Contenido

Modelos de pronóstico para la tuberculosis

Datos utilizados

Se utilizan los datos abiertos para la tuberculosis en Colombia, publicados por el SIVIGILA en su portal web. Se descargan y estructuran los datos por departamentos y capitales. Para determinar el número de casos para toda Colombia, se agregan los datos de los departamentos.

En SIVIGILA los datos están reportados por año-semana y discriminan entre los diferentes tipos de tuberculosis reportados: para el 2020 y 2019, tuberculosis resistente y sensible (2 tipos) y para el 2018 hacia atrás, meningitis tuberculosa, tuberculosis extrapulmonar, pulmonar y farmacorresistente (4 tipos distintos). Sin embargo, para nuestros análisis, basados en los análisis descriptivos y la recomendación del experto, se decidió agregarlos por periodo epidemiológico (cada 4 semanas continuas, teniendo así, 13 periodos en el año) y sin discriminar entre los diferentes tipos de tuberculosis existentes. Generando así, 13 datos por año para cada entidad, ya sea toda Colombia, los departamentos o las capitales. Esta decisión, mejoró considerablemente la precisión de los modelos generados, al disminuir la varianza de los datos.

Por último, aunque se recolectaron datos desde el 2015 en adelante, experimentalmente se decidió trabajar solo con los años comprendidos entre el 2017 al 2020, ya que el comportamiento y la media de los datos de los años 2015 y 2016, para la tuberculosis, es muy distinta a la de los años 2017 en adelante.

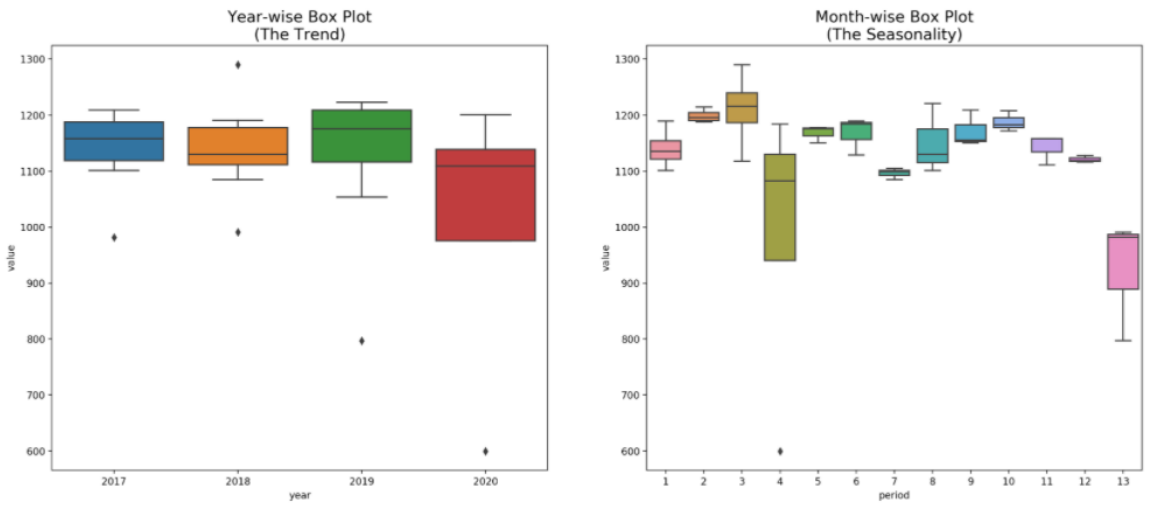

Figura 1. Boxplot con la distribución de los datos por cuartiles por años y por periodos epidemiológicos.

En la figura 1, se observa en 2 *bloxplot* la distribución de los datos seleccionados (a partir del 2017) por años y por periodos epidemiológicos. Se observa la similitud entre la media y la desviación estándar en los datos de los años del 2017, 2018 y 2019. Claramente, los datos del 2020 son muy distintos, ya sea porque aún no se han reportado todos los casos ocurridos durante esos meses o porque hubo una disminución significativa de los casos en el 2020 debido al COVID-19. Esto sugirió, que los datos del 2020 no serían incluidos para los análisis de pronósticos sin COVID-19.

Creación de los modelos

Se usa un enfoque basado en series de tiempo para crear y seleccionar los modelos que reproducen el comportamiento de la cantidad de casos de tuberculosis en Colombia y en los departamentos, por lo tanto, inicialmente, se realizan pruebas estadísticas de estacionariedad (prueba de raíz unitaria de Dickey-Fuller aumentada), también se realizan análisis de correlación y autocorrelación de los datos y por último se comprueba la descomposición estacional; todas estas pruebas se realizan con el objetivo de entender el comportamiento y naturaleza de la serie de tiempo en análisis, y tener criterios de selección de parámetros, en caso de tener que escoger el modelo optimo entre un grupo de modelos con precisiones similares.

La línea base se calcula usando regresiones polinómicas (grado 3 y grado 4). Posteriormente, se exploran principalmente 2 algoritmos: SARIMA, de la familia de Box-Jenkins y, Holt-Winters, que es un método de pronóstico de triple exponente suavizado. Para ambos métodos, se usa la técnica de búsqueda de rejilla o búsqueda exhaustiva (Grid Search), para encontrar la combinación de parámetros para los cuales el modelo ofrece el mejor rendimiento (el menor error posible sin caer en el sobreajuste). Las métricas usadas para la selección del mejor modelo son el error porcentual absoluto medio (MAPE), la raíz del error cuadrático medio (RMSE), 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.

Experimentalmente se obtuvo que, para la enfermedad tuberculosis el método SARIMA ofrece mejores resultados que el método Holt-Winters, y estos a su vez, que la regresión polinómica, sin embargo, es mucho más fácil seleccionar el mejor modelo para Holt-Winters, que para el método SARIMA, ya que no basta con seleccionar el que tenga el menor MAPE o RMSE, sino, el que tenga el mejor error posible, pero a la vez, no caiga en el sobreajuste, para lo cual usamos los indicadores AIC y BIC (en caso de empate).

Con respecto a los datos usados para crear los modelos, se usan los datos reportados por el SIVIGILA semanalmente, durante el 2017 y 2020, agregados por periodos epidemiológicos (4 semanas continuas). Para el caso específico del pronóstico sin COVID-19, no se incluyen los datos del 2020 para el entrenamiento y validación del modelo, ya que modifica notablemente (hacia la baja) la tendencia del pronóstico y la precisión del modelo.

Debido a la cantidad de datos utilizados para crear las series de tiempo (aproximadamente 39), se utiliza una distribución de 80% de los datos para el entrenamiento (aproximadamente 31 datos) y 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.

Por último, se utiliza la técnica de señal de rastreo del pronóstico, para medir y estimar la cantidad de periodos para los cuales el pronóstico realizado es relevante (menor o igual a 3 desviaciones estándares).

Parámetros de los modelos

Para la regresión polinomial se crean modelos de grado 3 y grado 4 y, se calcula la pendiente y coeficientes para los cuales la regresión ofrece el mejor ajuste.

Para el método SARIMA, que es un método ARIMA con componentes de temporada, se itera entre los parámetros p, d, q, P, D, Q y, para cada parámetro, se evalúan valores entre {0, 1, 2}, generando así 729 modelos, de los cuales, se selecciona el mejor. A continuación, la descripción de cada parámetro:

- p: orden autorregresivo.
- d: orden de diferenciación.
- q: orden del promedio móvil.
- P o Sp: orden autorregresivo estacional.
- D o Sd: orden de diferenciación estacional.
- Q o Sq: orden del promedio móvil estacional.
- f: frecuencia anual de los datos (se utiliza 13, debido a que hay 13 periodos epidemiológicos por año).

Para le método Holt-Winters, específicamente, para el de triple exponente suavizado, se itera entre los parámetros alpha, beta, gamma, phi, trend, damped, seasonal, seasonal periods y boxcox. Este modelo cuenta con hiperparámetros que controlan la naturaleza del exponencial realizado para la serie, tendencia y estacionalidad. A continuación, el detalle de los hiperparámetros:

- alpha o smoothing level: el coeficiente de suavizado para el nivel.
- beta o smoothing slope: el coeficiente de suavizado de la tendencia.
- gamma o smoothing seasonal: el coeficiente de suavizado para el componente estacional.
- phi o damping slope: el coeficiente de la tendencia amortiguada.
- trend: el tipo de componente de tendencia ya sea como "add" para aditivo o "mul" para multiplicativo. El modelado de la tendencia se puede deshabilitar configurándolo en Ninguno.
- damped: si el componente de tendencia debe amortiguarse o no, ya sea Verdadero o Falso.
- seasonal: El tipo de componente estacional, ya sea como "add" para aditivo o "mul" para multiplicativo. El modelado del componente estacional se puede deshabilitar configurándolo en Ninguno.
- seasonal periods: el número de pasos de tiempo en un período estacional. Se itera entre 0, 4 y 13.
- boxcox: si realizar o no una transformación de potencia de la serie (Verdadero / Falso) o especificar la lambda para la transformación.

Resultados

Para los datos de tuberculosis en Colombia, específicamente, para el escenario sin COVID-19, primero se creó un modelo usando una regresión polinomial de grado 3, el cual tuvo un RMSE de 96.32, un MAPE de 6.723 %, un AIC de 402.82 y un BIC de 411.63. Aunque el MAPE obtenido es muy bueno, las regresiones lineales o polinomiales reproducen la tendencia de los datos, pero no el comportamiento variable de los mismos. Por lo tanto, este modelo solo se usó como línea base, más no como candidato a solución.

Luego, se crearon diversos modelos SARIMA, los cuales fueron entrenados y validados con una distribución de 80-20 % de los datos respectivamente y se seleccionó el que mejor comportamiento

ofrecía, el cual tiene un RMSE de 85.05, un MAPE de 5.88 %, un AIC de 252.26 y un BIC de 255.53. Este modelo tanto visualmente como basado en las métricas de validación, genera muy buenos resultados.

Posteriormente, se crearon diversos modelos Holt-Winters, también entrenados y validados con una distribución de 80-20 % de los datos, y se seleccionó el que menor MAPE tenía, el cual fue de 6.16%, además de tener un RMSE de 83.46, AIC de 344.48 y BIC de 374.88.

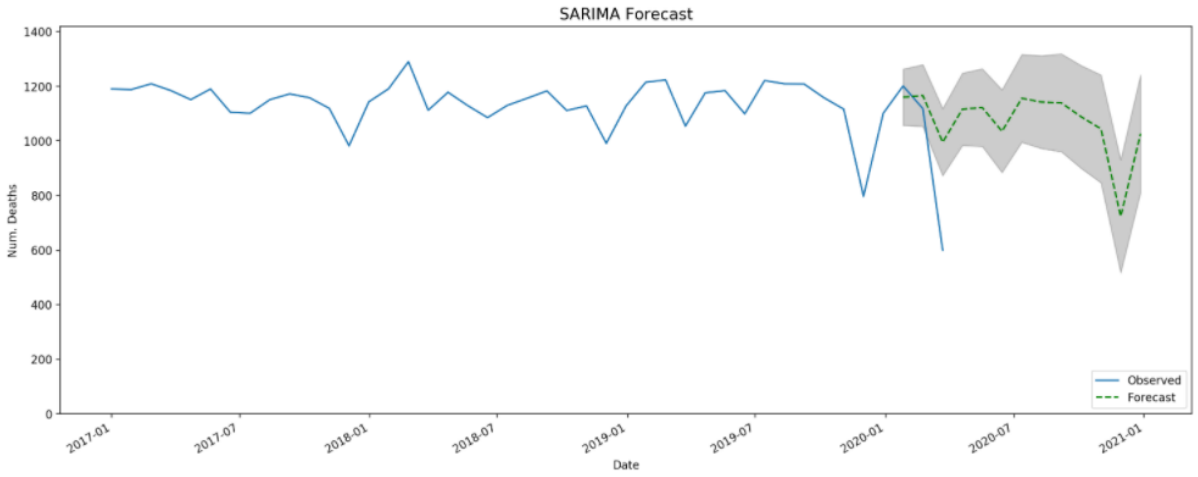

Figura 2. Pronostico parala tuberculosis en Colombia usando el método SARIMA.

Finalmente, se seleccionó el modelo generado con SARIMA como mejor solución al problema (figura 2), ya que ofrece el pronóstico con menor MAPE de todos y, visualmente, un comportamiento esperado, reproduciendo los picos inferiores observador en noviembre y diciembre de cada año. En la figura 3 se puede observar la comparación visual entre las métricas de los mejores modelos para cada método utilizado, en donde se confirma y justifica la selección del modelo proveniente por SARIMA.

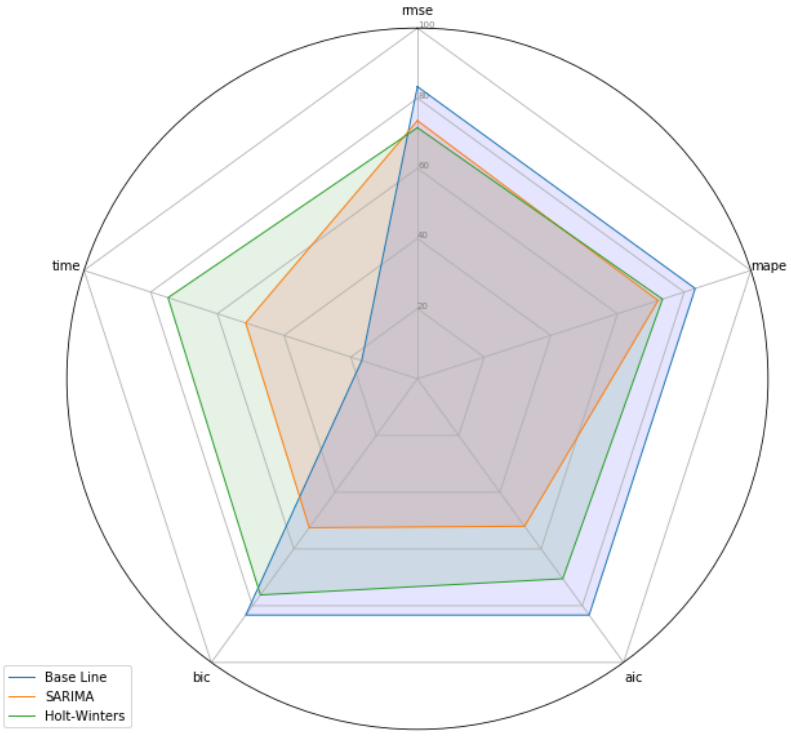

Figura 3. Comparación visual entre los mejores modelos por metodo.

Con respecto, a la señal de rastreo, se calculó para el pronóstico obtenido por el modelo de SARIMA en la fase de validación, sugiriendo el análisis confiabilidad en los 3 primeros periodos de pronóstico, más no en toda la serie, ya que el límite establecido era máximo 3 desviaciones estándares.

Tabla 1. Resultado de la señal de rastreo para la calibración de los datos para Colombia.

Periodo	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
		1115	**1131**	**58**	**5.99%**							

Modelos por departamentos y capitales

Para crear los modelos por departamentos y capitales, se realizó primero un análisis de similitud y de agrupación (clusterización), con el objetivo de no crear 64 nuevos modelos (32 para los departamentos y 32 para las capitales), sino, un máximo de 10 nuevos modelos, que pudieran reproducir los comportamientos de las 64 entidades mencionadas. La justificación es clara, entre menos modelos se tengan que crear, calibrar y mantener, la solución final será más mantenible y eficiente en el tiempo.

La agrupación se realiza usando el algoritmo k-means, con 2 enfoques, el primero basado en la distancia euclidiana de los datos agregados por periodos epidemiológicos (superponiendo los datos por años) y el segundo, basado en la distancia de deformación dinámica del tiempo (o *dynamic time warping)* entre los valores atómicos de las series de tiempo. Ambos análisis dieron resultados similares, no son concluyentes en cuanto a cuál enfoque es mejor, pero sí, en cuanto a valor agregado de generar subgrupos de entidades que se usarán para alimentar los modelos a crear. Se exploró justificar los grupos resultantes con las regiones geográficas de Colombia o con variables sociodemográficas de los departamentos (población, área, PIB, cantidad de nacimientos en el último año, porcentaje de nacimiento de mujeres), sin embargo, no se encontró una relación o causalidad entre las variables y los grupos resultantes.

Una vez se definen los elementos similares que conforman cada clúster, ya sean departamentos o capitales, se procede a crear 1 nuevo modelo que reproduzca a las series de tiempo contenidas en dicho grupo.

NOTA: ESTE PUNTO AUN ESTÁ EN DESARROLLO