Universidad de San Carlos de Guatemala
Escuela de Ciencias Físicas y Matemáticas
Departamento de Física

MODELADO DE LA TASA DE PROPAGACIÓN DEL COVID-19 EN GUATEMALA MEDIANTE MODELOS ARIMA

Carlos Roberto Salazar Mateo

Asesorado por M.Sc. Juan Diego Chi-Wen Chang Santizo
y Ph.D. Juan Adolfo Ponciano Castellanos

Guatemala, Septiembre de 2023

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA



ESCUELA DE CIENCIAS FÍSICAS Y MATEMÁTICAS

MODELADO DE LA TASA DE PROPAGACIÓN DEL COVID-19 EN GUATEMALA MEDIANTE MODELOS ARIMA

TRABAJO DE GRADUACIÓN
PRESENTADO A LA JEFATURA DEL
DEPARTAMENTO DE FÍSICA
POR

CARLOS ROBERTO SALAZAR MATEO

ASESORADO POR M.SC. JUAN DIEGO CHI-WEN CHANG SANTIZO
Y PH.D. JUAN ADOLFO PONCIANO CASTELLANOS

AL CONFERÍRSELE EL TÍTULO DE

LICENCIADO EN FÍSICA APLICADA

GUATEMALA, SEPTIEMBRE DE 2023

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

ESCUELA DE CIENCIAS FÍSICAS Y MATEMÁTICAS



CONSEJO DIRECTIVO INTERINO

Director	M.Sc. Jorge Marcelo Ixquiac Cabrera
Representante Docente	Arqta. Ana Verónica Carrera Vela
Representante Docente	M.A. Pedro Peláez Reyes
Representante de Egresados	Lic. Urías Amitaí Guzmán García
Representante de Estudiantes	Elvis Enrique Ramírez Mérida
Representante de Estudiantes	Oscar Eduardo García Orantes
Secretario	Ing. Edgar Damián Ochoa Hernández

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

Director	M.Sc. Jorge Marcelo Ixquiac Cabrera
Examinador	M.Sc. Osmar Obdulio Hernández Aguilar
Examinador	M.Sc. Erick Estuardo Hernández García
Examinador	M.Sc. Ricardo Enrique Contreras Folgar
Secretario	Ing. Edgar Damián Ochoa Hernández

Escuela de Ciencias Físicas y Matemáticas

Ref. D.DTG. 002-2023
Guatemala 13 de julio de 2023

El Director de la Escuela de Ciencias Físicas y Matemáticas de la Universidad de San Carlos de Guatemala, luego de conocer la aprobación por parte del Coordinador de la Licenciatura en Física Aplicada, al trabajo de graduación titulado: "**Modelado de la Tasa de Propagación del COVID-19 en Guatemala mediante Modelos ARIMA",** presentado por el estudiante universitario Carlos Roberto Salazar Mateo, autoriza la impresión del mismo.

IMPRÍMASE.

"ID Y ENSEÑAD A TODOS"

M.Sc. Jorge Marcelo Ixquiac Cabrera
Director

JMIC/Paola

AGRADECIMIENTOS

A mis papás Julio Salazar y Eulalia Mateo por creer en mí y apoyarme en todo momento a pesar de las dificultades. A Kevin, mi hermano, por su incondicional amistad y porque jugábamos de niños a ser científicos. A mi tío Gonzalo quien siempre me apoyó y estuvo pendiente de mi carrera.

A todos mis amigos con los que alguna vez tuve conversaciones que generaban en mí un mayor interés por continuar estudiando física.

A mi maestra de primaria Deyli Gómez por despertar mi curiosidad por la matemática, cuando enseñó cierta fórmula que a sus estudiantes nos parecía muy divertida, ahora sé que se trataba de la definición del número de Euler.

A mi maestra de diversificado, Vilma, quien me habló de la existencia de la Licenciatura en Física.

A mis compañeros y profesores de la ECFM que dedicaron tiempo a organizar los congresos de física y matemática de dónde también obtenía mucha inspiración para continuar estudiando.

A los organizadores del curso Programación para Científicos, de Física sin Fronteras en 2018, ya que fue de gran importancia para mí pues impulsó significativamente mi carrera profesional.

A mis asesores Ph.D. Juan Ponciano y M.Sc. Juan Chang por su guía en esta investigación, y por compartirme su conocimiento a lo largo de la carrera.

DEDICATORIA

A mis papás, Julio Salazar y Eulalia Mateo, y a mi hermano Kevin Salazar.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS v

ÍNDICE DE TABLAS vii

LISTA DE SÍMBOLOS ix

OBJETIVOS xi

INTRODUCCIÓN xiii

JUSTIFICACIÓN xv

1. MODELO SIR 1

1.1. Análisis de Estabilidad 2

1.1.1. Relación de Tamaño Final 2

1.2. Determinación de la Tasa de Contacto β 9

2. REDES NEURONALES 11

2.1. Agrupación de Datos 13

2.2. Función de Costo 14

2.3. Algoritmo de Descenso de Gradiente 15

2.4. Descenso de Gradiente Estocástico 16

3. PROCESOS ESTOCÁSTICOS 17

3.1. Procesos Estocásticos Estacionarios 17

3.1.1. Prueba ADF 18

3.1.2. Modelos de Series de Tiempo Estocásticas 18

3.1.2.1. Modelos Autorregresivos 18

3.1.2.2. Modelos de Medias Móviles 22

3.1.3. Determinación del Orden de los Modelos Autoregresivos y de Medias Móviles 23

3.1.4. Modelos ARMA y ARIMA 26

4. METODOLOGÍA PARA PRONOSTICAR β APLICANDO REDES NEURONALES 27

5. METODOLOGÍA PARA PRONOSTICAR β APLICANDO MODELOS ARIMA 29

6. DETERMINACIÓN DE β UTILIZANDO REDES NEURONALES 31

7. MODELADO DE β COMO PROCESO ESTOCÁSTICO. 39

7.1. Tercera Ola de Infectados 41

7.2. Cuarta Ola de Infectados 49

8. DISCUSIÓN DE RESULTADOS 57

CONCLUSIONES 61

RECOMENDACIONES 63

ANEXOS 73

BIBLIOGRAFÍA 79

ÍNDICE DE FIGURAS

1.1. Función g(x) para estudiar la solución de la relación de tamaño final. 6
1.2. Primera amplificación de la gráfica de g(x), con $R_0 > 1$. 6
1.3. Segunda amplificación de la gráfica de g(x), con $R_0 > 1$. 7
1.4. Función g(x) con $R_0 < 1$. 7
1.5. Primera amplificación de la gráfica de g(x) con $R_0 < 1$. 8
1.6. Segunda amplificación de la gráfica de g(x) con $R_0 < 1$. 8

2.1. Neurona artificial. 12
2.2. Red neuronal. 12
2.3. Red neuronal, segunda parte. 13
2.4. Función de costo. 15

3.1. Ejemplo de serie obtenida a partir de distribución normal. 23
3.2. Funciones de autocorrelación y autocorrelación parcial. 24
3.3. Movimiento browniano. 25
3.4. Funciones de autocorrelación y autocorrelación parcial del movimiento browniano. 25

4.1. Red neuronal utilizada para predecir β. 28

6.1. Función de costo de la red neuronal. 31
6.2. Predicciones de β para diferentes días. 31
6.3. Distribución de pronósticos de β. 32
6.4. Comparación de los pronósticos de la red neuronal con los valores observados de β. 34
6.5. Comparación de δA observado en la muestra con las predicciones realizadas. . . 35
6.6. Comparación de los pronósticos de acumulados con los valores observados. . . 36
6.7. Comparación de los pronósticos de los susceptibles con los valores observados. 37
6.8. Comparación de los pronósticos de diferentes variables con los valores observados. 38

7.1. Primeras olas de infectados. 40

7.2. Duración de la ola suponiendo β constante. 40
7.3. Duración de la ola suponiendo dos valores de β diferentes. 41
7.4. Test ADF en la serie previa a la tercera ola. 41
7.5. Test ADF en la nueva serie. 42
7.6. Funciones de autocorrelación y autocorrelación parcial de la nueva serie. 42
7.7. Predicciones de β para diferentes modelos ARIMA. 43
7.8. Proyecciones de β y δA. 44
7.9. ACF y PACF para β del departamento de Guatemala. 45
7.10. Error cuadrático medio en ARIMA(2,0,q). 45
7.11. Modelos ARIMA(2,0,7) y ARIMA(2,0,13). 46
7.12. Predicciones de ARIMA(2,0,7). 47
7.13. Predicciones de β para ARIMA(2,0,13). 48
7.14. Estimación de A mediante ARIMA(2,0,13). 49
7.15. Test ADF en serie previa a la cuarta ola. 49
7.16. ACF y PACF para serie previa a cuarta ola. 50
7.17. Error cuadrático medio para ARIMA(3,0,q). 50
7.18. Predicciones de β y δA mediante ARIMA(3,0,14). 51
7.19. Predicciones de β y δA mediante ARIMA(3,0,14), segunda parte. 52
7.20. Predicciones de I y S mediante ARIMA(3,0,14). 53
7.21. Predicciones de I y S mediante ARIMA(3,0,14), segunda parte. 54
7.22. Predicciones de R y A mediante ARIMA(3,0,14). 55
7.23. Predicciones de R y A mediante ARIMA(3,0,14), segunda parte. 56

8.1. Test ADF en la serie del departamento de Santa Rosa. 58
8.2. Error cuadrático medio para modelo ARIMA(1,0,q). 59
8.3. Estimación de β según ARIMA(1,0,8) para el departamento de Santa Rosa. . . . 59
8.4. Pronóstico de A mediante ARIMA(1,0,8) para el departamento de Santa Rosa. . 60
8.5. Pronóstico de A mediante ARIMA(1,0,8) para el departamento de Santa Rosa, segunda parte. 60

8.6. Puntos de inicio y fin de algunas olas. 63
8.7. Simulación de β como función triangular. 64
8.8. β asociado a los nuevos reportes diarios. 65
8.9. Modelado de β como una función triangular. 66
8.10. Pronóstico de A y S suponiendo a β triangular. 67
8.11. Pronóstico de I y R suponiendo a β triangular. 67
8.12. δA para diferentes β y w. 68

8.13. δA para diferentes β y w, segunda parte. 69
8.14. Influencia de la duración de la fluctuación β en el máximo δA. 70
8.15. Ajuste lineal de $ln(\delta A_{max}$ normalizado$)$ vs. $ln(C_p)$. 72
8.16. Ajuste lineal de $ln(\delta A_{max}$ normalizado$)$ vs. $ln(C_s)$. 72

8.17. Script para la determinación de β. 73
8.18. Script para manipulación de datos de entrenamiento y de prueba. 74
8.19. Script para la creación de la red neuronal. 74
8.20. Script para la creación de la función de costo. 74
8.21. Script para la ejecución de la red neuronal. 75
8.22. Parámetros del modelo ARIMA(2,0,13). 76
8.23. Parámetros del modelo ARIMA(3,0,14). 77
8.24. Parámetros del modelo ARIMA(1,0,8). 78

ÍNDICE DE TABLAS

1.1. Ejemplo del cálculo de las variables δA, A, I y R. 9

8.1. Departamentos cuya serie $\beta(t)$ es no estacionaria. 58

8.2. β observado al alacanzar el máximo de reportes diarios en cada ola. 65

8.3. Promedio de β en el que se espera alcanzar el máximo de reportes diarios. . . . 66

8.4. Proyecciones de β para una de las corridas de la red. 75

LISTA DE SÍMBOLOS

Símbolo	Significado
β	Tasa de propagación de la infección, o tasa de contacto efectiva.
S	Individuos susceptibles a infectarse.
I	Individuos infectados.
R	Individuos removidos.
A	Individuos infectados acumulados.
δA	Nuevos individuos infectados reportados cada día.
N	Población total de un departamento.
γ	Tasa de recuperación de la infección.
R_0	Número de reproducción básico.
L	Operador de retardos.
γ_0	Varianza de una serie.
γ_τ	Autocovarianza con un desfase de τ unidades.
ρ_τ	Autocorrelación, equivale a γ_τ / γ_0.
ϵ_k	Término de ruido blanco al tiempo k.
σ_ϵ	Desviación estándar de la distribución de ruido blanco.
ϕ_i	$i-$ésimo coeficiente de un modelo autorregresivo.
θ_j	$j-$ésimo coeficiente de un modelo de medias móviles.
ACF	Siglas en inglés para la función de autocorrelación.
PACF	Siglas en inglés para la función de autocorrelación parcial.

OBJETIVOS

General

- Realizar un modelo matemático para proyecciones de los casos acumulados de Covid-19 para el departamento de Guatemala.

Específicos

- Calcular las tasas de contagio y proponer métodos para pronosticar los valores futuros.
- Utilizar los pronósticos de las tasas de contagio para proyectar la evolución del total de infectados acumulados y de individuos susceptibles.

INTRODUCCIÓN

La pandemia de Covid-19 ha tenido un gran impacto en todo el mundo, y Guatemala no ha sido la excepción. La comprensión de la propagación de la enfermedad es crucial para poder tomar medidas efectivas con el fin de controlarla. En este contexto, los modelos compartimentales como el modelo SIR se han utilizado ampliamente para estudiar la evolución de la Covid-19. Sin embargo, la tasa de propagación de la enfermedad no es constante, sino que varía con el tiempo y las circunstancias, como épocas festivas en las que se relajan las restricciones, lo que dificulta la descripción precisa del desarrollo de la pandemia.

Para abordar esta limitación y bajo la suposición de que la tasa de propagación es un proceso estocástico, se propone determinar su variación mediante la combinación de los modelos de autorregresión lineal y de medias móviles.

Otro método para modelar la variación de la tasa de propagación es el de redes neuronales artificiales, una técnica de aprendizaje automático que ha demostrado ser efectiva en muchos campos. En el caso de la Covid-19, se ha demostrado que las redes neuronales artificiales pueden usarse para pronosticar el número de nuevos casos de infectados en diferentes países, como se puede observar en [2].

La tasa de propagación para diferentes días se calcula a partir de los nuevos casos diarios reportados por el Ministerio de Salud de Guatemala. Los pronósticos realizados por los modelos ARIMA y por la red neuronal se sustituyen en el modelo SIR para estimar las proyecciones de los casos acumulados. Finalmente se comparan estas estimaciones con los datos registrados.

JUSTIFICACIÓN

La modelación matemática de epidemias ha sido un campo de investigación activo durante varios siglos, y ha sido utilizado para comprender y predecir la propagación de enfermedades infecciosas en todo el mundo. Uno de los primeros en modelar matemáticamente la evolución de una epidemia fue el físico y matemático Daniel Bernoulli en el siglo XVIII cuando analizó la mortalidad causada por la viruela tratando de modelar la cantidad de sujetos que a cierta edad aún no habían contraído la enfermedad [3], lo que inició todo un campo de investigación en modelado epidemiológico.

Desde entonces, se han desarrollado varios modelos para entender la propagación de enfermedades infecciosas, incluyendo el modelo SIR (para estimar las variables susceptibles, infectados y recuperados) y sus variantes, que han sido utilizados para comprender la evolución de la pandemia de COVID-19. El modelo SIR es particularmente efectivo para simular la fase inicial de la enfermedad, cuando aún no han hecho efecto las medidas de restricción. Ya que en tales casos se puede considerar que el número de reproducción básico R_0 y por lo tanto la tasa de contacto de propagación β, una cantidad proporcional a R_0, son constantes. Por ejemplo, en [5] se determina el número de reproducción básico R_0 de diversos países para los primeros días de la infección. Sin embargo, este modelo deja de dar resultados adecuados para el resto de la evolución de la epidemia, sobreestimando el tiempo de duración de las olas. Por lo tanto, determinar las fluctuaciones de β es crucial para estimar el desarrollo de la enfermedad y conducir a la adopción de medidas más efectivas para controlarla.

1. MODELO SIR

El SIR es un modelo compartimental que fue propuesto por R. Ross [15]. Describe la evolución de los individuos susceptibles, infectados y removidos. Las ecuaciones de este modelo son

$$\begin{cases} \frac{dS}{dt} = -\frac{\beta IS}{N} \\ \frac{dI}{dt} = \frac{\beta IS}{N} - \gamma I \\ \frac{dR}{dt} = \gamma I \end{cases} \tag{1.1}$$

donde la tasa de propagación está representada con β y la tasa de recuperación con γ, en este proyecto se supone que $\gamma = 1/21$ ya que el IGSS considera que un individuo infectado se recupera luego de 21 días [7].

El modelo SIR en series finitas y suponiendo $\Delta t = 1$ es

$$\begin{cases} S_j = S_{j-1} - \frac{\beta_{j-1} I_{j-1} S_{j-1}}{N} \\ I_j = I_{j-1} + \frac{\beta_{j-1} I_{j-1} S_{j-1}}{N} - \gamma I_{j-1} \\ R_j = R_{j-1} + \gamma I_{j-1} \end{cases} \tag{1.2}$$

note que β corresponde exactamente a la tasa de personas contagiadas al introducir un individuo infectado entre la población inicialmente sana.

Por la condición de restrición

$$N = S + I + R, \tag{1.3}$$

solo se necesitan las primeras dos ecuaciones del sistema 1.1. Las condiciones iniciales son $S(0) = S_0$, $I(0) = I_0$ y $S_0 + I_0 = N$.

1.1. Análisis de Estabilidad

$S(t)$ es una función decreciente ya que $S'(t) < 0$ para todo t. Por otro lado $I'(t) > 0$ si

$$\frac{\beta}{\gamma}\frac{S}{N} > 1. \tag{1.4}$$

La condición 1.4 no puede cumplirse indefinidamente debido a que la población no es infinita y como $S(t)$ es decreciente, en algún momento I' decrece también hasta

$$\lim_{t\to\infty} I(t) = 0. \tag{1.5}$$

Al inicio de la epidemia $S(0) \approx N$ por lo que, partiendo de 1.4, se obtiene la cantidad

$$R_0 = \frac{\beta}{\gamma}, \tag{1.6}$$

conocida como número de reproducción básico. Si $0 < R_0 < 1$ luego $I' < 0$ por lo que cesa la propagación de la infección. Mientras que si $R_0 > 1$ entonces $I' > 0$ y se encuentra en curso una epidemia.

1.1.1. Relación de Tamaño Final

La cantidad final de individuos susceptibles S_∞, es mayor que cero y está dada por

$$ln\left(\frac{S_0}{S_\infty}\right) = R_0\left(1 - \frac{S_\infty}{N}\right). \tag{1.7}$$

Demostración: Del sistema 1.1

$$(S + I)' = -\gamma I, \tag{1.8}$$

luego $S + I \geq 0$ es una función decreciente. Debido a 1.5 se tiene que

$$\lim_{t\to\infty}(S + I) = S_\infty, \tag{1.9}$$

y la integral de 1.8 es

$$\int_0^\infty (S+I)'dt = -\gamma \int_0^\infty I dt, \tag{1.10}$$

$$S_\infty + I_\infty - S_0 - I_0 = -\gamma \int_0^\infty I dt, \tag{1.11}$$

$$S_\infty - N = -\gamma \int_0^\infty I dt, \tag{1.12}$$

$$\int_0^\infty I dt = \frac{1}{\gamma}(N - S_\infty). \tag{1.13}$$

Por otro lado:

$$S' = -\frac{\beta}{N} SI, \tag{1.14}$$

$$\int_0^\infty \frac{S'}{S} = -\frac{\beta}{N} \int_0^\infty I dt, \tag{1.15}$$

$$\int_0^\infty \frac{ds}{S} = -\frac{\beta}{N} \int_0^\infty I dt, \tag{1.16}$$

$$ln\left(\frac{S_0}{S_\infty}\right) = \frac{\beta}{N} \int_0^\infty I dt, \tag{1.17}$$

y por 1.13

$$ln\left(\frac{S_0}{S_\infty}\right) = \frac{\beta}{\gamma N}(N - S_\infty), \tag{1.18}$$

$$ln(\frac{S_0}{S_\infty}) = \frac{\beta}{\gamma}\left(1 - \frac{S_\infty}{N}\right), \tag{1.19}$$

$$ln\left(\frac{S_0}{S_\infty}\right) = R_0\left(1 - \frac{S_\infty}{N}\right), \tag{1.20}$$

que es la misma expresión que 1.7 y se conoce como la relación de tamaño final. Note que $N - S_\infty$ es la cantidad de individuos que fueron infectados en todo el tiempo que duró la epidemia.

Al porcentaje equivalente $1 - \frac{S_\infty}{N}$ se le conoce como *attack rate* o tasa de ataque (aunque no es una tasa, sino más bien una proporción).

De manera similar, integrando de 0 a t:

$$\int_0^t (S+I)'dt = -\gamma \int_0^t I dt, \tag{1.21}$$

$$S(t) + I(t) - S_0 - I_0 = -\gamma \int_0^\infty I dt, \tag{1.22}$$

$$S(t) + I(t) - N = -\gamma \int_0^t I dt, \tag{1.23}$$

$$\int_0^\infty I dt = \frac{1}{\gamma}(N - S(t) - I(t)), \tag{1.24}$$

por otro lado:

$$S' = -\frac{\beta}{N} SI, \tag{1.25}$$

$$\int_0^t \frac{S'}{S} = -\frac{\beta}{N} \int_0^t I dt, \tag{1.26}$$

$$\int_0^t \frac{ds}{S} = -\frac{\beta}{N} \int_0^t I dt, \tag{1.27}$$

$$ln\left(\frac{S_0}{S(t)}\right) = \frac{\beta}{N} \int_0^t I dt, \tag{1.28}$$

además, por 1.24

$$ln(S_0) - ln(S(t)) = \frac{\beta}{\gamma N}(N - S(t) - I(t)), \tag{1.29}$$

$$S(t) + I(t) - \frac{\gamma N}{\beta} ln(S(t)) = S_0 + I_0 - \frac{\gamma N}{\beta} ln(S_0). \tag{1.30}$$

Las órbitas en el plano (S, I), solución del sistema 1.1 están dadas por 1.30.

En la expresión 1.7 si $S_\infty = 0$ el lado izquierdo tendería al infinito y el derecho sería finito por lo que ésta no podría ser una solución. Si $S_\infty = N$ significaría que no hubo epidemia, luego $0 < S_\infty < N$. Por lo que de acuerdo a este modelo, una parte de la población queda sin ser infectada hasta el final de la epidemia. Hay un único valor solución S_∞ a la relación de tamaño final 1.7, como se demuestra a continuación:

Defina

$$g(x) = ln\left(\frac{S_0}{x}\right) - R_0\left(1 - \frac{x}{N}\right), \tag{1.31}$$

en los valores cercanos a 0 por la derecha $g(0^+) > 0$ y cercanos a N por la izquierda $g(N^-) < 0$ por lo que la solución está en $0 < x < N$. Será útil conocer el intervalo en el que

$$g'(x) = \frac{R_0}{N} - \frac{1}{x} < 0, \tag{1.32}$$

este intervalo es

$$0 < x < \frac{N}{R_0}, \tag{1.33}$$

Note que esto tiene sentido cuando $R_0 \geq 1$ ya que si $R_0 < 1$ no restringe que $x > N$, lo que no es coherente ya que supera a la población total. Por lo que si $R_0 \geq 1$, $g(x)$ es monótonamente decreciente desde $x = 0^+$ hasta $x = \frac{N}{R_0}$. Luego crece a un valor negativo cerca de N. Así que

$$0 < S_\infty < \frac{N}{R_0}. \tag{1.34}$$

Como ejemplo para observar el comportamiento de $g(x)$, utilice los parámentros $R_0 = 1.1$, $S_0 = 10$ y $N = 1000$.

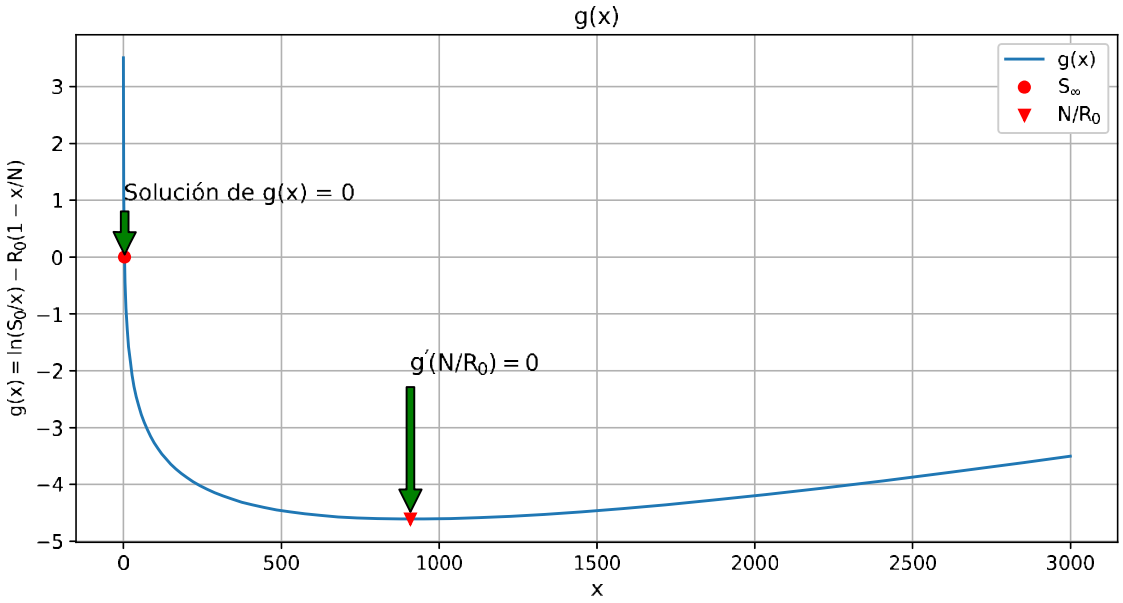


Figura 1.1. Comportamiento de $g(x)$, dadas las condiciones iniciales $R_0 = 1.1$, $S_0 = 10$, $N = 1000$.

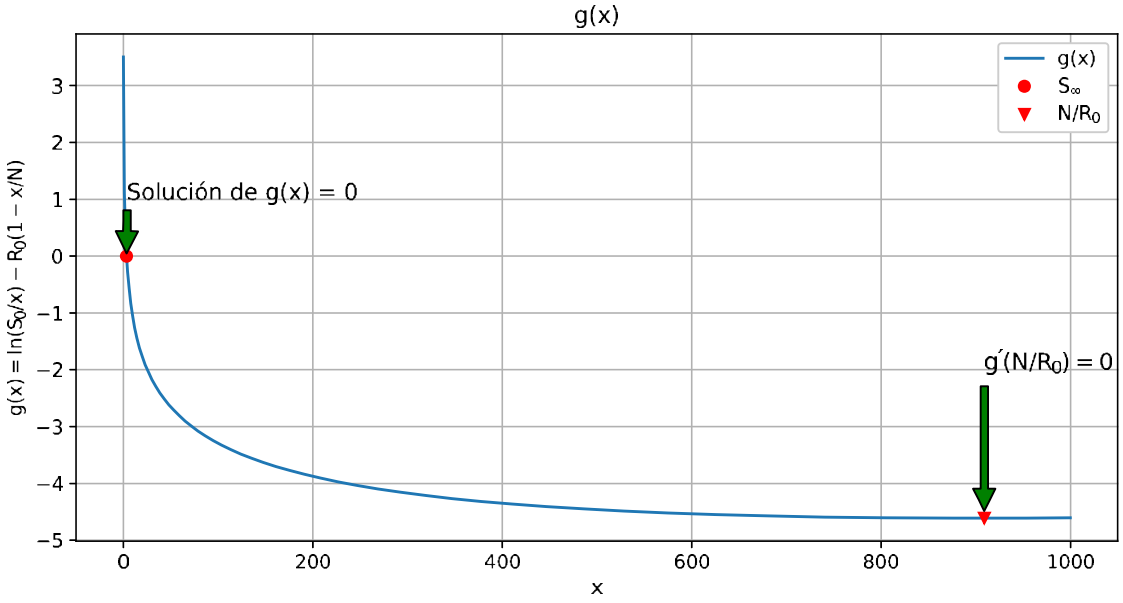


Figura 1.2. Ampliando la figura 1.1 se observa que puede delimitarse la solución al intervalo $x \in (0, N/R_0)$.

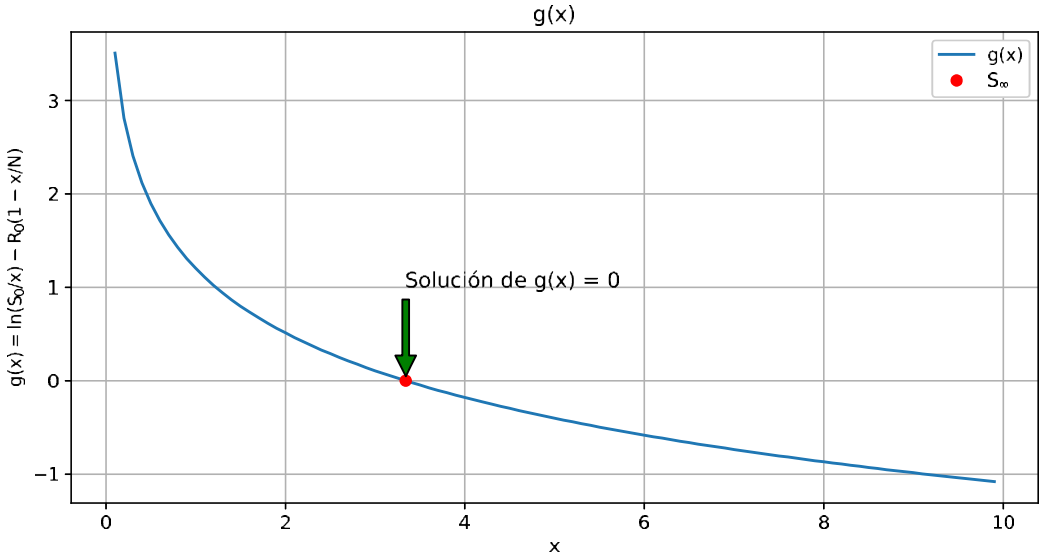


Figura 1.3. Ampliación de la región en la que se encuentra la solución.

Si por el contrario $R_0 < 1$ resulta que $g(x)$ es monótonamente decreciente desde $x = 0^+$ hasta $x = N$. Por lo que hay un único cero S_∞ de $g(x)$ con $S_\infty < N$. Considere el ejemplo anterior, suponiendo ahora que $R_0 = 0.8$ mientras que los otros parámetros continúan siendo $S_0 = 10$, $N = 1000$.

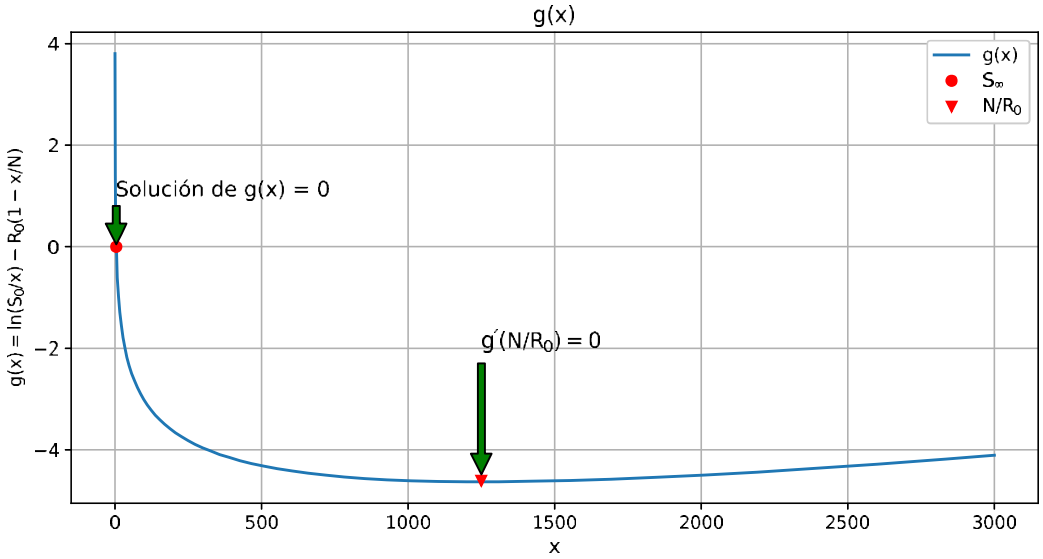


Figura 1.4. Comportamiento de $g(x)$ con $R_0 < 1$, note que aunque la función es monótonamente decreciente hasta $x = N/R_0$, este valor sería superior al de la población, considerada constante.

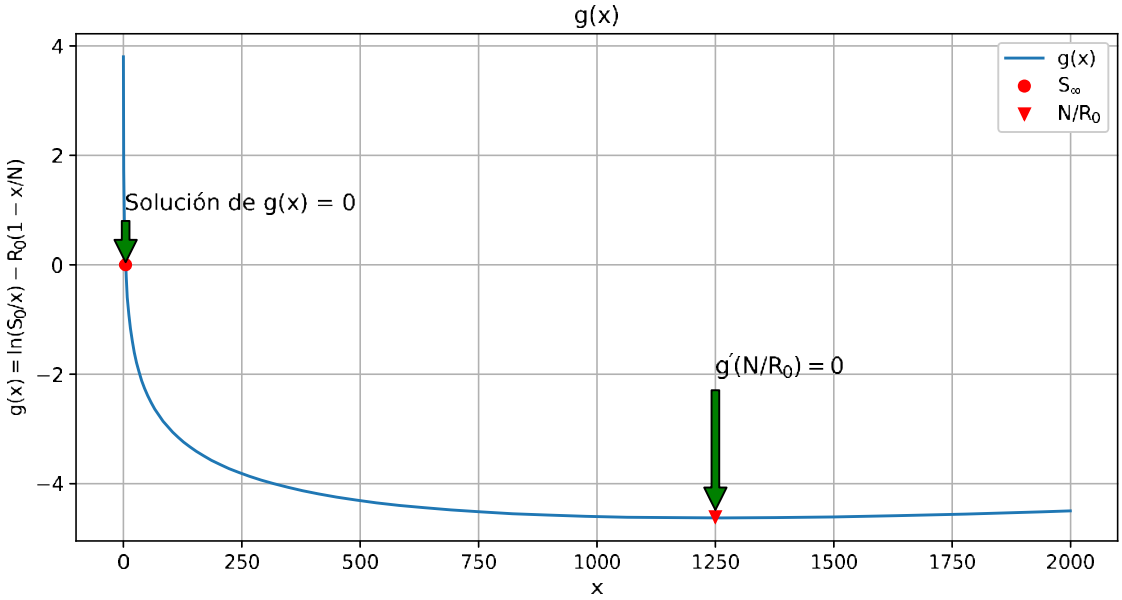


Figura 1.5. Acá puede observarse con mayor facilidad que $x = N/R_0$ supera a N, por lo que solo se puede decir que la solución está acotada en el intervalo $x \in (0, N)$.

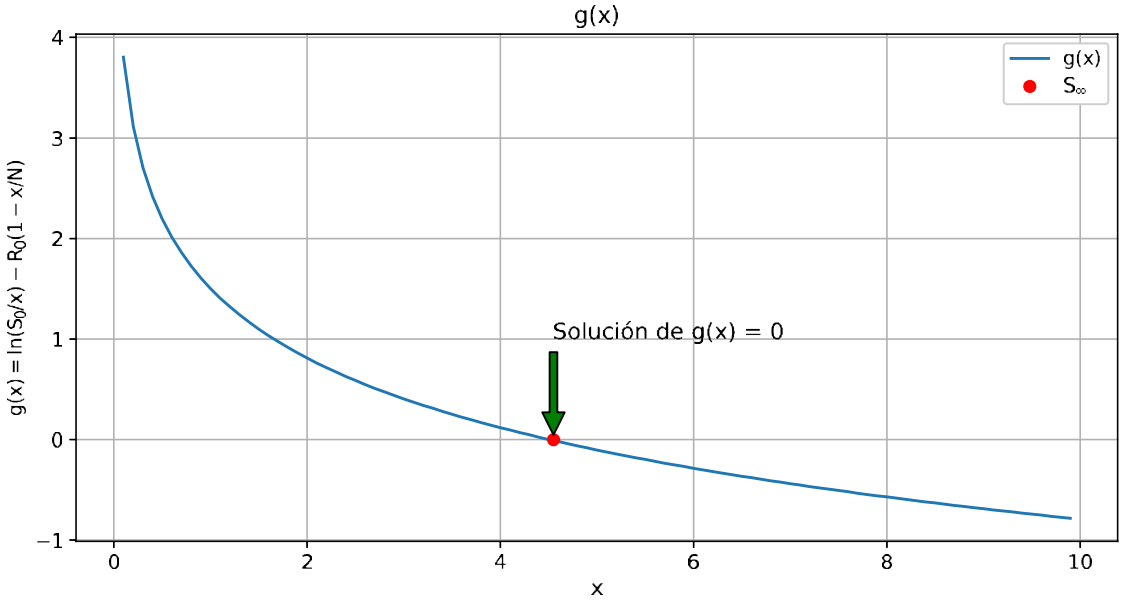


Figura 1.6. Ampliación de la región en la que se encuentra la solución.

1.2. Determinación de la Tasa de Contacto β

En el modelo compartimental se utiliza el símbolo I para representar a los infectados. Pero para realizar el cómputo será preferible clasificar a los infectados en las siguientes categorías. Se representa con $\delta A(t)$ a los nuevos casos de infectados detectados exactamente en el tiempo t. Desde su detección los individuos infectados entran al compartimento $I(t)$ donde permanecen $1/\gamma$ días. También es útil la cantidad $A(t)$ que contabiliza el total de personas que se han infectado desde el inicio de la epidemia hasta el tiempo t. Por ejemplo suponga que los primeros tres días se han reportado los siguientes infectados: primer día, 5 infectados, segundo y tercer día 7 y 4 infectados respectivamente. Las cantidades como las hemos definido arriba, suponiendo que la infección dura $1/\gamma = 2$ días quedarían como:

$\delta A(1) = 5$	$A(1) = 5$	$I(1) = 5$	$R(1) = 0$
$\delta A(2) = 7$	$A(2) = 12$	$I(2) = 12$	$R(2) = 0$
$\delta A(3) = 4$	$A(3) = 16$	$I(3) = 11$	$R(3) = 5$

Tabla 1.1. Ejemplo del cálculo de las variables δA, A, I y R suponiendo $\gamma = 2$.

Los datos disponibles para esta investigación son $\delta A(t)$ y la población N de cada departamento. A partir de estas cantidades es posible calcular las otras variables del modelo, por ejemplo:

$$A(t = n) \equiv A_n = \sum_i^n \delta A_i, \tag{1.35}$$

y suponiendo que hay inmunidad permanente luego de salir de la infección se tiene que

$$S(t = n) \equiv S_n = N - A_n \tag{1.36}$$

mientras que quienes fueron infectados γ días atrás pasan al compartimento de los removidos

$$R(t = n) \equiv R_n = \delta A_{n-\gamma} + R_{n-1}, \tag{1.37}$$

por lo que la cantidad de infectados $I(t)$ en $t = n$ es

$$I_n = A_n - R_n, \tag{1.38}$$

conociendo estas variables es posible determinar β_n, de la siguiente forma

$$\beta_n = N\frac{\delta A_{n+1}}{I_n S_n}. \tag{1.39}$$

2. REDES NEURONALES

Suponer que β es otra variable del modelo conlleva la necesidad de tener una ecuación más en 1.1, a falta de tal ecuación se utilizarán redes neuronales para pronosticar los futuros valores de la tasa de propagación de la infección. La entrada de la red son los valores de β de los n días anteriores al que se quiere estimar. Cada neurona artificial de las que conforman la red realiza las siguientes acciones [12]:

1. Lleva a cabo una combinación lineal de los parámetros de entrada β_i de la siguiente forma:

$$z = \sum_i (w_i \beta_i) + b \tag{2.1}$$

   donde el peso w_i indica la importancia que la neurona asigna a la i$-$ésima entrada y b es un sesgo (bias) análogo al voltaje de encendido de una neurona natural.

2. A la combinación lineal le aplica la función de activación ReLU (Unidad Lineal Rectificada), definida por:

$$f(z) = \begin{cases} 0 \text{ si } z < 0 \\ mz \text{ si } z \geq 0 \end{cases} \tag{2.2}$$

   donde z será la combinación lineal 2.1, y $f(z)$ el valor final que devuelve la neurona. Los parámetros b y w_i pueden variarse para ajustar ese valor final, al valor dado por la muestra de entrenamiento.

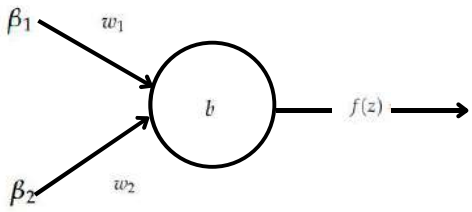


Figura 2.1. Neurona artificial.

Con neuronas como la de 2.1 se construye una red neuronal densa. Al tener más neuronas se dispone de más parámetros (pesos y bias) para ajustar el resultado de la red al valor observado.

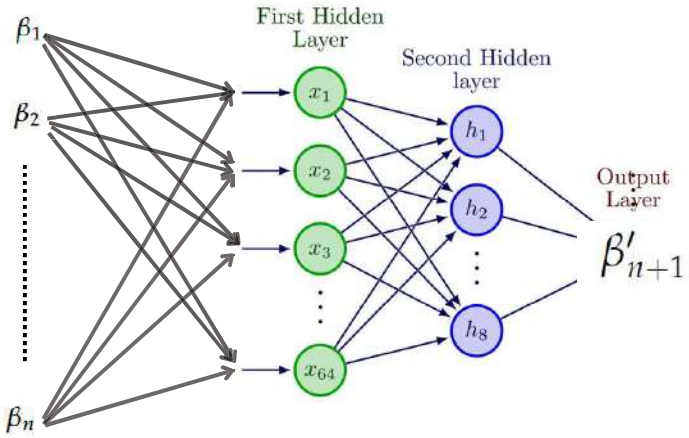


Figura 2.2. Red neuronal estimando β_{n+1}.

Se utilizarán los valores de n días para predecir el valor del día $n+1$. De esta forma decimos que β'_{n+1} es la predicción de β_{n+1} que es el valor observado.

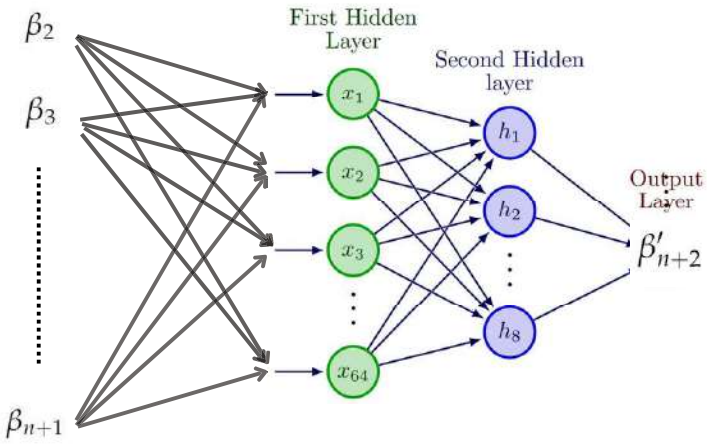


Figura 2.3. Red neuronal estimando β_{n+2}.

Por lo que en general, los datos de entrenamiento tendrán la forma:

$$[\beta_j, \beta_{j+1}, ..., \beta_{j+n-1}] \longrightarrow \beta_{j+n} \tag{2.3}$$

2.1. Agrupación de Datos

Suponga que el listado de datos disponible es $[\beta_0, \beta_1, ..., \beta_{k-1}]$, se definen las siguientes cantidades:

- $E \equiv$ Train size, tamaño del conjunto de datos de entrenamiento.
- $P \equiv$ Test size, tamaño del conjunto de datos de evaluación.
- $k = E + P$, tamaño del conjunto de datos completo.
- $D \equiv$ Días a pronosticar.
- $n \equiv$ Intervalo de tiempo que determina el valor futuro.

Se separa el conjunto original de la siguiente manera:

$$[\beta_0, \beta_1, ..., \beta_{E-1}] \cup [\beta_E, \beta_{E+1}, ..., \beta_{k-1}], \tag{2.4}$$

el conjunto a la izquierda corresponde a los datos de entrenamiento y el de la derecha es el conjunto de evaluación, con una longitud

$$P = D + n. \tag{2.5}$$

Ahora los datos de entrenamiento serán separados también con los parámetros E', P', D' (basta conocer D' y n para determinar E', P'),

$$[\beta_0, \beta_1, ..., \beta_{E'-1}] \cup [\beta_{E'}, \beta_{E'+1}, ..., \beta_{E-1}], \tag{2.6}$$

los P' datos en el conjunto de la derecha serán agrupados de la siguiente forma:

$$[\beta_{E'}, \beta_{E'+1}, ..., \beta_{E'+n-1}] \to \beta_{E'+n}, \tag{2.7}$$
$$[\beta_{E'+1}, \beta_{E'+2}, ..., \beta_{E'+n}] \to \beta_{E'+n+1}, \tag{2.8}$$
$$\vdots \to \vdots \tag{2.9}$$
$$[\beta_{E-2-n-1}, ..., \beta_{E-2}] \to \beta_{E-1}. \tag{2.10}$$

2.2. Función de Costo

Los parámetros w_i y b de cada neurona en la red pueden ajustarse hasta alcanzar un β'_n cercano al valor de entrenamiento β_n. Para cuantificar qué tan cerca del valor observado quedan las predicciones definimos la siguiente función de costo

$$C = \frac{1}{2N} \sum_{j=1}^{N} (\beta'_{n+j} - \beta_{n+j})^2, \tag{2.11}$$

conforme los parámetros w_i y b se van ajustando para mejorar la predicción C va disminuyendo, a este proceso se le llama aprendizaje de la red.

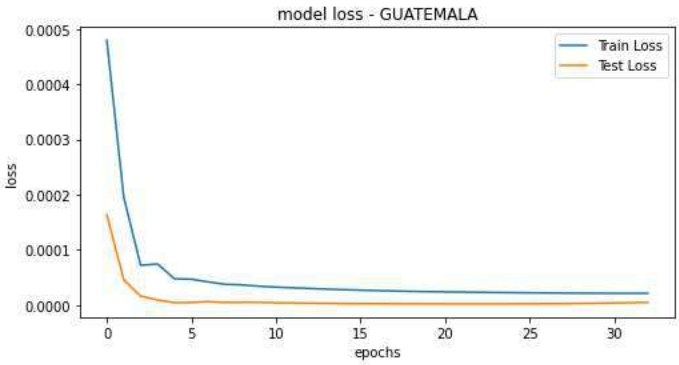


Figura 2.4. Función de costo o *loss function* para alguna de las corridas de la red.

Cuando $C \approx 0$, la red guarda los parámetros w_i y b que son el resultado final de su aprendizaje. Ahora podemos usar un conjunto diferente de parámetros de entrada η_1, η_2, ..., η_n para predecir η_{n+1}.

2.3. Algoritmo de Descenso de Gradiente

La función de costo se calcula sobre los N datos de entrenamiento

$$[\beta_1, \beta_2, ..., \beta_n] \rightarrow \beta_{n+1}, \tag{2.12}$$

$$[\beta_2, \beta_3, ..., \beta_n] \rightarrow \beta_{n+2}, \tag{2.13}$$

$$\vdots \rightarrow \vdots \tag{2.14}$$

$$[\beta_N, \beta_{N+1}..., \beta_{N+n-1}] \rightarrow \beta_{n+N}, \tag{2.15}$$

y está dada por

$$C = \frac{1}{2N} \sum_{i=n+1}^{n+N} (\beta_i' - \beta_i)^2. \tag{2.16}$$

Las predicciones β_i' dependen de los parámetros $(w_{l,k}, b_l)$ por lo que se puede escribir

$$C = C(w_{l,k}, b_l) \tag{2.17}$$

El aprendizaje de la red consiste en determinar los $(w_{l,k}, b_l)$ que minimizan la función de costo. Es decir que se busca minimizar $C(f(z))$ con z definida en 2.1. Con ese fin se parte de alguna posición aleatoria $(w_{l,k}, b_l)$, estos parámetros determinan toda una curva $f(z)$ que corresponde a las predicciones. Al variar los parámetros, la variación de C es:

$$\Delta C = \frac{\partial C}{\partial w_{l,k}} \Delta w_{l,k} + \frac{\partial C}{\partial bl} \Delta b_l, \tag{2.18}$$

donde se está utilizando la notación de sumas de Einstein. La expresión anterior se puede escribir en términos del producto interno del gradiente de C y la variación de los parámetros.

$$\Delta C = \nabla C \cdot \Delta V, \tag{2.19}$$

donde

$$\nabla C = \left(\frac{\partial C}{\partial w_{l,k}}, \frac{\partial C}{\partial bl} \right), \tag{2.20}$$

y

$$\Delta v = (\Delta w_{l,k}, \Delta b_l). \tag{2.21}$$

La elección $\Delta v = -\eta \nabla C$ con $\eta > 0$ garantiza que nos movemos hacia un valor menor de C ya que se tiene

$$\Delta C = -\eta \nabla C \cdot \nabla C = -\eta |\nabla C|^2 \leq 0, \tag{2.22}$$

lo que permite acercarce al mínimo a cualquier distancia arbitraria.

2.4. Descenso de Gradiente Estocástico

El conjunto de entrenamiento puede ser tan grande que calcular la función de costo podría requerir bastante tiempo de cómputo. Una forma más eficiente de calcularla es proceder de la siguiente forma: elegir M datos aleatoriamente de los N que hay en el conjunto de entrenamiento y con ellos calcular $\bar{C}_M$. Si M es suficientemente grande $\bar{C}_M \approx \bar{C}$ es decir que se supone que

$$\frac{1}{M} \sum_{i=1}^{M} C_i \approx \frac{1}{N} \sum_{i=1}^{N} C_i. \tag{2.23}$$

Se hace variar $(w_{k,l}, b_l)$ y se calcula $\bar{C}$ con otro conjunto de M datos elegidos aleatoriamente, cada uno de estos conjuntos es conocido como un *mini batch*. Se continúa variando $(w_{k,l}, b_l)$ y calculando $\bar{C}_M$ hasta utilizar todos los datos de entrenamiento. Cuando esto ocurre se dice que se ha completado una época. El proceso se repite con más épocas hasta alcanzar aproximaciones satisfactorias.

3. PROCESOS ESTOCÁSTICOS

Todorovic [16] define un proceso estocástico de la siguiente forma:

Un proceso estocástico es una familia de variables aleatorias $X(t), t \in T$, definidas en un espacio de probabilidad común $\{E, \mathcal{F}, P\}$, donde el parámetro T es un subconjunto de la línea real R.

Para comprender a qué hace referencia el espacio de probabilidad $\{E, \mathcal{F}, P\}$ será de utilidad revisar los axiomas formulados por Kolmogorov en [9] para describir un *campo de probabilidad* que se transcriben a continuación:

Sea E una colección de elementos $\xi, \eta, \zeta, ...$, a los que llamaremos eventos elementales, y $\mathcal{F}$ un conjunto de subconjuntos de E; los elementos del conjunto $\mathcal{F}$ son llamados eventos aleatorios.

1. *$\mathcal{F}$ es un campo de conjuntos.*
2. *$\mathcal{F}$ contiene a E.*
3. *A cada conjunto $A \in \mathcal{F}$ se le asigna un número real no negativo $P(A)$. Este número $P(A)$ es llamado la probabilidad del evento A.*
4. *$P(E) = 1$.*

Un sistema de conjuntos, $\mathcal{F}$, junto con una asignación definitiva de números $P(A)$, satisfaciendo los axiomas $1 - 4$ es llamado campo de probabilidad.

3.1. Procesos Estocásticos Estacionarios

Un proceso estocástico es estacionario cuando el promedio y la varianza de su serie se mantienen constantes y además la autocovarianza cumple con

$$COV(y_0, y_h) = COV(y_t, y_{t+h}). \tag{3.1}$$

3.1.1. Prueba ADF

Prueba de Dickey-Fuller Aumentada, se utiliza para determinar si una serie corresponde a un proceso estacionario o no. Para aplicarla se empleará la función *adfuller* de la biblioteca *statsmodels* [1] de Python.

3.1.2. Modelos de Series de Tiempo Estocásticas

3.1.2.1. Modelos Autorregresivos

Un modelo autoregresivo de orden p se escribe $AR(p)$, y es de la siguiente forma [8]:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \epsilon_t, \tag{3.2}$$

donde ϵ_t es un término de error o ruido blanco que sigue una distribución normal con promedio 0 y que cumple con

$$E[\epsilon_t y_{t-\tau}] = 0 \forall \tau > 0, \tag{3.3}$$

lo que significa que el error no influye en los valores de la serie ocurridos en el pasado.

En el caso $p = 1$ se tiene el modelo AR(1):

$$y_t = \phi_1 y_{t-1} + \epsilon_t, \tag{3.4}$$

que puede reescribirse como

$$(1 - \phi_1 L) y_t = \epsilon_t, \tag{3.5}$$

donde L es el operador de retardos [8]. La raíz del polinomo característico

$$1 - \phi_1 L = 0 \tag{3.6}$$

es

$$L_1 = 1/\phi_1, \tag{3.7}$$

y para tratarse de un proceso estacionario debe cumplir con

$$|L_1| > 1, \tag{3.8}$$

[1] Específicamente *adfuller* es una función del submódulo *statsmodels.tsa.stattools*

o bien

$$|\phi_1| < 1, \tag{3.9}$$

como se podrá ver en breve.

Suponga que el proceso 3.4 se inició en el período $-N$, por sustituciones sucesivas se tiene que

$$\begin{aligned} y_t &= \phi_1 y_{t-1} + \epsilon_t \\ &= \phi_1(\phi_1 y_{t-2} + \epsilon_{t-1}) + \epsilon_t \\ &= \phi_1^2 y_{t-2} + \phi_1 \epsilon_{t-1} + \epsilon_t \\ &\vdots \\ &= \phi_1^{t+N} y_{-N} + \sum_{m=0}^{t+N-1} \phi_1^m \epsilon_{t-m}, \end{aligned} \tag{3.10}$$

y debido a que el promedio de la distribución del ruido blanco es cero, se tiene que el valor esperado de y_t es:

$$E[y_t] = \phi_1^{t+N} y_{-N}, \tag{3.11}$$

por lo que la serie se puede clasificar en las siguientes tres categorías dependiendo del valor de ϕ_1:

- Si $|\phi_1| > 1$ El valor esperado de la serie crece de manera indefinida.
- Si $|\phi_1| = 1$ El valor esperado de la serie es constante o tiene alternancia de signos.
- Si $|\phi_1| < 1$ El valor esperado de la serie tiende a cero (o a una constante) en el infinito.

Cuando se tiene el caso $|\phi_1| < 1$ se dice que se trata de una serie estacionaria o estable. En ese caso el valor esperado cuando $-N = -\infty$ es

$$E[y_t] = 0, \tag{3.12}$$

o bien cierta constante [2] μ. La varianza de un proceso $AR(1)$ iniciado en $-N$ es

$$\begin{aligned}
\gamma_0 &= E[(y_t - E[y_t])^2] \\
&= E[(\phi_1^{t+N} y_{-N} + \sum_{m=0}^{t+N-1} \phi_1^m \epsilon_{t-m} - \phi_1^{t+N} y_{-N})^2] \\
&= E[(\sum_{m=0}^{t+N-1} \phi_1^m \epsilon_{t-m})^2] \\
&= E[\sum_{m=0}^{t+N-1} \phi_1^{2m} \epsilon_{t-m}^2] \\
&= \sigma_\epsilon^2 \sum_{m=0}^{t+N-1} \phi_1^{2m}.
\end{aligned} \tag{3.13}$$

Como se considera que la serie es estacionaria, $\phi_1 < 1$, la varianza en 3.13 es la suma de una serie geométrica con razón ϕ_1^2. Los términos primero y último y la razón son:

$$\begin{aligned}
a_1 &= \sigma_\epsilon^2, \\
a_n &= \sigma_\epsilon^2 \phi_1^{2(t+N-1)}, \\
r &= \phi_1^2,
\end{aligned} \tag{3.14}$$

y la suma en 3.13 puede reexpresarse como:

$$\gamma_0 = \sigma_\epsilon^2 \left(\frac{1 - \phi_1^{2(t+N-1)}}{1 - \phi_1^2} \right). \tag{3.15}$$

Ya que el proceso es estacionario y se inició en $-\infty$, la expresión 3.15 se simplifica como

$$\gamma_0 = \left(\frac{\sigma_\epsilon^2}{1 - \phi_1^2} \right), \tag{3.16}$$

la expresión 3.16 da la varianza de la serie. Para calcular la covarianza[3] de la serie con una versión de sí misma desplazada τ unidades procedemos de la siguiente forma:

[2] Esto ocurre si al proceso 3.4 se le agrega la constante μ, es decir: $y_t = \mu + \phi_1 y_{t-1} + \epsilon_t$

[3] A este tipo de covarianza se le conoce como *autocovarianza*

$$\begin{aligned}\gamma_\tau &= E[y_t y_{t-\tau}] \\ &= E[(\phi_1 y_{t-1} + \epsilon_t) y_{t-\tau}] \\ &= \phi_1 E[y_{t-1} y_{t-\tau}] \\ &= \phi_1 \gamma_{\tau-1},\end{aligned} \tag{3.17}$$

la expresión 3.17 es una ecuación en diferencias, homogénea de primer orden cuya solución es

$$\gamma_\tau = A\lambda^\tau = \gamma_0 \phi_1^\tau. \tag{3.18}$$

Las autocorrelaciones se definen como

$$\rho_\tau = \frac{\gamma_\tau}{\gamma_0}, \tag{3.19}$$

por lo que

$$\rho_\tau = \phi_1^\tau. \tag{3.20}$$

La ecuación 3.17 puede generalizarse [8] para un modelo $AR(p)$:

$$\gamma_\tau = \phi_1 \gamma_{\tau-1} + \phi_2 \gamma_{\tau-2} + ... + \phi_p \gamma_{\tau-p}. \tag{3.21}$$

Debido a que calcular analíticamente la autocovarianza inicial [4] γ_0 para modelos de orden $p > 1$ resulta muy complicado, es más práctico trabajar con las autocorrelaciones definidas en 3.19, ya que $\rho_0 = 1$ y

$$\rho_\tau = \phi_1 \rho_{\tau-1} + \phi_2 \rho_{\tau-2} + ... + \phi_p \rho_{\tau-p}. \tag{3.22}$$

La ecuación en diferencias 3.22 da lugar al siguiente sistema de ecuaciones lineales

$$\begin{aligned}\rho_1 &= \phi_1 \rho_0 + \phi_2 \rho_1 + ... + \phi_p \rho_{1-p}, \\ \rho_2 &= \phi_1 \rho_1 + \phi_2 \rho_0 + ... + \phi_p \rho_{2-p}, \\ &\vdots \\ \rho_p &= \phi_1 \rho_{p-1} + \phi_2 \rho_{p-2} + ... + \phi_p \rho_0,\end{aligned} \tag{3.23}$$

[4] O simplemente varianza

o bien definiendo

$$\vec{\rho} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_p \end{bmatrix}, \ \vec{\phi} = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{bmatrix}, \ M_\rho = \begin{bmatrix} \rho_0 & \rho_1 & \cdots & \rho_{1-p} \\ \rho_1 & \rho_0 & \cdots & \rho_{2-p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p-1} & \rho_{p-2} & \cdots & \rho_0, \end{bmatrix} \tag{3.24}$$

la ecuación 3.23 se puede expresar de forma más compacta como

$$\vec{\rho} = M_\rho \vec{\phi}. \tag{3.25}$$

Sin embargo en la práctica es sencillo calcular los valores de autocorrelación entre una serie y su versión retardada. El interés está más bien en calcular los coeficientes ϕ_i para determinar el modelo $AR(p)$, lo cual se consigue de la siguiente forma:

$$\vec{\phi} = M_\rho^{-1} \vec{\rho}. \tag{3.26}$$

3.1.2.2. Modelos de Medias Móviles

Un modelo de medias móviles de orden q se escribe $MA(q)$, y es de la siguiente forma:

$$y_t = \epsilon_t - \theta_1 \epsilon_{t-1} - ... - \theta_q \epsilon_{t-q}, \tag{3.27}$$

donde los ϵ_i son valores de ruido blanco, independientes entre sí y que siguen una distribución normal. En el caso $q = 1$ el modelo MA(1) es

$$y_t = \epsilon_t - \theta_1 \epsilon_{t-1}, \tag{3.28}$$

y con el operador de retardos puede expresarse como

$$y_t = (1 - \theta_1 L)\epsilon_t. \tag{3.29}$$

La covarianza es

$$\begin{aligned} E[y_t y_{t-\tau}] &= E[(\epsilon_t - \theta_1 \epsilon_{t-1})(\epsilon_{t-\tau} - \theta_1 \epsilon_{t-\tau-1})] \\ &= E[\epsilon_t \epsilon_{t-\tau}] - E[\epsilon_t \theta_1 \epsilon_{t-\tau-1}] - E[\epsilon_{t-1} \theta_1 \epsilon_{t-\tau}] + \theta_1^2 E[\epsilon_{t-1} \theta_1 \epsilon_{t-\tau-1}], \end{aligned} \tag{3.30}$$

como el ruido blanco ϵ_t sigue una distribución normal y es independiente de $\epsilon_{t'}$ si $t \neq t'$, en $\tau = 0$ y en $\tau = 1$ se tiene

$$\begin{aligned} \gamma_0 &= E[y_t y_t] = \sigma_\epsilon^2(1 + \theta_1^2), \\ \gamma_1 &= E[y_t y_{t-1}] = -\theta_1 \sigma_\epsilon^2, \end{aligned} \tag{3.31}$$

mientras que para $\tau > 1$

$$\gamma_\tau = 0. \tag{3.32}$$

Note que mientras en el proceso AR(1) las covarianzas 3.18 son una función suave decreciente, en el proceso MA(1) cambian abruptamente a cero luego de $\tau = 1$. Note también que el proceso es estacionario sin importar el valor del coeficiente θ_1 .

3.1.3. Determinación del Orden de los Modelos Autoregresivos y de Medias Móviles

Primero considere el proceso trivial que consiste en una serie de valores aleatorios independientes entre sí. Estos valores se obtienen de una distribución normal por lo que el proceso es estacionario al tener un promedio constante de 0 y una desviación estándar, también constante, igual a 5. El código para generar esta gráfica y las demás de esta sección se encuentra en https://github.com/CarlosSalazarR/Thesis-Scripts.

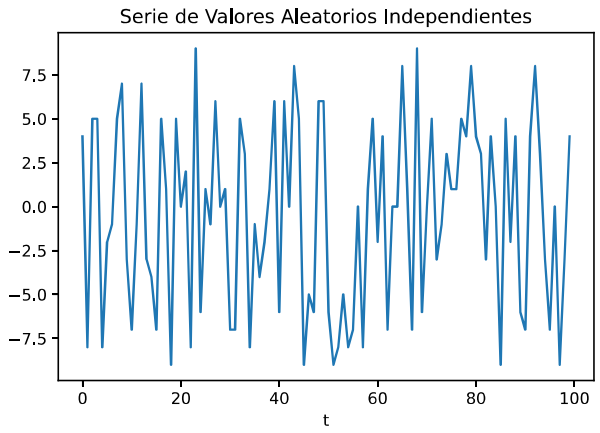


Figura 3.1. Ejemplo de serie obtenida a partir de distribución normal.

Las gráficas de las funciones de autocorrelación ACF y de autocorrelación parcial PACF de la serie en Figura 3.1 son

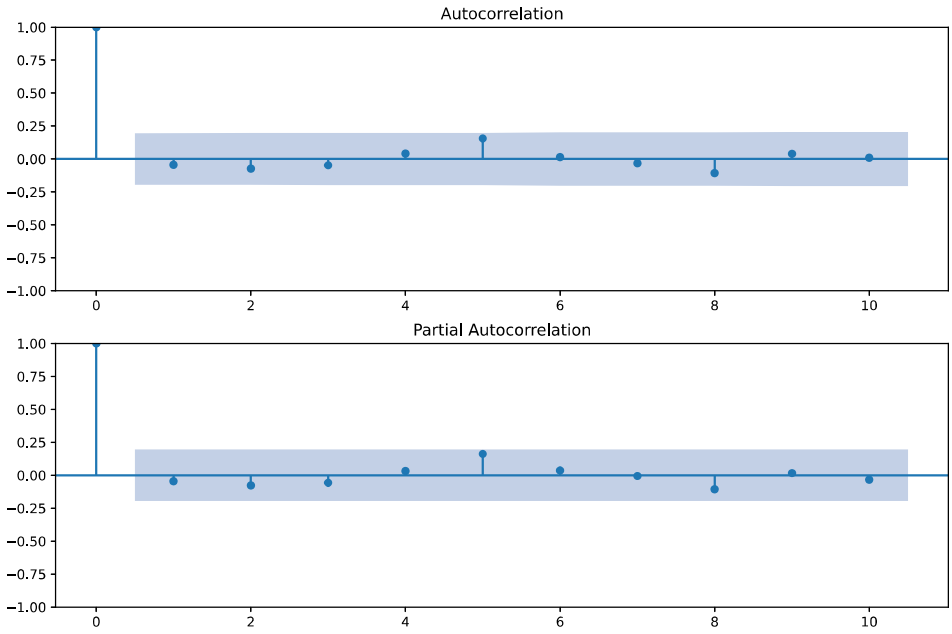


Figura 3.2. Las ACF y ACFP indican que los valores de la serie son independientes entre sí.

Las autocorrelaciones que quedan dentro de la banda celeste son aquellas que no son estadísticamente significativas, es decir que los respectivos retrasos no influyen en el actual valor de la serie. En este caso el único valor significativo corresponde al retraso 0, que consiste en la serie misma sin retraso. Por lo que ninguno de los valores anteriores influye en el actual como se esperaba pues son independientes entre sí.

Ahora considere el movimiento browniano, un proceso estocástico que describe, entre otras cosas, el movimiento de pequeñas partículas golpeadas por moléculas de agua. En este ejemplo el movimiento browniano es unidimensional y partiendo de cierta posición inicial la partícula puede avanzar un paso hacia adelante o hacia atrás en cada iteración, por lo que es un proceso AR(1) como en 3.4 donde los ϵ_i pueden tomar únicamente los valores 1 y -1, y donde $\phi_1 = 1$. La gráfica de la serie resultante es como la mostrada en la Figura 3.3.

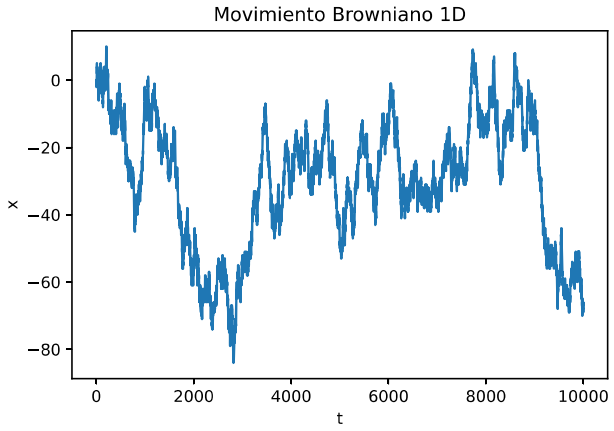


Figura 3.3. Una realización del movimiento browniano descrito por el proceso $y_t = y_{t-1} + \epsilon_t$, donde cada ϵ_i puede valer solamente 1 ó -1.

La PACF en la Figura 3.4 indica que los valores de la serie dependen únicamente del valor anterior. Esto se debe a que el nuevo valor se obtiene de desplazamientos a partir del valor previo. Por otro lado la ACF muestra una dependencia de una mayor cantidad de valores anteriores puesto que toma en cuenta la dependencia indirecta de los retrasos mientras que la función de autocorrelación parcial mide solamente la dependencia directa. Así que para determinar el orden del modelo de autoregresión lineal se debe observar la PACF.

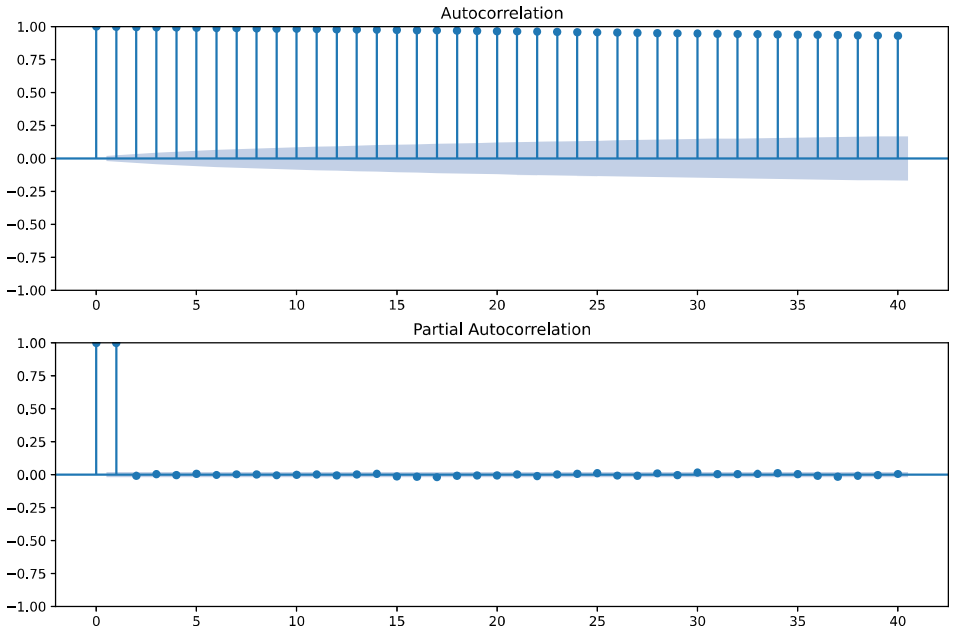


Figura 3.4. Funciones de autocorrelación y autocorrelación parcial del movimiento browniano.

Debido a que los errores siguen una distribución normal

$$E[\epsilon_\tau \epsilon_{\tau'}] = 0, \tag{3.33}$$

la función de autocorrelación permite determinar el orden q de un modelo de medias móviles, ya que en este caso no existe influencia indirecta entre los errores de las diferentes predicciones.

3.1.4. Modelos ARMA y ARIMA

Un modelo ARMA(p,q) supone que el valor actual de la variable depende de sus p valores pasados y de q valores ϵ_i que son la diferencia entre una observación y la estimación del modelo autorregresivo

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - ... - \theta_q \epsilon_{t-q}. \tag{3.34}$$

Si el test ADF indica que la serie no es estacionaria puede realizarse la transformación que consiste en la diferencia entre valores consecutivos. Este proceso se realiza d veces hasta que el test indique que la nueva serie sea estacionaria, y el modelo a utilizar es ARIMA(p,d,q).

4. METODOLOGÍA PARA PRONOSTICAR β APLICANDO REDES NEURONALES

Se considera a $\beta = \beta(t)$ como una variable dependiente del tiempo por lo que se necesitaría agregar una ecuación más en el sistema (1.1) para resolverlo. Se carece de tal ecuación así que se recurre a las redes neuronales.

La fuente de datos consultada es el sitio web del Ministerio de Salud de Guatemala [1], el intervalo de tiempo de las observaciones empleadas en esta sección va desde el 06 de junio de 2020 hasta el 30 de agosto de 2021.

Los datos diponibles son:

- Los nuevos infectados $\delta A = \frac{\beta IS}{N}$ cada día por departamento, los cuales aportan un término en 1.2.

- La población N en cada departamento.

Conocemos los δA al $j-$ ésimo día de donde despejamos

$$\beta_{j-1} = N\frac{\delta A_j}{I_{j-1}S_{j-1}}. \tag{4.1}$$

Una parte de los β calculados con (4) se utiliza para entrenar la red neuronal que consta de 3 capas, la primera de 64 neuronas, la segunda de 8 y la tercera de 1.

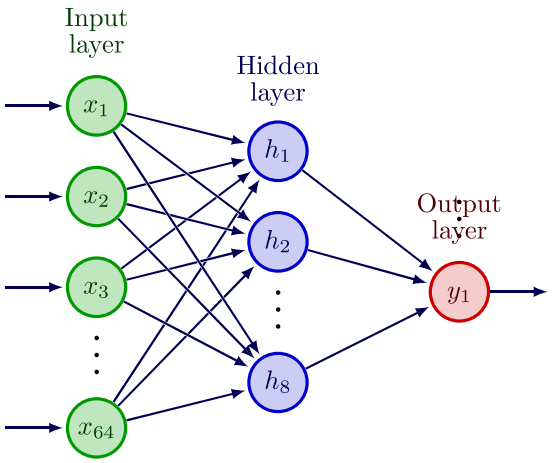


Figura 4.1. Red neuronal utilizada para predecir β.

1. Calcular β de los datos disponibles y utilizar una parte de ellos para entrenar una red neuronal.
2. Sustituir los β estimados por la red en el modelo SIR para predecir las variables S, I y R utilizando diferencias finitas.

5. METODOLOGÍA PARA PRONOSTICAR β APLICANDO MODELOS ARIMA

Se propone que las fluctuaciones de β son un proceso estocástico que se puede modelar como un proceso ARIMA (modelo lineal autoregresivo integrado y de medias móviles). Para ello se implementará la librería *statsmodels* de Python con el fin de determinar los coeficientes ϕ_i y θ_i del modelo ARIMA(p,d,q) que describe la serie β. La biblioteca también se utilzará para realizar la prueba de Dickey-Fuller aumentada para determinar si la serie es estacionaria.

1. Determinar si la serie en cierto intervalo de tiempo es estacionaria, para ello ejecutar la función *adfuller* de la librería *statsmodels*.
2. En caso de que la serie no sea estacionaria, realizar d transformaciones hasta obtener una serie estacionaria.
3. Si la serie es estacionaria desde el principio $d = 0$.
4. Calcular la función de autocorrelación parcial y aquellos valores para los que sea estadísticamente significativa corresponden al orden p de la parte de autorregresión lineal e indican que cada valor de la variable puede determinarse mediante una combinación lineal de los p valores anteriores.
5. Calcular la función de autocorrelación para determinar el orden q de la parte de medias móviles.
6. Implementar la función *ARIMA(serie, order=(p,d,q))* de la librería *statsmodels* para elaborar el modelo ARIMA.
7. Si se presentan diferentes opciones de p o q se elige aquél para el cual el error cuadrático medio de las primeras diez predicciones sea el menor.

6. DETERMINACIÓN DE β UTILIZANDO REDES NEURONALES

Para predecir β se utilizaron redes de 3 capas de 64, 8 y 1 neuronas, 100 epochs y batch size igual 30. Se corrió la red neuronal 100 veces por cada día que se desea pronosticar la tasa de contagio y se propuso que los resultados siguen la distribución de Poisson, con el promedio representado por λ,

$$P(x = k, \lambda) = \frac{e^{-\lambda}\lambda^k}{k!}. \tag{6.1}$$

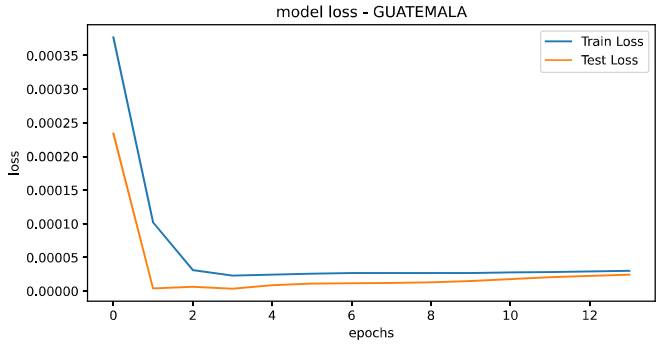


Figura 6.1. Función de costo de la red neuronal para el departamento de Guatemala.

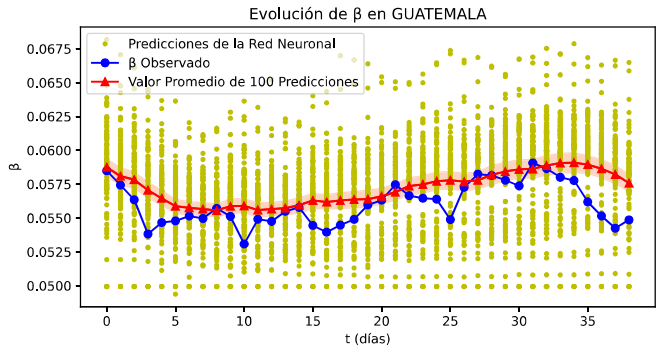


Figura 6.2. Los puntos amarillos corresponden a las 100 predicciones de β cada día para el departamento de Guatemala. El promedio es la β reportada.

La suposición de la distribución de β es utilizada después para definir los intervalos de confianza. Entrenar y probar la red neuronal da funciones de costo como en Figura 6.1.

Distribuciones de los pronósticos para varios departamentos y diferentes días se muestran en la Figura 6.3

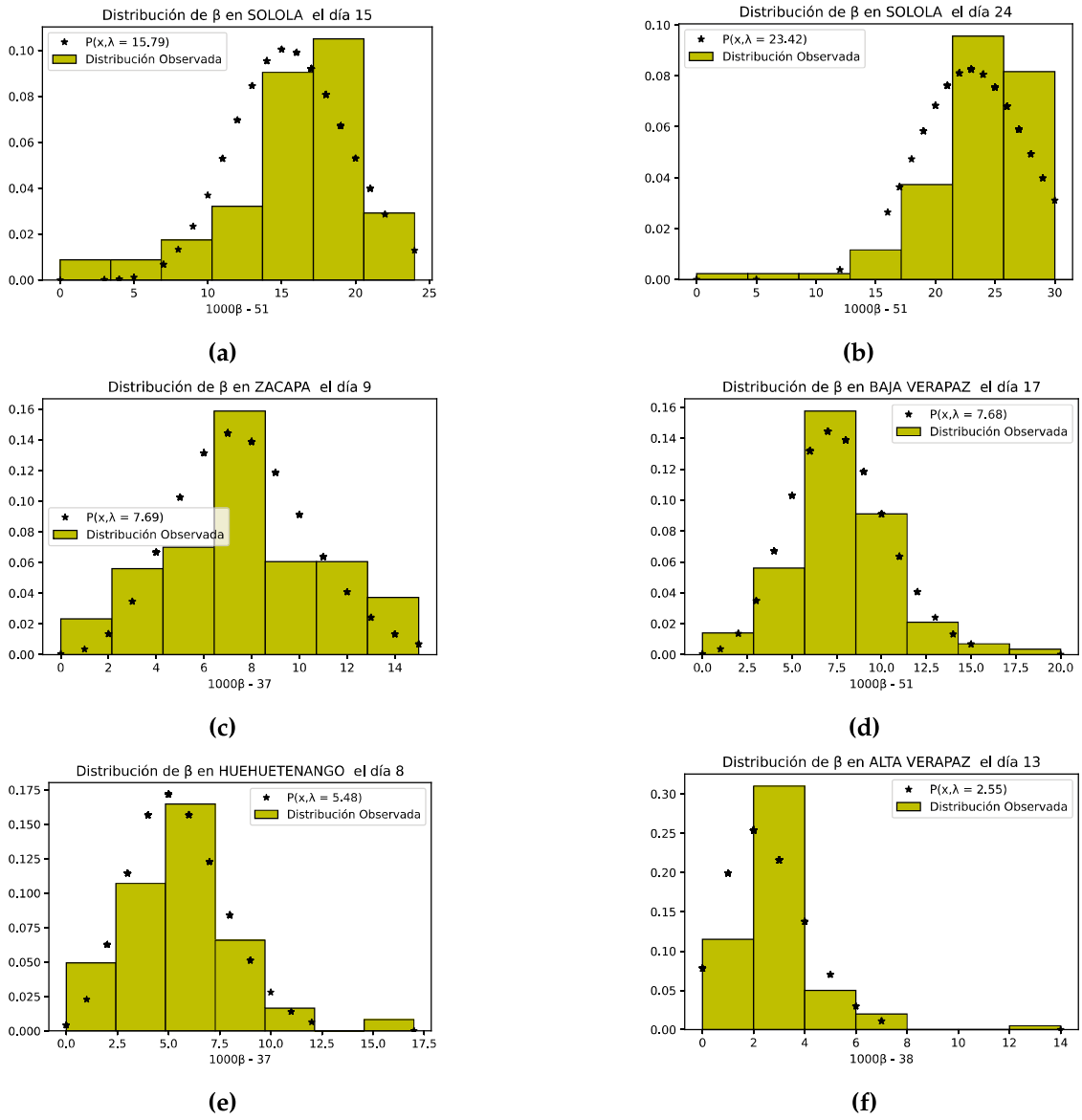


Figura 6.3. Distribución de los 100 resultados de β obtenidos con la red neuronal para diferentes días y departamentos.

Notar que β no es un valor entero por lo que se realizó una transformación para aplicar la distribución (6.1): Se multiplicó cada predicción de la red por 1000, estos datos se aproximan al entero más pequeño y se guardan en un listado, el valor mínimo en este listado se resta a cada uno de los datos en la lista. Las anteriores transformaciones solamente modifican la escala y no afectan el tipo de distribución que caracteriza a los valores originales. Las modificaciones para observar el tipo de distribución fueron las de amplificación, discretización y luego traslación.

La desviación estándar σ y los intervalos de confianza dependen de la distribución que siguen los resultados, por ejemplo para la distribución normal se tiene

$$\sigma_t = \sqrt{\frac{\sum |x_{i,t} - \mu_t|^2}{N}}, \tag{6.2}$$

donde $x_{i,t}$ es el $i-$ésimo resultado en el $t-$ésimo día y μ_t es el promedio de los resultados en el $t-$ésimo día, mientras que para la distribución de Poisson

$$\sigma_t = \sqrt{\mu_t}, \tag{6.3}$$

con el fin de calcular los intervalos de confianza de β, suponiendo que sigue una distribución de Poisson, basta realizar la transformación de amplificación a las predicciones para utilizar 6.3. Luego se hace la operación opuesta a la amplificación (dividir entre la constante de amplificación, 1000 en este caso) para regresar β a la escala original. El intervalo de 90 por ciento de confianza se calcula de la siguiente forma:

$$C_{90\,\%} = 1.645\sigma/\sqrt{n}, \tag{6.4}$$

donde n es el número de predicciones de la red neuronal, 100 en este caso.

Algunos resultados de la red neuronal para la tasa de propagación se muestran a continuación:

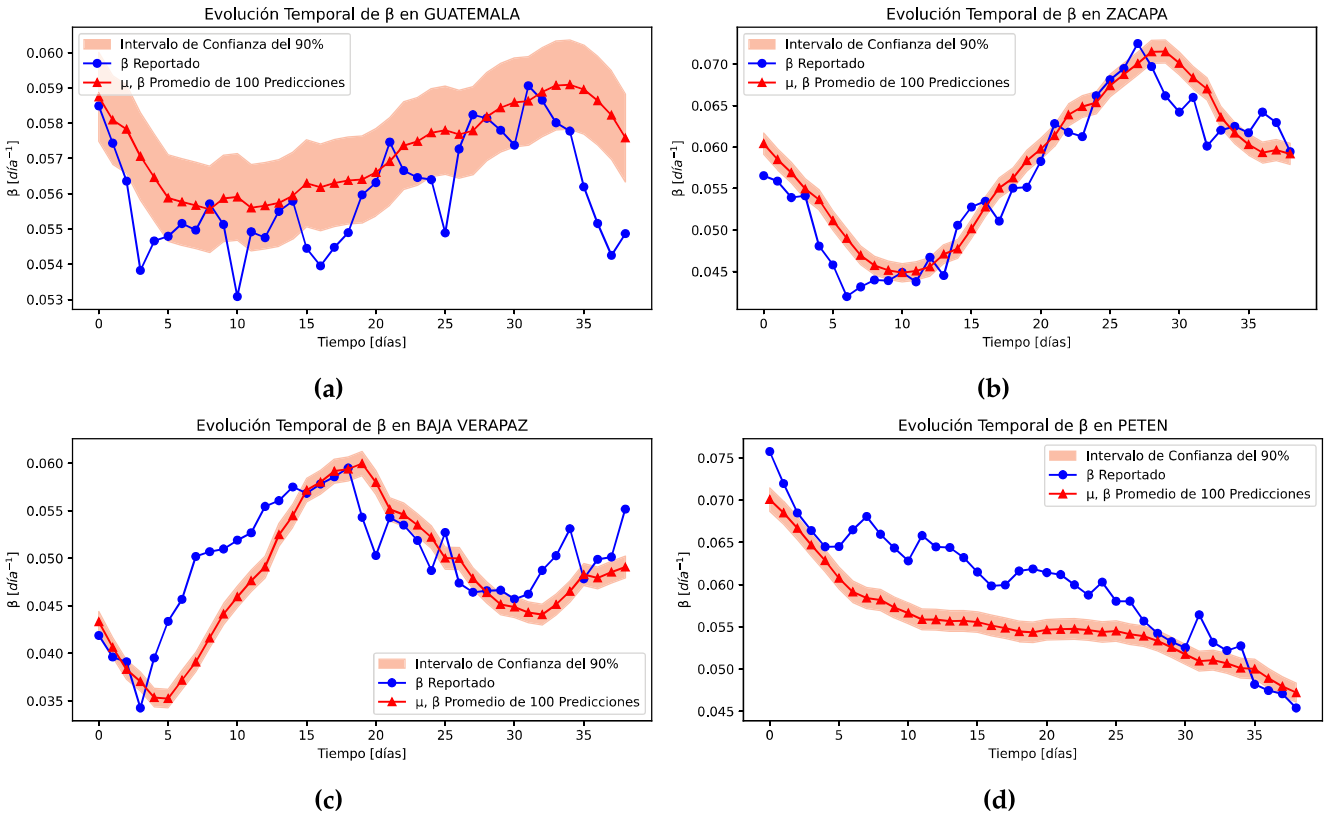


Figura 6.4. Comparación de β observado en la muestra con las predicciones realizadas por la red neuronal en la Figura 4.1 para los departamentos de Guatemala, Zacapa, Baja Verapaz y Petén.

Los valores pronosticados de β son substituidos en el sistema (1.2), con el fin de obtener la evolución en el tiempo de los individuos susceptibles, de los nuevos casos reportados de individuos infectados y del total de acumulados.

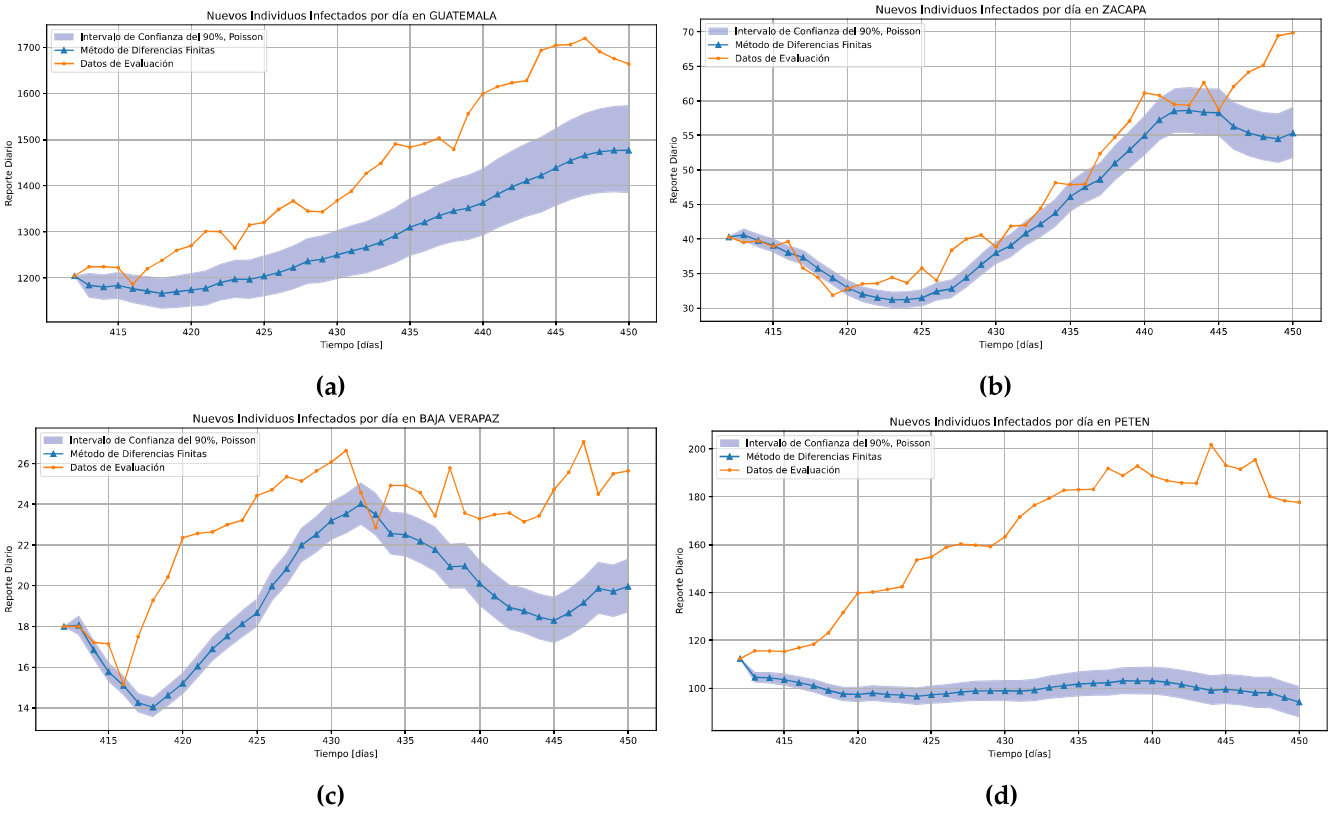


Figura 6.5. Comparación de δA observado en la muestra con las predicciones realizadas.

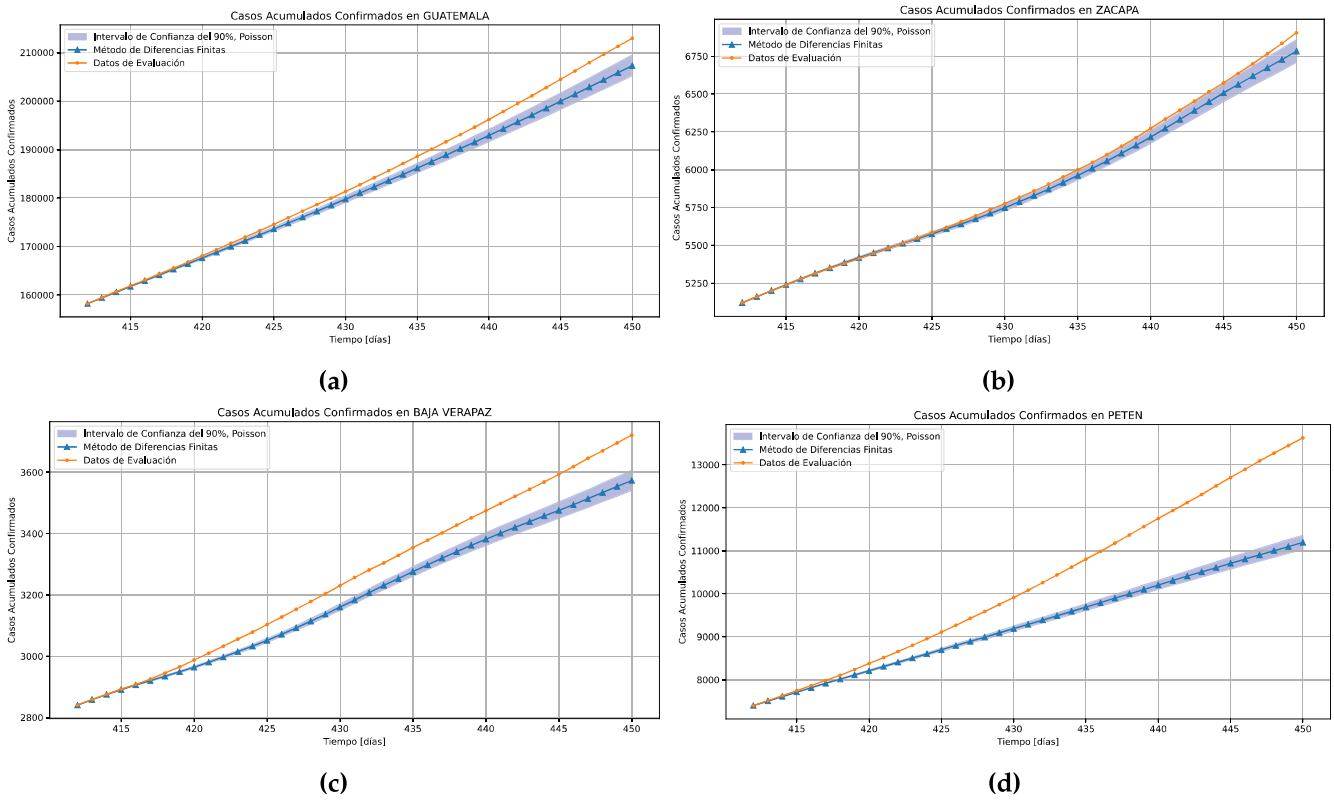


Figura 6.6. Comparación de los casos acumulados observados en la muestra con las predicciones realizadas.

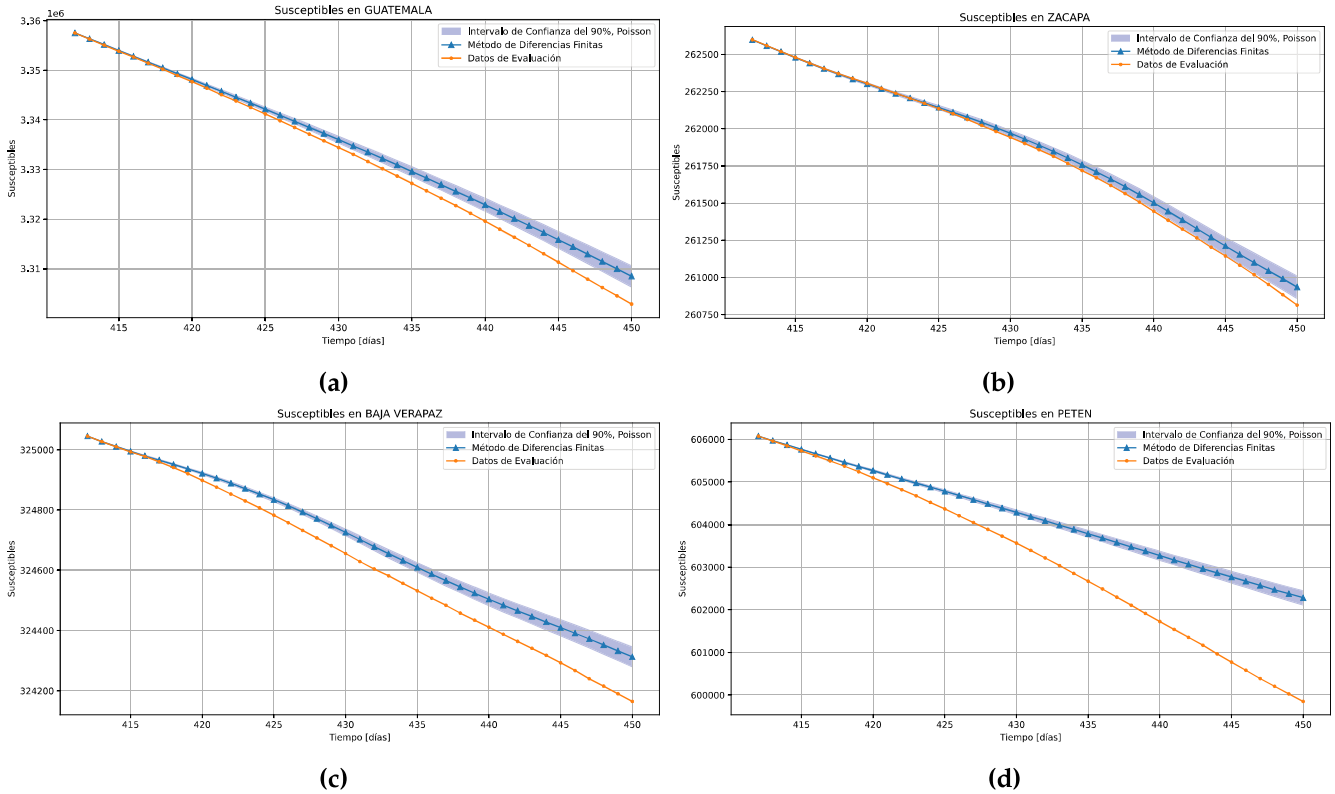


Figura 6.7. Comparación de los susceptibles estimados en la muestra con las predicciones realizadas.

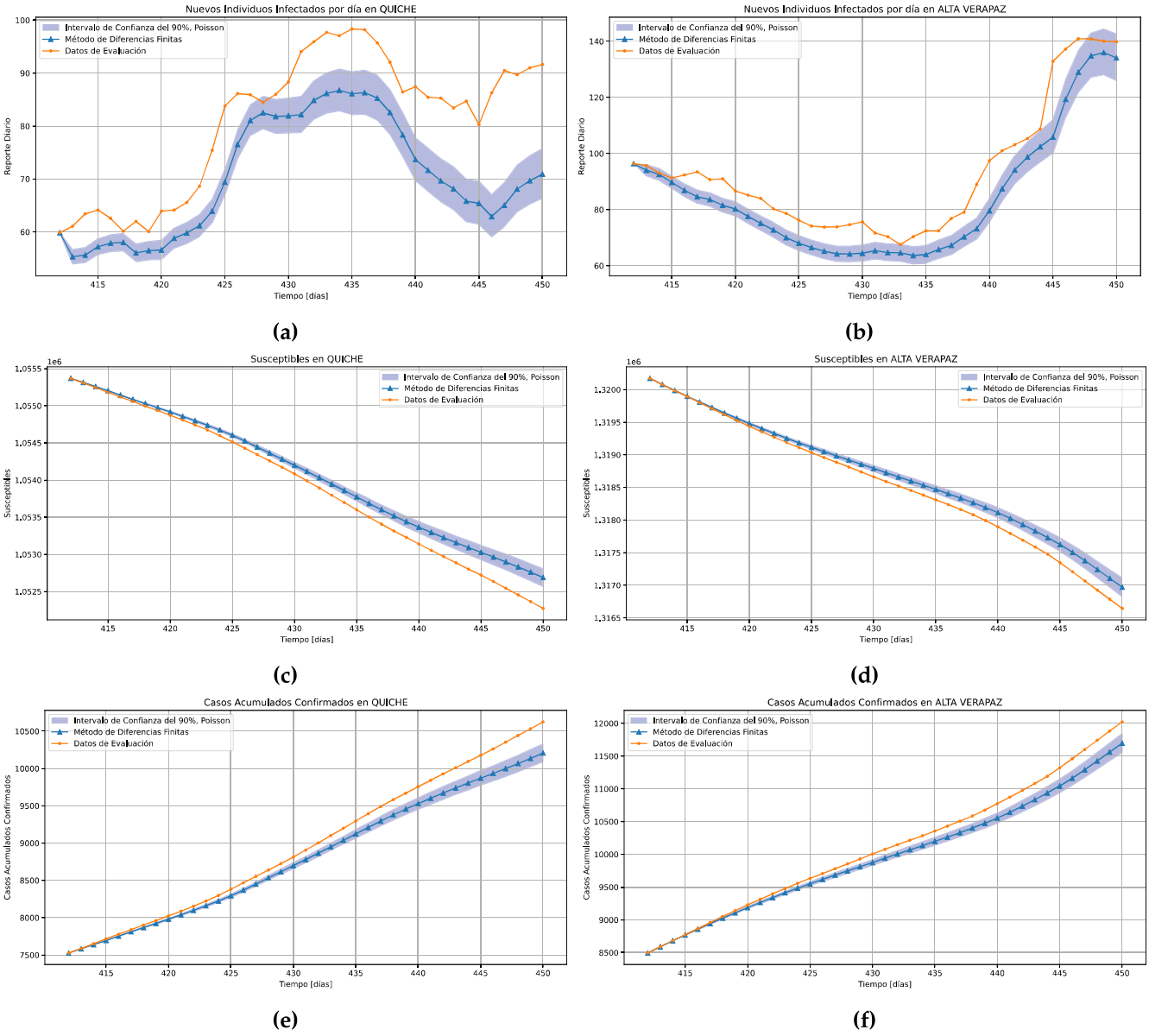


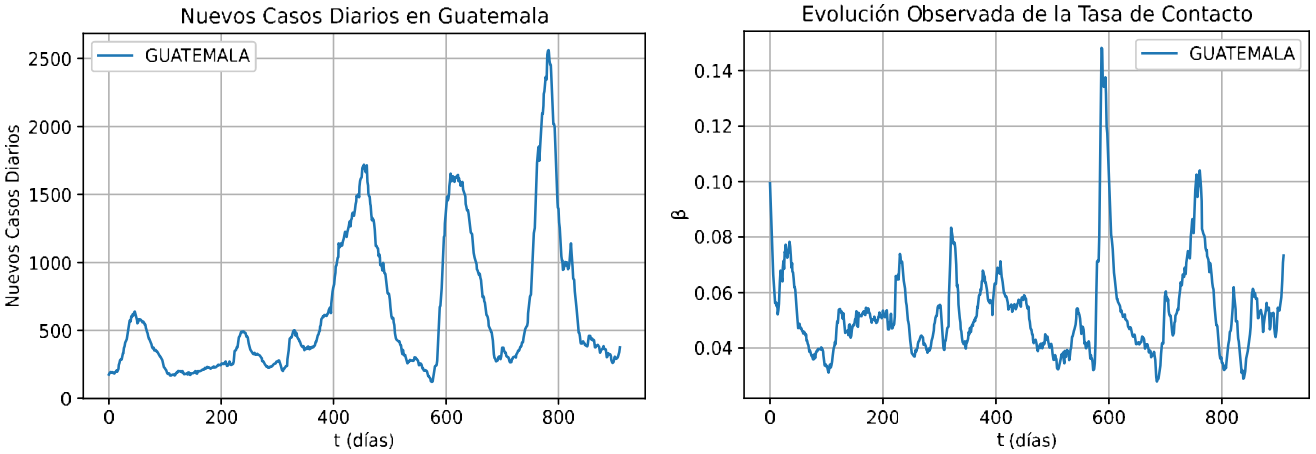
Figura 6.8. Pronóstico de varias variables para los departamentos de Quiché y Alta Verapaz.

7. MODELADO DE β COMO PROCESO ESTOCÁSTICO.

Se investiga la capacidad de los modelo ARIMA para simular la tercera y cuarta ola. Se supone que la tercera ola empieza cuando $\beta > \gamma$, correspondiendo a $t = 583$ días equivalente al 20 de diciembre de 2021, por lo que para entrenar el modelo se emplean los datos previos a esta fecha.

Las redes neuronales brindaron resultados satisfactorios, pero a la vez se limitan a predecir el valor a solamente un día a futuro. Por otro lado, suponer a β constante produce resultados inadecuados, como se verá más adelante, pues predice que la ola de infecciones tendrá una duración mucho mayor de lo que se ha verificado en la realidad a través de los datos recopilados por el Ministerio de Salud.

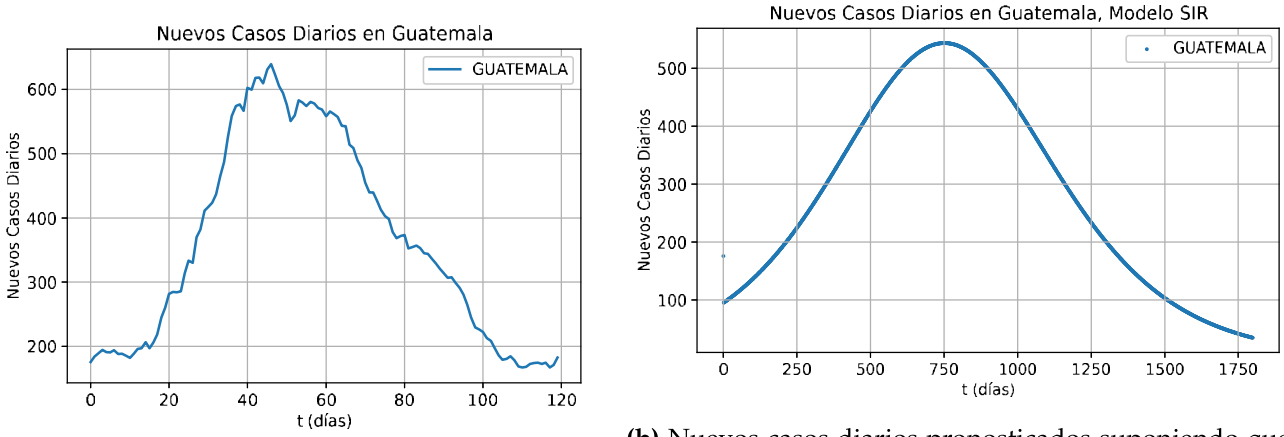
El primer caso se dio a conocer en Guatemala el 13 de marzo de 2020 pero la base de datos inicia desde el 14 de febrero reportando 0 casos en cada departamento. Al inicio de los primeros casos hay demasiado ruido estadístico por lo que se empieza a utilizar para este estudio los casos a partir del día 80 de registro que corresponde al 16 de mayo de 2020. Este será el día 0 en los siguientes gráficos.



(a) **(b)**

Figura 7.1. a) Se distinguen cuatro olas principales de nuevos infectados durante el periodo de mayo 2021 a noviembre de 2022. b) La evolución de β en el intervalo correspondiente al gráfico de la izquierda.

Tome como ejemplo la primera ola con condiciones iniciales $\delta A(0) = 176$, $A(0) = 1992$ $S(0) = 3513783$, $I(0) = 1852$, $R(0) = 139$. En esta ola el promedio de la tasa de contacto es $\bar{\beta} = 0.05138$, sustituyendo este valor y las condiciones iniciales en el modelo SIR se tiene



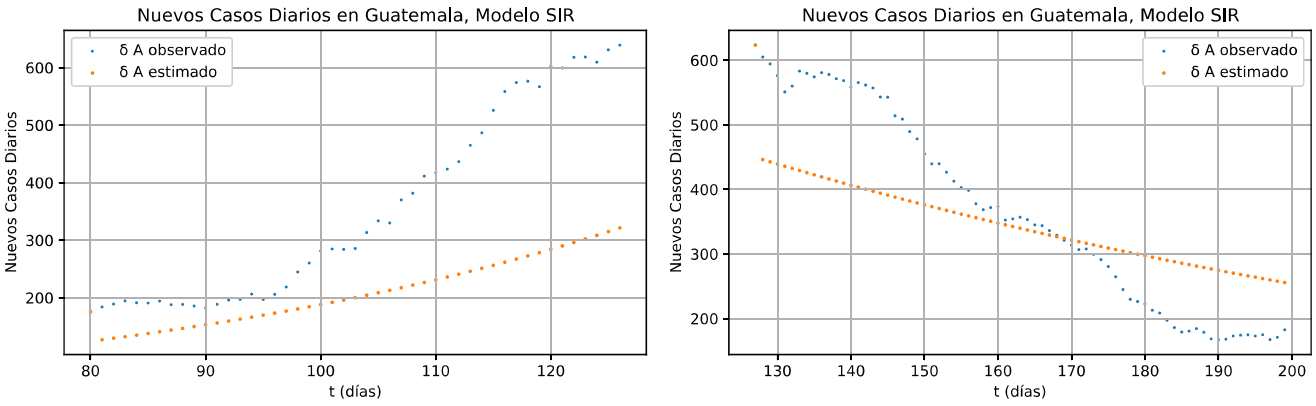
(a) Nuevos casos diarios reportados.

(b) Nuevos casos diarios pronosticados suponiendo que la tasa de contacto es constante y de valor $\beta = 0.05138$.

Figura 7.2. Comparación de los nuevos casos diarios reportados y estimados, con β constante durante todo el intervalo de tiempo.

Vemos que el tiempo de duración de la ola considerando β constante en el modelo es de más de 10 veces el observado en realidad. La misma situación ocurre para las otras tres olas principales. Observe qué sucede si se utiliza un valor de la tasa de contagio cuando los reportes diarios están incrementando β_{up} y otro valor cuando están disminuyendo β_{down}, los

que corresponden a los promedios de β en la subida y en la bajada respectivamente. En [10] se propone utilizar cuatro valores de β para diferentes intervalos durante la ola.



(a) **(b)**

Figura 7.3. Comparación de los nuevos casos diarios reportados y estimados, con β constante dentro de cada intervalo, pero de diferente valor entre un intervalo y otro.

La figura anterior muestra una mejora del modelo SIR describiendo la evolución de la epidemia, pero esta descripción aún es un análisis *retrospectivo*, ya que no estamos prediciendo la evolución de la infección, solo estamos estudiando cómo se ha comportado. Queda clara la importancia de estudiar la evolución de $\beta(t)$ para estimar adecuadamente el tiempo de duración de las olas de la epidemia.

7.1. Tercera Ola de Infectados

Ahora se estudia la fluctuación en β que produce la tercera ola de infectados. A partir de $t = 583$ días se tiene $\beta(t) \geq \gamma$ por lo que para entrenar el modelo ARIMA se considera la serie desde el inicio hasta este tiempo que corresponde a la fecha 20/12/2021. y se intenta predecir el resto de la fluctuación.

```
In [27]: adfuller_test(Beta[0][:583])

         ADF Test Statistic : -2.846191400666716
         p-value : 0.05197255986678822
         #Lags Used : 15
         Number of Observations : 567
         weak evidence against null hypothesis,indicating it is non-stationary

Out[27]: (-2.846191400666716,
          0.05197255986678822,
          15,
          567,
          {'1%': -3.441935806025943,
           '5%': -2.8666509204896093,
           '10%': -2.5694919649816947},
          -5943.04081053161)
```

Figura 7.4. El test ADF indica que la serie no es estacionaria en el intervalo anterior a la tercera ola.

```
In [31]: serie_nueva = Diferencias(Beta[0][:583])

In [32]: adfuller_test(serie_nueva)

         ADF Test Statistic : -8.139002023007132
         p-value : 1.0403721659877836e-12
         #Lags Used : 14
         Number of Observations : 567
         strong evidence against the null hypothesis(Ho), reject the null hypothesis. Data is stationary

Out[32]: (-8.139002023007132,
          1.0403721659877836e-12,
          14,
          567,
          {'1%': -3.441935806025943,
           '5%': -2.8666509204896093,
           '10%': -2.5694919649816947},
          -5939.837095794135)
```

Figura 7.5. El test ADF indica que luego de diferenciar una vez la serie, se consigue una serie estacionaria por lo que $d = 1$ en el modelo ARIMA(p,d,q).

El test ADF en Figura 7.4indica que la serie $\beta(t)$ no es estacionaria para el intervalo $t \in [0, 582]$ por lo que realizamos una diferenciación, a esta nueva serie le aplicamos el test con el resultado mostrado en Figura 7.5. Así que el modelo a utilizar será ARIMA, con $d = 1$. Mediante las gráficas de autocorrelación y autocorrelación parcial determinamos los órdenes p y q correspondientes a AR y MA.

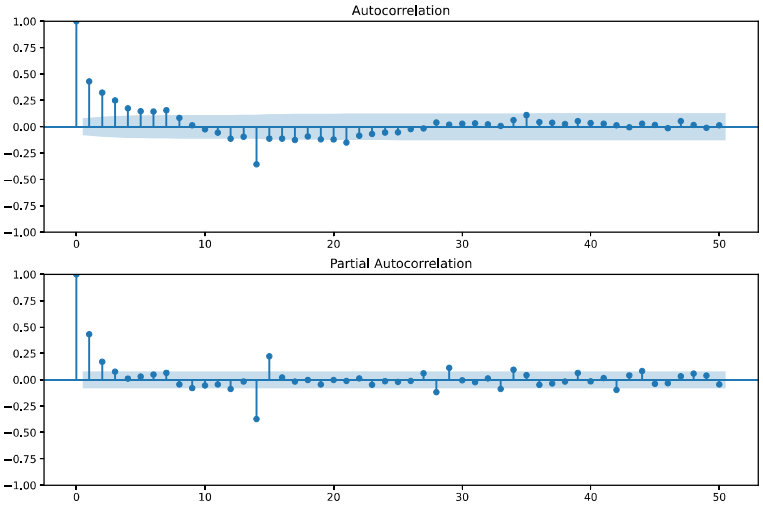


Figura 7.6. Funciones de autocorrelación y de autocorrelación parcial de la primera diferenciación de β del departamento de Guatemala para determinar los parámetros del modelo ARIMA.

Las gráficas de las funciones de autocorrelación y de autocorrelación parcial sugieren que $p = 2$ y $3 \leq q \leq 7$ por lo que emplearemos inicialmente un modelo ARIMA(2,1,3) a un ARIMA(2,1,7).

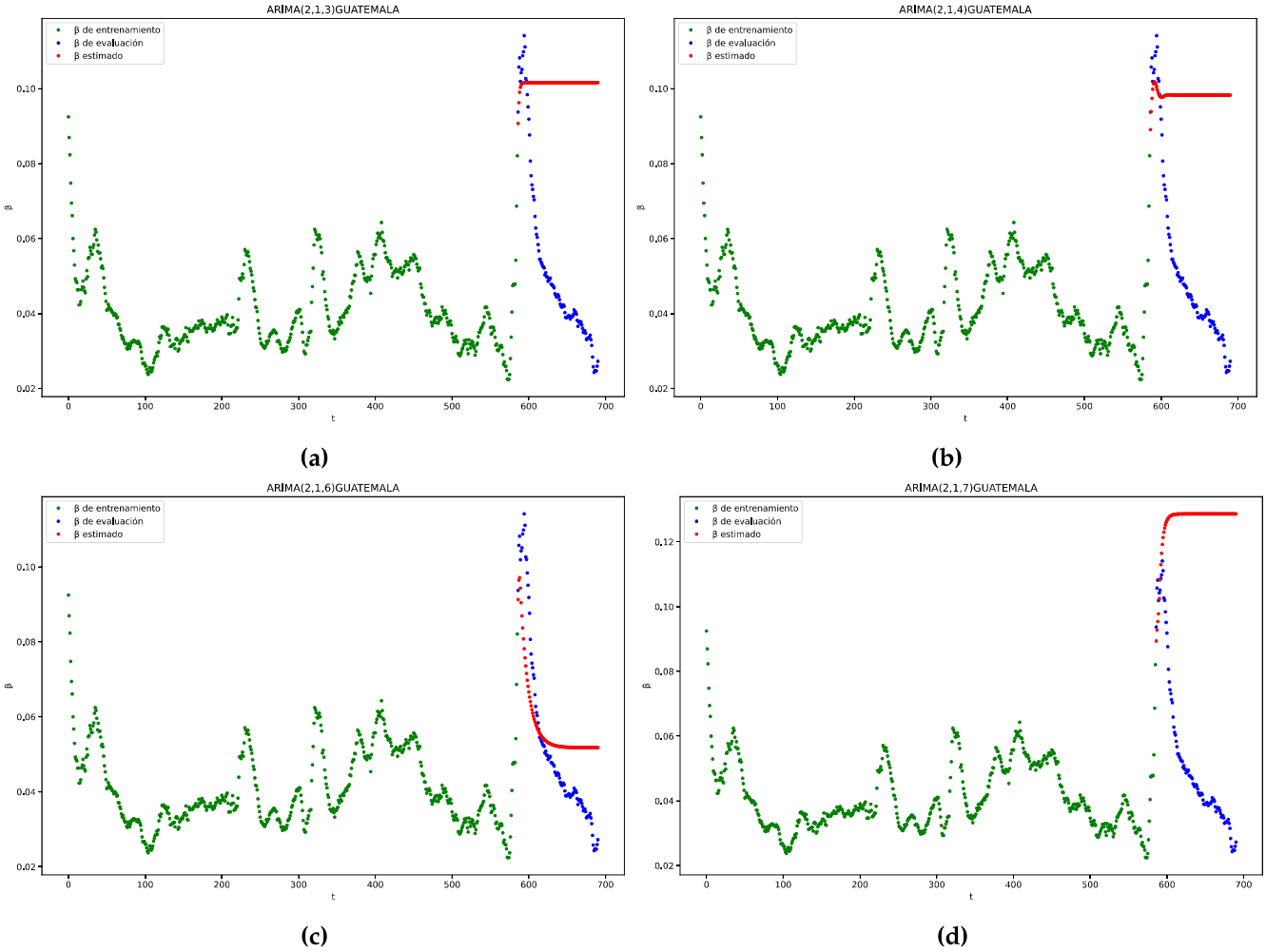


Figura 7.7. Predicciones de β para diferentes modelos ARIMA.

Las mejores predicciones a partir del día 586 corresponden a ARIMA(2,1,6), en Figura 7.13 (c). Se evalúa este modelo partiendo de diferentes días. Será importante considerar que el promedio de los datos de entrenamiento es $\beta = 0.0409$, mientras que las predicciones de ARIMA(2,1,6) en la Figura 7.13 finalizan en 0.0518. Iniciando las predicciones en el intervalo $t \in [583, 589]$ tenemos las siguientes proyecciones de β y δA.

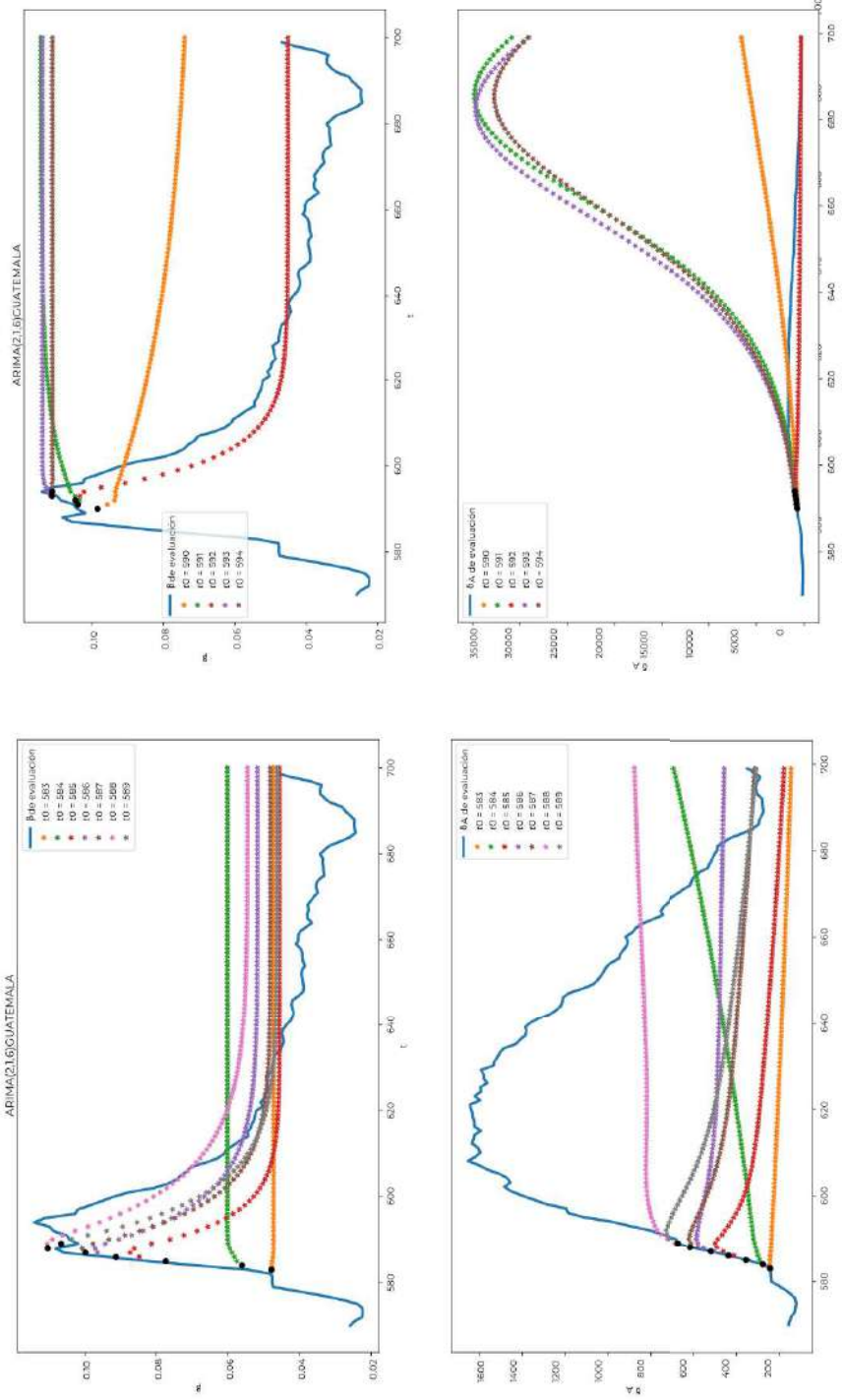


Figura 7.8. Proyecciones de β y δA para el intervalo $t \in [582, 589]$ en la parte inferior y $t \in [590, 595]$ en la parte superior.

En el segundo intervalo de tiempo las predicciones no solo no siguen la tendencia sino

que divergen completamente de los valores observados. Nótese que se ha utilizado $d = 1$ porque el test ADF indicó que la serie se convertía en estacionaria solo al diferenciarla una vez. Supongamos que la serie $\beta(t)$ es ya estacionaria, por lo que $d = 0$ y a continuación determinamos p y q. De la Figura 7.9 proponemos utilizar un modelo ARIMA(2,0,q), con q a ser determinado, observando el error cuadrático medio en la Figura 7.10 para diferentes valores de q durante los primeros diez días.

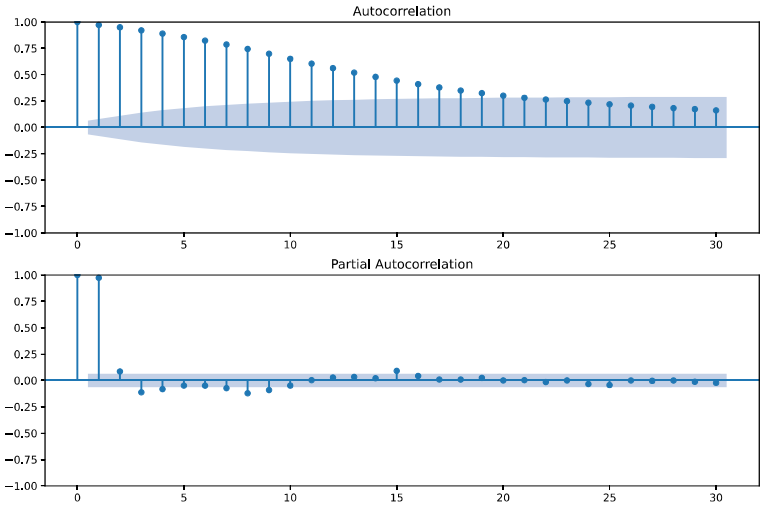


Figura 7.9. Funciones de autocorrelación y de autocorrelación parcial para β del departamento de Guatemala para determinar los parámetros del modelo ARIMA.

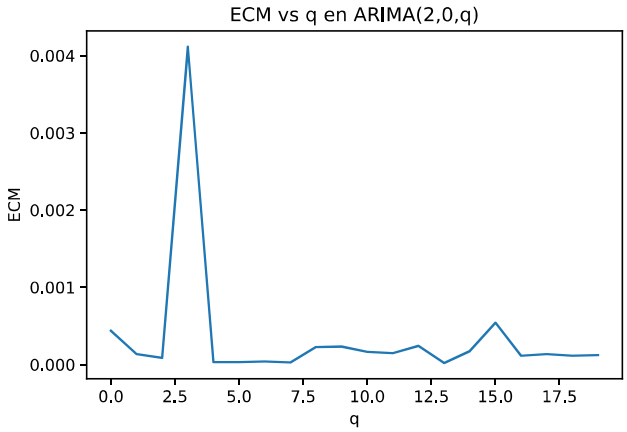
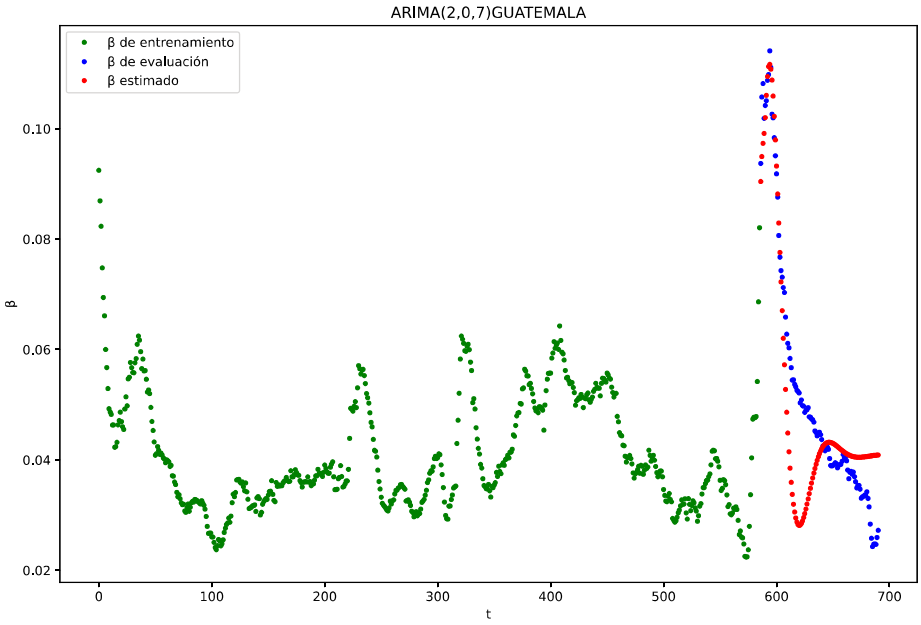
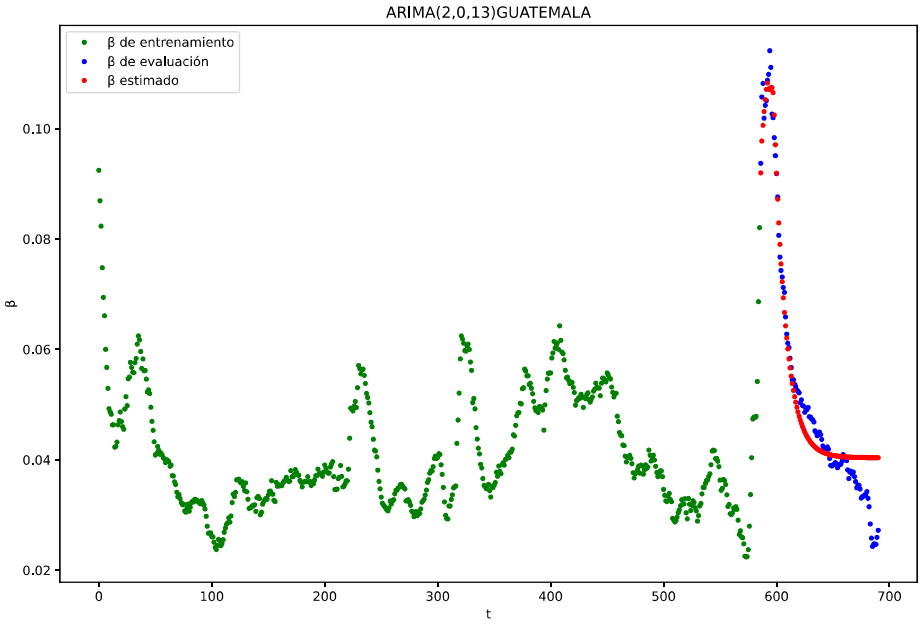


Figura 7.10. Error cuadrático medio para diferentes valos de q en el modelo ARIMA(2,0,q).

Se obtiene el menor error cuadrático medio cuando utilizamos $q = 7$ y $q = 13$.



(a)



(b)

Figura 7.11. Modelos ARIMA(2,0,7) (a) y ARIMA(2,0,13) (b).

Las predicciones de ARIMA(2,0,7) y ARIMA(2,0,13) llegan a un mínimo de 0.0408 y 0.0403 respectivamente, a diferencia de ARIMA(2,1,6) que no era capaz de volver al menos al promedio de la serie de entrenamiento.

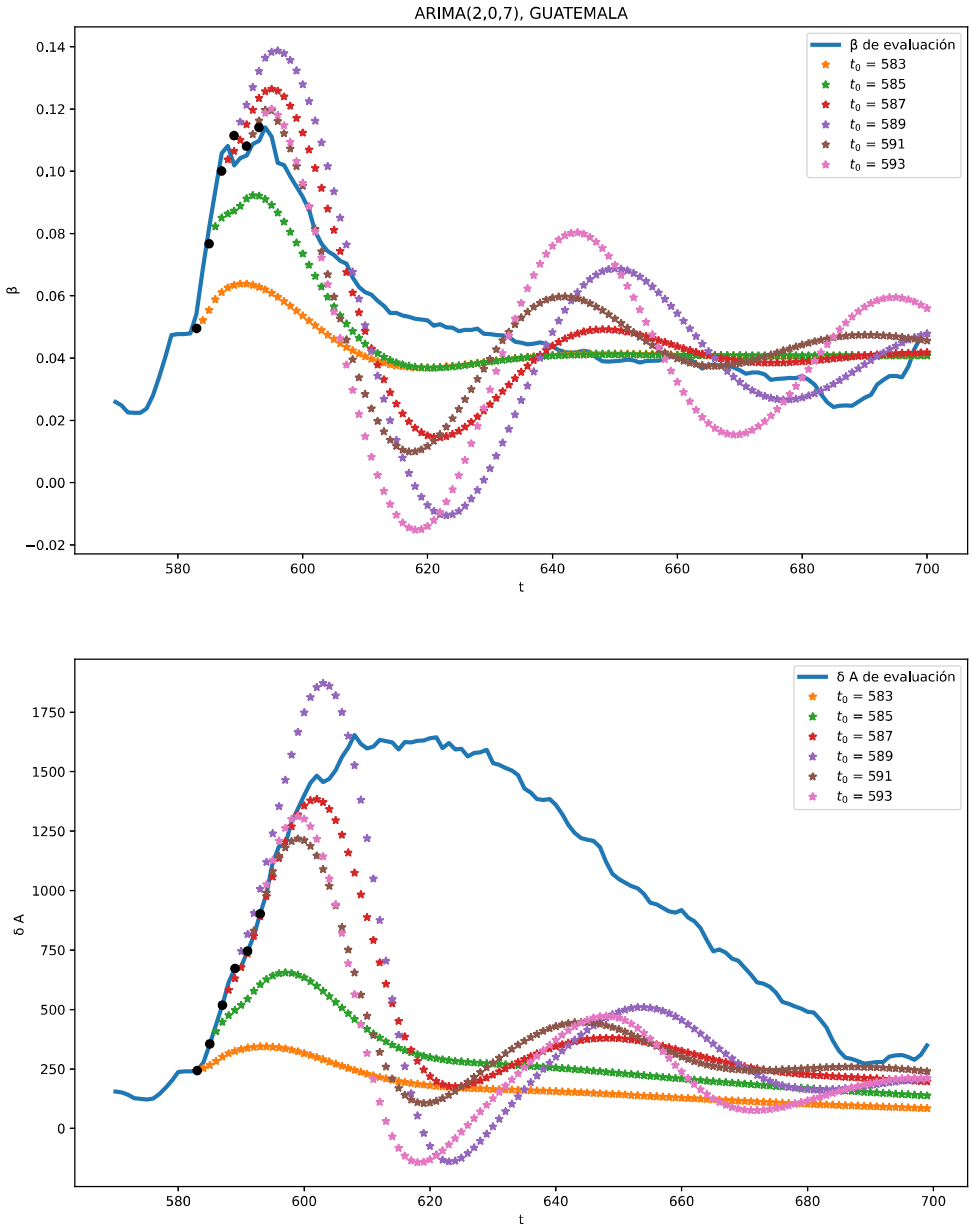
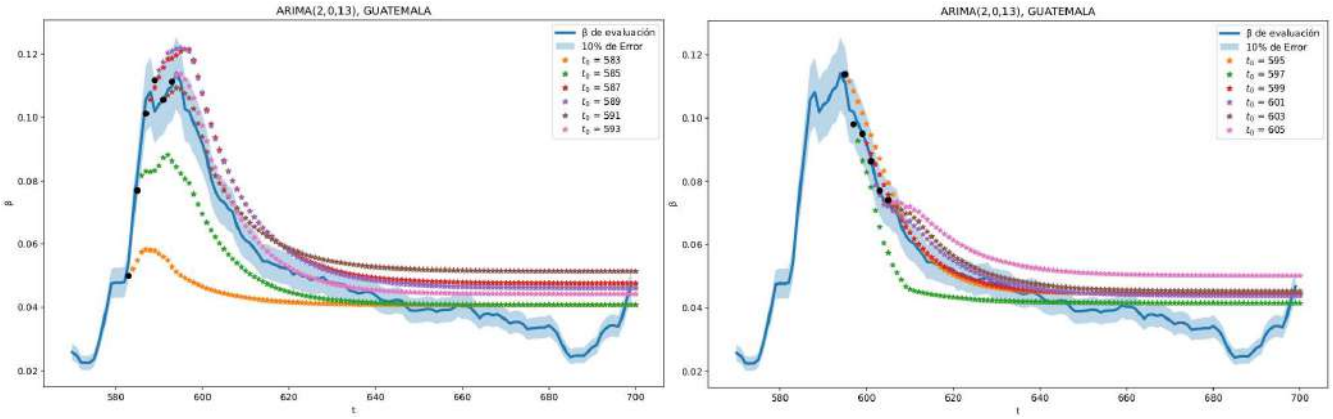


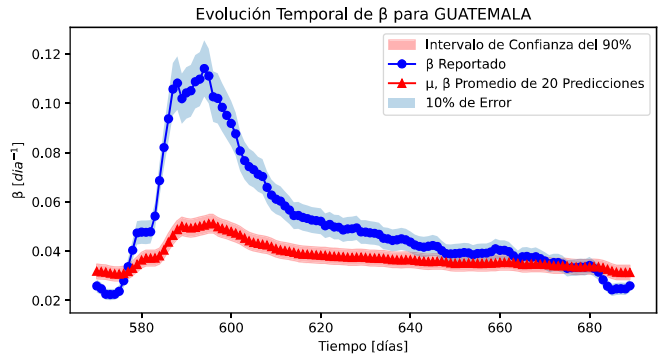
Figura 7.12. Predicciones de ARIMA(2,0,7).

Observe que las variables que se están estudiando no son negativas, lo que debe ser una restricción en las predicciones. Para $q = 7$ no se cumple con dicha restricción. Veamos qué ocurre con $q = 13$.

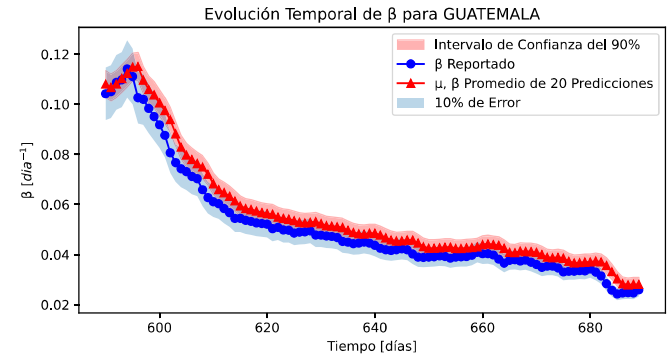


(a) ARIMA(2,0,13), $t_0 \in [583, 593]$

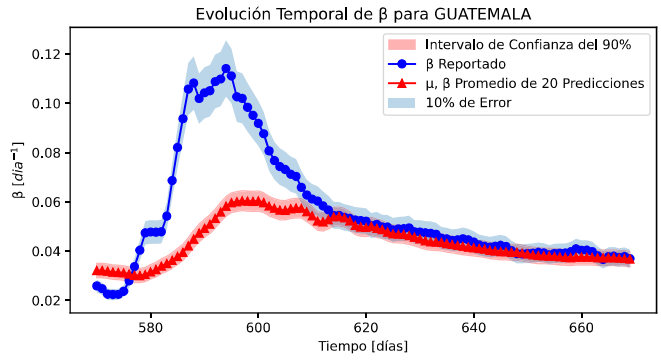
(b) ARIMA(2,0,13), $t_0 \in [595, 605]$



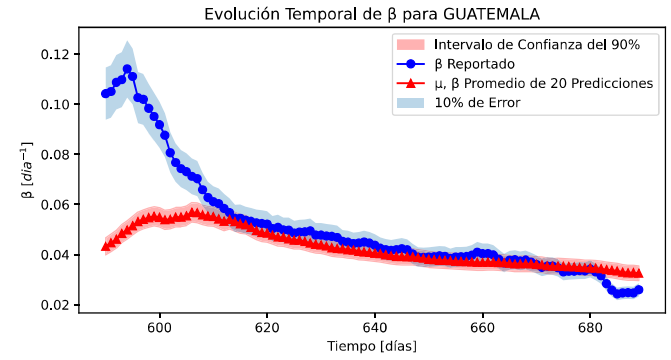
(c) Red Neuronal, $t_0 = 570$, look back $= 2$



(d) Red Neuronal, $t_0 = 590$, look back $= 2$



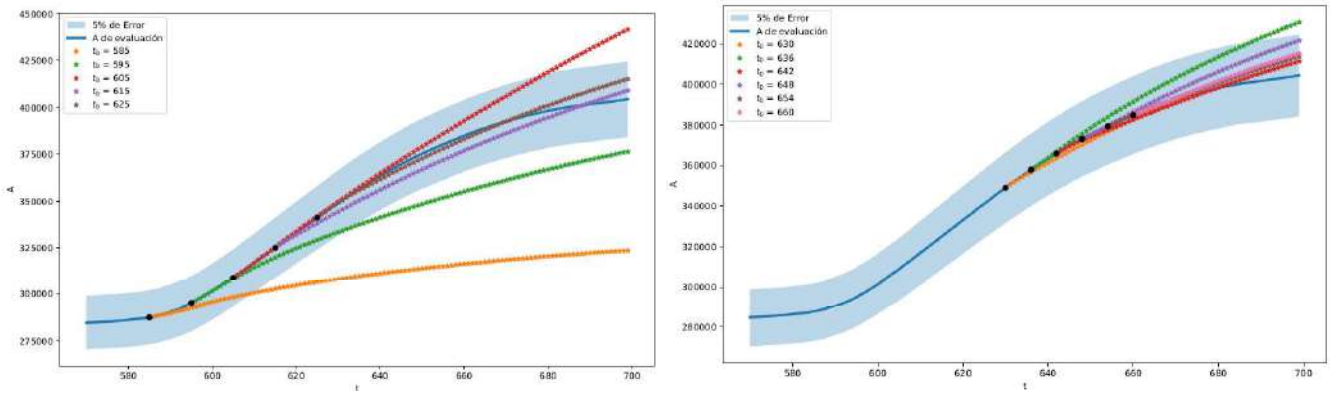
(e) Red Neuronal, $t_0 = 570$, look back $= 30$



(f) Red Neuronal, $t_0 = 590$, look back $= 30$

Figura 7.13. Predicciones de β para ARIMA(2,0,13) y red neuronal con diferente *look back*, que es la cantidad de datos utilizada para pronosticar el valor al día siguiente.

Note que las predicciones de β en Figura 7.13a siguen la tendencia de los valores observados, sin embargo tienen una discrepancia mayor al 10 %. Continúa presente el inconveniente de que las predicciones no disminuyen lo suficiente presentando un comportamiento asintótico que no es observado en la realidad. El orden de la parte autorregresiva del modelo ARIMA es $p = 2$, lo que sugiere utilizar un *look back* $= 2$ en la red neuronal como en la Figura 7.13d, compare con la Figura 7.13f donde se ha utilizado *look back* $= 30$ como en la sección 6.



(a) (b)

Figura 7.14. Estimación de los casos acumulados en la tercera ola mediante el modelo ARIMA(2,0,13) para diferentes condiciones iniciales.

Conforme se actualiza el modelo se obtienen estimados dentro del 5 % de error para los acumulados durante hasta 80 días como se muestra en la Figura 7.14.

7.2. Cuarta Ola de Infectados

Se realizarán las predicciones a partir de $t = 727$ días, correspondiente al 13/05/2022 pues a partir de esta fecha $\beta > \gamma$. El test ADF indica que la serie hasta el día anterior a la predicción es estacionaria por lo que $d = 0$.

```
In [5]: adfuller_test(Beta[0][:727])

        ADF Test Statistic : -3.475149042145694
        p-value : 0.008648732829846168
        #Lags Used : 20
        Number of Observations : 706
        strong evidence against the null hypothesis(Ho), reject the null hypothesis. Data is stationary

Out[5]: (-3.475149042145694,
         0.008648732829846168,
         20,
         706,
         {'1%': -3.439646367660705,
          '5%': -2.8656425177031375,
          '10%': -2.5689546724554404},
         -7269.375504427285)
```

Figura 7.15. El test ADF indica que la serie de datos es estacionaria en el intervalo que precede a la cuarta ola.

Ahora se debe observar las funciones de autocorrelación y de autocorrelación parcial.

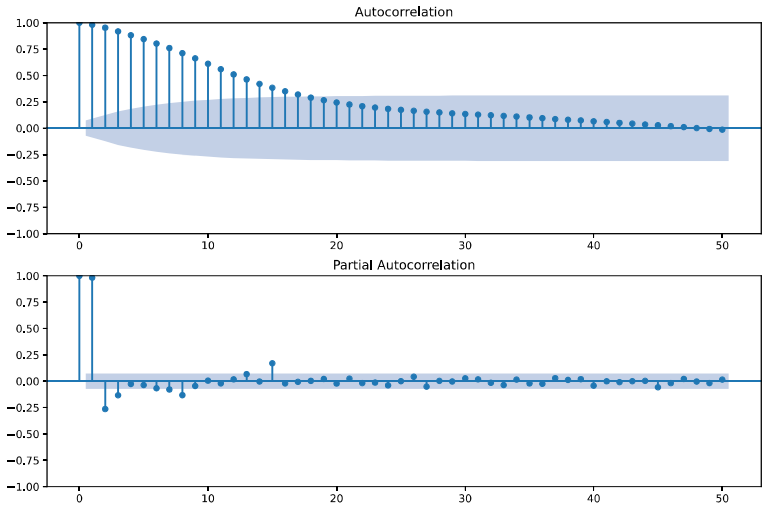


Figura 7.16. Funciones de autocorrelación y autocorrelación parcial para determinar los parámetros del modelo ARIMA.

Lo que indica que $p = 3$ y $1 \leq q \leq 17$, por lo que evaluamos ARIMA(3,0,q) para diferentes valores de q, en las predicciones de los primeros 10 días. La gráfica del error cuadrático medio es

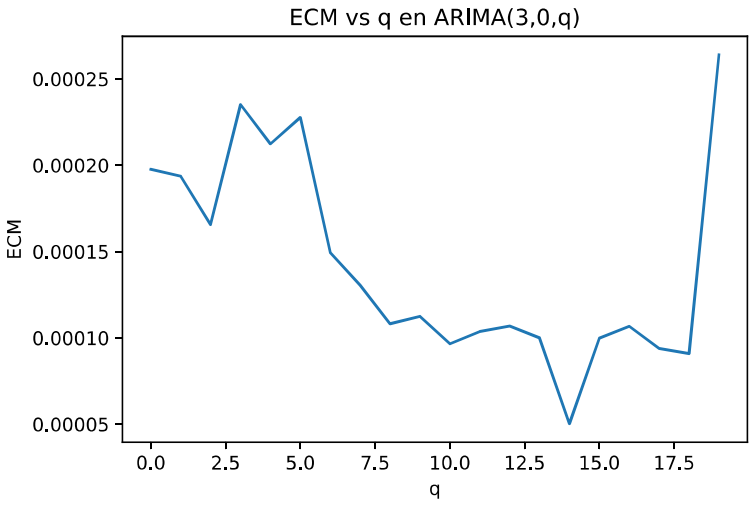


Figura 7.17. Error cuadrático medio para diferentes valos de q en el modelo ARIMA(3,0,q).

El mínimo error cuadrático medio en la Figura 7.17 corresponde a $q = 14$ lo que sugiere utilizar un modelo ARIMA(3,0,14).

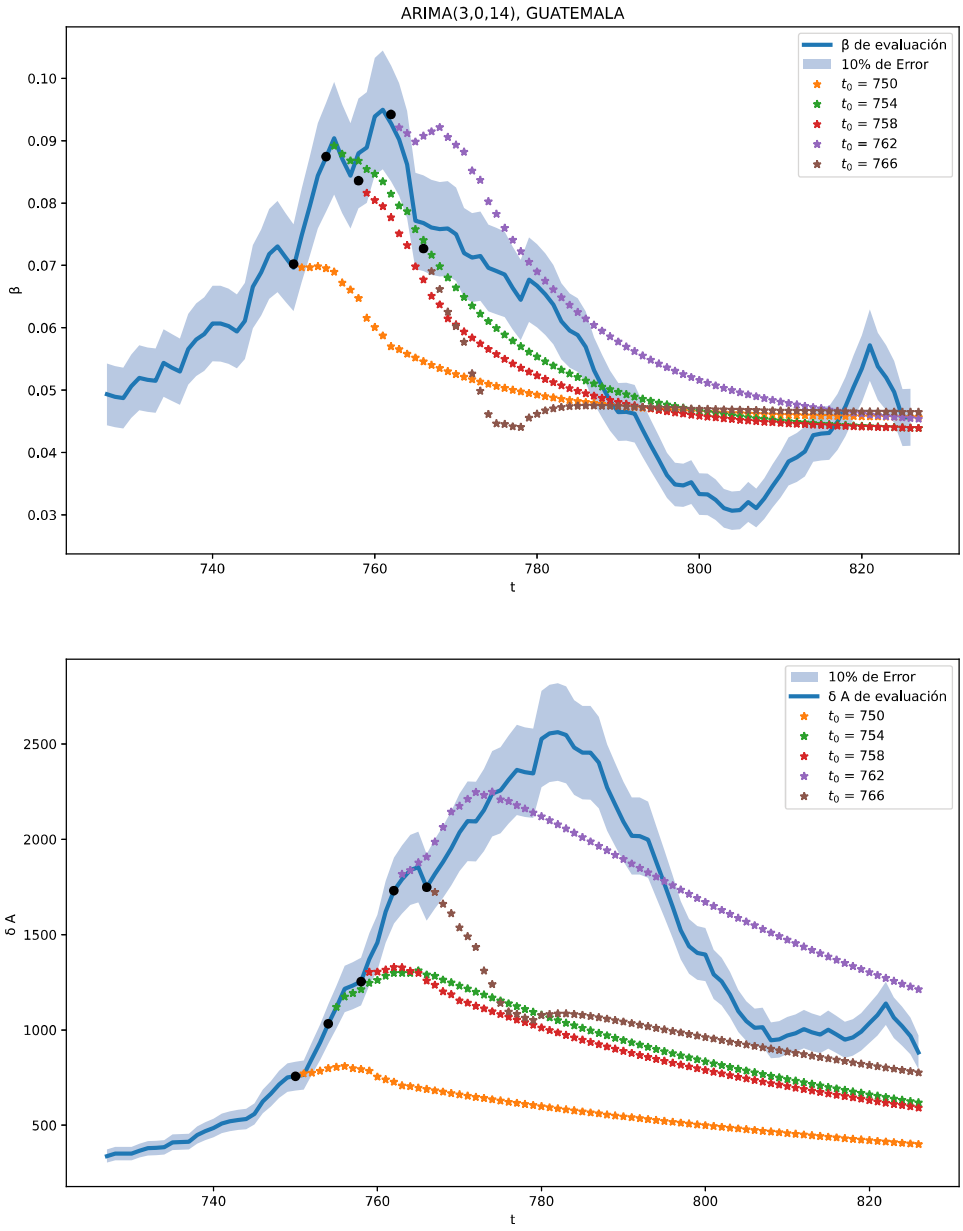


Figura 7.18. Predicciones del modelo ARIMA(3,0,14) para β y δA con condiciones iniciales en la región de incremento de β.

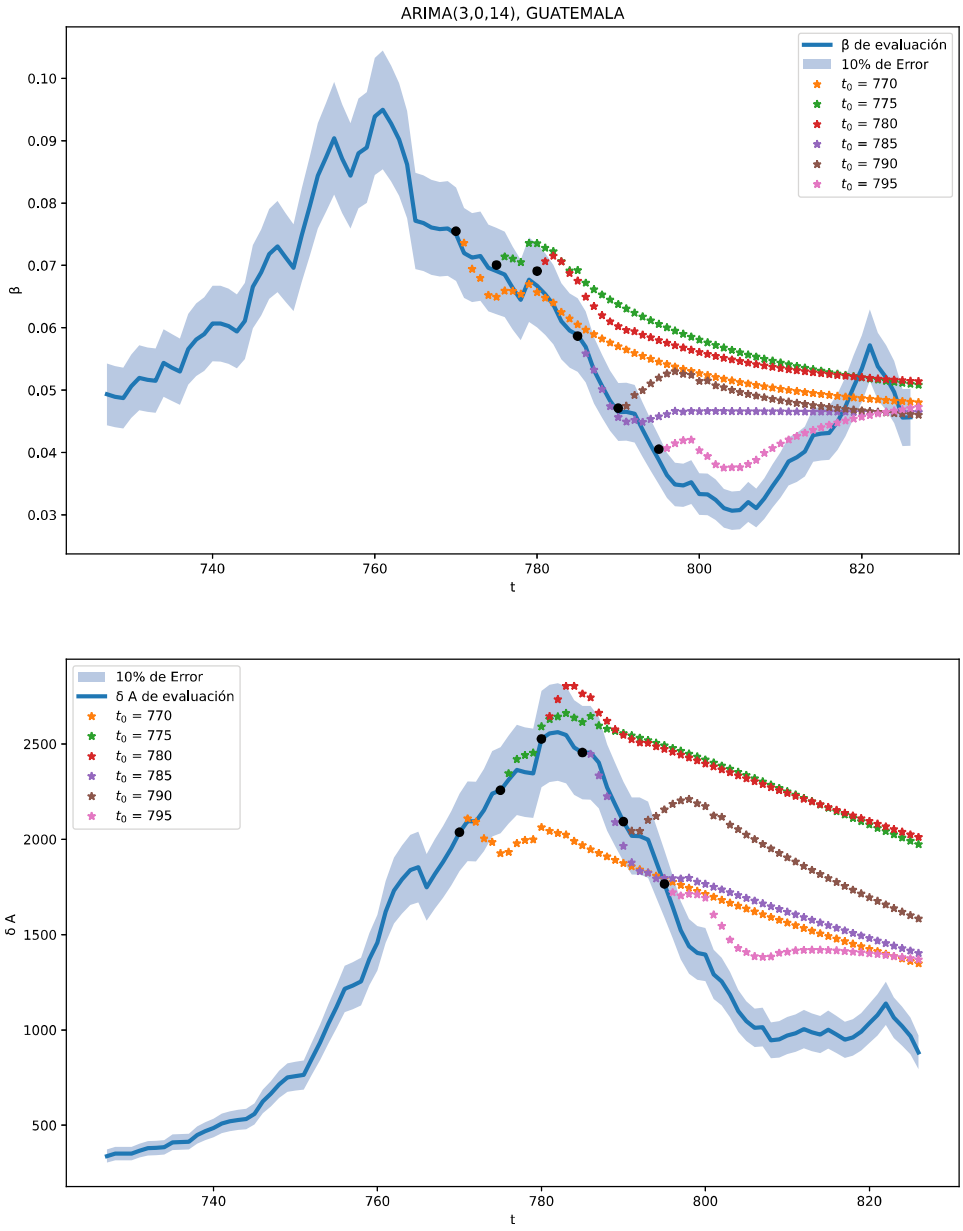


Figura 7.19. Predicciones del modelo ARIMA(3,0,14) para β y δA con condiciones iniciales en la región de disminución de β.

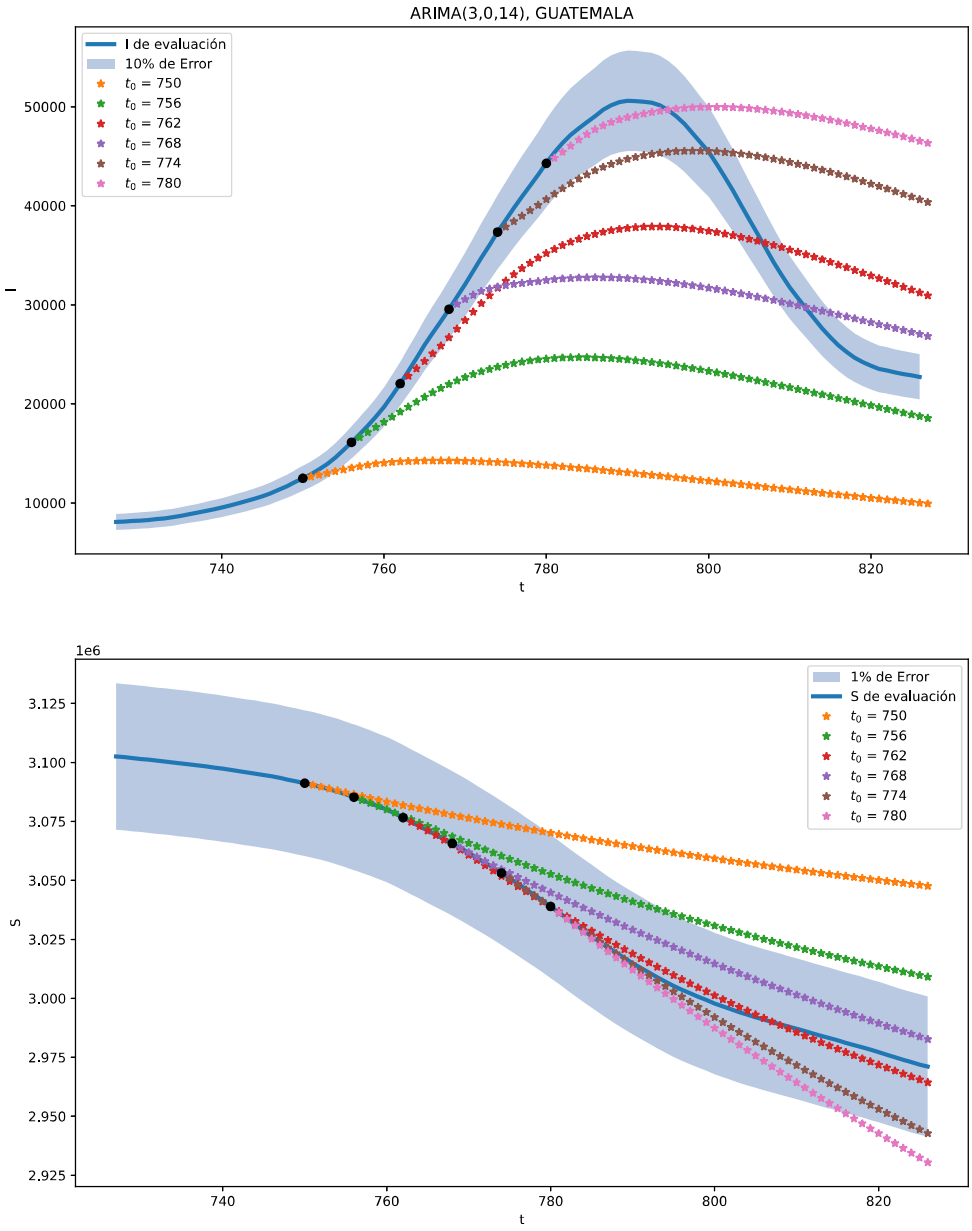


Figura 7.20. Predicciones de I y S y región de incerteza de 10 % y 1 % respectivamente.

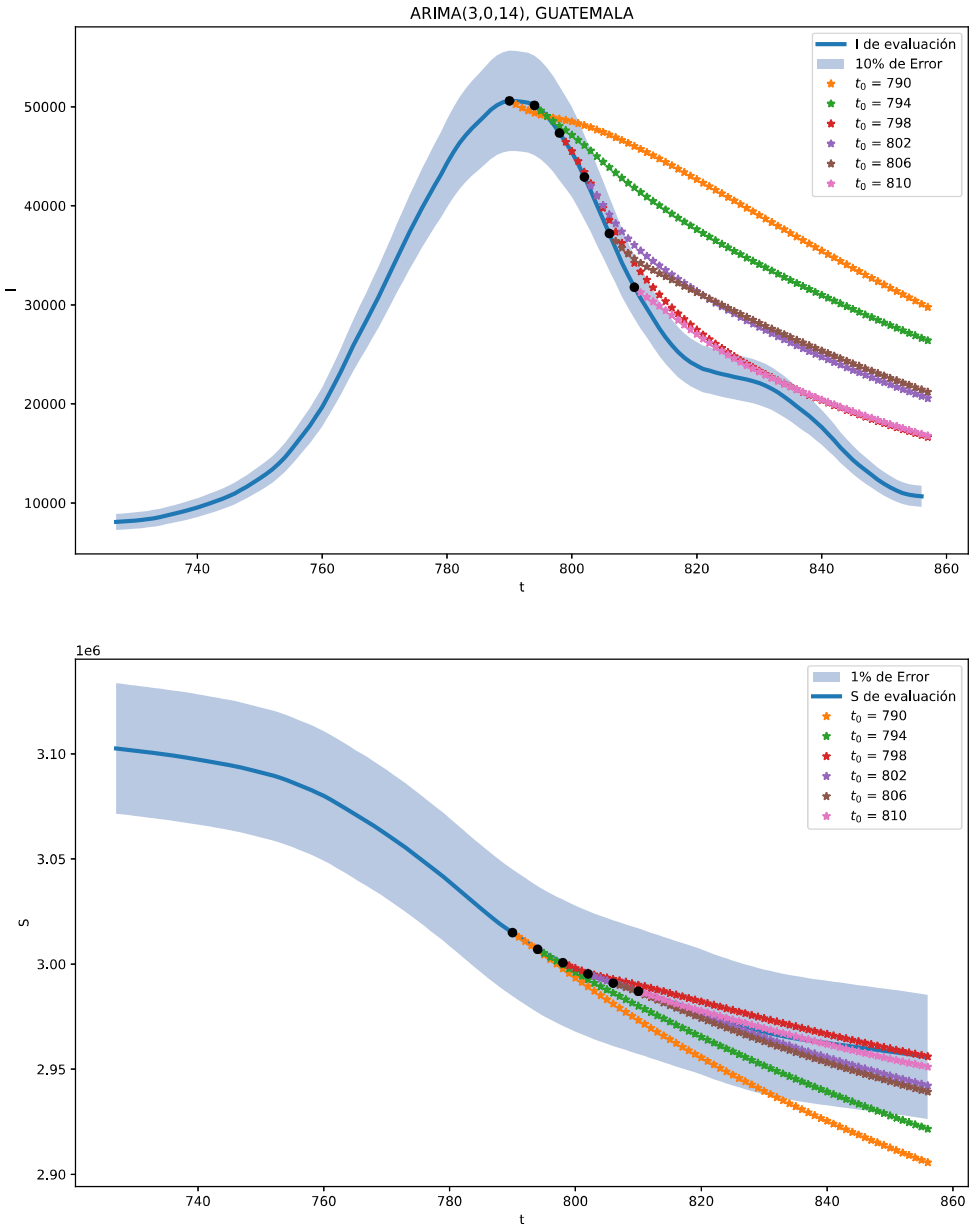


Figura 7.21. Continuación de las predicciones de I y S y región de incerteza de 10 % y 1 % respectivamente.

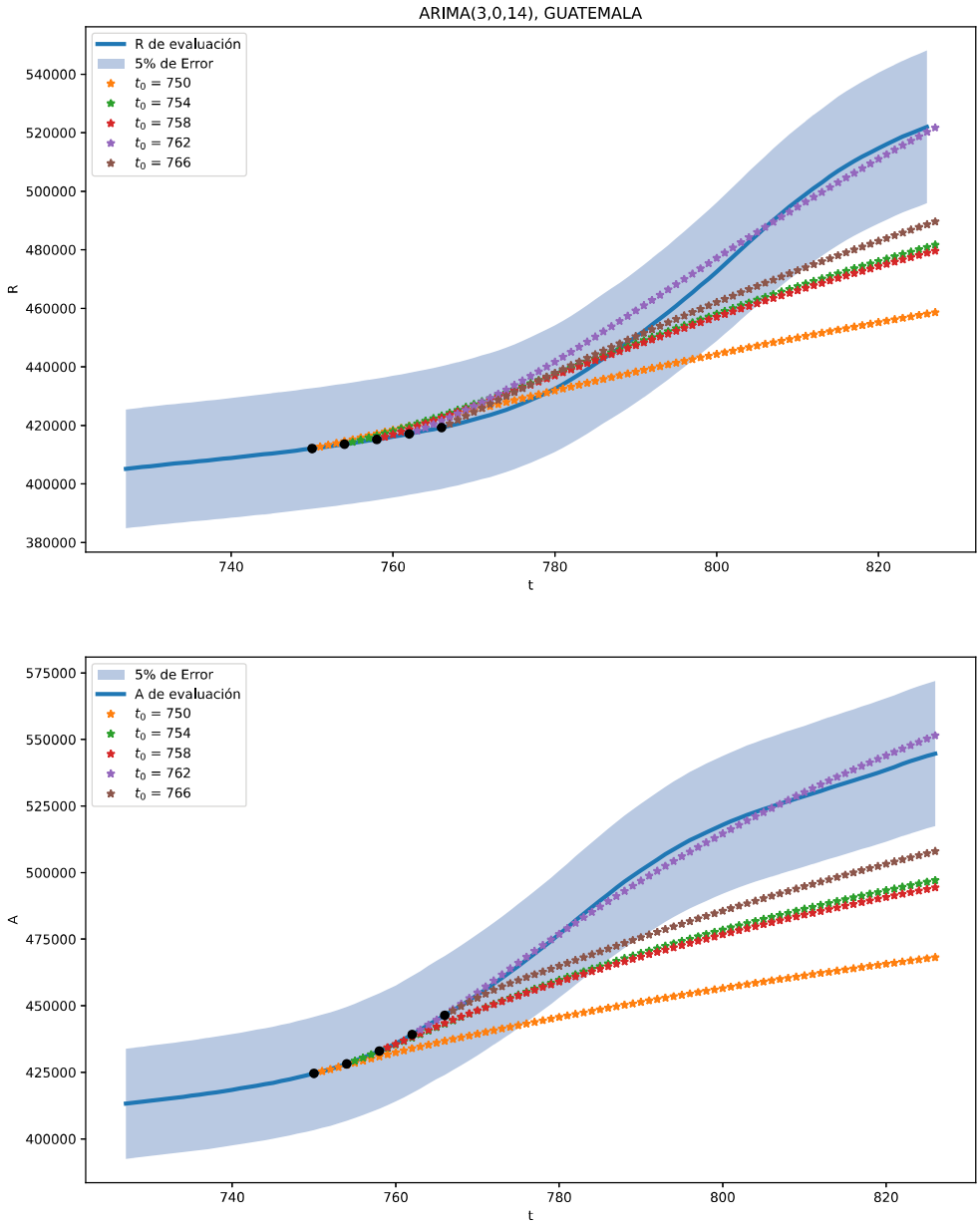


Figura 7.22. Removidos y Acumulados con región de incerteza del 5 % cada uno.

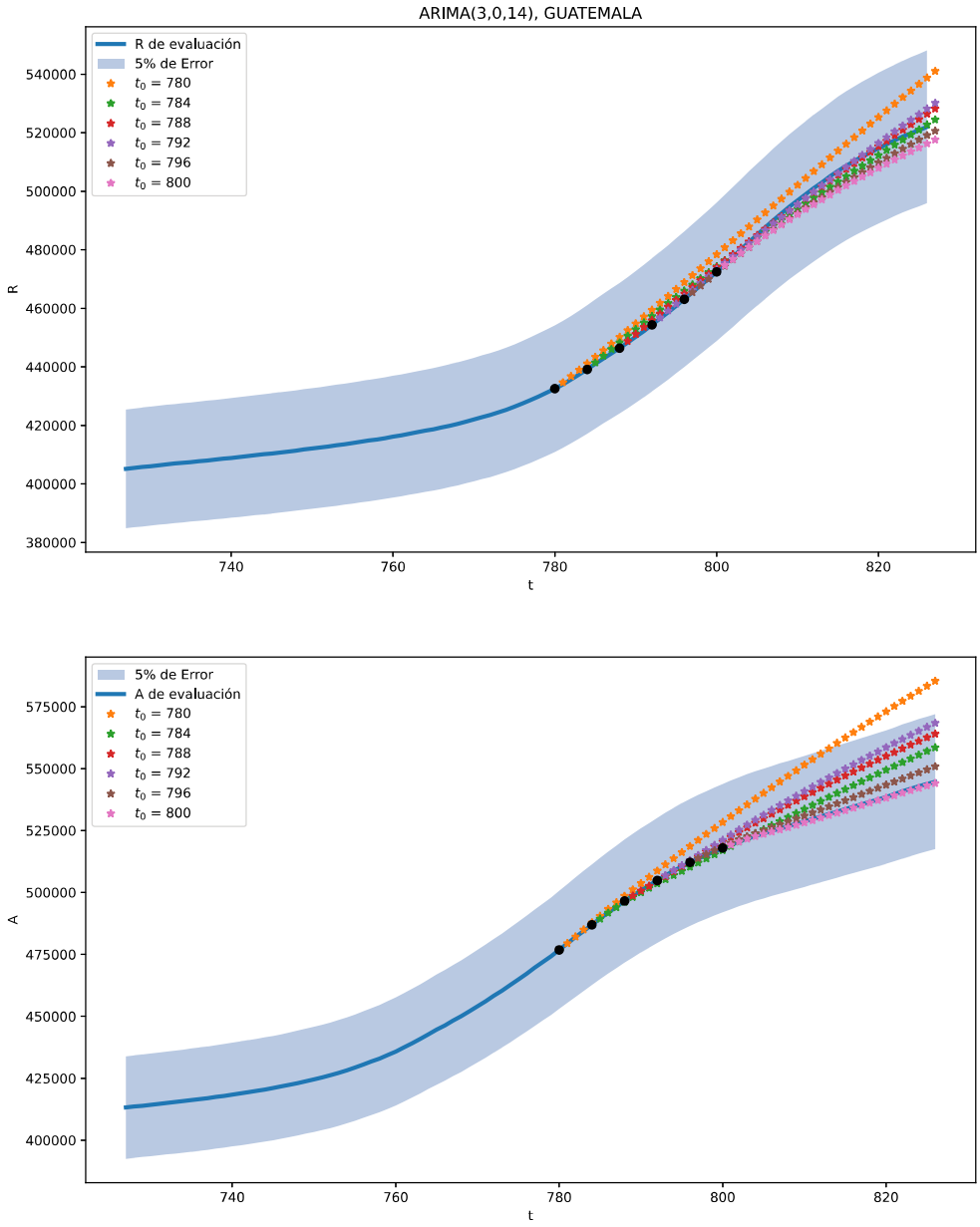


Figura 7.23. Continuación de removidos y acumulados con región de incerteza del 5 % cada uno.

8. DISCUSIÓN DE RESULTADOS

Aunque el test ADF indicó que la serie de datos anterior a la tercera ola corresponde a un proceso no estacionario, los resultados muestran que el modelo realiza mejores pronósticos cuando se supone que la serie es estacionaria. Se propone que esta situación se debe al intervalo en el que se determina el coeficiente ϕ_1 de la parte autorregresiva. En 8.22 se observa que, dentro de dos desviaciones estándar, una estimación de ϕ_1 corresponde al caso estacionario mientras que otra parte corresponde al caso no estacionario.

Suponer que el proceso $\beta(t)$ es un proceso estacionario permitió obtener una estimación de los casos acumulados de hasta 75 días dentro del 5 % de error, sin embargo aplicar el modelo ARIMA en series no estacionarias puede provocar una sobreestimación del orden de parámetros como en el caso de la tercera ola donde $q = 13$, frente al correspondiente $p = 2$, lo que afecta la interpretabilidad del modelo. Un modelo con menos parámetros sería preferible pues puede ser más fácil de interpretar. En este proyecto se ha conseguido precisión a expensas de interpretabilidad.

Además, aplicar modelos ARIMA a series no estacionarias puede disminuir la confiabilidad del pronóstico por lo que podría ser útil comparar con un caso estacionario y observar si se alcanza un modelo de orden p, q similar y si los pronósticos son también similares, para esto se clasifican las series $\beta(t)$ de cada departamento en estacionarias y no estacionarias mediante el test ADF. Alrededor de la tercera parte queda determinada como no estacionaria, correspondiendo la serie del departamento de Guatemala a esta categoría.

En el Cuadro 8.1 se muestran los 7 departamentos para los que el test ADF indica no estacionariedad. De 22 departamentos, 15 de ellos presentan a β como un proceso estacionario.

Número	Departamento
0	Guatemala
7	Jutiapa
10	Sololá
12	Suchitepéquez
14	Baja Verapaz
15	Retalhuleu
18	Jalapa

Tabla 8.1. Departamentos cuya serie $\beta(t)$ es no estacionaria en el intervalo $[0, 583]$.

Con el fin de verificar cómo queda el modelo ARIMA en el caso estacionario, se toma como ejemplo el departamento de Santa Rosa.

```
In [74]: NOMBREDEPTO[6]
Out[74]: 'SANTA ROSA'

In [75]: adfuller_test(Beta[6][:583])
         ADF Test Statistic : -4.254203940525635
         p-value : 0.00053270889690627 96
         #Lags Used : 15
         Number of Observations : 567
         strong evidence against the null hypothesis(Ho), reject the null hypothesis. Data is stationary
Out[75]: (-4.254203940525635,
          0.0005327088969062796,
          15,
          567,
          {'1%': -3.441935806025943,
           '5%': -2.8666509204896093,
           '10%': -2.5694919649816947},
          -4854.224929282059)
```

Figura 8.1. Santa Rosa presenta a β como un proceso estacionario, previo a la tercera ola.

La función de autocorrelación parcial indica $p = 1$ ó $p = 2$ y $q \in [0, 9]$. Considerando $p = 1$, según la Figura 8.2 el menor ECM para los primeros diez días ocurre cuando $q = 1, 8$ ó 9. Utilizando ARIMA(1,0,8) se obtienen los parámetros mostrados en la Figura 8.24, donde se puede ver que la estimación de ϕ_1 queda dentro del intervalo correspondiente a un proceso estacionario, no existe ambigüedad como en el caso del departamento de Guatemala.

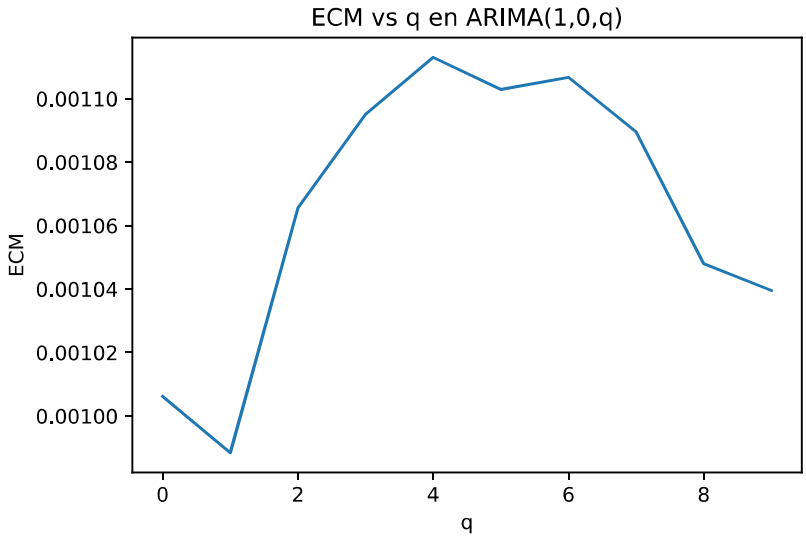


Figura 8.2. Error cuadrático medio durante los primeros diez días de predicción del modelo ARIMA(1,0,q) para diferentes valores de q.

Note que en las figuras 8.4 y 8.5 el pronóstico de los acumulados, en el caso estacionario del departamento de Santa Rosa, queda dentro del 5 % hasta durante 80 días, similar a lo estimado para el departamento de Guatemala. También es similar la cantidad de parámetros necesarios en los modelos de ambos departamentos. Estas dos observaciones mantienen la confianza en el modelo determinado para el departamento de Guatemala.

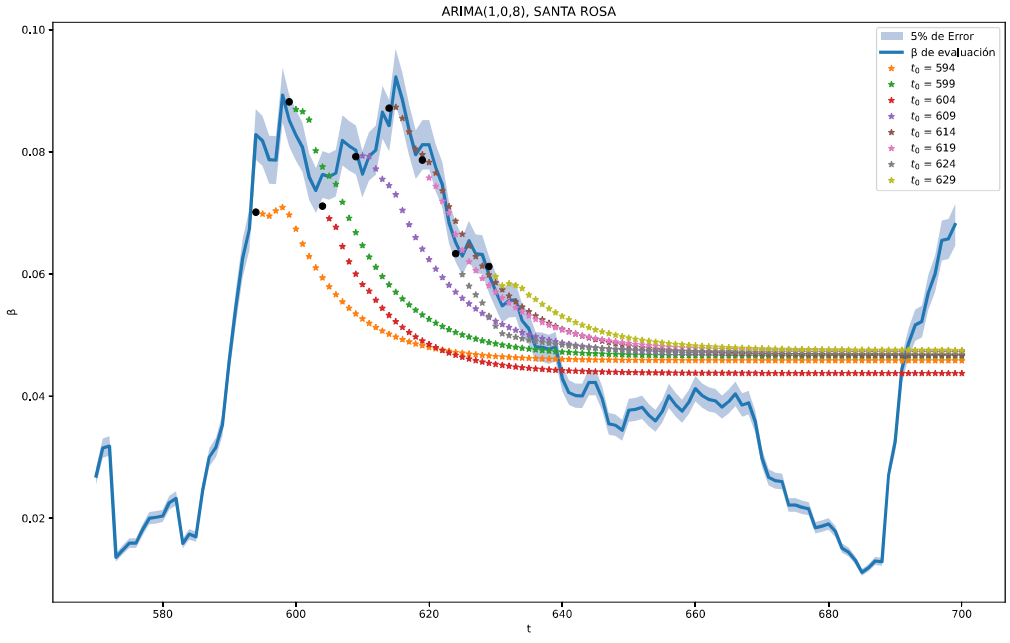


Figura 8.3. Comparación de los pronósticos de la tasa de contagio β para diferentes condiciones iniciales en el intervalo $t \in [594, 630]$, según ARIMA(1,0,8) para el departamento de Santa Rosa.

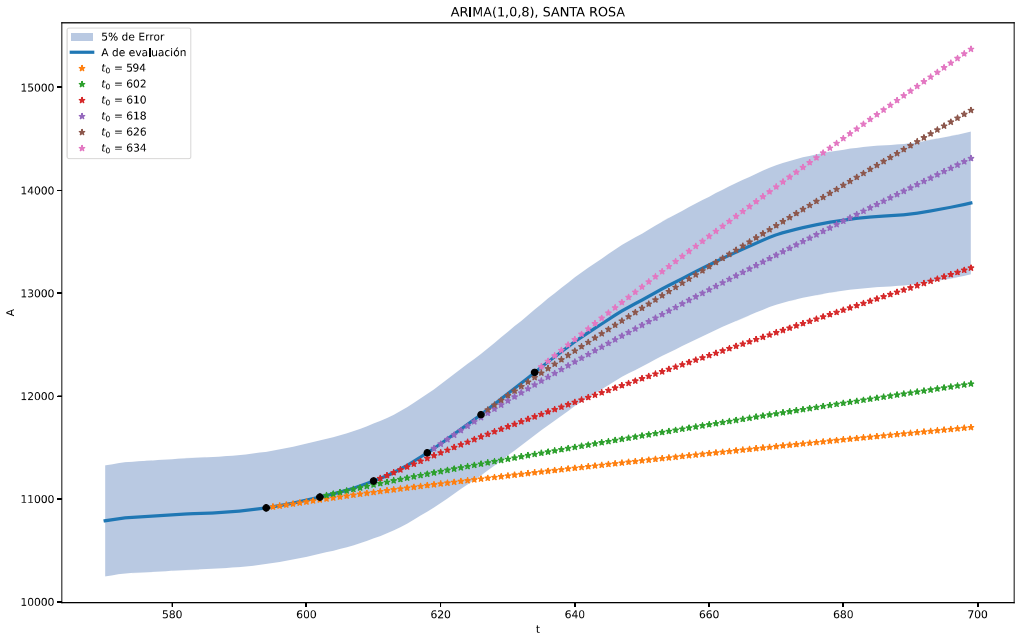


Figura 8.4. Comparación de los pronósticos de los casos acumulados para diferentes condiciones iniciales en el intervalo $t \in [594, 629]$, según ARIMA(1,0,8) para el departamento de Santa Rosa.

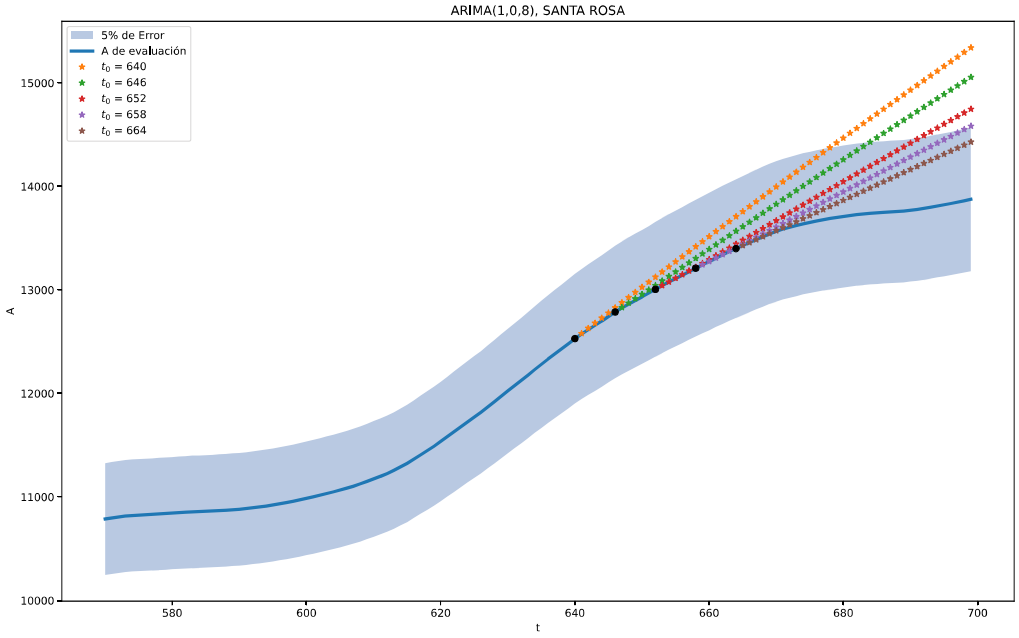


Figura 8.5. Comparación de los pronósticos de los casos acumulados para diferentes condiciones iniciales en el intervalo $t \in [640, 664]$, según ARIMA(1,0,8) para el departamento de Santa Rosa.

CONCLUSIONES

- La estrategia de utilizar redes neuronales para estimar la evolución de β ha dado resultados satisfactorios para compensar la falta de una ecuación diferencial en el sistema (1.1). Aunque la red presentada en este proyecto se limita a pronosticar un solo día a futuro.
- Los modelos ARIMA resultaron particularmente útiles para pronosticar los casos acumulados dentro de un 5 % del valor observado en los siguientes 30 a 80 días aproximadamente, como se muestra en la sección 7. De forma similar pronostica los susceptibles, en los siguientes 30 a 60 días, dentro de un 1 % del valor estimado por el Ministerio de Salud.
- Los modelos obtenidos resultaron prácticos para realizar los pronósticos, sin embargo el orden de la parte de medias móviles dificulta la interpretabilidad de los parámetros.

RECOMENDACIONES

Para complementar esta investigación se proponen dos proyectos: El primero relacionado a ajustar funciones a la forma de las fluctuaciones de la tasa de propagación y el segundo enfocado en evaluar cómo influye la conectividad de una ciudad con sus ciudades vecinas en la dinámica de la epidemia de Covid-19.

Modelado de β como una Función Triangular.

De las figuras en 7.1 vemos que las olas de nuevas infecciones están relacionadas con fluctuaciones en β que sobresalen en la curva:

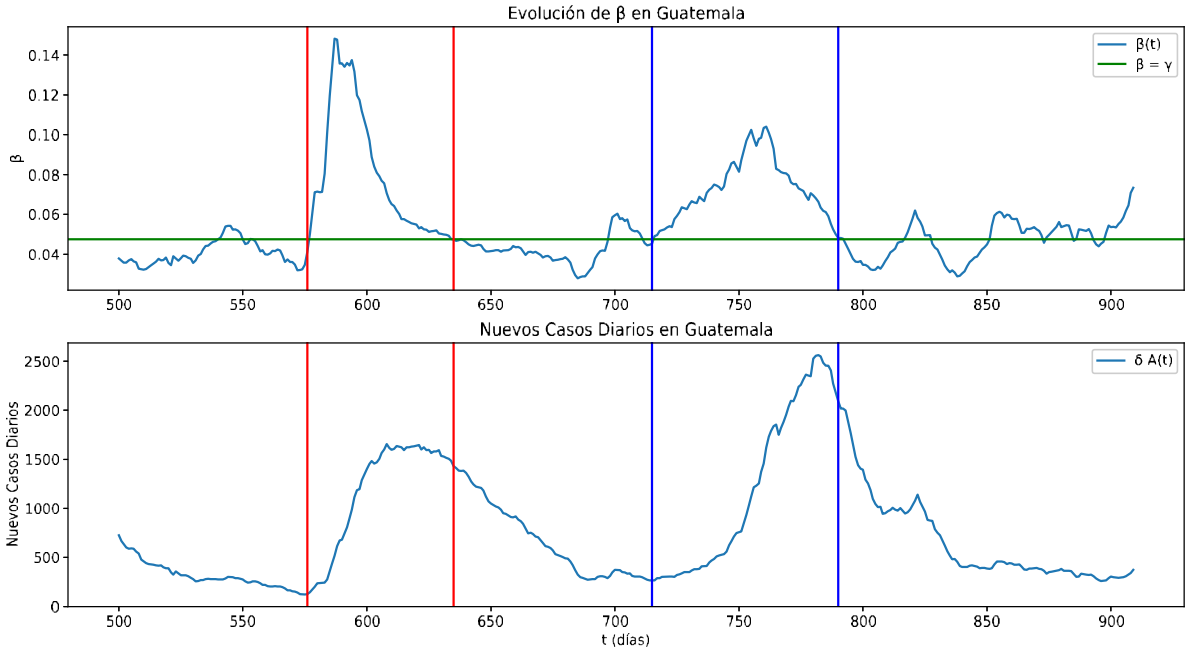


Figura 8.6. Las líneas de cada color delimitan desde que $\beta = \gamma$ en la región de incremento, hasta que $\beta = \gamma$ nuevamente en la región de disminución.

Propuesta: Las fluctuaciones de β durante las olas pueden aproximarse como funciones

triangulares de la forma

$$\beta(t) = \begin{cases} mt + \gamma & \text{si } \gamma \leq \beta \leq \beta_{max} \\ \beta_{max} - mt & \text{si } \beta_{max} > \beta \geq \gamma \end{cases} \tag{8.1}$$

Advierta que es necesario conocer la primera parte de $\beta(t)$, es decir, conocer β_{max} (el próximo proyecto también debería resolver el problema de no saber cuando se encuentra en el máximo pues no se trata de una curva suave y además es estocástica) para determinar la pendiente m de la recta. Así que únicamente podemos pronosticar $\beta(t)$ en la región en la va disminuyendo de nuevo a $\beta = \gamma$.

Ejemplo: De ser adecuada la descripción de $\beta(t)$ mediante la función en 8.1, se estudia qué características deberían esperarse en la evolución de la pandemia. Se realizará mediante el siguiente ejemplo:

Suponga que una ciudad tiene una población de 10^6 habitantes, con las condiciones iniciales $A_0 = 900$, $\delta A_0 = 10$, $S_0 = 10^6 - 900$, $I_0 = 100$, $R_0 = 800$. La fluctuación de β que produce esta ola tiene un $\beta_{max} = 0.2$ y una duración $w = 140$ días.

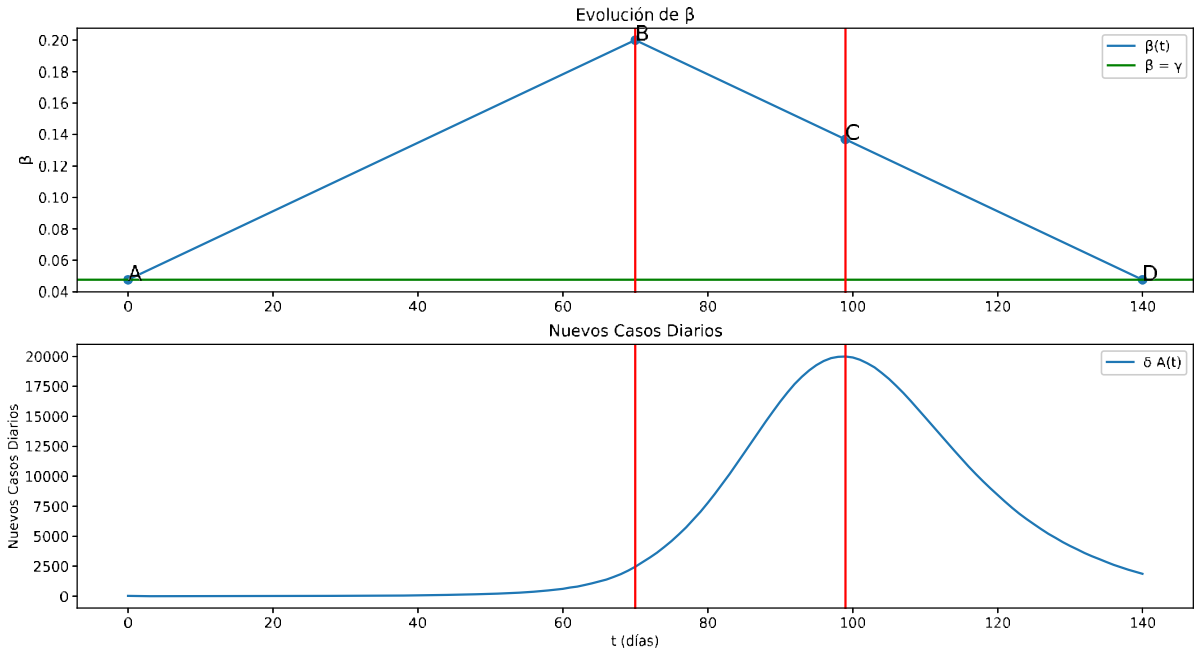


Figura 8.7. Las líneas verticales delimitan la región desde que $\beta = \beta_{max}$, hasta que $\delta A = \delta A_{max}$.

Necesitamos conocer los datos en el intervalo de (A, B) para estimar $\beta(t)$ en el intervalo (B, D). La posición de C depende de las condiciones iniciales de cada caso, en este ejemplo se

encuentra en $(t, \beta) = (99, 0.137)$. Por lo que, podríamos haber sido capaces de determinar que el máximo ocurriría en 30 días aproximadamente.

En la siguiente gráfica de los nuevos reportes diarios para el departamento de Guatemala se indica el valor de β mediante colores para buscar algún patrón. Específicamente queremos estimar la posición de C como en el ejemplo de la Figura 8.7 para conocer con cuánto tiempo de anticipación podríamos pronosticar el máximo de cada ola.

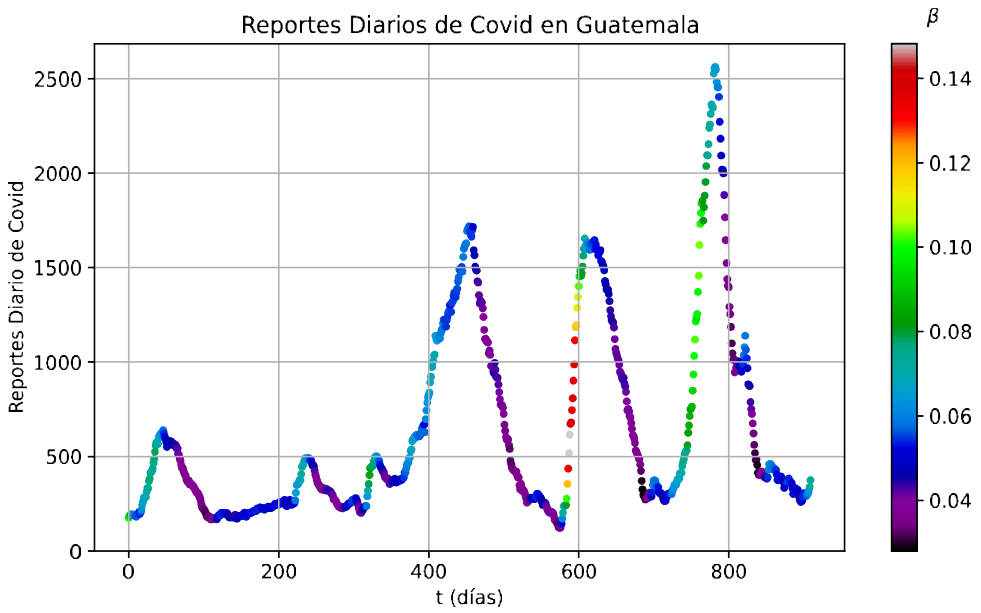


Figura 8.8. β asociado a los nuevos reportes diarios.

Observamos que los reportes diarios empiezan a disminuir cuando $0.04 < \beta < 0.08$. Lo que tiene sentido pues para que la infección se propague es necesario cumplir con $\beta > \gamma$ que en este caso corresponde a $\gamma = 1/21 \approx 0.048$. La siguiente tabla muestra el valor de β cuando se alcanza el máximo de reportes diarios en cada ola.

No. de Ola	δA_{max}	t_{max}	$\beta(t_{max})$
1	639	46	0.058
2	1719	454	0.056
3	1654	608	0.071
4	2562	782	0.066

Tabla 8.2. Máximo de reportes diarios en cada ola y su correspondiente β.

Por lo que en el mejor de los casos[1] se puede determinar el máximo con Δt días de anticipación donde $\Delta t = t(\delta A_{max}) - t(\beta_{max})$ y $t_{max} \equiv t(\delta A_{max})$ es tal que

$\beta(t_{max})$
0.06275 ± 0.006

Tabla 8.3. Promedio de β en el que se espera alcanzar el máximo de reportes diarios.

A continuación se modela, utilizando 8.1, la fluctuación de β que provoca la cuarta ola en la Figura 8.8.

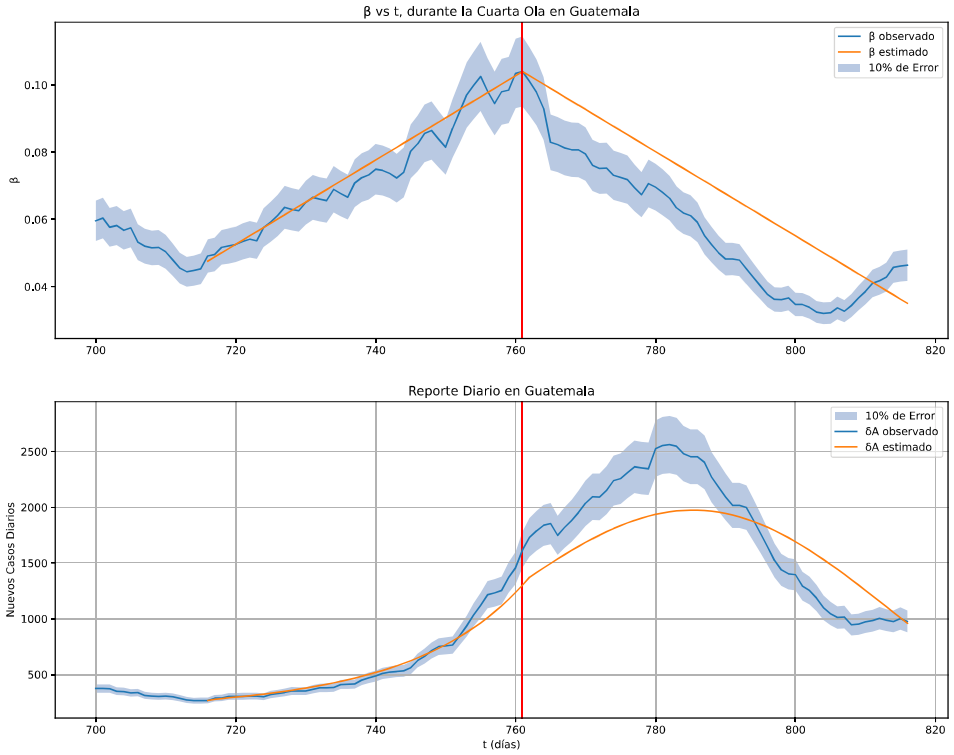


Figura 8.9. Modelado de la fluctuación de β como función triangular y simulación de los nuevos infectados.

[1]El mejor de los casos ocurre cuando posicionamos el máximo de la función triangular en el máximo de la fluctuación observada, Pero hasta que no lleguemos al máximo, no podemos estar seguros de haberlo alcanzado, por la naturaleza estocástica de β.

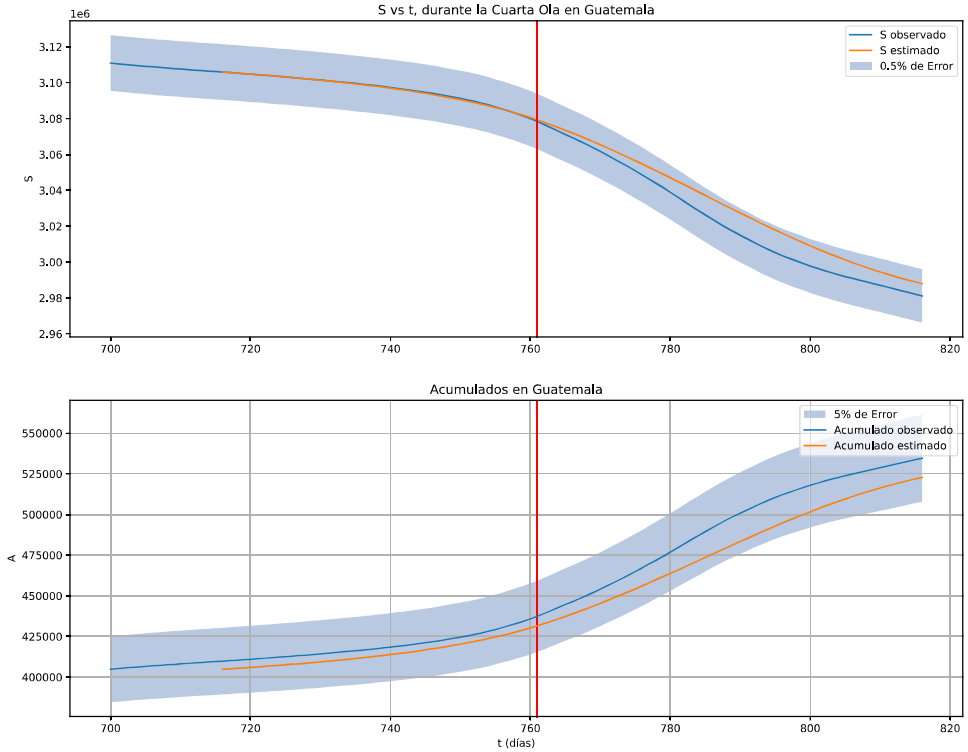


Figura 8.10. Simulación de los casos acumulados y susceptibles relacionados a la Figura 8.9.

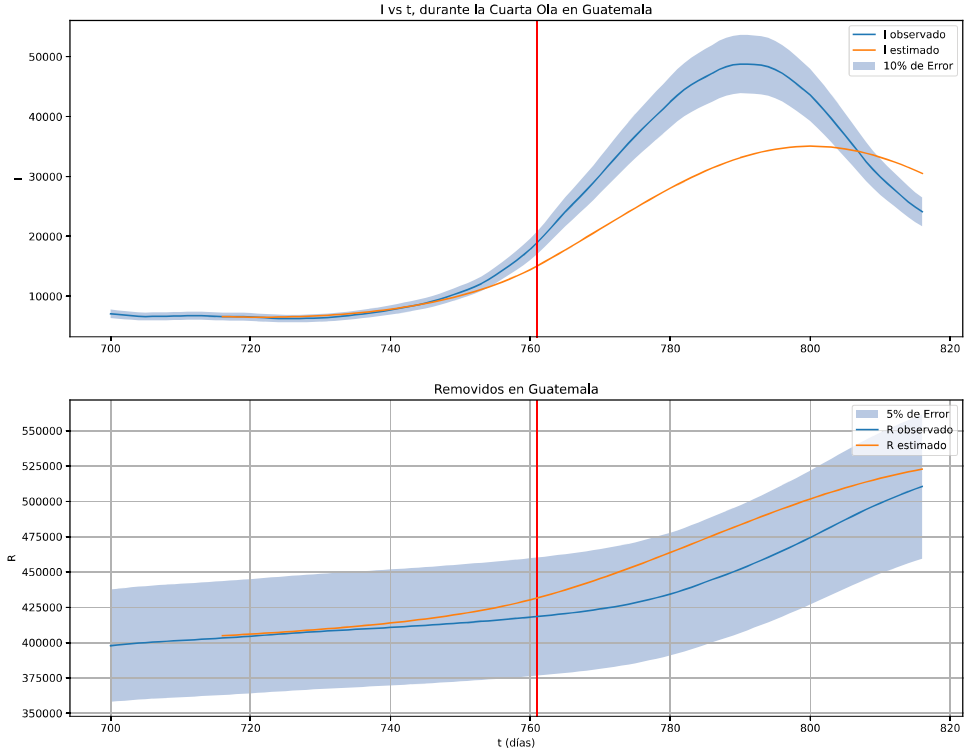
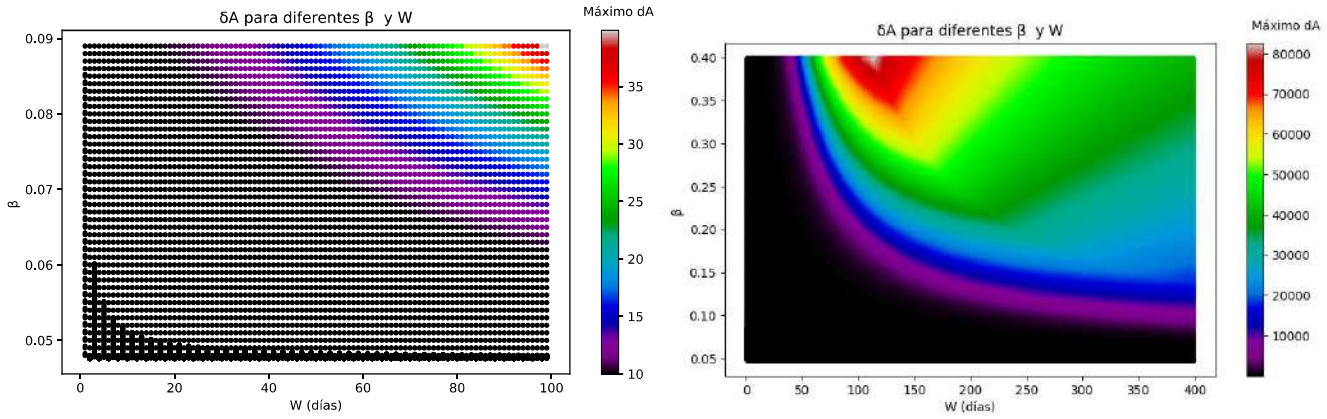


Figura 8.11. Simulación de los casos removidos e infectados relacionados a la Figura 8.9.

Donde se han utilizado las condiciones iniciales correspondientes al día 02 de mayo de 2022: $t_0 = 716$, $\beta(t_0) = \gamma$, población $= 3515775$, $A_0 = 404834$, $\delta A_0 = 267$, $S_0 = 3105946$, $I_0 = 6617$, $R_0 = 403211$. La simulación determina que el máximo de los nuevos infectados ocurre alrededor del día 785 esto corresponde al día 24 de predicciones, mientras que el máximo observado ocurrió en el día 780, el día 19 de predicciones. Por lo que el error en el estimado del tiempo en ocurrir el máximo es del 26 %.

$$E = (24 - 19)/19 \approx 0.26 \tag{8.2}$$

Regresando al ejemplo y partiendo de las mismas condiciones iniciales, estudiaremos qué ocurre con el máximo de la ola cuando variamos el ancho de la función triangular, es decir que para cada β_{max} tendremos diferentes tiempos en alcanzarlo.



(a) **(b)**

Figura 8.12. División de regiones de máximos de nuevos infectados para diferentes valores de β y w.

Note en Figura 8.12 (a) que para cada β existe cierto w debajo del cual la infección se desvanece. Ahora consideremos condiciones iniciales similares a las anteriores excepto que la población es de 10^5 habitantes, por ser conveniente computacionalmente.

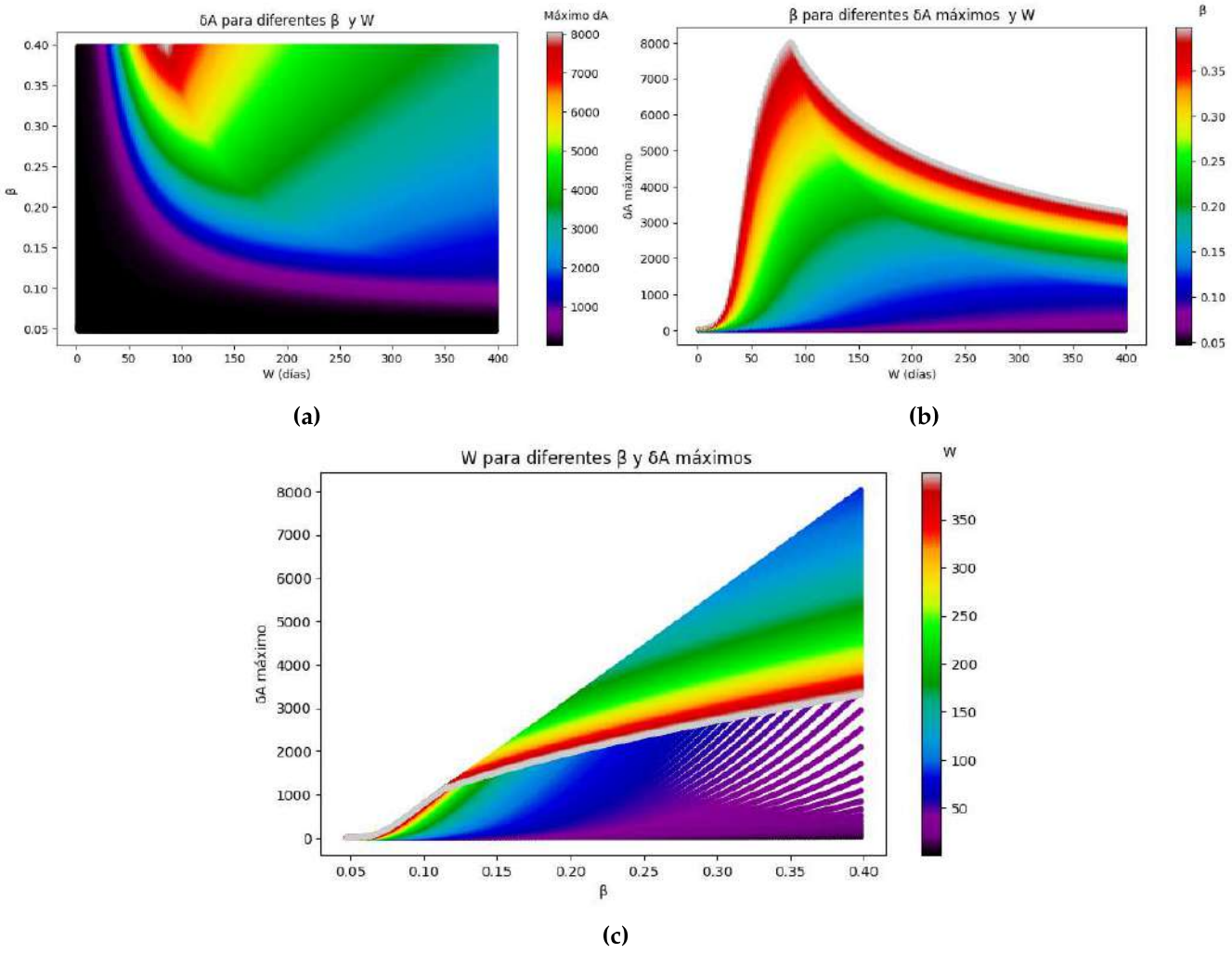


Figura 8.13. Diferentes formas de observar la relación entre las variables β, δA_{max} y w.

A continuación se irá mostrando la evolución de la Figura 8.13 para diferentes tiempos de duración de la fluctuación:

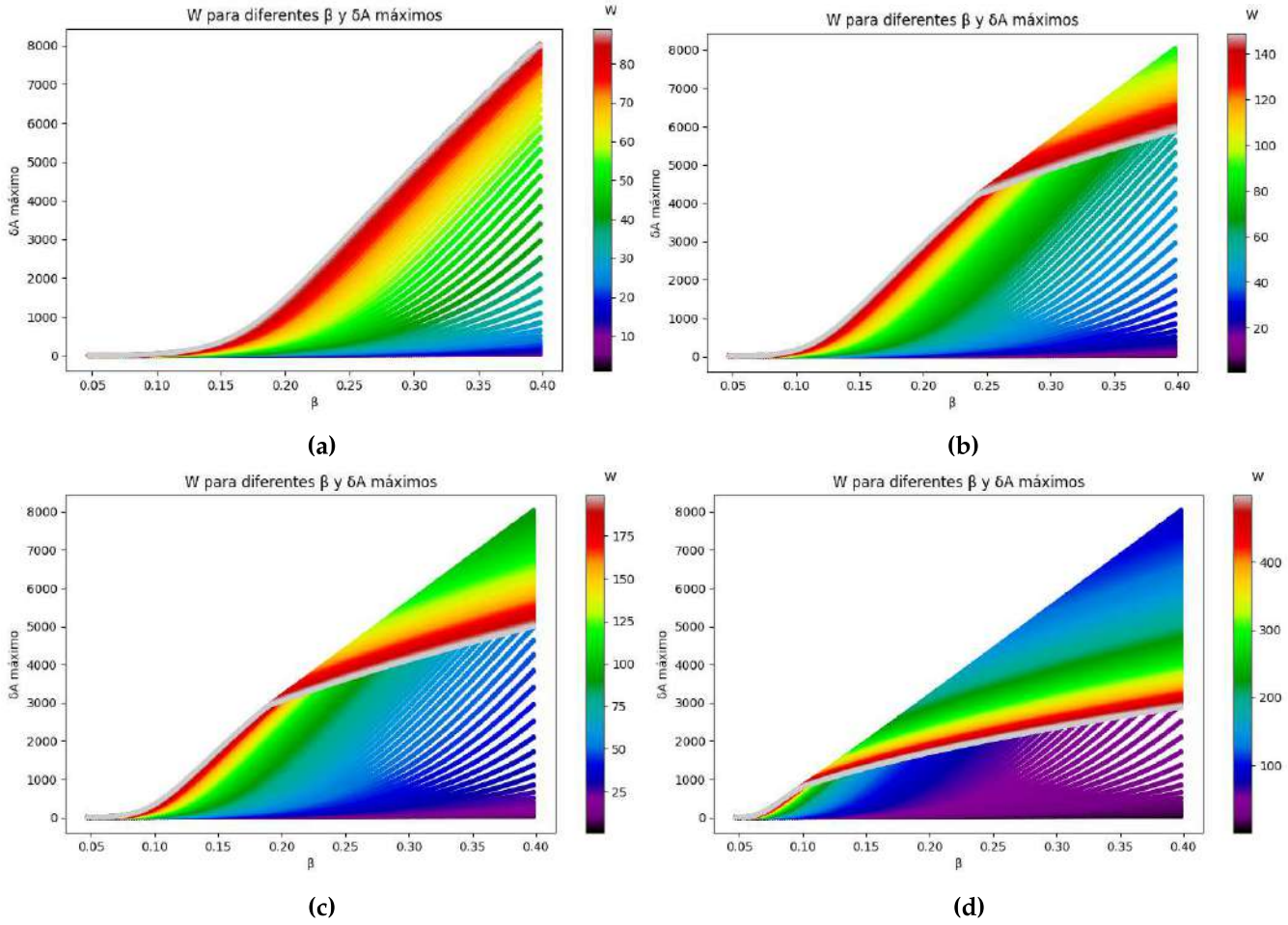


Figura 8.14. Gráficas que resultan de incrementar w para ciertos δA_{max} y β.

La gráfica en la Figura 8.14 parece doblarse sobre sí misma como consecuencia de un límite en los δA_{max} para cada β cuando se hace crecer w indefinidamente.

El proyecto también debería agregar una componente aleatoria a la función triangular, por lo que la ecuación 8.1 quedaría

$$\beta(t) = \begin{cases} mt + \gamma + \delta\beta & \text{si } \gamma \leq \beta \leq \beta_{max} \\ \beta_{max} - mt + \delta\beta & \text{si } \beta_{max} > \beta \geq \gamma \end{cases} \tag{8.3}$$

donde $\delta\beta$ es la componente aleatoria y sigue cierta distribución probabilística que debería ser determinada también en futuras investigaciones.

Modelo de Percolación

En trabajos previos como en [13] y en [6] se ha propuesto a la teoría de percolación como una herramienta para estudiar la influencia de la movilidad entre ciudades sobre la dinámica de la propagación de infecciones. Por lo que para este estudio se propone examinar la relación entre el número máximo de casos nuevos de COVID-19 reportados en cada departamento de Guatemala y la conectividad con sus departamentos vecinos. Con el fin de medir la conectividad se puede utilizar la biblioteca OSMnx de Python para obtener información de la red vial descargada de *OpenStreetMap*. Dicha biblioteca permite identificar las intersecciones de grado 1, que corresponden al número de conexiones que tiene la ciudad con el exterior.

Pueden ser útiles las cantidades *densidad superficial* y *densidad perimetral* de las conexiones que son definidas más adelante. Los resultados iniciales indican que el logaritmo natural del número máximo de nuevos contagios es proporcional al logaritmo natural de la densidad superficial de vías que conectan a cada departamento con sus vecinos.

La librería OSMnx permite elegir la porción de la red de carreteras que llega hasta la frontera de la ciudad. Se considera que los nodos que han quedado conectados a una sola carretera representan las entradas a la ciudad. Esta información se obtiene mediante la función *basic_stats* de OSMnx que permite obtener información de la distribución de las carreteras en la red. La cantidad de nodos mencionada anteriormente se emplea para proponer las siguientes medidas de conectividad de la red, de la siguiente forma:

$$C_s = C/S, \tag{8.4}$$

$$C_p = C/P, \tag{8.5}$$

donde C_s y C_p son la densidad superficial de conexiones y la densidad perimetral de conexiones, respectivamente. S y P son el área y el perímetro del departamento, mientras que C es la cantidad de conexiones con sus departamentos vecinos.

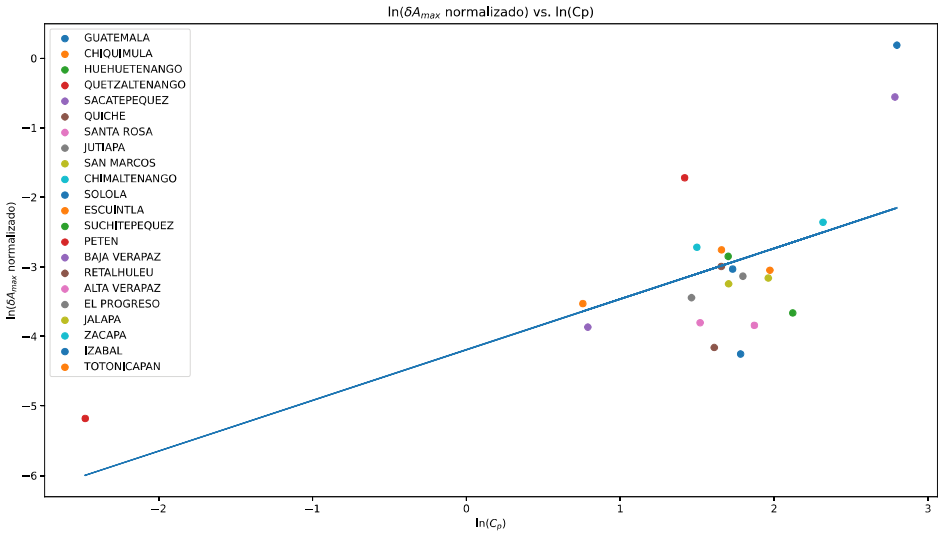


Figura 8.15. El ajuste lineal de $ln(\delta A_{max}$ normalizado$)$ vs. $ln(C_p)$ presenta un coeficiente de determinación $R^2 = 0.64$.

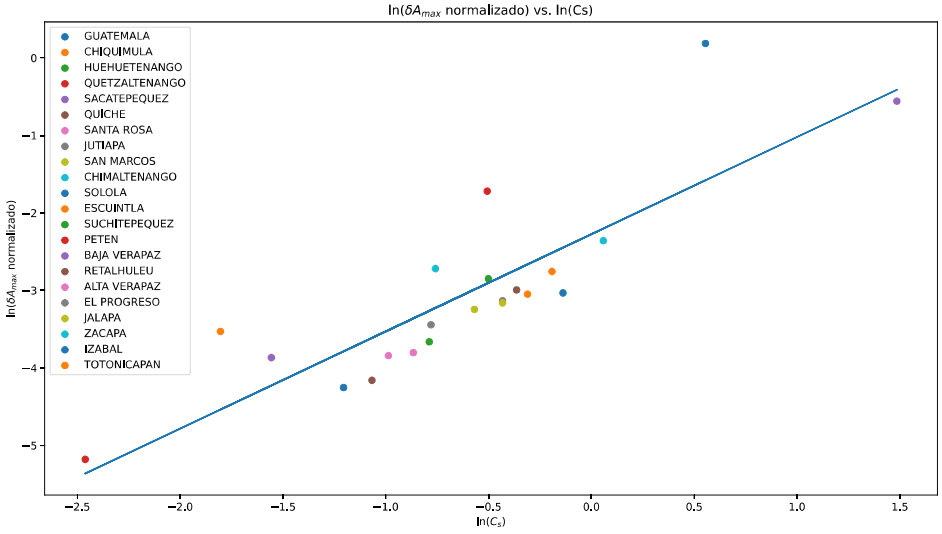


Figura 8.16. El ajuste lineal de $ln(\delta A_{max}$ normalizado$)$ vs. $ln(C_s)$ presenta un coeficiente de determinación $R^2 = 0.85$.

ANEXOS

Implementación en Python

El código completo empleado en cada una de las secciones se encuentra en https://github.com/CarlosSalazarR/Thesis-Scripts. En esta sección se muestra solo alguna parte de lo que se puede encontrar en el enlace anterior.

In [6]:

```
#=============================================================================
#ESTIMACIONES DE BETA y OTRAS VARIABLES DEL MODELO SIR
#=============================================================================
#INICIALIZACION DE VARIABLES
gamma = 1./21.
A = acumuladosReportados #TOTAL DE INDIVIDUOS QUE SE HAN INFECTADO.
dA = POSITIVOSDEPTO #CASOS NUEVOS REPORTADOS CADA DIA POR CADA DEPARTAMENTO
R = np.zeros((len(A),len(A[0])))
I = A - R
S = np.zeros((len(A),len(A[0])))
Beta = np.zeros((len(A),len(A[0])))
Rt = np.zeros((len(A),len(A[0])))
R0 = np.zeros((len(A),len(A[0])))

inicio = 100

for i in range(len(A)):
    for j in range(len(A[0])):
        S[i][j] = Poblacion[i] - A[i][j] #ECUACION (4.37)

for i in range(len(A)):
    N = Poblacion[i]
    for j in range(len(A[0])-1):
        if j>20: #porque empezamos desde j=0.
            R[i][j] = dA[i][j-21] + R[i][j-1] #ECUACION (4.38)
        else:
            R[i][j] = 0

        if A[i][j] - R[i][j] <= 0:
            I[i][j] = I[i][j-1]
        else:
            I[i][j] = A[i][j] - R[i][j] #ECUACION (4.39)
        S[i][j] = N - I[i][j] - R[i][j]
        Beta[i][j] = dA[i][j+1]/(S[i][j]*I[i][j])*N ##ECUACION (5.1)
        #mas adelante quitar el ultimo de beta porque es cero.
        Rt[i][j] = Beta[i][j]/gamma * S[i][j]/N
        R0[i][j] = Beta[i][j]/gamma

        if A[i][j] <= 0:
            A[i][j] = A[i][j-1]
```

Figura 8.17. Implementación de la sección 1.2

```
In [7]: #-------------------------------------------------------------------------
            #RED NEURONAL y OTRAS FUNCIONES
        #-------------------------------------------------------------------------


        #MODIFICA EL ARREGLO DE DATOS PARA QUE TENGA LA FORMA DE LA EXPRESIÖN (4.41)
        def convert2matrix(data_arr, look_back):
            X, Y =[], []
            for i in range(len(data_arr)-look_back):
                d=i+look_back
                X.append(data_arr[i:d,0])
                Y.append(data_arr[d,0])
            return np.array(X), np.array(Y)


        #trainX, trainY = convert2matrix(train, look_back)
        #testX, testY = convert2matrix(test, look_back)
        ###############################################################################
```

Figura 8.18. Implementación de la expresión 2.3.

```
from keras.models import Sequential
from keras.layers import Dense
from keras.callbacks import EarlyStopping
import seaborn as sns

#CREACION DE LA RED DE LA FIGURA (4.7)
def model_dnn(look_back):
    # Dense Neural Network significa que estan conectadas todas las neuronas de una capa
    #con todas las neuronas de las capas vecinas.
    model=Sequential()
    model.add(Dense(units=64, input_dim=look_back, activation='relu'))
    model.add(Dense(8, activation='relu'))
    model.add(Dense(1))
    model.compile(loss='mean_squared_error',  optimizer='adam',metrics = ['mse', 'mae'])
    return model
```

Figura 8.19. Creación de la red de la figura 4.1.

```
#GRAFICA LA FUNCION DE COSTO DE LA EXPRESION (4.42)
def model_loss(history):
    plt.figure(figsize=(8,4))
    plt.plot(history.history['loss'], label='Train Loss')
    plt.plot(history.history['val_loss'], label='Test Loss')
    plt.title('model loss - '+NOMBREDEPTO[n])
    plt.ylabel('loss')
    plt.xlabel('epochs')
    plt.legend(loc='upper right')
    plt.savefig('GraficasProyDNN/'+NOMBREDEPTO[n]+'Loss.pdf', bbox_inches = 'tight')
    #plt.show();
```

Figura 8.20. Cálculo de la función de costo, sección 2.2.

```
In [9]: #GUARDANDO LAS PREDICCIONES EN ARCHIVOS CSV: DEPARTAMENTO_NUMERO DE EJECUCION
L = len(A[0])-1
for n in range(15,22):
#for n in range(len(departamentos)):
    df1 = pd.DataFrame({'Fecha':fechas[inicio:L],'Beta':Beta[n][inicio:L], 'I':I[n][inicio:L],
                        'S':S[n][inicio:L], 'R':R[n][inicio:L]})
    df1 = df1.set_index('Fecha')
    train_size = int(len(Rt[n][inicio:])-70)
    #Data frame de entrenamiento y prueba
    trainB, testB = df1.values[0:train_size,0:1],df1.values[train_size:len(df1.values),0:1]

    #separa cada data frame en x y y
    trainXB, trainYB = convert2matrix(trainB, look_back)
    testXB, testYB = convert2matrix(testB, look_back)
    for m in range(100):
        model = model_dnn(look_back) #modelo de red densa
        #ajustando modelo a la data
        history = model.fit(trainXB,trainYB,epochs=100,batch_size=30,
                            verbose=0,validation_data=(testXB,testYB),
                            callbacks=[EarlyStopping(monitor='val_loss',patience=10)],shuffle=False)
        train_score = model.evaluate(trainXB, trainYB, verbose=0)
        test_score = model.evaluate(testXB, testYB, verbose=0)
        model_loss(history)
        test_predictB = model.predict(testXB)
        #testpredict es lo que usare como coeficientes Beta para las series finitas
        #tambien se guardara en archivo de texto.
        ##prediction_plot(testYB,test_predictB,n)
        data = {"BetaPred": test_predictB.transpose()[0], "Beta Real": testYB}
        data2 = pd.DataFrame.from_dict(data)
        data2.to_csv(NOMBREDEPTO[n] + 'BetaPrediccion'+str(m)+'.csv',
                     header=["Beta Prediccion", "Beta Real" ], index=False)
```

Figura 8.21. Puesta en acción de la red neuronal. Las últimas líneas guardan las predicciones en un archivo de extensión *csv*.

Cada archivo generado contiene una tabla con las proyecciones de 40 días, como la siguiente

Predicción de β $[dia^{-1}]$	**β observada** $[dia^{-1}]$
0.0709679	0.0624266
0.0661554	0.0612805
.	.
.	.
.	.
0.063060	0.06012599

Tabla 8.4. Proyecciones de β para una de las corridas de la red.

Por cada departamento se generaron 100 archivos de la forma del Cuadro 8.4. De esta manera se tienen 100 predicciones para cada día y se reporta el promedio de ellas.

Parámetros de los Modelos ARIMA

La tercera ola fue modelada mediante ARIMA(2,0,13), utilizando el código:

```
model=ARIMA(Beta[0][:585],order=(2,0,13))
model_fit=model.fit()
model_fit.summary()
```

Obteniendo los parámetros siguientes:

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0408	0.004	11.499	0.000	0.034	0.048
ar.L1	0.9497	0.170	5.580	0.000	0.616	1.283
ar.L2	-0.0395	0.163	-0.242	0.809	-0.359	0.280
ma.L1	0.4206	0.177	2.382	0.017	0.075	0.767
ma.L2	0.2094	0.082	2.566	0.010	0.049	0.369
ma.L3	0.1618	0.046	3.500	0.000	0.071	0.252
ma.L4	0.1803	0.038	4.719	0.000	0.105	0.255
ma.L5	0.2126	0.053	4.030	0.000	0.109	0.316
ma.L6	0.3726	0.046	8.150	0.000	0.283	0.462
ma.L7	0.5194	0.060	8.715	0.000	0.403	0.636
ma.L8	0.4614	0.088	5.267	0.000	0.290	0.633
ma.L9	0.3138	0.084	3.743	0.000	0.150	0.478
ma.L10	0.2786	0.064	4.352	0.000	0.153	0.404
ma.L11	0.1904	0.066	2.866	0.004	0.060	0.321
ma.L12	0.1560	0.055	2.861	0.004	0.049	0.263
ma.L13	0.2617	0.047	5.523	0.000	0.169	0.355

Figura 8.22. Parámetros de ARIMA(2,0,13) para modelar β en la tercera ola.

De forma similar para la cuarta ola utilizando ARIMA(3,0,14) se obtiene

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0455	0.003	13.040	0.000	0.039	0.052
ar.L1	1.3740	0.228	6.016	0.000	0.926	1.822
ar.L2	-0.2282	0.157	-1.455	0.146	-0.536	0.079
ar.L3	-0.1693	0.124	-1.365	0.172	-0.412	0.074
ma.L1	0.1306	0.224	0.582	0.561	-0.309	0.570
ma.L2	0.0058	0.240	0.024	0.981	-0.465	0.476
ma.L3	0.0616	0.174	0.354	0.723	-0.279	0.402
ma.L4	-0.0158	0.148	-0.107	0.915	-0.306	0.274
ma.L5	-0.0695	0.108	-0.641	0.522	-0.282	0.143
ma.L6	0.1716	0.073	2.352	0.019	0.029	0.315
ma.L7	0.3202	0.096	3.351	0.001	0.133	0.507
ma.L8	0.1492	0.150	0.993	0.321	-0.145	0.444
ma.L9	0.0461	0.168	0.275	0.783	-0.282	0.375
ma.L10	0.0241	0.146	0.164	0.870	-0.263	0.311
ma.L11	-0.0251	0.115	-0.219	0.826	-0.250	0.200
ma.L12	-0.0808	0.088	-0.916	0.360	-0.254	0.092
ma.L13	0.1691	0.060	2.810	0.005	0.051	0.287
ma.L14	-0.3714	0.084	-4.446	0.000	-0.535	-0.208

Figura 8.23. Parámetros de ARIMA(3,0,14) para modelar β en la cuarta ola.

Para el departamento de Santa Rosa para la tercera ola se tienen los siguientes parámetros con un modelo ARIMA(1,0,9). Este caso se trata de un proceso estacionario:

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0439	0.004	11.409	0.000	0.036	0.051
ar.L1	0.8723	0.020	42.610	0.000	0.832	0.912
ma.L1	0.3278	0.043	7.605	0.000	0.243	0.412
ma.L2	0.2275	0.051	4.463	0.000	0.128	0.327
ma.L3	0.2612	0.052	5.028	0.000	0.159	0.363
ma.L4	0.1736	0.057	3.026	0.002	0.061	0.286
ma.L5	0.1521	0.055	2.764	0.006	0.044	0.260
ma.L6	0.1865	0.049	3.818	0.000	0.091	0.282
ma.L7	0.2016	0.052	3.900	0.000	0.100	0.303
ma.L8	0.1834	0.053	3.440	0.001	0.079	0.288
ma.L9	0.1171	0.048	2.443	0.015	0.023	0.211

Figura 8.24. Parámetros de ARIMA(1,0,8) para modelar β en la tercera ola, correspondiente al departamento de Santa Rosa.

BIBLIOGRAFÍA

[1] Covid board. `http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm`. Accessed: 2022-08-15.

[2] F. Almomani. Forecast of the outbreak of covid-19 using artificial neural network: Case study qatar, spain, and italy. *Results in Physics*, 27, 2021.

[3] Daniel Bernoulli. An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it. *Reviews in Medical Virology*, 2004.

[4] Zhilan Feng Fred Brauer, Carlos Castillo-Chavez. *Mathematical Models in Epidemiology*. Springer.

[5] K. Gabriel. Global convergence of covid-19 basic reproduction number and estimation from early-time sir dynamics. *PLoS One*, 15(9), 2020.

[6] D. Herrera. Estrategias de movilidad basadas en la teorıa de percolación para evitar la diseminación de enfermedades: Covid-19. *Revista Mexicana De Fisica*, 68, 2022.

[7] IGSS. *Aislamiento y Manejo de Casos Leves/Asintomáticos que Cursan con COVID-19 en los Hoteles durante la Pandemia*. IGSS, 2020.

[8] Ezequiel Jiménez. *Análisis de Series Temporales, Modelos Arima*. Paraninfo, 1985.

[9] A. N. Kolmogorov. *Foundations of the theory of probability*. Chelsea Publishing Co., 1950.

[10] Shiva Moein. Inefficiency of sir models in forecasting covid-19 epidemic: a case study of isfahan. *Scientific Reports*, 11:1–9, 2021.

[11] N.Al-Salti. Sir model with time-varying contact rate. *International Journal of Biomathematics*, 14, 2021.

[12] Michael A. Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015.

[13] J. Raimbault. Multi-dimensional urban network percolation. *Journal of Interdisciplinary Methodologies and Issues in Science*, 5, 2019.

[14] Kowalik J. M. Shay D. K. Ridenhour, B. Unraveling R0: Considerations for public health applications. *American Journal of Public Health, 108(Suppl 6), S445–S454.*, 108(02):123–456, February 2018.

[15] R. Ross. Discussion: The kermack-mckendrick epidemic treshold theorem. *Bulletin of mathematical biology.*, 53:3–32, 1991.

[16] Petar Todorovic. *An Introduction to Stochastic Processes and Their Applications*. Springer Series in Statistics, 1992.