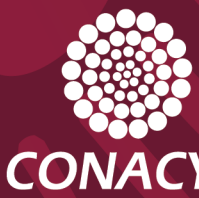
;


CONACYT


CIMAT


México
2021
Año de la
Independencia

Metodología para el análisis de tendencias de curvas epidémicas de casos nuevos de síndrome COVID-19 y de ocupación de camas y defunciones registrados en la Red IRAG

Coordinación de Servicios Tecnológicos

CONACYT CIMAT México 2021 Año de la Independencia

Reconocimientos

El contenido del presente documento fue elaborado por personal de la Coordinación de Servicios Tecnológicos del CIMAT.

- M.C. Judith Esquivel Vázquez
- Dr. Joaquín Peña Acevedo
- M.C. Domingo Iván Rodríguez González
- M.C. Ivete Sánchez Bravo
- M.T.B. Juan Luis Salazar Villanueva

Centro de Investigación en Matemáticas, A.C.

Julio 2021

Tabla de Contenidos

;

1. Introducción

1.1 Descripción general

Una serie de tiempo es una secuencia de datos obtenidos de observaciones recolectadas secuencialmente en el tiempo.

Los cambios sistemáticos en una serie de tiempo que no corresponden a un comportamiento periódico constituyen lo que se denomina como la *tendencia*, y corresponden a incrementos o decrementos prolongados en la serie.

El trabajo desarrollado consiste en el análisis de la tendencia que tienen las siguientes variables de observación relacionados con la epidemia de COVID-19 en México:

- Número de casos nuevos diarios de síndrome COVID-19
- Número de decesos nuevos diarios de la Red IRAG
- Ocupación diaria de camas IRAG

El análisis se hace para las series de tiempo de cada entidad federativa y zonas metropolitanas del país y los resultados son utilizados como uno de los insumos en el Semáforo de Riesgo epidemiológico [1] para transitar hacia una nueva normalidad, el cual es un sistema de monitoreo para la regulación del uso del espacio público de acuerdo con el riesgo de contagio de COVID-19 regulado por la Secretaría de Salud.

El propósito de este documento es detallar la metodología que se aplica para analizar la tendencia en las series de tiempo antes mencionadas, cuantificar la tendencia para dar un indicador y categorizar este indicador en cinco niveles que indican la manera en que los datos crecen o decrecen, con la finalidad de que esta información sea utilizada para elaborar un reporte semanal en el que se presenta la situación del país en fechas recientes a la publicación de los resultados.

A continuación se describen las variables de observación y algunas caraterísticas que tienen las series de tiempo de estas variables.

[1] Ver detalles en `https://coronavirus.gob.mx/semaforo/`

1.2 Variables de observación

- **Número de casos nuevos diarios de síndrome COVID-19**: Corresponde a la totalidad de posibles casos COVID-19 registrados en el Sistema de Vigilancia Epidemiológica de Enfermedades Respiratorias (SISVER) y que presentan signos y síntomas de COVID-19. Incluye los casos a los cuales se ha tomado la muestra independientemente del resultado y aquellos a los que no se tomó muestra. En los reportes se estudian los casos nuevos de Síndrome COVID-19 por día por cada entidad federativa o zona metropolitana según sea el caso. Estos corresponden a todos los registros de la base de datos sin y con muestra de laboratorio y sin importar el resultado del mismo.
- **Número de decesos nuevos diarios de la Red IRAG**: Corresponde a la cantidad de defunciones nuevas por día que se encuentran registradas en la Red IRAG (Infección Respiratoria Aguda Grave).
- **Ocupación diaria de camas IRAG**: Corresponde a la cantidad de personas que se encuentran hospitalizadas por día distribuidas por entidad federativa, esta cantidad es notificada y registrada en la base de datos de la red IRAG (Infección Respiratoria Aguda Grave) por los hospitales correspondientes. Se contabilizan los casos de hospitalización general y de hospitalizados con intubación.

1.2.1 Fuentes de información

- La base de datos del sistema nacional de vigilancia epidemiólogica de enfermedades respiratoria (SISVER), contiene los registros de casos que presentan signos y sintomas asociados a COVID-19 en México. [2] Los datos registrados son proporcionados por el gobierno de cada entidad federativa.
- La ocupación de camas asociadas a una enfermedad respiratoria y las defunciones, es proporcionada por la Secretaría de Salud a través del Sistema de Información de la Red IRAG (Infección Respiratoria Aguda Grave). Los datos registrados son proporcionados directamente por cada hospital por medio de su Clave Única de Establecimientos de Salud (CLUES).
- La población de cada nivel territorial, corresponde a la proyección al 2020 que emite el Consejo Nacional de Población (CONAPO). [3]
- La delimitación de las 74 zonas metropolitanas en México, corresponden a los resultados de la delimitación de zonas metropolitanas 2015, que derivan de la información de la Encuesta Intercensal 2015 por la Secretaría de Desarrollo Agrario, Territorial y Urbano (SEDATU), el Consejo Nacional de Población (CONAPO) y el Instituto Nacional de Estadística y Geografía (INEGI). [4]

Para hacer una comparación justa entre las diversas localidades, se analizan los datos escalados a casos nuevos por cada 100,000 habitantes y camas ocupadas por cada 100,000 habitantes.

[2] Base de datos disponible en: `https://www.gob.mx/salud/documentos/datos-abiertos-152127`

[3] Base de datos disponible en: `https://datos.gob.mx/busca/dataset/proyecciones-de-la-poblacion-de-mexico-y-de-las-entidades-federativas-2016-2050`

[4] Base de datos disponible en: `https://www.gob.mx/conapo/documentos/delimitacion-de-las-zonas-metropolitanas-de-mexico-2015`

1.3 Análisis del rezago entre fecha de inicio de síntomas y fecha de ingreso a unidad médica para la determinación del parámetro *D*

Es importante notar que existe un rezago entre la fecha en la que una persona presenta la sintomatología sospechosa de COVID-19 y la fecha en la que ingresa a una unidad médica para recibir atención, y por consiguiente ser registrada en la base de datos abiertos. Este rezago se debe principalmente al desarrollo natural de la enfermedad respiratoria que determina el tiempo que una persona espera para acudir a recibir atención médica, a partir de que inició a presentar síntomas.

Caracterizar el rezago es muy importante para la estimación de la tendencia del síndrome COVID-19, puesto que éste representa una demora en la disponibilidad de la información que ocasiona un decremento aparente en el número de casos en los días más recientes, que no corresponde con un decremento real de los contagios. La figura 1.1 muestra este efecto al comparar las series de tiempo de casos nuevos por día del 12 de octubre y del 26 de octubre, en donde puede notarse que aparentemente había un descenso en la primera serie de tiempo, que no corresponde con los datos observados 14 días más tarde.

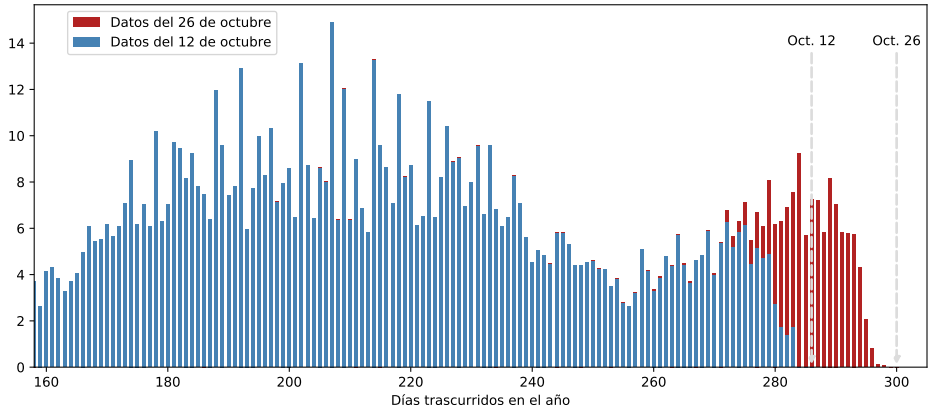


Fig. 1.1: Comparación de las series de tiempo de casos nuevos por día para las fechas del 12 y 26 de octubre, en donde se hace evidente el efecto del rezago en los datos.

Con base en el análisis del rezago es posible determinar el valor de un parámetro *D* que corresponde a los días que deben descartarse al final de la serie de tiempo de casos nuevos diarios, para que la metodología presentada en este reporte represente con mayor precisión la tendencia real del síndrome COVID-19.

Para determinar el tiempo de rezago se realizó una estimación de la función de densidad de los días que transcurren entre las fechas críticas de la enfermedad (inicio de síntomas y fecha de ingreso), por medio de funciones kérnel de tipo gaussiano [Weglarczyk 18]. Una vez determinada la función de densidad, es posible calcular los valores estadísticos de la mediana y los percentiles 10 y 90 del número de días que transcurren entre estas fechas, para establecer un valor adecuado del parámetro *D*.

Para la tendencia del Síndrome COVID-19, solo es relevante el retraso entre la fecha de síntomas y la fecha de ingreso, que representa el tiempo que tarda una persona en buscar atención médica, desde que presenta síntomas hasta que es registrada en la base de datos abiertos, sin importar si se le tomará una muestra para prueba de laboratorio y cuál será el resultado de la misma. Por otro lado, la ocupación hospitalaria reportada en la Red IRAG no presenta ningún rezago, por lo que se omite de este análisis.

1.3.1 Densidad del rezago a nivel nacional

Las siguientes gráficas muestran la densidad del rezago al día 13 de Diciembre de 2020, considerando los días que transcurren entre las fechas de inicio de síntomas y la fecha de ingreso. Las densidades fueron estimadas mediante funciones kérnel de tipo gaussiano con un ancho de banda de 0.6 días.

Las figuras 1.2 y 1.3 muestran las funciones de densidad y distribución estimadas a nivel nacional. Se observa que la mediana del rezago es de 3.72 días, por lo que se estima que en el transcurso de esta cantidad de días solo la mitad de la población que ha presentado síntomas habrá buscado atención médica. Mientras que el percentil 90 indica que al transcurrir 8.41 días el 10% de la población no se habría presentado aún a una unidad médica para recibir atención, y por consiguiente no estaría aún registrada en el SISVER.

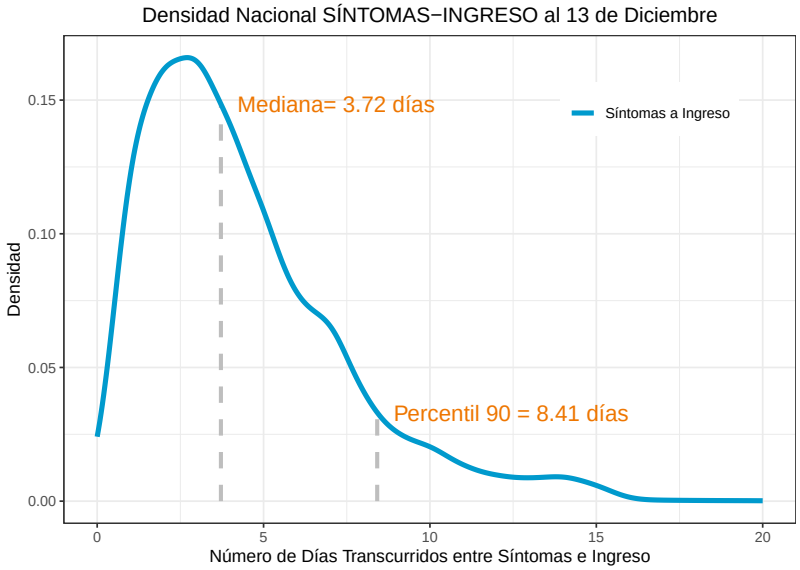


Fig. 1.2: Función de densidad del rezago a nivel nacional al 13 de diciembre de 2020.

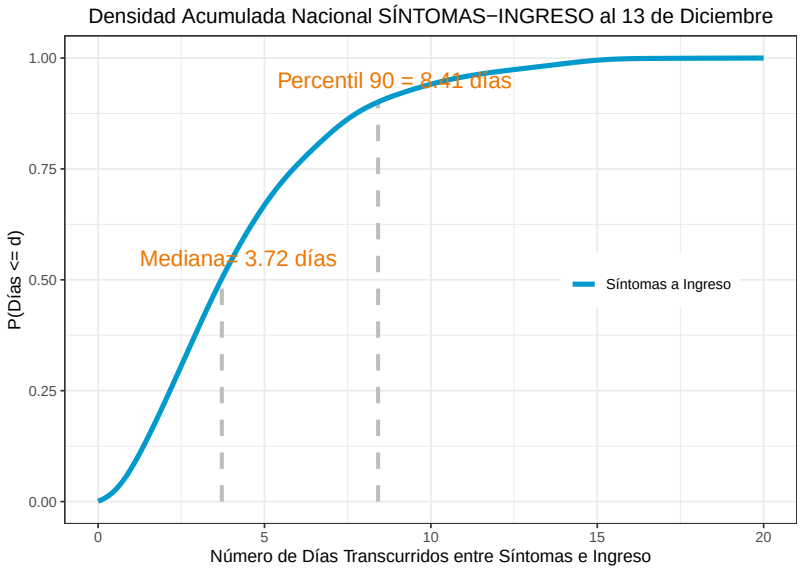


Fig. 1.3: Función de distribución del rezago a nivel nacional al 13 de diciembre de 2020.

Debe notarse que no se tomó un valor del parámetro $D = 8.41$ días, debido a que este es el caso nacional y se han observado diferencias considerables entre las funciones de densidad de los estados, como se muestra en la siguiente sección. Debe considerarse, además, que existen estados con un número significativamente mayor de casos que la mayoría y que tienen una contribución mayor a esta función de densidad. Por tal motivo, se realiza el análisis a nivel estado para determinar un valor del parámetro D de tal manera que el rezago no tenga un efecto significativo en la determinación de la tendencia para ningún estado.

1.3.2 Densidad del rezago por Estado

Se estimaron también las densidades del rezago para cada uno de los estados. Se consideraron únicamente los datos de las últimas 4 semanas previas al 13 de diciembre, y se utilizó un ancho de banda de 1.5 días.

La tabla 1.1 muestra la lista de estados con los valores de la mediana y el percentil 90 de su densidad del rezago entre fecha de síntomas e ingreso. Se tienen los siguientes valores estadísticos del percentil 90:

- Mediana del Percentil 90: 8.3 días
- Mínimo del Percentil 90: 5.7 días
- Máximo del Percentil 90: 10.9 días

Estado	Mediana	Percentil 90	Estado	Mediana	Percentil 90
AGUASCALIENTES	3.91	9.35	MORELOS	3.09	7.12
BAJA CALIFORNIA	3.95	8.92	NAYARIT	3.29	6.65
BAJA CALIFORNIA SUR	2.78	6.14	NUEVO LEÓN	3.87	8.45
CAMPECHE	5.13	9.28	OAXACA	3.37	6.42
CHIAPAS	3.44	7.12	PUEBLA	4.31	9.04
CHIHUAHUA	5.17	10.96	QUERÉTARO	4.11	8.26
CIUDAD DE MÉXICO	3.99	8.85	QUINTANA ROO	3.09	6.81
COAHUILA	4.54	9.67	SAN LUIS POTOSÍ	4.46	9.00
COLIMA	3.29	6.73	SINALOA	4.27	8.81
DURANGO	3.91	8.96	SONORA	3.64	7.28
GUANAJUATO	3.72	7.40	TABASCO	3.91	8.14
GUERRERO	4.27	7.79	TAMAULIPAS	4.42	8.26
HIDALGO	4.27	8.61	TLAXCALA	3.91	8.73
JALISCO	3.72	8.30	VERACRUZ	3.52	7.59
MÉXICO	4.23	9.16	YUCATÁN	2.70	5.71
MICHOACÁN	4.50	9.16	ZACATECAS	4.46	8.26

Tabla 1.1: Resumen de los valores de la mediana y el percentil 90 para los estados.

Las figuras 1.4 y 1.5 muestran las gráficas de las funciones de densidad para los 32 estados.

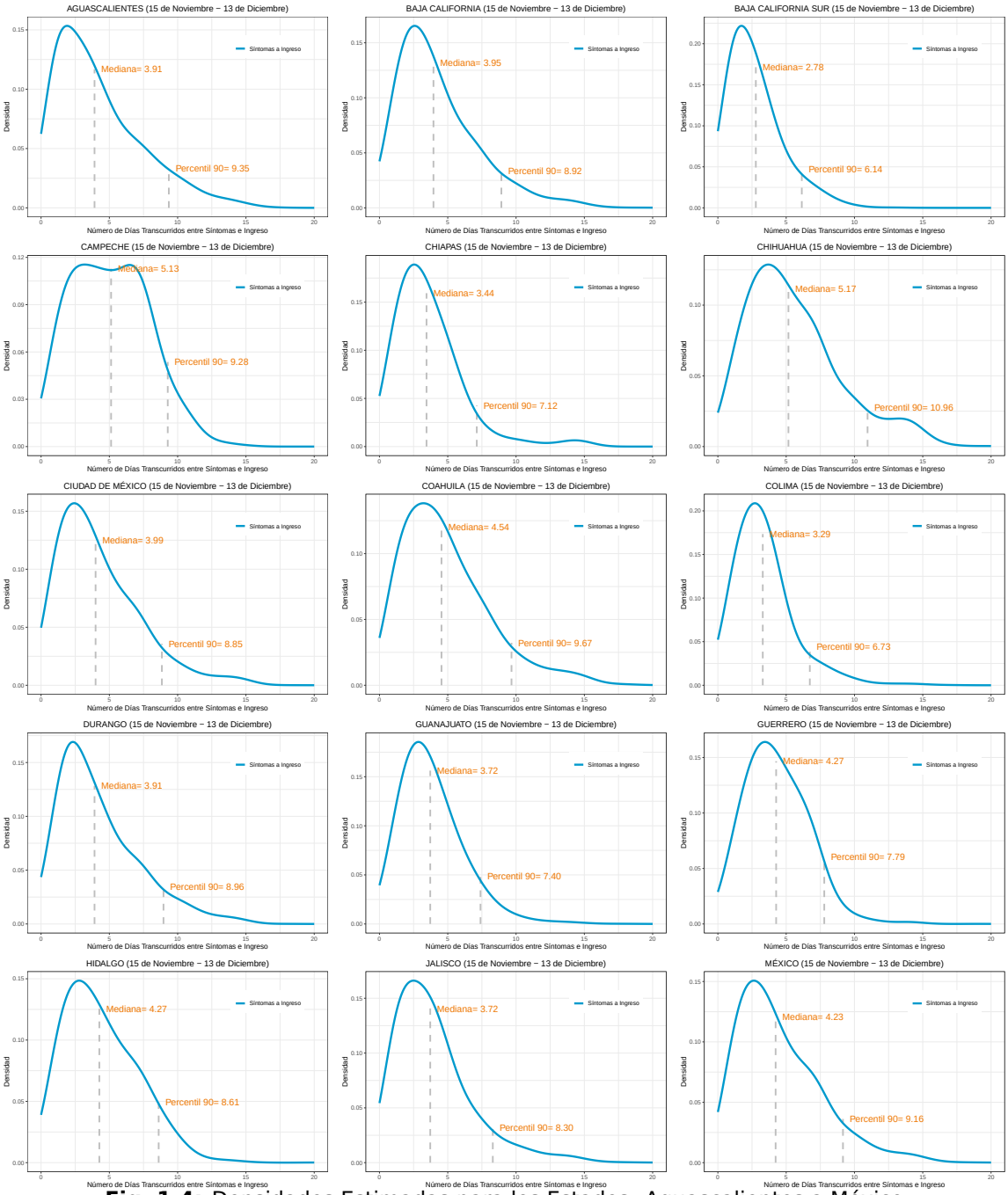


Fig. 1.4: Densidades Estimadas para los Estados: Aguascalientes a México

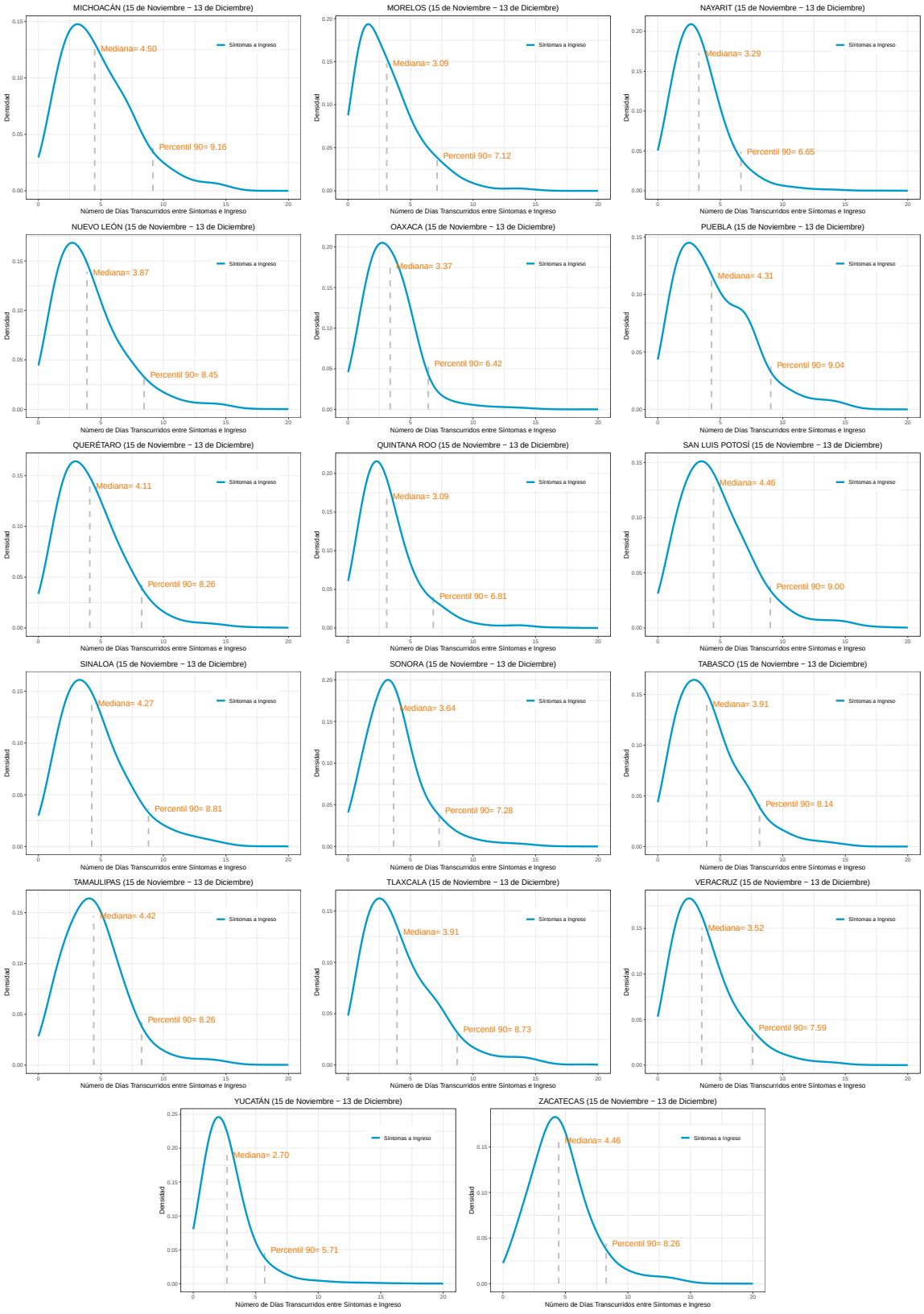


Fig. 1.5: Densidades Estimadas para los Estados: Michoacán a Zacatecas

1.3.3 Selección del Parámetro *D*

Para determinar el valor de *D* se tomó en cuenta el percentil 90 de las densidades de cada estado. A partir de la tabla 1.1 se observa que 31 de los 32 estados, tienen un percentil 90 inferior a 10 días. Por lo que se tomó un valor general del parámetro $D = 10$ para la determinación de la tendencia de casos nuevos de síndrome COVID-19. La figura 1.6 muestra la comparación de este valor con el percentil 90 de todos los estados.

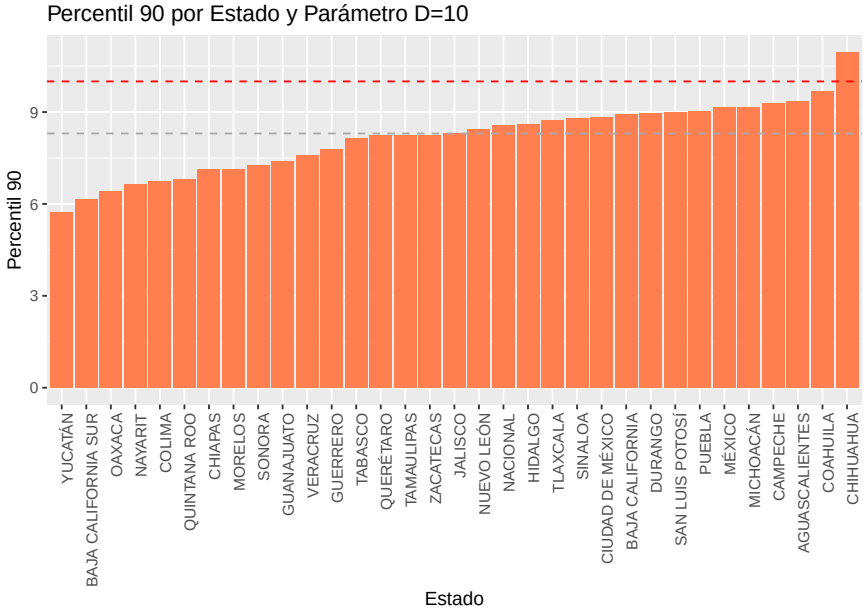


Fig. 1.6: Estados en orden ascendente por el valor del Percentil 90 de la densidad del rezago. Se muestra en línea punteada en color gris la mediana del percentil 90 de todos los estados, y en línea punteada en color naranja el valor de 10 días del parámetro *D*.

La selección de $D = 10$ significa que los últimos 10 días de la serie de casos nuevos de síndrome COVID-19 son descartados para el cálculo de la tendencia para todos los estados. La figura 1.7 muestra la comparación de las series de tiempo del 12 y 26 de octubre y la representación de los días descartados de acuerdo al parámetro *D*.

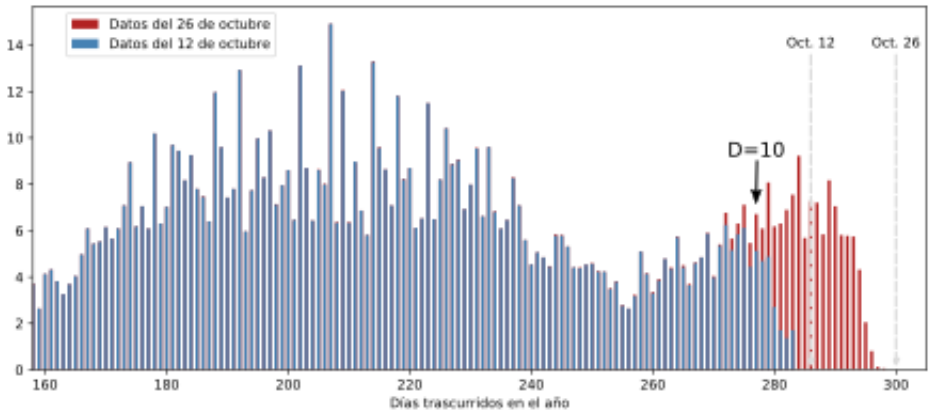


Fig. 1.7: Comparación de las series de tiempo de casos nuevos de síndrome COVID-19 para el 12 y 26 de octubre, y los 10 días descartados para la serie del 12 de octubre de acuerdo con el parámetro D.

2. Metodología para el análisis de la tendencia

2.1 Datos para el análisis

En esta sección se describen varios aspectos que se tomaron en cuenta para analizar las series de tiempo y generar la información que se presenta en el reporte que semanalmente se emite sobre la tendencia que mantienen los datos de cada entidad federativa y zonas metropolitanas.

Una serie de tiempo está formada por una secuencia $y_0, y_1, \ldots, y_n$ de valores que toma una variable en los instantes de tiempo $t_0 < t_1 < \ldots < t_n$. Se hace el supuesto de que hay una función desconocida $\mu(t)$ para la cual se cumple que

$$y_i = \mu(t_i) + \epsilon_i, \quad i = 0, 1, \ldots, n, \tag{2.1}$$

donde los ϵ_i son los errores aleatorios no correlacionados con media cero y tienen la misma varianza. El objetivo de los métodos de regresión es estimar $\mu(t)$ a partir de los datos y de ciertas suposiciones adicionales, pues $\mu(t)$ explica el comportamiento de los datos en el pasado y puede ser usada para hacer predicciones de corto plazo.

Las series de tiempo que se analizan están formadas por las observaciones $y_0, y_1, \ldots, y_n$ registradas en días consecutivos a partir de una fecha inicial f_0, de modo que $t_0 = 0, t_1 = 1, \ldots, t_n = n$ indican la cantidad de días transcurridos y $f_0 + t_n$ es la fecha más reciente de la que se tiene información.

Debido a la prioridad que tiene la publicación la información que se tiene disponible cada día, algunas de las series de tiempo tienen el problema de que los valores reportados en fechas recientes a la de la publicación pueden ir cambiando en fechas posteriores, conforme cada entidad federativa reporta su información, como se explicó en la Sección 1.3.

Por ejemplo, en la Figura 2.1 las barras azules indican la cantidad de casos nuevos de pacientes contagiados por cada 100,000 habitantes en una entidad federativa, de acuerdo con los datos publicados el 7 de diciembre de 2020. La disminución de casos al final de la serie de tiempo se debe al retraso de la información y al hacer la estimación de la función $\mu(t)$ (línea naranja) con estos datos, $\mu(t)$ refleja esta caída, por lo que la estimación de la tendencia será incorrecta dado que estos datos cambian en los siguientes días conforme se actualice la información.

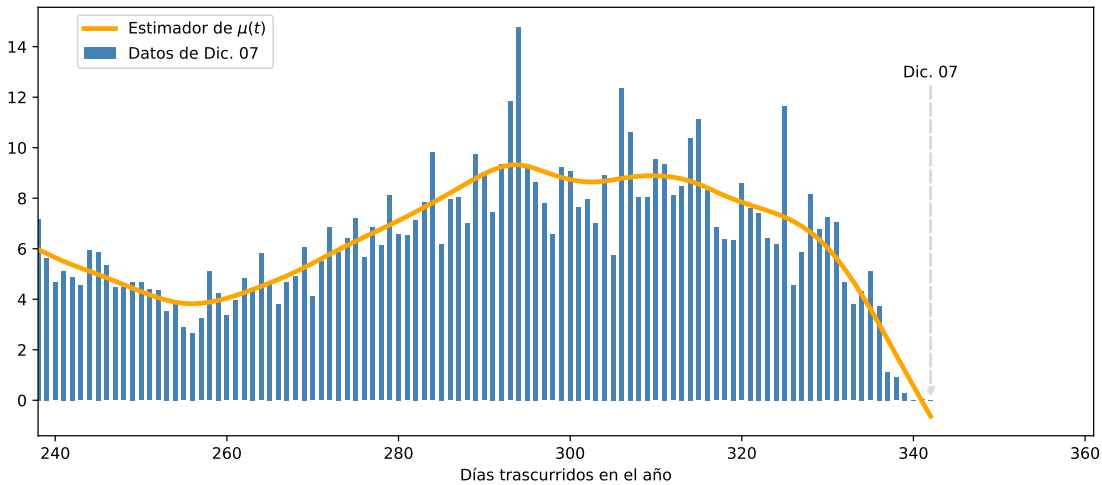


Fig. 2.1: Las barras azules corresponden a los datos publicados el 7 de diciembre de casos nuevos por cada 100,000 habitantes para una entidad federativa. La curva naranja corresponde a la estimación de $\mu(t)$ en (2.1).

Para reducir el efecto que tiene el rezago de la información, se introduce el parámetro

D que indica la cantidad de datos que se remueven al final de la serie de tiempo. En la Figura 2.2 se grafican con barras azules los datos del 7 de diciembre y con barrar rojas los datos publicados el 21 de diciembre de la misma entidad federativa, para constatar que la disminución de alturas de las barras azules es debida del retraso en la información. La zona amarilla indica el intervalo que corresponde a los últimos $D = 10$ días de los datos del 7 de diciembre y al estimar la función $\mu(t)$ sólo con datos que están antes de la zona amarilla, se obtiene la curva naranja que ya no muestra una pronunciada caída. La curva verde corresponde a la estimación de la función $\mu(t)$ usando los datos del 21 de diciembre y se ve que aún hay diferencias entre las dos estimaciones, pero es menor en comparación con la estimación mostrada en la Figura 2.1.

El propósito de este trabajo es proponer una metodología que permita cuantificar la tendencia de las series de tiempo y categorizarla, que sea robusta al efecto del rezago en la información, de modo que la categoría que describe la tendencia en una determinada fecha no cambie en la mayoría de casos cuando se actualicen los datos, y que si llega a cambiar, lo haga en una categoría consecutiva a la estimada inicialmente.

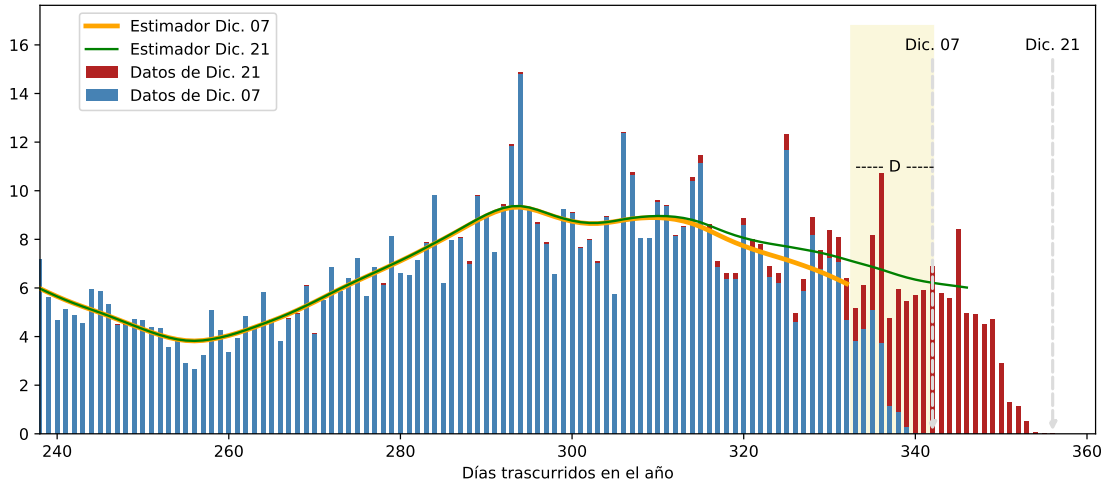


Fig. 2.2: Las barras azules corresponden a los datos publicados de casos nuevos por cada 100,000 habitantes para una entidad federativa. La línea naranja es la estimación de $\mu(t)$ con las barras azules que están fuera del rectángulo amarillo, mientras que la curva verde es la estimación de $\mu(t)$ usando las barras rojas.

2.2 Suavizamiento de datos y ajuste del modelo

Los modelos de regresión paramétrica pueden presentar varias restricciones para algunas aplicaciones. El enfoque clásico es aproximar $\mu(t)$ por un polinomio de bajo grado y calcular sus coeficientes resolviendo un problema de mínimos cuadrados lineales [Fahrmeir 13, capítulo 3]. Aunque es muy popular, la regresión polinomial tiene algunos inconvenientes como establecer un criterio para seleccionar los tiempos t_i que se usarán como nodos de interpolación y que sólo se pueden utilizar polinomios con grado no muy alto porque de lo contrario el polinomio puede tener variaciones no deseadas en intervalos entre los nodos. También puede ocurrir que los datos en ciertas posiciones pueden influir de manera significativa en el comportamiento del polinomio en partes lejanas de la curva. El uso de un modelo que es inapropiado para el problema bajo estudio puede conducir a conclusiones erróneas. Removiendo la restricción de que la función de regresión pertenezca a determinada familia paramétrica, permite el uso de técnicas de estimación no paramétrica de funciones [Wu 06], las cuales proporcionan diagnósticos importantes y herramientas de

inferencia para el análisis de datos [Eubank 94]. Dentro de estas técnicas se encuentra el uso de *smoothing splines* (Apéndice B), que tiene la desventaja de que no se puede calcular mediante una fórmula cerrada que sea fácilmente interpretable, pero por las características de este método, no hace falta aplicar alguna técnica de suavizamiento de datos para reducir los errores ϵ_i en (2.1) para obtener una aproximación de $\mu(t)$, y permite controlar el grado de ajuste al conjunto de datos y las variaciones que puede tener el estimador de $\mu(t)$, es decir, se puede controlar su complejidad. Esto se hace fijando un valor del parámetro de suavizamiento λ en (B.3).

Para elegir el valor del parámetro de suavizamiento λ se han realizado

- Análisis de los residuales $e_i = y_i - \mu_\lambda(t_i)$.
- Estimaciones con el método de validación cruzada generalizada.

Las técnicas antes mencionadas están pensadas para determinar el mejor valor para una sola serie. El problema principal en la elección del valor del parámetro λ es que se quiere usar un mismo valor para todas las series, para no hacer más complicado al método y tener que justicar un valor diferente para cada entidad federativa o zona metropolitana. Mediante las pruebas realizadas se validó que $\lambda = 199$ funciona de manera adecuada para todos los casos.

2.3 Categorización de la tendencia

El smoothing spline $\mu_\lambda(t)$ describe el comportamiento general de los datos en cada instante de tiempo t_i. Para cuantificar los cambios en la variable de observación se utiliza la derivada $\mu'_\lambda(t)$, pues su valor indica la rapidez con la cual los datos crecen o decrecen en determinado tiempo y puede usarse para predecir el comportamiento del spline a corto plazo. Por ejemplo, dado que la interpretación geométrica de la derivada es que ésta corresponde a la pendiente de la recta tangente a la curva en un punto, en la Figura 2.3 se muestran algunos segmentos magenta que son tangentes al spline, por lo que su pendiente coincide con la derivada del spline en dicho punto. Al cuantificar esta inclinación de los segmentos se puede establecer la tendencia de la serie en cada instante.

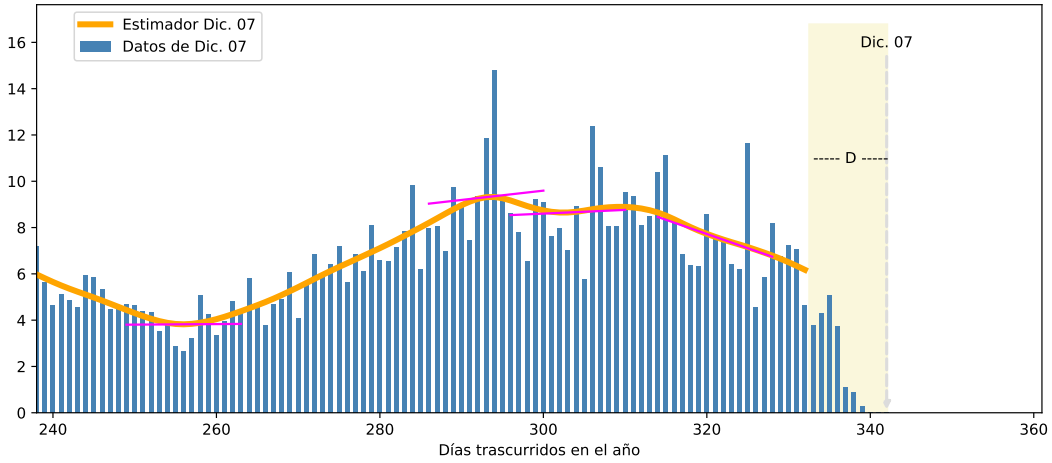


Fig. 2.3: Ilustración de la derivada del spline en algunos instantes de tiempo, representada por los segmentos magenta que corresponden a la rectas tangentes al spline.

Únicamente con el signo de la derivada del spline en un instante de tiempo se puede saber si un incremento o decremento de los valores, pero para especificar si estos cambios son moderados o no, hay que cuantificar la magnitud del valor de la derivada. Para simplificar su interpretación, se establece un intervalo de referencia y éste se divide en

subintervalos que corresponden a las siguientes categorías de la tendencia que indican que tan rápido crecen o decrecen los datos:

- *Ascendente*, que indica un crecimiento rápido de los valores.
- *Ascendente moderada*.
- *Estable*, que indica un cambio lento (incremento o decremento) de los valores por lo que permanecen en el mismo rango en un corto plazo.
- *Descendente moderada*.
- *Descendente*, que indica un decrecimiento rápido de los valores.

Como intervalo de referencia se elige a $[0, 1]$ y para mapear el valor de la derivada $m = \mu'_\lambda(t_i)$ a este intervalo se usa la función sigmoide, cuya gráfica se muestra en la Figura 2.4 y su expresión es

$$\text{Índice}_t(m) = \frac{1}{1 + \exp(-\alpha m + \beta)}, \tag{2.2}$$

donde los parámetros α y β se ajustan para controlar la forma en que crece esta función.

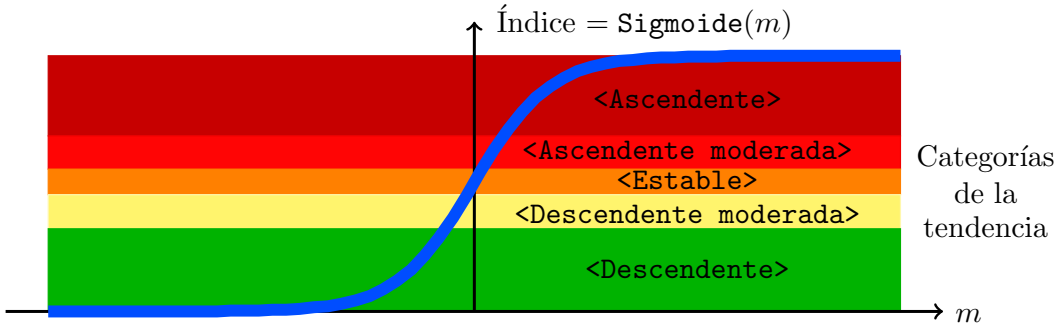


Fig. 2.4: La línea azul corresponde a la gráfica de la función sigmoide y los rectángulos de color indican cada rango de valores que corresponden a una de las categorías que se usan para indicar el tipo de tendencia, y dependiendo del valor de la derivada m se asigna la categoría correspondiente.

En la Figura 2.5 se colorean los puntos $\mu_\lambda(t_i)$ del spline ajustado en los datos, de acuerdo a la categoría de la tendencia que se le asigna a la derivada $\mu'_\lambda(t_i)$. Las líneas punteadas verticales indican la transición de una secuencia de datos de una categoría a otra para indicar los intervalos en la que los datos pertenecen a una misma categoría, y permite describir la evolución de los datos: después de un incremento moderado (primer intervalo), ocurrió un aumento en la rapidez en la ocurrencia del número de casos (segundo intervalo), posteriormente aunque se mantuvo la tendencia al alza se redujo la rapidez de contagios (tercer intervalo) hasta que se estabilizó la cantidad de número de casos diarios (cuarto intervalo), hasta que inició un descenso moderado (quinto intervalo).

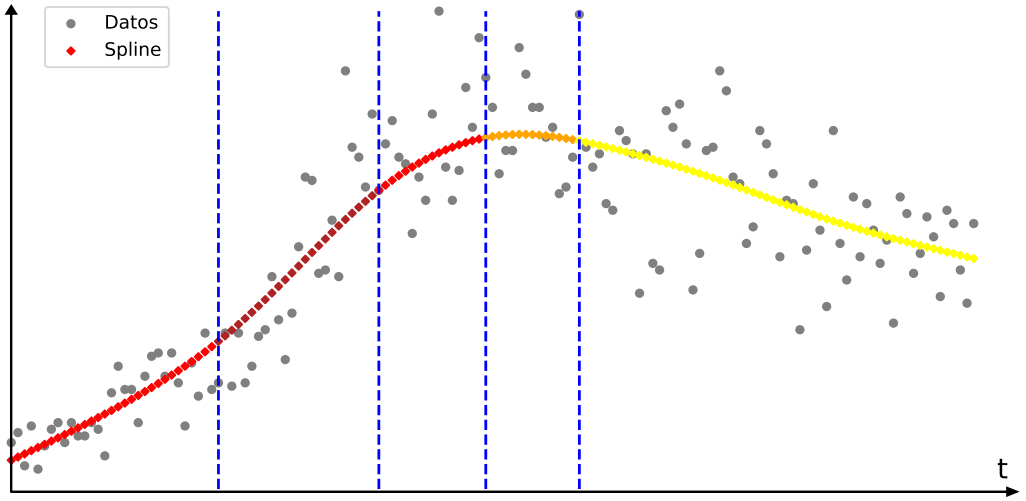


Fig. 2.5: Los colores asignados a los puntos sobre el spline son acordes a la categoría de la tendencia que se le asigna a la derivada, según se indica en la Figura 2.4.

2.4 Tendencia reportada

El propósito del reporte semanal es indicar para cada entidad federativa y zona metropolitana cual es la tendencia de cada una de variables estudiadas (número de casos, mortalidad y camas ocupadas) usando los resultados de las fechas más próximas a la fecha en la que se emite el reporte.

Debido al retraso en la información en algunas de las variables, como ya se explicó previamente, no es conveniente aplicar el análisis en los últimos D días de la serie de tiempo. De los días en los que sí se puede realizar el análisis, es complicado fijar un día particular para que la tendencia en ese día sea la que aparezca como conclusión en el reporte para todas las entidades federativas, porque el retraso en la información va cambiando en el tiempo, a veces se incrementa en algunas entidades o se observa que los registros en los fines de semana es menor que cuando inicia la semana, etc.

La determinación que se tomó fue la siguiente:

1. Calcular la derivada del spline en los los últimos siete días en los que sí se puede realizar el análisis de la tendencia:

$$\mu'_\lambda(t_{n_f-6}), \mu'_\lambda(t_{n_f-5}), \ldots, \mu'_\lambda(t_{n_f}).$$

2. Calcular la mediana m de estos valores.
3. Calcular $\text{Índice}_t(m)$ de acuerdo con la expresión (2.2).
4. Reportar como resultado el valor $\text{Índice}_t(m)$ y la categoría de la tendencia que corresponde a este valor.

Como la mediana es un estimador robusto, debe ser un buen representante de la tendencia en ese intervalo.

Para ilustrar esto, en la Figura 2.6 el rectángulo gris indica el intervalo en el que se removieron los últimos D datos de la serie. El rectángulo azul indica el intervalo que contiene los últimos 7 datos en donde se puede realizar el análisis de la tendencia. Para comparar visualmente las derivadas, en la parte derecha de la figura se muestran unos segmentos que tienen como pendiente los valores de las derivadas $\mu'_\lambda(t_i)$ en los puntos seleccionados. El segmento azul es el que tiene como pendiente a la mediana, así que ese es valor que se toma para calcular el índice (2.2) y la categoría del tipo de tendencia y eso es lo que reporta como la información más reciente.

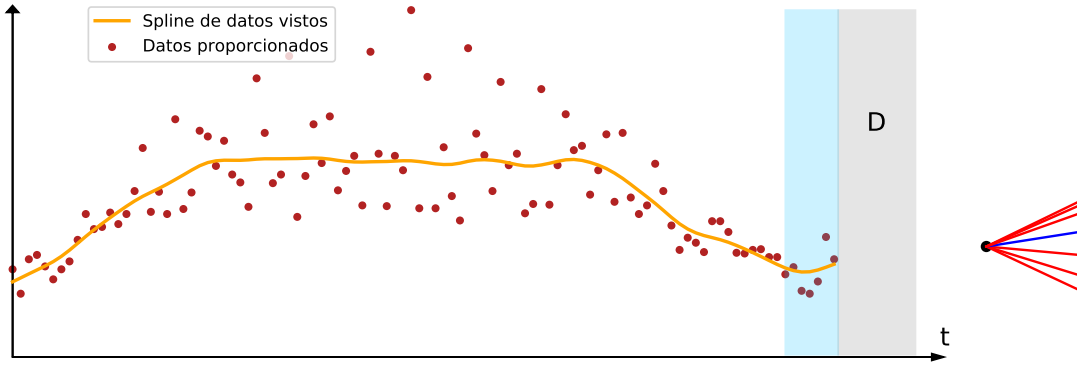


Fig. 2.6: A la izquierda se señala el intervalo de análisis (en color azul) y a la derecha se muestran unos segmentos que tienen como pendientes los valores de las derivadas de los puntos en el intervalo de análisis.

2.5 Resumen de la metodología

De acuerdo con lo explicado en las secciones anteriores, la metodología aplicada se resume en los siguientes pasos.

1. Eliminar los últimos D días de la serie de tiempo.
2. Aplicar la técnica de smoothing splines para ajustar un spline a los datos restantes.
3. Calcular las derivadas $\mu'_\lambda(t_i)$ del spline en cada instante de tiempo.
4. Evaluar el índice $\text{Índice}_t(m)$ en las derivadas en cada instante de tiempo t_i, y asignar la categoría de la tendencia.
5. Para el reporte semanal, la tendencia que se reporta para cada entidad federativa y zona metropolitana corresponde a la mediana de las derivadas en los últimos siete días en donde se pudo realizar el análisis, como se explica en la Sección 2.4.

3. Reporte semanal sobre las tendencias de curvas epidémicas

3.1 Reporte Semanal

Cada semana se emite un reporte al público y a la Secretaría de Salud, con la finalidad de proporcionar información de la tendencia más actualizada de la pandemia por COVID-19 por entidad federativa y por cada zona metropolitana. Dichos indicadores actualmente son tres de los diez indicadores de riesgo que definen el Semáforo de Riesgo epidemiológico [1] para transitar hacia una nueva normalidad, el cual es un sistema de monitoreo para la regulación del uso del espacio público de acuerdo con el riesgo de contagio de COVID-19 regulado por la Secretaría de Salud.

Como se ha explicado en el presente documento, debido a que los casos no son registrados en tiempo real, es decir, no se registran en la plataforma exactamente el día en el cual las personas presentan signos y sintomas, sino hasta que el paciente acude a una unidad de salud. Existe por lo tanto un rezago en el conocimiento de la información, el cual imposibilita conocer en tiempo real los casos de la pandemia. En los gráficos del reporte semanal, se presenta la información para casos nuevos diarios y decesos nuevos diarios de síndrome COVID-19 eliminando los $D = 10$ últimos días. Para ocupación hospitalaria y defunciones no se presenta rezago por lo que $D = 0$

El reporte es procesado con la información con corte los días lunes a las 9 hrs del Sistema de Vigilancia Epidemiológica de Enfermedades Respiratorias (SISVER) y con corte a las 00 horas del mismo día del sistema de Información de la Red IRAG (Infección Respiratoria Aguda Grave).

El reporte contiene información seccionada por cada entidad federativa y las zonas metropolitanas asociadas a dicha entidad, de cada región se reporta lo siguiente:

- Gráfico de tendencias de casos nuevos diarios de síndrome COVID-19.
- Gráfico de tendencias de decesos nuevos diarios de de la Red IRAG.
- Gráfico de tendencias de ocupación diaria de camas de la Red IRAG.
- Tabla con indicadores de riesgos de la localidad.

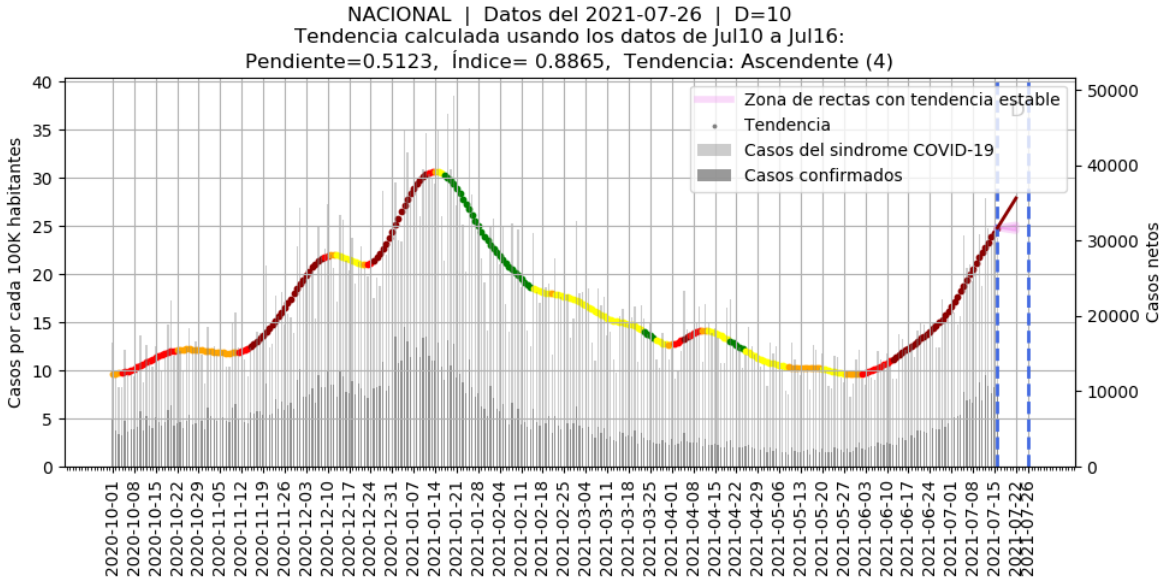


Fig. 3.1: Gráfico de tendencias de **casos nuevos diarios de síndrome COVID-19** con datos de la plataforma SISVER, 26-07-2021

[1] Ver detalles en `https://coronavirus.gob.mx/semaforo/`

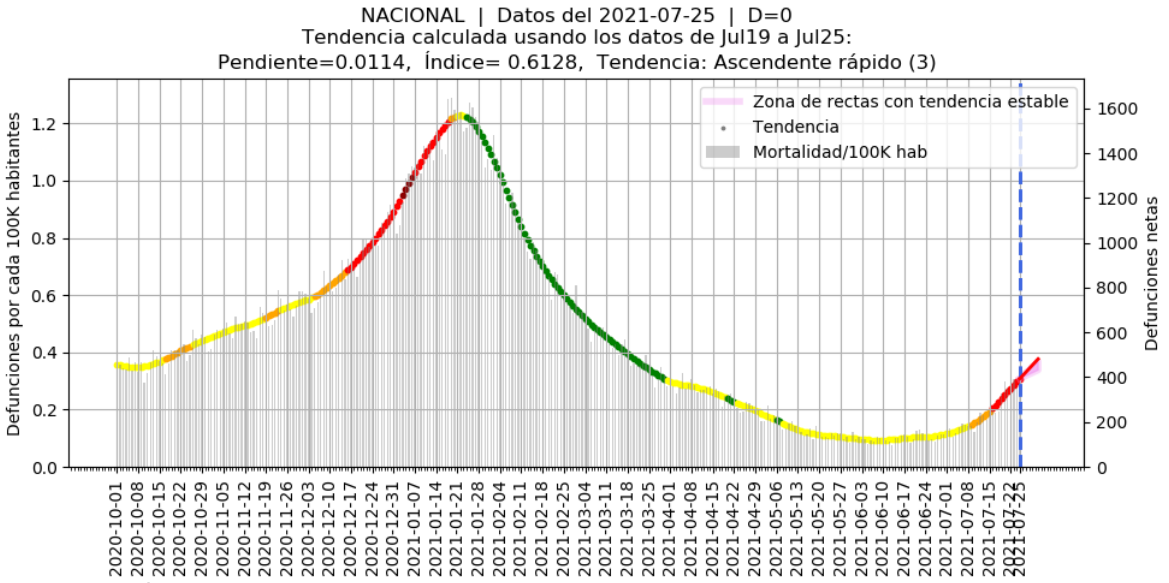


Fig. 3.2: Gráfico de tendencias de **decesos nuevos diarios**s con datos del sistema de la Red IRAG, 26-07-2021

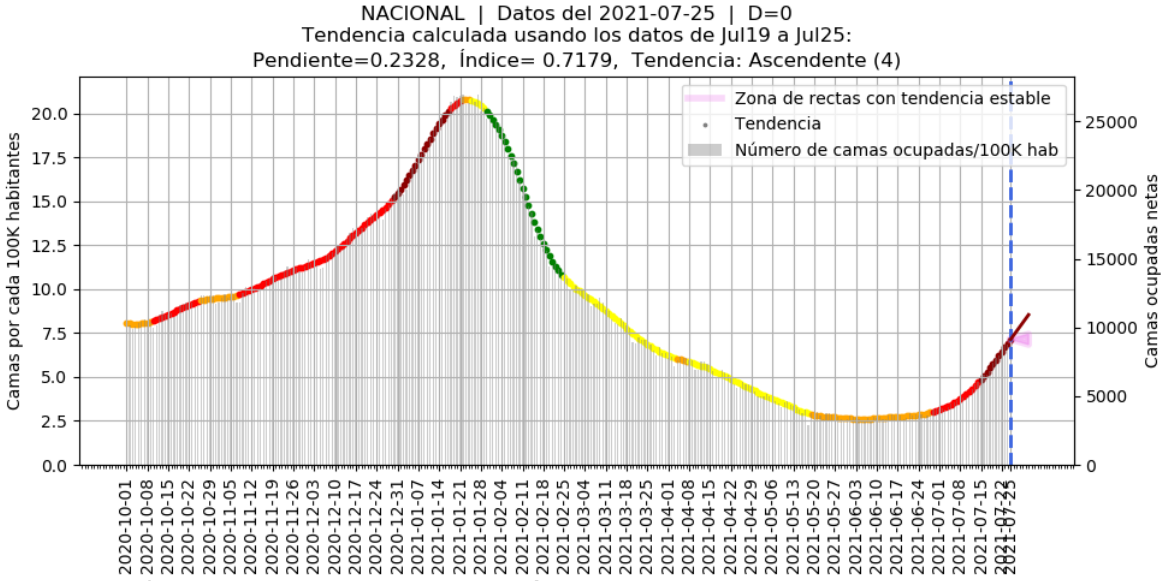


Fig. 3.3: Gráfico de tendencias de **ocupación hospitalaria** con datos del sistema de la Red IRAG, 26-07-2021

RESUMEN NACIONAL	
Tendencia de casos nuevos de síndrome COVID-19	Ascendente
Tendencia de decesos nuevos de la Red IRAG	Ascendente rápido
Tendencia Camas Ocupadas IRAG	Ascendente
Casos activos estimados	101348.0
Tasa de incidencia de casos estimados	79.3
Defunciones estimadas 14 días	2360.0
Tasa mortalidad estimada (14 dias)	1.8
% Positividad de casos COVID-19 semana 28	34.46 %
% Positividad de decesos COVID-19 semana 28	69.89 %
Retraso de síntomas a registro	$[q_{10} = 1, q_{50} = 4, q_{90} = 9]$

Tabla 3.1: Tabla de **indicadores de riesgo** de México con datos de la plataforma SISVER e IRAG, 26-07-2021

3.1.1 Interpretación de los gráficos

- Se grafica información a partir del 1 de Octubre de 2020 para efectos de visualización.
- La fecha en el gráfico, corresponde a la fecha de consulta de la base de datos de la plataforma SISVER o red IRAG según corresponda.
- Las barras gris oscuro, corresponden a los casos o decesos nuevos diarios confirmados distribuidos por fecha de signos y síntomas, según corresponda.
- Las barras gris claro, corresponde a la totalidad de posibles casos de COVID-19 registrados en la plataforma SISVER independiente del resultado obtenido, incluye los casos a los que no se tomó una muestra para la prueba PCR, y se distribuyen en el tiempo por fecha de signos y síntoma.
- La zona delimitada por las líneas punteadas azules indican el valor del parámetro D, es decir, corresponde a los días que se eliminan para el análisis, no se muestran datos ya que no son tomados en cuenta en el procesamiento.
- Del lado izquierdo, se visualiza la escala en casos por cada 100,000 habitantes. Del lado derecho se visualiza los casos netos, con la finalidad de dimensionar los casos en la zona territorial de análisis.
- En los gráficos de ocupación hospitalaria de camas IRAG y defunciones, no es necesario eliminar días de análisis y solo se presentan las barras con la totalidad de camas ocupadas y defunciones en cada región de análisis.
- El valor de la **pendiente**, el **índice**, la categoría de la **tendencia** y su color reportado en cada gráfico, corresponden al resultado de tomar la mediana del análisis de la tendencia en los últimos siete días de la serie de tiempo.

3.1.2 Interpretación de la tabla

- Además de los gráficos, se presenta una tabla con un resumen que contiene:
  - *Tendencia de casos nuevos de síndrome COVID-19*
  - *Tendencia de decesos nuevos de la Red IRAG*
  - *Tendencia de camas ocupadas IRAG*
  - *Casos activos estimados*
  - *Casos activos estimados por cada 100,000 habitantes, Tasa de incidencia de casos*
  - *Decesos estimados de los últimos 14 días*
  - *Decesos de los últimos 14 días estimados por cada 100,000 habitantes, Tasa de mortalidad estimada*
  - *Porcentaje de positividad de casos COVID-19*
  - *Porcentaje de positividad de decesos COVID-19*
  - *Retraso de síntomas a registro*

Definición de los indicadores de la tabla:

- **Tendencia de casos nuevos de síndrome COVID-19**, corresponde a la tendencia de los nuevos casos diarios de posibles contagios COVID-19 registrados en el Sistema de Vigilancia Epidemiológica de Enfermedadades Respiratorias (SISVER) distribuidos por fecha de inicio de signos y síntomas COVID-19.
- **Tendencia de decesos nuevos de la Red IRAG**, corresponde a la cantidad de defunciones nuevas por día registradas en la red IRAG (Infección Respiratoria Aguda Grave) en la región de análisis.

POR CASOS POR DIA POR CADA 100K HAB (m)		
Tendencia	Pendiente (m)	Valor
Ascendente Acelerado	≥ 0.020	4
Ascendente Rápido	$[0.010, 0.020)$	3
Ascendente Moderado	$[0.005, 0.010)$	2
Estable	$[-0.005, 0.005)$	1
Descendente	< -0.005	0

Tabla 3.2: Categorización de la pendiente de la recta tangente a la curva en un punto determinado. Parámetros para mortalidad reportada en la Red IRAG

- **Tendencia de camas ocupadas IRAG**, corresponde a la tendencia de ocupación de camas de la red IRAG (Infección Respiratoria Aguda Grave) en la región de análisis.

POR CASOS POR DIA POR CADA 100K HAB (M)		
Tendencia	Pendiente (m)	Valor
Ascendente	≥ 0.200	4
Ascendente moderada	$[0.05, 0.200)$	3
Estable	$[-0.05, 0.05)$	2
Descendente moderada	$[-0.200, -0.05)$	1
Descendente	< -0.200	0

Tabla 3.3: Categorización de la pendiente de la recta tangente a la curva en un punto (día) determinado. Parámetros para ocupación hospitalaria y casos nuevos asociados a síndrome COVID-19

- **Casos activos estimados**, corresponde a los nuevos casos cuya fecha de inicio de signos y síntomas está reportada en los últimos 14 días, para mitigar efectos del desconocimiento de registro, se suman los casos sospechosos multiplicados por el índice de positividad de la semana epidemiológica estudiada.
- **Casos activos estimados por cada 100,000 habitantes**, también conocida como **Tasa de incidencia de casos**, corresponde a los casos activos estimados escalados a casos por cada 100,000 habitantes. El valor de la *Incidencia promedio diaria estimada* se representa además por un color definido en los intervalos siguientes:

Valor	Rango	Color
4	$(46, \infty)$	
3	$(36, 46]$	
2	$(30, 36]$	
1	$(14, 30]$	
0	$[0, 14]$	

Tabla 3.4: Rango de valores de la Tasa de incidencia de casos para cada valor

- **Decesos estimados de los últimos 14 días**, corresponde a la cantidad de decesos cuya fecha de defunción está reportada en los últimos 14 días, para mitigar efectos del desconocimiento de registro, se suman los casos sospechosos multiplicados por el índice de positividad de mortalidad de la semana epidemiológica estudiada. Origen: SISVER
- **Decesos de los últimos 14 días estimados por cada 100,000 habitantes**, también conocida como **Tasa de mortalidad estimada**, corresponde a los Decesos estimados de los últimos 14 días escalados a decesos por cada 100,000 habitantes. El valor de la *Tasa de mortalidad estimada* se representa además por un color definido en los intervalos siguientes:

Valor	Rango	Color
4	$[8.5, \infty)$	
3	$[7.0, 8.5)$	
2	$[4.0, 7.0)$	
1	$[2.5, 4.0)$	
0	$[0, 2.5)$	

Tabla 3.5: Rango de valores de la Tasa de mortalidad estimada para cada valor

- **Porcentaje de positividad de casos**, El porcentaje de positividad de casos confirmados COVID-19 se calcula sobre la segunda semana epidemiológica anterior. Es decir, se elimina la última semana epidemiológica de la fecha de análisis. De la semana anterior, se contabilizan los casos por fecha de síntomas cuyo resultado sea positivo y se divide entre el total de casos positivos y negativos.
- **Porcentaje de positividad de mortalidad**, El porcentaje de positividad de decesos se calcula sobre la segunda semana epidemiológica anterior. Es decir, se elimina la última semana epidemiológica de la fecha de análisis. De la semana anterior, se contabilizan los casos por fecha de defunción, cuyo resultado sea positivo y se divide entre el total de defunciones con confirmación COVID-19 positivos y negativos.
- **Retraso de síntomas a registro**, Se presentan los percentiles 10, 50 y 90 de la distribución de día que pasan de que una persona presenta síntomas de COVID-19 hasta que se realiza el registro en la plataforma SISVER. Se consideraron los datos de las últimas cuatro semanas, del 22 de noviembre al 13 de diciembre de 2020, por fecha de resultados. Se utilizó el archivo 201213COVID19MEXICO.csv correspondiente al 13 de diciembre de 2020.

Este formato de reporte puede ser consultado a partir del 22 de de marzo de 2021 de forma semanal en la página oficial de coronavirus del CONACYT [2]

[2] Proyecto oficial `https://coronavirus.conacyt.mx/proyectos/tendencia.html`

Apéndices

A. Splines polinomiales

Una manera de crear un interpolador para un conjunto de datos es mediante la construcción de un *spline* [1], que es una curva como la que se muestra en la Figura A.1. Para el problema de interpolación de un conjunto de puntos

$$\{(t_0, y_0), (t_1, y_1), \ldots, (t_n, y_n)\},$$

en los que $t_0 < t_1 < \ldots < t_n$, se considera que en cada subintervalo $[t_j, t_{j+1}]$ el spline está definido por un polinomio $S_j(t)$ de bajo orden

$$S(t) = \begin{cases} S_0(t) & t \in [t_0, t_1], \\ S_1(t) & t \in [t_1, t_2], \\ \vdots & \\ S_{n-1}(t) & t \in [t_{n-1}, t_n]. \end{cases}$$

El spline interpola los datos, por lo que $S(t_j) = S_j(t_j) = y_j$ para $j = 0, 1, \ldots, n$. Además se impone alguna condición de suavidad para "unir" esos pedazos en intervalos consecutivos. Por lo menos debe ser una curva continua, así que

$$S_{j-1}(t_j) = S_j(t_j) \qquad j = 1, 2, \ldots, n-1.$$

Cuando se quiere usar una curva suave, lo usual es utilizar un spline cúbico, el cual es un spline en el que cada polinomio $S_j(t) = a_3 t^3 + a_2 t^3 + a_1 t + a_0$ es un polinomio cúbico, y al que se le pide que forme una curva continua y que su primera y segunda derivada también sean continuas en los puntos t_j. Así, el spline cúbico que interpola los puntos $\{(t_0, y_0), \ldots, (t_n, y_n)\}$ cumple con

- $S(t) = S_j(t)$ si $t \in [t_j, t_{j+1}]$, donde $S_j(t)$ es un polinomio cúbico.

[1] Este término se usa por la semejanza con aquellas piezas de metal o de madera que se podían flexionar para trazar curvas en, por ejemplo, planos que indican el diseño de objetos o estructuras.

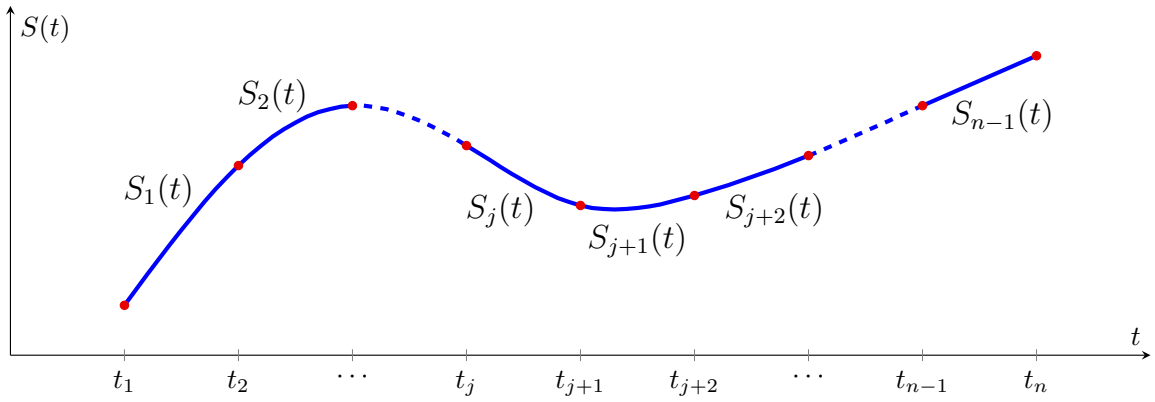


Fig. A.1: Un spline polinomial es una función en la que en cada subintervalo $[t_{i-1}, t_i]$ se define por un polinomio de modo que las gráficas de estos polinomios se "unan" en los nodos t_i para formar una curva continua y que además se le puede exigir que la curva tenga cierta suavidad en los puntos t_i.

- Para $j = 1, \ldots, n-1$, para los nodos en el interior se tiene que cumplir que

$$\begin{array}{rcccl} S(t_j) & = & S_{j-1}(t_j) & = & S_j(t_j) \quad = \quad y_j \\ S'(t_{j+1}) & = & S'_{j-1}(t_j) & = & S'_j(t_j) \\ S''(t_{j+1}) & = & S_{j-1}{}''(t_j) & = & S_j{}''(t_j) \end{array} \tag{A.1}$$

Se puede ver que se necesitan dos condiciones adicionales a las anteriores para poder determinar los coeficientes de los polinomios $S_j(t)$ [Quarteroni 00]. Es usual fijar a 0 los valores de la segunda derivada del spline en los extremos del intervalo, es decir, a la lista anterior se agregan las condiciones

$$S''(t_0) = S''(t_n) = 0.$$

Cuando se usan las condiciones anteriores para determinar los coeficientes de todos los polinomios que definen al spline en el intervalo $[t_0, t_n]$, se dice que se construye un *spline natural cúbico*. Este tipo de splines son los se van a utilizar en la metodología.

En la Figura A.2 se muestra un ejemplo de un spline natural cúbico que interpola 21 puntos. La curva es suave y en el interior de cada subintervalo la curva no tiene fluctuaciones innecesarias. Por estas características se usan frecuentemente los splines en diseño asistido por computadora.

Cuando se modifica el valor y_j en algún nodo, el cambio del spline es local, a diferencia de cuando se usa un solo polinomio para interpolar todos los datos.

Por otra parte, cuando los valores $y_0, y_1, \ldots, y_n$ varían demasiado debido a que hay ruido en las mediciones, no se quiere construir un interpolador para no incorporar al ruido en el modelo que describe a las observaciones. Cuando esto ocurre, se quiere construir una función cuya gráfica esté cerca de los datos aunque no pase por todos los puntos. Para este tipo problemas también se puede usar un spline pero hay que cambiar la formulación de su construcción como se explica en el Apéndice B.

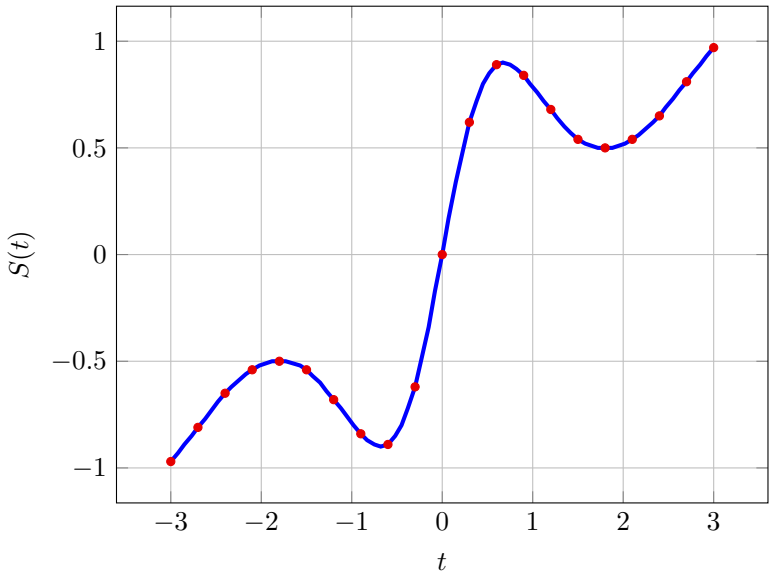


Fig. A.2: Spline cúbico natural que interpola los 21 puntos rojos.

B. Smoothing splines

Tenemos un conjunto de observaciones $y_0, y_1, \ldots, y_n$ que ocurren en los instantes de tiempo $t_0 < t_1 < \ldots < t_n$. Se hace la suposición de que las observaciones provienen del modelo

$$y_i = \mu(t_i) + \epsilon_i, \quad i = 0, 1, \ldots, n,$$

donde los ϵ_i son los errores aleatorios no correlacionados con media cero y tienen la misma varianza y $\mu(t)$ es la función de regresión desconocida [Eubank 94].

Dada una función $f(t)$, una medida típica del ajuste que tiene f a los datos es el promedio de la suma de residuales al cuadrado,

$$E_1(f) = \frac{1}{n+1} \sum_{i=0}^{n} [y_i - f(t_i)]^2 . \tag{B.1}$$

Si suponemos que la función $\mu(t)$ no tiene grandes fluctuaciones en intervalos cortos, tendríamos que todas las funciones f candidatas para estimar a μ deberían tener esta propiedad. Una medida natural de las variaciones que tiene una función f que es m veces diferenciable es

$$E_2(f) = \int_{t_0}^{t_n} \left[f^{(m)}(t) \right]^2 \, dt. \tag{B.2}$$

Intuitivamente para una función que tiene pocas fluctuaciones su m-ésima derivada toma valores relativamente menores a los de una función que sí tiene variaciones significativas, y por eso es que se espera que $E_2(f)$ pueda indicar esta característica, por lo que se dice que $E_2(f)$ mide la "rugosidad" de f ya que entre más plana es la gráfica de f, es menor es el valor de E_2 y $E_2(f) = 0$ si f tiene pendiente constante.

Así, una manera global de evaluar la calidad de un estimador candidato f está dada por la suma convexa de las expresiones (B.1) y (B.2):

$$E(f) = sE_1(f) + (1-s)E_2(f) = \frac{s}{n+1}\sum_{i=1}^{n}[y_i - f(t_i)]^2 + (1-s)\int_{t_0}^{t_n}\left[f^{(m)}(t)\right]^2 dt,$$

para algún $0 < s < 1$. Un estimador óptimo puede ser obtenido al minimizar este funcional. Fijando $\lambda = (1-s)/s$ y si aceptamos que $m = 2$ es suficiente para medir las fluctuaciones una función, entonces podemos estimar $\mu(t)$ mediante la función $\mu_\lambda(t)$ que resulta de resolver el problema

$$\min_{f \in C^2[t_0,t_n]} E_\lambda(t) = E_1(f) + \lambda E_2(f) = \frac{1}{n+1}\sum_{i=1}^{n}[y_i - f(t_i)]^2 + \lambda\int_{t_0}^{t_n}\left[f^{(2)}(t)\right]^2 dt, \tag{B.3}$$

sobre el espacio $C^2[t_0, t_n]$ de funciones con segunda derivada continua en el intervalo $[t_0, t_n]$. Cuando restringimos que la función sea un spline cúbico natural (Apéndice A), el estimador de $\mu(t)$ se llama *smoothing spline*.

El parámetro λ controla el compromiso entre las fluctuaciones de la curva y el ajuste de los datos, por esta razón usualmente se denomina a λ como el *parámetro de suavizamiento*. Cuando λ es grande se da prioridad a que la curva no tenga fluctuaciones significativas y los estimadores potenciales cuya segunda derivada es grande son fuertemente penalizados, por lo que no son seleccionados. Para valores pequeños de λ, es equivalente a poner mayor énfasis en el ajuste de los datos, y si $\lambda = 0$, se obtiene un estimador que interpola los datos, es decir, la curva pasa por todos los puntos (t_i, y_i).

Por ejemplo, en la Figura B.1 se muestran los datos con ruido y_i (triángulos rojo) y los datos sin ruido $\mu(t_i)$ (puntos azules). El objetivo es que a partir de los datos con ruido $y_0, \ldots, y_n$ se obtenga un estimador $\mu_\lambda(t)$ de $\mu(t)$ minimizando (B.3), para algún $\lambda > 0$, tal que $\mu_\lambda(t_i)$ aproxime a los datos sin ruido $\mu(t_i)$. En la Figura B.2 se muestra la gráfica del smoothing spline $\mu_\lambda(t)$ que se obtiene al minimizar (B.3) con $\lambda = 0.042$. Se puede ver que se obtiene una curva suave, que aunque no coincide del todo con $\mu(t)$, es una aproximación que sí logra describir su comportamiento general. En la Figura B.3 se muestra el histograma de los residuales $y_i - \mu_\lambda(t_i)$. Se puede ver del histograma que los residuales no tienen algún tipo de sesgo y que la mayoría de ellos son pequeños, por lo que sin conocer los verdaderos valores, el histograma indica que se hizo un buen ajuste.

Si el histograma hubiera mostrado algún sesgo o que los residuales no provienen de una distribución normal, habría que probar con otro valor para el parámetro λ.

En la Figura B.4 se muestran diferentes smoothing splines obtenidos para diferentes valores λ. En la parte superior de la Figura B.4 se puede ver que para $\lambda = 0$ el smoothing spline es un interpolador, es decir, su gráfica pasa a través de todos los datos, por lo que el error en el ajuste de datos (B.1) es $E_1 = 0$, pero tiene muchas fluctuaciones y por eso E_2 en (B.2) tiene el valor mayor de los cuatro casos. Este ajuste no es útil porque el smoothing spline explica tanto el modelo subyacente $\mu(t)$ como el ruido ϵ_i. Conforme λ aumenta se le da más peso al término E_2 por lo que al minimizar (B.3), se obtiene un smoothing spline en el que aumenta el valor del error E_1 del ajuste a los datos, pero va disminuyendo el valor E_2, lo cual indica que las fluctuaciones de la curva son menos significativas. El caso extremo es cuando λ es muy grande en el que el spline básicamente coincide con el resultado que se obtiene al ajustar una recta mediante el método de mínimos cuadrados lineales. De este modo, una parte importante de este método es la elección del valor del parámetro de suavizamiento λ y hay varios criterios que se pueden usar para seleccionarlo, como la técnica de validación cruzada, el criterio de información de Akaike o el criterio de información bayesiana [Wu 06].

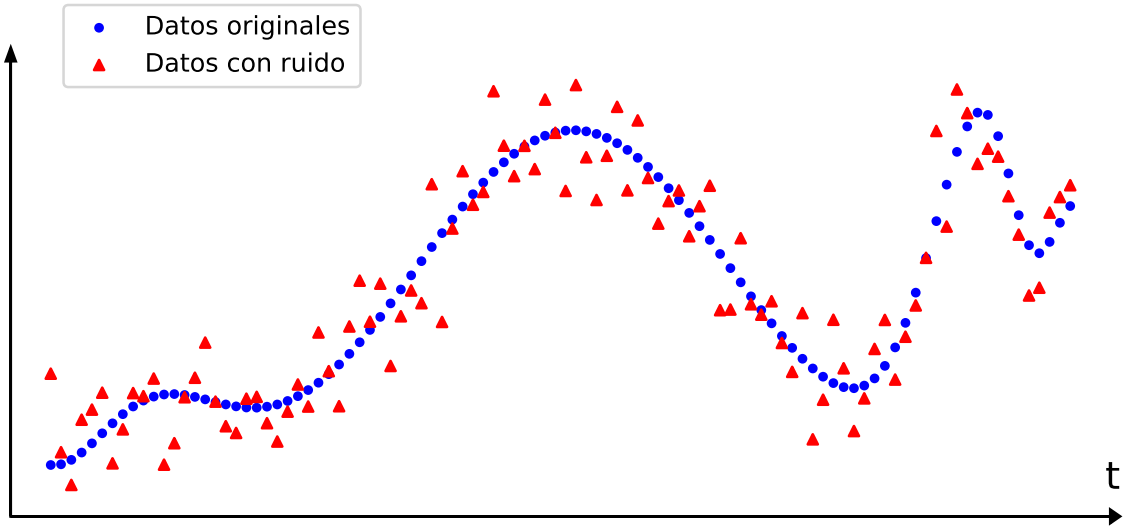


Fig. B.1: Los puntos azules son los datos originales $\mu(t_0), \ldots, \mu(t_n)$, mientras que los puntos rojos son los datos con ruido $y_i = \mu(t_i) + \epsilon_i$ para $i = 0, \ldots, n$.

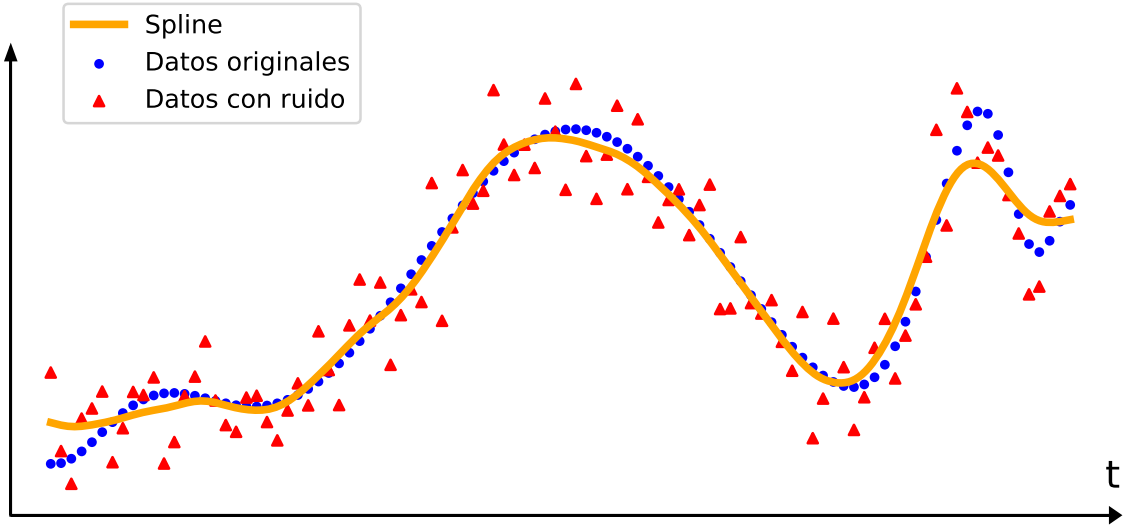


Fig. B.2: La línea naranja representa al smoothing spline, que es el estimador de $\mu(t)$. Si bien no pasa a través de todos los datos originales (puntos azules), sí logra pasar cerca de ellos.

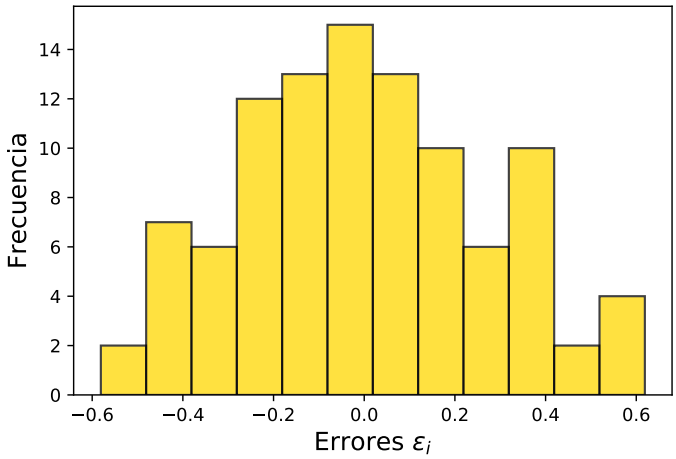


Fig. B.3: Histograma de los residuales $y_i - \mu_\lambda(t_i)$ para $i = 0, 1, \ldots, n$.

Fijo el valor de λ, el smoothing spline cúbico se obtiene al resolver un sistema de ecuaciones lineales. La construcción de este sistema es descrito en [Green 94].

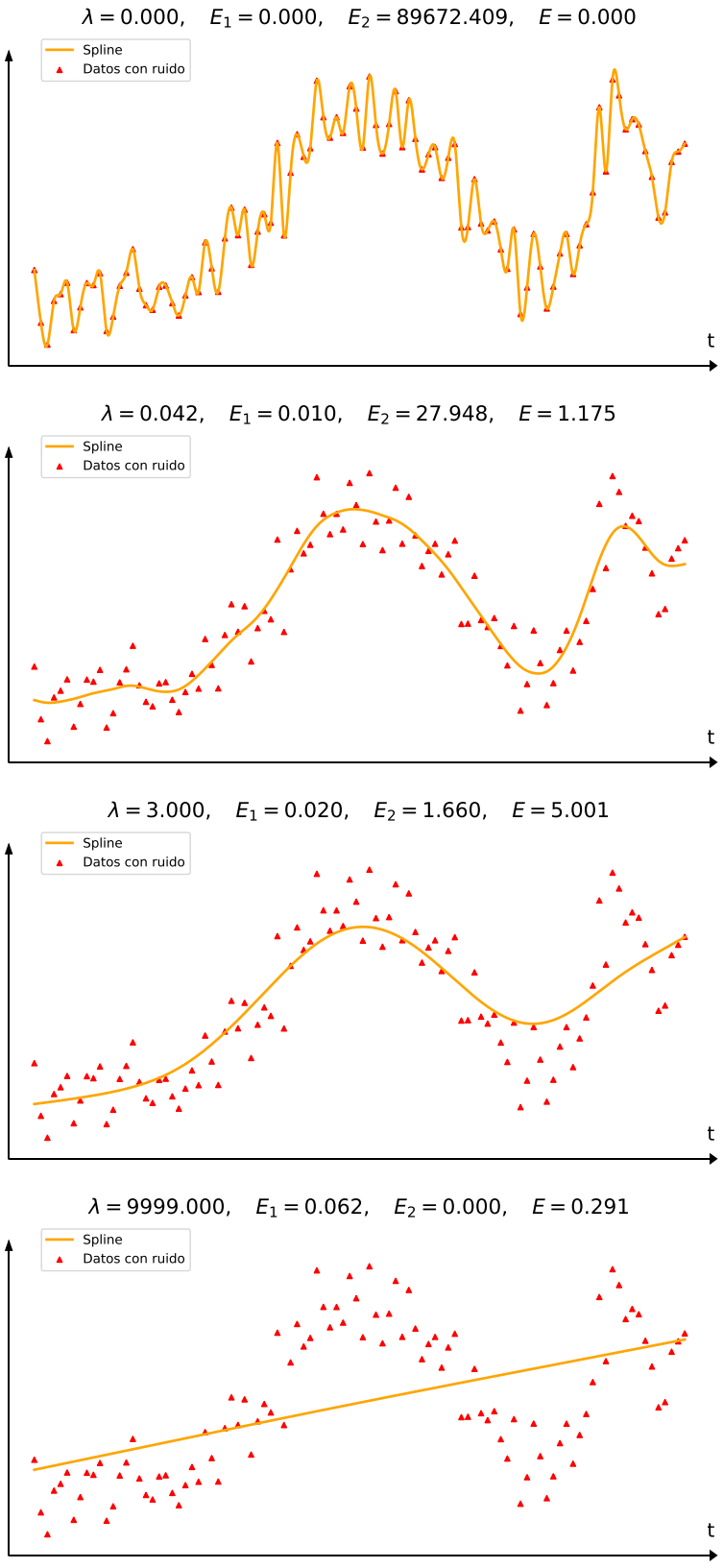


Fig. B.4: Diferentes smoothing splines obtenidos al usar diferentes valores de λ.

Bibliografía

[Eubank 94] R. L. Eubank. *A Simple Smoothing Spline*. The American Statistician, vol. 48, no. 2, pages 103–106, May 1994.

[Fahrmeir 13] L. Fahrmeir, T. Kneib, S. Lang & B. Marx. Regression: Models, methods and applications. Springer–Verlag, 2013.

[Green 94] P.J. Green & B.W. Silverman. Nonparametric regression generalized linear models: a roughness penalty approach. Monographs on Statistics and Applied Probability 58. Springer Science+Business Media, B.V., 1994.

[Quarteroni 00] Alfio Quarteroni. Numerical mathematics. Springer Verlag, 2000.

[Wu 06] H. Wu & Zhang J.T. Nonparametric regression methods for longitudinal data analysis. John Wiley & Sons, Inc., 2006.

[Weglarczyk 18] Weglarczyk, Stanislaw. Kernel density estimation and its application. ITM Web of Conferences, 2018.