argentina. Se consideran las mismas propuestas de matriz de conectividad que en el caso homogéneo, al igual que los mismos indicadores del ajuste. Sin embargo, se remueve el valor de β/γ, dado que hay un único valor de la tasa de contagio. En cambio, analizaremos los valores β_i en otra sección a continuación. Con estas consideraciones hechas, los resultados de los nuevos ajustes se presentan en la tabla 7.2.

Modelo	χ^2	AIC	BIC	$1/\gamma$
Neutral	3.9811e+10	98176.4607	98351.7677	3.154
Distancia	2.5358e+10	95351.0877	95526.3946	2.726
Población	3.9644e+11	112573.650	112748.957	15
Distancia + Población	2.0796e+10	94108.6567	94283.9636	3.045
Control (sin conectividad)	5.8460e+10	100581.149	100749.713	2.508

Tabla 7.2: Valores de χ^2,AIC,BIC y $1/\gamma$ obtenidos del ajuste de los modelos presentados, asumiendo una tasa de contagio variable entre provincias.

En este caso, se tiene que el modelo de matriz de conectividad únicamente dependiente de las poblaciones se ajusta peor que el control sin conectividad correspondiente. En cambio, el ajuste con mejores indicadores corresponde a la propuesta dependiente de las distancias y las poblaciones. Para el modelo de solamente poblaciones influyendo en la matriz A_{ij}, el período de recuperación vale 15 días. Sin embargo, el ajuste de este modelo resulta ser inadecuado, como puede observarse de los indicadores. Este modelo también presentaba un ajuste gráfico claramente poco adecuado. En cambio, para el resto de los modelos, las tasas de recuperación corresponden nuevamente a un periodo de entre 2 y 3 días.

7.6. Análisis del ajuste óptimo

Hasta ahora, hemos realizado ajustes para los diferentes modelos, con distinta propuestas de matrices de conectividad. Es interesante observar que el ajuste con los mejores indicadores corresponde al modelo de tasa de contagio variable entre provincias tal que la matriz de conectividad depende de la distancia y de la población. Este modelo no solo resulta ajustarse mejor entre los modelos de β_i variable, sino que también es mejor que todos los modelos de tasa de contagio homogénea.

Por ello, nos dedicaremos a analizar en profundidad los parámetros que resultan del mismo. Previamente, ya hemos analizado el valor obtenido de $1/\gamma$. En este caso, se obtiene que $1/\gamma = 3{,}045$, lo cual es consistente con lo que puede hallarse en la bibliografía. En cambio, en esta sección estudiaremos los valores de β_i asignados a este

modelo. También, se presentan los gráficos del ajuste y de los datos, lo cual permite realizar un análisis gráfico de la bondad del ajuste.

7.6.1. Análisis de los valores de β_i

En la Figura 7.10 presentamos los valores de β_i con sus errores correspondientes, en función de las provincias. Se observa que la mayoría de estos valores se encuentran en el rango $0{,}3 - 0{,}4$, con la excepción de algunas jurisdicciones. Entre ellas, se encuentran las provincias de Catamarca, Formosa, La Pampa y Misiones, que tienen errores muy grandes para sus tasas de contagio. En cambio, los errores son muy pequeños para el resto del país. Esto puede deberse a la cantidad y la calidad de los datos reportados por estas provincias. Por ejemplo, Catamarca, Formosa y Misiones son las tres provincias con menor cantidad de casos en el período de tiempo contemplado.

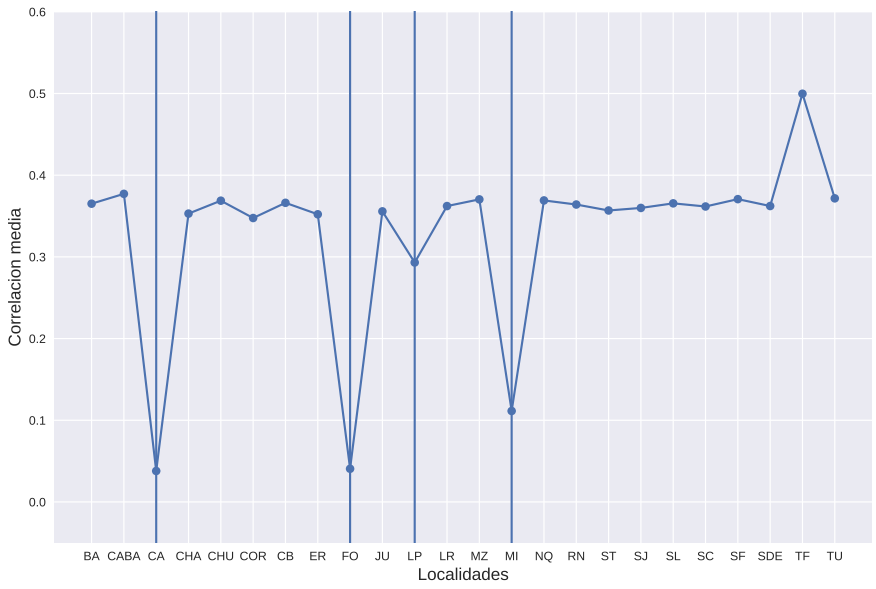
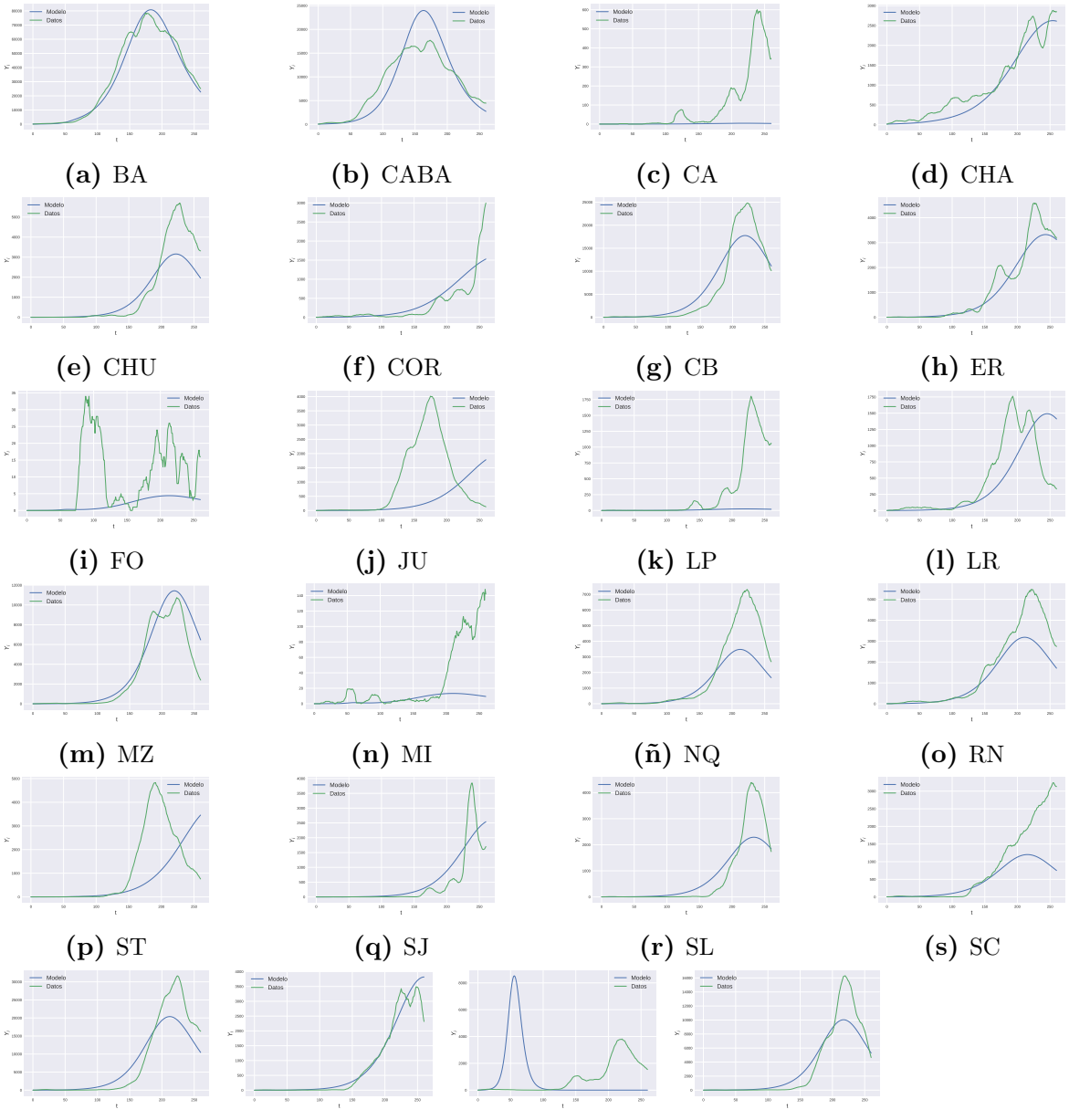


Figura 7.10: Valores de β_i obtenidos para el ajuste de matriz conectividad dependiente de la población y de la distancia entre provincias.

Por otra parte, cabe destacar que la provincia de Tierra del Fuego tiene la tasa de contagio más elevada del país. Una posible hipótesis para explicar esto es que su baja conexión con el resto del país, dado su bajo número de habitantes y su distancia con el resto de las provincias. De esta manera, solo una tasa de contagio tan elevada puede justificar la existencia de un brote tan importante en esta provincia, en el marco de nuestros modelos.

7.6.2. Gráficos resultantes

Finalmente, en la Figura 7.11 se grafican las series temporales provinciales.



(a) BA **(b)** CABA **(c)** CA **(d)** CHA

(e) CHU **(f)** COR **(g)** CB **(h)** ER

(i) FO **(j)** JU **(k)** LP **(l)** LR

(m) MZ **(n)** MI **(ñ)** NQ **(o)** RN

(p) ST **(q)** SJ **(r)** SL **(s)** SC

(t) SF **(u)** SDE **(v)** TF **(w)** TU

Figura 7.11: Datos oficiales de cada una de las provincias, junto con el ajuste correspondiente al modelo con tasas de contagio inhomogéneas, con una matriz de conectividad dependiente de las poblaciones y las distancias interprovinciales.

Adicionalmente, presentamos los ajustes correspondientes al modelo en estudio para cada una de las jurisdicciones. En la mayoría de los casos, se tiene que las curvas ajustadas siguen cualitativamente a los datos oficiales. Si bien no llegan a capturar pequeñas fluctuaciones de infectados, se ajustan razonablemente. Por ejemplo, en el caso de la Provincia de Buenos Aires o Mendoza. En otros casos, las curvas ajustadas no son cuantitativamente exactas, mas siguen una tendencia cualitativa. Esto puede

visualizarse en el caso de CABA o de la provincia de Río Negro. Sin embargo, las curvas del ajuste fallan totalmente a la hora de realizar predicciones con exactitud y con precisión en algunas provincias como Jujuy, La Pampa o Tierra del Fuego. Por un lado, puede pensarse que La Pampa tiene tanto pocos casos como pocos testeos y que Tierra del Fuego es particular por su posición geográfica. En el caso de Jujuy, esto podría deberse al ingreso de personas por fronteras informales, como hemos destacado previamente.

7.7. Conclusiones

En este capítulo, el objetivo principal ha sido introducir modelos para hallar posibles explicaciones para las características de los lags y de las correlaciones provinciales. De esta manera, hemos dedicado varias secciones al estudio de modelos metapoblacionales del tipo SIR. Así, hemos postulado modelos de diferentes subpoblaciones conectadas por una matriz de conectividad A_{ij}, con tasas de contagio homogéneas y variables. Adicionalmente, hemos tenido que indagar sobre posibles matrices A_{ij}. Para aplicar estos modelos, hemos asociado las diferentes subpoblaciones de un modelo metapoblacional con las provincias argentinas. A partir de esta asociación, nos hemos dedicado a analizar dos características centrales señaladas en el capítulo 6: los valores de correlación y del lag promedios; y la relación de la correlación y el lag con la distancia. En ambos casos, hemos encontrado que este tipo de modelos puede reproducir cualitativamente estos comportamientos para una población de 24 jurisdicciones.

Adicionalmente, se han realizado ajustes de todos los modelos propuestos a los datos oficiales. El modelo que más se ajusta a los datos es un modelo de tasa de contagios heterogénea, con una matriz de conectividad que depende tanto de la población como de la distancia entre provincias. Uno de los parámetros más relevantes que surgen de este ajuste es un valor de $1/\gamma = 3{,}045$, es decir, un periodo estimado de recuperación de 3 días después del inicio de síntomas. Si bien esto puede parecer erróneo, hay evidencia en la bibliografía de que este es el período correspondiente a la mayor contagiosidad de un infectado. Finalmente, hemos analizado tanto los valores de β_i obtenidos, como los gráficos de las curvas de ajuste en comparación con los datos reales. Hemos observado que el ajuste es cualitativamente bueno en la mayoría de los casos. Sin embargo, sería interesante continuar la investigación de modelos de este tipo. En particular, para poder reproducir la dinámica de algunas provincias, como Tierra del Fuego. Finalmente, como un trabajo a continuación, sería interesante introducir modificaciones para que estos modelos puedan aplicarse al resto de los brotes de la COVID-19 en el país, que fueron observados durante el año 2021.

Códigos utilizados

Todo el código utilizado para la optimización de pesos con RN del capítulo 4 fue de elaboración propia para esta tesis, tras el aprendizaje del funcionamiento de las RN y de las librerías específicas de Python utilizadas, como Keras y Tensorflow. Por otra parte, para la resolución de los sistemas de ecuaciones del capítulo 5 el código también fue de desarrollo propio para este trabajo y para ello se utilizó nuevamente el lenguaje de programación Python. Finalmente, se hizo uso, previo aprendizaje, de herramientas de desarrollo colaborativo de software, específicamente Github. El link del repositorio donde se procedió a almacenar los códigos de esta tesis y tener un control de sus diferentes versiones es: https://github.com/denisecammarota/tesis-ib.

Apéndice A

Abreviaturas de poblaciones del capítulo 4

A.1. Abreviaturas de provincias.

En la tabla A.1 se presentan las abreviaciones y los nombres correspondientes de la provincias utilizadas en el análisis del retraso óptimo del capítulo 4, al igual que a lo largo los capítulos 6 y 7. Adicionalmente, presentamos las condiciones iniciales a las que hacemos referencias en el capítulo 7. Como hemos explicado, para cada provincia se considera como número inicial de infectados al valor de la incidencia a $t = 70$. Esto, en términos matemáticos, es la tasa $A^{14}_{t=70}$.

A.2. Abreviaturas de localidades.

En la tabla A.1 se presentan las abreviaciones y los nombres correspondientes de las localidades utilizadas en el análisis del retraso óptimo del capítulo 4. Específicamente, datos de estas localidades fueron empleados en la confección de las Figuras 4.1 y 4.2.

Abreviatura	Provincia	$A^{14}_{t=70}$
CABA	Ciudad Autónoma de Buenos Aires	57
BA	Provincia de Buenos Aires	59
CA	Provincia de Catamarca	0
CB	Provincia de Córdoba	7
COR	Provincia de Corrientes	3
CH	Provincia de Chaco	17
CHU	Provincia de Chubut	1
ER	Provincia de Entre Ríos	4
FO	Provincia de Formosa	0
JU	Provincia de Jujuy	1
LP	Provincia de La Pampa	0
LR	Provincia de La Rioja	1
MZ	Provincia de Mendoza	2
MI	Provincia de Misiones	0
NQ	Provincia de Neuquén	2
RN	Provincia de Río Negro	6
ST	Provincia de Salta	0
SJ	Provincia de San Juan	0
SL	Provincia de San Luis	1
SC	Provincia de Santa Cruz	3
SF	Provincia de Santa Fe	4
SDE	Provincia de Santiago del Estero	0
TF	Provincia de Tierra del Fuego, Antártida e Islas del Atlántico Sur	3
TU	Provincia de Tucumán	2

Tabla A.1: Abreviaturas de todas las provincias junto con sus nombres oficiales y las condiciones iniciales, correspondientes a la incidencia al día $t = 70$.

Abreviatura	Localidad
AV	Municipio de Avellaneda, Provincia de Buenos Aires
BB	Municipio de Bahía Blanca, Provincia de Buenos Aires
BRC	San Carlos de Bariloche, Provincia de Río Negro
CABA	Ciudad Autónoma de Buenos Aires
CM 8	Comuna 8, CABA
CBC	Ciudad de Córdoba, Provincia de Córdoba
GR	Municipio de General Roca, Provincia de Río Negro
LM	Municipio de La Matanza, Provincia de Buenos Aires
MZC	Ciudad de Mendoza, Provincia de Mendoza
SLC	Ciudad de Salta, Provincia de Salta
ORÁN	Orán, Provincia de Salta
SFC	Ciudad de Santa Fe, Provincia de Santa Fe
ROS	Rosario, Provincia de Santa Fe
TUC	Ciudad de Tucumán, Tucumán

Tabla A.2: Abreviaturas de localidades utilizadas en las Figuras 4.1 y 4.2 del capítulo 4.

Bibliografía

[1] Ben Hu, P. Z., Hua Guo, Shi, Z.-L. Characteristics of sars-cov-2 and covid-19. *Nature Reviews Microbiology (Online)*, pág. 1–14, 2020. URL https://www.frontiersin.org/article/10.3389/fpubh.2020.00383. xi, xi, 1, 2, 3

[2] Matt J. Keeling, P. R. Modeling Infectious Diseases in Humans and Animals. 1ª edón. Princeton University Press, 2007. xii, 6, 31, 43, 58, 61

[3] Chams, N., Chams, S., Badran, R., Shams, A., Araji, A., Raad, M., *et al.* Covid-19: A multidisciplinary review. *Frontiers in Public Health*, **8**, 383, 2020. 1

[4] OMS. Coronavirus disease (covid-19): How is it transmitted?, Dic 2020. URL https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19\-how-is-it-transmitted. 1

[5] Dirección Nacional de Epidemiología y Análisis de Situación de Salud. COVID-19. Casos registrados en la República Argentina. Datos Abiertos del Ministerio de Salud., May 2020. URL http://datos.salud.gob.ar/dataset/covid-19-casos-registrados-en-la-republica-argentina. 5

[6] Laneri, K. F., Kolton, A. B. Diagramas de riesgo: Significado, implementación y uso, Aug 2020. URL https://droyktton.github.io/loscoihues/tutorial/tutorial.pdf. 5, 6, 8

[7] Computational Biology and Complex Systems. BIOCOMSC. Risk diagrams, Jan 2021. URL https://biocomsc.upc.edu/en/covid-19/Risk%20Diagrams. 8

[8] Pathria R.K., B. P. Statistical mechanics. 3ª edón. Elsevier, 2011. 14

[9] Stanley, H. E. Introduction to Phase Transitions and Critical Phenomena. Monographs on Physics, first edition edón. Oxford University Press, 1971. 14

[10] Chollet, F., *et al.* Keras, 2015. URL https://keras.io. 23

[11] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., *et al.* TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL http://tensorflow.org/, software available from tensorflow.org. 23

[12] Chollet, F., *et al.* Keras API Reference / Layers API / Core layers / Dense layer, 2015. URL https://keras.io/api/layers/core_layers/dense/. 25

[13] Chollet, F., *et al.* Keras API reference / Layers API / Core layers / Lambda layer, 2015. URL https://keras.io/api/layers/core_layers/lambda/. 25

[14] Chollet, F., *et al.* Keras API reference / Layers API / Merging layers / Multiply layer, 2015. URL hhttps://keras.io/api/layers/merging_layers/multiply/. 25

[15] Chollet, F., *et al.* Keras API Reference/ Optimizers/ Adam, 2015. URL https://keras.io/api/optimizers/adam/. 25

[16] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., *et al.* Array programming with NumPy. *Nature*, **585** (7825), 357–362, sep. 2020. 25

[17] Manrubia, S., Zanette, D. H. Individual risk-aversion responses tune epidemics to critical transmissibility ($r = 1$), 2021. 34

[18] Rearte, A., Baldani, A. E. M., Barbeira, P. B., Domínguez, C. S., Laurora, M. A., Pesce, M., *et al.* Características epidemiológicas de los primeros 116 974 casos de covid-19 en argentina, 2020. *Rev. Argent. Salud Pública*, **12** (1), 5, oct. 2020. 42

[19] Amnistía Internacional. Formosa: Amnistía Internacional advierte sobre el uso desmedido de cuarentenas obligatorias en centros de aislamiento, 2021. URL https://amnistia.org.ar/formosa-amnistia-internacional-advierte-sobre-el-uso-desmedido-de-\cuarentenas-obligatorias-en-centros-de-aislamiento/, accedido el 7 de enero de 2022. 42

[20] Instituto Geográfico Nacional. Distancias entre Ciudades Capitales, 2022. URL https://www.ign.gob.ar/NuestrasActividades/Geografia/DatosArgentina/DistanciasCiudades, accedido el 13 de enero de 2022. 58, 64

[21] Brozak, S., Pant, B., Safdar, S., Gumel, A. Dynamics of covid-19 pandemic in India and Pakistan: A metapopulation modelling approach. *Infectious Disease Modelling*, **6**, 10 2021. 61

[22] Coletti, P., Libin, P., Petrof, O., Willem, L., Abrams, S., Herzog, S., *et al.* A data-driven metapopulation model for the belgian covid-19 epidemic: assessing the impact of lockdown and exit strategies, 07 2020. 61

[23] Calvetti, D., Hoover, A. P., Rose, J., Somersalo, E. Metapopulation network models for understanding, predicting, and managing the coronavirus disease covid-19. *Frontiers in Physics*, **8**, 2020. URL https://www.frontiersin.org/article/10.3389/fphy.2020.00261. 61

[24] Arenas, A., Cota, W., Gómez-Gardeñes, J., Gomez, S., Granell, C., Matamalas, J., *et al.* Modeling the spatiotemporal epidemic spreading of covid-19 and the impact of mobility and social distancing interventions. *Physical Review X*, **10**, 12 2020. 61

[25] Ilkka A. Hanski, M. E. G. Metapopulation Biology: Ecology, Genetics and Evolution. 1ª ed$^{\text{ón}}$. Academic Press, 1997. 61

[26] Instituto Nacional de Estadística y Censos (INDEC). Proyecciones nacionales, 2022. URL https://www.indec.gob.ar/indec/web/Nivel4-Tema-2-24-84, accedido el 19 de enero de 2022. 64

[27] Newville, M., Otten, R., Nelson, A., Ingargiola, A., Stensitzki, T., Allan, D., *et al.* lmfit/lmfit-py 0.9.12, nov. 2018. URL https://doi.org/10.5281/zenodo.1699739. 72

[28] Cevik, M., Marcus, J., Buckee, C., Smith, T. Sars-cov-2 transmission dynamics should inform policy. *Clinical Infectious Diseases*, **73**, 09 2020. 73

[29] He, X., Lau, E., Wu, P., Deng, X., Wang, J., Hao, X., *et al.* Temporal dynamics in viral shedding and transmissibility of covid-19. *Nature Medicine*, **26**, 672–675, 02 2020. 73

[30] Centers for Disease Control and Prevention (CDC). CDC Updates and Shortens Recommended Isolation and Quarantine Period for General Population, 2021. URL https://www.cdc.gov/media/releases/2021/s1227-isolation-quarantine-guidance.html, accedido el 19 de enero de 2021. 73

[31] Cevik, M., Marcus, J., Buckee, C., Smith, T. Sars-cov-2 transmission dynamics should inform policy. *Clinical Infectious Diseases*, **73**, 09 2020.