



Predicción de series temporales de incidencia de tuberculosis pulmonar a partir de algoritmos de inteligencia computacional

Iván Camilo Sánchez Vega

Universidad Antonio Nariño
Facultad de Ingeniería Mecánica, Electrónica y Biomédica - FIMEB
Programa de Maestría en Instrumentación y Automatización
Bogotá D.C., Colombia
2023

Predicción de series temporales de incidencia de tuberculosis pulmonar a partir de algoritmos de inteligencia computacional

Iván Camilo Sánchez Vega

Trabajo Integral de Grado presentado como requisito para optar al título de:
Magister en Instrumentación y Automatización

Director:

Ing. Andrés Leonardo Jutinico Alarcón PhD

Codirector:

Ing. Álvaro David Orjuela Cañón PhD

Línea de Investigación:

Inteligencia Computacional

Grupo de investigación en Bioinstrumentación, Control, Inteligencia Computacional y Energías Alternativas (GIBIO)

Universidad Antonio Nariño

Facultad de Ingeniería Mecánica, Electrónica y Biomédica - FIMEB

Programa de Maestría en Instrumentación y Automatización

Bogotá D.C., Colombia

2023

Nota de Aceptación

Firma del presidente del Jurado

Firma del Jurado 1

Firma del Jurado 2

Bogotá D.C. _____

A mi papá Fabio Sánchez y mi mamá Maricela Vega que con tanto esfuerzo y dedicación me enseñaron buenas costumbres desde pequeño, a ellos quienes me vieron caer y ayudaron a levantarme; les dedico todos mis triunfos y agradezco todo lo que hacen por mí, gracias por brindarme tanto amor y comprensión. A toda mi familia y a mis hermanos Andrés, Fabio Jr., Linda, Aurum y Elroi que son pieza clave en mi vida, ustedes merecen los frutos de lo que hemos sembrado juntos; gracias por aportar madurez, por su apoyo incondicional y por motivarme cada día a ser mejor persona. A mi amada esposa Claudia Lorena quien continuamente me brinda su amor y comprensión sin importar la circunstancia, gracias por tu apoyo incondicional, mereces mucho más de lo que algún día pueda llegar a ofrecerte

Iván Camilo Sánchez Vega

Agradecimientos

Y él me enseñaba, y me decía: “Retenga tú corazón mis razones, guarda mis mandamientos, y vivirás. Adquiere sabiduría, adquiere inteligencia; no te olvides ni te apartes de las razones de mi boca”. **Proverbios 4:4-5**

Porque Jehová da la sabiduría, y de su boca viene el conocimiento y la inteligencia. **Proverbios 2:6**

Al Ingeniero Andrés Leonardo Jutinico quién siempre se mostró atento y con la mejor disposición para tomar lo mejor de mis capacidades y materializarlas en un trabajo impecable.

A todo el cuerpo colegiado de la Facultad de Ingeniería Mecánica, Electrónica y Biomédica (FIMEB) en la Universidad Antonio Nariño. A cada uno de los profesores que me dictó clases, agradezco los conocimientos compartidos; más aún por la paciencia y dedicación para explicar a detalle cada tema y convertir esos momentos en espacios de enriquecimiento intelectual.

Resumen

En concordancia con el plan estratégico del Ministerio de Salud de Colombia: “Hacia el fin de la tuberculosis”, se concibe esta iniciativa que busca estrategias participativas que cooperen con la predicción de casos nuevos de tuberculosis pulmonar. Entre las limitaciones que afronta Colombia en el sector de la salud se encuentran las causadas por la escasez de modelos computacionales que apoyen en la predicción de enfermedades. Un paso significativo será la implementación de algoritmos de inteligencia computacional que faciliten a los profesionales de la salud, las herramientas para el pronóstico temprano de la Tuberculosis. En esta investigación se diseñaron e implementaron algoritmos basados en redes neuronales artificiales, el filtro de Kalman y modelos autorregresivos, para realizar la predicción de los posibles nuevos casos de Tuberculosis en los departamentos de Colombia. Dichos algoritmos han sido entrenados mediante los registros de las bases de datos de SIVIGILA, los cuales contienen casos confirmados de Tuberculosis en un periodo de más de 10 años. Se observó, que los algoritmos implementados, arrojaron resultados aceptables en la predicción de casos de TB. Para medir el porcentaje de eficiencia se implementaron los índices de error MAE, RMSE y TRS, por ejemplo para el índice MAE en el proceso de validación con la Red Neuronal Artificial en los departamentos de Valle del Cauca, Antioquia y Santander se obtuvo 4.755, 6.791 y 3.204 respectivamente. Los algoritmos que presentaron mejor resultado tomando como base los índices MAE y RMSE fueron el ANN y el RKF. Para el caso del índice TRS, se obtuvo un mejor resultado con los algoritmos N4SID y ARIMA. Los índices MAE y RMSE nos permitieron conocer que tan dispersos estaban los datos obtenidos con los algoritmos en comparación a los datos de la serie real. Por su parte el TRS nos indicó que tan distantes se encontraban los datos obtenidos con relación a los picos de la señal original. Esto nos permitió evaluar integralmente el rendimiento de los algoritmos. Este estudio se desarrolló en el marco del proyecto de investigación, “Generación de modelos alternativos basados en inteligencia computacional para tamización y diagnóstico de Tuberculosis pulmonar”, el cual fue financiado por el Ministerio de Ciencia Tecnología e innovación (Minciencias) y ejecutado por la Universidad Antonio Nariño.

Palabras Claves: Tuberculosis pulmonar (PTB), estrategias, modelos, predicción, ANN, KF, ARIMA, algoritmos computacionales.

Abstract

In accordance with the strategic plan of the Ministry of Health of Colombia: "Towards the end of tuberculosis", this initiative is designed to seek participatory strategies that cooperate with the prediction of new cases of pulmonary tuberculosis. Among the limitations faced by Colombia in the health sector are those caused by the scarcity of computational models that support disease prediction. A significant step will be the implementation of computational intelligence algorithms that will provide health professionals with tools for the early prognosis of tuberculosis. In this research, algorithms based on artificial neural networks, the Kalman filter and autoregressive models were designed and implemented to predict possible new cases of Tuberculosis in the departments of Colombia. It became evident that the algorithms implemented yielded acceptable results in the prediction of TB cases. To measure the efficiency percentage, the MAE, RMSE, and TRS error indicators were implemented. For example, for the MAE index in the validation process with the Artificial Neural Network in the departments of Valle del Cauca, Antioquia, and Santander, 4,755, 6,791, and 3,204, respectively, were obtained. The algorithms that presented the best results based on the MAE and RMSE indexes were ANN and RKF. For the TRS index, a better result was obtained with the N4SID and ARIMA algorithms. The MAE and RMSE indices allowed us to know how dispersed the data obtained with the algorithms were compared to the data of the real series. On the other hand, the TRS indicated how far the data obtained were from the peaks of the original signal. This allowed us to comprehensively evaluate the performance of the algorithms. These algorithms have been trained using the records of the SIVIGILA databases, which contain confirmed cases of Tuberculosis over a period of more than 10 years. This study was developed within the framework of the research project, "Generation of alternative models based on computational intelligence for screening and diagnosis of pulmonary tuberculosis", which was funded by the Ministry of Science, Technology and Innovation (Minciencias) and executed by the Universidad Antonio Nariño (UAN).

Keywords: Pulmonary tuberculosis (PTB), strategies, models, prediction, ANN, KF, ARIMA, computational algorithms.

Tabla de contenidos

Agradecimientos	9
Resumen	11
Abstract	13
Lista de Figuras	23
Lista de Tablas	23
1 Introducción	23
1.1 Planteamiento del problema	26
1.2 Justificación	27
1.3 Objetivos	29
1.3.1 Objetivo general	29
1.3.2 Objetivos específicos	29
1.4 Estado del Arte	30
2 Marco Teórico	33
2.1 Red Neuronal Artificial (RNA)	33
2.2 Modelo Autorregresivo Integrado de Media Móvil (ARIMA)	34
2.3 Modelo autorregresivo N4SID	35
2.4 El Filtro de Kalman (KF)	35
2.5 El Filtro de Kalman Robusto (RKF)	36
3 Desarrollo Metodológico	39
3.1 Generación de la base de datos	39
3.2 Generación de los modelos computacionales para el entrenamiento	40
3.2.1 Modelo RNA	40
3.2.2 Modelo ARIMA	41
3.2.3 Modelo Autorregresivo N4SID	41
3.2.4 Modelo del Filtro de Kalman	42
3.2.5 Modelo del Filtro de Kalman Robusto	42
3.3 Evaluación de la efectividad de los algoritmos implementados	43
3.4 Número máximo de semanas para las cuales se logra una predicción aceptable	45

4	Resultados	47
4.1	Valle del Cauca	48
4.1.1	Resultados de la RNA para Valle del Cauca	48
4.1.2	Resultados del modelo ARIMA para Valle del Cauca	48
4.1.3	Resultados del modelo N4SID para Valle del Cauca	49
4.1.4	Resultados del Filtro de Kalman para Valle del Cauca	49
4.1.5	Resultados del Filtro de Kalman Robusto para Valle del Cauca	50
4.1.6	Resultados de los algoritmos computacionales para Valle del Cauca	51
4.2	Antioquia	53
4.2.1	Resultados de la RNA para Antioquia	53
4.2.2	Resultados del modelo ARIMA para Antioquia	54
4.2.3	Resultados del modelo N4SID para Antioquia	54
4.2.4	Resultados del Filtro de Kalman para Antioquia	55
4.2.5	Resultados del Filtro de Kalman Robusto para Antioquia	56
4.2.6	Resultados de los algoritmos computacionales para Antioquia	56
4.3	Santander	58
4.3.1	Resultados de la RNA para Santander	58
4.3.2	Resultados del modelo ARIMA para Santander	59
4.3.3	Resultados del modelo N4SID para Santander	60
4.3.4	Resultados del Filtro de Kalman para Santander	60
4.3.5	Resultados del Filtro de Kalman Robusto para Santander	61
4.3.6	Resultados de los algoritmos computacionales para Santander	62
4.4	Bogotá	63
4.4.1	Resultados de la RNA para Bogotá	63
4.4.2	Resultados del modelo ARIMA para Bogotá	64
4.4.3	Resultados del modelo N4SID para Bogotá	65
4.4.4	Resultados del Filtro de Kalman para Bogotá	65
4.4.5	Resultados del Filtro de Kalman Robusto para Bogotá	66
4.4.6	Resultados de los algoritmos computacionales para Bogotá	67
4.5	Meta	69
4.5.1	Resultados de la RNA para Meta	69
4.5.2	Resultados del modelo ARIMA para Meta	70
4.5.3	Resultados del modelo N4SID para Meta	71
4.5.4	Resultados del Filtro de Kalman para Meta	71
4.5.5	Resultados del Filtro de Kalman Robusto para Meta	72
4.5.6	Resultados de los algoritmos computacionales para Meta	73
4.6	Análisis de los Resultados	74
5	Conclusiones	81

Bibliografía

Lista de Figuras

3-1	Casos de TB en Amazonas de 2007 a 2021	44
3-2	Mapa de calor de los valores RMSE para Valle del Cauca	45
3-3	Mapa de calor de los valores MAE para Valle del Cauca	46
3-4	Mapa de calor de los valores TRS para Valle del Cauca	46
4-1	Casos TB Pulmonar en Valle del Cauca de 2007 a 2021	47
4-2	Resultados RNA Valle del Cauca	48
4-3	Resultados modelo ARIMA Valle del Cauca	49
4-4	Resultados modelo autorregresivo N4SID Valle del Cauca	50
4-5	Resultados Filtro de Kalman Valle del Cauca	50
4-6	Resultados Filtro de Kalman Robusto Valle del Cauca	51
4-7	Resultados Entrenamiento Algoritmos Computacionales Valle del Cauca	51
4-8	Resultados Validación Algoritmos Computacionales Valle del Cauca	52
4-9	Resultados RNA Antioquia	53
4-10	Resultados modelo ARIMA Antioquia	54
4-11	Resultados modelo autorregresivo N4SID Antioquia	55
4-12	Resultados Filtro de Kalman Antioquia	55
4-13	Resultados Filtro de Kalman Robusto Antioquia	56
4-14	Resultados Entrenamiento Algoritmos Computacionales Antioquia	57
4-15	Resultados Validación Algoritmos Computacionales Antioquia	57
4-16	Resultados RNA Santander	59
4-17	Resultados modelo ARIMA Santander	59
4-18	Resultados modelo autorregresivo N4SID Santander	60
4-19	Resultados Filtro de Kalman Santander	61
4-20	Resultados Filtro de Kalman Robusto Santander	61
4-21	Resultados Entrenamiento Algoritmos Computacionales Santander	62
4-22	Resultados Validación Algoritmos Computacionales Santander	62
4-23	Resultados RNA Bogotá	64
4-24	Resultados modelo ARIMA Bogotá	64
4-25	Resultados modelo autorregresivo N4SID Bogotá	65
4-26	Resultados Filtro de Kalman Bogotá	66
4-27	Resultados Filtro de Kalman Robusto Bogotá	66
4-28	Resultados Filtro de Kalman Robusto Bogotá	67

4-29 Resultados Entrenamiento Algoritmos Computacionales Bogotá	68
4-30 Resultados Validación Algoritmos Computacionales Bogotá	68
4-31 Resultados RNA Meta	70
4-32 Resultados modelo ARIMA Meta	70
4-33 Resultados modelo autorregresivo N4SID Meta	71
4-34 Resultados Filtro de Kalman Meta	72
4-35 Resultados Filtro de Kalman Robusto Meta	72
4-36 Resultados Entrenamiento Algoritmos Computacionales Meta	73
4-37 Resultados Validación Algoritmos Computacionales Meta	73
4-38 Resultados obtenidos con el índice MAE	75
4-39 Resultados obtenidos con el índice RMSE	75
4-40 Resultados obtenidos con el índice TRS	76
4-41 Resultados Globales obtenidos con el índice RMSE en el proceso de validación	78
4-42 Resultados Globales obtenidos con el índice MAE en el proceso de validación	78
4-43 Resultados Globales obtenidos con el índice TRS en el proceso de validación	79

Lista de Tablas

2-1	Lista de Entradas y Salidas (RNA)	34
2-2	Filtro de Kalman Robusto Recursivo	37
3-1	Organización datos de entrada RNA - Casos de TB en Bogotá	40
3-2	Parámetros del Filtro de Kalman Robusto.	43
4-1	Departamentos con mayor índice de TB en Colombia en 2021	47
4-2	Resultados de los algoritmos computacionales en el entrenamiento para el departamento de Valle del Cauca	52
4-3	Resultados de los algoritmos computacionales en la validación para el departamento de Valle del Cauca	53
4-4	Resultados de los algoritmos computacionales en el entrenamiento para el departamento de Antioquia	58
4-5	Resultados de los algoritmos computacionales en la validación para el departamento de Antioquia	58
4-6	Resultados de los algoritmos computacionales en el entrenamiento para el departamento de Santander	63
4-7	Resultados de los algoritmos computacionales en la validación para el departamento de Santander	63
4-8	Resultados de los algoritmos computacionales en el entrenamiento para Bogotá	69
4-9	Resultados de los algoritmos computacionales en la validación para Bogotá .	69
4-10	Resultados de los algoritmos computacionales en el entrenamiento para el departamento de Meta	74
4-11	Resultados de los algoritmos computacionales en la validación para el departamento de Meta	74
4-12	Ubicación de los departamentos en las gráficas de los índices de evaluación .	77

1 Introducción

La tuberculosis (TB) es una enfermedad transmisible y una de las 10 principales causas de muerte en todo el mundo. Además, es considerada la principal causa de muerte por un solo agente contaminante, inclusive por encima del VIH/SIDA. En 2019, cerca de 10 millones de personas desarrollaron tuberculosis y 1,4 millones murieron [1]. La tuberculosis es una enfermedad infecciosa crónica que puede afectar cualquier órgano o tejido y es causada por el *Mycobacterium tuberculosis*. La forma más común de la enfermedad es la pulmonar, por otro lado, cuando la infección afecta órganos diferentes al pulmón, esta se denomina tuberculosis extrapulmonar [2]. Las manifestaciones clínicas de la tuberculosis pueden variar según la edad, el estado inmunológico, la presencia de comorbilidades y los órganos afectados [3]. La principal fuente de incubación de esta enfermedad es el hombre, quien puede presentar o no la infección. En este sentido, es importante resaltar, que se considerará sintomático respiratorio a toda persona que presente tos y expectoración por más de 15 días. La transmisión de la TB se da principalmente a través del aire, debido al esparcimiento de diminutas partículas de saliva o secreción nasal que contienen los bacilos de la TB [4]. El periodo de incubación de la tuberculosis no es preciso y no ha sido claramente establecido, algunos autores [5] indican que depende de varios factores, se estima que el periodo promedio desde el momento que se contagia el paciente hasta la aparición de las primeras lesiones está entre la segunda y la décima semana, sin embargo hay casos donde la enfermedad permanece oculta toda la vida; sin manifestarse externamente [3]. Existen diversos factores que aumentan el riesgo de contraer la enfermedad, algunos están asociados a niveles socioeconómicos como el estado nutricional y aglomeración de individuos o hacinamiento. Un 10 % de los infectados por tuberculosis puede desarrollar la enfermedad, mayormente aquellos pacientes que presentan infección por VIH ya que dicha patología aumenta considerablemente el riesgo. Luego que una persona se infecta de TB existe la probabilidad que se desarrolle la enfermedad incluso varios años después de que se haya producido el contagio. En cada ser humano se puede desarrollar la tuberculosis de una manera diferente, según el Instituto Nacional de Salud, se puede determinar médicamente con base en los siguientes síntomas: fiebre, sudoración nocturna excesiva, pérdida de peso, cansancio, ausencia total de fuerza física u otras manifestaciones [3]. Existen métodos de diagnóstico denominados "pruebas de tamizaje", que se utilizan para determinar quiénes corren el riesgo de desarrollar una determinada enfermedad y quiénes no. Para precisar el estado de salud de un paciente que presenta un cuadro sintomático, es necesario realizar los respectivos exámenes médicos a través de muestras de sangre o a través de muestras de tejidos, donde se determina la presencia o ausencia del *Mycobacterium tuberculosis*

[6]. Si bien es cierto que los medicamentos son utilizados para prevenir y tratar infecciones en los seres humanos, existen casos en los cuales el cuerpo presenta resistencia debido al uso indebido de los antimicrobianos. Los medicamentos con propiedades antimicrobianas, como antibióticos, antivirales, antifúngicos y antiparasitarios, se suelen utilizar para combatir y controlar las infecciones en seres humanos, animales y plantas [7]. En las dos últimas décadas se ha generado un progreso importante en la lucha contra la tuberculosis. Sin embargo, la resistencia a los antimicrobianos por diferentes factores como el mal uso de los tratamientos médicos denigra los progresos alcanzados. En 2016, la Organización Mundial de la Salud (OMS) estimó que había 10,4 millones de casos incidentes de tuberculosis. Ese mismo año surgieron unos 600.000 casos de tuberculosis resistente a la rifampicina. Un factor esencial de los programas mundiales de control, reducción y contención de la enfermedad es la vigilancia de la tuberculosis farmacorresistente, esto genera grandes retos para la contención de la bacteria y supone un desafío para los entes de control [8]. En Colombia en el año 2020 se enfermaron de TB 11.390 personas, de las cuales 10.632 fueron casos nuevos, revelando que en el país la tasa de incidencia de tuberculosis está en 20,88 por cada 100.000 habitantes [9]. En Colombia el ente encargado de velar por el bienestar y la salud de los habitantes es el Instituto Nacional de Salud (INS) el cual tiene como objetivo el desarrollo y la gestión del conocimiento científico en salud y biomedicina. Una de las dependencias del INS es el Sistema Nacional de Vigilancia en Salud Pública (SIVIGILA). La información contenida en las bases de datos del SIVIGILA donde se almacena el seguimiento y control que se realiza a los pacientes que presentan PTB, puede utilizarse para estimar el contagio promedio, planificar el diagnóstico, preparar el tratamiento y supervisar la eficacia de los medicamentos suministrados [3], [10]. Algunos artículos relacionados con la predicción de casos de TB, son los siguientes: Chu-Chang Ku et al. [11], desarrolló un modelo de predicción con relación al impacto del envejecimiento de la población en la incidencia de tuberculosis en Taiwán. El modelo desarrollado indica que para 2035 la tasa de incidencia de TB disminuirá en un 54 %. Sin embargo, hace énfasis en que tendrán mayores tasas de mortalidad por TB para los adultos mayores. Zheng Y et al. [12], realizó el análisis de series temporales bajo la metodología de Box-Jenkins (el cual aplica modelos de ARMA o ARIMA) y Red Neuronal Recurrente (ElmanNN) en Kashgar (China) donde el índice de incidencia de TB en 2017 fue de 465,33 por cada 100.000 habitantes. El cálculo de la efectividad del estudio fue realizado con base en el error porcentual absoluto medio (MAPE) el cual es un indicador de desempeño usado frecuentemente para elaborar pronósticos. En el conjunto de datos de ajuste, el MAPE fue de 0,2858 para el método de Box-Jenkins y 0,1837 para el método de ElmanNN. En el conjunto de datos de pronóstico, el MAPE fue de 0,2029 para el método de Box-Jenkins y 0,2006 para el método de ElmanNN. Los autores concluyen que el modelo autorregresivo es eficaz y puede ser usado para la predicción de incidencia de TB. Liu Qiao et al. [13], realizó un modelo de predicción de casos de TB en el sureste de China en la provincia de Jiangsu, donde la tasa de incidencia de TB en 2015 fue de 79,36 por cada 100.000 habitantes. Estos investigadores utilizaron el Modelo Autorregresivo Integrado de Media Móvil (ARIMA) y

el modelo de red Neuronal de propagación hacia atrás (BPNN), para medir la efectividad del estudio se empleó el error porcentual absoluto medio (MAPE). En el rendimiento de los datos de ajuste, el MAPE fue de 6,0498 para el método de ARIMA y 6,0113 para el método de BPNN. En el rendimiento de los datos de pronóstico, el MAPE fue de 4,6041 para el método de ARIMA y 3,2172 para el método de BPNN. Ambos modelos arrojaron resultados satisfactorios y los autores consideran que estos algoritmos pueden ser empleados como métodos de ajuste óptimo para la predicción de la TB.

1.1. Planteamiento del problema

Dentro de las responsabilidades del sistema de salud pública (SIVIGILA) se encuentran los procesos de observación y análisis objetivo, sistemático y constante de los eventos en salud, bajo estas premisas se sustenta el planeamiento, seguimiento y evaluación de las buenas prácticas que se adoptan en la salud pública. En este sentido, se debe promover el estudio de algoritmos computacionales que aporten significativamente a los entes regulatorios que están encargados del bienestar en materia de salud, para la predicción de la tendencia de PTB. La predicción de las series de TB pulmonar puede contribuir al trabajo que han adelantado las entidades de salud y los entes gubernamentales en torno a la contención en la propagación del *Mycobacterium tuberculosis*. No obstante, en la actualidad no existen herramientas computacionales para realizar la predicción de esta enfermedad en los departamentos de Colombia, por lo cual sería ideal tener un modelo computacional que permitiera a los gobernantes y profesionales de la salud, tomar decisiones con base en el registro médico del paciente y la predicción del número de casos de incidencia de TB pulmonar en cada región. Esto aportaría al control y disminución en los índices de contagio de la enfermedad, lo cual sería un aporte importante al programa “End TB” de la OMS. El Ministerio de Salud y Protección Social ha comenzado a adoptar algunos algoritmos para el diagnóstico de la TB de acuerdo al grupo social en el cual se encuentra caracterizado el paciente, esto con el fin de encarar el diagnóstico de la tuberculosis activa pulmonar [3], [10], [14]. Sin embargo, el desarrollo de estas técnicas computacionales, que influyan en la toma de decisiones para minimizar el efecto de la TB en nuestro país, aún están en fase de investigación. Un ejemplo de esto es el proyecto de investigación que busca generar modelos para tamización y diagnóstico de PTB que actualmente está siendo desarrollado por el grupo de investigación GIBIO-UAN en conjunto con investigadores de la Subred Integrada de Servicios de Salud Centro Oriente y la Universidad del Rosario, con apoyo de Minciencias. En el proyecto de investigación mencionado anteriormente se estudian algoritmos de computación inteligente para la predicción y detección de tuberculosis pulmonar (PTB). Por lo tanto, en este trabajo integral de posgrado (TIP) se propone realizar la predicción de series temporales en la incidencia de tuberculosis pulmonar en los departamentos de Colombia a partir de algoritmos de inteligencia computacional, haciendo uso de la información que se encuentra en las bases de datos del SIVIGILA. Con base en lo anterior surge la siguiente pregunta investigativa, ¿Es factible el uso de modelos computacionales para realizar la predicción de los casos de TB pulmonar en Colombia?

1.2. Justificación

En 2014 la OMS dio a conocer la estrategia: “End TB” la cual tiene como propósito el disminuir los índices de muerte e incidencia de TB en el mundo. Muchos países como por ejemplo Brasil han implementado estrategias alineadas a tal objetivo, entre ellas la investigación e innovación en sistemas para la predicción y contención de la TB [15]. En Colombia, en la página web del SIVIGILA, están publicados los eventos que afectan o pueden afectar la salud de la población colombiana, como por ejemplo los casos de TB que se presentaron en los últimos años. En las bases de datos, está clasificada la información por departamento, municipio, eventos (todas las enfermedades) y casos que se reportan de ciudadanos extranjeros. En el año 2020 en Colombia, se reportaron 11.304 casos de tuberculosis y 267 casos de tuberculosis farmacorresistente. En el año 2021, se han reportado hasta la semana epidemiológica 20, 4.723 casos de tuberculosis y 103 casos de tuberculosis farmacorresistente. Adicionalmente, se tiene información de las enfermedades que padecen los ciudadanos extranjeros que residen en el territorio nacional. El informe de SIVIGILA indica que, en el año 2020, ciudadanos de 7 países presentaron 163 casos de tuberculosis y 1 caso de tuberculosis farmacorresistente. El 88,96 % de los casos de TB de las poblaciones extranjeras es aportado por Venezuela, así como también el 100 % de los casos de TB farmacorresistente. En la semana epidemiológica 38 del año 2021, se reportaron 894 casos de TB en Colombia, aproximadamente el 76,5 % de los casos los aportan los estratos socioeconómicos 1 y 2. Del total de casos reportados, el 50 % de los individuos se encuentran en el régimen de salud subsidiado y el 6 % no se encuentra asegurado. Cabe resaltar que en el artículo 49 de la Constitución Política de Colombia, se menciona que la atención en materia de salud es un servicio público que está a cargo del estado y que se debe garantizar a todas las personas el acceso a la protección y recuperación de la salud, así mismo, existen algunas condiciones en las cuales la atención básica es totalmente gratuita y obligatoria para los habitantes. La predicción de la incidencia de PTB busca que el sistema de salud tenga un tiempo prudente para el desarrollo de acciones que estén orientadas a la contención del virus, tales como campañas de vacunación y decisiones administrativas que estén enfocadas en la minimización de los índices de contagio. En el mundo se han desarrollado diversos modelos para la predicción de TB, mayormente en países como China, que tiene una carga de tuberculosis del 8,4 % global (Aproximadamente 840.000 personas de 10 millones reportados en 2019) [1], [12]. Por ejemplo, Wang K. W. et al. [16], implementaron un sistema híbrido que incorpora el modelo Autorregresivo Integrado de Media Móvil (ARIMA) y Redes Neuronales Autorregresivas no lineales (NAR) para superar las limitaciones que representa el uso individual de las técnicas, los datos que se utilizaron en este estudio fueron recolectados desde el año 2007 hasta el año 2016 (aproximadamente 9 años). Como índice de desempeño del estudio se empleó el error porcentual absoluto medio (MAPE). En el rendimiento de los datos de ajuste y pronóstico, el MAPE fue de 0,0406 para el método híbrido ARIMA-NAR con una reducción porcentual de 41.33 % inferior al modelo ARIMA. Los resultados indican que ambas técnicas son efec-

tivas por separado, sin embargo, el modelo híbrido (ARIMA-NAR) se desempeñó mejor en la predicción de incidencia de TB. En Colombia, el grupo de Bioinstrumentación y Control de la Universidad Antonio Nariño participa activamente en proyectos de investigación donde se hace uso de inteligencia computacional y el procesamiento digital de señales, áreas sobre las cuales se fundamenta esta investigación. Es importante resaltar que tanto el estudiante autor del presente TIP como el proyecto en sí, serán vinculados a un proyecto de investigación financiado por Minciencias, que actualmente se encuentra en desarrollo. Adicionalmente es relevante mencionar, que se realizó la búsqueda en Google Académico, ScienceDirect y SciELO para revisar los artículos que se han publicado respecto a la predicción de series de tiempo de PTB en Colombia, de la siguiente manera: “Tuberculosis times series forecasting in Colombia”, la búsqueda arrojó 11.500, 133 y 0 resultados, respectivamente. Tres de los resultados obtenidos con Google Académico son artículos publicados por el Ingeniero Andrés L. Jutinico. Dos de estas investigaciones están enfocadas en la predicción de TB en Bogotá y una investigación se enfocó en la predicción de casos de TB en el departamento de Antioquia, dicho artículo se denomina: “Towards Tuberculosis Incidence Time Series Forecasting in Colombian Regions: The Antioquia Case”, el cual hace parte integral de este TIP. Por lo tanto, este proyecto se enfocará en la predicción de casos de PTB para los departamentos de Colombia haciendo uso de la información proporcionada por SIVIGILA.

1.3. Objetivos

1.3.1. Objetivo general

Predecir series temporales de incidencia de tuberculosis pulmonar a partir de algoritmos de inteligencia computacional con base en la información suministrada por el Sistema Nacional de Vigilancia en Salud Pública.

1.3.2. Objetivos específicos

- Generar una base de datos de los casos de incidencia de TB Pulmonar desde el año 2007 hasta 2021 considerando los 32 departamentos del territorio nacional.
- Entrenar modelos computacionales que permitan la predicción de casos de TB Pulmonar en por lo menos 4 semanas hacia el futuro considerando Redes Neuronales Artificiales (ANN), el Filtro de Kalman (KF) y el Modelo autorregresivo integrado de media móvil (ARIMA).
- Evaluar la efectividad de los algoritmos implementados mediante los índices de Error Porcentual Absoluto Medio (MAPE), Error Cuadrático Medio (RMSE) y Error Absoluto Medio (MAE).
- Determinar el número máximo de semanas para las cuales se logra predecir la incidencia de TB Pulmonar en los departamentos de Colombia a partir de los algoritmos de inteligencia computacional evaluados.

1.4. Estado del Arte

En Colombia, en el año 2020, aproximadamente 11.390 personas se enfermaron de TB, de las cuales 10.632 fueron casos nuevos, lo que revela que en el país, la tasa de incidencia de tuberculosis es de 20,88 por cada 100.000 habitantes [9]. La forma mas conocida de esta enfermedad afecta a los pulmones, contrayéndose por vía respiratoria cuando las personas infectadas con TB tosen y expulsan la bacteria. La predicción de la tendencia epidémica a partir de los datos que recopila el SIVIGILA le puede servir al sistema sanitario para prevenir y controlar la aparición de casos de TB. Existen varios métodos de predicción de enfermedades infecciosas [12]. Sin embargo, la predicción de la TB sigue siendo un reto dados los cambios bruscos de tendencia y la dispersión de la incidencia de la TB con relación a cada región o país. Obsérvese que el comportamiento de esta epidemia depende de varios factores: la enfermedad es contagiosa, se suele presentar notoriamente según los factores socioeconómicos y adicionalmente, la importancia del estado para velar por la salud de las personas en las poblaciones. Algunos de estos factores son aleatorios y variables en el tiempo, por ejemplo, el aumento de la población inmigrante, que representa una parte importante de las epidemias de TB [17]. En Kashgar (China), Zheng et al., realizó el análisis de series temporales con el método híbrido Box-Jenkins y la red Elman Neural Network para predecir la incidencia de TB, donde hay una alta tasa de aparición de TB. Concluyen que este método híbrido AR es eficaz y puede predecir la incidencia de TB en este lugar [12]. En Taiwan, Chu-Chang Ku et al., desarrolló un modelo de predicción sobre el impacto del envejecimiento de la población de Taiwán en la incidencia de la tuberculosis. El modelo desarrollado indica que en 2035 la tasa de incidencia de la tuberculosis disminuirá un 54 % respecto a 2015. No obstante, destacan que tendrán mayores tasas de mortalidad por tuberculosis en los adultos mayores [11]. Liu et al., abordó la predicción de casos de TB pulmonar en el sureste de China, en la provincia de Jiangsu, donde las tasas de incidencia de TB son altas. Los autores utilizaron el modelo de media móvil integrada autorregresiva (ARIMA) y la red neuronal de retropropagación (BPNN). Ambos métodos fueron evaluados para predecir la estacionalidad y la tendencia mediante índices de error, donde la BPNN presentó mejores resultados en la predicción [18]. Otro estudio realizado en Jiangsu, arrojó resultados similares utilizando un sistema híbrido que incorpora el modelo ARIMA y las redes neuronales autorregresivas no lineales (NAR), los datos que se utilizaron para este estudio corresponden a los casos mensuales de TB entregados por la Oficina de Salud de la provincia de Jiangsu en China desde enero de 2007 hasta marzo de 2016. En esta investigación el modelo ARIMA presentó mejores resultados y los autores concluyen que este algoritmo se puede implementar para series de tiempo con características de estacionalidad y tendencia [19]. En la Provincia del Cabo (Sudáfrica), los autores compararon dos métodos para predecir la incidencia de la tuberculosis y analizaron su estacionalidad. Estos métodos son la media móvil integrada autorregresiva estacional (SARIMA) y la autorregresiva de red neuronal (SARIMA-NNAR) [20]. En 2020 ArunKumar et al., realizó un estudio para predecir las tendencias epidemiológicas de la

pandemia de COVID-19 para los 16 países que contenían el 70 % o 80 % de los casos totales de dicha enfermedad, entre ellos Colombia; para ello, los autores utilizaron los modelos ARIMA y SARIMA [21]. Los índices de TB a nivel mundial siguen siendo altos, esto debido a que aún hay personas que se infectan con la bacteria y nunca son diagnosticadas y por ende no reciben el respectivo tratamiento [22], [23]. El índice de contagios de los habitantes de calle a nivel nacional en Estados Unidos es diez veces el índice de contagio general de TB [24]. En Nepal un grupo de investigadores realizó una investigación para verificar si la estrategia de encontrar casos de forma activa, separar con seguridad y tratar con eficacia (FAST por sus siglas en inglés), era viable. Concluyen afirmando que la estrategia es viable siempre y cuando se den las condiciones económicas y organizacionales en los centros médicos [25]. En Colombia, los estudios relacionados con el pronóstico de series temporales se dirigen principalmente a otro tipo de enfermedades. En [26] se estudió el pronóstico de la enfermedad del Dengue, aplicando modelos lineales generalizados dinámicos jerárquicos bayesianos y se trataron casos de incidencia semanal desde 2008 hasta 2015. En [27] se ha tratado el pronóstico de la enfermedad del Zika en la región de Antioquia utilizando el modelo generalizado de Richards. La tuberculosis es una enfermedad que puede ser curada desde hace 70 años si se proporciona el tratamiento adecuado a tiempo [7], [28]. Sin embargo, según la Organización Mundial de la Salud (WHO por sus siglas en inglés) al menos 3 millones de personas con TB al rededor del mundo, no reciben el respectivo tratamiento por parte del sistema de salud [7]. En el momento en que un individuo infectado con la TB (sin tener conocimiento que es portador) busca tratamiento para los síntomas que están padeciendo, ya ha sufrido los efectos de la bacteria por un largo periodo de tiempo y la transmisión del virus ya ha dado lugar a si círculo cercano y comunidad [29], [30]. Para detener la transmisión de TB, es necesario prevenir la aparición de nuevos casos y posibles nuevas infecciones. En los últimos años ha sido posible y se ha demostrado con modelos matemáticos, la búsqueda de casos activos de TB [31]. La transmisión de TB de un individuo infectado a uno sano, puede ser detenida en el momento en que se comience el tratamiento médico [32]. En [31], los autores confirman que con la inversión económica adecuada, la voluntad y la disposición política, es posible disminuir sustancialmente los índices de contagio. Para realizar la predicción de los casos de TB, lo primero es considerar los modelos que se van a utilizar y los porcentajes de los datos de entrenamiento y test. Al obtener los resultados, se evalúa la precisión de la predicción con los índices de desempeño y para finalizar se presentan las conclusiones.

2 Marco Teórico

En este capítulo abordaremos las definiciones de los modelos computacionales utilizados en el presente TIP, con el fin de dar a entender su definición matemática y estructura lógica. Se consideran modelos computacionales para predecir el número de casos de TB en los departamentos de Colombia. La serie de tiempo que representa la incidencia de TB está conformada por los casos confirmados, reportados por el Sistema de Vigilancia en Salud Pública de Colombia (SIVIGILA) desde el año 2007 hasta el año 2021.

2.1. Red Neuronal Artificial (RNA)

En 1988, Lang y Hinton trabajaron por primera vez con Redes Neuronales Artificiales con retardos de tiempo. Dichas redes hacen parte de las arquitecturas de retroalimentación [33]. Una de las redes más usadas para resolver problemas relacionados con la ingeniería y la ciencia son las Redes Perceptron Multicapa ó Multilayer Perceptron Networks (MLP por sus siglas en inglés), las mismas, hacen parte de las redes neuronales artificiales (RNA) totalmente conectadas. Pueden ser implementadas en diferentes áreas del conocimiento, por ejemplo en la salud, química, física, economía, ecología, psicología, biología, y en un rango de aplicaciones como: predicción de mercados, predicción de consumo energético, predicción de infecciones, etc. Una de las tres clases de problemas que reúne la mayoría de las aplicaciones en las cuales podemos implementar redes MLP tiene que ver con los sistemas dinámicos o aquellos cuyo estado evoluciona en el tiempo [33]. Según [34], las salidas de los sistemas dinámicos, en cualquier momento, dependerán de los valores de entrada y salida previos. La predicción de valores después de un tiempo k se calcula con respecto a valores anteriores, lo que significa:

$$x(k) = f(x(k-1), x(k-2), \dots, x(k-n_p)), \quad (2-1)$$

Donde n_p es el orden del predictor, es decir, el número de muestras pasadas que serán necesarias para la estimación del valor $x(k)$. En la terminología utilizada en el área de identificación de sistemas, el modelo presentado en (2-1) también es conocido como autorregresivo (AR), cuya función $f(\cdot)$ será implementada por la red MLP. En la Tabla **2-1**, presentamos el modelo matemático implementado para la organización y clasificación de la RNA, se presentan datos de los casos de TB reportados por SIVIGILA para la ciudad de Bogotá.

Tabla 2-1: Lista de Entradas y Salidas (RNA)

	X1	X2	X3	Salida deseada
t = 4	13	0	0	22
t = 5	22	13	0	7
t = 6	7	22	13	17
t = 7	17	7	22	17
t = 8	17	17	7	8
t = 9	8	17	17	15
t = 10	15	8	17	13
t = 11	13	15	8	15
t = 12	15	13	15	14
t = ⋮	⋮	⋮	⋮	⋮
t = 771	26	30	36	19

Los retardos de tiempo sirven como memoria, asegurando que las muestras anteriores, que reflejan el comportamiento temporal de la serie, siempre se inserten en la red sin necesidad de retroalimentar sus salidas [33].

2.2. Modelo Autorregresivo Integrado de Media Móvil (ARIMA)

El Modelo ARIMA fue introducido en 1976 por los estadísticos Box y Jenkins, es una metodología econométrica que utiliza variaciones y regresiones de datos con el fin de encontrar patrones para realizar una predicción hacia el futuro [35]. Factores biológicos, económicos, sociales y demográficos hacen parte de los modelos estadísticos de los cuales podemos obtener series temporales. El modelo se ajusta a los datos históricos para predecir los valores futuros mediante la identificación de patrones en la serie de tiempo, combina los modelos de autorregresión (AR) y de media móvil (MA). La ecuación general de un modelo ARIMA es:

$$(1 - \phi_1 L - \dots - \phi_p L^p)(1 - L)^d y_t = (1 + \theta_1 L + \dots + \theta_q L^q) \varepsilon_t \quad (2-2)$$

Donde y_t son los datos de la serie temporal en el tiempo t , L es el operador de retardo, ϕ_1, \dots, ϕ_p y $\theta_1, \dots, \theta_q$ son los coeficientes autorregresivos y de media móvil, respectivamente, y d es el grado de diferenciación. El término ε_t representa el componente de error de ruido blanco. El modelo ARIMA puede especificarse utilizando tres parámetros: p , d y q . El parámetro p representa el número de términos autorregresivos, el parámetro d representa el grado de diferenciación y el parámetro q representa el número de términos de media móvil.

2.3. Modelo autorregresivo N4SID

Utilizamos la técnica de estimación de modelos de espacio de estado utilizando un método de subespacio (N4SID) [36]. Las ecuaciones del modelo en (2-3)-(2-4) se describen de forma canónica, lo que permite pronosticar los estados según el orden del sistema. En este modelo $x_k \in \mathbb{R}^n$ es el vector de estado que se debe estimar con sus correspondientes variables de estado (χ), $y_k \in \mathbb{R}^p$ es la señal de medida, $\omega_k \in \mathbb{R}^{m_2}$ es el ruido de estado y $v_k \in \mathbb{R}^t$ es el ruido de salida. Las matrices nominales F , G y C tienen dimensiones apropiadas.

$$\begin{bmatrix} \chi_{k+1} \\ \chi_{k+2} \\ \vdots \\ \chi_{k+n-1} \\ \chi_{k+n} \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ f_0 & f_1 & f_2 & \cdots & f_{n-1} \end{bmatrix}}_F \begin{bmatrix} \chi_k \\ \chi_{k+1} \\ \chi_{k+2} \\ \vdots \\ \chi_{k+n-1} \end{bmatrix} + \underbrace{\begin{bmatrix} g_{n-1} \\ g_{n-2} \\ g_{n-3} \\ \vdots \\ g_0 \end{bmatrix}}_G \omega_k \quad (2-3)$$

$$y_k = \underbrace{\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \end{bmatrix}}_C \underbrace{\begin{bmatrix} \chi_k \\ \chi_{k+1} \\ \chi_{k+2} \\ \vdots \\ \chi_{k+n-1} \end{bmatrix}}_{x_k} + v_k. \quad (2-4)$$

2.4. El Filtro de Kalman (KF)

El Filtro de Kalman tiene por objeto encontrar las estimaciones óptimas de los estados \hat{x}_k^* , \hat{x}_{k+1}^* reduciendo al mínimo el criterio cuadrático (2-5) sujeto al modelo (2-3)-(2-4).

$$\min_{\hat{\omega}_k, \hat{v}_k, \hat{x}_k, \hat{x}_{k+1}} \left\{ \|\hat{x}_k - \hat{x}_{k|k-1}\|_{P_{k|k-1}^{-1}}^2 + \begin{bmatrix} \hat{\omega}_k \\ \hat{v}_k \end{bmatrix}^T \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix}^{-1} \begin{bmatrix} \hat{\omega}_k \\ \hat{v}_k \end{bmatrix} \right\} \quad (2-5)$$

Las matrices $P_{k|k} \succ 0$, $Q \succ 0$ y $R \succ 0$ son la covarianza de la estimación a posteriori, la covarianza del ruido del proceso y la covarianza del ruido de la observación, respectivamente. La actualización de la matriz $P_{k|k}$ está dada por la ecuación de Riccati (2-6)-(2-7), y la ganancia de Kalman L_k está dada por (2-8).

$$P_{k|k} = P_{k|k-1} - P_{k|k-1} C^T (R + C P_{k|k-1} C^T)^{-1} C P_{k|k-1}, \quad (2-6)$$

$$P_{k+1|k} = Q + F P_{k|k} F^T, \quad (2-7)$$

$$L_k = P_{k|k-1} C^T (R + C P_{k|k-1} C^T)^{-1}. \quad (2-8)$$

Por lo tanto, los estados actuales del modelo (2-3) se estiman mediante el filtro de Kalman de la siguiente manera:

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + L_k(y_k - C\hat{x}_{k|k-1}). \quad (2-9)$$

Los lectores pueden encontrar más detalles sobre el Filtro de Kalman en [37] y [38].

2.5. El Filtro de Kalman Robusto (RKF)

El Filtro de Kalman Robusto fue desarrollado por Cerri en su tesis doctoral [39] y publicado en 2015 junto a Ishihara et al. [40]. Considere el modelo en (2-10)-(2-11) que es una representación del modelo (2-3)-(2-4), pero teniendo en cuenta las incertidumbres del modelo,

$$x_{k+1} = (F_k + \delta F_k)x_k + G_k w_k, \quad (2-10)$$

$$y_k = (C_k + \delta C_k)x_k + v_k, \quad k \geq 0. \quad (2-11)$$

Las matrices inciertas $\delta F_k \in \mathbb{R}^{n \times n}$ y $\delta C_k \in \mathbb{R}^{p \times n}$ se definen del siguiente modo,

$$\delta F_k = M_k \Delta_k E_{F_k}, \quad \delta C_k = N_k \Delta_k E_{C_k}, \quad \|\Delta_k\| \leq 1, \quad (2-12)$$

Donde, E_{F_k} y E_{C_k} deben tener las dimensiones adecuadas, M_k y N_k son matrices distintas de cero, y Δ_k es una contracción arbitraria. Supongamos que x_0 , w_k y v_k son variables aleatorias gaussianas de media cero mutuamente independientes con varianzas $\mathbb{E}\{x_0 x_0^T\} = \Pi_0 \succ 0$, $\mathbb{E}\{w_0 w_0^T\} = Q_k \succ 0$ y $\mathbb{E}\{v_0 v_0^T\} = R_k \succ 0$, respectivamente. El RKF se deduce a partir de la solución del problema de optimización (2-13)-(2-14). El objetivo es obtener un filtrado robusto que minimice la influencia del ruido de estado ω_k y ruido de medición v_k en la estimación de los estados del modelo (2-10)-(2-11), con $e_k = \hat{x}_k - \hat{x}_{k|k-1}$.

$$\min_{\hat{\omega}_k, \hat{v}_k, \hat{x}_k, \hat{x}_{k+1}} \max_{\delta F_k, \delta C_k} \{\mathcal{J}_k\}, \quad (2-13)$$

$$\mathcal{J}_k := \begin{bmatrix} e_k \\ \hat{\omega}_k \\ \hat{v}_k \\ \hat{x}_k \\ \hat{x}_{k+1} \end{bmatrix}^T \begin{bmatrix} P_{k|k-1}^{-1} & 0 & 0 & 0 & 0 \\ 0 & Q_k^{-1} & 0 & 0 & 0 \\ 0 & 0 & R_k^{-1} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} e_k \\ \hat{\omega}_k \\ \hat{v}_k \\ \hat{x}_k \\ \hat{x}_{k+1} \end{bmatrix} + \left\{ \begin{bmatrix} -I & 0 & 0 & I & 0 \\ 0 & G_k & 0 & F_k + \delta F_k & -I \\ 0 & 0 & I & C_k + \delta C_k & 0 \end{bmatrix} \begin{bmatrix} e_k \\ \hat{\omega}_k \\ \hat{v}_k \\ \hat{x}_k \\ \hat{x}_{k+1} \end{bmatrix} - \begin{bmatrix} x_{k|k-1} \\ 0 \\ y_k \end{bmatrix} \right\} \mu \{\bullet\}. \quad (2-14)$$

El esquema para describir el RKF en forma filtrada puede verse en la Tabla 2-2 y sus matrices auxiliares en (2-15)-(2-18). Para garantizar una respuesta óptima, el parámetro $\alpha \geq 1$ está seleccionado.

Tabla 2-2: Filtro de Kalman Robusto Recursivo

Considere la ecuación (2-10), con $\Pi_0 \succ 0$, $Q_k \succ 0$, y $R_k \succ 0$.

Condiciones iniciales: $P_{0|-1} = \Pi_0$ y $\hat{x}_{0|-1} = 0$.

Valor $k \geq 0$: Actualiza $\{\hat{x}_{k+1|k}; P_{k+1|k}\}$ y $\{\hat{x}_{k|k}; P_{k|k}\}$.

$$\begin{bmatrix} \hat{x}_{k|k} & \hat{x}_{k+1|k} \\ P_{k|k} & * \\ * & P_{k+1|k} \end{bmatrix}^T = \begin{bmatrix} 0 \\ 0 \\ 0 \\ I \end{bmatrix}^T \begin{bmatrix} \mathfrak{P}_k & 0 & I & 0 \\ 0 & \mathfrak{S}_k & \mathfrak{A}_k & \mathfrak{B}_k \\ I & \mathfrak{A}_k^T & 0 & 0 \\ 0 & \mathfrak{B}_k^T & 0 & 0 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 0 \\ \mathfrak{Z}_k & 0 \\ 0 & 0 \\ 0 & -I \end{bmatrix},$$

$$\mathfrak{P}_k = \begin{bmatrix} P_{k|k-1} & 0 & 0 \\ 0 & Q_k & 0 \\ 0 & 0 & R_k \end{bmatrix}, \quad \mathfrak{S}_k = \begin{bmatrix} \mu^{-1}I & 0 & 0 \\ 0 & \mathcal{W}_{1,k} & 0 \\ 0 & 0 & \mathcal{W}_{2,k} \end{bmatrix},$$

$$\mathfrak{A}_k = \begin{bmatrix} I & 0 & 0 \\ 0 & \hat{G}_k & 0 \\ 0 & 0 & \hat{D}_k \end{bmatrix}, \quad \mathfrak{B}_k = \begin{bmatrix} -I & 0 \\ \hat{F}_k & -\hat{I} \\ \hat{C}_k & 0 \end{bmatrix}, \quad \mathfrak{Z}_k = \begin{bmatrix} -\hat{x}_{k|k-1} \\ 0 \\ \mathcal{Y}_k \end{bmatrix}.$$

$$\hat{\lambda}_k = (1 + \alpha) \left\| \begin{bmatrix} \mu M_k^T M_k & 0 \\ 0 & \mu N_k^T N_k \end{bmatrix} \right\|, \quad (2-15)$$

$$\mathcal{W}_{1,k} = \begin{bmatrix} (\mu^{-1}I - \hat{\lambda}_k^{-1} M_k M_k^T) & 0 \\ 0 & \hat{\lambda}_k^{-1} I \end{bmatrix}, \quad (2-16)$$

$$\mathcal{W}_{2,k} = \begin{bmatrix} (\mu^{-1}I - \hat{\lambda}_k^{-1} N_k N_k^T) & 0 \\ 0 & \hat{\lambda}_k^{-1} I \end{bmatrix}, \quad (2-17)$$

$$\hat{I} = \begin{bmatrix} I \\ 0 \end{bmatrix}, \quad \hat{F}_k = \begin{bmatrix} F_k \\ E_{F_k} \end{bmatrix}, \quad \hat{G}_k = \begin{bmatrix} G_k \\ 0 \end{bmatrix},$$

$$\mathcal{Y}_k = \begin{bmatrix} y_k \\ 0 \end{bmatrix}, \quad \hat{C}_k = \begin{bmatrix} C_k \\ E_{C_k} \end{bmatrix}, \quad \hat{D}_k = \begin{bmatrix} I \\ 0 \end{bmatrix}. \quad (2-18)$$

3 Desarrollo Metodológico

En este capítulo se presenta como se abordaron los componentes de cada tarea que se realizó durante el desarrollo de la investigación, esto involucró el desarrollo integral de los algoritmos computacionales y la documentación del proceso.

3.1. Generación de la base de datos

Los archivos están segmentados por año y la información está disponible desde el año 2007 hasta 2021. En el siguiente link, se puede descargar la información que es de conocimiento público y hace parte de la información rutinaria recolectada por SIVIGILA:

- <http://portalsivigila.ins.gov.co/Paginas/Vigilancia-Rutinaria.aspx/>

La información publicada por SIVIGILA, hace parte de los datos recolectados a nivel nacional y que han sido confirmados por las entidades territoriales. De la página web, se descargaron 15 bases de datos; desde 2007 hasta 2021. Al momento de abrir un archivo, se observa que la información está categorizada así:

- **Departamento:** Cantidad de eventos ocurridos en los departamentos, clasificados en las 52 semanas del año.
- **Municipio:** Cantidad de eventos ocurridos en los Municipios de Colombia, clasificados en las 52 semanas del año.
- **Evento:** Enfermedades, infecciones, epidemias e insuficiencias, segmentadas por departamento y clasificados en las 52 semanas del año.
- **Exterior:** Eventos ocurridos para ciudadanos Colombianos en el exterior.

La información que se encuentra en cada una de las bases de datos, contiene el número total de eventos ocurridos a nivel nacional. Posterior a la descarga de la información, se procedió a segmentar los datos de los casos de TB Pulmonar para los 32 departamentos y el distrito capital de Bogotá, con un total de 33 bases de datos que pueden ser descargadas del siguiente repositorio:

- <https://onx.la/1f8f3/>

Cabe aclarar que se realizó la segmentación de tal forma que nos permitiera asegurar que los datos presentados son fidedignos con la información entregada por SIVIGILA.

3.2. Generación de los modelos computacionales para el entrenamiento

Se consideraron cinco modelos computacionales: RNA, ARIMA, N4SID, KF y RKF. Para cada modelo, se escribió un código que permite predecir los casos de TB que se pueden presentar en los departamentos de Colombia en por lo menos 7 semanas hacia el futuro. La metodología de entrenamiento es similar para cada uno de los cinco modelos computacionales implementados. En cada caso, se introduce el 70 % de los datos como entrenamiento y el 30 % restante se toma para la validación. A continuación se resume cómo se implemento cada algoritmo en particular.

3.2.1. Modelo RNA

Para la implementación del modelo computacional por RNA se considera una red perceptron multicapa (MLP) con retardo de tiempo (TDNN), la cual hace parte de las arquitecturas feedforward. Para el entrenamiento y validación se utilizó el método backpropagation con el algoritmo Levenberg-Marquardt. Se tomó como base la teoría sobre redes MLP con entradas atrasadas presentada por Iván Nunes Da Silva en [33] para configurar un modelo autorregresivo, por lo cual se procede a organizar los datos de entrada de la red de la siguiente manera:

Tabla 3-1: Organización datos de entrada RNA - Casos de TB en Bogotá

Muestra	X1	X2	X3	X4	X5	X6	Salida deseada
t = 7	17	0	0	0	0	0	8
t = 8	8	17	0	0	0	0	15
t = 9	15	8	17	0	0	0	13
t = 10	13	15	8	17	0	0	15
t = 11	15	13	15	8	17	0	14
t = 12	14	15	13	15	8	17	19
t = 13	19	14	15	13	15	8	9
t = 14	9	19	14	15	13	15	10
t = 15	10	9	19	14	15	13	14
t = 16	14	10	9	19	14	15	14
t = ⋮	⋮	⋮	⋮	⋮			
t = 771	19	26	30	36	30	20	25

Los datos se organizaron según lo presentado en la ecuación 2-1. El modelo de la RNA se estructuró con 10 neuronas y una capa oculta dado que se evaluaron diferentes posibles configuraciones, de una a cien neuronas, y con esta configuración se obtuvo mejores resultados.

La validación cruzada permitió realizar una evaluación eficaz segmentando los datos en dos subconjuntos (entrenamiento y validación). Con cada iteración, se calculó el valor futuro haciendo uso de los valores previos.

3.2.2. Modelo ARIMA

El modelo se ajustó a los datos históricos para predecir los valores futuros mediante la identificación de patrones en la serie de tiempo, a partir de un modelo lineal, como se describe en la ecuación 3-1

$$(1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p)(1 - L)^d y_t = (1 + \theta_1 L + \theta_2 L^2 + \dots + \theta_q L^q) \epsilon_t \quad (3-1)$$

Donde:

- y_t es la serie de tiempo en el tiempo t ,
- ϵ_t es el término de error en el tiempo t ,
- L es el operador de retraso, definido como $Ly_t = y_{t-1}$,
- p es el orden del componente AR (autoregresivo) = 5,
- d es el orden de integración de la serie de tiempo = 2,
- q es el orden del componente MA (media móvil) = 3,
- ϕ_i son los coeficientes AR,
- θ_i son los coeficientes MA.

3.2.3. Modelo Autorregresivo N4SID

Este modelo se ha utilizado para predicción de series temporales en múltiples sectores, como el consumo de energía de una ciudad, el pronóstico del clima, la predicción de mercados financieros, la transmisión de enfermedades, etc [33]. Se realizó el análisis y tratamiento de la serie de tiempo entregada por SIVIGILA teniendo en cuenta que debido a la naturaleza de la información, los datos fueron tratados como sistemas dinámicos. A continuación, se presenta el modelo matemático que se implementó para el modelo N4SID, en forma canónica:

$$X_{k+1} = A_k X_k + B_k u_k + K_k (y_{k+1} - \hat{y}_{k+1}), \quad \hat{y}_{k+1} = C_k X_k + D_k u_k \quad (3-2)$$

Donde:

- X_k es un vector de estado en el tiempo k ,
- u_k es la entrada al sistema en el tiempo k ,

- y_k es la salida del sistema en el tiempo k ,
- \hat{y}_{k+1} es la salida estimada del sistema en el tiempo $k + 1$,
- A_k, B_k, C_k , y D_k son matrices definidas para cada instante de tiempo k ,
- K_k es la ganancia del observador en el tiempo k .

3.2.4. Modelo del Filtro de Kalman

Para conseguir las estimaciones óptimas de los estados \hat{x}_k^* , \hat{x}_{k+1}^* se minimizó el criterio cuadrático (2-5) sujeto al modelo (2-3)-(2-4). Los parámetros del Filtro de Kalman implementado son:

- $P_{k|k} \succ 0$, Corresponde a la covarianza de la estimación a posteriori
- $Q \succ 0$, Corresponde a la covarianza del ruido del proceso (Varianza en los estados)
- $R \succ 0$, Corresponde a la covarianza del ruido de la observación (Varianza en las mediciones)

La actualización de la matriz $P_{k|k}$ está dada por la ecuación de Riccati (2-6)-(2-7), y la ganancia de Kalman L_k está dada por (2-8). Los estados actuales del modelo (2-3) se estimaron mediante (3-3), es importante resaltar que dado que este modelo se encuentra en forma canónica es posible realizar estimaciones hacia el futuro, en correspondencia con el orden del sistema.

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + L_k(y_k - C\hat{x}_{k|k-1}). \quad (3-3)$$

Adicionalmente, la incorporación del Filtro de Kalman en las predicciones permite la corrección del error medio en función del modelo y la corrección de errores dinámicos asociados mediante las covarianzas Q y R . Para las predicciones realizadas en esta investigación, se tomó $R = 0.1$ y $Q = 400$, asumiendo una desviación estándar de 20 casos de TB por semana.

3.2.5. Modelo del Filtro de Kalman Robusto

Para la implementación del RKF se tuvo en cuenta la metodología presentada en [41] en la predicción de series de tiempo de TB en Bogotá. Este algoritmo se deduce a partir de la solución del problema de optimización (2-13)-(2-14). El objetivo es obtener un filtrado robusto que minimice la influencia del ruido de estado ω_k y ruido de medición v_k en la estimación de los estados del modelo (2-10)-(2-11), con $e_k = \hat{x}_k - \hat{x}_{k|k-1}$. El esquema para describir el RKF en forma filtrada puede verse en la Tabla **2-2** y sus matrices auxiliares en (2-15)-(2-18). Para garantizar una respuesta óptima, el parámetro $\alpha \geq 1$ está seleccionado. El RKF utiliza el mismo modelo nominal y covarianzas que el filtro de Kalman. Además,

Tabla 3-2: Parámetros del Filtro de Kalman Robusto.

$$E_{F_k} = [f_0 \ f_1 \ f_2 \ \cdots \ f_{n-1}] \cdot 10^{-3}, \quad E_{C_k} = C \cdot 10^{-3},$$

$$M_k = [1 \ 1 \ \cdots \ 1 \ 1000]^T \cdot 10^{-3}, \quad N_k = 0.5,$$

las matrices inciertas se definen con los parámetros que se muestran en la Tabla **3-2**. Los parámetros auxiliares son $\alpha = 2$ y $\mu = 1 \cdot 10^5$. Al incorporar el RKF para realizar las predicciones de los casos de TB, también se permite la corrección del error medio en función del modelo y la corrección de errores dinámicos del modelo asociados mediante las covarianzas Q y R . No obstante, este algoritmo permite considerar que el modelo es impreciso y definido con incertidumbres, obteniendo de esta manera predicciones diferentes a la obtenidas con el Filtro de Kalman nominal.

3.3. Evaluación de la efectividad de los algoritmos implementados

Para evaluar la efectividad de los modelos de predicción implementados, empleamos los índices estadísticos:

- Error Absoluto Medio (MAE)
- Error Cuadrático Medio (RMSE)
- Señal de Seguimiento (TRS)

Estos índices se emplean habitualmente para evaluar los modelos de predicción [42], [43] y se calculan así:

$$MAE = \frac{1}{\mathcal{N}} \sum_{k=1}^{\mathcal{N}} |e_k|, \quad (3-4)$$

$$RMSE = \sqrt{\frac{1}{\mathcal{N}} \sum_{k=1}^{\mathcal{N}} e_k^2}, \quad (3-5)$$

$$TRS = \frac{\sum_{k=1}^{\mathcal{N}} e_k}{MAE}, \quad (3-6)$$

Donde \mathcal{N} es el número máximo de pasos o puntos en la secuencia y $e_k = y_k - \hat{y}_k$ es el error entre la serie real y la serie obtenida con el modelo computacional en cuestión. Los índices

MAE y RMSE se utilizaron para obtener el valor promedio de la serie de tiempo y así evaluar que tan dispersos estaban los datos de este valor. El índice TRS lo utilizamos para evaluar que tan lejanos estaban los datos obtenidos de los picos de la serie original.

A pesar de que el índice de Error Porcentual Absoluto Medio (MAPE), presentado en la ecuación 3-7, fue propuesto en uno de los objetivos específicos para evaluar el desempeño de los algoritmos, finalmente no fue considerado. Lo anterior, debido a que algunos de los reportes semanales de la incidencia de casos de TB en algunos departamentos del país en el periodo de 2007 a 2021 fue cero, lo cual genera una respuesta indeterminada para este índice.

$$MAPE = \frac{1}{N} \sum_{k=1}^N \left| \frac{e_k}{y_k} \right| 100 \%, \quad (3-7)$$

Un ejemplo que muestra claramente que el índice MAPE no es adecuado para evaluar los algoritmos propuestos, se observa en la Figura 3-1, la cual presenta los casos de TB reportados por SIVIGILA en el departamento de Amazonas de 2007 a 2021.

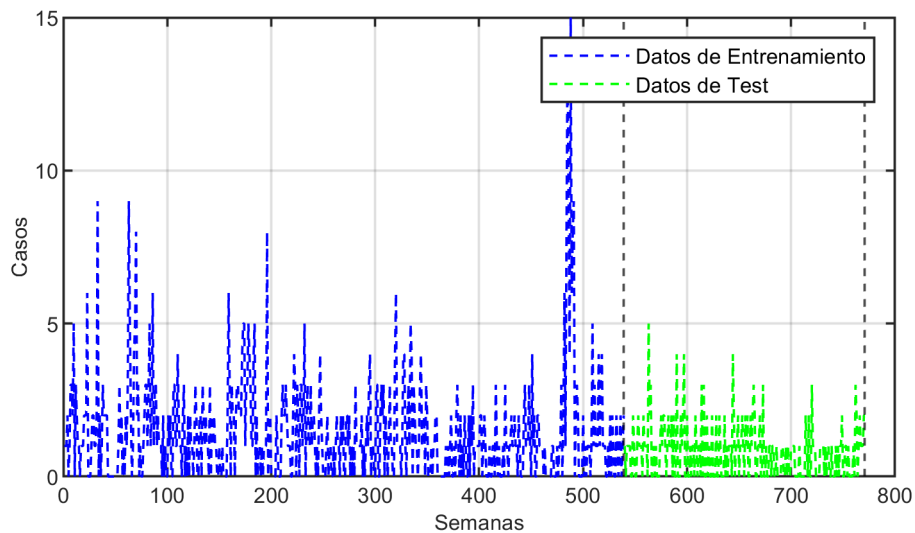


Figura 3-1: Casos de TB en Amazonas de 2007 a 2021

Para el caso de Amazonas, son 275 semanas en las que no se reportaron casos de TB, esto representa el 35,67% de la serie de tiempo, son más de cinco años de los cuales no tenemos información. Por lo tanto se evaluó y consideró otro índice de desempeño como reemplazo al MAPE.

3.4. Número máximo de semanas para las cuales se logra una predicción aceptable

La información de los casos de TB en los departamentos de Colombia, se usó de tal manera, que permitió obtener un modelo de espacios de estado, entrenando con el 70% de las series de tiempo. Para definir el modelo se tuvo en cuenta principalmente el número de pasos que se logran predecir así como también el orden del sistema. Por lo tanto, se consideraron modelos AR con diferente orden n para predecir la incidencia de TB entre una y diez semanas. Además, se empleó un Filtro de Kalman con covarianza $Q = 400$ y $R = 0.1$ para la estimación de los datos. Lo anterior, dado que asumimos perturbaciones desconocidas relativas a imperfecciones del proceso de identificación. Por lo tanto, se considera una desviación estándar de 20 casos de TB.

A modo de ejemplo se presentan los resultados obtenidos con los índices de evaluación RMSE, MAE y TRS para el departamento de Valle del Cauca. En este sentido, realizamos tres mapas de calor considerando varios tamaños del sistema y diferentes periodos de tiempo hacia el futuro. La Figura 3-2 muestra el mapa de calor para los valores RMSE, el mapa relaciona la predicción de paso adelante m , y la variable α , que es proporcional con el orden del sistema, donde $n = m + \alpha - 1$.

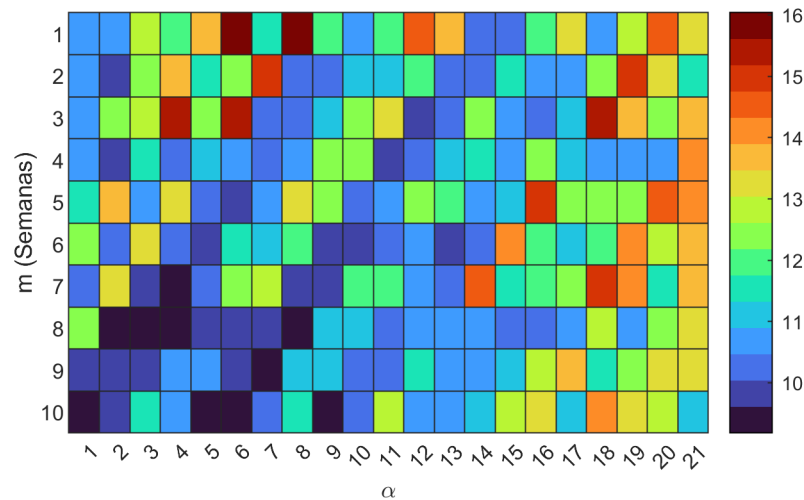


Figura 3-2: Mapa de calor de los valores RMSE para Valle del Cauca

Observe que los resultados con menor porcentaje de error en la Figura 3-2, tienen un color azul oscuro, por lo tanto podemos afirmar que el mejor resultado tiene los parámetros $m = 7$ y $\alpha = 4$.

Las Figuras 3-3 y 3-4 muestran el mapa de calor para los valores MAE y TRS, respectivamente. Donde n es el orden del sistema, m es el paso de predicción hacia adelante, y α es una variable relacionada, donde $n = m + \alpha - 1$.

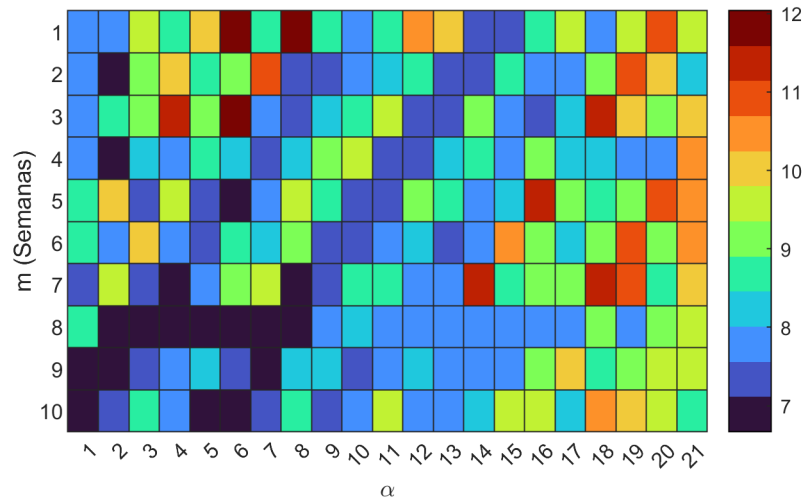


Figura 3-3: Mapa de calor de los valores MAE para Valle del Cauca

Una respuesta similar se obtuvo con el índice MAE, como se observa en la Figura 3-3, los datos con menor porcentaje de error tienen un color azul oscuro, lo que nos permite afirmar que para la predicción en más de 4 semanas hacia el futuro, el mejor resultado tiene los parámetros $m = 7$ y $\alpha = 4$.

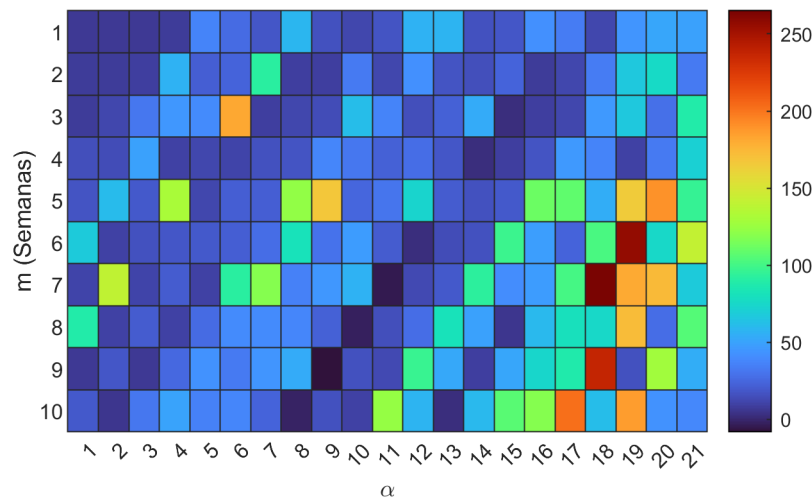


Figura 3-4: Mapa de calor de los valores TRS para Valle del Cauca

Debido a que cada población tiene condiciones únicas, como factores económicos, políticos, ambientales, sociales y culturales, se considera necesario evaluar cada algoritmo individualmente para cada uno de los departamentos con el fin de obtener el mejor resultado posible, determinando así el número máximo de semanas que se pueden predecir con el menor error.

4 Resultados

Para presentar los resultados, se tomó como base la información publicada por SIVIGILA en la Semana 38 del calendario epidemiológico 2021. En la tabla 4-1 se observan los cinco departamentos con mayor cantidad de casos de TB reportados, los cuales representan el 61,3% de los casos de TB confirmados en ese periodo de tiempo.

Tabla 4-1: Departamentos con mayor índice de TB en Colombia en 2021

Departamento	Casos Reportados
Valle del Cauca	259
Antioquia	92
Santander	81
Bogotá	76
Meta	38

En la Figura 4-1 se presenta a modo de ejemplo la serie de tiempo que representa los casos reportados de TB por SIVIGILA para el departamento de Valle del Cauca desde el año 2007 hasta el año 2021.

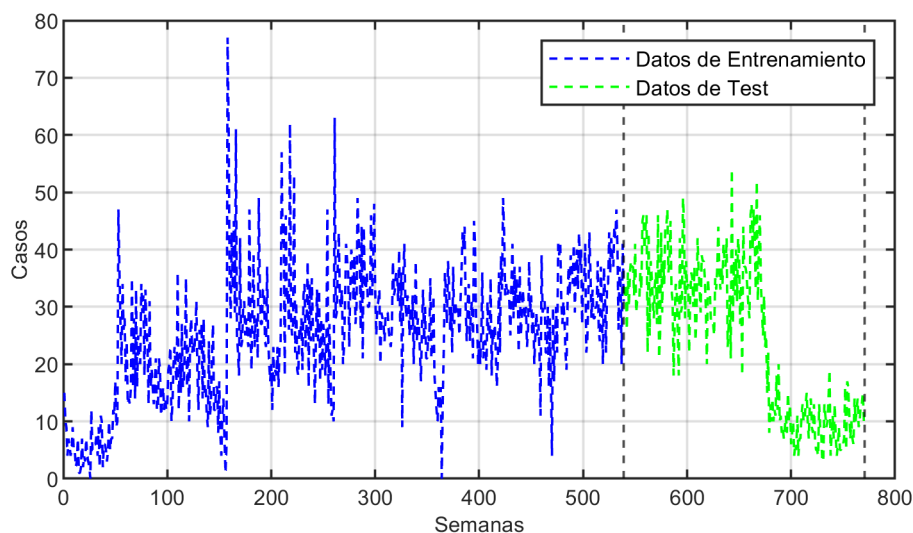


Figura 4-1: Casos TB Pulmonar en Valle del Cauca de 2007 a 2021

Donde, el 70% de los datos se usó para identificación de los modelos de predicción y el 30% de los datos se usó para validación. Es de resaltar que en la Figura 4-1 los datos de entrenamiento y de test se encuentran separados por una línea punteada de color negro. El entrenamiento de los modelos computacionales se realizó considerando los casos de TB ocurridos entre la semana 1 y la semana 539. A continuación se presentan los resultados obtenidos para los departamentos que se presentaron en la Tabla 4-1.

4.1. Valle del Cauca

4.1.1. Resultados de la RNA para Valle del Cauca

A continuación, se presentan los resultados que se obtuvieron al predecir los casos de TB, en pasos de siete semanas hacia adelante, con la RNA para Valle del Cauca. En la Figura 4-2, se observa que, el algoritmo logra predecir la tendencia de la serie, sin embargo no llega hasta los picos cuando la dispersión es alta.

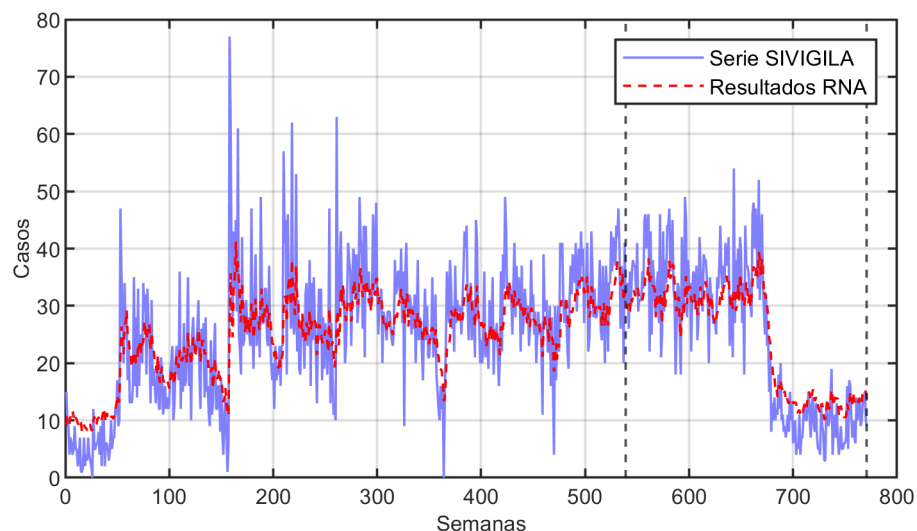


Figura 4-2: Resultados RNA Valle del Cauca

En los resultados de entrenamiento se obtuvo para los índices de error RMSE, MAE y TRS: 6.9678, 5.1829 y 40.315, respectivamente. En los resultados de validación se obtuvo para los índices de error RMSE, MAE y TRS: 5.9073, 4.7555 y -9.2319, respectivamente.

4.1.2. Resultados del modelo ARIMA para Valle del Cauca

A continuación, se presentan los resultados que se obtuvieron al predecir los casos de TB, en pasos de siete semanas hacia adelante, con el modelo ARIMA para Valle del Cauca. En la Figura 4-3, se observa que en los resultados de entrenamiento y validación, el algoritmo

tiende a presentar una serie de datos lejana a la realidad, los valores están siempre en los extremos y en pocas circunstancias se encuentran en la media.

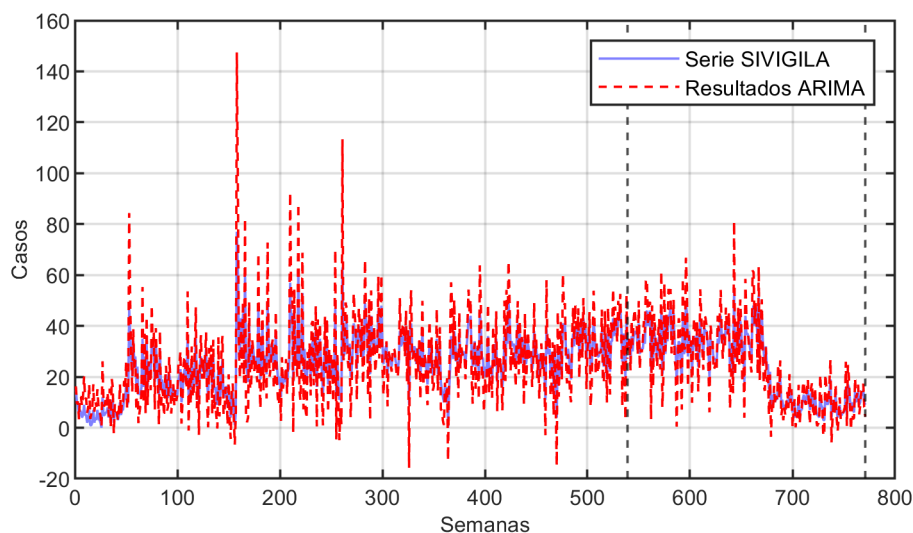


Figura 4-3: Resultados modelo ARIMA Valle del Cauca

En los resultados de entrenamiento se obtuvo para los índices de error RMSE, MAE y TRS: 8.9865, 6.4992 y -18.137, respectivamente. En los resultados de validación se obtuvo para los índices de error RMSE, MAE y TRS: 6.6147, 5.0594 y -14.544, respectivamente.

4.1.3. Resultados del modelo N4SID para Valle del Cauca

En la Figura 4-4, se presentan los resultados que se obtuvieron al predecir los casos de TB, en pasos de siete semanas hacia adelante, con el modelo autorregresivo N4SID para Valle del Cauca. La señal que se obtuvo arrojó un resultado favorable sin embargo los datos siempre se encontraron en la media. En los resultados de entrenamiento se obtuvo para los índices de error RMSE, MAE y TRS: 8.8745, 6.3256 y 51.024, respectivamente. En los resultados de validación se obtuvo para los índices de error RMSE, MAE y TRS: 7.3605, 5.9189 y -53.607, respectivamente.

4.1.4. Resultados del Filtro de Kalman para Valle del Cauca

Con el KF se obtuvieron resultados acertados, donde se logra predecir la tendencia de la serie, la señal se ajusta a la media y en ocasiones alcanza los picos de la señal. En la Figura 4-5, se observan los resultados de entrenamiento y validación.

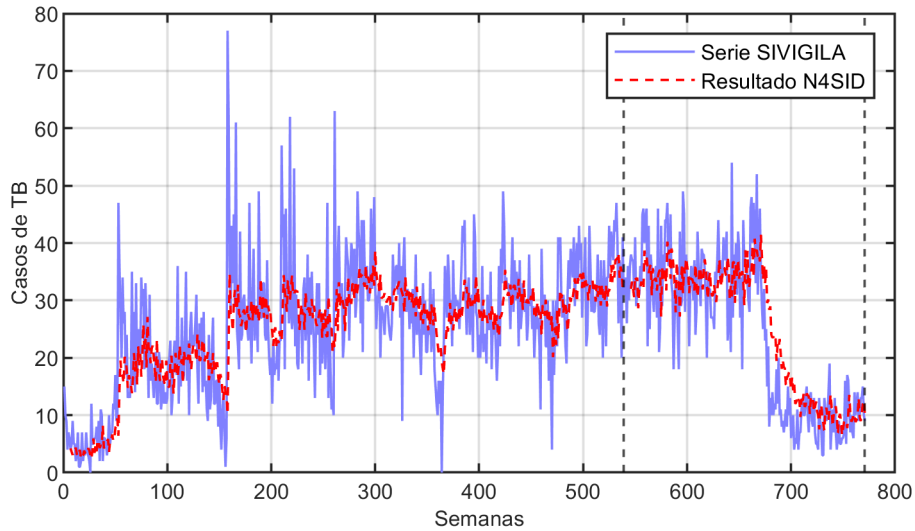


Figura 4-4: Resultados modelo autorregresivo N4SID Valle del Cauca

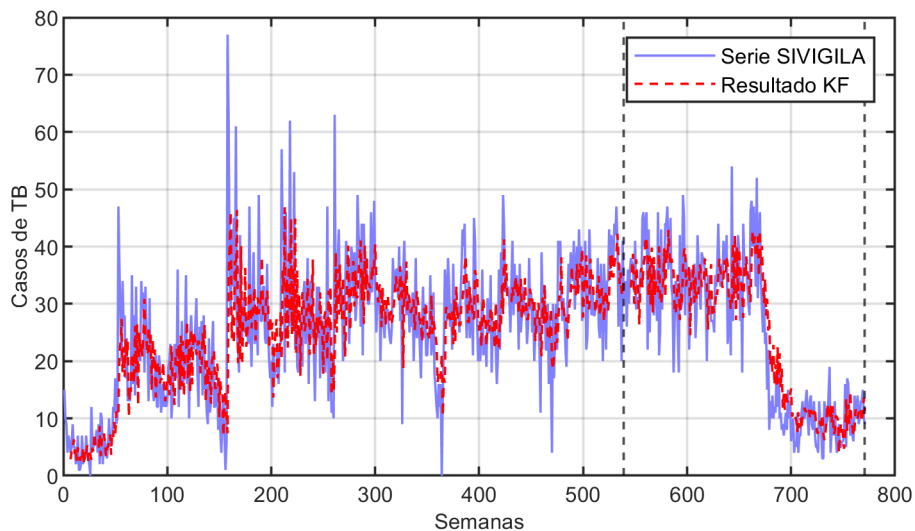


Figura 4-5: Resultados Filtro de Kalman Valle del Cauca

En los resultados de entrenamiento se obtuvo para los índices de error RMSE, MAE y TRS: 9.318, 6.7027 y 23.152, respectivamente. En los resultados de validación se obtuvo para los índices de error RMSE, MAE y TRS: 7.1334, 5.5712 y -33.015, respectivamente.

4.1.5. Resultados del Filtro de Kalman Robusto para Valle del Cauca

En la Figura 4-6 se presentan los resultados que se obtuvieron al predecir los casos de TB, en pasos de siete semanas hacia adelante, con el Filtro de Kalman Robusto para Valle del Cauca. En este caso, obtuvimos una señal que logra acertar a la tendencia de la serie original tanto en los valores medios como en los picos.

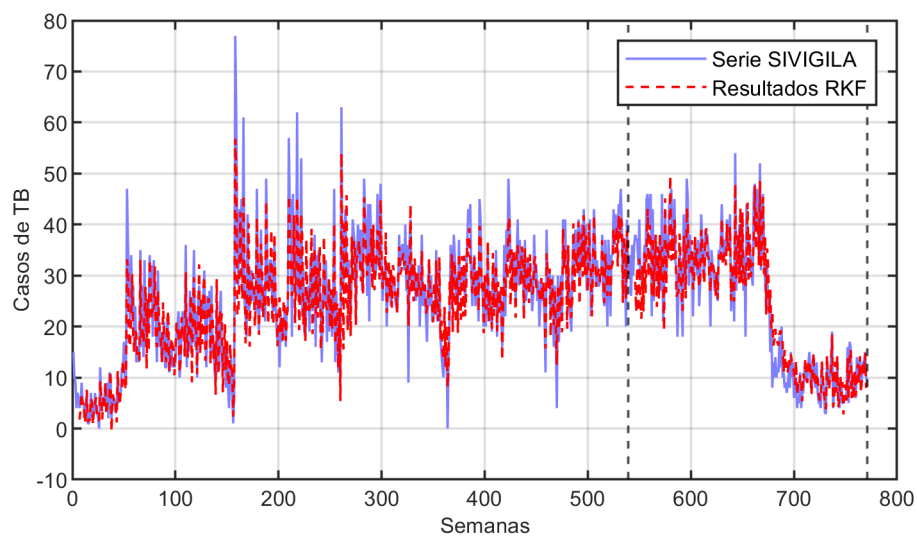


Figura 4-6: Resultados Filtro de Kalman Robusto Valle del Cauca

En los resultados de entrenamiento se obtuvo para los índices de error RMSE, MAE y TRS: 5.3336, 4.0363 y 88.722, respectivamente. En los resultados de validación se obtuvo para los índices de error RMSE, MAE y TRS: 4.785, 3.631 y 19.787, respectivamente.

4.1.6. Resultados de los algoritmos computacionales para Valle del Cauca

Se realizó la comparativa de la serie de tiempo original con las cinco series de tiempo que arrojaron los algoritmos computacionales implementados para Valle del Cauca. En la Figura 4-7, se observan los resultados de entrenamiento.

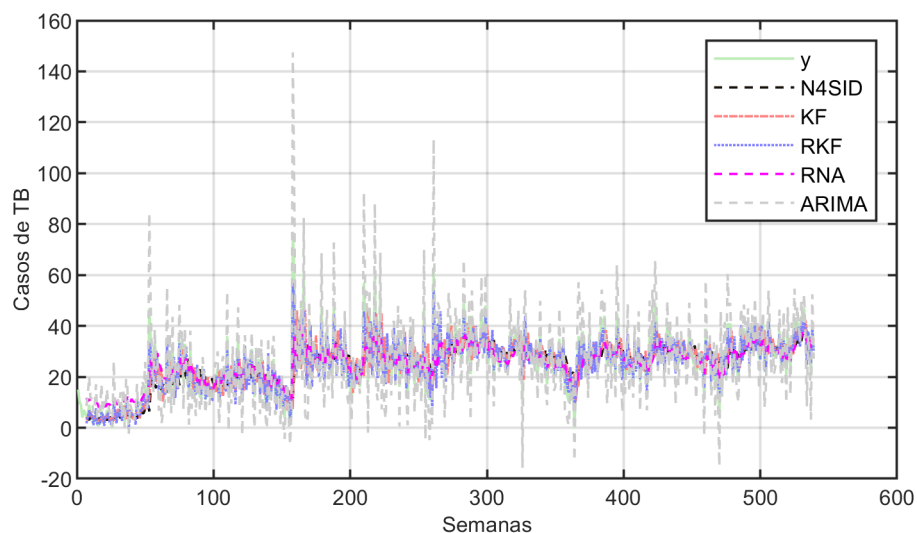


Figura 4-7: Resultados Entrenamiento Algoritmos Computacionales Valle del Cauca

En la Figura 4-8, se observan los resultados de validación.

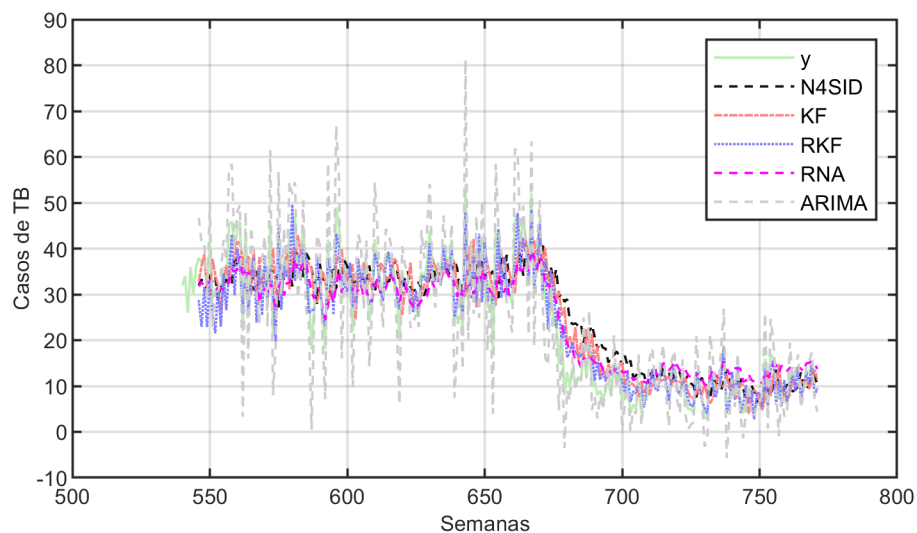


Figura 4-8: Resultados Validación Algoritmos Computacionales Valle del Cauca

En la tabla 4-2, se presentan los resultados obtenidos con los índices de error RMSE, MAE y TRS para los cinco algoritmos implementados en el proceso de entrenamiento.

Tabla 4-2: Resultados de los algoritmos computacionales en el entrenamiento para el departamento de Valle del Cauca

Algoritmo	RMSE	MAE	TRS
Modelo Autorregresivo N4SID	8.8745	6.3256	51.024
Filtro de Kalman	9.318	6.7027	23.152
Red Neuronal Artificial	6.9678	5.1829	40.315
Modelo ARIMA	8.9865	6.4992	-18.137
Filtro de Kalman Robusto	5.3336	4.0363	88.722

En la tabla 4-3, se presentan los resultados obtenidos con los índices de error RMSE, MAE y TRS para los cinco algoritmos implementados en el proceso de validación.

Tabla 4-3: Resultados de los algoritmos computacionales en la validación para el departamento de Valle del Cauca

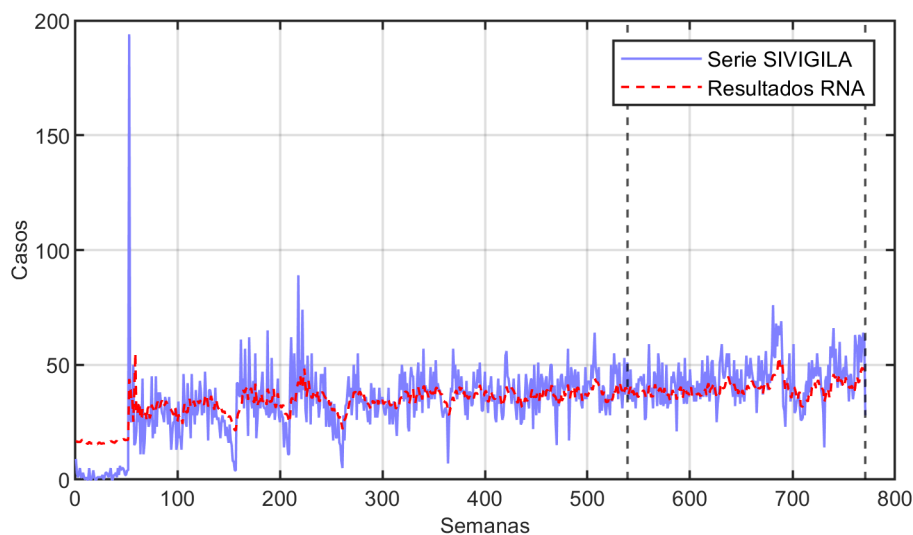
Algoritmo	RMSE	MAE	TRS
Modelo Autorregresivo N4SID	7.3605	5.9189	-53.607
Filtro de Kalman	7.1334	5.5712	-33.015
Red Neuronal Artificial	5.9073	4.7555	-9.2319
Modelo ARIMA	6.6147	5.0594	-14.544
Filtro de Kalman Robusto	4.785	3.631	19.787

En el caso de Valle del Cauca, los resultados obtenidos con los cinco algoritmos son satisfactorios, esta eficiencia se confirma con los índices de desempeño. Indudablemente el mejor resultado lo presentó el Filtro de Kalman Robusto que logra seguir a detalle la serie original.

4.2. Antioquia

4.2.1. Resultados de la RNA para Antioquia

En la Figura 4-9 se presentan los resultados que se obtuvieron al predecir los casos de TB, en pasos de siete semanas hacia adelante, con la RNA para Antioquia.

**Figura 4-9:** Resultados RNA Antioquia

En los resultados de entrenamiento se obtuvo para los índices de error RMSE, MAE y TRS: 11.561, 7.7322 y -78.235, respectivamente. En los resultados de validación se obtuvo para los índices de error RMSE, MAE y TRS: 8.6545, 6.7919 y 126.09, respectivamente.

4.2.2. Resultados del modelo ARIMA para Antioquia

En la Figura 4-10 se presentan los resultados que se obtuvieron al predecir los casos de TB, en pasos de siete semanas hacia adelante, con el modelo ARIMA para Antioquia. En los

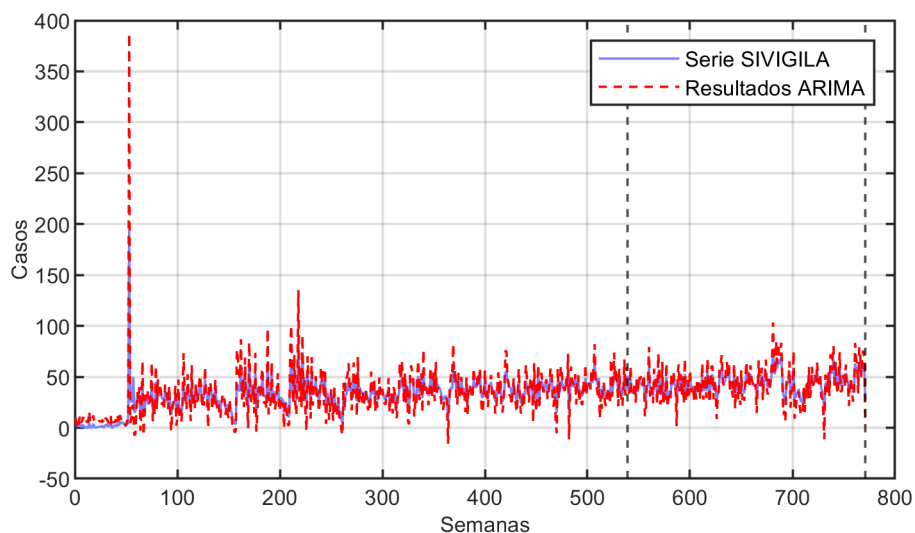


Figura 4-10: Resultados modelo ARIMA Antioquia

resultados de entrenamiento se obtuvo para los índices de error RMSE, MAE y TRS: 12.748, 7.7462 y -24.223, respectivamente. En los resultados de validación se obtuvo para los índices de error RMSE, MAE y TRS: 8.6207, 6.8759 y -5.7907, respectivamente.

4.2.3. Resultados del modelo N4SID para Antioquia

En la Figura 4-11 se presentan los resultados que se obtuvieron al predecir los casos de TB, en pasos de siete semanas hacia adelante, con el modelo autorregresivo N4SID para Antioquia.

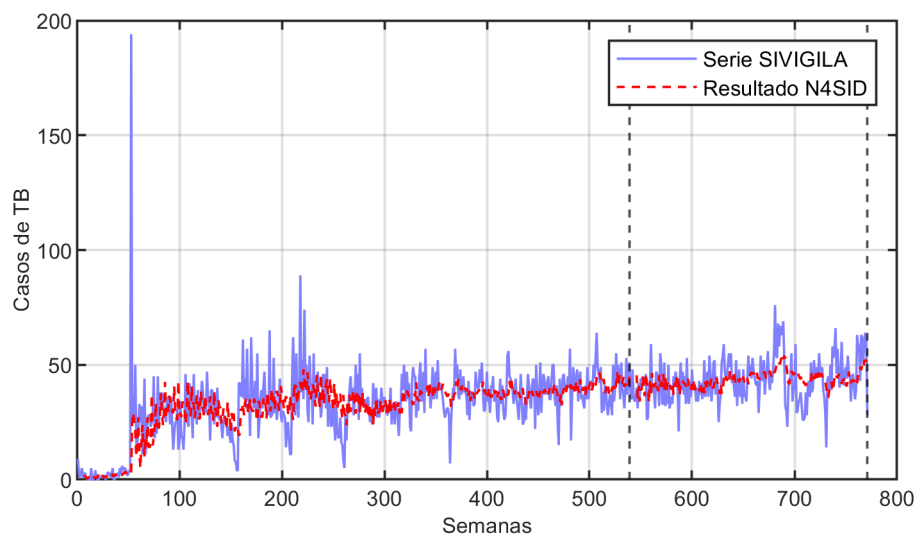


Figura 4-11: Resultados modelo autorregresivo N4SID Antioquia

En los resultados de entrenamiento se obtuvo para los índices de error RMSE, MAE y TRS: 13.076, 7.8265 y 35.725, respectivamente. En los resultados de validación se obtuvo para los índices de error RMSE, MAE y TRS: 9.2594, 7.178 y 5.2422, respectivamente.

4.2.4. Resultados del Filtro de Kalman para Antioquia

En la Figura 4-12 se presentan los resultados que se obtuvieron al predecir los casos de TB, en pasos de siete semanas hacia adelante, con el Filtro de Kalman para Antioquia.

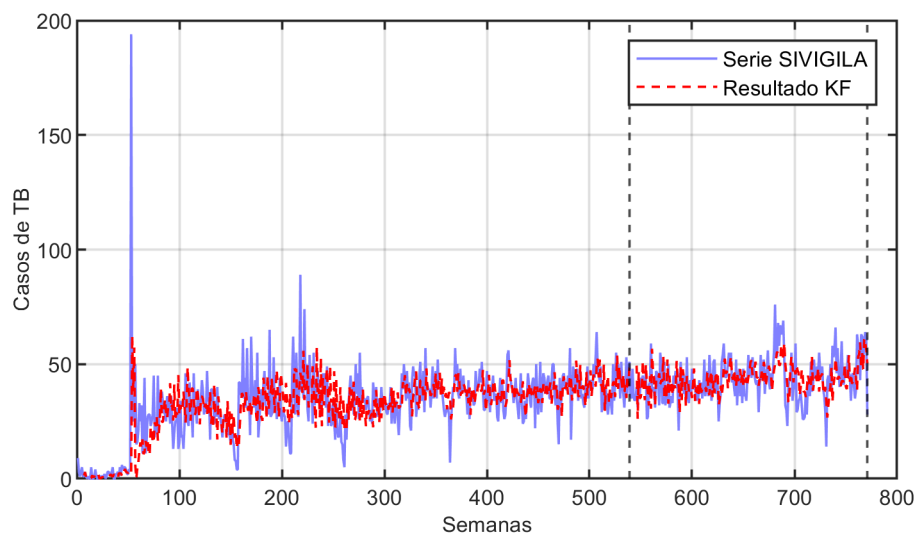


Figura 4-12: Resultados Filtro de Kalman Antioquia

En los resultados de entrenamiento se obtuvo para los índices de error RMSE, MAE y TRS:

13.765, 8.3917 y 32.791, respectivamente. En los resultados de validación se obtuvo para los índices de error RMSE, MAE y TRS: 9.6446, 7.6178 y 0.43447, respectivamente.

4.2.5. Resultados del Filtro de Kalman Robusto para Antioquia

En la Figura 4-13 se presentan los resultados que se obtuvieron al predecir los casos de TB, en pasos de siete semanas hacia adelante, con el Filtro de Kalman Robusto para Antioquia.

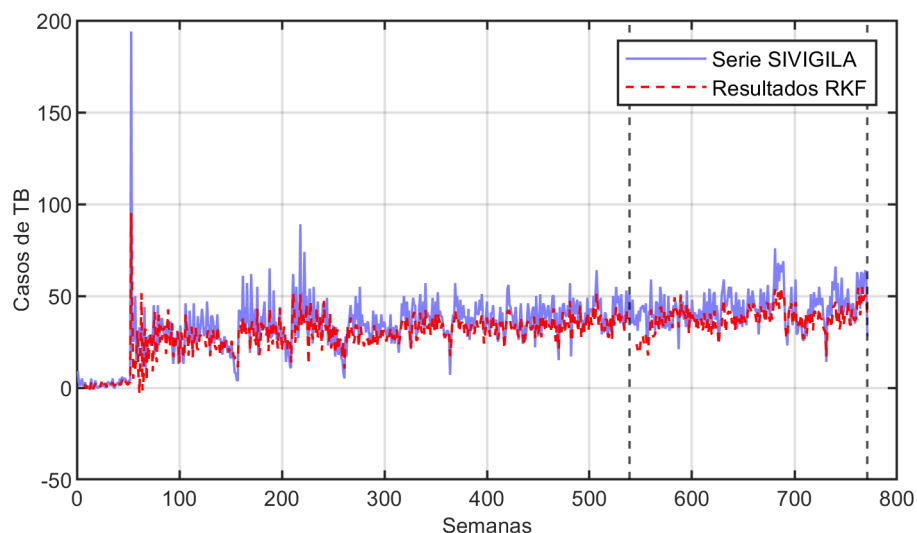


Figura 4-13: Resultados Filtro de Kalman Robusto Antioquia

En los resultados de entrenamiento se obtuvo para los índices de error RMSE, MAE y TRS: 9.4318, 6.3084 y 346.54, respectivamente. En los resultados de validación se obtuvo para los índices de error RMSE, MAE y TRS: 8.5811, 6.9884 y 185.03, respectivamente.

4.2.6. Resultados de los algoritmos computacionales para Antioquia

Se realizó la comparativa de la serie de tiempo original con las cinco series de tiempo que arrojaron los algoritmos computacionales implementados para Antioquia. En la Figura 4-14, se observan los resultados de entrenamiento.

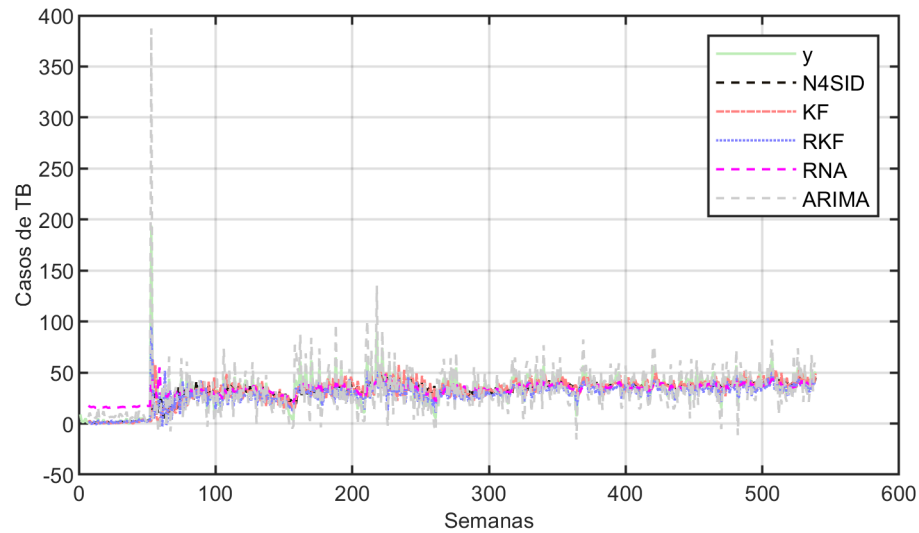


Figura 4-14: Resultados Entrenamiento Algoritmos Computacionales Antioquia

En la Figura 4-15, se observan los resultados de validación.

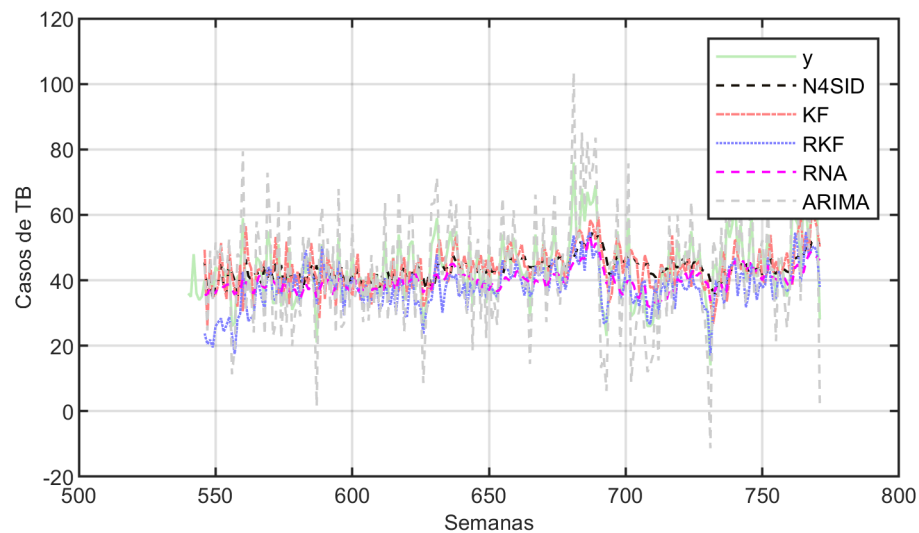


Figura 4-15: Resultados Validación Algoritmos Computacionales Antioquia

En la tabla 4-4, se presentan los resultados obtenidos con los índices de error RMSE, MAE y TRS para los cinco algoritmos implementados en el proceso de entrenamiento. En la tabla 4-5, se presentan los resultados obtenidos con los índices de error RMSE, MAE y TRS para los cinco algoritmos implementados en el proceso de validación.

Tabla 4-4: Resultados de los algoritmos computacionales en el entrenamiento para el departamento de Antioquia

Algoritmo	RMSE	MAE	TRS
Modelo Autorregresivo N4SID	13.076	7.8265	35.725
Filtro de Kalman	13.765	8.3917	32.791
Red Neuronal Artificial	11.561	7.7322	-78.235
Modelo ARIMA	12.748	7.7462	-24.223
Filtro de Kalman Robusto	9.4318	6.3084	346.54

Tabla 4-5: Resultados de los algoritmos computacionales en la validación para el departamento de Antioquia

Algoritmo	RMSE	MAE	TRS
Modelo Autorregresivo N4SID	9.2594	7.178	5.2422
Filtro de Kalman	9.6446	7.6178	0.43447
Red Neuronal Artificial	8.6545	6.7919	126.09
Modelo ARIMA	8.6207	6.8759	-5.7907
Filtro de Kalman Robusto	8.5811	6.9884	185.03

En Antioquia, hay un caso particular, alrededor de la semana 53 se presentaron 194 casos de TB. Este valor supera por más de 100 el promedio de casos de TB en toda la serie de tiempo. Sin embargo los valores obtenidos con los índices de desempeño son coherentes con la realidad. Definitivamente el mejor resultado se obtuvo con el Filtro de Kalman, que fue fidedigno a la serie original.

4.3. Santander

4.3.1. Resultados de la RNA para Santander

En la Figura 4-16 se presentan los resultados que se obtuvieron al predecir los casos de TB, en pasos de siete semanas hacia adelante, con la RNA para Santander.

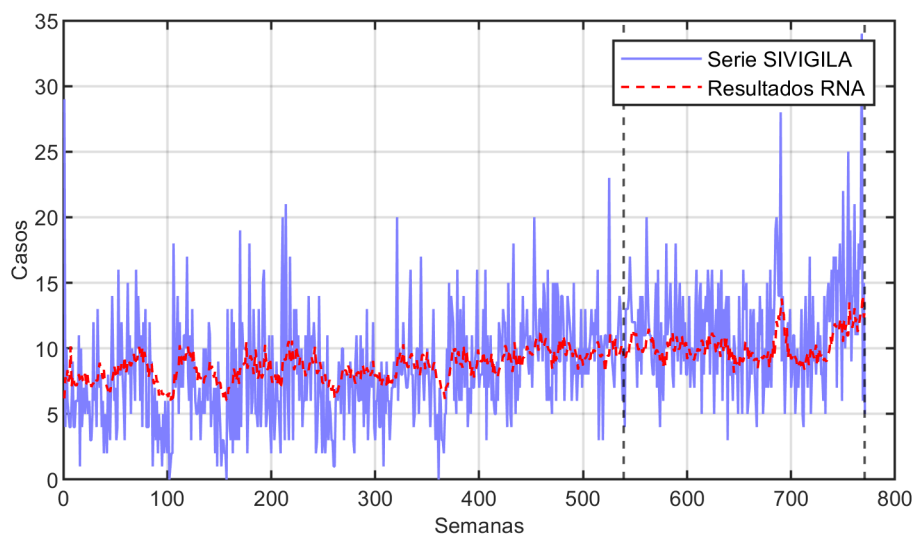


Figura 4-16: Resultados RNA Santander

En los resultados de entrenamiento se obtuvo para los índices de error RMSE, MAE y TRS: 3.7024, 2.899 y -91.603, respectivamente. En los resultados de validación se obtuvo para los índices de error RMSE, MAE y TRS: 4.1629, 3.2042 y 87.515, respectivamente.

4.3.2. Resultados del modelo ARIMA para Santander

En la Figura 4-17 se presentan los resultados que se obtuvieron al predecir los casos de TB, en pasos de siete semanas hacia adelante, con el modelo ARIMA para Santander.

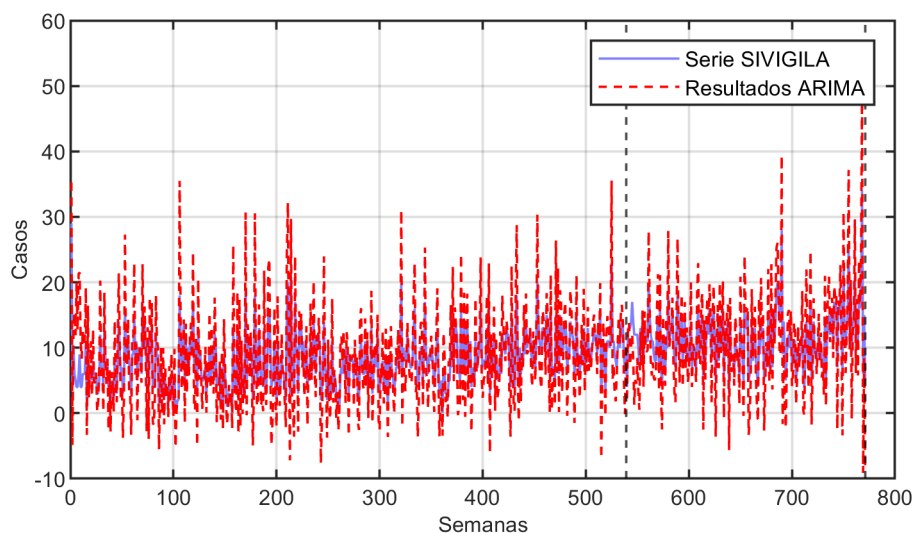


Figura 4-17: Resultados modelo ARIMA Santander

En los resultados de entrenamiento se obtuvo para los índices de error RMSE, MAE y TRS:

4.3309, 3.3171 y -15.929, respectivamente. En los resultados de validación se obtuvo para los índices de error RMSE, MAE y TRS: 4.6335, 3.6257 y 10.685, respectivamente.

4.3.3. Resultados del modelo N4SID para Santander

En la Figura 4-18 se presentan los resultados que se obtuvieron al predecir los casos de TB, en pasos de siete semanas hacia adelante, con el modelo autorregresivo N4SID para Santander.

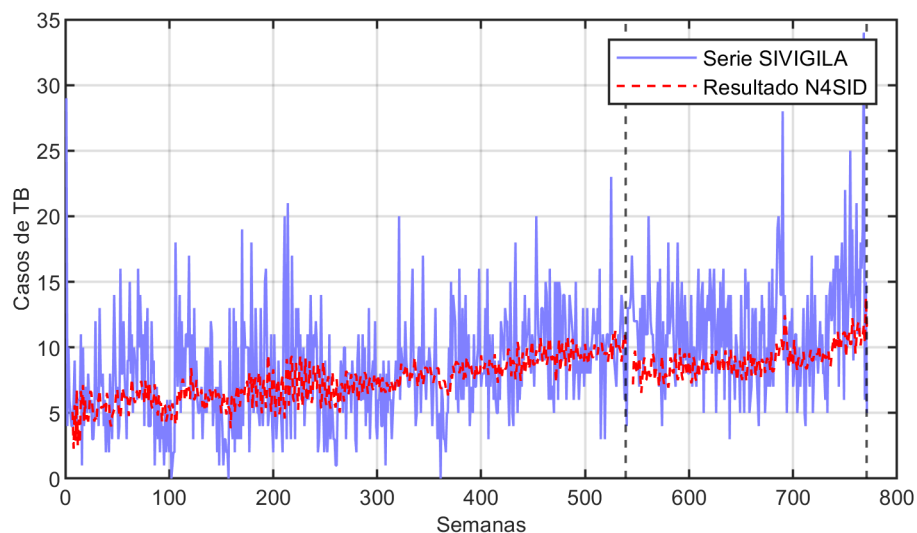


Figura 4-18: Resultados modelo autorregresivo N4SID Santander

En los resultados de entrenamiento se obtuvo para los índices de error RMSE, MAE y TRS: 3.8268, 2.9476 y 115.16, respectivamente. En los resultados de validación se obtuvo para los índices de error RMSE, MAE y TRS: 4.7213, 3.6282 y 142.07, respectivamente.

4.3.4. Resultados del Filtro de Kalman para Santander

En la Figura 4-19 se presentan los resultados que se obtuvieron al predecir los casos de TB, en pasos de siete semanas hacia adelante, con el Filtro de Kalman para Santander.

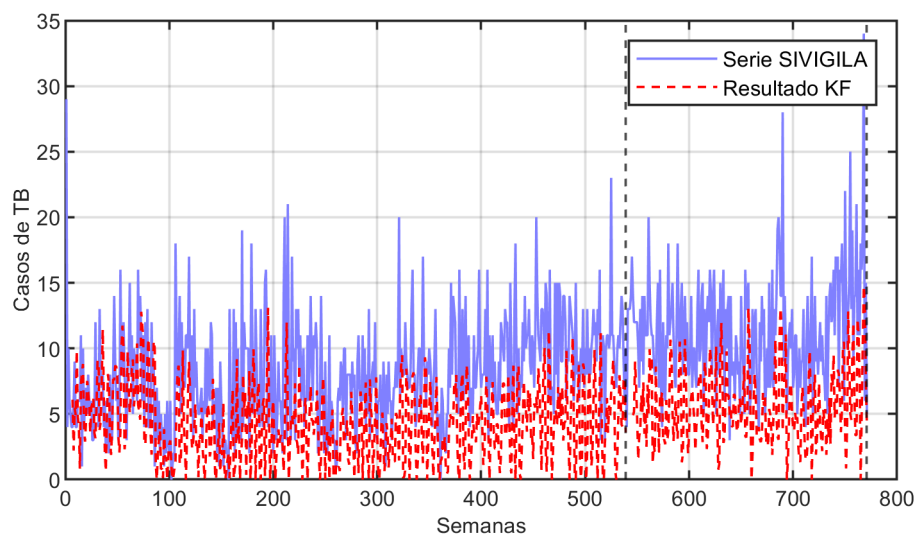


Figura 4-19: Resultados Filtro de Kalman Santander

En los resultados de entrenamiento se obtuvo para los índices de error RMSE, MAE y TRS: 6.1209, 4.8532 y 419.84, respectivamente. En los resultados de validación se obtuvo para los índices de error RMSE, MAE y TRS: 7.5434, 6.2034 y 210.43, respectivamente.

4.3.5. Resultados del Filtro de Kalman Robusto para Santander

En la Figura 4-20 se presentan los resultados que se obtuvieron al predecir los casos de TB, en pasos de siete semanas hacia adelante, con el Filtro de Kalman Robusto para Santander.

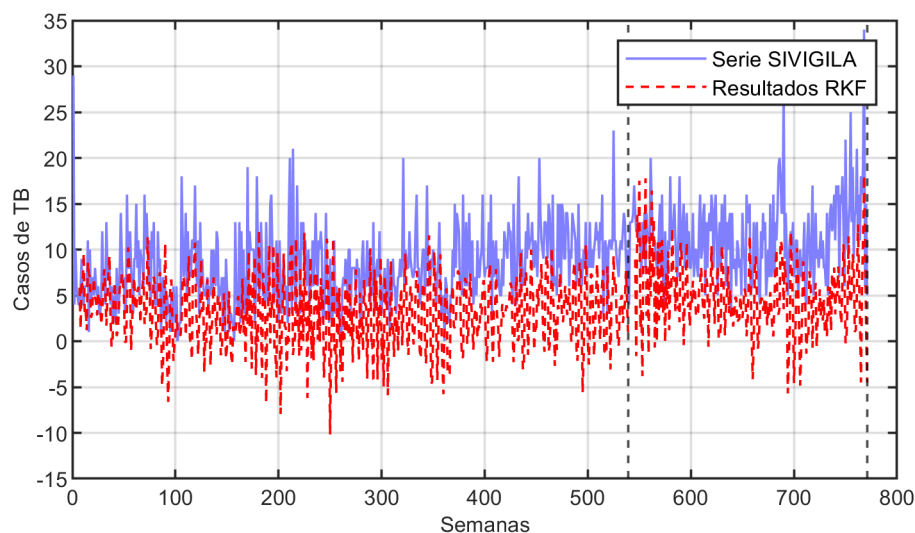


Figura 4-20: Resultados Filtro de Kalman Robusto Santander

En los resultados de entrenamiento se obtuvo para los índices de error RMSE, MAE y TRS:

7.252, 5.8401 y 453.23, respectivamente. En los resultados de validación se obtuvo para los índices de error RMSE, MAE y TRS: 8.4349, 6.9623 y 212.55, respectivamente.

4.3.6. Resultados de los algoritmos computacionales para Santander

Se realizó la comparativa de la serie de tiempo original con las cinco series de tiempo que arrojaron los algoritmos computacionales implementados para Santander. En la Figura 4-21, se observan los resultados de entrenamiento.

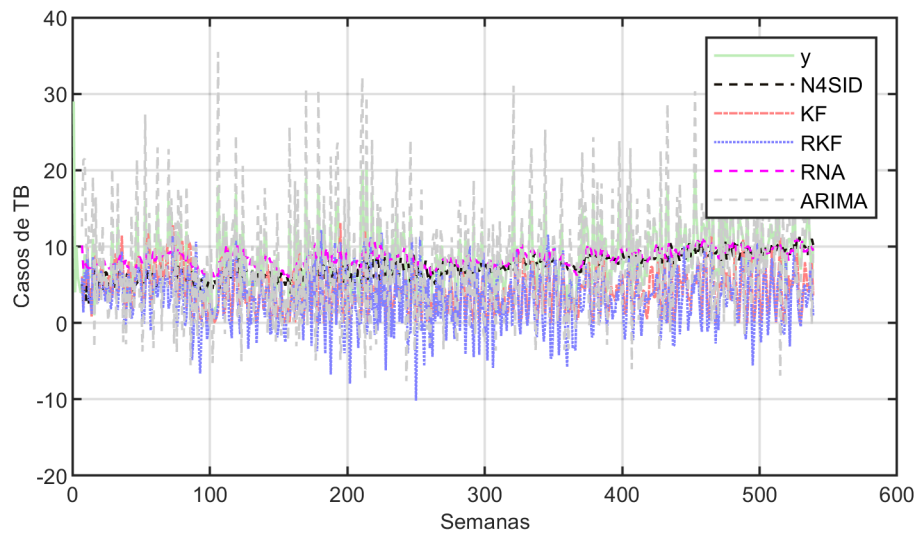


Figura 4-21: Resultados Entrenamiento Algoritmos Computacionales Santander

En la Figura 4-22, se observan los resultados de validación.

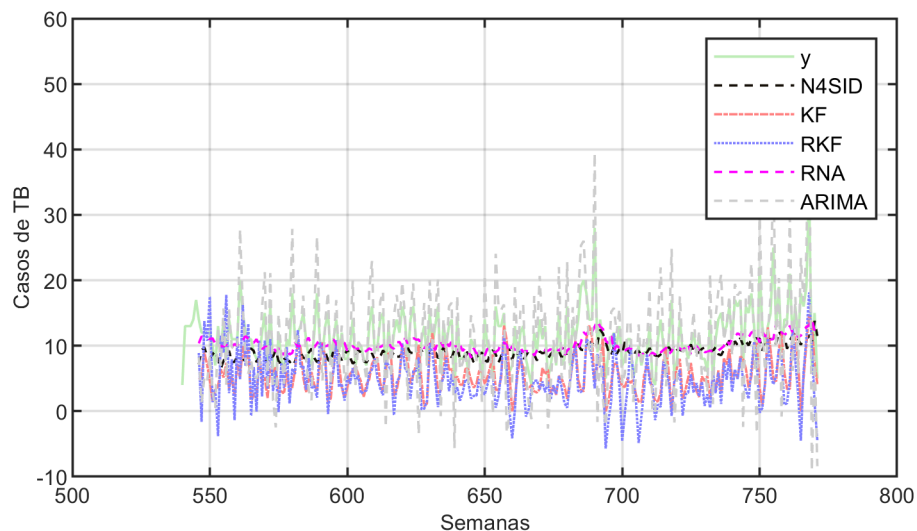


Figura 4-22: Resultados Validación Algoritmos Computacionales Santander

En la tabla 4-6, se presentan los resultados obtenidos con los índices de error RMSE, MAE y TRS para los cinco algoritmos implementados en el proceso de entrenamiento.

Tabla 4-6: Resultados de los algoritmos computacionales en el entrenamiento para el departamento de Santander

Algoritmo	RMSE	MAE	TRS
Modelo Autorregresivo N4SID	3.8268	2.9476	115.16
Filtro de Kalman	6.1209	4.8532	419.84
Red Neuronal Artificial	3.7024	2.899	-91.603
Modelo ARIMA	4.3309	3.3171	-15.929
Filtro de Kalman Robusto	7.252	5.8401	453.23

En la tabla 4-7, se presentan los resultados obtenidos con los índices de error RMSE, MAE y TRS para los cinco algoritmos implementados en el proceso de validación.

Tabla 4-7: Resultados de los algoritmos computacionales en la validación para el departamento de Santander

Algoritmo	RMSE	MAE	TRS
Modelo Autorregresivo N4SID	4.7213	3.6282	142.07
Filtro de Kalman	7.5434	6.2034	210.43
Red Neuronal Artificial	4.1629	3.2042	87.515
Modelo ARIMA	4.6335	3.6257	10.685
Filtro de Kalman Robusto	8.4349	6.9623	212.55

A pesar de que en Santander los resultados no fueron favorables para el Filtro de Kalman y el RKF debido a los cambios bruscos de la serie, podemos afirmar que se obtuvo un buen resultado con el modelo ARIMA, que procuró seguir la tendencia en los picos y mantenerse en la media.

4.4. Bogotá

4.4.1. Resultados de la RNA para Bogotá

En la Figura 4-23 se presentan los resultados que se obtuvieron al predecir los casos de TB, en pasos de siete semanas hacia adelante, con la RNA para Bogotá.

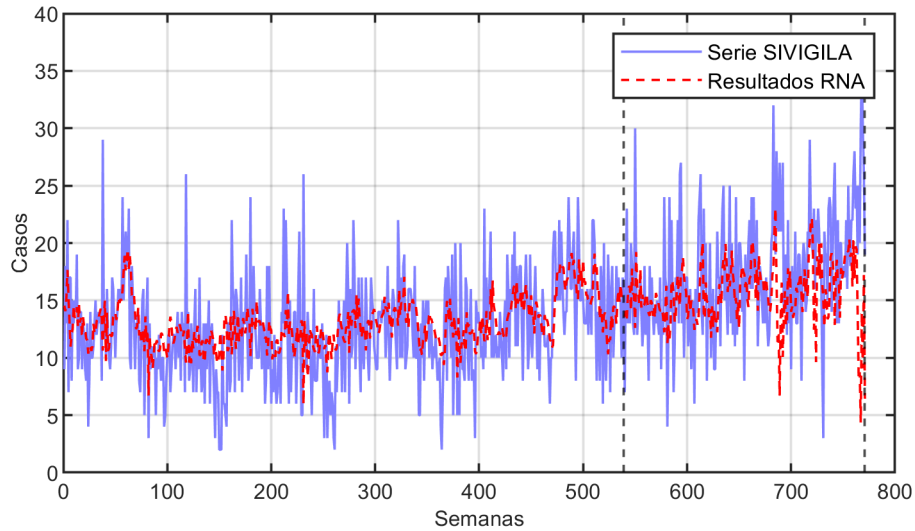


Figura 4-23: Resultados RNA Bogotá

En los resultados de entrenamiento se obtuvo para los índices de error RMSE, MAE y TRS: 4.1314, 3.2041 y -106.79, respectivamente. En los resultados de validación se obtuvo para los índices de error RMSE, MAE y TRS: 6.0637, 4.3623 y 89.621, respectivamente.

4.4.2. Resultados del modelo ARIMA para Bogotá

En la Figura 4-24 se presentan los resultados que se obtuvieron al predecir los casos de TB, en pasos de siete semanas hacia adelante, con el modelo ARIMA para Bogotá.

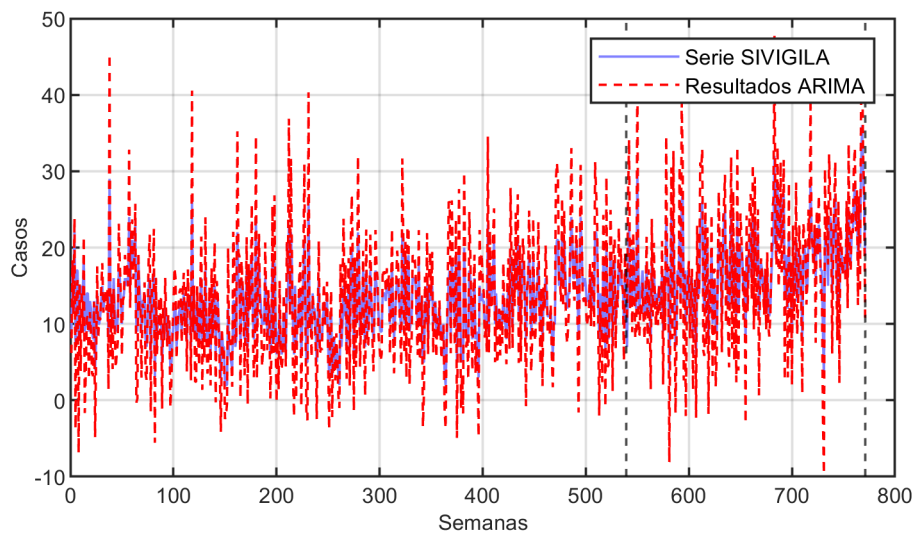


Figura 4-24: Resultados modelo ARIMA Bogotá

En los resultados de entrenamiento se obtuvo para los índices de error RMSE, MAE y TRS:

4.3927, 3.4259 y 18.978, respectivamente. En los resultados de validación se obtuvo para los índices de error RMSE, MAE y TRS: 5.3206, 4.1839 y 6.1169, respectivamente.

4.4.3. Resultados del modelo N4SID para Bogotá

En la Figura 4-25 se presentan los resultados que se obtuvieron al predecir los casos de TB, en pasos de siete semanas hacia adelante, con el modelo autorregresivo N4SID para Bogotá.

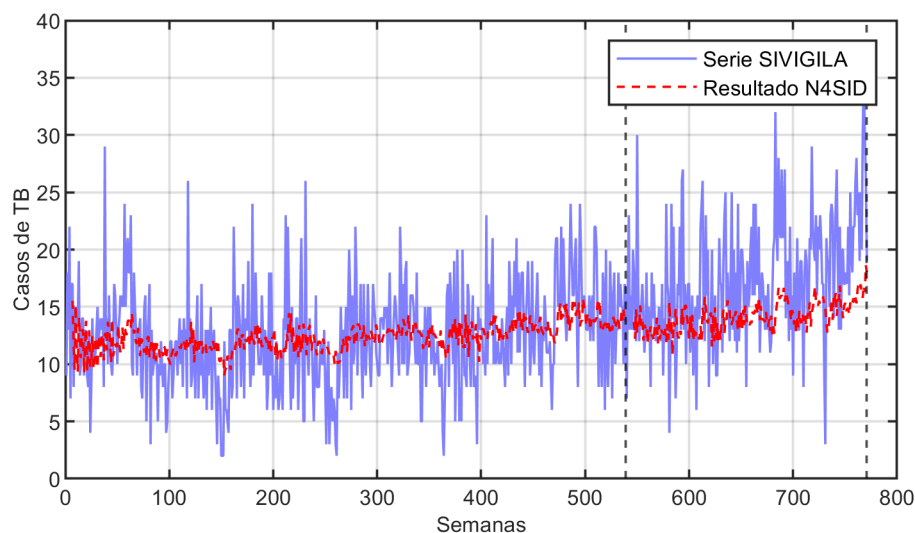


Figura 4-25: Resultados modelo autorregresivo N4SID Bogotá

En los resultados de entrenamiento se obtuvo para los índices de error RMSE, MAE y TRS: 4.2056, 3.3077 y -7.5275, respectivamente. En los resultados de validación se obtuvo para los índices de error RMSE, MAE y TRS: 5.9513, 4.6293 y 157.8, respectivamente.

4.4.4. Resultados del Filtro de Kalman para Bogotá

En la Figura 4-26 se presentan los resultados que se obtuvieron al predecir los casos de TB, en pasos de siete semanas hacia adelante, con el Filtro de Kalman para Bogotá.

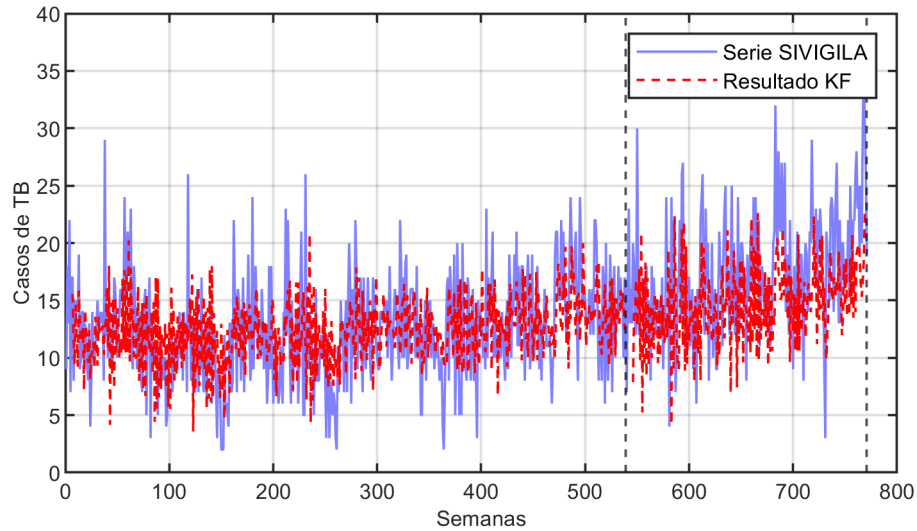


Figura 4-26: Resultados Filtro de Kalman Bogotá

En los resultados de entrenamiento se obtuvo para los índices de error RMSE, MAE y TRS: 4.8901, 3.876 y -4.2587, respectivamente. En los resultados de validación se obtuvo para los índices de error RMSE, MAE y TRS: 6.4364, 5.2929 y 112.2, respectivamente.

4.4.5. Resultados del Filtro de Kalman Robusto para Bogotá

En la Figura 4-27 se presentan los resultados que se obtuvieron al predecir los casos de TB, en pasos de siete semanas hacia adelante, con el Filtro de Kalman Robusto para Bogotá.

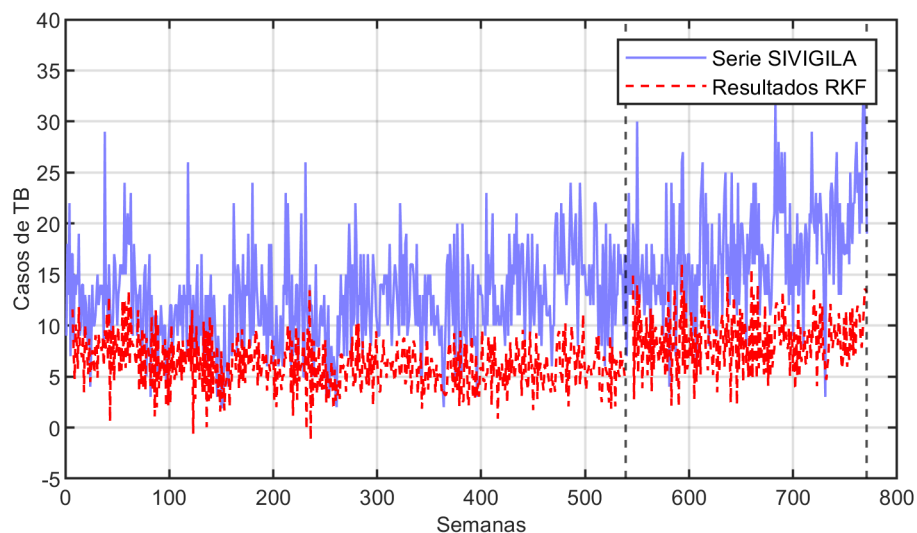


Figura 4-27: Resultados Filtro de Kalman Robusto Bogotá

En los resultados de entrenamiento se obtuvo para los índices de error RMSE, MAE y TRS: 7.7196, 6.4966 y 508.14, respectivamente. En los resultados de validación se obtuvo para los

índices de error RMSE, MAE y TRS: 10.311, 8.9854 y 221.13, respectivamente. El resultado publicado en la Figura 4-27 corresponde a una varianza de 20 casos de TB. Es importante destacar que el resultado no fue el esperado, dado que la señal obtenida no se acerca a la media ni está en las vecindades de la serie de tiempo original. Obtuvimos mejores resultados al reducir la covarianza $Q = 4$ y las semanas que se logran predecir $m = 6$, como se observa en la Figura 4-28.

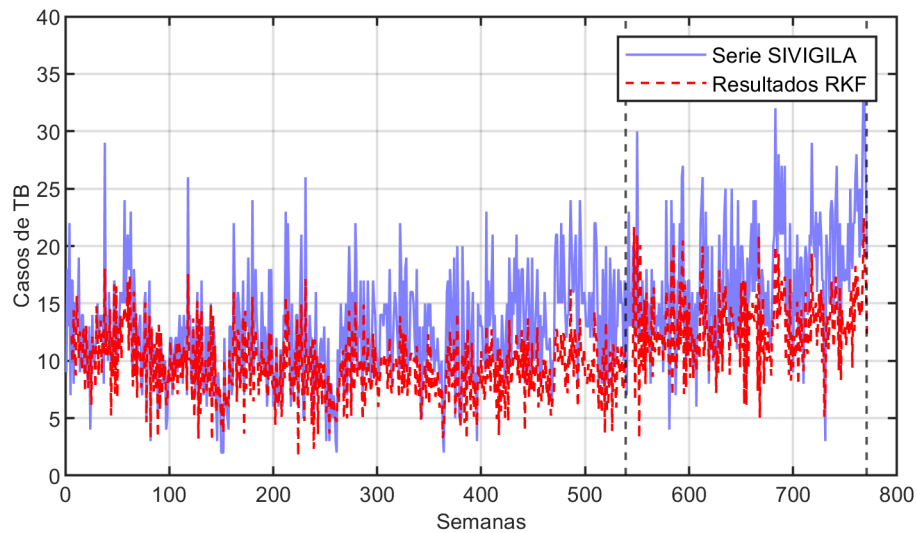


Figura 4-28: Resultados Filtro de Kalman Robusto Bogotá

4.4.6. Resultados de los algoritmos computacionales para Bogotá

Se realizó la comparativa de la serie de tiempo original con las cinco series de tiempo que arrojaron los algoritmos computacionales implementados para Bogotá. En la Figura 4-29, se observan los resultados de entrenamiento.

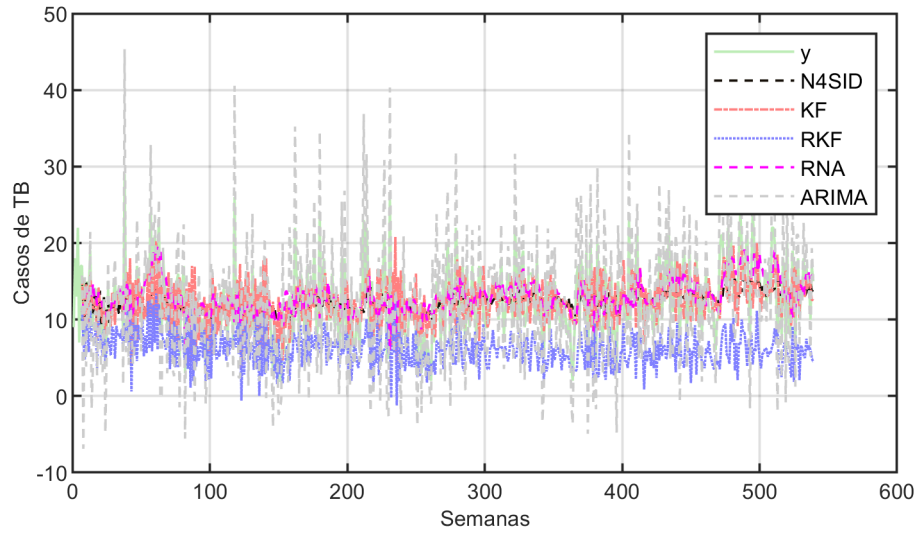


Figura 4-29: Resultados Entrenamiento Algoritmos Computacionales Bogotá

En la Figura 4-30, se observan los resultados de validación.

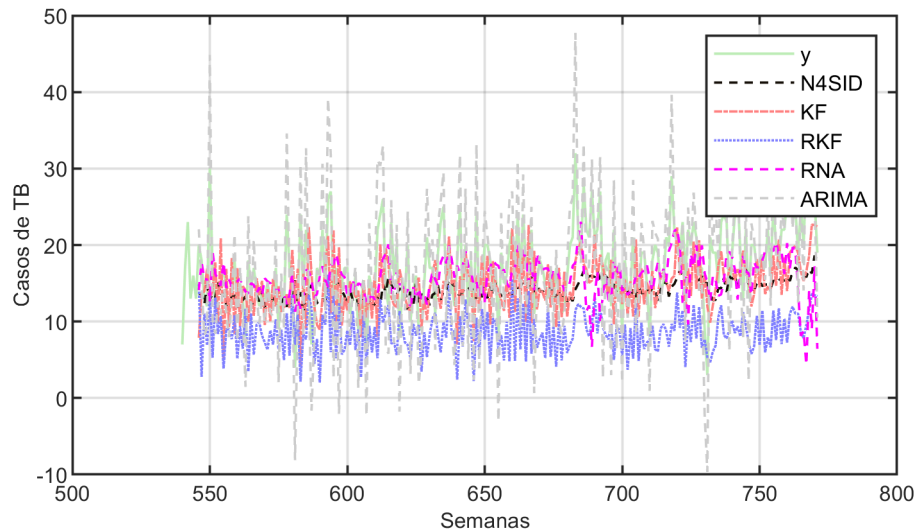


Figura 4-30: Resultados Validación Algoritmos Computacionales Bogotá

En la tabla 4-8, se presentan los resultados obtenidos con los índices de error RMSE, MAE y TRS para los cinco algoritmos implementados en el proceso de entrenamiento.

Tabla 4-8: Resultados de los algoritmos computacionales en el entrenamiento para Bogotá

Algoritmo	RMSE	MAE	TRS
Modelo Autorregresivo N4SID	4.2056	3.3077	-7.5275
Filtro de Kalman	4.8901	3.876	-4.2587
Red Neuronal Artificial	4.1314	3.2041	-106.79
Modelo ARIMA	4.3927	3.4259	18.978
Filtro de Kalman Robusto	7.7196	6.4966	508.14

En la tabla 4-9, se presentan los resultados obtenidos con los índices de error RMSE, MAE y TRS para los cinco algoritmos implementados en el proceso de validación.

Tabla 4-9: Resultados de los algoritmos computacionales en la validación para Bogotá

Algoritmo	RMSE	MAE	TRS
Modelo Autorregresivo N4SID	5.9513	4.6293	157.8
Filtro de Kalman	6.4364	5.2929	112.2
Red Neuronal Artificial	6.0637	4.3623	89.621
Modelo ARIMA	5.3206	4.1839	6.1169
Filtro de Kalman Robusto	10.311	8.9854	221.13

Para el caso de Bogotá, se logró mayor exactitud con el Filtro de Kalman, donde la serie obtenida tuvo un desempeño satisfactorio y muy superior a los demás algoritmos. Por lo tanto consideramos que es un algoritmo que se puede utilizar para la predicción de series de tiempo en Bogotá.

4.5. Meta

4.5.1. Resultados de la RNA para Meta

En la Figura 4-31 se presentan los resultados que se obtuvieron al predecir los casos de TB, en pasos de siete semanas hacia adelante, con la RNA para Meta.

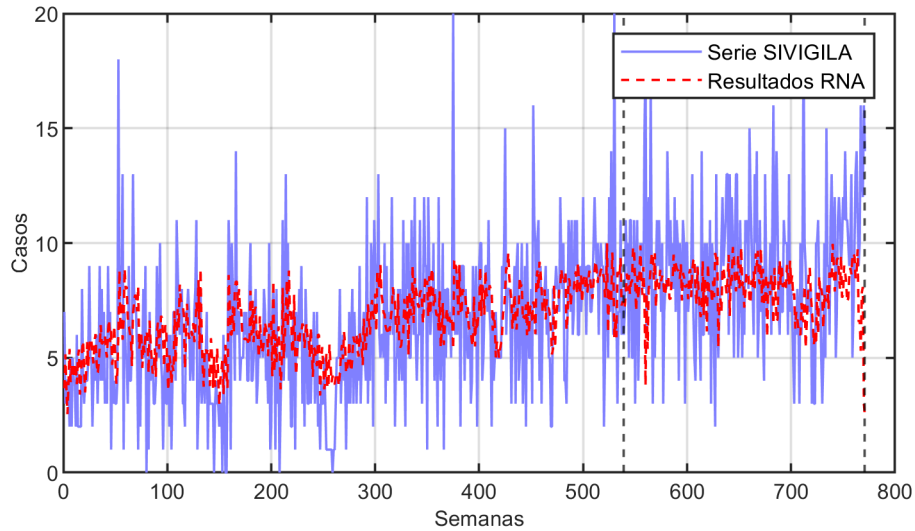


Figura 4-31: Resultados RNA Meta

En los resultados de entrenamiento se obtuvo para los índices de error RMSE, MAE y TRS: 2.7894, 2.1519 y -104.44, respectivamente. En los resultados de validación se obtuvo para los índices de error RMSE, MAE y TRS: 3.5958, 2.7384 y 74.712, respectivamente.

4.5.2. Resultados del modelo ARIMA para Meta

En la Figura 4-32 se presentan los resultados que se obtuvieron al predecir los casos de TB, en pasos de siete semanas hacia adelante, con el modelo ARIMA para Meta.

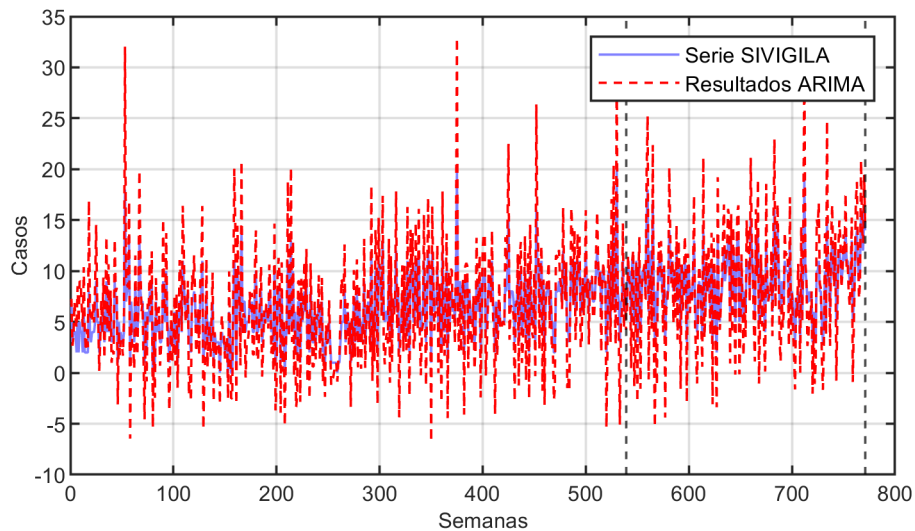


Figura 4-32: Resultados modelo ARIMA Meta

En los resultados de entrenamiento se obtuvo para los índices de error RMSE, MAE y TRS:

3.0088, 2.3327 y -17.354, respectivamente. En los resultados de validación se obtuvo para los índices de error RMSE, MAE y TRS: 3.2213, 2.4878 y 13.094, respectivamente.

4.5.3. Resultados del modelo N4SID para Meta

En la Figura 4-33 se presentan los resultados que se obtuvieron al predecir los casos de TB, en pasos de siete semanas hacia adelante, con el modelo autorregresivo N4SID para Meta.

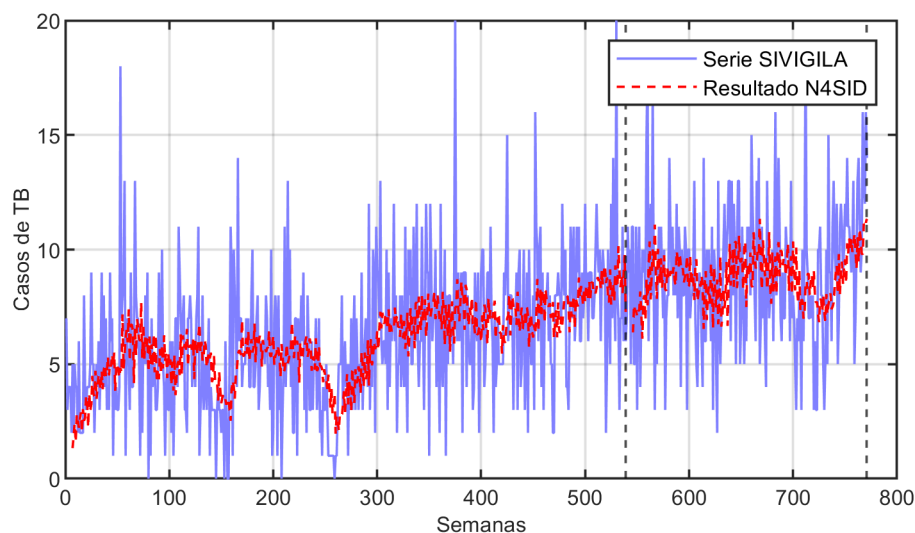


Figura 4-33: Resultados modelo autorregresivo N4SID Meta

En los resultados de entrenamiento se obtuvo para los índices de error RMSE, MAE y TRS: 2.8681, 2.1919 y 16.389, respectivamente. En los resultados de validación se obtuvo para los índices de error RMSE, MAE y TRS: 3.172, 2.4877 y 11.924, respectivamente.

4.5.4. Resultados del Filtro de Kalman para Meta

En la Figura 4-34 se presentan los resultados que se obtuvieron al predecir los casos de TB, en pasos de siete semanas hacia adelante, con el Filtro de Kalman para Meta.

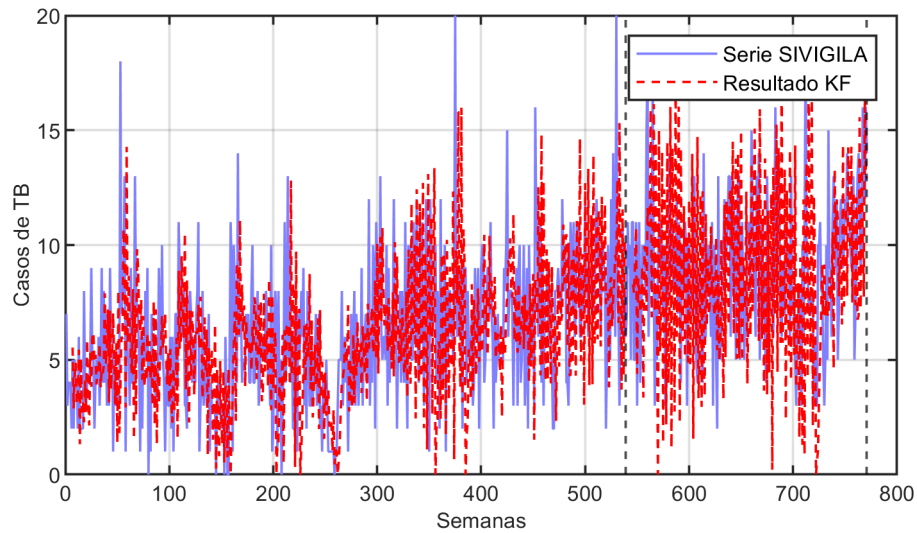


Figura 4-34: Resultados Filtro de Kalman Meta

En los resultados de entrenamiento se obtuvo para los índices de error RMSE, MAE y TRS: 3.5627, 2.7449 y 18.757, respectivamente. En los resultados de validación se obtuvo para los índices de error RMSE, MAE y TRS: 5.0851, 4.2206 y 3.4606, respectivamente.

4.5.5. Resultados del Filtro de Kalman Robusto para Meta

En la Figura 4-35 se presentan los resultados que se obtuvieron al predecir los casos de TB, en pasos de siete semanas hacia adelante, con el Filtro de Kalman Robusto para Meta.

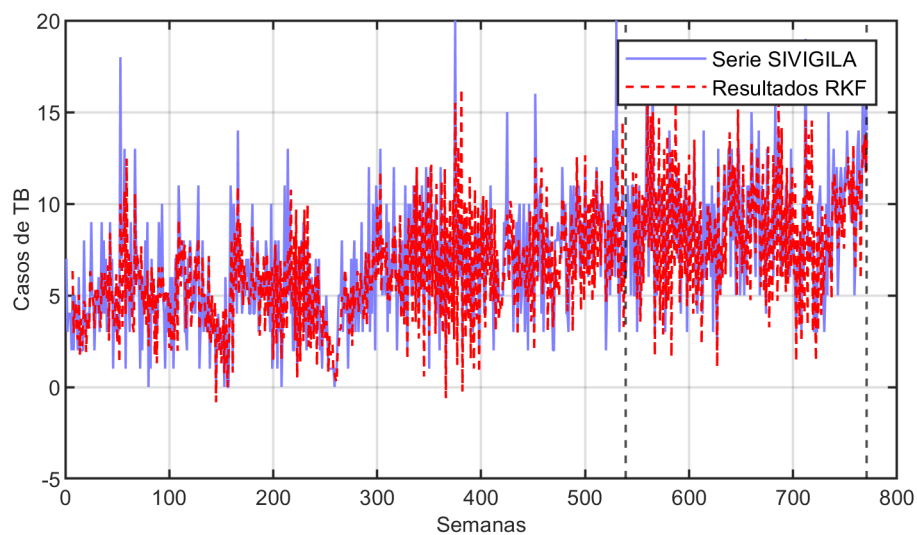


Figura 4-35: Resultados Filtro de Kalman Robusto Meta

En los resultados de entrenamiento se obtuvo para los índices de error RMSE, MAE y TRS:

2.5783, 2.0125 y 46.307, respectivamente. En los resultados de validación se obtuvo para los índices de error RMSE, MAE y TRS: 3.3013, 2.7112 y 22.33, respectivamente.

4.5.6. Resultados de los algoritmos computacionales para Meta

Se realizó la comparativa de la serie de tiempo original con las cinco series de tiempo que arrojaron los algoritmos computacionales implementados para Meta. En la Figura 4-36, se observan los resultados de entrenamiento.

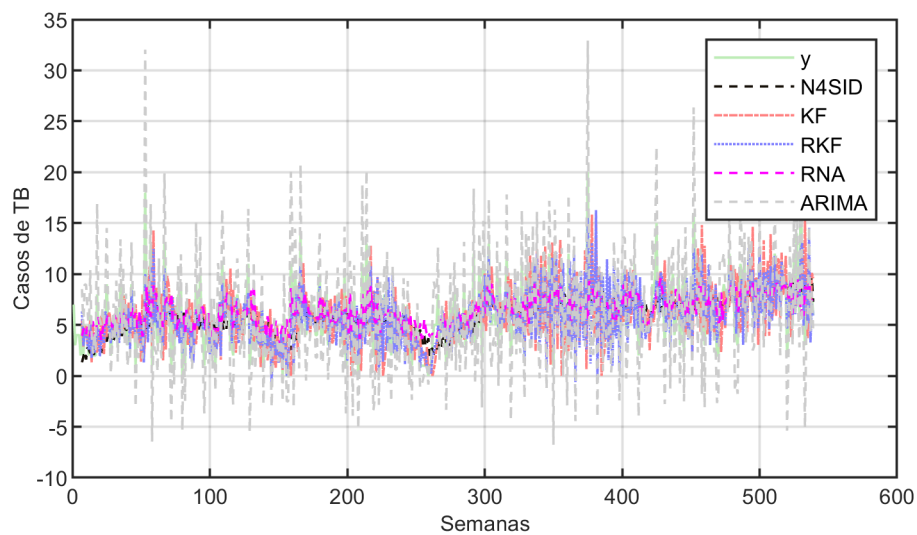


Figura 4-36: Resultados Entrenamiento Algoritmos Computacionales Meta

En la Figura 4-37, se observan los resultados de validación.

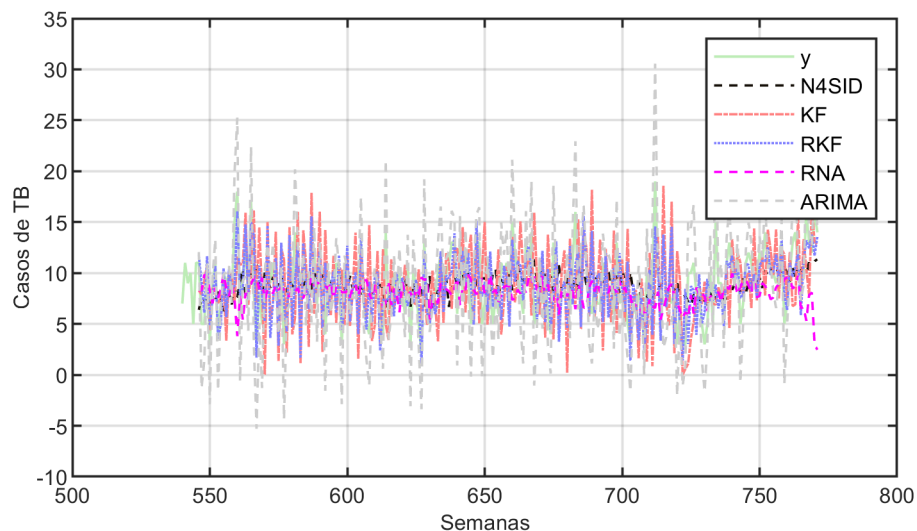


Figura 4-37: Resultados Validación Algoritmos Computacionales Meta

En la tabla 4-10, se presentan los resultados obtenidos con los índices de error RMSE, MAE y TRS para los cinco algoritmos implementados en el proceso de entrenamiento.

Tabla 4-10: Resultados de los algoritmos computacionales en el entrenamiento para el departamento de Meta

Algoritmo	RMSE	MAE	TRS
Modelo Autorregresivo N4SID	2.8681	2.1919	16.389
Filtro de Kalman	3.5627	2.7449	18.757
Red Neuronal Artificial	2.7894	2.1519	-104.44
Modelo ARIMA	3.0088	2.3327	-17.354
Filtro de Kalman Robusto	2.5783	2.0125	46.307

En la tabla 4-11, se presentan los resultados obtenidos con los índices de error RMSE, MAE y TRS para los cinco algoritmos implementados en el proceso de validación.

Tabla 4-11: Resultados de los algoritmos computacionales en la validación para el departamento de Meta

Algoritmo	RMSE	MAE	TRS
Modelo Autorregresivo N4SID	3.172	2.4877	11.924
Filtro de Kalman	5.0851	4.2206	3.4606
Red Neuronal Artificial	3.5958	2.7384	74.712
Modelo ARIMA	3.2213	2.4878	13.094
Filtro de Kalman Robusto	3.3013	2.7112	22.33

Aunque los resultados obtenidos con la RNA y el Modelo N4SID para Meta, siempre se mantuvieron en la media, se observa que se logra seguir la tendencia de la serie. A diferencia del KF y RKF que presentaron datos muy dispersos y con tendencia a los picos. Se recomienda el uso del modelo ARIMA para la predicción de series de tiempo en la Región de la Orinoquía.

4.6. Análisis de los Resultados

En la Figura 4-38 se presentan los resultados obtenidos con el índice MAE en el proceso de validación con los algoritmos computacionales, cada uno de los valores representados por β corresponden a los departamentos de Colombia, la lógica de organización se observa en la Tabla 4-12.

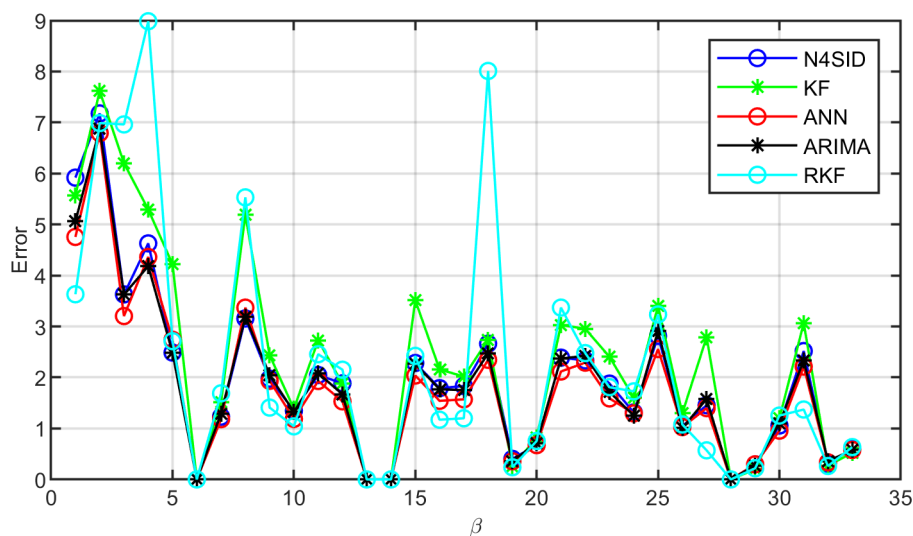


Figura 4-38: Resultados obtenidos con el índice MAE

Aunque en la Figura 4-38 los resultados con menor error absoluto medio son los más próximos a cero, podemos asegurar que este índice por si solo no es un factor definitivo para afirmar que se tuvo una mejor respuesta en un departamento en particular.

En la Figura 4-39 se presentan los resultados obtenidos con el índice RMSE en el proceso de validación con los algoritmos computacionales, cada uno de los valores representados por β corresponden a los departamentos de Colombia, la lógica de organización se observa en la Tabla 4-12.

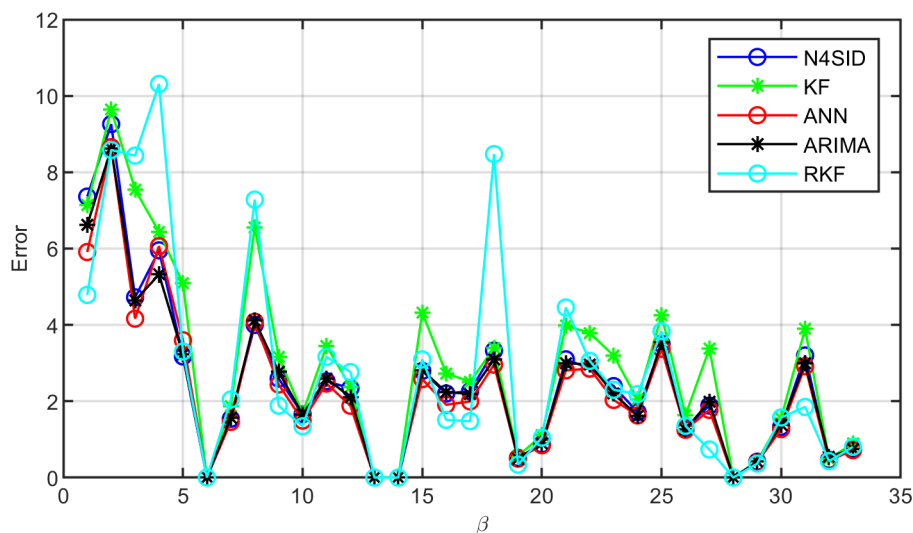


Figura 4-39: Resultados obtenidos con el índice RMSE

Al analizar a detalle las Figuras 4-38 y 4-39 se observa que los datos son muy similares y quizá a simple vista el único factor que nos permite asegurar que son dos gráficos diferentes

es la escala en el eje Y (Error). Esto se debe a que los índices MAE y RMSE son muy parecidos, sin embargo, el MAE es menos sensible a valores extremos en los datos, ya que trata los errores de manera lineal. Consideramos que esta puede ser una de las razones por las cuales se obtiene un menor error con el MAE que con el RMSE.

En la Figura 4-40 se presentan los resultados obtenidos con el índice TRS en el proceso de validación con los algoritmos computacionales, cada uno de los valores representados por β corresponden a los departamentos de Colombia, la lógica de organización se observa en la Tabla 4-12.

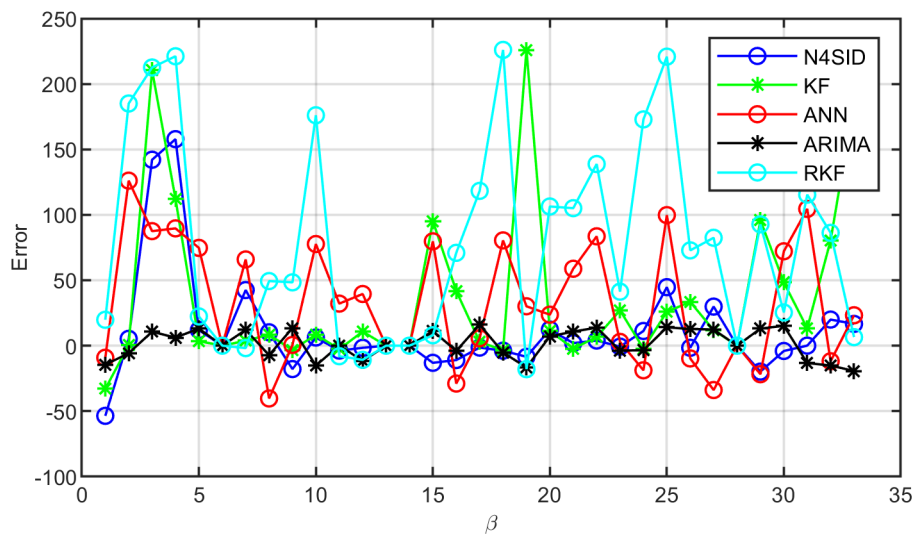


Figura 4-40: Resultados obtenidos con el índice TRS

Aunque la respuesta obtenida con el MAE, RMSE y TRS para algunos departamentos, no fue la esperada, podemos garantizar que se logró el mejor resultado para el departamento de Valle del Cauca. Donde la señal que se obtuvo en cada uno de los algoritmos es lo más cercano posible en comparación a la señal original.

En la Tabla 4-12, se detalla la organización que se tuvo en cuenta para representar las Figuras 4-38, 4-39 y 4-40.

Tabla 4-12: Ubicación de los departamentos en las gráficas de los índices de evaluación

Departamento	β
Valle del Cauca	1
Antioquia	2
Santander	3
Bogotá	4
Meta	5
Amazonas	6
Arauca	7
Atlántico	8
Bolívar	9
Boyacá	10
Caldas	11
Caquetá	12
Casanare	13
Cauca	14
Cesar	15
Chocó	16
Córdoba	17
Cundinamarca	18
Guainía	19
Guaviare	20
Huila	21
La Guajira	22
Magdalena	23
Nariño	24
Norte de Santander	25
Putumayo	26
Quindío	27
Risaralda	28
San Andrés y Providencia	29
Sucre	30
Tolima	31
Vaupés	32
Vichada	33

En la Figura 4-41 se presentan los resultados obtenidos con el índice RMSE. Esta representación nos permite obtener una estadística visual de la respuesta de los algoritmos compu-

tacionales implementados. Observe que los resultados medios son similares y se mantienen en un rango entre 2 y 3. Los algoritmos ANN y RKF, fueron los que obtuvieron un mejor resultado con este índice de evaluación.

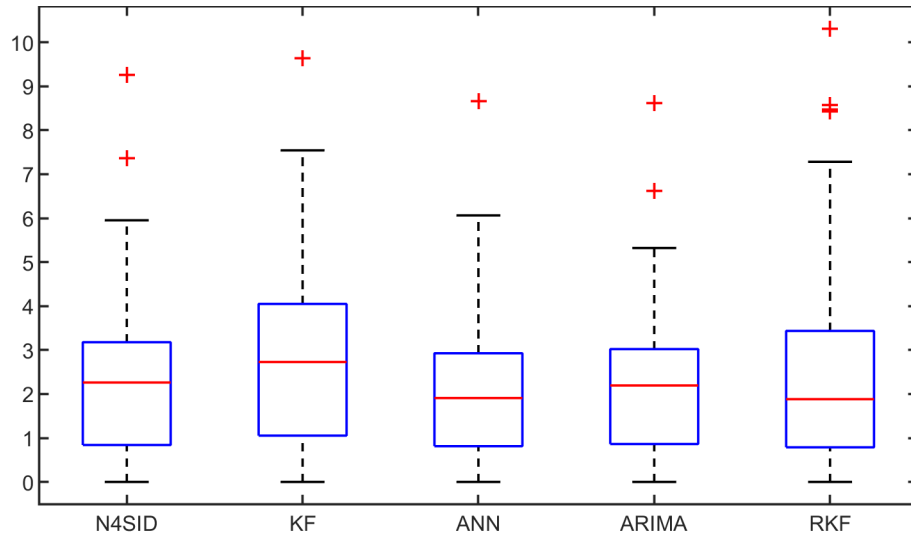


Figura 4-41: Resultados Globales obtenidos con el índice RMSE en el proceso de validación

En la Figura 4-42 se presentan los resultados obtenidos con el índice MAE. Observe que los resultados medios son similares y se mantienen en un rango entre 1 y 2. Al igual que con el índice RMSE, los algoritmos ANN y RKF, fueron los que obtuvieron un mejor resultado con el índice MAE.

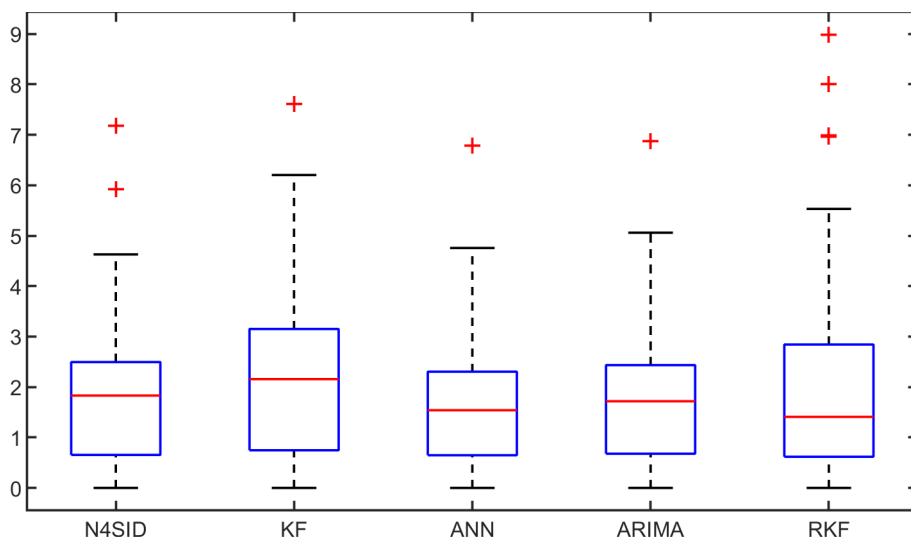


Figura 4-42: Resultados Globales obtenidos con el índice MAE en el proceso de validación

En la Figura 4-43 se presenta a través de un diagrama de caja, los resultados obtenidos con el índice TRS para los algoritmos computacionales implementados. Se puede observar que los

algoritmos N4SID y ARIMA presentaron un mejor desempeño en este índice. Es importante aclarar que el propósito es cuantificar que tan distantes están los datos obtenidos con relación a los picos de la señal original.

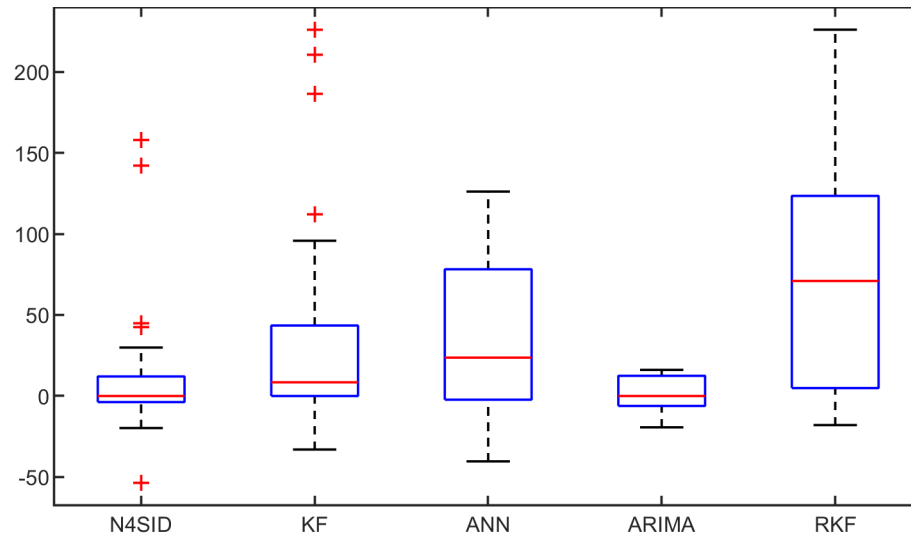


Figura 4-43: Resultados Globales obtenidos con el índice TRS en el proceso de validación

5 Conclusiones

Podemos afirmar con certeza que se ha logrado el propósito de predecir los casos de TB en más de 4 semanas hacia el futuro. Sin embargo, debido a la gran dispersión de los datos de las series temporales de casos de TB reportados por el SIVIGILA en Colombia, la información obtenida con los algoritmos suele mantenerse en la media y en algunos casos la señal obtenida presenta mucho ruido y se pierde en los picos.

Consideramos que los resultados obtenidos son favorables y si se pueden implementar modelos computacionales para la predicción de TB en Colombia, teniendo en cuenta algunas recomendaciones. Aunque se utilice la misma serie de datos para cada uno de los algoritmos computacionales, se involucran parámetros y variables exógenas que permiten tener un mejor desempeño en un periodo de tiempo determinado. Por lo tanto, es necesario evaluar cada algoritmo individualmente para obtener el mejor resultado con el fin de determinar el número de semanas que se pueden predecir con el menor porcentaje de error.

Consolidando el trabajo realizado, se presentan los puntos claves de esta investigación:

- Se generaron 33 bases de datos con los casos de TB en Colombia desde 2007 a 2021, la información se encuentra en el repositorio indicado en la sección 3.1
- Los algoritmos computacionales implementados (RNA, ARIMA, N4SID, KF y RKF) nos permitieron predecir la tendencia de la serie con relación a la media. En algunos casos logramos predecir el comportamiento de los picos, en los periodos donde se presentó un número muy bajo o muy alto de casos de TB, sin embargo los modelos no son exactos dados los cambios bruscos en la serie.
- Para calcular el desempeño de los algoritmos, implementamos el Error Absoluto Medio (MAE), el Error Cuadrático Medio (RMSE) y el Índice de Seguimiento de la Señal (TRS). Lo que nos permitió evaluar la precisión y exactitud de la serie obtenida con relación a la serie real.
- Luego de realizar pruebas exhaustivas para determinar el número máximo de semanas que se logran predecir los casos de TB, se llegó a la conclusión que los algoritmos presentan resultados diferentes cada vez que se compila el código y depende de los valores asignados a cada una de las variables arbitrarias utilizadas. Cabe resaltar que el desempeño de la computadora utilizada para este procesamiento de datos también juega un papel determinante toda vez que la ejecución del programa puede tardar más de 45 minutos para cada departamento en cada intento realizado.

El trabajo se continuará hasta conseguir mejores resultados con los índices de evaluación para los algoritmos implementados, esto nos permitirá determinar el número de semanas en las cuales la predicción es precisa y exacta. A futuro se trabajará con otro tipo de métodos computacionales como las redes LSTM, Redes Convolucionales (CNN) y Redes Neuronales de Retropropagación (BPNN).

Bibliografía

- [1] World Health Organization. *Global tuberculosis report 2020*. Geneva: World Health Organization, 2020.
- [2] Anastasia Koch and Valerie Mizrahi. Mycobacterium tuberculosis. *Trends in Microbiology*, 26(6):555–556, 2018.
- [3] Santiago Fadul-Pérez, Lisette Andrea Bermudez-Pinzon, Claudia Llerena-Polo, Martha Lucía Ospina-Martínez, Franklyn Edwin Prieto-Alvarado, Diana Marcela Walteros-Acero, and Hernán Quijada-Bonilla. *Protocolo de vigilancia en salud pública - Tuberculosis Farmacoresistente*. 2020.
- [4] Martha Lucía Ospina-Martínez, Carolina Flórez, Pilar Zambrano, Claudia Llerena-Polo, and Angie Paola Zabaleta-Vanegas. *Guia para la vigilancia por laboratorio de Tuberculosis*. 2020.
- [5] Ivan Camilo Sanchez-Vega, Alvaro David Orjuela-Cañon, Carlos Enrique Awad Garcia, Erika Vergara, Maria Angelica Palencia, and Andres L. Jutinico. Towards tuberculosis incidence time series forecasting in colombian regions: The antioquia case. *CS IT - CSCP*, pages 117–126, 11 2022.
- [6] Astrid López-Arias. ¿qué son y por qué se hacen las pruebas de tamizaje? *Facultad de Salud - Grupo de Comunicaciones, Universidad del Valle*, 2015.
- [7] World Health Organization. *Global tuberculosis report 2021*. Geneva: World Health Organization, 2021.
- [8] Matteo Zignol, Andrea Maurizio Cabibbe, and Anna S. Dean. Genetic sequencing for surveillance of drug resistance in tuberculosis in highly endemic countries; a multi-country population-based surveillance study. *Lancet Infectious Diseases*, 18(6):675–683, 2018. NPP.
- [9] Martha Luica Ospina-Martinez, Frnaklyn Edwin Prieto-Alvarado, Diana Walteros, and Hernan Quijada-Bonilla. *Comportamiento de la Vigilancia de Tuberculosis, Colombia, 2020*. Instituto Nacional de Salud, 2021.
- [10] Colombia, Ministerio de salud y protección social. Lineamiento técnico operativo del programa Nacional Tuberculosis. Dirección de promoción y Prevención. Subdirección

de enfermedades Transmisibles. *Resolución número 227 del 20 de febrero de 2020*. Colombia, Ministerio de salud y protección social, 2020.

- [11] Chu-Chang Ku and Peter J Dodd. Forecasting the impact of population ageing on tuberculosis incidence. *PLOS ONE*, 14(9):1–13, September 2019.
- [12] Yanling Zheng, Xueliang Zhang, Xijiang Wang, Kai Wang, and Yan Cui. Predictive study of tuberculosis incidence by time series method and elman neural network in kashgar, china. *BMJ Open*, 11(1), 2021.
- [13] Ji Y Martinez L Zia UH Javaid A Lu W Wang J. Liu Q, Li Z. Forecasting the seasonality and trend of pulmonary tuberculosis in jiangsu province of china using advanced statistical time-series analyses. *Infection and Drug Resistance*, 12:2311–2322, 07 2019.
- [14] Colombia, Ministerio de salud y protección social. *Circular 007 del 26 de febrero de 2015*. Colombia, Ministerio de salud y protección social, 2015.
- [15] Afranio Kritski, Margareth Pretti Dalcolmo, Fernanda Carvalho Queiroz Mello, Anna Cristina Calçada Carvalho, Denise Rossato Silva, Martha Maria de Oliveira, and Julio Croda. The role of the brazilian tuberculosis research network in national and international efforts to eliminate tuberculosis. *Jornal Brasileiro de Pneumologia*, 44(J. bras. pneumol., 2018 44(2)):77–81, Mar 2018.
- [16] K. W. WANG, C. DENG, J. P. LI, Y. Y. ZHANG, X. Y. LI, and M. C. WU. Hybrid methodology for tuberculosis incidence time-series forecasting based on arima and a nar neural network. *Epidemiology amp; Infection*, 145(6):1118–1129, 2017.
- [17] Matthew Kavanagh, Lawrence O. Gostin, and John Stephens. Tuberculosis, human rights, and law reform: Addressing the lack of progress in the global tuberculosis response. *PLoS Medicine*, 17(10):1–9, 2020.
- [18] Q Liu, Z Li, Y J, L Martinez, UH Zi, A Javaid, W Lu, and J Wang. Forecasting the seasonality and trend of pulmonary tuberculosis in jiangsu province of china using advanced statistical time-series analyses. *Infection and Drug Resistance*, 12:2311–2322, 2019.
- [19] KW Wang, C Deng, JP Li, YY Zhang, XY L, and MC Wu. Hybrid methodology for tuberculosis incidence time-series forecasting based on ARIMA and a NAR neural network. *Epidemiol Infect*, 145(6):118–1129, 2017.
- [20] A Azeez, D Obaromi, A Odeyemi, J Ndege, and R Muntabayi. Seasonality and trend forecasting of tuberculosis prevalence data in eastern cape, south africa, using a hybrid model. *International journal of environmental research and public health*, 13(8):1–13, 2016.

- [21] K.E. ArunKumar, Dinesh V. Kalaga, Ch. Mohan Sai Kumar, Govinda Chilkoor, Masahiro Kawaji, and Timothy M. Brenza. Forecasting the dynamics of cumulative covid-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: Auto-regressive integrated moving average (ARIMA) and seasonal auto-regressive integrated moving average (SARIMA). *Applied Soft Computing*, 103:107–161, 2021.
- [22] Vincent C.C. Cheng, Jasper F.W. Chan, Kelvin K.W. To, and K.Y. Yuen. Clinical management and infection control of sars: Lessons learned. *Antiviral Research*, 100(2):407–419, 2013.
- [23] Jonathan Golub, C Mohan, G Comstock, and R Chaisson. Active case finding of tuberculosis: Historical perspective and future prospects. *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease*, 9:1183–203, 12 2005.
- [24] Sapna Bamrah, R. Woodruff, K Powell, Smita Ghosh, Steve Kammerer, and M Haddad. Tuberculosis among the homeless, united states, 1994-2010. *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease*, 17:1414–1419, 11 2013.
- [25] Sagun Paudel, Siwi Padmawati, Ashmita Ghimire, Choden Yonzon, and Yodi Mahendradhata. Feasibility of find cases actively, separate safely and treat effectively (fast) strategy for early diagnosis of tb in nepal: An implementation research. *PLOS ONE*, 16:e0258883, 10 2021.
- [26] DA Martinez-Bello, A Lopez-Quilez, and A Torres-Prieto. Bayesian dynamic modeling of time series of dengue disease case counts. *PLoS Negl Trop Dis*, 11(7):1–19, 2017.
- [27] G Chowell, D Hincapie-Palacio, J Ospina, B Pell, A Tariq, S Daha, S Moghadas, A Smirnova, L Simonsen, and C Viboud. Using phenomenological models to characterize transmissibility and forecast patterns and final burden of zika epidemics. *PLoS Curr*, 8:1–21, 2016.
- [28] Katrina Ortblad, Rafael Lozano, and Christopher Murray. An alternative estimation of tuberculosis incidence from 1980 to 2010: methods from the global burden of disease 2010. *The Lancet*, 381:S104, 06 2013.
- [29] J. Almeida Santos, Andreia Leite, Patricia Soares, Raquel Duarte, and Carla Nunes. Delayed diagnosis of active pulmonary tuberculosis - potential risk factors for patient and healthcare delays in portugal. *BMC Public Health*, 21, 11 2021.
- [30] Fentabil Getnet, Meaza Demissie, Alemayehu Worku, Tesfaye Gobena, Rea Tschopp, Michael Girmachew, Gebeyehu Mitku, and Berhanu Seyoum. Delay in diagnosis of

pulmonary tuberculosis increases the risk of pulmonary cavitation in pastoralist setting of ethiopia. *BMC Pulmonary Medicine*, 19, 11 2019.

- [31] Courtney Yuen, Farhana Amanullah, Ashwin Dharmadhikari, Edward Nardell, James Seddon, Irina Vasilyeva, Yanlin Zhao, Salmaan Keshavjee, and Mercedes Becerra. Turning off the tap: Stopping tuberculosis transmission through active case-finding and prompt effective treatment. *Lancet (London, England)*, 386, 10 2015.
- [32] Ashwin Dharmadhikari, Matsie Mphahlele, Kobus Venter, Anton Stoltz, Rirhandzu Mathebula, Thabiso Masotla, Martha Van der Walt, Martie, Marcello Pagano, Paul Jensen, and Edward Nardell. Rapid impact of effective treatment on transmission of multidrug resistant tuberculosis. *The International Journal of Tuberculosis and Lung Disease*, 18, 09 2014.
- [33] Ivan Nunes da Silva, Danilo Hernane Spatti, Rogerio Andrade Flauzino, Luisa Helena Bartocci Liboni, and Silas Franco dos Reis Alves. *Artificial Neural Networks, A Practical Course*. 2017.
- [34] Naresh K. Sinha. System identification - theory for the user : Lennart Ijung. *Autom.*, 25(3):475–476, 1989.
- [35] Alvaro David Orjuela-Cañón, Andres Leonardo Jutinico, Mario Enrique Duarte González, Carlos Enrique Awad García, Erika Vergara, and María Angélica Palencia. Time series forecasting for tuberculosis incidence employing neural network models. *Heliyon*, 8(7):e09897, 2022.
- [36] Peter Van Overschee and Bart De Moor. N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems. *Automatica*, 30(1):75–93, 1994. Special issue on statistical signal processing and control.
- [37] Ali H. Sayed. *Adaptive Filters*. John Wiley & Sons, Inc, 2008.
- [38] Mohinder S. Grewal and Angus P. Andrews. *Kalman Filtering: Theory and Practice Using Matlab*. John Wiley & Sons, Inc, 2014.
- [39] J. P. Cerri. *Control and filtering for uncertain discrete-time Markovian jump linear systems*. PhD thesis, University of São Paulo at São Carlos School of Engineering, Brazil, 2013.
- [40] Joao Yoshiyuki Ishihara, Marco Henrique Terra, and Joao Paulo Cerri. Optimal robust filtering for systems subject to uncertainties. *Automatica*, 52:111 – 117, November 2015.
- [41] Andres L. Jutinico, Erika Vergara, Carlos Enrique Awad García, Maria Angélica Palencia, and Alvaro David Orjuela-Cañón. Robust kalman filter for tuberculosis incidence time series forecasting. *IFAC-PapersOnLine*, 54(15):424–429, 2021.

-
- [42] Fuad Dwi Hanggara. Forecasting car demand in indonesia with moving average method. *Journal of Engineering Science and Technology Management (JES-TM)*, 1(1):1–6, Apr. 2021.
- [43] Maxim Shcherbakov, Adriaan Brebels, N.L. Shcherbakova, Anton Tyukov, T.A. Janovsky, and V.A. Kamaev. A survey of forecast error measures. *World Applied Sciences Journal*, 24:171–176, 01 2013.