DESARROLLO Y EVALUACIÓN DE MODELOS MATEMÁTICOS Y EPIDEMIOLÓGICOS QUE APOYEN LA TOMA DE DECISIONES EN ATENCIÓN A LA EMERGENCIA POR SARS-COV2 Y OTROS AGENTES CAUSALES DE IRA EN COLOMBIA UTILIZANDO DATA ANALYTICS Y MACHINE LEARNING

Descripción de modelos analíticos del módulo: Impacto de COVID-19 en otros eventos de interés a nivel de salud pública

Mayo – diciembre 2020


Alianza
CAOBA
Centro de Excelencia y Apropiación
en Big Data y Data Analytics

Contenido

1. Introducción

Este documento corresponde al reporte técnico relacionado con el objetivo general de "Evaluar el impacto de la COVID-19 en otras atenciones de salud, para hacer recomendaciones de política que permitan tomar decisiones buscando atender los casos que se están represando por la crisis actual", especificado dentro del contrato entre Minciencias y la Pontificia Universidad Javeriana a favor del Instituto Nacional de Salud (INS), en el contexto del proyecto "Desarrollo y evaluación de modelos matemáticos y epidemiológicos que apoyen la toma de decisiones en atención a la emergencia por SARS-Cov2 y otros agentes causales de IRA en Colombia utilizando Data Analytics y Machine Learning".

El presente reporte describe el proceso seguido para el desarrollo e implementación de modelos analíticos descriptivos y de pronóstico, y presenta cada uno de los modelos analíticos propuestos para responder al objetivo de evaluar el impacto de eventos No COVID-19 a la situación de salud actual.

Este documento está estructurado de la siguiente forma: El capítulo 2 presenta la selección de los eventos a trabajar y la descripción general de las fuentes de datos a utilizar. El capítulo 3 describe la estrategia utilizada para la construcción de los modelos analíticos, tanto descriptivos como de pronóstico, detalla los indicadores de impacto utilizados para la medición de los resultados de los modelos de pronóstico y de forma general, el despliegue de modelos en la herramienta de visualización desarrollada. El capítulo 4 presenta los modelos descriptivos y da un ejemplo para Tuberculosis, uno de los eventos analizados. Los modelos de pronóstico son presentados en el capítulo 5, utilizando la Tuberculosis para ejemplificar los modelos, al igual que el capítulo anterior. Finalmente, el reporte se cierra en el capítulo 6 con las conclusiones y recomendaciones.

2. Selección de eventos y sus fuentes de datos

2.1. Selección de eventos

Las epidemias como el COVID-19, ponen en tensión los servicios de salud. La congestión derivada por la atención de contagios por coronavirus, el miedo de la población a acudir a los servicios de salud y la dificultad de acceso a estos, generada por medidas como las cuarentenas, afecta la normal atención a pacientes con enfermedades crónicas y la adopción y seguimiento de medidas preventivas en salud. Esto motiva el desarrollo de este proyecto, y fue la razón por la que los eventos seleccionados para su análisis, responden a los siguientes criterios: (1) trazadores de otras enfermedades a riesgo epidémico, (2) enfermedades crónicas que son comorbilidades frente al COVID-19, (3) eventos que son trazadores de problemas de salud mental y, (4) eventos cuyo incremento o reducción se puede presentar como consecuencia de las medidas de autocuidado para prevenir el COVID-19 o por la no atención o atención tardía.

Considerando los anteriores criterios, se selecciona la Tuberculosis que se puede confundir o mezclar con COVID-19, la Diabetes, donde se puede ver disminuida la consulta y, en consecuencia,

aumentar la demanda de servicios de urgencias, los intentos de suicidio como un trazador de problemas en salud mental y, la EDA donde se espera una disminución por las medidas de autocuidado (lavado de manos). Adicionalmente, se incluye la mortalidad infantil, que se puede incrementar por problemas mentales o falta de atención oportuna. Finalmente, se analiza el indicador de exceso de mortalidad, directamente afectado por el COVID-19, por enfermedades no relacionadas con COVID-19 como el accidente cerebrovascular y por una saturación de los servicios de urgencias.

2.2. Fuentes de datos

Para los eventos objeto del presente estudio se consideran las fuentes oficiales donde se registran los números de casos de cada evento. Así, se tomaron las bases de datos del Sistema Integral de Información de la Protección Social (SISPRO), del Sistema Nacional de Vigilancia en Salud Pública (SIVIGILA) y del Departamento Administrativo Nacional de Estadísticas (DANE). Específicamente, la información de EDA, Tuberculosis, Mortalidad infantil e Intento de suicidio se tomó del SIVIGILA; la de Diabetes mellitus de SISPRO; y el Exceso de mortalidad del DANE y de datos abiertos de Colombia (datos.gov.co). Para completar el análisis, se tomó información sobre la población existentes por departamento y municipio, del Censo de población y vivienda del 2018. La Tabla 1 presenta para cada evento, la base de datos de donde se extrajo la información.

Fuente de información	**Eventos analizados**
SIVIGILA	EDA, Tuberculosis, Mortalidad infantil, intento de suicidio
SISPRO (MinSalud)	Diabetes mellitus (hospitalizaciones)
Censo Nacional de Población y Vivienda DANE 2018	Número de personas por departamento y municipio
Estadísticas vitales DANE	Exceso de mortalidad.
Casos positivos de COVID-19 en Colombia – Datos.gov.co	Exceso de mortalidad. Muertes por COVID-19

Tabla 1: Fuentes de datos de los eventos analizados

Las últimas actualizaciones de estas bases de datos se dieron en marzo 2020 para el caso de SIVIGILA y SISPRO y, diciembre 2019 para las bases de datos del DANE.

A continuación, se describe cada una de las bases de datos.

2.2.1. Sistema Integral de Información de la Protección Social – SISPRO

Bodega de datos gestionada por el Ministerio de Protección Social que integra información del sector sobre oferta y demanda de servicios de salud, calidad de los servicios (según Resolución 256 de 2016 y Resolución 1446 de 2006), aseguramiento, financiamiento y promoción social.

Sobre los **prestadores de servicio** se encuentra información sobre la oferta de prestadores en salud. La información relativa al C**iudadano, s**e enlaza con un portal web en donde se puede consultar los datos básicos de salud, riesgos laborales, pensiones y subsidios, al profesional de salud la verificación de su estatus para la prescripción, y la realización de trámites electrónicos de afiliación al Sistema General de Seguridad Social en Salud, de traslados y movilidad entre regímenes, asignación de plazas del Servicio Social Obligatorio-SSO y de otros trámites

Adicionalmente, se encuentra registrada información sobre recursos de financiamiento para actividades y entidades del sector, indicadores sobre saneamiento básico, indicadores de calidad del agua y el registro de estratificación.

El proveedor de esta base de datos es el Ministerio de Salud y Protección Social y la fuente está disponible en este enlace: https://www.sispro.gov.co/Pages/Home.aspx

2.2.2. SIVIGILA

El Sistema Nacional de Vigilancia en Salud Pública -SIVIGILA, se creó para realizar la provisión en forma sistemática y oportuna, de vigilancia y análisis del riesgo de eventos de interés en salud pública (EISP) en el país. Los eventos considerados como de interés esta agrupan según los siguientes ejes temáticos: Enfermedades transmitidas por vectores, Zoonosis, Inmunoprevenibles, Vigilancia nutricional, Enfermedades crónicas no transmisibles, Maternidad segura, Eventos de factores de riesgo ambiental y sanitario, Infecciones de transmisión sexual, Mycobacterias, Infecciones asociadas a la atención en salud, Salud Mental y lesiones de causa externa, Enfermedades emergentes, Muertes en menores de 5 años por EDA-IRA y DN, Morbilidad por EDA y por IRA.

La información es notificada semanalmente por las entidades territoriales (ET) al Instituto Nacional de Salud (INS) a través del Sistema de vigilancia en salud pública (SIVIGILA). En la página web del Instituto Nacional de Salud se encuentran publicados los informes de los EISP, donde se incluye un análisis descriptivo e indicadores de incidencia, prevalencia, mortalidad y letalidad (según aplique), para el nivel nacional y departamental; adicionalmente, también se encuentran publicadas las estadísticas de vigilancia rutinaria donde se presentan los casos notificados de eventos de interés en salud pública por departamento y municipio.

En el Software SIVIGILA se registra información vinculada con la caracterización de las Unidades Primarias Generadoras de Datos – UPGD y las notificaciones individuales de casos.

El proveedor de esta fuente es el Instituto Nacional de Salud, dependencia SIVIGILA. La fuente está disponible en este enlace: https://www.ins.gov.co/Direcciones/Vigilancia/Paginas/SIVIGILA.aspx

2.2.3. Base de datos del DANE

El Departamento Administrativo Nacional de Estadística (DANE), organiza la información disponible en tres categorías: Economía, Sociedad y Territorio. Para este estudio, se tomó información de la base de datos de sociedad, la cual está organizada en ocho subcategorías, Cultura, Demografía y

población, Educación, Gobierno, Género, Pobreza y condiciones de vida, Seguridad y defensa, y Salud. En esta última se toman las estadísticas vitales, como nacimientos y defunciones por semana a nivel nacional y departamental.

El proveedor de esta base de datos es el DANE y la fuente está disponible en este enlace: https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/nacimientos-y-defunciones

2.2.1. Datos abiertos de Colombia

Los datos abiertos de Colombia son un portal web que busca facilitar la interacción entre el Estado y el ciudadano quitando barreras de acceso y consolidando en un solo sitio web trámites, servicios, información y ejercicios de participación. En particular este portal brinda información para investigar, desarrollar aplicaciones, crear visualizaciones e historias y lo hace a nivel de diferentes categorías, en las cuales se incluye cultura, educación, función pública, trabajo, transportes y salud y protección social entre otras.

La fuente consultada en este proyecto fue la de muertes COVID-19 que está disponible en este enlace: https://www.datos.gov.co/Salud-y-Protecci-n-Social/Casos-positivos-de-COVID-19-en-Colombia/gt2j-8ykr

3. Estrategia para la construcción de los modelos analíticos

La estrategia definida para estandarizar el desarrollo de los modelos descriptivos (análisis descriptivos) y de pronósticos, parte de la recolección y preparación de los datos en las fuentes seleccionadas (etapa I), hasta llegar a la etapa IV. de despliegue de los modelos, como se muestra en la Figura 1. A continuación, se describen las etapas que conforman la estrategia y, el detalle de las estrategias por cada tipo de modelo, se da respectivamente en los capítulos 4 y 5 de este documento.

3.1. Recolección y preparación de datos

La recolección y preparación de los datos tiene como propósito organizar la información proveniente de las bases de datos, en un único archivo estándar que sirva para alimentar los modelos de análisis y facilitar los análisis que se hagan sobre esos resultados. A partir de los datos disponibles y seleccionados de las fuentes de datos de SIVIGILA, SISPRO y DANE, se generan dos instrumentos que facilitan tanto la interpretación de los resultados obtenidos de los modelos y verificar su validez, como el desarrollo de los mismos.

El primer instrumento es una ficha técnica para un evento objeto de análisis. La Figura 2 muestra un ejemplo de ficha para el evento tuberculosis. La ficha registra información de la fecha en la que se obtuvo la información de la fuente de datos, el último periodo reportado, los códigos de los eventos

consultados en las bases de datos, en este caso que utiliza SIVIGILA para el evento, y el número de casos utilizados en el análisis por año. De igual manera se muestran las variables seleccionadas para la extracción, cómo se manejarán los datos faltantes, los sucesos que pueden afectar el número de casos reportados, las metas asociadas a incidencias o muertes definidas por entidades referentes, si existen resultados previos realizados por entidades como el Observatorio Nacional de Salud, comparables con el trabajo realizado y finalmente, los indicadores calculados para el análisis del impacto.

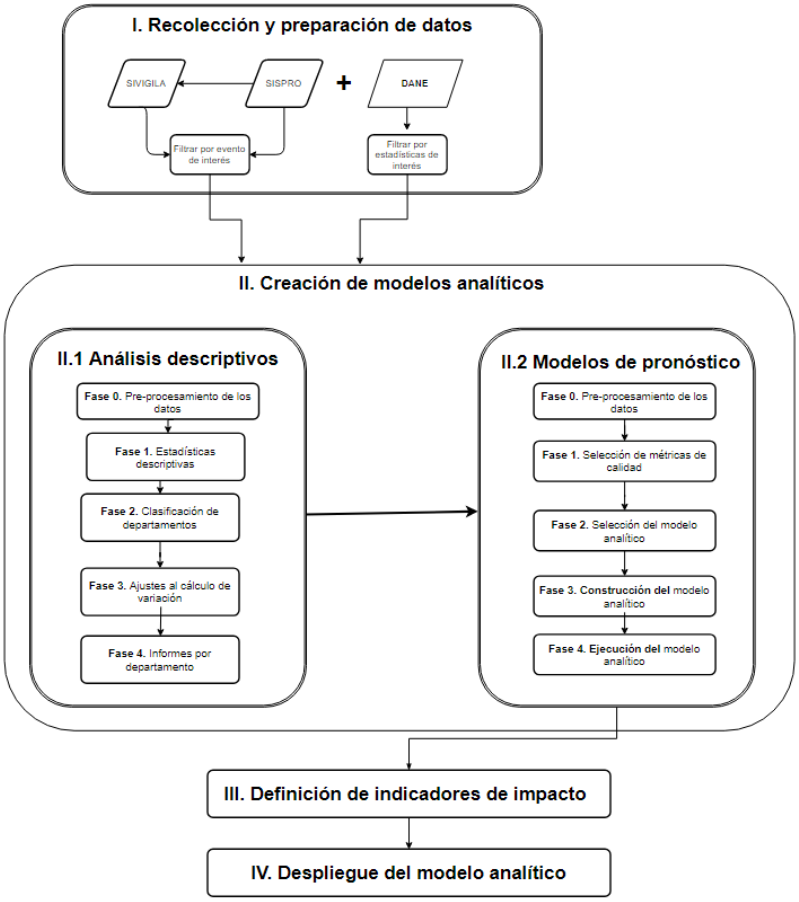


Figura 1 Estrategia general de los análisis descriptivos y modelos de pronóstico

Tuberculosis	
Fuente de datos	Sivigila (Portal Web - datos abiertos)
Fecha de consulta de la fuente	19/08/2020
Último periodo reportado	01/03/2020
Criterio de selección SIVIGILA: (Hace referencia a los códigos de los eventos que se analizan)	Año 2019 – 52 semanas: • TUBERCULOSIS - Resistente (COD 813): 322 cases • TUBERCULOSIS - Sensible (COD 813): 14466 cases • TOTAL: 14788 casos Año 2020 – 8 semanas: • TUBERCULOSIS - Resistente (COD 813): 49 cases • TUBERCULOSIS - Sensible (COD 813): 2007 cases • TOTAL: 2056 casos
Variables seleccionadas (indicado el nivel geográfico)	Número de casos por semana epidemiológica por Departamentos y capitales, No hay datos con valores en cero.
Manejo de datos faltantes != datos con valor cero	Datos faltantes serán manejados con un valor igual a -1. En los análisis descriptivos no se incluirán estos valores. En los modelos de pronóstico el -1 será reemplazado por cero
Sucesos/situaciones a partir de Enero 2020 (e.g., Confinamiento -, Apertura)	Inicio de cuarentena: 22 de marzo, fase 1: 25 de marzo- 26 de abril. Fase 2: 27 de abril - 31 de agosto, Fase 3: 1 de septiembre -
Metas asociadas a los evento por entidades referentes	"Metas dirigidas al cumplimiento de las definidas por los Objetivos de Desarrollo Sostenible de poner fin a la tuberculosis a 2030. Como meta de impacto se propone la reducción de al menos el 35% de las muertes por tuberculosis (letalidad) a 2025 247 muertes), en comparación con el 2015 (987 muertes)." De igual manera se tiene una serie de proyecciones sobre incidencias ymuertes como por ejemplo, a 2020 10.772 casos y 678 muertes. Para 2021 9.894 casos y 590 muertes. https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/INEC/INTOR/Plan-estrategico-fin-tuberculosis-colombia-2016-2025.pdf.
Resultados de análisis previos en el INS	No se tiene información de resultados previos.
Indicadores de impacto	Diferencia acumulada respecto al año anterior Diferencia acumulada respecto al año vigente Indicador indirecto con Covi19 (%) Indicador indirecto con Covi19 (No.)

Figura 2 Ficha del evento

El segundo instrumento, es la generación de un formato estándar para facilitar el manejo de los datos de los diferentes eventos y que sirva para la ingesta de los datos a los modelos de pronóstico. En este formato, se guarda en un archivo que se nombra, de forma general "`nombreEvento_dataset.csv`", para el caso de la tuberculosis, el archivo se denomina "`tuberculosis_dataset.csv`". Este archivo, en el formato que se muestra en la Figura 3, contiene el reporte por fecha (representado en la figura por las columnas ***date***, ***year***, ***month*** y ***week***), a nivel del periodo epidemiológico (***period***), el número de casos asociados al evento (***value***), con información de geografía, a nivel de país (nacional), departamentos y capitales (contenido en la columna ***entity***).

date	entity	year	month	week	period	value
1/01/2017	COLOMBIA	2017	1	1	1	298
1/08/2017	COLOMBIA	2017	1	2	1	256
1/15/2017	COLOMBIA	2017	1	3	1	347
1/22/2017	COLOMBIA	2017	1	4	1	289
1/29/2017	COLOMBIA	2017	1	5	2	312
1/01/2017	AMAZONAS	2017	1	1	1	0
1/08/2017	AMAZONAS	2017	1	2	1	1
1/15/2017	AMAZONAS	2017	1	3	1	0
1/22/2017	AMAZONAS	2017	1	4	1	2
1/29/2017	AMAZONAS	2017	1	5	2	0
1/01/2017	BARRANQUILLA	2017	1	1	1	14
1/08/2017	BARRANQUILLA	2017	1	2	1	5
1/15/2017	BARRANQUILLA	2017	1	3	1	10
1/22/2017	BARRANQUILLA	2017	1	4	1	9
1/29/2017	BARRANQUILLA	2017	1	5	2	13
1/01/2017	CUCUTA	2017	1	1	1	9
1/08/2017	CUCUTA	2017	1	2	1	8
1/15/2017	CUCUTA	2017	1	3	1	13
1/22/2017	CUCUTA	2017	1	4	1	7
1/29/2017	CUCUTA	2017	1	5	2	7

Figura 3 Datos sobre el reporte de casos de un evento, a nivel país, departamento y capital utilizados para los modelos analizados

La estandarización de fuentes incluye datos tomados del DANE como división política y población, que fueron organizados como se muestra en las figuras Figura 4 y Figura 5. De igual manera, la estandarización se realiza para fuentes adicionales con el fin de facilitar labores de configuración de los modelos a lo largo del tiempo. Estas fuentes adicionales corresponden a los archivos de configuración de cada tipo de modelo, como lo muestran respectivamente las figuras

```
"event_list": [
    {
      "analysis_list": [ "PARTIAL", "FULL" ],
      "ci_alpha": 0.9,
      "enabled": true,
      "full_init_date": "2017-01-01",
      "mape_threshold": 4.0,
      "n_forecast": 13,
      "name": "TUBERCULOSIS",
      "partial_end_date": "2019-12-27",
      "perc_test": 0.20,
      "ts_tolerance": 4.0
    },
```

Figura 7 y Figura 7.

Adicionalmente, para los análisis descriptivos el archivo de configuración incluye entre otros parámetros, la frecuencia de agrupación para realizar el análisis, el nombre del evento, la unidad de medida del evento (tasa o número de casos) y los años a excluir del análisis, si así se requiere. Por

su parte, el archivo de configuración para modelos de pronóstico, incluye los modelos a construir Sin COVID-19 (*Partial*) o con COVID-19 (*Full*), fechas de inicio y fin de la serie de tiempo a analizar (*full_init_date*, *partial_end_date*), porcentaje de registros utilizado para generar el conjunto de prueba, y valores de umbrales utilizados en la selección de los modelos, entre otros.

entity	divipola	type
CARTAGENA	13001	capital
COLOMBIA	0	country
BOLIVAR	13000	department
BOYACA	15000	department
CALDAS	17000	department
CAQUETA	18000	department

Figura 4 Datos sobre la división geográfica de Colombia provistos por el DANE

divipola	entity	type_entity	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
5001	Medellín	capital	2343049	2368282	2393011	2417325	2441123	2464322	2486723	2508452	2529403	2549537	2569007
8001	Barranquilla	capital	1186640	1193952	1200820	1207264	1213246	1218737	1223967	1228621	1232766	1236489	1239804
11001	Bogotá, D.C.	capital	7363782	7467804	7571345	7674366	7776845	7878783	7980001	8080734	8181047	8281030	8380801
13001	Cartagena	capital	944250	955569	967051	978574	990151	1001680	1013454	1025086	1036412	1047321	1057767
5	Antioquia	department	6065846	6143709	6221742	6299886	6378069	6456207	6534764	6613063	6690977	6768362	6845057
8	Atlántico	department	2314447	2344140	2373680	2403027	2432145	2461001	2489709	2518096	2546138	2573816	2601116
0	Colombia	country	45508205	46043696	46581372	47120770	47661368	48202617	48747632	49291925	49834727	50375194	50912429

Figura 5 Datos sobre proyecciones de población total por división política provistos por el DANE

```
"event_list": [
    {
      "enabled": true,
      "frequency": "weekly",
      "name": "TUBERCULOSIS",
      "rate_enable": true,
      "skip_years": []
    },
```

Figura 6 Parámetros requeridos para los modelos descriptivos – archivo desc_config.json

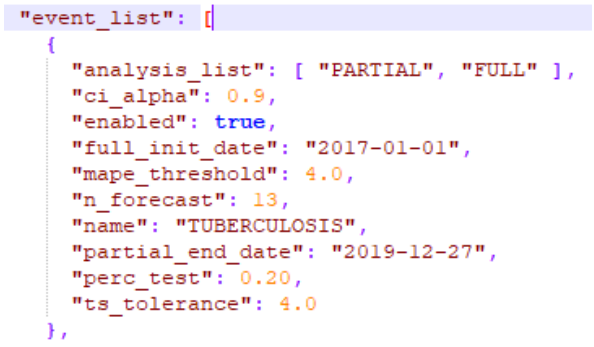


Figura 7 Parámetros requeridos para los modelos de pronóstico – archivo pred_config.json

3.2. Análisis descriptivos

Los análisis descriptivos buscan identificar el comportamiento histórico del número de casos, para cada uno de los eventos seleccionados. Adicionalmente, se busca determinar la pertinencia del uso de las medidas de tendencia central, así como la relevancia de agrupar esta información por unidad de tiempo (días, semanas, periodo epidemiológico, meses) y por unidad geográfica (nacional, departamental, municipal). El propósito es poder llegar a la unidad de análisis más pequeña que permita a partir de los diferentes modelos, sacar conclusiones relevantes sobre el comportamiento del evento. De esta manera, se podrá cuantificar el posible impacto de las mediadas asociadas con la COVID-19 en el comportamiento de dichos eventos. El detalle de estos análisis se presenta en el Capítulo 4. Análisis descriptivo

3.3. Modelos analíticos de pronóstico

El objetivo de los modelos de pronóstico, es a partir de una fecha dada, estimar el número de casos de un evento, que se presentarán durante cada uno de los siguientes 13 periodos epidemiológicos. Estos modelos se realizan para dos escenarios: el primero en el cual incluye la información hasta el último periodo epidemiológico de 2019, que se utiliza para determinar el número de casos que se presentaría si no se hubiera presentado la pandemia de COVID-19 (sin COVID-19). El segundo escenario, con COVID-19 pronostica el número casos incluyendo los datos reportados durante el año 2020. Estos datos son la base para calcular los indicadores de impacto de cada uno de los eventos, descritos en la sección 3.4. El detalle de estos modelos se presenta en el Capítulo 5.

3.4. Definición de indicadores de impacto

Considerando que los datos a los que se tuvo acceso, tenían como última actualización marzo de 2020, y que para esa fecha el número de casos COVID-19 reportados por día no ascendía a 120, la medición del impacto se ve limitada por el acceso a información actualizada.

Basados en el informe sobre exceso de mortalidad presentado por el DANE en el 2020, y haciendo el paralelo con los eventos estudiados; se plantea, dos tipos de mediciones de impacto, para un evento en salud, el primer indicador, llamado **`indicador indirecto Sin COVID-19`**, que mide el impacto entre enero y marzo del 2020, se estima como la diferencia entre los casos reportados y los casos proyectado por el modelo de pronóstico Sin COVID-19. Este indicador se puede estimar en número de casos (No.) o en porcentaje (%), tomando como referencia, los casos reportados. De forma similar, se plantea el **`indicador indirecto Con COVID-19`**, que, a diferencia del anterior, toma como casos reportados desde marzo y como proyectados, los reportados por el modelo de pronóstico Con COVID-19.

El quinto indicador es el denominado **`Diferencia acumulada respecto al año anterior`**, que se estima como la diferencia entre el total de casos pronosticado para el año anterior y el acumulado de casos reportados del año vigente.

Finalmente, se tiene el indicador **`Diferencia acumulada respecto al año vigente`**, que se estima como la diferencia entre el total de casos pronosticado del año vigente y el acumulado de casos reportados del año vigente.

Los resultados obtenidos para estos indicadores, para el evento tuberculosis, toman como insumos los casos reportados y los pronósticos realizados para el año 2020. Se estimaron los indicadores de indirectos a nivel nacional, departamental y para cada capital, para el año 2020. Estas estimaciones son a manera de ejemplo, considerando que la última actualización de datos a la que se tuvo acceso fue en marzo de 2020, y a esta fecha se habían reportado no más de 1000 casos de COVID-19 en todo el país.

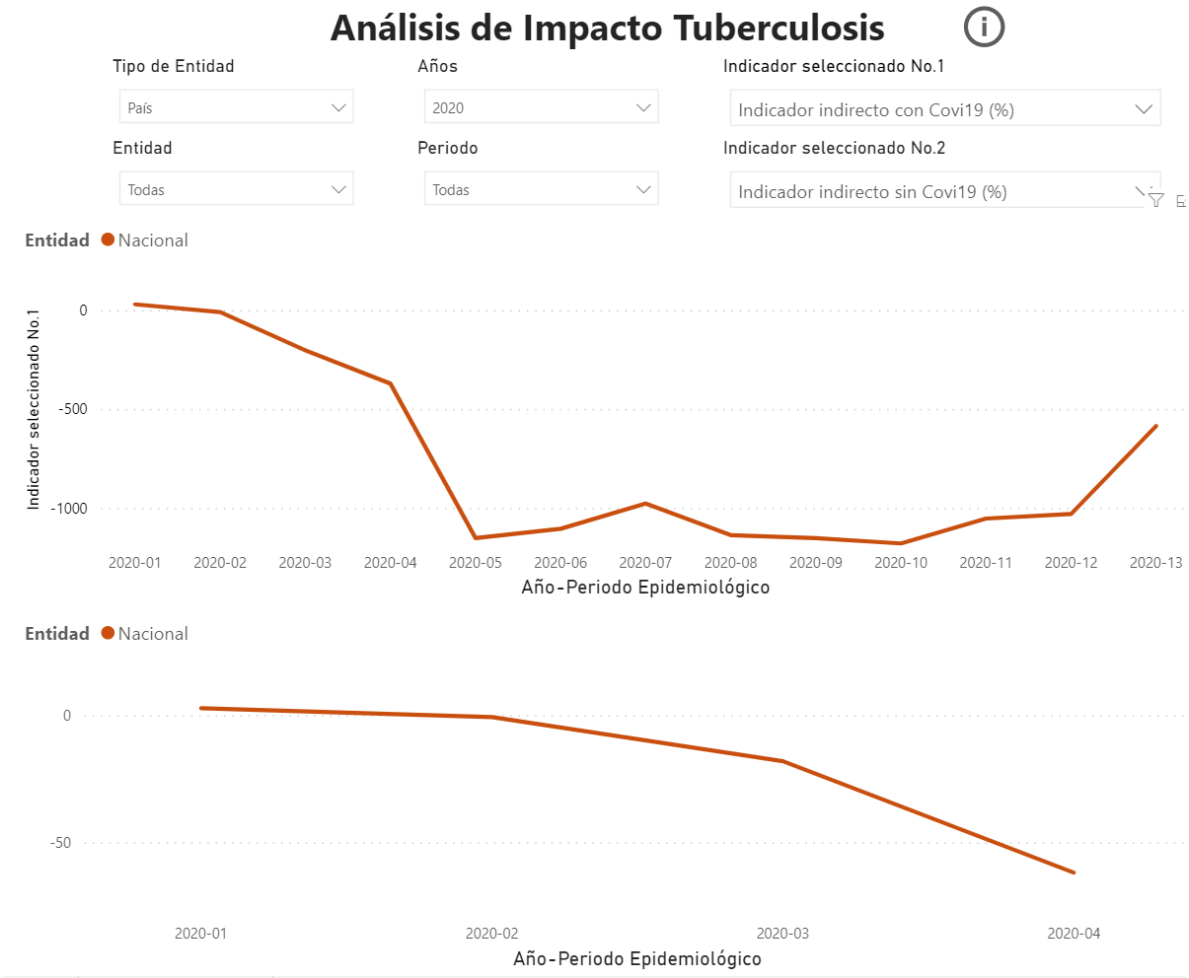


Figura 8: Indicadores indirectos con y sin COVID-19 en porcentaje a nivel nacional.

La Figura 8 muestra para el evento de tuberculosis, los indicadores indirectos de impacto con y sin COVID-19 en %, a nivel nacional para cada periodo epidemiológico. De las gráficas se aprecia que, de acuerdo con los datos, hay una tendencia a que se presente un sub-registro de casos partiendo del quinto periodo epidemiológico, para el caso del indicador indirecto con COVID-19. Para el indicador indirecto sin COVID-19, el cual se estimó para los cuatro primeros periodos epidemiológicos, presenta, igualmente una tendencia a la baja, aunque los órdenes de magnitud son de las decenas, mientras que en el caso de indicador indirecto con COVID-19 es de las centenas o inclusive de los miles.

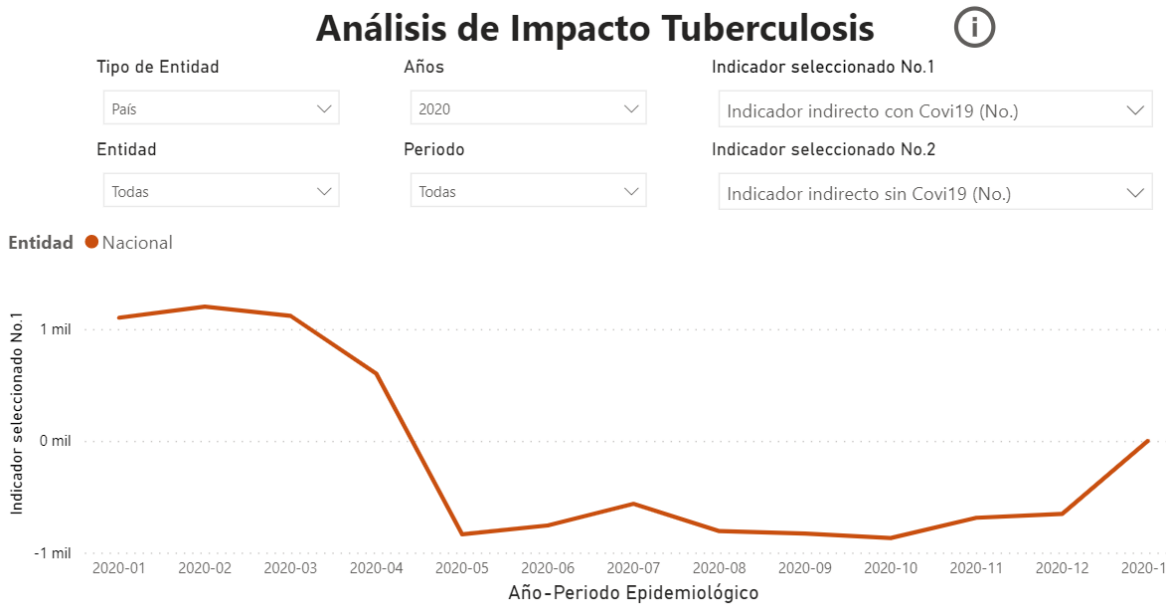


Figura 9: Indicadores indirectos con y sin COVID-19 en número de casos a nivel nacional.

La Figura 9 muestra los indicadores en número de casos, validando los resultados obtenidos en los indicadores en porcentajes, presentados en la Figura 8. En cuanto al indicador de diferencias con respecto al año anterior y al año vigente, resultados que se muestran en la Figura 10, estos también reflejan el sub-registro que se ha venido presentando en los eventos de tuberculosis para este año, en ambos indicadores.

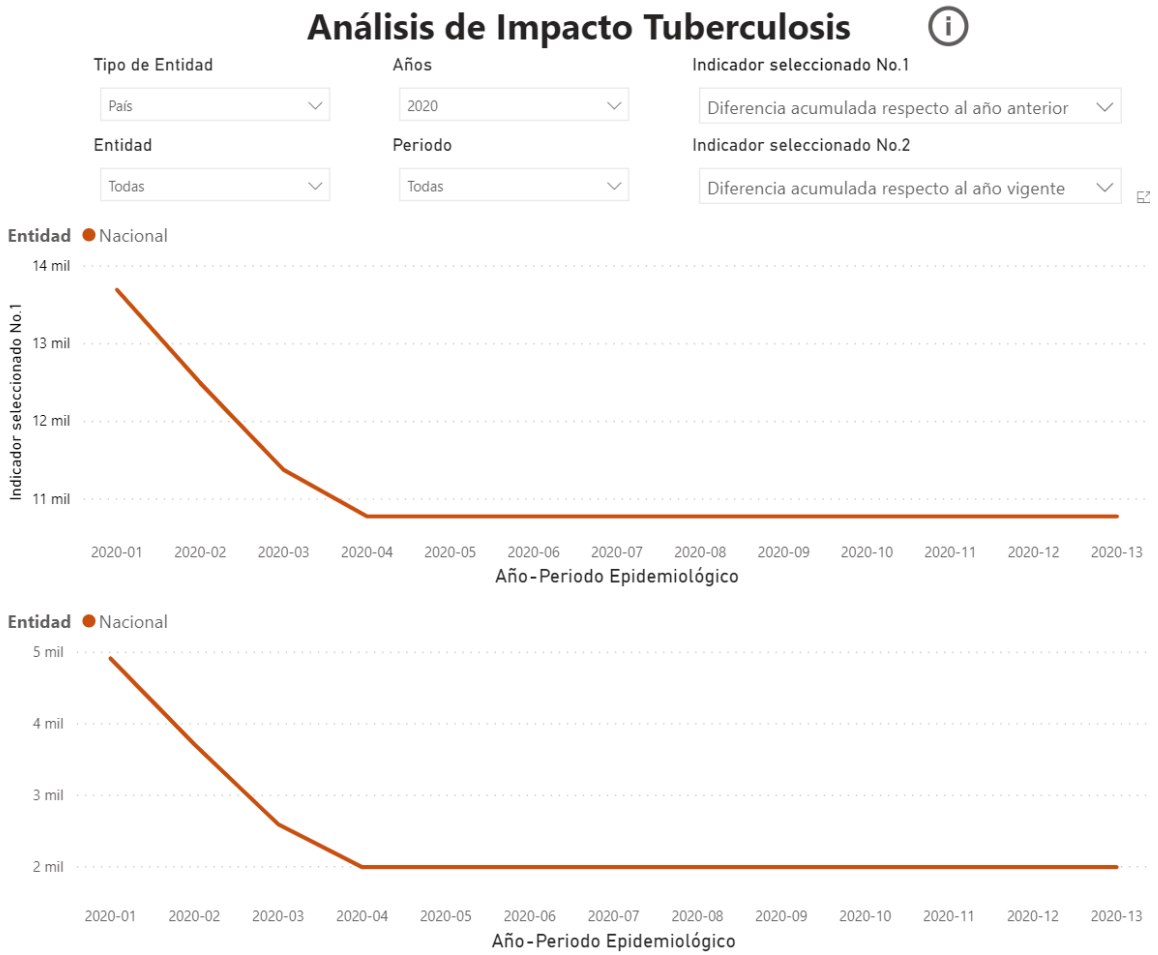


Figura 10: Indicador de impacto Diferencia acumulada respecto al año anterior y al año vigente.

3.5. Despliegue de los modelos

Los motores que generan los modelos analíticos (descriptivos y de pronóstico) para todos los eventos, fueron desplegados en una instancia de cómputo intensivo (c6g.4xlarge) de AWS, siguiendo una arquitectura de nube. Allí los modelos podrán ser sincronizados, cada vez que se reporten nuevos datos. De ser necesario, los modelos también podrán ser actualizados (bajo demanda) a partir de nuevos hiperparámetros de configuración. Sin embargo, en la medida que se incrementen los tiempos de este proceso, se debe ajustar esta decisión y sincronizarlos cuando el modelo de pronóstico sea obsoleto.

Los resultados de los modelos analíticos y de los indicadores de impacto, pueden visualizarse en la página http://caobacovidepv.org/#/politicas-publicas/.

4. Análisis descriptivo

El análisis descriptivo, parte del supuesto que cada evento puede ser tratado como una serie de tiempo. En consecuencia, se efectúan análisis independientes, para cada combinación evento – entidad territorial. Teniendo en cuenta que las condiciones de prestación del servicio pueden cambiar, en el análisis de los departamentos se excluyen los datos de las capitales y estas se presentan de manera independiente.

4.1. Estrategia general

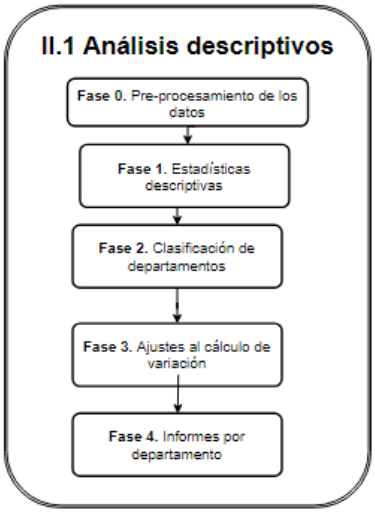


Figura 11: Metodología de análisis.

Para cada evento, se sigue la metodología descrita en la Figura 11. En la fase 0 se toman los datos de formato común y se llevan a un formato propio para facilitar los análisis descriptivos. En particular se realizan los siguientes pasos:

1. Se completan los archivos empleados para incluir datos desde el 2009 en adelante.

2. Se aplica un análisis de calidad de datos a las entidades de los datos (País, Departamento y Capital), para estandarizar y homologar los nombres de las entidades.

3. Se pivotean los datos verticales (entidad como filas y semanas como columnas) para que la manipulación, lectura y posterior procesamiento fuera más sencillo.

4. Se completan con el valor “-1”, las semanas-entidad para las cuales no se reportan datos.

5. Se manejan casos puntuales por evento, como fue el caso para EDA, en el cual los datos de los años 2013 y 2018 estaban registrados en SIVIGILA de forma no homogénea. Es decir, para la mayoría de las entidades los registraron solo en 1 o 2 semanas del año. Por lo tanto, para esa enfermedad, no se tomaron en cuenta esos años, ya que cualquier operación que se hiciera para redistribuir los datos iba a alterar los resultados del análisis descriptivo.

6. El posterior agrupamiento de los datos de semanas a periodos epidemiológicos, se realizó directamente en el motor utilizado para construir el análisis descriptivo.

En la fase 1 se generan estadísticas descriptivas (media, desviación estándar, mínimo, máximo, coeficiente de variación y los tres cuartiles (25%, 50% y 75%), por departamento, para el número de casos reportados durante el periodo de análisis. Estas estadísticas son usadas, en la fase 2, para hacer una clasificación de los departamentos en cinco grupos presentados en la tabla 2. La clasificación permite identificar los municipios en los que las medidas de tendencia central pueden ser usadas para describir el comportamiento histórico del evento (grupo 5). Esta clasificación permitirá al tomador de decisiones identificar si el departamento analizado tiene altos o bajos niveles de variación y, por lo tanto, darle la debida relevancia a las conclusiones.

Grupo	Descripción
1	Departamentos en los que el porcentaje de semanas en cero supera el 40%
2	Departamentos en los que existe tendencia en la serie de datos 2009-2019
3	Departamentos con coeficientes de variación alto
4	Departamentos con coeficientes de variación medio
5	Departamentos con coeficientes de variación bajo

Tabla 2: Clasificación de los departamentos

Adicionalmente, se calcula la variación porcentual, en el número de casos reportados, comparando el mismo periodo del 2020 y del 2019, con el fin de identificar los cambios en el número de casos reportados de un evento antes y durante el año de la pandemia, dando un contexto del indicador y así, identificar si estos cambios son debido o no a la pandemia. Finalmente, se elaboran informes por Departamento que son usados como insumo en un análisis cualitativo.

La metodología descrita en la Figura 11 fue validada, de manera preliminar, con dos eventos: Enfermedad Diarreica Aguda (EDA) y VIH[1]. El objetivo es encontrar posibles necesidades de ajuste,

[1] Se considera el VIH como evento para validar, pues de acuerdo con expertos, este es uno de los eventos que será afectado por el COVID-19. Inicialmente se consideró incluir el VIH como evento a analizar, pero debido a la baja disponibilidad de información, y los posibles usuarios de las herramientas desarrolladas en este proyecto, no los dieron entre las prioridades de estudio.

a la luz de la información disponible en SIVIGILA, para dos eventos en los que se espera un comportamiento diferente. Esta validación permitió identificar que:

- Usando información pública de SIVIGILA existe el riesgo de incluir variaciones artificiales. Las fechas de los eventos, reportadas en estos datos, pueden verse afectadas por el proceso de registro en el sistema de información. En consecuencia, es necesario tener acceso a los datos fuente para verificar si es posible eliminar dicha variación. Por ahora, se debe ser cuidadoso en el uso de las estadísticas para formular conclusiones.
- Los coeficientes de variación calculados son superiores a lo esperado. Para eventos en los que no se proyectan comportamientos estacionales, se esperaba que las medidas de tendencia central permitieran describir el comportamiento histórico del número de casos. Por ende, se plantea la construcción de intervalos de confianza para la media del número de casos, con el fin de cuantificar el impacto de la pandemia. Sin embargo, los altos niveles de variabilidad observados hacen que esta medida no sea suficiente.
- Después de analizados los datos disponibles, se espera que un porcentaje muy alto de municipios registren cero casos nuevos, por periodo epidemiológico. En consecuencia, se propone la unidad de análisis sean los Departamentos

Finalmente, considerando la disponibilidad de la información, el reporte en cero de varios municipios y que las inconsistencias en el registro eran mayores cuando se analizaban por municipio y por semana, se decide que, las unidades de análisis serán a nivel nacional, departamental y para cada departamento se analizará por separado su capital. En cuanto a la unidad temporal, se utilizará el periodo epidemiológico, teniendo 13 periodos por año.

4.2. Análisis descriptivos caso tuberculosis

Esta sección presenta el análisis de la tuberculosis como el valor agregado de múltiples eventos, reportados de manera independiente en los informes de vigilancia rutinaria de SIVIGILA. Para los años 2009-2018, se sumaron los nuevos casos de: Meningitis tuberculosa (código 530 en SIVIGILA), Tuberculosis extra pulmonar (810), Tuberculosis pulmonar (820) y Tuberculosis fármaco resistente (825). Así mismo, a partir del año 2019, se suman los casos de Tuberculosis sensible y Tuberculosis resistente (813).

La

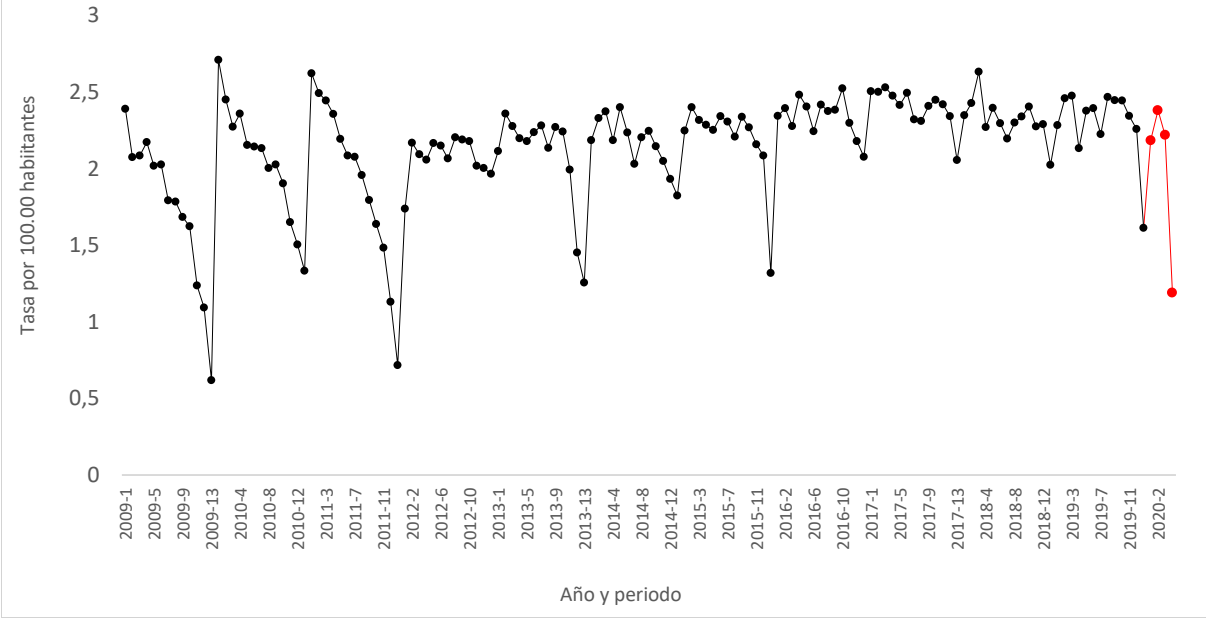


Figura 12 presenta la tasa de nuevos casos, por cada 100.00 habitantes, a nivel nacional. Los valores en rojo representan los cuatro primeros periodos epidemiológicos del año 2020. Como puede observarse, a partir del año 2015, el comportamiento es más estable. Adicionalmente, se mantiene el componente de estacionalidad en el que los últimos periodos del año tienen valores inferiores. Para el 2020, el cuarto periodo epidemiológico presenta una reducción atípica que será examinada por cada Departamento.

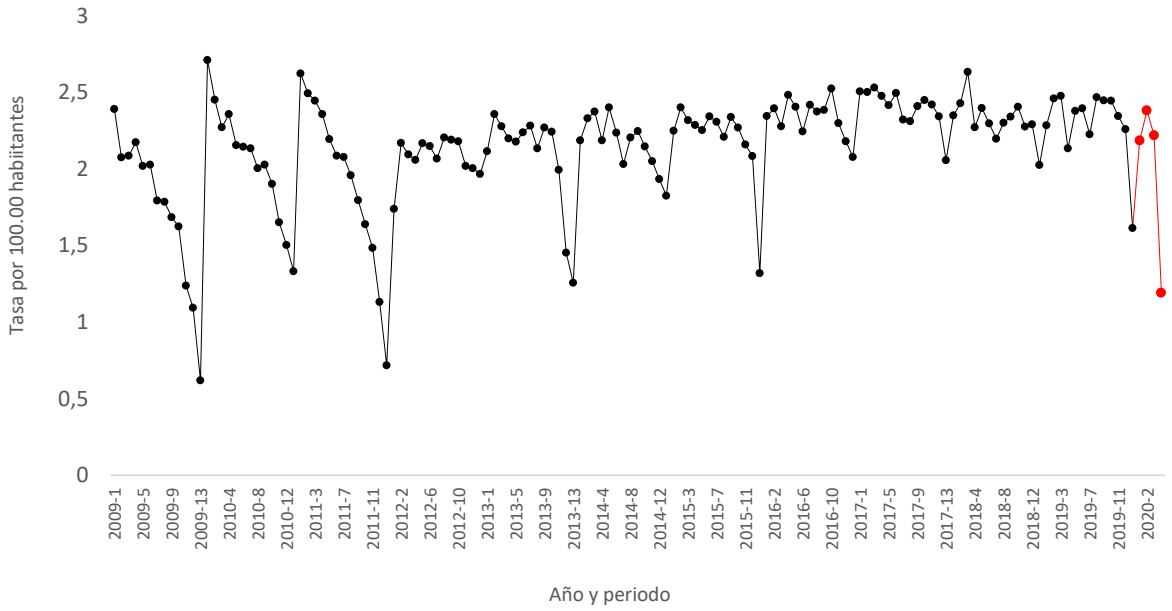
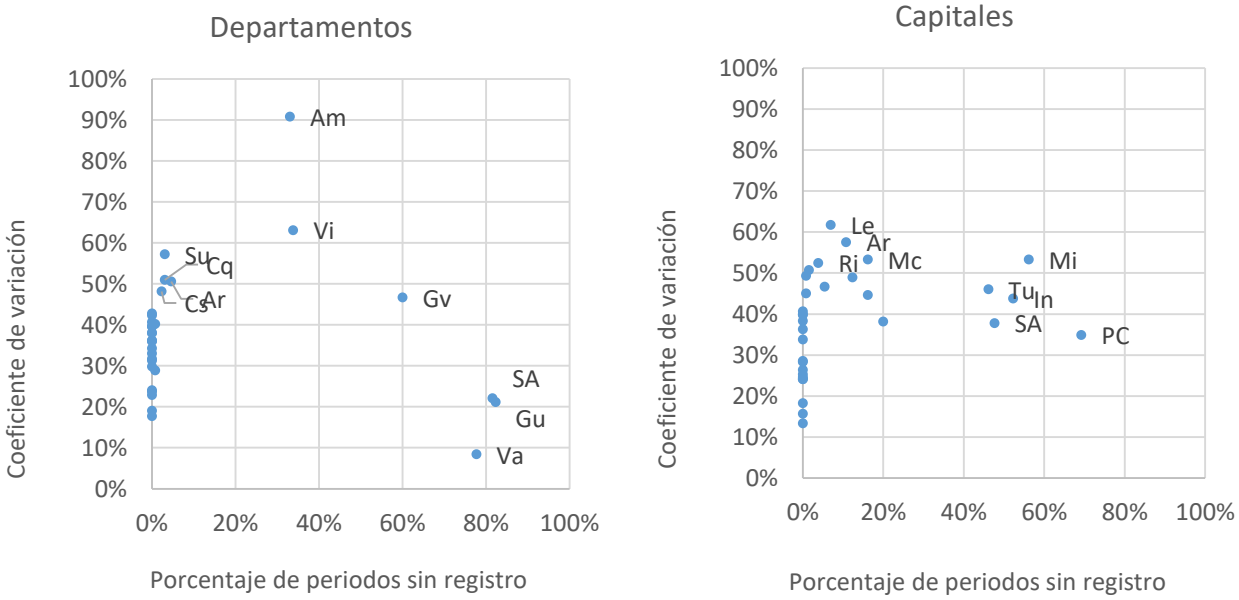


Figura 12: Nuevos casos de tuberculosis, a nivel nacional, expresados como tasa en cada periodo epidemiológico.

Adicionalmente, la Figura 13 presenta los coeficientes de variación y el porcentaje de periodos epidemiológicos sin registro de nuevos casos, para los años 2009-2019, en cada departamento y capital, información útil para la interpretación de las estadísticas descriptivas. Por ejemplo, al analizar los cambios porcentuales en el número de casos nuevos, se debe tener en cuenta que en



departamentos como Vaupés (Va), Guainía (Gu), San Andres (SA) y Guaviare (Gv), se observó un número muy alto de los periodos epidemiológicos sin registros. Adicionalmente, Amazonas (Am) y Vichada (Vi) tienen coeficientes de variación superiores al 60%. De la misma forma, en Puerto Carreño (PC), Mitú (Mi), Tunja (Tu), Inírida (In) y San Andrés (SA) más del 50% de los periodos epidemiológicos no tienen reporte de casos nuevos. Finalmente, Leticia (Le), Arauca (Ar), Mocoa (Mc) y Riohacha (Ri) tienen coeficientes de variación superiores al 50%

Figura 13: Información disponible por Departamento y Ciudad Capital.

Finalmente, en la Figura 14 se comparan los valores observados para el cuarto periodo epidemiológico, entre los años 2019 y 2020. Con el objetivo de entender si la variación es atípica, se analizaron las variaciones anuales observadas entre 2015 y 2019. Después de excluir las variaciones en las que uno de los años no tenía casos reportados, se seleccionaron los valores máximo y mínimo presentados, por departamento. Usando la información de la Figura 13, y excluyendo los seis departamentos en los que el coeficiente de variación o el número de periodos sin reporte es alto, en 12 de los 26 departamentos analizados, existe una disminución atípica para el cuarto periodo epidemiológico. Para Antioquia (An), Atlántico (At), Boyacá (Bo), Caldas (Cl), Cauca (Cu), Cundinamarca (Cn), La Guajira (LG), Risaralda (Ri) y Valle del Cauca (VC) la disminución es mayor al 50% de los casos reportados en el año 2019. Un análisis similar, por capital, se presenta en la Figura 15. La disminución del 100% en el caso de Popayán (Po) significa que no existen casos reportados en el cuarto periodo epidemiológico de 2020. En doce capitales, la disminución es mayor a lo observado entre 2015 y 2019 y en nueve la disminución es del más del 50%: Medellín (Me), Barranquilla (Ba), Cartagena (Cl), Manizales (Ma), Popayán (Po), Mocoa (Mo), Vichada (Vi) y Pereira (Pe).

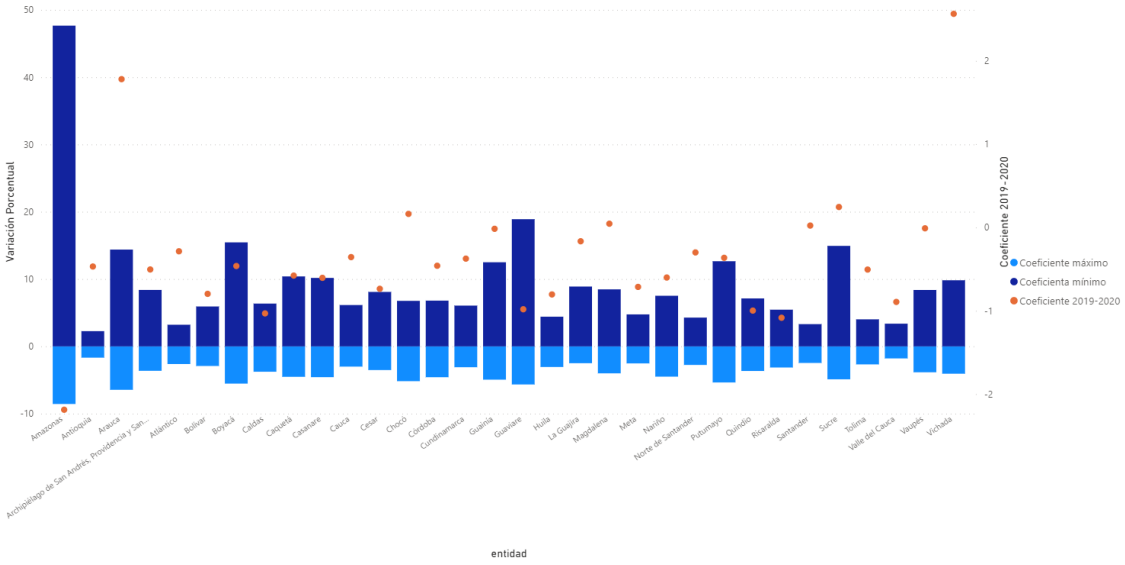


Figura 14: Variación porcentual en la tasa de casos reportados para el periodo epidemiológico 4, por Departamento[2].

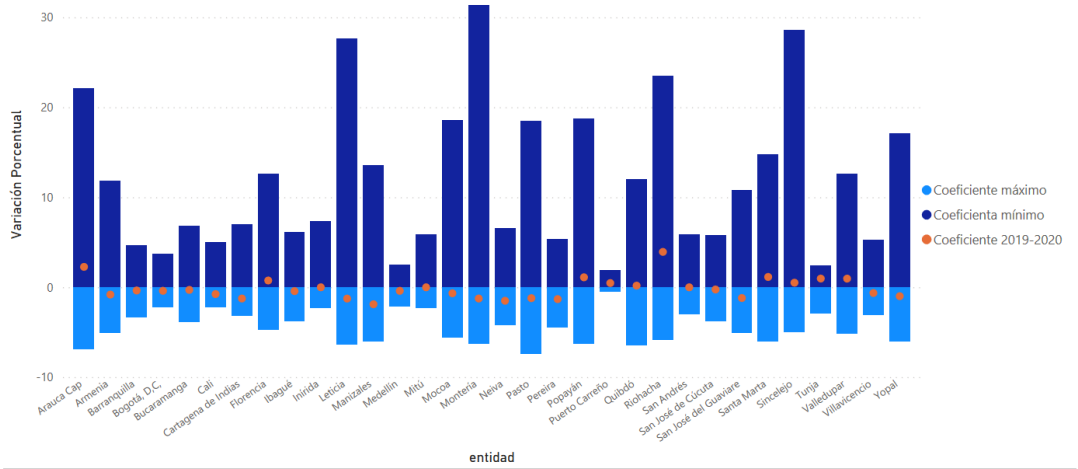


Figura 15: Variación porcentual en la tasa de casos reportados para el periodo epidemiológico 4, por Capital.

En este contexto, debido a los niveles de variación discutidos anteriormente, usar una medida de tendencia central para analizar los cambios en el número casos reportados, es insuficiente. En consecuencia, se propone el cálculo del impacto indirecto, que tuvo la COVID-19 en el reporte de nuevos casos de tuberculosis, usando un modelo de pronóstico. Se calcula el exceso (o defecto) de nuevos casos como la diferencia entre el número de casos reportados, en un periodo epidemiológico, y el valor pronosticado para el mismo periodo. La Figura 16 resume información

[2] Las gráficas son extraídas de la herramienta. Las etiquetas de coeficiente máximo y mínimo corresponden a los valores máximo y mínimo de la variación porcentual.

sobre este indicador de impacto indirecto para las capitales y los departamentos. El Eje X presenta el indicador medido como un porcentaje de cambio respecto al pronóstico y el Eje Y presenta la diferencia medida en número de nuevos casos reportados.

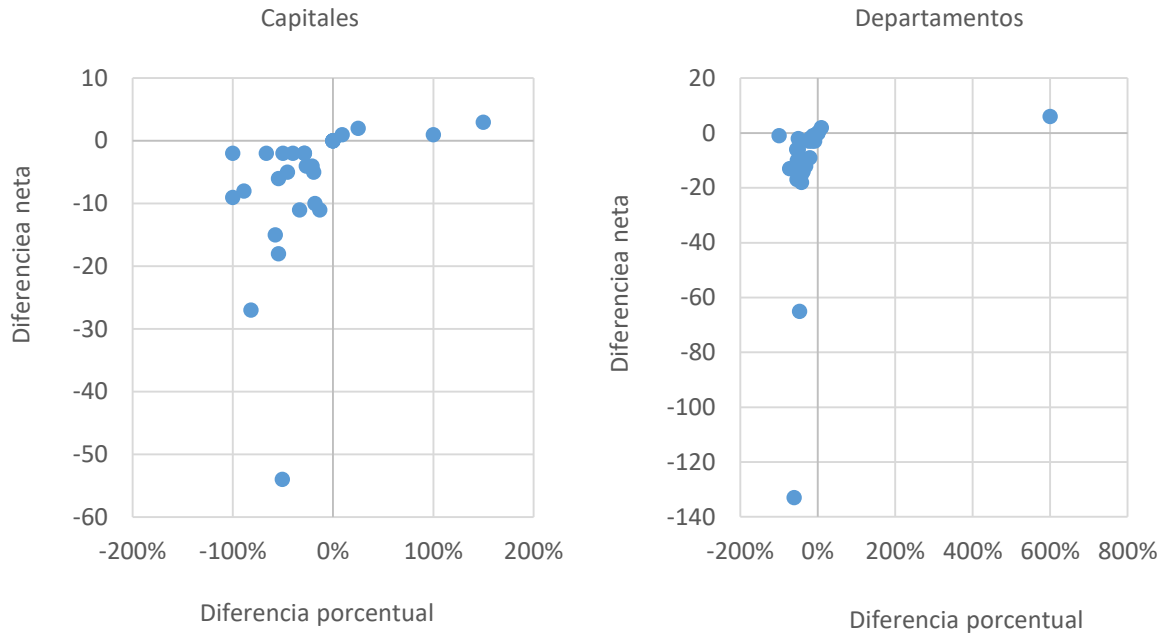


Figura 16: Impacto indirecto en el periodo epidemiológico 4.

Como puede observarse en la Figura 16, para la mayoría de las capitales se encontró una disminución de menos de 20 casos en el número de reportes para el periodo 4. En promedio, la disminución observada es del 46%.

4.3. Análisis descriptivos caso exceso de muertes

Esta sección presenta el análisis del exceso de muerte, por causas naturales, durante las semanas 10 y 34 del año 2020. La Figura 17 compara la estimación del exceso de muertes por semana, a nivel nacional, y las muertes registradas por COVID-19.

El análisis parte del supuesto de que, si no se presenta la pandemia, el número de fallecidos en cada semana del 2020 se hubiera podido estimar usando el promedio de fallecidos reportados en los cinco años anteriores (2015-2019). A nivel nacional, el número de fallecidos por semana en el periodo 2015-2019 presenta variaciones menores con coeficientes que oscilan entre el 3% y el 8%. En consecuencia, para la estimación del exceso de muertes, se construye un intervalo de confianza (al 95%) del promedio de fallecidos por semana. Posteriormente, se calcula la diferencia entre los límites de dicho intervalo y el número de muertes naturales reportadas durante 2020. En

la Figura 17 se aprecia cómo a partir de la semana 30 el número de muertes registradas por COVID-19 es menor a la estimación del exceso de muertes por causas naturales.

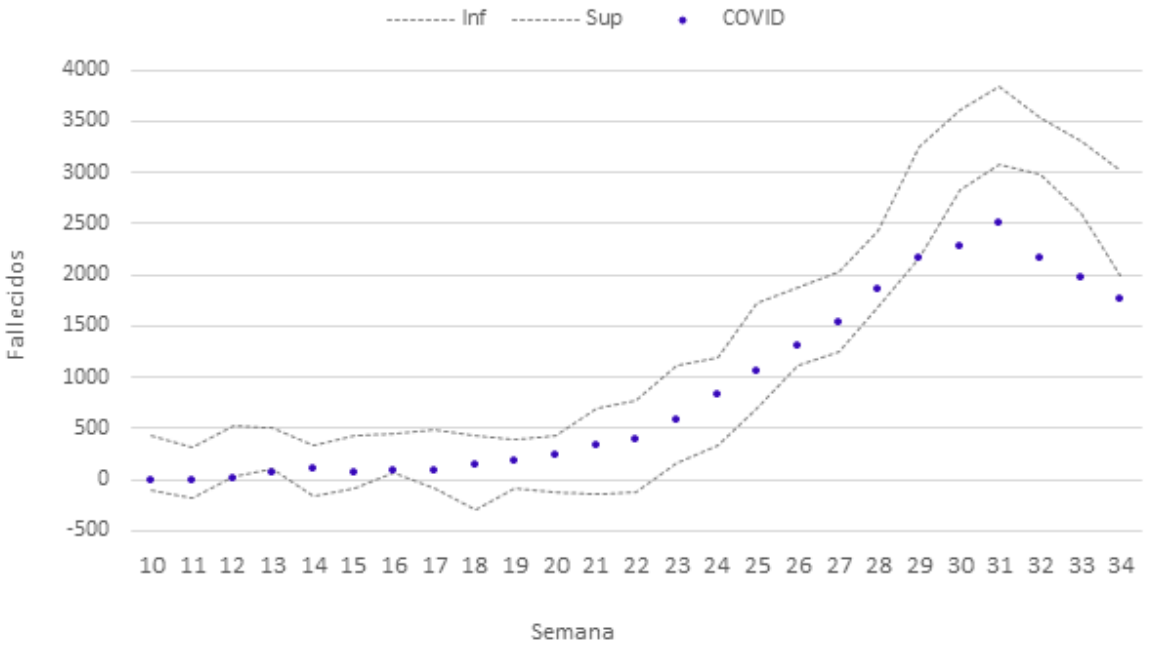


Figura 17: Intervalo de confianza del 95% para el exceso de mortalidad.

La Figura 18, resume el número de fallecimientos por semana, en cada departamento, entre el 2015 y el 2019. Mientras que en el eje X se presenta el promedio de los coeficientes de variación, en el eje Y se presenta el promedio de fallecimientos por semana. En la mayoría de los departamentos, el número de fallecidos semanales tiene niveles de variación bajos, con coeficientes inferiores al 20%. En consecuencia, el exceso de mortalidad de una semana puede estimarse como la diferencia entre el número de fallecidos en 2020 y el promedio observado, para la misma semana, durante los cinco años anteriores. Adicionalmente, once departamentos tienen coeficientes de variación promedio superiores al 20%: Chocó (Ch, 22%), Casanare (Cs, 22%), Putumayo (Pu, 23%), La Guajira (LG, 26%), Arauca (Ar, 25%), San Andrés (SA, 45%), Guaviare (Gv, 47%), Vichada (Vi, 48%), Amazonas (Am, 50%), Guainía (Gu, 75%) y Vaupés (Va, 76%).

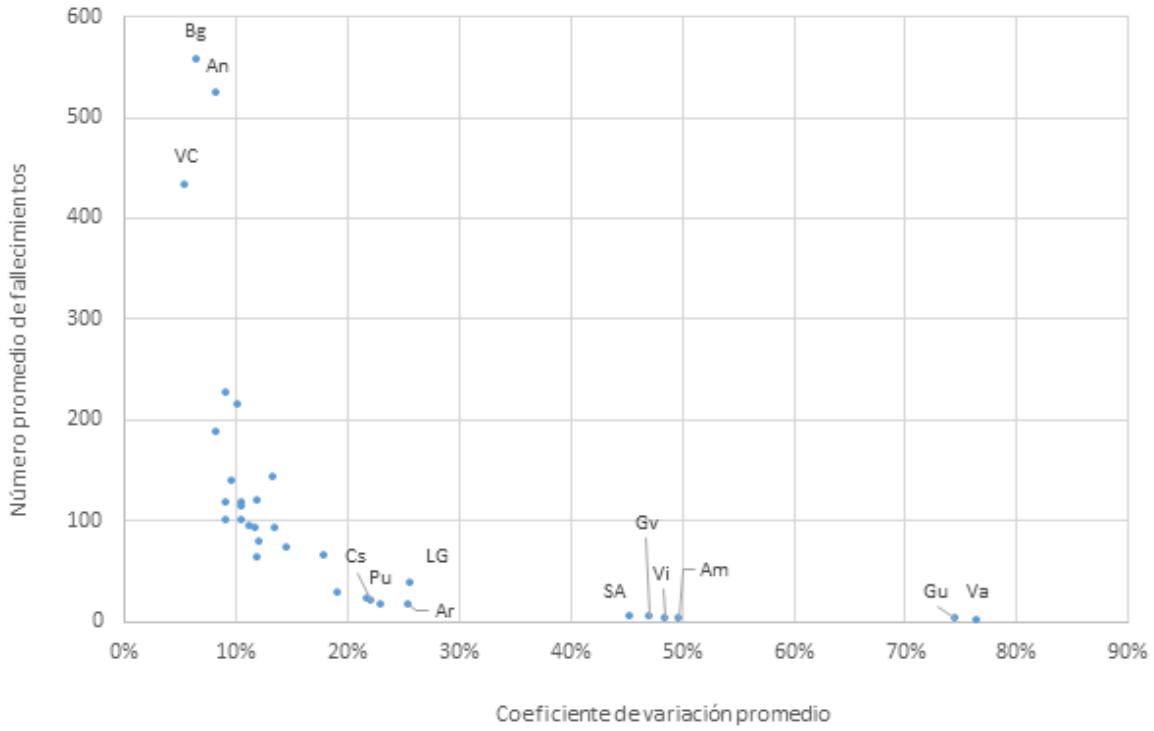


Figura 18: Promedio de los coeficientes de variación entre el 2015 y el 2019 vs. número de fallecimientos por semana, en cada departamento.

La Figura 19 presenta los rangos de los coeficientes de variación de cada departamento. Dado que el análisis es semanal, para cada departamento se calculan 25 coeficientes de variación. Se plantea una clasificación en tres grupos usando el valor máximo de los coeficientes de variación. Si bien en 13 departamentos el mayor coeficiente de variación es inferior al 20%, en ocho departamentos este indicador varía entre el 41% (Putumayo, Pu) y el 141% (Vaupés, Va).

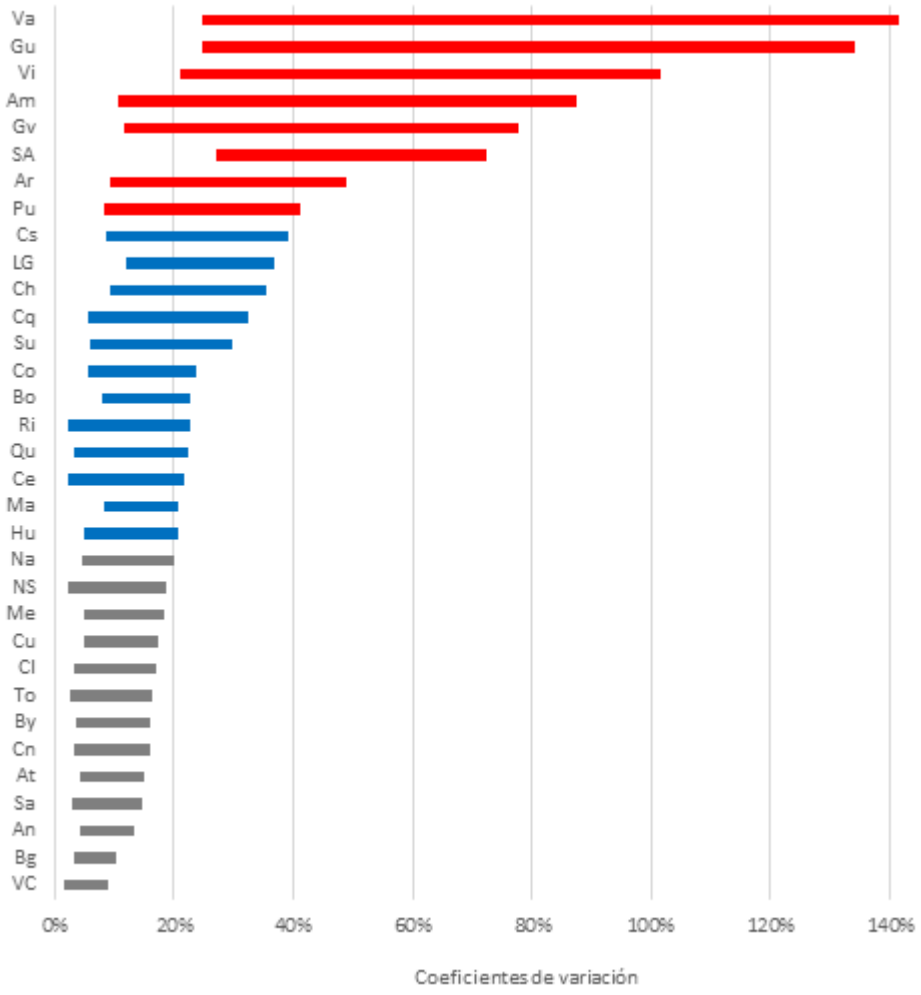


Figura 19: Coeficientes de variación por departamento.

Usando esta información como base, la Figura 20 presenta el exceso de muertes, acumulado a la semana 34, para cada departamento. Los promedios y el indicador de exceso de muertes se calculan semanalmente. En el caso de Bogotá, por ejemplo, entre la semana 10 y la semana 34 del 2020 se habían reportado 5973 fallecimientos adicionales. En seis departamentos, este valor es negativo: Quindío (Qu, -122), Vaupés (Va, -10), Boyacá (Bo, -8), Guainía (Gu, -3), Caldas (Ca, -2) y San Andrés (SA, -2). Esto quiere decir que, en esos departamentos, el número de fallecimientos durante las 25 semanas analizadas en 2020 fue menor que el promedio histórico para el mismo periodo. Adicionalmente, en seis departamentos el exceso de muerte es inferior a 100 fallecidos: Vichada (Vi, 5), Guaviare (Gv, 9), Casanare (Cs, 29), Arauca (Ar, 41), Tolima (To, 88) y Risaralda (Ri, 91).

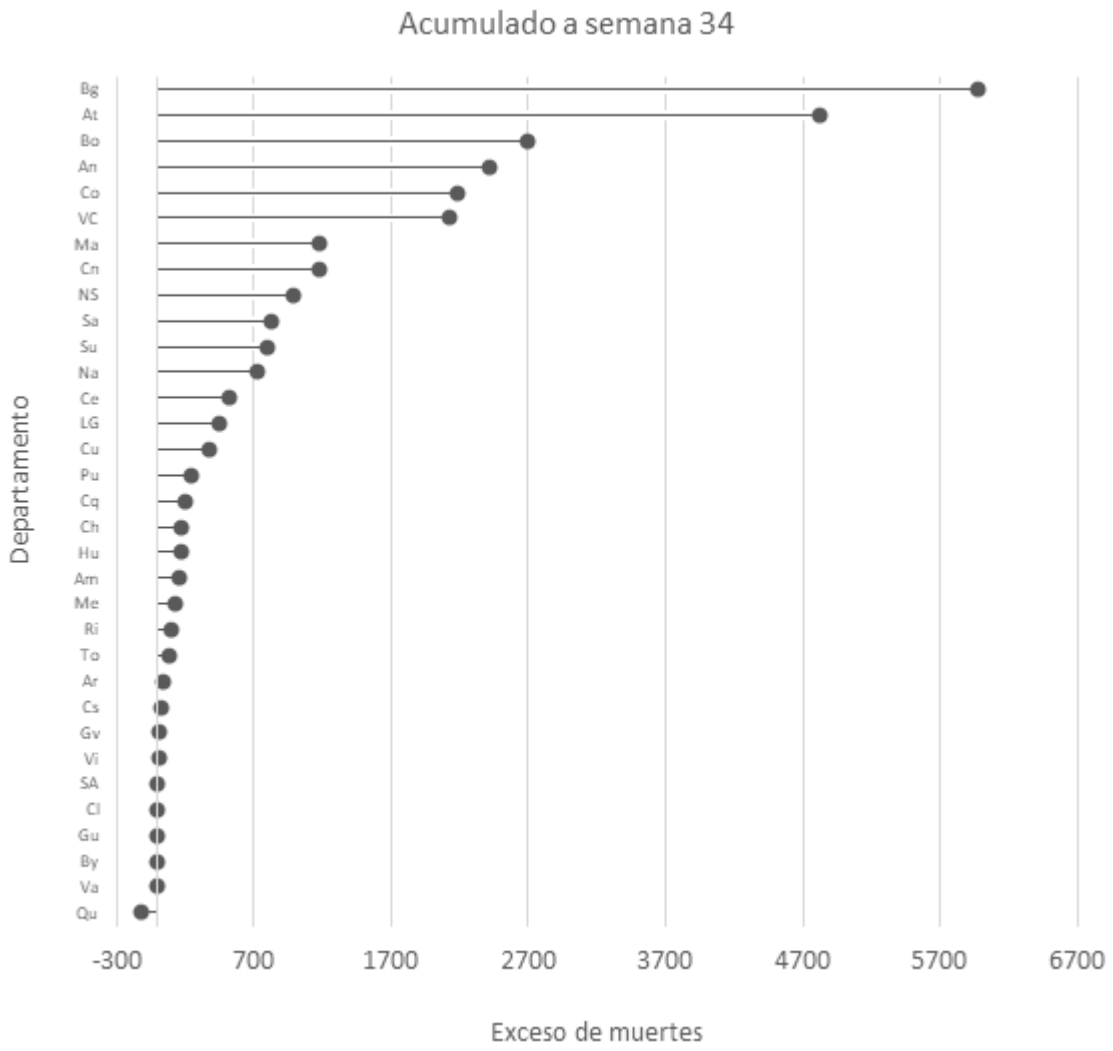


Figura 20: Exceso de muertes acumulado hasta la semana 34

La Figura 21 compara el exceso de muertes, acumulado a la semana 34, con el número de fallecidos por COVID-19, en cada departamento. Por ejemplo, en Bogotá, el exceso de muertes a la semana 34 está entre 4500 y 7446. Adicionalmente, 5844 de estas muertes están asociadas a la COVID-19. Como puede verse, en tres departamentos el número de fallecidos por COVID-19 es inferior a la estimación del exceso de muerte. Esto puede implicar que, en Atlántico, Bolívar y Córdoba hubo un aumento en las muertes naturales no atribuibles al virus. Finalmente, la Figura 22 presenta el exceso de muertes, medido a partir del valor promedio de fallecidos entre 2015 y 2019, y los fallecidos por COVID-19 en cada departamento.

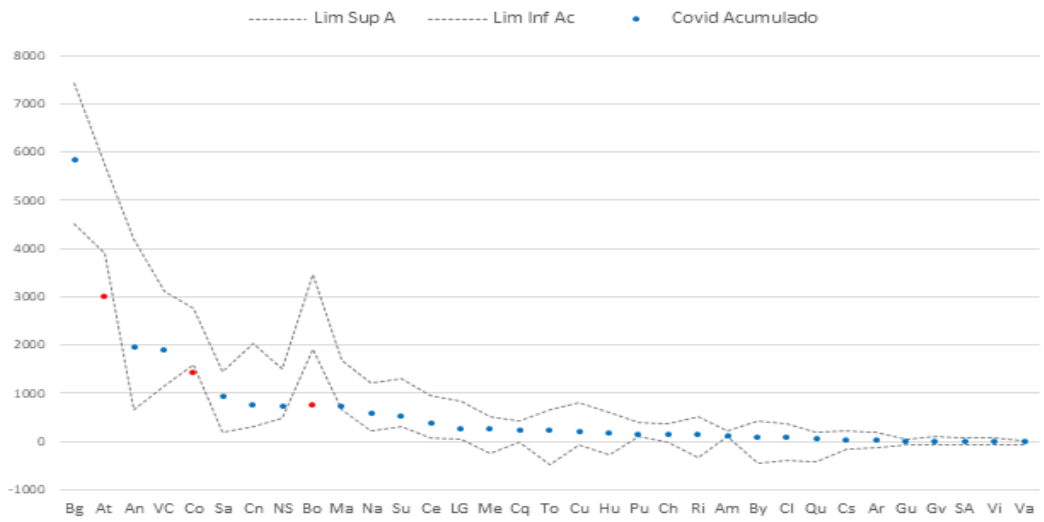


Figura 21: Intervalo de confianza para el exceso de muertes y número de fallecidos por COVID-19 .

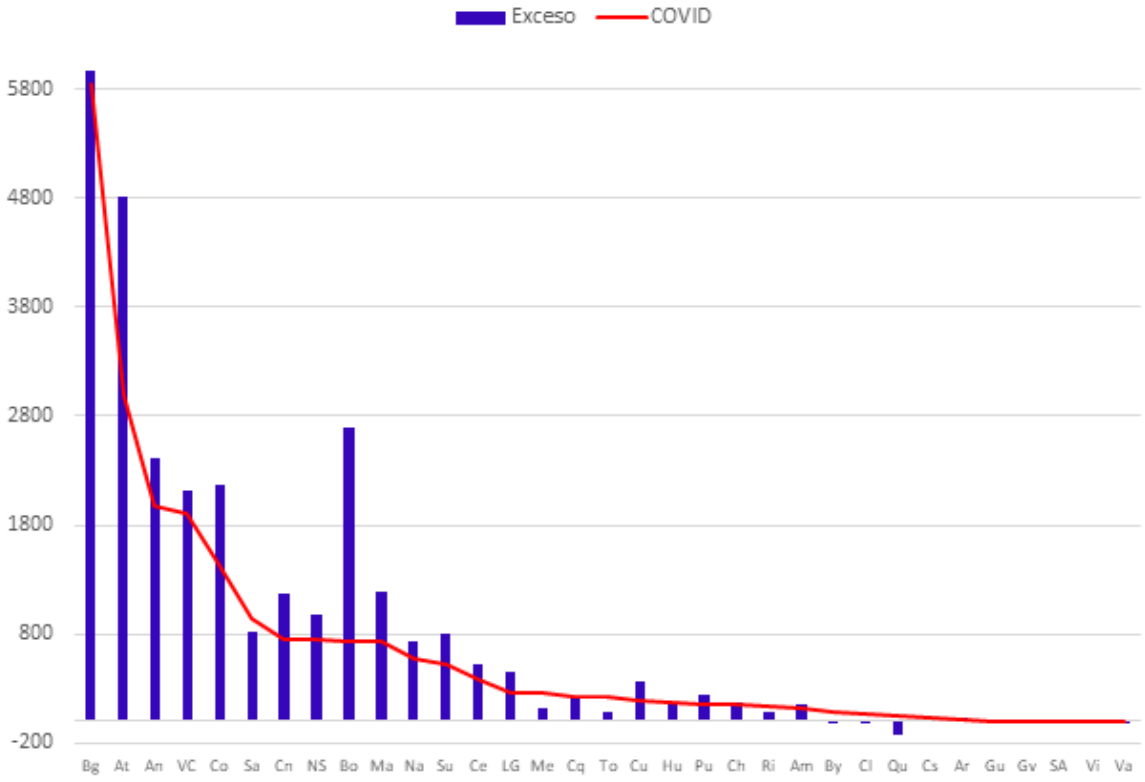


Figura 22: Exceso de muertes, usando el valor promedio, y fallecidos por COVID-19.

5. Modelos de pronóstico

Teniendo como punto de partida las decisiones arrojadas por los análisis descriptivos con respecto a la unidad de agrupación de tiempo y geografía que son respectivamente periodos epidemiológicos y, departamentos, capitales y municipios, se plantea la estrategia para la construcción de los modelos de pronóstico, que se muestra en la Figura 23. Esta estrategia se describe a continuación y se utiliza para cada uno de los eventos de análisis, con unas pequeñas modificaciones para el caso de exceso de mortalidad.

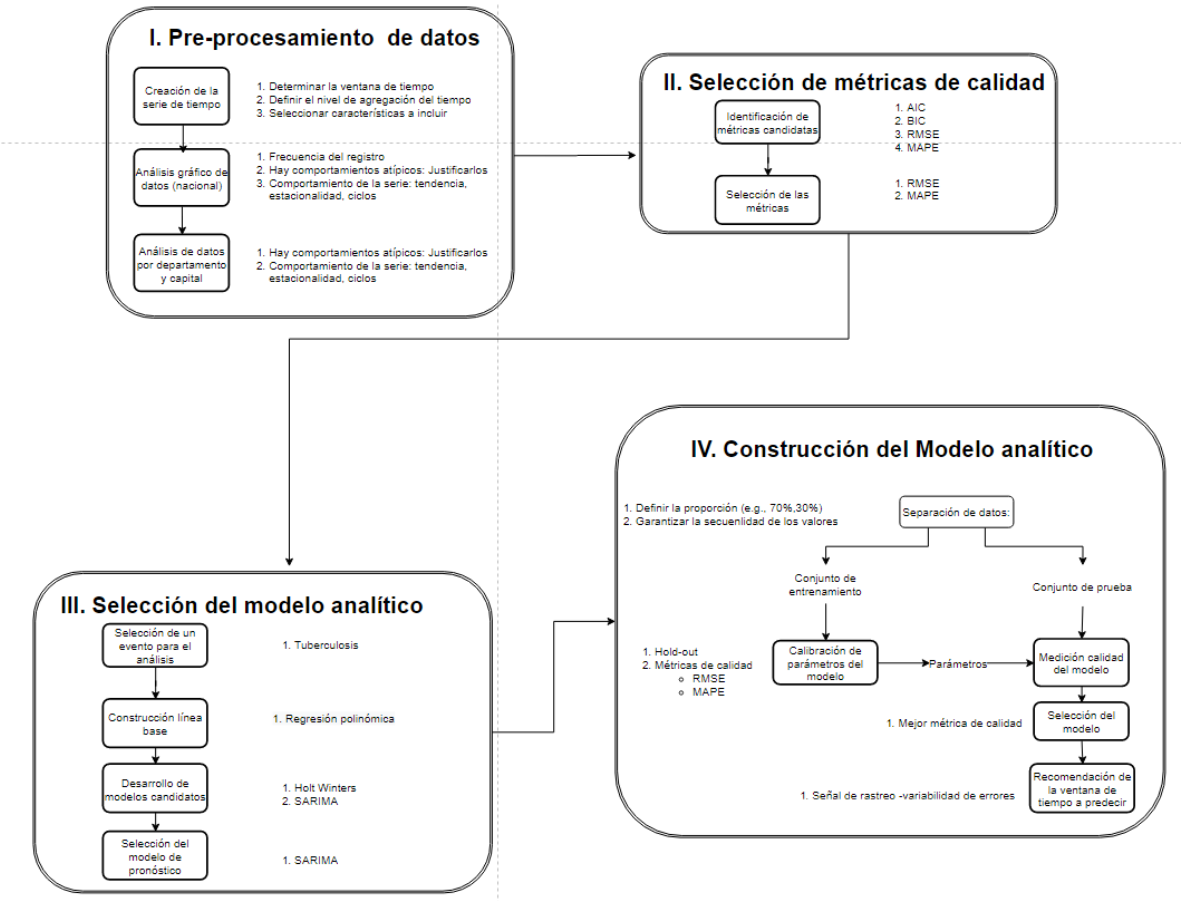


Figura 23 Estrategia de desarrollo de modelos de pronóstico

5.1 Pre-procesamiento de datos

En esta etapa se valida, para un evento, el nivel de agrupación temporal y geográfica, con el fin de identificar grupos para generar los modelos. Las tareas asociadas a esta etapa corresponden a: i. graficar la serie de tiempo, ii. analizar el gráfico de los datos y iii. Analizar los datos por país, departamento y capital.

i. **Graficar la serie de tiempo**: en una primera instancia se grafica la serie a nivel nacional, en un gráfico de líneas, para identificar los años a incluir en los modelos, el comportamiento de la serie, y validar la granularidad en el tiempo con la que se analiza (diario, semanal, mensual o anual).
   Con base en los análisis descriptivos, presentados en el capítulo 4. Análisis descriptivoy de acuerdo con la recomendación del médico experto del grupo y, del equipo del INS, los datos en todos los eventos, se agrupan por periodo epidemiológico, cada 4 semanas continuas, teniendo así, 13 periodos en el año.
ii. **Análisis del gráfico de datos**: En esta etapa la serie se analiza para identificar comportamientos atípicos, la ausencia de información o los valores nulos de la serie, así como el comportamiento general de la serie (estacionaria, tendencial, estacionalidad, ciclos).
iii. **Análisis de datos por departamento y capital**: Una vez identificado el comportamiento de la serie a nivel de país, se procede a realizar el análisis, esta vez por departamento y su capital, tomando los resultados dado por los análisis descriptivos. Igual que en la etapa anterior, se identifican el comportamiento general de la serie y los comportamientos atípicos.

5.2. Selección de métricas de calidad

Las métricas candidatas para la selección del mejor modelo son el error porcentual absoluto medio (MAPE, de las iniciales en inglés), la raíz del error cuadrático medio (RMSE, de las iniciales en inglés), el criterio de información de Akaike (AIC), el criterio de información bayesiano (BIC) y en caso de un posible empate, el tiempo empleado en la calibración del modelo. A continuación, se presentan las definiciones asumidas en el proyecto.

- **MAPE**: Error porcentual absoluto medio que indica el porcentaje de error que tiene el modelo de pronóstico.
- **RMSE:** Raíz del error cuadrático que corresponde a la medida de la variabilidad del modelo de pronóstico.
- **AIC** y **BIC**: Criterios de información que permiten evaluar el ajuste del modelo de pronóstico y penalizarlo por el número de parámetros utilizados. Permite decidir entre dos modelos el mejor y el más simple (con menos parámetros estimados). AIC utiliza el criterio de información de Akaike y el BIC usa el criterio de información Bayesiano (o de Schwarz). Entre más pequeños sean los valores, el modelo de pronóstico será mejor.

Adicionalmente, se construye el indicador de Señal de rastreo para cada valor pronosticado. Este indicador es una medida de desempeño del pronóstico, permite medir para un periodo, la desviación del pronóstico respecto a variaciones en la demanda. Análogamente se puede leer

como el número de MAD (Desviación Media Absoluta o *Mean Absolute Deviation*) en el que un valor pronosticado se aleja de la realidad. Para el objeto del presente estudio, se reporta el número de periodos que la señal de rastreo de un pronóstico, permanece dentro de los límites de ±4 MAD, que es el valor recomendado por la literatura.

Tomando como base estas medidas de desempeño, se utilizan el MAPE y la cantidad de periodos en la señal de rastreo para la calibración de los modelos de pronóstico. Adicionalmente, se definieron como hiperparámetros para la ejecución de los modelos, un umbral mínimo del MAPE (para evitar el sobre ajuste –*overfitting*- de los modelos), y un número mínimo de periodos de señal de rastreo. Es así como se tomó 4.0 para el umbral del MAPE de Tuberculosis, Mortalidad Infantil, Intento de Suicidio y EDA y, de 5.0 para Exceso de Mortalidad y Diabetes Mellitus. Estos valores son configurables para cada enfermedad desde el archivo `config.json` que sirve como entrada al motor predictivo. Los valores de ese hiperparámetro se seleccionaron empíricamente en la fase de entrenamiento y pruebas. A nivel de periodos de señal de rastreo, se definió que debe ser un valor mayor a 0, es decir, mínimo 1 periodo disponible de tolerancia.

5.3 Selección de los modelos de pronóstico a utilizar

El enfoque usado para la construcción de los modelos de pronósticos es basado en series de tiempo. Inicialmente, como lo indica el procedimiento general definido, para entender el comportamiento y naturaleza de la serie de datos en el tiempo, se realizan las pruebas estadísticas de estacionalidad (prueba de raíz unitaria de Dickey-Fuller aumentada), el análisis de correlación y autocorrelación de los datos y, por último, se comprueba la descomposición estacional.

La estrategia inicia con la selección de un evento que será utilizado para tomar decisiones como el modelo a construir, en este caso tuberculosis, y continúa con la construcción de un modelo de línea base que permite identificar y seleccionar el modelo más apropiado para los diferentes eventos.

La línea base se calcula usando regresiones polinómicas (grado 3 y grado 4). Para el ajuste de los parámetros se calcula la pendiente y coeficientes para los cuales la regresión ofrece el mejor ajuste. La razón por la cual se seleccionan estos modelos como de línea base, es debido a que las regresiones lineales o polinomiales reproducen la tendencia de los datos, pero no el comportamiento variable de los mismos. En el caso de los eventos que se analizan, es importante también capturar la variabilidad de los datos.

Una vez se tiene el modelo línea base se comparan los métodos de Holt-Winters o de suavizado exponencial triple con el de SARIMA para seleccionar el más apropiado entre estos.

El método de Holt-Winters es un método de pronóstico de series de tiempo que tiene en cuenta la tendencia y la estacionalidad de las series. Este método y específicamente, el de triple exponente suavizado, requiere los parámetros *alpha, beta, gamma, phi, trend, damped, seasonal, seasonal periods y boxcox,* los cuales fueron calculados utilizando una búsqueda exhaustiva, que permita identificar la combinación que genere el mejor rendimiento.

El modelo SARIMA pertenece a la familia de Box-Jenkis y se basa en los modelos ARIMA (De las iniciales en inglés: ***Autoregressive integrated moving average***), los cuales utilizan la correlación que existe entre los datos. Al igual que el método de Holt-winters, este modelo incluye un componente estacional.

Los modelos SARIMA cuentan con siete parámetros así que deben ser ajustados para obtener el mejor modelo. De la misma forma que para el ajuste de los parámetros en el modelo Holt-Winters, en el modelo SARIMA se utiliza una búsqueda exhaustiva variando cada parámetro entre los valores {0, 1, 2}, generando así 729 ajustes diferentes, de los cuales se selecciona el que maximice la medida de rendimiento (MAPE y el número de periodos de la señal de rastreo).

Durante la fase de pruebas, nos dimos cuenta de que los modelos generados por Holt-Winters eran buenos, pero rara vez superaban a los creados por SARIMA, al compararlos con respecto a las métricas MAPE, RMSE, AIC y BIC. Es así como se decide trabajar con modelos SARIMA.

5.4 Construcción de los modelos de pronóstico

En este punto se parte de la decisión de construir modelos SARIMA, para los cual los datos de análisis, se dividen en dos grupos, los datos para el entrenamiento y los datos para la prueba. En este caso haciendo variaciones del 70%-30%, 80%-20% y 85%-15% con el fin de determinar la mejor partición para todos los eventos de análisis.

Con los datos para entrenamiento se calibra el modelo y se utilizan las métricas de calidad de RMSE (raíz del error cuadrático medio) y MAPE (Error porcentual absoluto medio) para comparar los resultados. El modelo seleccionado, es aquel cuyos parámetros arrojen las mejores métricas de calidad en el conjunto de entrenamiento.

Con el ánimo de establecer los periodos epidemiológicos para los cuales el pronóstico realizado tiene mayor nivel de certidumbre, se propone utilizar un gráfico de señal de rastreo. El valor utilizado en el proyecto para la señal de rastreo, indica el número de valores pronosticados que están dentro de un rango de 3,2 desviaciones estándar, antes de que el primer pronóstico salga de este rango.

5.5 Modelos de pronóstico para el caso de tuberculosis

En esta sección se describe la etapa III que se muestra en la Figura 23. De acuerdo con la estrategia descrita en secciones previas, se seleccionó la Tuberculosis como evento para analizar en detalle y tomar decisiones de parámetros y modelos a construir. En este punto, a diferencia de los datos utilizados para el análisis descriptivo, se tomó la información semanal desde el año 2015 al primer trimestre del 2020. Esta información se encuentra discriminada entre meningitis tuberculosa (código 530), tuberculosis extra-pulmonar (código 810), pulmonar (código 820) y fármaco-resistente (código 825), para los datos del 2015 al 2018, mientras que para el 2019 y 2020, se presentan dos tipos, la tuberculosis resistente y sensible (código 813).

Siguiendo la recomendación del médico experto del grupo, y después de compartir con el equipo del Instituto Nacional de Salud (INS) los tipos de tuberculosis a incluir en los modelos, se decide que para analizar el evento de tuberculosis se agruparán los cuatro tipos reportados para los años 2015 a 2018 y los dos tipos reportados, para los años 2019 y 2020.

A pesar de que para el caso de Tuberculosis se poseen datos desde el 2015, el comportamiento y la media de los datos de los años 2015 y 2016 es muy distinta a la de los años 2017 en adelante, por lo cual, se decidió que para este evento se consideraría únicamente la información de los años comprendidos entre el 2017 al 2020. Un análisis similar, se realiza con cada evento para determinar los años que se deben incluir en los pronósticos.

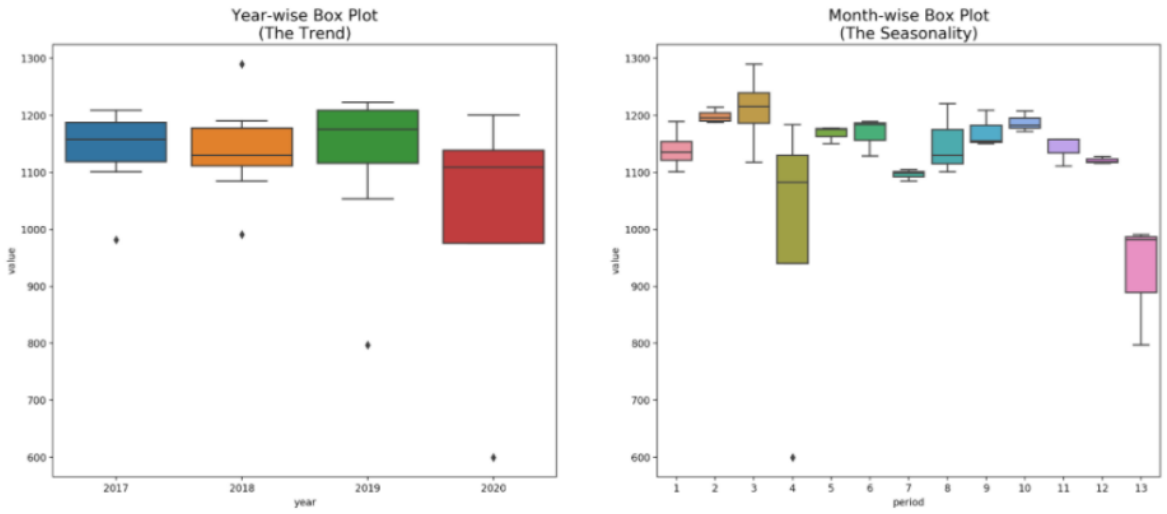


Figura 24 Boxplot con la distribución de los datos por cuartiles por años y por periodos epidemiológicos.

En la Figura 1Figura 24, se observa la distribución de los datos, comparándolos por años y por periodo epidemiológico, a partir del 2017. En la gráfica de la izquierda, se aprecia la similitud de la media y la desviación estándar para los periodos observados, 2017, 2018 y 2019, con valores respectivamente de (`1146, 57.68`), (`1140, 65.56`), (`1137, 110`). Los datos del 2020 reflejan una disminución del 11,6% con respecto al año inmediatamente anterior, esto puede ser por diferentes razones: (1) porque del 2020 aún no se han reportado todos los casos ocurridos durante el primer trimestre, o (2) pudo presentarse una disminución de los casos debido al COVID-19. Esto sugirió, que los datos del 2020 no serían incluidos para los análisis de pronósticos sin COVID-19.

Finalmente, se revisa el comportamiento de la serie a nivel nacional, departamental y municipal siguiendo los resultados del análisis descriptivo y se valida la calidad de información para ese nivel de agregación geográfica. Con esto, se construye el modelo de línea base y los modelos candidatos para seleccionar el mejor modelo a construir para este y los demás eventos incluidos en el análisis.

Debido a la cantidad de datos utilizados para crear las series de tiempo (aproximadamente 39), se probó utilizar el 80% de los datos para el entrenamiento (aproximadamente 31 datos) y el 20% para la validación (8 datos). Se exploró utilizar una distribución de 90-10 % pero se observó una subestimación del error porcentual absoluto, ya que solo se validaban los modelos contra los últimos 4 periodos epidemiológicos del 2019. Esto permitió concluir, que para los diferentes eventos se utilizaría una proporción 80%-20% para los conjuntos de datos de entrenamiento y prueba.

Para los datos de tuberculosis en Colombia (agregados de los departamentos) en el escenario sin COVID-19, se probaron los tres modelos con las diferentes combinaciones de parámetros. Las Figura 25 a Figura 27 presentan el detalle de los tres modelos construidos y, la Tabla 3 resume los resultados obtenidos. Hay que recordar que, el modelo de regresión polinomial es el modelo para construir la línea base.

Utilizando al MAPE (menor valor), como criterio de selección y revisando visualmente, el resultado del pronóstico para el caso de SARIMA, este reproduce los picos inferiores observados en noviembre y diciembre de cada año, por lo que se decide realizar todo el proceso de construcción de modelos utilizando los modelos SARIMA.

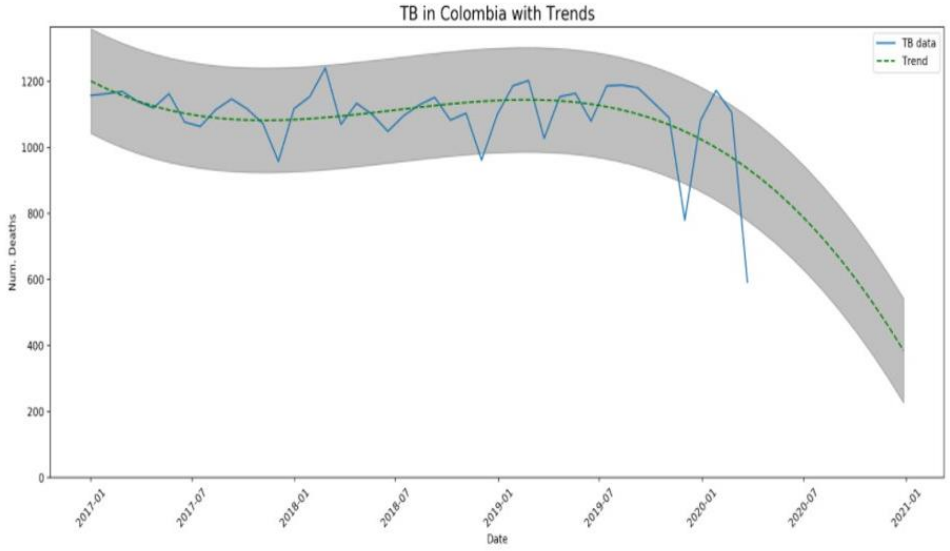


Figura 25 Modelo de línea base – regresión

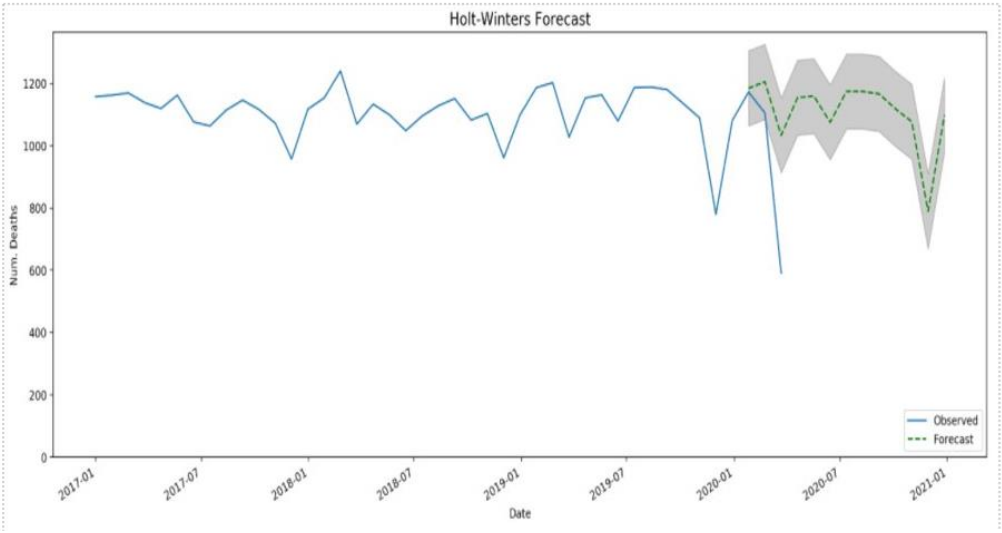
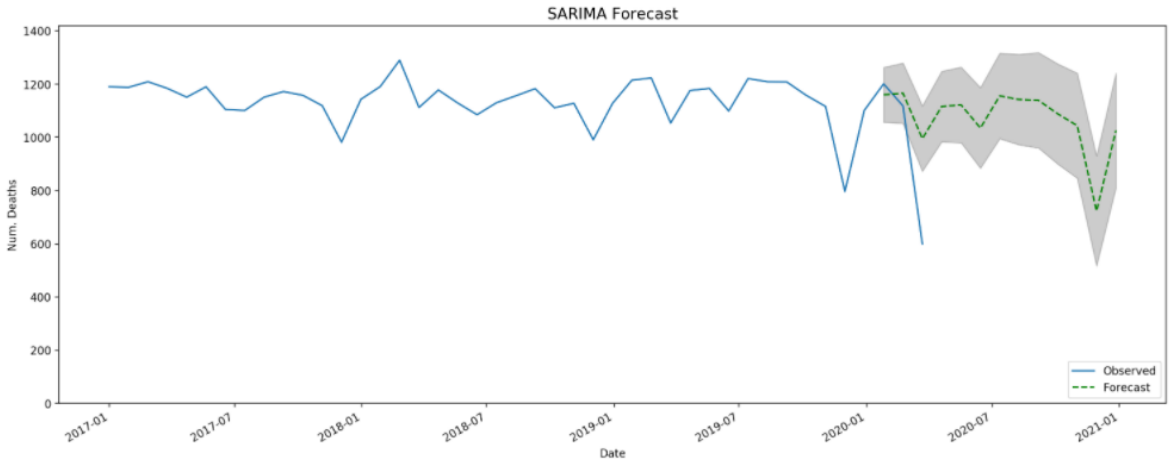


Figura 26 Modelo de pronóstico– Holt-Winters



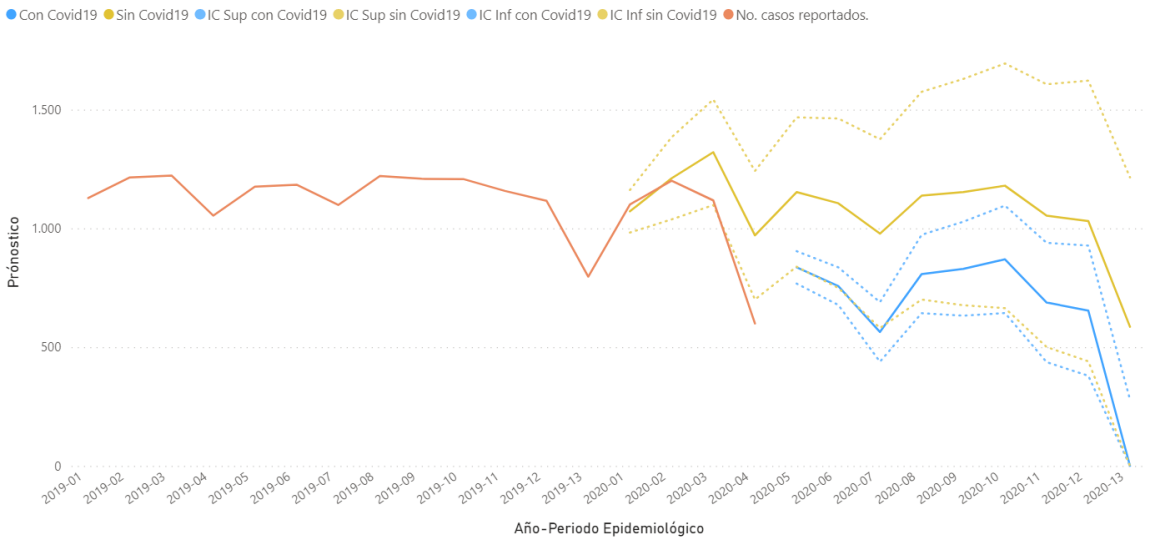


Figura 27 Pronóstico para tuberculosis en Colombia usando el método SARIMA.

	Regresión polinomial de grado 3 (Línea base)	Holt-Winters	SARIMA
MAPE	6,72%	6,16%	***5,88%***
RMSE	96,32	***83,46***	85,05
AIC	402,82	344,48	252,26
BIC	411,63	374,88	255,53

Tabla 3 Resultados de las medidas de desempeño para cada método de pronóstico

En la Figura 28 se puede observar las métricas obtenidas para los diferentes métodos usados, con los mejores ajustes de parámetros. Para facilitar la comparación se usa un diagrama de radar, donde se valida y justifica la selección del modelo SARIMA.

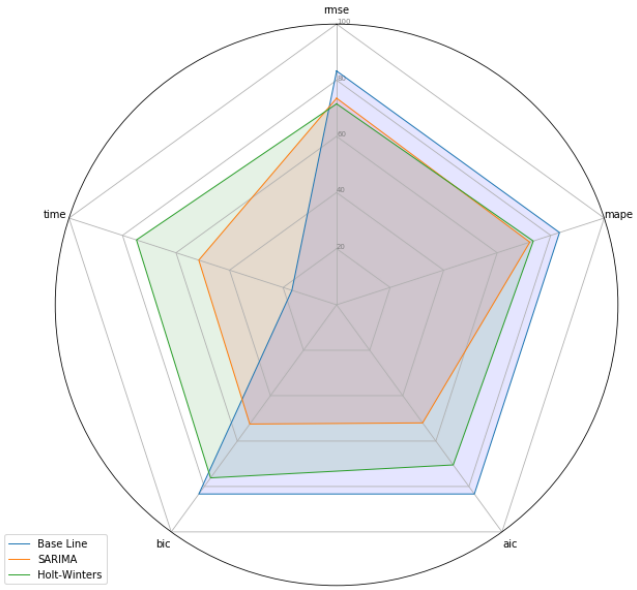


Figura 28 Comparación visual entre los mejores modelos por método

Los resultados de los modelos obtenidos están directamente relacionados con los parámetros utilizados. Es así como siguiendo la estrategia propuesta para la selección de los mejores parámetros, se obtienen los resultados similares a los presentados en la Tabla 4. En este caso, por ejemplo, utilizando el MAPE como criterio de selección, se observa que los valores apropiados para los parámetros p, d, q, Sp, Sd y Sq son respectivamente 2.0, 0.0, 2.0, 2.0, 0.0 y 2.0.

method	p	d	q	Sp	Sd	Sq	freq	rmse	mape	aic	bic	var_coef _diff
SARIMA	2.0	0.0	2.0	2.0	0.0	2.0	13.0	10.267	14.122	795.104	852.619	0.0073
SARIMA	1.0	0.0	1.0	0.0	1.0	1.0	13.0	1.558	18.125	685.755	714.077	0.2797
SARIMA	2.0	0.0	2.0	2.0	0.0	2.0	13.0	15.731	33.125	856.507	914.022	0.0731
SARIMA	2.0	1.0	0.0	2.0	0.0	1.0	13.0	0.3536	41.667	330.921	369.264	0.8083
SARIMA	1.0	1.0	2.0	1.0	0.0	2.0	13.0	741.314	59.808	1.566.868	1.606.414	0.0748
SARIMA	1.0	0.0	2.0	2.0	0.0	1.0	13.0	33.983	70.167	962.062	1.016.143	0.1001
SARIMA	2.0	1.0	2.0	2.0	0.0	2.0	13.0	67.024	79.737	1.067.583	1.118.428	0.0537

Tabla 4 Resultados de modelos de pronóstico – SARIMA para seleccionar el valor de los parámetros

Adicional a los parámetros descritos previamente, en la Tabla 4, se observa la columna **var_coeff_diff**, que representa la diferencia entre el coeficiente de variación de los datos usados para entrenamiento vs el coeficiente de variación de los datos pronosticados. El objetivo de esta métrica es validar si se mantiene la tendencia de los datos originales, por lo cual permite evaluar entre dos modelos de pronóstico, aquel con menor diferencia. Estas tablas se generan durante la

construcción de los modelos, por lo cual, en el futuro los responsables de esta etapa del proceso, podrán explorar nuevos criterios de selección modelos y sus parámetros.

Continuando con la etapa III de la Figura 23, una vez construido el modelo SARIMA, se procede a calcular la señal de rastreo. En este caso específico, los resultados obtenidos sugirieren confiabilidad en los tres primeros periodos de pronóstico, más no en toda la serie, ya que el límite establecido era máximo 3,2 desviaciones estándares del error.

Period	Fecha	Real	Pronóstico	Error Abs.	MAPE	Sum Error Abs.	DMA	Error	Sum Error	Señal Rastreo	UCI	LCI
1.0	6/16/2019	1099	1097	2	0.18%	2.0	2.0	2	2	1.0	3.0	-3.0
2.0	7/14/2019	1221	1144	77	6.31%	79.0	39.5	77	79	2.0	3.0	-3.0
3.0	8/11/2019	1221	1171	50	4.10%	129.0	43.0	50	129	3.0	3.0	-3.0
4.0	9/8/2019	1208	1200	8	0.66%	137.0	34.3	8	137	**4.0**	3.0	-3.0
5.0	10/6/2019	1158	1129	29	2.50%	166.0	33.2	29	166	5.0	3.0	-3.0
6.0	11/3/2019	1116	1147	31	2.78%	197.0	32.8	-31	135	4.1	3.0	-3.0
7.0	12/1/2019	797	1012	215	26.98%	412.0	58.9	-215	-80	-1.4	3.0	-3.0
8.0	12/29/2019	1101	1150	49	4.45%	461.0	57.6	-49	-129	-2.2	3.0	-3.0

Tabla 5. Resultado de las medidas de desempeño para la calibración de los datos para Colombia.

Sobre los resultados obtenidos de la señal de rastreo, reportados en la

Period	Fecha	Real	Pronóstico	Error Abs.	MAPE	Sum Error Abs.	DMA	Error	Sum Error	Señal Rastreo	UCI	LCI
1.0	6/16/2019	1099	1097	2	0.18%	2.0	2.0	2	2	1.0	3.0	-3.0
2.0	7/14/2019	1221	1144	77	6.31%	79.0	39.5	77	79	2.0	3.0	-3.0
3.0	8/11/2019	1221	1171	50	4.10%	129.0	43.0	50	129	3.0	3.0	-3.0
4.0	9/8/2019	1208	1200	8	0.66%	137.0	34.3	8	137	**4.0**	3.0	-3.0
5.0	10/6/2019	1158	1129	29	2.50%	166.0	33.2	29	166	5.0	3.0	-3.0
6.0	11/3/2019	1116	1147	31	2.78%	197.0	32.8	-31	135	4.1	3.0	-3.0
7.0	12/1/2019	797	1012	215	26.98%	412.0	58.9	-215	-80	-1.4	3.0	-3.0
8.0	12/29/2019	1101	1150	49	4.45%	461.0	57.6	-49	-129	-2.2	3.0	-3.0

Tabla 5, se concluye que de ocho periodos pronosticados, el pronóstico del periodo 5 se encuentra por fuera del intervalo de ±4MAD y los otros siete periodos están en el intervalo de tolerancia. Esto indica que el modelo planteado está en control y que puede usarse para predecir el comportamiento de la serie.

De acuerdo con los resultados obtenidos para este evento en la construcción de los modelos, se tendrá: un modelo a nivel nacional por cada evento para el caso Con COVID-19 (siete modelos, uno por evento) y, un modelo para seis de los eventos, excluido exceso de mortalidad, para el caso Sin COVID -19, para un total de 13 modelos a nivel nacional. A nivel de departamentos y capitales se construye un modelo para cada uno de ellos (`(7x32)*2=448`) para el caso con COVID -19 y (`(6x32)*2=384`) para el caso Sin COVID -19, para un total de `845` modelos de pronóstico construidos.

6. Conclusiones y recomendaciones

En este proyecto se plantea una estrategia de análisis de la información disponible sobre eventos en saludo, con el ánimo de evaluar el efecto del COVID-19 en eventos de salud diferentes a la enfermedad por coronavirus.

La estrategia planteada incluye cuatro etapas generales: 1. Recolección y preparación de datos, 2. Desarrollo de modelos analíticos, 3. Medición de indicadores de impacto, y 4. Despliegue del modelo. El desarrollo de modelos analíticos incluye el análisis descriptivo y el desarrollo de modelos de pronóstico.

El modelo planteado fue alimentado con información de diferentes bases de datos disponibles al público para generar una serie de análisis descriptivos y de pronóstico aplicados a seis eventos distintos (Tuberculosis, mortalidad infantil, EDA, intento de suicidio, Diabetes mellitus y exceso de mortalidad), entre los cuales se incluye un análisis por exceso de mortalidad.

La estrategia planteada, reflejó ser flexible y ajustable no solo a los eventos de salud analizados en este proyecto, lo que permite considerar la inclusión de otros eventos en salud en el futuro.

- Aunque cada uno de los eventos en salud analizados tiene un comportamiento diferente y propio a sus características, en este proyecto se propone una única estrategia para la construcción de modelos de pronósticos que permite incluir estas características y obtener medidas de error e indicadores de desempeño más que aceptables (MAPE inferiores al 20% y Señales de rastreo estables superiores a un periodo de pronóstico).
- Los modelos y resultados presentados cubren los seis eventos analizados con datos reportados hasta marzo de 2020, sin embargo, el software desarrollado para apoyar la estrategia general de construcción de modelos, permite incluir nuevos datos asociados a esos eventos, con el fin de mantener actualizado el resultado del proyecto. De igual manera, es posible utilizarlos para incluir nuevos eventos, para los cuales se recomienda seguir la estrategia que guía las decisiones a tomar en cada etapa, con el fin de mejorar la calidad de los resultados obtenidos.

Se proponen siete indicadores indirectos de impacto que, a pesar de no tener valores contundentes en la actualidad, por la ausencia de información, son de fácil estimación e interpretación.

Las contribuciones de este proyecto se concentran en dos aspectos:

- El uso de información existente y disponible en bases de datos públicas
- El análisis estadístico de los datos permitió identificar, que la información existente permite realizar pronósticos y descripciones a nivel de departamento y capitales, por periodo epidemiológico, agregando valor a las estadísticas actualmente reportadas a nivel nacional y por año.

Los resultados de los análisis descriptivos, modelos de pronóstico e indicadores indirectos de impacto, se visualizan en una herramienta desarrollada en el marco del proyecto, que permite hacer análisis comparativos a nivel nacional, departamental y por capitales, en periodos epidemiológicos.

Recomendaciones

- Si bien la opción de utilizar información pública, en particular del SIVIGILA permitió el desarrollo de este proyecto, existe el riesgo de incluir variaciones artificiales en los datos. Las fechas de los eventos, reportadas en estos datos, pueden verse afectadas por el proceso de registro en el sistema de información. En consecuencia, es necesario tener acceso a los datos fuente para verificar si es posible eliminar dicha variación. Por lo cual, los resultados obtenidos en los análisis descriptivos y modelos de pronóstico deben ser usados de manera cautelosa para formular conclusiones.
- El funcionamiento correcto de los modelos provistos requiere de una transformación de las fuentes originales a un formato estándar, la cual debe ser realizada de forma manual por parte de los funcionarios responsables de mantener la aplicación.