TRABAJO FIN DE MÁSTER

Máster en Bioestadística

ANÁLISIS DE SERIES TEMPORALES PARA LA DETECCIÓN DE BROTES DE SALMONELOSIS

SEPTIEMBRE 2021

Izan Hoyuela Iglesias (UCM – Facultad de Estudios Estadísticos)

Miguel Ángel Moreno Romo (UCM – Facultad de Veterinaria) *Tutor*

Julio Álvarez Sánchez (UCM – Facultad de Veterinaria, VISAVET) *Tutor*

ÍNDICE

RESUMEN

Antecedentes: La salmonelosis es la primera causa de brotes de enfermedades de origen alimentario de los países desarrollados y, según algunas estimaciones, en los EE. UU podrían producirse anualmente más de un millón de casos. Las infecciones por este agente patógeno son de obligada notificación a las autoridades sanitarias en multitud de países, por lo que está disponible una gran cantidad de información sobre la frecuencia y distribución de la enfermedad**.**

Objetivo: Este trabajo pretende evaluar la capacidad de varios tipos de análisis para capturar la variabilidad en el número de casos esporádicos de salmonelosis, de modo que puedan predecirse los casos esperados con antelación y detectar así desviaciones que puedan ser debidas a la presencia de brotes.

Métodos: Se analizó una serie temporal, comprendida entre los años 2005-2017, del número de casos semanales definidos como esporádicos por el departamento de salud de Minnesota. La capacidad de los modelos para detectar brotes se evaluó generando predicciones sobre cada semana del año 2018 y comparando posteriormente estas con los casos totales reales observados, incluyendo los procedentes de brotes conocidos. Adicionalmente, se generaron curvas ROC y se fijaron puntos de corte alternativos para optimizar la sensibilidad y la especificidad de la diferencia entre casos totales observados (esporádicos y pertenecientes a brote) y casos esporádicos predichos como herramienta para detectar brotes activos.

Resultados: Los mejores modelos ajustados a partir de las diferentes técnicas empleadas fueron un modelo autorregresivo integrado de media móvil estacional (SARIMA) (1,0,1) x (0,1,1)52 con constante, una regresión armónica dinámica con 10 términos de Fournier y un modelo ARIMA (0,1,1) para los errores, además el mejor modelo de redes neuronales autorregresivas (NNAR) tuvo parámetros (14,1,8) y las variables precipitación y temperatura como predictoras. Las mejores predicciones de los casos de salmonelosis esporádicos semanales del año 2018 fueron, según el error absoluto medio (2,740) y la raíz del error cuadrático medio (4,388), las obtenidos a través de la regresión armónica dinámica. Las predicciones puntuales de los casos semanales de salmonelosis humanas, una vez aplicados los puntos de corte óptimos, permitieron obtener una sensibilidad y especificidad para la detección de brotes activos en las semanas de 2018, del 76,67% y 72,73% en el caso de las obtenidas a partir de la regresión armónica dinámica, del 83,33% y 63,63% con las del SARIMA y del 93,33% y 40,90% con las del NNAR.

Conclusiones: Los modelos SARIMA, NNAR y las regresiones dinámicas armónicas fueron capaces de modelizar correctamente la distribución temporal de los casos esporádicos semanales de salmonelosis y de generar predicciones aceptables. Además, el método desarrollado, basado en la comparación entre predicciones puntuales semanales de casos esporádicos de salmonelosis con los totales observados, permite detectar semanas en las que hay brotes activos con una sensibilidad elevada y una especificidad aceptable.

Palabras clave: Salmonelosis, Biovigilancia, SARIMA, Regresión armónica dinámica, Redes neuronales, NNAR.

ABSTRACT

Background: Salmonellosis is the leading cause of foodborne illness outbreaks in developed countries, and by some estimations, more than one million cases may occur in the US annually. Notification to health authorities of cases caused by this pathogen is mandatory in many countries, so a large amount of information is available on the frequency and distribution of the disease.

Objective: This work aims to evaluate the capacity of various types of analysis to capture the variability in the number of sporadic cases of salmonellosis, in a way that the expected cases can be predicted in advance and thus detect deviations that may be due to the presence of outbreaks.

Methods: A time series, between the years 2005-2017, of the number of weekly cases defined as sporadic by the Minnesota Department of Health was analyzed. The ability of the models to detect outbreaks was assessed by generating predictions for each week of 2018 and then comparing these predictions with the actual total observed cases, including those from known outbreaks. Additionally, ROC curves were generated, and alternative cut-off points were set to optimize the sensitivity and specificity of the difference between total observed cases (sporadic and belonging to an outbreak) and predicted sporadic cases as a tool to detect active outbreaks.

Results: The best models adjusted from the different techniques used were a seasonal moving average integrated autoregressive model (SARIMA) (1,0,1) x (0,1,1) 52 with constant, a dynamic harmonic regression with 10 Fournier terms and an ARIMA model (0,1,1) for the errors, in addition the best autoregressive neural network model (NNAR) had parameters (14,1,8) and the variables precipitation and temperature as predictors. The best predictions of the weekly sporadic salmonellosis cases in 2018 were, according to the mean absolute error (2.740) and the root mean square error (4.388), those obtained through dynamic harmonic regression. The punctual predictions of the weekly cases of human salmonellosis, once the optimal cut-off points were applied, allowed obtaining a sensitivity and specificity for the detection of active outbreaks in the weeks of 2018, of 76.67% and 72.73% in the in the case of those obtained from the dynamic harmonic regression, 83.33% and 63.63% with those of the SARIMA and 93.33% and 40.90% with those of the NNAR.

Conclusions: SARIMA, NNAR and dynamic harmonic regressions were able to correctly model the temporal distribution of sporadic weekly cases of salmonellosis and to generate acceptable predictions. In addition, the method developed, based on the comparison between weekly punctual predictions of sporadic cases of salmonellosis with the observed totals, makes it possible to detect weeks in which there are active outbreaks with high sensitivity and acceptable specificity.

Keywords: Salmonellosis, Biovigilance, SARIMA, Dynamic harmonic regression, Neural networks, NNAR.

1. INTRODUCCIÓN

La Autoridad Europea de Seguridad Alimentaria (EFSA) estima que la salmonelosis humana supone una carga económica global anual que podría ascender a los 3.000 millones de euros (1), produciendo alrededor del mundo según algunas estimaciones más de 93 millones de casos y más de 155.000 muertes anuales (2). Por estas razones es necesario seguir desarrollando y mejorando sistemas de vigilancia epidemiológica que permitan reducir el impacto de las enfermedades de transmisión alimentaria, y en particular las producidas por *Salmonella*. (2).

1.1Salmonella

1.1.1 Organismo

Salmonella es un género bacteriano, perteneciente a la familia *Enterobacteriaceae*, de distribución universal. Está constituido por bacilos gran negativos, anaerobios facultativos, con flagelos peritricos que les confieren motilidad. El hábitat natural de estas bacterias es el intestino de animales homeotermos. Es un agente productor de zoonosis, es decir, es un agente patógeno que se transmite naturalmente entre los animales y el ser humano (zooantroponosis) (3). La principal fuente de infección para el hombre es el consumo de alimentos contaminados, destacando los huevos, las carnes crudas o poco cocinadas, especialmente las de aves de corral, así como la leche y productos lácteos no sometidos a tratamientos térmicos de esterilización. Las salmonelas también pueden infectar a las personas a través de la ingesta de aguas con presencia de estos agentes patógenos, así como a través de frutas u hortalizas contaminadas por estas (3,4).

Salmonella puede causar dos tipos diferentes de enfermedad en función del serotipo causante, salmonelosis y fiebres tifoideas. La salmonelosis generalmente suele ser autolimitante en personas sanas con el sistema inmunitario intacto, aunque también puede causar cuadros graves en este tipo de sujetos (3). La fiebre tifoidea es una enfermedad más grave con un peor pronóstico, pero también es mucho menos habitual y está controlada en los países desarrollados (4).

1.1.2 Salmonelosis

Esta enfermedad tiene una letalidad en la población general menor de un 1%. En brotes producidos en residencias de ancianos y hospitales la letalidad puede llegar a ser de un 3,6% (3). El periodo de incubación suele ser de entre 6 y 72 horas, y la dosis infectiva puede ser de 10^5 a 10^8 UFC (unidades formadoras de colonias), pudiendo llegar a ser suficiente tan solo 1 UFC en individuos con sistemas inmunitarios comprometidos (4). Los síntomas más habituales son náuseas, vómitos, calambres abdominales, fiebre y dolor de cabeza. La duración de los síntomas varía en función de las características propias del hospedador, de la dosis ingerida y de las propias características de la cepa implicada, durando habitualmente de 4 a 7 días, y presentándose los más agudos entre el primer y el segundo día. Además, se pueden producir complicaciones como la deshidratación y/o un desbalance electrolítico producto de la diarrea y los vómitos, los cuales pueden llevar a la muerte tanto a individuos jóvenes como ancianos e individuos inmunocomprometidos si el tratamiento no es adecuado (3,4).

1.1.3 Notificación obligatoria y sistemas de vigilancia

Las infecciones por *Salmonella* son enfermedades de notificación obligatoria en la mayoría de los países por su especial relevancia para la salud pública. Este es el caso de EE. UU. (5) y de 22 estados de la UE (6), entre ellos España, en la cual la salmonelosis humana es de notificación obligatoria desde el año 2015 (7,8). La notificación obligatoria implica que cualquier facultativo o institución de salud que detecte un caso de salmonelosis debe notificarlo al organismo de vigilancia epidemiológica correspondiente (6). Estos organismos, u otras entidades sanitarias, publican los criterios para la definición y clasificación de los casos para facilitar la armonización de los datos y la notificación, siendo en la mayoría de los países necesaria una confirmación por cultivo para clasificar un caso como confirmado (5,6) además también es habitual que existan criterios clínicos y epidemiológicos para la clasificación de los casos (9). La salmonelosis es una enfermedad en la que gran parte de los casos forman parte de brotes, estando los criterios para la asociación de un caso a un brote, así como para la definición de un brote también especificados por las autoridades sanitarias (5,6). Un brote generalmente es definido como la existencia de dos o más casos asociados a una exposición pasada común a la misma fuente de contagio (7,9). El resto de los casos (no asociados a un brote) se consideran esporádicos (6). La notificación obligatoria produce gran cantidad de datos, cuyo análisis permite obtener un conocimiento sobre la incidencia y evolución de la

enfermedad brindando así la oportunidad de desarrollar correctamente políticas de salud pública que ayuden a reducir su impacto a través de la prevención (5).

Los sistemas de vigilancia de salmonelosis pueden ser pasivos o activos en función del papel que juegue el organismo de vigilancia epidemiológica. En el primero son las propias instituciones de salud las que entregan los datos de cada caso a los organismos de vigilancia correspondientes, lo cual abarata costes en detrimento de un mayor riesgo de infranotificación y de producción de datos incompletos. En los sistemas de vigilancia activa son los propios organismos de vigilancia epidemiológica los que proactivamente recopilan la información contactando a las instituciones de salud, favoreciendo una mayor precisión en los datos y una menor infranotificación, a costa de una mayor inversión en recursos (10,11). Existe otra forma de vigilancia, llamada centinela, la cual es especialmente útil cuando el objetivo es recoger la información para evaluar la presencia de tendencias más que controlar los casos individuales. En este tipo de sistema se realiza una vigilancia activa de una muestra de la población a través de la selección de una cohorte de instituciones de salud, o de sanitarios, de manera aleatoria o en base a ciertos criterios habitualmente geográficos (10,11,12). En algunos documentos científicos a esta forma de vigilancia también se le considera como de vigilancia activa (12).

En EE. UU la vigilancia de la salmonelosis humana se desarrolla principalmente a través de tres sistemas de vigilancia pasiva, el Laboratory-based Enteric Disease Surveillance (LEDS) system, el National Notifiable Diseases System (NNDSS) y el National Outbreak Reporting System (NORS), encargado de la recolección de la información relativa a brotes (5), y uno de vigilancia centinela, el Foodborne Disease Active Surveillance Network (FoodNet) (5,13). FoodNet abarca a 48 millones de personas, el 15% de la población estadounidense. En los estados de Connecticut, Georgia, Maryland, New Mexico, Oregon, Tennessee y Minnesota el sistema cubre al 100% de la población permitiendo tener datos de excelente calidad sobre estas regiones (13).

En la UE los sistemas de vigilancia epidemiológica de salmonelosis varían entre los estados miembros. En el año 2014, según el Centro Europeo para la Prevención y Control de Enfermedades (ECDC), tres países tenían sistemas de vigilancia activa de salmonelosis humana mientras que el resto tenían sistemas de vigilancia pasiva (14). En España el sistema de vigilancia es pasivo y es realizado por la Red Nacional de Vigilancia Epidemiológica (RENAVE), coordinada por el Instituto de Salud Carlos III (ISCIII) (7). La implementación del sistema de vigilancia de *Salmonella* en las CCAA está siendo

paulatino. En el año 2018, el último en el que el ISCIII publicó datos sobre *Salmonella* (15), no se reportaron datos en Andalucía, Asturias, Baleares, Galicia y Murcia, siendo la cobertura de vigilancia del 66,6%, abarcando a 31.578.578 de personas. Otros países europeos en los cuales el sistema de vigilancia no llegó al 100% de la población en el 2019 fueron Francia y Países Bajos con un 48 y 64% respectivamente (6).

1.1.4 Frecuencia de la enfermedad

UE

La EFSA y el ECDC elaboran un informe anual llamado "One health zoonoses Report", en el cual se integran y analizan los datos del monitoreo de zoonosis realizados por los sistemas de vigilancia epidemiológica de los estados miembros (6,16). Según el último informe publicado por la EFSA y el ECDC correspondiente al año 2019, en el conjunto de la UE hubo un total de 87.923 casos confirmados (20,0 por cada 100.000 habitantes), y se dispuso de información sobre el estatus de hospitalización del 44,5% de los casos confirmados (39.126 casos), de los cuales el 42,5% requirió de hospitalización (16.628), es decir el 18,9% del total. Diecisiete países proporcionaron información sobre el desenlace de la enfermedad, notificándose 140 muertes y una letalidad del 0,22%. Según el citado informe en el año 2019 fueron identificados 926 brotes alimentarios causados por *Salmonella* con 9.169 casos asociados que provocaron 1.915 hospitalizaciones y 7 muertes. Considerando solo los casos originados en la UE, se calcula que en torno al 11,6% de los casos detectados podrían venir asociados a brotes de origen alimentario, siendo el resto esporádicos (6).

En la UE, *Salmonella* es la segunda causa de infección gastrointestinal de la que más casos se reportan (26,6%), la primera causa de hospitalizaciones (40,8%) y la segunda causa de muertes por detrás de *Listeria* (24,8%) (año 2019). Además, es la primera causante de brotes de origen alimentario (17,8%), la segunda en el número de casos totales derivados de brotes (18,5%) por detrás de Norovirus (causante de cuadros clínicos leves generalmente), la primera de hospitalizaciones (49,6%) y la segunda en muertes (11.6%) por detrás de *Listeria* (6).

España

En el mismo informe de la EFSA y la ECDC se indica que en España se registraron 5.103 casos confirmados, si bien es cierto que debido a que no todas las comunidades autónomas

notificaron datos por la crisis de la COVID19 (el informe se publicó este año) no se ofrece información sobre la incidencia por desconocimiento del porcentaje de población al que dan cobertura las notificaciones (6). Los datos completos más recientes publicados sobre *Salmonella* en España se pueden encontrar en un informe del Instituto de Salud Carlos III elaborado con datos de RENAVE de los años 2017 y 2018 (15), también se puede encontrar información en los informes de la EFSA y la ECDC del 2018. El ISCIII notificó ese año 8.872 casos causados por *Salmonella* con una incidencia de 27,77 por 100.000 habitantes y una letalidad de 0,12%, reportándose 11 muertes. Es preciso reseñar otra vez que el sistema de vigilancia no da cobertura a todas las comunidades autónomas. En este informe (informe RENAVE) no se comunican las hospitalizaciones ni la información relativa a brotes, aunque el propio ISCIII cuenta con una base de datos de brotes y el MSSSI posee la información sobre hospitalizaciones en el Mínimum Basic Data (MBDS) (7,17). Los datos relativos a los brotes se pueden encontrar en el informe de la EFSA y de la ECDC, según el informe publicado por estos organismos en España hubo un total de 229 brotes con 1.829 casos asociados, de un total de 8.730. Conviene aclarar que las diferencias en el número de casos notificados por organismos nacionales y europeos son comunes debido a diferencias en la definición de los casos y a desfases en los periodos de extracción y envío de los datos (6). El último dato de hospitalizaciones por *Salmonella* en España encontrado se corresponde con en el año de 2015: en ese año se produjeron 3.776 hospitalizaciones y un total de 9.069 casos notificados (7).

EE. UU

Los CDC, en un informe realizado a partir de los datos recopilados por el sistema de vigilancia LEDS, señalan que en el año 2016 se notificaron 46.623 casos confirmados de salmonelosis en EE. UU., lo cual indica una incidencia por cada 100.000 habitantes de 14,51 casos (18). En este informe no se reporta el número de hospitalizaciones ni de muertes; sin embargo, en un informe realizado también por los CDC con datos de 2016 provenientes del sistema de vigilancia centinela FoodNet (el cual da cobertura a menos población que LEDS) se incluye información sobre las hospitalizaciones. FoodNet registró un total de 7.554 casos confirmados con una incidencia de 15,4 casos por cada 100.000 habitantes (incidencia similar a la obtenida a partir de LEDS), de los que 2.163 necesitaron hospitalización (29% del total). Un total de 39 casos finalizaron en muerte suponiendo un 0,4% del total (19). En ese mismo año fueron reportados por los CDC, a

través del Foodborne Disease Outbreak Surveillance System, 132 brotes causados por *Salmonella* con 3.047 casos asociados y un total de 456 hospitalizaciones (20).

En EE. UU. las enfermedades por *Salmonella* son las que más reportes de casos confirmados totales provocan (38.9%), las que más hospitalizaciones causan (46,7%) y las que más muertes producen (40%) de las relacionadas con los alimentos (19). Además, son la principal causa de brotes de origen alimentario (21%), de casos asociados a estos (25%) y de hospitalizaciones provocadas por dichos brotes (54%) (20).

Infranotificación e impacto real de la salmonelosis

Según el ECDC los casos notificados son solo la punta del iceberg de las infecciones producidas por *Salmonella* (21,22). Los modelos desarrollados a partir de mediciones de anticuerpos en diferentes países europeos indican que la "seroincidencia" de *Salmonella* podría ir desde 0,07 (IC 95% 0,020 – 0,139) y 0,08 (IC 95% 0,020 – 0,139) infecciones por persona y año en Finlandia y Dinamarca respectivamente, hasta las 0,61 (0,435 – 0,789) y 0,55 (0,382 – 0,735) de España y Polonia. Este indicador epidemiológico no se corresponde exactamente con la definición de la incidencia clínica, pero sirve para dar una idea de la fuerza de transmisión y la situación epidemiológica de la salmonelosis en Europa (22). Un estudio destacado como de referencia por los CDC realizó una estimación del verdadero alcance de las enfermedades de transmisión alimentaria en los EE. UU. teniendo en cuenta el infradiagnóstico y la infranotificación de los casos. El resultado fue una estimación de 1.027.561 casos totales de salmonelosis (IC 90% 644.786–1.679.667), 55.961 (39.534–75.741) hospitalizaciones y 378 muertes (0-1.1011). *Salmonella* según este estudio es la segunda causa de infecciones de transmisión alimentaria (11%), solo por detrás de *Norovirus*, de curso mucho más leve, la primera que más hospitalizaciones causa (35%) y la primera que más muertes causa (28%) (23).

El primer objetivo de la vigilancia de brotes de enfermedades transmitidas por alimentos debe ser la pronta identificación de conglomerados de casos relacionados espacial o temporalmente (24) de manera que pueda aplicarse una intervención precoz que limite el alcance de dichos brotes. En el caso de la salmonelosis, la primera causa de brotes en países desarrollados (16,20), cobra especial importancia este objetivo debido a la gran proporción de casos asociados a brotes, siendo la detección precoz de estos una herramienta que puede limitar de manera ostensible su impacto. En los últimos años ha crecido el interés en la aplicación métodos estadísticos para la detección de aumentos

inesperados de casos de enfermedades infecciosas lo cual puede servir al objetivo de limitar su alcance (25).

1.2 Series temporales

Desde la primera década de los dos mil hay una gran actividad investigadora relacionada con la detección prospectiva de brotes de enfermedades infecciosas, una disciplina habitualmente conocida como biovigilancia. Algunos de los métodos más útiles para este fin tienen como base el análisis de series temporales (25).

Una serie de tiempo es una colección de observaciones realizadas secuencialmente a lo largo del tiempo. Hay ejemplos de estas estructuras de datos en numerosas disciplinas, desde las ciencias de la vida hasta la ingeniería, pasando por la economía (26). Los sistemas de vigilancia epidemiológica habitualmente proporcionan datos de esta naturaleza y su análisis puede servir para diferentes fines como la descripción, la explicación o la predicción de la evolución de una enfermedad, el campo más importante de la detección de brotes (27).

Una de las características claves de las series temporales es la ausencia de independencia entre medidas consecutivas y en ocasiones incluso entre valores separados por más de una medición, o lo que es lo mismo, por más de un retardo. Al igual que ocurre con los estudios longitudinales, en los que tenemos sujetos con medidas repetidas y necesitamos de enfoques que permitan considerar esa relación entre medidas de un mismo individuo como los modelos mixtos, en el análisis de series temporales los enfoques comunes como la regresión simple no son apropiados. En última instancia lo que distingue los datos de medidas repetidas y los datos de series temporales es que en los primeros varios individuos suelen estar involucrados, habiendo pocas medidas de cada uno, mientras que en los segundos los datos corresponden habitualmente a un único individuo sobre el que se realizan gran cantidad de mediciones. En cualquier caso, la idea detrás del análisis de ambos es la misma: considerar la correlación entre valores (28).

1.2.1 Modelos de series temporales

Los modelos más utilizados en el análisis de series temporales en multitud de disciplinas son los modelos autorregresivos integrados de media móvil (ARIMA) los cuales pueden extenderse para considerar posibles estacionalidades en las observaciones (SARIMA) (26,29). No obstante, existen otros menos conocidos y menos empleados en la biovigilancia y en las ciencias de la vida que ya han demostrado su potencial en este

campo (y especialmente en otros como la economía), como los modelos autorregresivos de redes neuronales (NNAR) (30), las regresiones dinámicas armónicas (DHR) (31) y las regresiones dinámicas con errores ARIMA. Todas estas técnicas permiten modelizar series temporales con componente estacional o lo que es lo mismo, series temporales con variaciones periódicas a lo largo de un periodo de tiempo concreto (29), y por ello resultan prometedoras para analizar series temporales de datos de salmonelosis, la cual, tiene un claro aumento de los casos en verano frente a los meses más fríos.

1.2.1.1 Modelos autorregresivos integrados de media móvil (ARIMA)

Los modelos ARIMA con parámetros (p, d, q) combinan dos tipos de componentes para explicar dos posibles dinámicas de las series temporales, uno autorregresivo (p) y otro de medias móviles (q), a los que se les pueden añadir diferenciaciones (d) para lograr la estacionariedad (32).

"La correlación es a la regresión, lo que la autocorrelación es a la autorregresión". Esto quiere decir que de igual manera que una variable que esté relacionada con la variable dependiente puede utilizarse para su predicción o explicación en una regresión, en una variable autocorrelada pueden utilizarse los valores pasados como predictores de valores futuros. Esta última es la razón de ser del parámetro (p) del modelo ARIMA, el cual indica el número de valores de observaciones anteriores, o retardos, que se utilizan para predecir una presente. Un ARIMA (1,0,0) utilizaría la observación inmediatamente anterior para explicar la presente (33). La ecuación que define este proceso es:

$$y_t = \phi\,(y_{t-1}) + e_t \qquad (1)$$

donde "ϕ" es el coeficiente autorregresivo, "y_{t-1}" la observación anterior y "e_t" el error. De manera general estos modelos vendrían dados por:

$$y_t = \phi_1(y_{t-1}) + \phi_2(y_{t-2}) + \cdots + \phi_p(y_{t-p}) + e_t \qquad (2)$$

El componente de Medias móviles (q), por su parte indica el número de valores anteriores cuyos errores se utilizan para predecir un valor presente. Para comprender este componente hay que explicar la razón de ser de las diferenciaciones (d). Los ARIMA solo pueden emplearse para explicar o modelizar procesos estocásticos o estacionarios en los cuales sus propiedades (media, varianza y estructura de autocorrelación) no dependen

del tiempo en el que se observan. Diferenciar la serie (restar a los valores presentes el inmediatamente anterior) puede servir para eliminar tendencias que alteren esas propiedades, indicando "d" el número de veces que se repite ese proceso (o el número de diferenciaciones que se aplican). En consecuencia, los procesos estocásticos están determinados por dos componentes o fuerzas principales: sus propios valores previos (explicados por su componente autorregresivo (AR) y las variaciones impredecibles producidas por multitud de variables que cambian e interaccionan a lo largo del tiempo de tal manera que son ostensiblemente aleatorias. Es el efecto de estas variaciones impredecibles lo que busca explicar el componente de medias móviles (*Moving Averages*, MA). Conceptualmente los modelos ARIMA $(0, 0, q)$ o MA (q) buscan explicar un valor presente a través de una función que lo relacione con las variaciones impredecibles de "q" valores anteriores, o, en otras palabras, con el error de los términos de "q" tiempos anteriores, de tal manera que un ARIMA (0,0,1) utilizaría el error de la observación inmediatamente anterior para explicar el valor presente y vendría definido por la ecuación (29,33):

$$y_t = (\theta\, e_{t-1}) + e_t \qquad (3)$$

Donde "e_{t-1}" es el error aleatorio de la observación anterior y "θ" su coeficiente. De manera general estos modelos vienen dados por la siguiente ecuación:

$$y_t = \theta_1(e_{t-1}) + \theta_2(e_{t-2}) + \cdots + \theta_p(e_{t-q}) + e_t \qquad (4)$$

En definitiva, los modelos ARIMA permiten estimar un valor concreto como una función lineal de las observaciones anteriores y de los errores de estas. De manera general si el modelo necesita una diferenciación para lograr la estacionariedad el modelo vendría dado por:

$$y'_t = c + \phi_1(y'_{t-1}) + \cdots + \phi_p(y'_{t-p}) + \theta_1(e_{t-1}) + \cdots + \theta_q(e_{t-q}) + e_t \qquad (5)$$

Donde "y'_t" es la serie diferenciada. La constante "c" puede estar o no en el modelo (29,33).

SARIMA

Los modelos ARIMA pueden extenderse para la modelización de datos cuando hay estacionalidad, mediante la adición de componentes que expliquen esas posibles tendencias periódicas de la serie. A estos modelos se les conoce como SARIMA y vienen dados por los parámetros:

$$(p, d, q)x\ (P, D, Q)m$$

Donde "p", "d" y "q" son los mismos parámetros de los modelos ARIMA no estacionales, m es el número de observaciones de cada ciclo estacional y "P", "D" y "Q" son parámetros de la misma naturaleza pero que se aplican a observaciones de m periodos anteriores a la presente (29).

Los modelos SARIMA están diseñados para ciclos de observaciones relativamente cortos (de frecuencia mensual o cuatrimestral). El problema con periodos más largos es que hay "m-1" parámetros a estimar para los estados estacionales iniciales y esto puede conllevar problemas computacionales relacionados con la memoria disponible. Por ello puede ser recomendable utilizar otras aproximaciones para frecuencias más pequeñas que las mensuales, como es nuestro caso. Una solución pueden ser las regresiones dinámicas armónicas, basadas en el enfoque de regresiones dinámicas con errores ARIMA (29,34).

Regresiones dinámicas con errores ARIMA

Los modelos ARIMA y SARIMA no contemplan la inclusión de otras covariables susceptibles de contribuir a explicar el proceso estudiado; sin embargo, modelos como las regresiones lineales sí permiten la inclusión de gran cantidad de covariables, aunque no son capaces de captar las dinámicas de la serie temporal que sí puede captar el modelo ARIMA (o SARIMA). Intuitivamente, mediante la extensión de los modelos ARIMA (o SARIMA) podemos conseguir incluir covariables en el modelo a través de una regresión lineal, a la vez que explicamos las dinámicas de las series mediante modelos ARIMA (o SARIMA).

Las regresiones dinámicas con errores ARIMA (o SARIMA) ajustan una regresión en la que la autocorrelación de los residuos es corregida a través de un modelo ARIMA (o SARIMA). Estos modelos de manera general vienen dados por la ecuación:

$$y_t = ß_0 + ß_1 x_{1,t} + \cdots ß_k x_{k,t} + \eta_t \qquad (6)$$

Donde "η_t" es el error de la regresión, que puede estar autocorrelado y que se modeliza mediante modelos ARIMA (o SARIMA), siendo el error del modelo ARIMA (o SARIMA) "e_t" el que debe de tener una estructura de ruido blanco.

Es importante que todas las variables incluidas en el modelo tengan un comportamiento estacionario, tanto la dependiente como las independientes. Es común diferenciar, es decir; restar a los valores presentes el inmediatamente anterior o en el caso de diferenciaciones estacionales, el valor de un ciclo estacional completo atrás a un valor presente; a todas las variables del modelo de la misma forma, para así mantener la relación entre ellas (29).

1.2.1.2 Regresiones armónicas dinámicas

La regresión armónica dinámica está basada en el principio de que una combinación de funciones de seno y de coseno, términos de Fourier, pueden modelizar cualquier función periódica. De esta manera el patrón estacional podría modelizarse, de manera general, mediante la siguiente ecuación de regresión:

$$y_t = ß_0 + \sum_{k=1}^{K} [\alpha_k s_k(t) + \gamma_k c_k(t)] + e_t \qquad (7)$$

Donde "$s_k(t) = \sin(\frac{\mathbf{2\pi kt}}{m})$", " $c_k(t) = \cos(\frac{\mathbf{2\pi kt}}{m})$", "$m$" es el periodo estacional y "α_k" e "γ_k" son los coeficientes de la regresión. "e_t" es el error, que se modeliza a través de un proceso ARIMA, considerándose así las tendencias a corto plazo de la serie (34).

Esta metodología es apropiada para manejar estacionalidades largas (cada ciclo tiene muchos periodos u observaciones); además, permite incluir más de una tendencia estacional si es necesario. Otra ventaja es que permite controlar el suavizado del patrón estacional mediante el número de "K" términos de Fourier incluidos; cuantos menos términos, más suavizado. La única desventaja de esta técnica en comparación con los modelos SARIMA es que asumen que la estacionalidad es fija a lo largo del tiempo y no puede captar cambios en su patrón. En estos modelos también se pueden incluir variables regresoras externas ya que su funcionamiento es similar a las regresiones dinámicas con errores ARIMA del apartado anterior (29,34).

1.2.1.3 Redes neuronales autorregresivas (NNAR)

Las redes neuronales artificiales son métodos de predicción formulados a partir de modelos matemáticos con una estructura que se asemeja a la conexión entre las neuronas humanas y que permiten captar relaciones no lineales complejas entre variables dependientes e independientes. Es en este último punto donde estos modelos suponen una ventaja frente a los modelos ARIMA que solo pueden captar relaciones lineales (29,33).

La arquitectura de los modelos de redes neuronales se basa en neuronas o nodos individuales agrupados en capas, donde las entradas (inputs) del modelo se encontrarían en la primera capa (o la más profunda) y las salidas (outputs) en la última capa (o la más superficial). Un modelo con tan solo dos capas (figura 1) sería similar a uno de regresión lineal en el cual a cada neurona se le atribuiría un peso que sería similar al concepto de coeficiente de una regresión lineal común. Estos pesos se generan a través de un algoritmo de aprendizaje que minimiza una función de coste que en definitiva compara la salida del modelo con el valor real y que es usada por el algoritmo de aprendizaje para optimizar los parámetros de la red neuronal para la predicción (33).

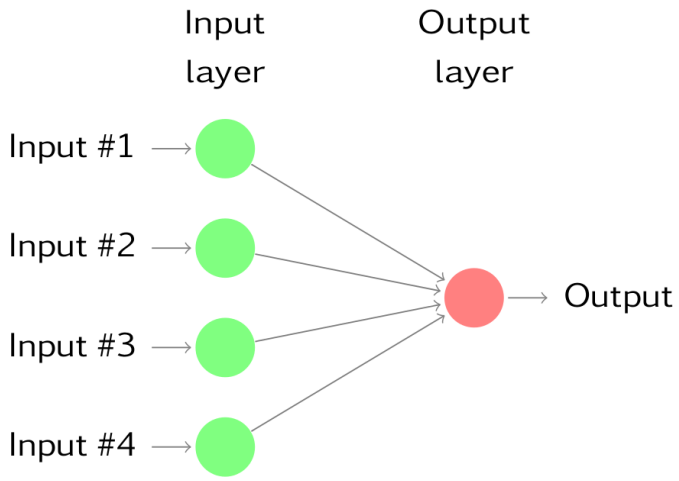

Figura 1. Estructura de red neuronal con dos capas (29).

En un caso como el de la figura 1 sería más eficiente un modelo como una regresión simple, pero una vez que incluimos una capa intermedia entre los inputs y los outputs, como por ejemplo en el modelo de la figura 2, la red neuronal puede establecer relaciones no lineales entre variables.

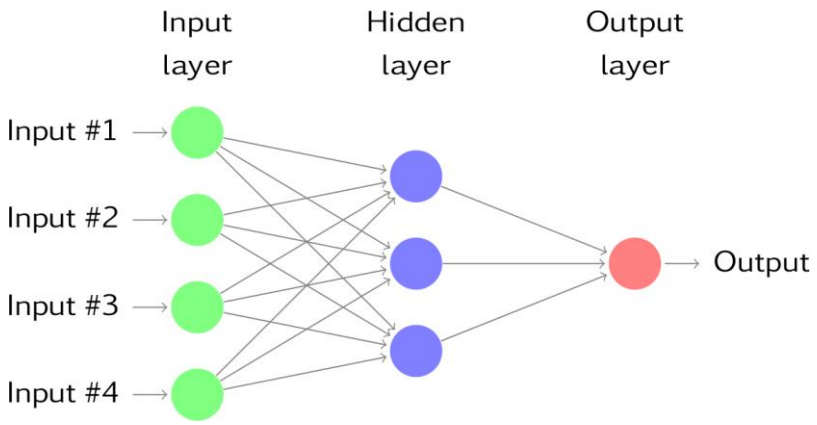

Figura 2. Estructura de red neuronal con una capa oculta (29).

En este tipo de estructuras el output de una capa anterior representa el input de una capa posterior y en cada neurona o nudo se combinan los inputs mediante combinaciones lineales ponderadas, las cuales son modificadas por funciones no lineales antes de producir el output. En el ámbito de las series temporales, en los últimos años han destacado las redes neuronales *feed-forward*, en las que la información se procesa únicamente en una dirección (29). Por ejemplo, en la figura 2 los inputs de la capa oculta vendrían dados por una combinación lineal del tipo:

$$z_j = b_j + \sum_{i=1}^{n} X_i w_{ij} \qquad (8)$$

Estos inputs serían transformados mediante funciones no lineales, las cuales se denominan funciones de activación; posteriormente la red neuronal pondera los outputs de la capa oculta para producir el output final. Los parámetros de la red neuronal son aprendidos a partir de los datos y es habitual que a los pesos utilizados en las ponderaciones se les pongan restricciones para que no sean demasiado grandes a través de parámetros "decay". Para comenzar el entrenamiento de la red neuronal se asignan valores iniciales a los pesos que posteriormente se van actualizando con los datos observados, lo que produce un factor de aleatoriedad a los pesos finales y por ello es habitual realizar varios entrenamientos desde puntos de partida diferentes para posteriormente promediar los resultados y obtener un resultado final (29,33).

Debido a que en el análisis de series temporales para el output final se utilizan valores de la propia variable sobre la que se predice, estos modelos reciben la denominación de redes neuronales autorregresivas (NNAR, por su acrónimo en inglés), con parámetros (p, k),

indicando "p" el número de retardos que se utilizan como inputs y "k" el número de neuronas de la capa oculta. Una ventaja adicional que ofrecen este tipo de modelos es que no tienen ningún tipo de condicionalidad ligada a la estacionariedad (29), lo cual limita la necesidad de transformaciones a ocasiones muy limitadas. Su ecuación general sería:

$$y_t = \alpha_0 + \sum_{j=1}^{k} \alpha_j g\left(ß_{0j} + \sum_{i=1}^{p} ß_{ij}\, y_{t-i}\right) + e_t \qquad (9)$$

Donde "y_t" es la salida final de la red, "α_j" y "$ß_{ij}$" los pesos, "p" el número de retardos utilizados como inputs, "k" el número de nodos de la capa oculta y "g" la función de activación.

En el caso de las series estacionales habría un parámetro adicional al modelo NNAR expresado como "P", el cual indica el número de retardos estacionales que se utilizan como input. Un modelo NNAR(p, k) es similar a un ARIMA $(p, 0, 0)$ pero con funciones no lineales, de igual manera que un modelo NNAR (p, P, k)se corresponde con un SARIMA (p,0,0) x (P,0,0) pero con funciones no lineales. Los intervalos de predicción pueden obtenerse a través de simulación mediante técnicas como el Bootstrap (29,33).

Estos modelos también pueden incluir regresores externos, añadiéndose una neurona más en la capa más profunda (la de los inputs) por cada covariable.

2. OBJETIVOS

El objetivo principal de este trabajo es evaluar la capacidad de varios tipos de análisis para capturar la variabilidad en el número de casos esporádicos de salmonelosis, de modo que puedan predecirse los casos esperados con antelación y detectar así desviaciones en el número de casos que puedan ser debidas a la presencia de brotes.

Los objetivos específicos son:

I. Estudiar la capacidad de diferentes técnicas de análisis de series temporales para predecir casos esporádicos de salmonelosis.

II. Estudiar la viabilidad de la comparación de las predicciones puntuales de casos esporádicos de salmonelosis con los casos totales observados como herramienta para la detección de aumentos inesperados en la notificación, sugestivos de la presencia de brotes activos.

3. MATERIAL Y MÉTODOS

3.1 Base de datos

Los datos analizados han sido facilitados por el departamento de salud de Minnesota, agencia encargada del sistema de notificación de casos en este estado de los EE. UU. La base de datos utilizada contiene datos relativos al número total de casos semanales de salmonelosis confirmados por cultivo en el estado de Minnesota del periodo 2005 – 2018; además estos casos están desglosados en aquellos asociados a un brote y aquellos definidos como esporádicos. La base de datos tiene información de 730 semanas con un total de 8.094 casos, de los cuales 6.850 son esporádicos y 1.244 se encuentran asociados a brotes. Los casos asociados a viajes fueron excluidos de esta base de datos.

Los criterios de definición de "caso confirmado" no han variado desde el año de inicio de la serie por lo que se asume que los sistemas de clasificación y de notificación no son una variable que afecte a posibles cambios en el número de casos notificados a lo largo del periodo de tiempo analizado.

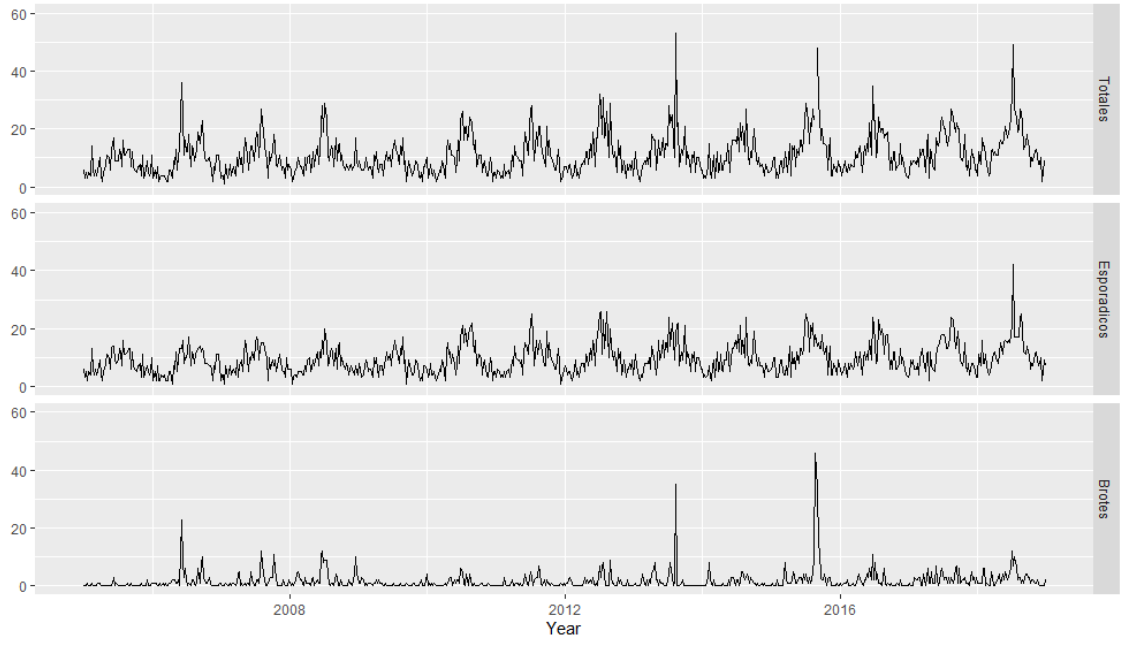

Figura 3. Serie temporal de los casos semanales de salmonelosis totales (arriba), esporádicos (centro) y asociados a brotes (abajo) registrados por el departamento de salud de Minnesota en el periodo 2005 – 2018.

Adicionalmente, se utilizaron también dos series temporales de datos de temperatura media máxima en grados centígrados y de precipitaciones en pulgadas del mismo periodo que la serie de salmonelosis, obtenidas a partir de los datos diarios de la estación meteorológica de New Hope (Station ID: 215838) (36).

3.2 Procedimiento realizado

En la figura 4 se puede observar un diagrama de flujo que sirve como resumen y explicación visual del proceso que se ha seguido para el ajuste y desarrollo de los modelos estudiados y para el paso final ligado a la detección de los brotes. Todos los modelos se ajustaron sobre la serie temporal de casos esporádicos semanales de salmonelosis del año 2005 al año 2017, dejándose los registros del año 2018 para evaluar la capacidad de los modelos ajustados para la predicción y la detección de brotes, definiéndose las semanas con más de un caso asociado a brote como semanas con brote activo. El paso final, que es probablemente el concepto clave para comprender la metodología, consiste en la predicción con cada modelo de los casos esporádicos semanales de salmonelosis del año 2018 para su posterior comparación con los casos totales semanales observados ese mismo año, de tal manera que un exceso en los casos totales semanales frente a los esporádicos predichos indicaría la presencia de un brote. Adicionalmente, mediante la fijación de un punto de corte, que maximizase la sensibilidad y especificidad del método propuesto, sobre la diferencia entre casos totales y esporádicos se buscó aumentar la calidad de los umbrales y la capacidad discriminatoria del método para clasificar correctamente a las semanas con brote y sin él. Por último, mediante la sensibilidad y especificidad alcanzada de cada modelo se comparó la utilidad de las predicciones de cada tipo de modelo para la detección de brotes.

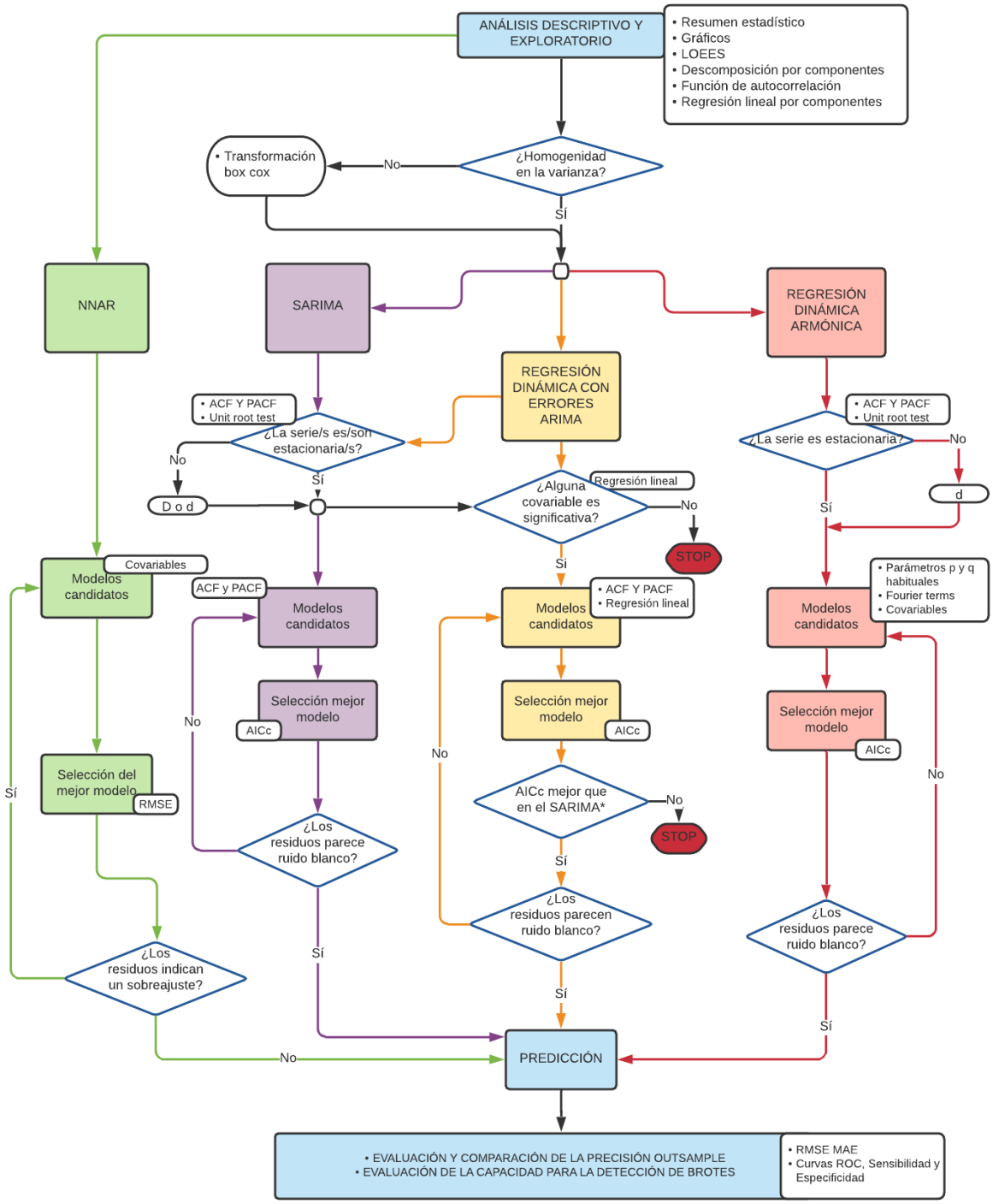

Figura 4. Diagrama de flujo del procedimiento desarrollado en este trabajo. (D = Diferenciación estacional, d = diferenciación no estacional o primeras diferencias)

Todos los análisis se realizaron mediante el software R (versión 4.0.2) a través de la interfaz de RStudio (versión 1.4.1106), utilizando los paquetes: “ggplot2” (37), “plyr” (38), “dplyr” (39), “lubridate” (40), “tseries” (41), “vars” (42), “forecast” (43), “RCurl” (44), “TSA” (45), “geoR” (46), “car” (47), “MuMIn” (48), “astsa” (49) y “feasts” (50).

3.2.1 Análisis descriptivo y exploratorio

La serie temporal utilizada para ajustar todos los modelos, correspondiente con el número de casos esporádicos semanales registrados por el departamento de salud de Minnesota en el periodo 2005-2017, tiene una media de 9,2 casos semanales con una desviación típica de 4,95 y una mediana de 8. El número mínimos de casos semanales registrados fue de 1 y el máximo de 26, correspondiéndose el primer cuartil a 5,25 casos y el tercero a 12 casos. En total fueron registrados en el periodo analizado 6.211 casos esporádicos.

En primera instancia se realizó una aproximación visual a la serie temporal analizada (2005 – 2017) con la que se exploraron sus posibles tendencias mediante su representación gráfica, al que se le añadió una curva suavizada de la evolución de los casos ajustada mediante LOESS. También se generaron gráficos de cajas en los que se agruparon los casos por semanas, meses y años. A partir de las visualizaciones se evaluó la homogeneidad de la varianza a lo largo de toda la serie temporal y la posible aplicación de una transformación Box-Cox para su estabilización. Además, se analizó visualmente la posible existencia de estacionalidad y tendencia, y por ende la posible necesidad de diferenciaciones de la serie para lograr su estacionariedad (27).

Regresión lineal.

Se estableció como variable dependiente el número de casos esporádicos por semana y como predictores el tiempo, para evaluar la tendencia, y una variable categórica correspondiente a la semana del año para evaluar la estacionalidad. Con este análisis de regresión se buscó ampliar el conocimiento proporcionado por los gráficos y la línea de tendencia ajustada por LOESS, obteniendo una primera información cuantitativa sobre los componentes de la serie (27).

Descomposición por componentes.

La serie temporal fue desagregada en cada uno de sus componentes con el fin de evaluarlos visualmente por separado. También se evaluó que el error se distribuyera de manera aleatoria, es decir, ser un ruido blanco. A continuación, se valoró la necesidad de descomponer la serie de manera aditiva (varianza constante en el tiempo) (27);

$$Y[t] = T[t] + S[t] + e[t] \quad (10)$$

O multiplicativa (varianza heterogénea);

$$Y[t] = T[t] * S[t] * e[t]$$

(11)

El proceso se realizó con la función decompose del paquete stats (41) la cual, en primer lugar, determina la tendencia a largo plazo mediante medias móviles y elimina este componente de la serie. Posteriormente computa la estacionalidad promediando por cada unidad temporal (en nuestro caso la semana), eliminándola también para finalmente evaluar el error. Con la visualización resultante se puede analizar cada componente y se obtiene más información sobre las características de la serie, facilitando la selección de futuras posibles transformaciones y estructuras para los modelos.

También se utilizó la representación gráfica de la función de autocorrelación (ACF) para evaluar la presencia de estacionalidad y estacionariedad. Las autocorrelaciones separadas por 52 retardos fueron interpretadas como evidencia de componente estacional.

3.2.2 Transformación Box-Cox

La varianza se estabilizó a través de una transformación Box-Cox antes de proceder a la selección de los parámetros del SARIMA, de la regresión dinámica armónica y de la regresión dinámica con errores SARIMA (29). La selección del parámetro lambda de la transformación se realizó a través de la función boxcoxfit del paquete geoR (46) que lo selecciona de manera automática.

Los modelos NNAR por su parte no tienen ningún tipo de requisito en lo que se refiere a la estacionariedad de la serie temporal ni en varianza ni en media, por lo que su modelización se realizó a partir de la serie original (29).

3.2.3 Modelización de los casos semanales esporádicos de salmonelosis

Se utilizaron las técnicas SARIMA, regresión dinámica con errores ARIMA, regresión armónica dinámica y NNAR.

3.2.3.1 SARIMA

Diferenciaciones

Tras los pasos anteriores se evaluaron la estacionariedad y la posible necesidad de diferenciaciones a nivel estacional o de primeras diferencias a través de la ACF de la serie transformada y a través de tests estadísticos como el Augmented Dickie Fuller (ADF) y

el Kwitakowski Philips Smith Sinn (KPSS) (32,51). Se utilizaron los dos tests porque ambos tienen las hipótesis alternativas y nulas intercambiados, considerándose útil utilizarlos de manera conjunta y confirmatoria. El ADF tiene como hipótesis nula que la serie no es estacionaria y como alternativa que, si lo es, mientras que en el KPSS las hipótesis tienen el rol opuesto; el p-valor para aceptar la hipótesis alternativa se fijó en 0,05 (32,51). Únicamente se consideró la serie estacionaria cuando ambas pruebas respaldaban dicha hipótesis.

En primer lugar, se visualizó el ACF, ante la más que segura presencia de estacionalidad (3,6) con las correlaciones más significativos siendo aquellas separadas por 52 semanas, se dio prioridad a una diferenciación estacional ya que de otra manera el modelo ajustado estaría sesgado al considerar que este patrón desaparecería con el tiempo. Posteriormente se evaluó la estacionariedad de la serie a través de una nueva visualización de la ACF y mediante los test estadísticos anteriormente mencionados, asumiéndose esta característica, si a partir de los tres métodos se obtienen las mismas conclusiones. Se mantuvo especial precaución en no caer en sobrediferenciaciones comprobando que la primera autocorrelación no fuese menor de -0.5 (52).

Selección de parámetros AR y MA

Analizando visualmente los correlogramas de la ACF y la autocorrelación parcial (PACF) se seleccionaron los modelos concordantes con los patrones observados (27), con diferentes combinaciones de parámetros de AR y MA (estacionales y no estacionales). También se evaluó la necesidad de incluir una constante basándose en el número de diferenciaciones aplicadas (52,53).

Los parámetros de los modelos candidatos fueron estimados mediante máxima verosimilitud a través de la función Arima del paquete forecast (43). El modelo final se seleccionó a través de la comparación de la calidad relativa de los modelos potenciales para explicar los datos de la serie utilizando el AICc (el de menor AICc) como recomiendan Hyndman y Athanasopoulos (29). El AICc es una versión corregida del AIC y para un modelo ARIMA viene dado por la siguiente ecuación:

$$AIC = -2\log(L) + 2(p + q + k + 1)$$

$$AICc = AIC + \frac{2(p + q + k + 1)(p + q + k + 2)}{T - p - q - k - 2}$$

(12)

Donde "L" es la verosimilitud de los datos, "p" el orden del componente AR, "q" el orden del componente MA, siendo "k" = 1 si no hay constante en el modelo y 0 si la hay. "T" es el número total de observaciones de la serie (29).

Regresión dinámica con errores ARIMA.

La regresión dinámica fue utilizada para comprobar si la inclusión de las variables temperatura media máxima en grados centígrados y precipitación en pulgadas podía ayudar a aumentar la calidad de las predicciones de los casos esporádicos de salmonelosis realizadas por el SARIMA (29).

Inicialmente se valoró la posible inclusión de las variables temperatura y precipitaciones en el modelo mediante una regresión simple utilizando un punto de corte de significación de 0,05. Las variables significativas fueron incluidas en el análisis mediante un modelo de regresión dinámica con errores ARIMA a través de la función Arima del paquete forecast. La estacionariedad de las series temporales de las covariables estudiadas fue evaluada de manera similar a la de los casos de salmonelosis; con el objetivo de no perder la relación entre las variables se aplicaron las mínimas diferenciaciones necesarias a todas las series para lograr la estacionariedad de todas ellas (29).

Con las covariables que resultaron significativas se ajustaron modelos de regresión dinámica con errores ARIMA con parámetros AR y MA similares a los utilizados en el SARIMA (29).

Se seleccionó el mejor modelo de entre los candidatos evaluados a través del AICc. Siempre y cuando se aplicasen las mismas diferenciaciones a las series, el modelo de regresión dinámica seleccionado se comparó mediante AICc con el mejor modelo SARIMA resultante del paso anterior, utilizándose en caso de igualdad el SARIMA por ser considerado más parsimonioso.

Evaluación de los residuos.

Los residuos del modelo final seleccionado fueron analizados para determinar su utilidad para la predicción de brotes. Para evaluar que dichos residuos se comportaran como ruido

blanco se generó un ACF de los residuos y se realizó una prueba Ljung-Box basada en el siguiente estadístico Q* (27,29):

$$Q^* = T(T+2)\sum_{k=1}^{h}(T-k)^{-1}\,r_k^2$$

(13)

Donde h es el número máximo de retardos analizados, T el número total de observaciones r_k el coeficiente de correlación. En este caso se consideraron 104 retardos (el doble de los que componen un ciclo) ya que la serie es estacional. Cuanta más correlación haya entre los residuos más alto será el valor de Q^* el cual, si las autocorrelaciones vienen dadas por una estructura de ruido blanco, seguirá una distribución $\chi 2$.

Mediante un histograma se comprobó si los residuos estaban centrados en 0 y si seguían una distribución normal. Adicionalmente se realizó un test McLeod-Li para comprobar si los residuos tenían una heterocedasticidad condicional autorregresiva (proceso ARCH), señal de una varianza no constante a lo largo de los residuos provocada por periodos de volatilidad (54). La existencia de este proceso provocaría que los intervalos de predicción no fuesen fiables, requiriendo el uso de modelos ARIMA-GARCH para considerar esa volatilidad en la predicción. El test de McLeodLi utiliza los residuos al cuadrado para evaluar la presencia de un proceso ARCH, el 5% de retardos significativos fue fijado como umbral para asumir la presencia de este (55,56).

3.2.3.2 Regresión armónica dinámica

Diferenciaciones.

Solo se consideraron primeras diferencias ya que la parte estacional se modelizó mediante los términos de Fournier en la parte de la regresión del modelo (34). La ACF puede resultar confusa por la ausencia de una diferenciación estacional, que por la naturaleza de las series analizadas se presupone necesaria. Por este motivo la estacionariedad y la necesidad de diferenciaciones se evaluaron a partir de los test ADF y KPSS, aunque también se visualizó el ACF de manera complementaria.

Selección de términos de Fourier y parámetros AR y MA.

Preliminarmente se realizó un proceso secuencial donde se comenzó por ajustar un modelo con 21 pares de senos y cosenos, el máximo posible para una serie de frecuencia semanal (34), y se fue disminuyendo el número hasta los 6 pares, el mínimo para no tener una curva excesivamente suavizada. La estructura de los errores de estos modelos preliminares se seleccionó a través de la función de selección automática de parámetros auto.arima del paquete forecast (43). El número de términos de Fournier a partir del cual el AICc dejó de disminuir fue seleccionado junto con el anterior más complejo y el inmediatamente posterior más sencillo y con ellos se estimaron los modelos ARIMA para los errores con los parámetros más habituales (53). El modelo con menor AICc de fue el escogido para la predicción. El objetivo de este proceso de dos fases es en primer lugar filtrar el número idóneo de pares de senos y cosenos para explicar el componente estacional y en segundo lugar seleccionar la mejor estructura de parámetros ARIMA para explicar las dinámicas no estacionales.

Adicionalmente se probó a incluir las variables que resultaron significativas en la regresión lineal del apartado anterior como variables regresoras a modelos con los mismos parámetros y términos de Fournier que los seleccionados en el paso anterior. Finalmente se comparó el AICc de estos con los modelos sin variables externas y se seleccionó el mejor para las predicciones de acuerdo con ese criterio.

Evaluación de los residuos.

Los residuos que deben de tener estructura de ruido blanco y sobres los que se realizó el diagnóstico, son los errores del modelo ARIMA y no los de la regresión dinámica (29). Las comprobaciones realizadas para estos fueron las mismas que en el modelo SARIMA.

3.2.3.3 NNAR

Para el desarrollo de estos modelos se utilizó la función nnetar del paquete forecast (43) la cual permite seleccionar automáticamente los parámetros p y P del modelo, determinándose el número de nodos de la capa oculta (k) de acuerdo con:0

$$k = (p + P + 1)/2 \tag{14}$$

redondeándose al número entero más cercano.

En primer lugar, se ajustó un modelo sin ninguna variable externa, y posteriormente se ajustaron modelos con todas las combinaciones de variables externas usadas en este

trabajo; temperatura, precipitación, y ambas. Por último, se compararon todos estos modelos mediante la raíz del error cuadrático medio (RMSE) para la muestra utilizada para el entrenamiento (insample) y se seleccionó el mejor (29,57). En el siguiente apartado se explica el concepto del RMSE.

Evaluación de los residuos.

Se evaluó que los residuos estuviesen normalmente distribuidos y centrados en 0, ya que de otra manera podrían existir problemas de sobreajuste (57).

3.2.4 Predicción de los casos semanales esporádicos de salmonelosis

Con cada modelo seleccionado se realizó una predicción de los casos esporádicos semanales de salmonelosis de 2018. La calidad de estas predicciones fue evaluada a través del error absoluto medio (MAE) y de la raíz del error cuadrático medio (RMSE); ambas medidas de precisión están basadas únicamente en los propios residuos del modelo y por tanto dependen de la escala de estos (29,51), resultando por ello apropiados para este trabajo en el que se quiere comparar entre series con las mismas unidades. La expresión matemática de ambas medidas de precisión es la siguiente:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|\hat{y}_t - y_t|$$

(15)

Donde $\hat{y}_t$ es el valor predicho, y_t el real y n el número de observaciones.

La diferencia entre el RMSE frente al MSE es que el primero es más sensible a desviaciones grandes de la predicción sobre el valor real (29).

3.2.5 Detección de brotes.

Las predicciones puntuales del número de casos esporádicos semanales fueron comparadas con el número de casos registrados en las mismas semanas de 2018 con el fin de dar respuesta al objetivo principal de desarrollar una metodología capaz de detectar la presencia de brotes. Las semanas con más de dos casos registrados como pertenecientes a un brote se definieron como “semanas con brote activo” (y por tanto objetivo del análisis) y aquellas semanas en las que el número de casos totales registrados superó la

predicción de los modelos se consideraron, en primera instancia, como “semanas de alarma” o definidas por la metodología como con un brote activo. Estas definiciones permiten clasificar las semanas en cuatro categorías: verdadero positivo, falso positivo, verdadero negativo y falso negativo, a través de las cuales se calcularon la sensibilidad y especificidad del modelo. Este procedimiento se desarrolló para cada modelo y se compararon los resultados obtenidos en base a las diferentes predicciones obtenidas a partir de estos.

Adicionalmente se generó una nueva variable, la diferencia entre casos totales y casos esporádicos predichos, que se utilizó para generar curvas ROC con el fin de evaluar la capacidad discriminatoria obtenida a partir de las predicciones de cada modelo para la detección de semanas con brote activo, comparándose estas a través del área bajo la curva (AUC). Mediante este análisis se seleccionaron puntos de corte para optimizar la sensibilidad y la especificidad del método, utilizándose para ello el criterio de Youden (58). Con los nuevos puntos de corte se calcularon otra vez la sensibilidad y la especificidad, comparándose de nuevo los resultados obtenidos a partir de los diferentes modelos y con las clasificaciones realizadas con el punto de corte 0 utilizado originalmente.

4. RESULTADOS

4.1 Análisis descriptivo y exploratorio

La serie temporal analizada corresponde a los casos esporádicos de *Salmonella* confirmados por el departamento de Salud de Minnesota entre los años 2005 y 2017 (figura 5).

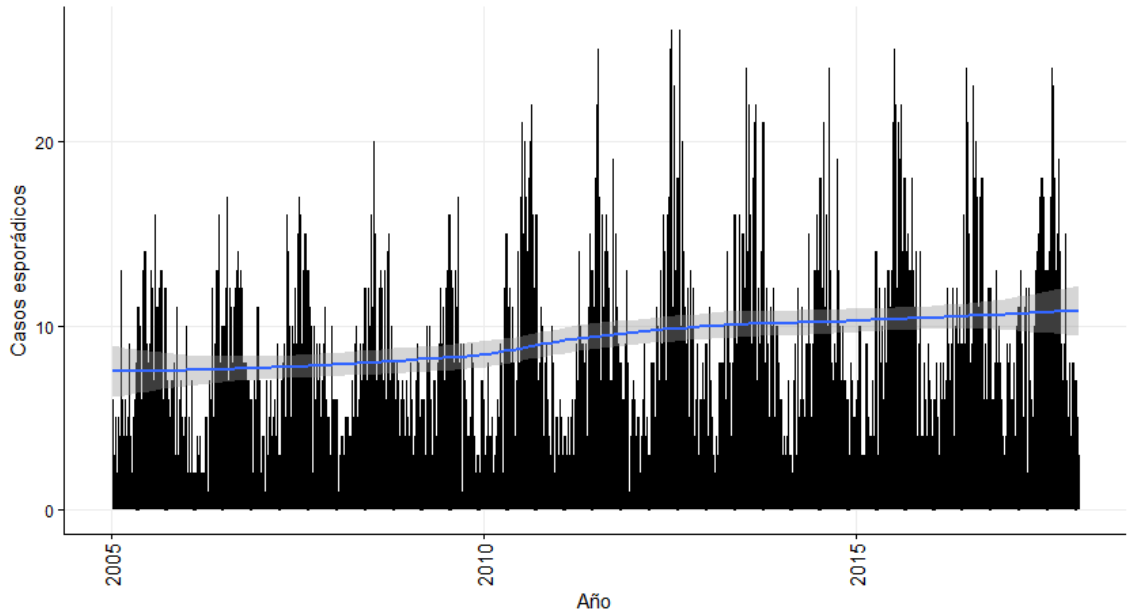

Figura 5. Número semanal de casos esporádicos de *Salmonella* notificados al Departamento de Salud de Minnesota, 2005 – 2017 y curva suavizada por LOEES.

En la figura 5 se observa cómo hay un claro componente estacional en el número de casos semanales esporádicos. Además, también se puede apreciar cómo estas oscilaciones estacionales no son homogéneas a lo largo de la serie, ya que a medida que pasan los años los picos son más pronunciados y la diferencia entre éstos y los valles es más grande, posible señal de la existencia de una varianza heterogénea que puede requerir de una transformación Box-Cox para lograr su homogenización. La existencia de ese componente estacional puede observarse también en la figura 6, donde se ve como la mediana de las cajas que agrupan al número de casos de las semanas centrales de todos los años (verano) tienen una mediana mayor que las semanas del principio y del final del año (invierno).

En la figura 5 también se aprecia, mediante la curva suavizada por LOESS representada en azul, como existe una aparente tendencia alcista en los casos de salmonelosis esporádicos semanales a lo largo del tiempo.

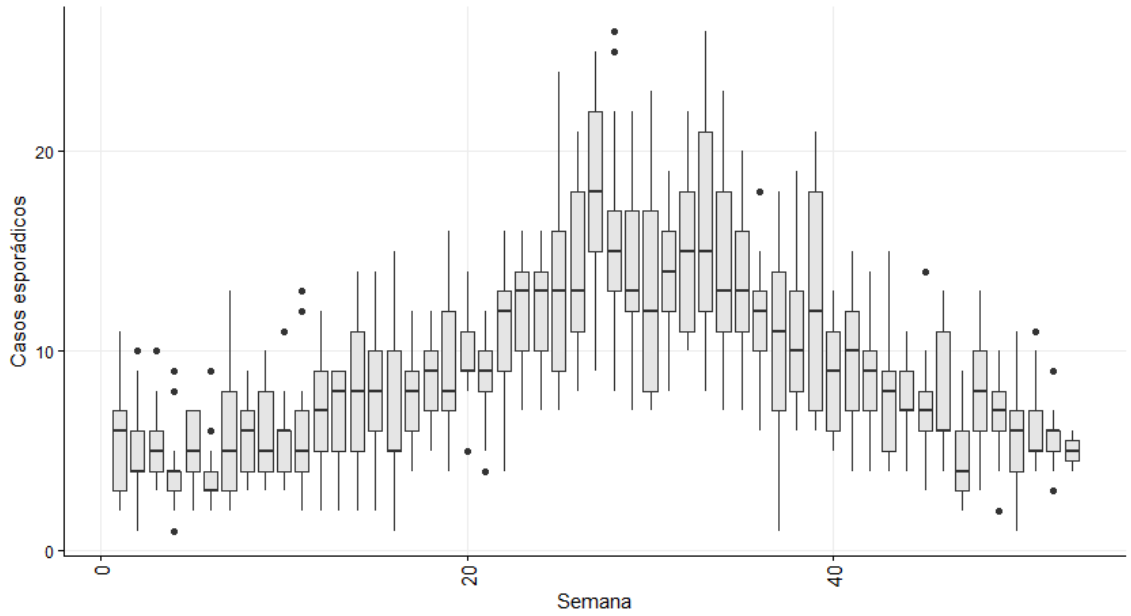

Figura 6. Diagrama de cajas del número semanal de casos esporádicos de Salmonella confirmados por el Departamento de Salud de Minnesota, 2005 – 2017. Cada caja representa la distribución del número de casos de salmonelosis producidos en la semana correspondiente (eje x) de todos los años que abarca la serie.

La tendencia alcista, así como el aumento de la varianza con los años, puede observarse de nuevo en la figura 7, donde se ve como la mediana y la amplitud de las cajas que agrupan el 50% central de la distribución de los casos semanales son mayores en los años más recientes.

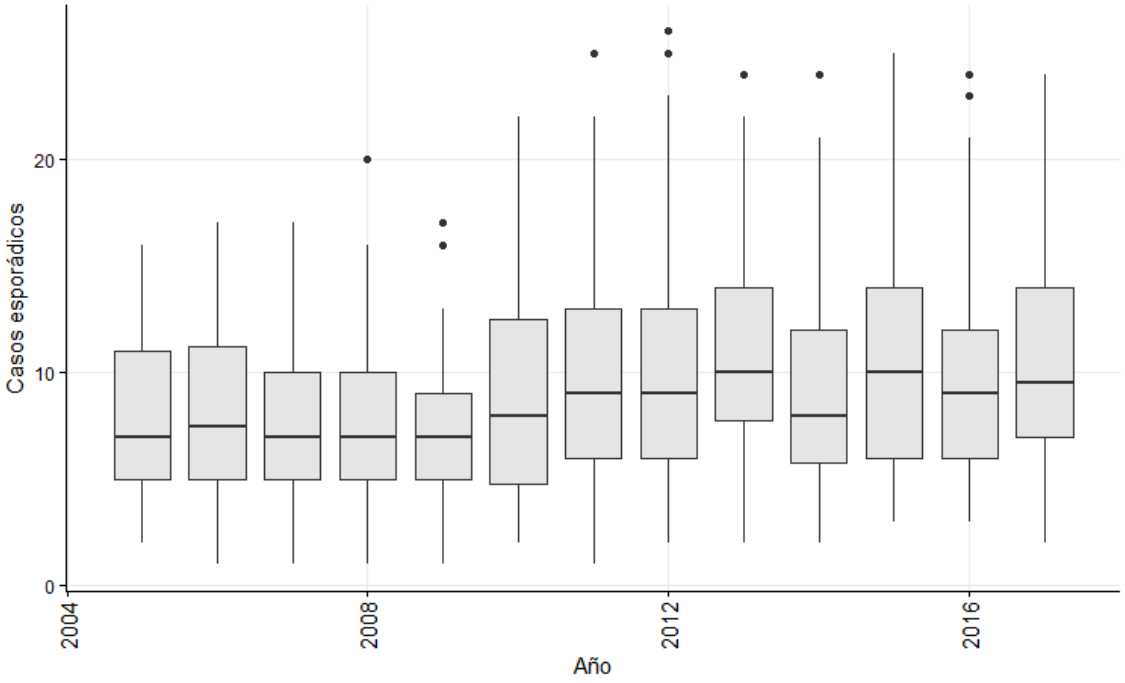

Figura 7. Diagrama de cajas del número semanal de casos esporádicos de Salmonella confirmados por el Departamento de Salud de Minnesota, 2005 – 2017. Cada caja representa la distribución del número de casos de salmonelosis semanales producidos del año correspondiente (eje x).

Las deducciones realizadas sobre la estacionariedad en varianza de la serie parecen firmes y por ello realizó una transformación Box-Cox antes de proceder a modelizar todos los modelos salvo el NNAR.

En el análisis de regresión lineal el tiempo en años resultó significativo indicando que por cada año el número de casos semanales de *Salmonella* aumenta en 0,275 casos (p valor < 0,001); además, las semanas centrales del año (de la 21 a la 42) son significativamente diferentes de la primera semana del año referencia (p valor < 0,05) establecida como de referencia señal de clara estacionalidad. Igualmente, la descomposición multiplicativa de la figura 8, muestra claramente la preminencia del componente estacional en la serie, así como una ligera tendencia alcista. La representación gráfica de la ACF (figura 9) también indica la presencia de componente estacional y la necesidad de algún tipo de diferenciación. Por ello se aplicó una diferenciación estacional a la serie para la modelización mediante SARIMA y primeras diferencias para la modelización mediante regresión armónica dinámica. En los apartados correspondientes se evaluó de nuevo la estacionariedad de las series tras las diferenciaciones indicadas.

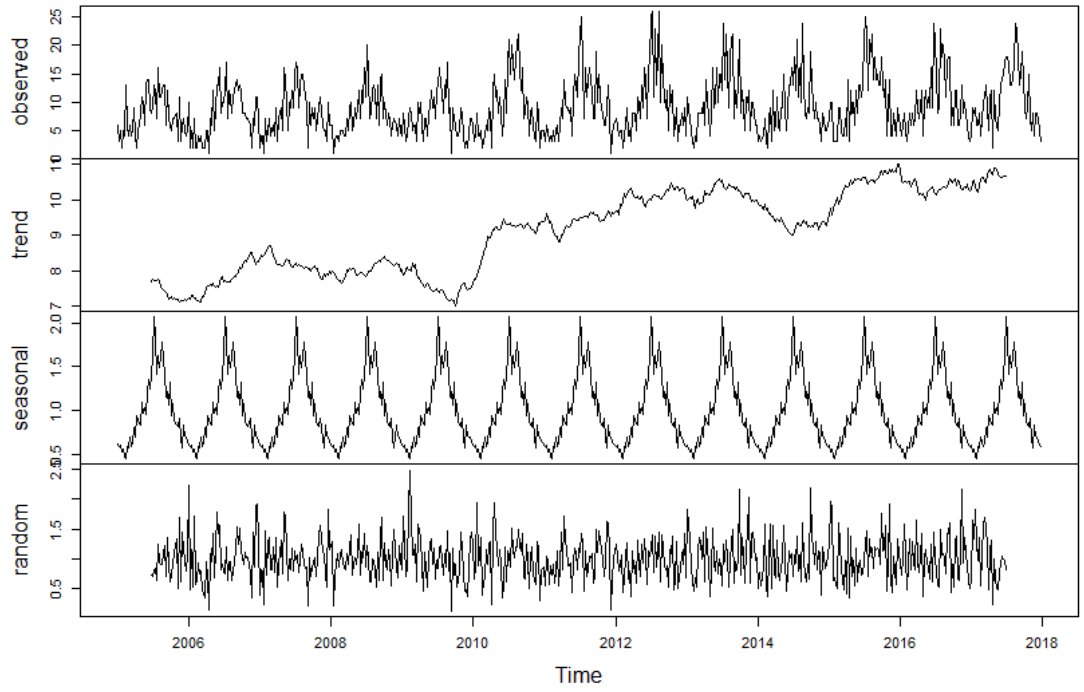

Figura 8. Componentes de la serie temporal de casos semanales de Salmonella esporádicos confirmados por el Departamento de Salud de Minnesota, 2005 – 2017. Las unidades del eje y son los casos esporádicos / semana; las unidades de este eje son diferentes por componente. En orden descendente, serie observada, tendencia de la serie, componente estacional y error aleatorio.

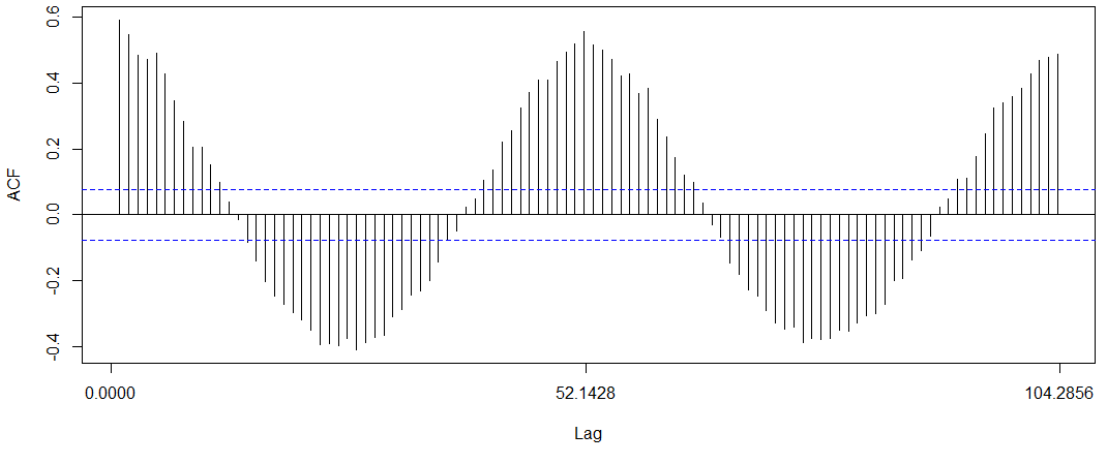

Figura 9. Representación gráfica de la función de autocorrelación (ACF) de la serie temporal de casos semanales de Salmonella esporádicos confirmados por el Departamento de Salud de Minnesota, 2005 – 2017.

4.2 Transformación Box-Cox

El análisis exploratorio reflejó la necesidad de una transformación Box-Cox para estabilizar la varianza y lograr su estacionariedad, que se aplicó con un parámetro lambda de 0,3233895 obtenido a través de la función *boxcoxfit*. En la figura 10 se muestran la

serie original y la transformada, observándose cómo en esta última la varianza es más estable ya que la diferencia entre los picos y los valles es más homogénea a lo largo de todo el periodo, en contraste con la original, donde hay una heterogeneidad más exacerbada.

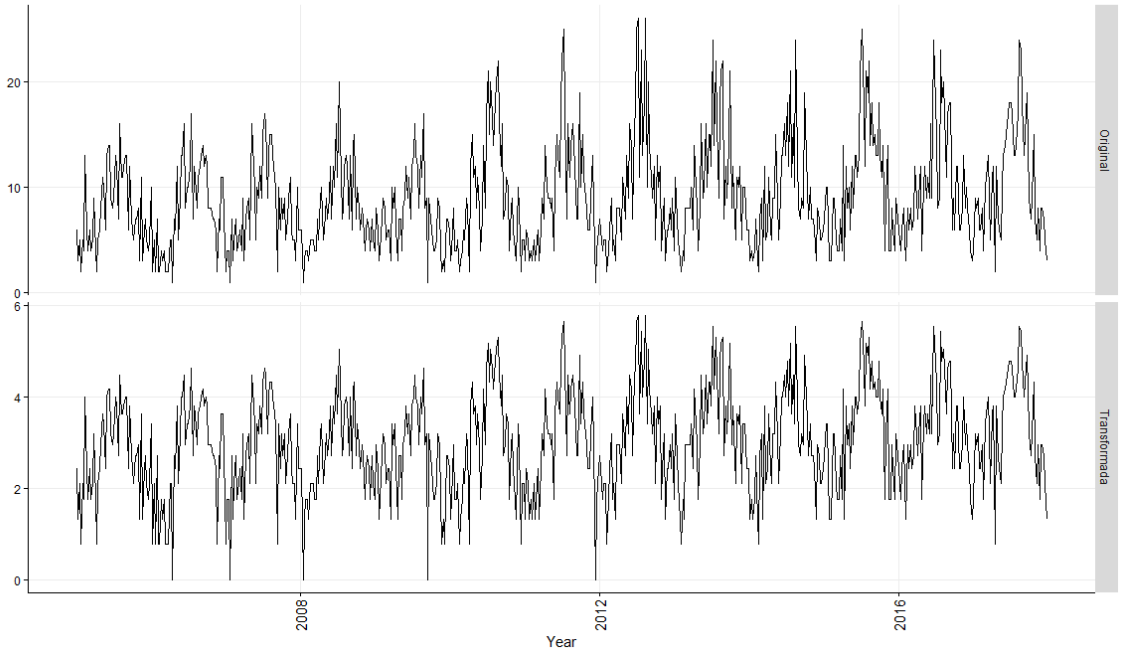

Figura 10. Serie original (arriba) y transformada (debajo) mediante el método Box-Cox con una lambda = 0,3233895, del número semanal de casos esporádicos de Salmonella confirmados por el Departamento de Salud de Minnesota, 2005 – 2017.

4.3 Modelización de los casos semanales esporádicos de salmonelosis

En este apartado se muestran los mejores modelos obtenidos a través de las diferentes técnicas estudiadas.

4.3.1 SARIMA

Diferenciaciones.

La ACF de la serie transformada, como cabría esperar, es similar a la de la serie no transformada, como puede observarse comparando las figuras 11 (serie transformada) y 10 (serie original). En base a lo observado en el análisis exploratorio y a los resultados de los test ADF y KPSS, se decidió realizar una diferenciación estacional y es que, aunque

el test ADF con un p valor menor que 0.01 llevó a aceptar la hipótesis alternativa de que la serie es estacionaria, el test KPSS con un p valor de 0.01 llevó a aceptar la hipótesis alternativa de que no lo es.

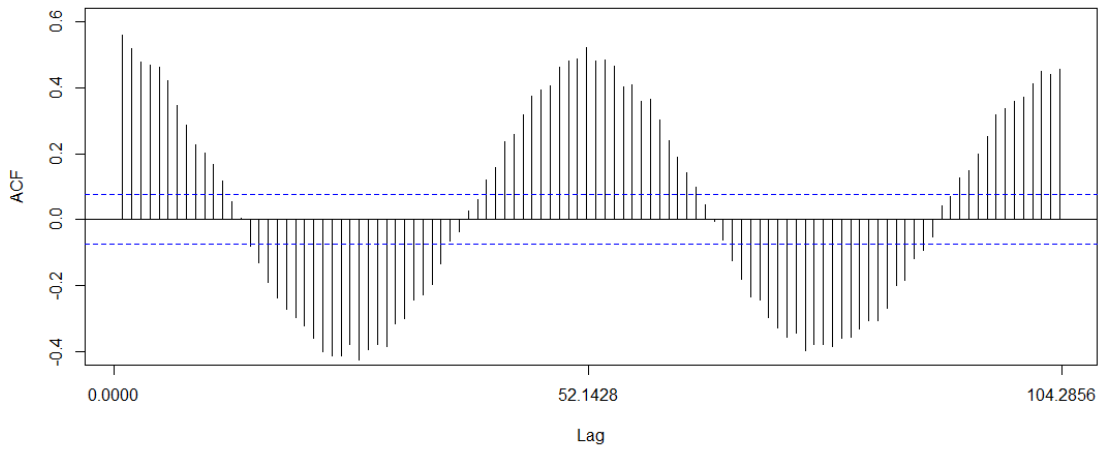

Figura 11. Representación gráfica de la función de autocorrelación (ACF), de la serie temporal transformada mediante el método Box-Cox con una lambda = 0,3233895 de los casos semanales de Salmonella esporádicos confirmados por el Departamento de Salud de Minnesota, 2005 – 2017.

El ADF y el KPSS, tras la aplicación de la diferenciación estacional, apuntan en la misma dirección, al aceptar a partir del ADF (p valor < 0,01) que la serie es estacionaria y no rechazar que esta lo fuese a partir del KPSS (p valor > 0,1). El ACF mostrado en la figura 12 también apunta hacia la estacionariedad de la serie al observarse como la única autocorrelación significativa es la del primer retardo (en la parte no estacional). Tampoco se aprecia existencia de sobrediferenciación al no tener el primer retardo una autocorrelación negativa menor de -0,5. En definitiva, la diferenciación aplicada parece correcta y suficiente.

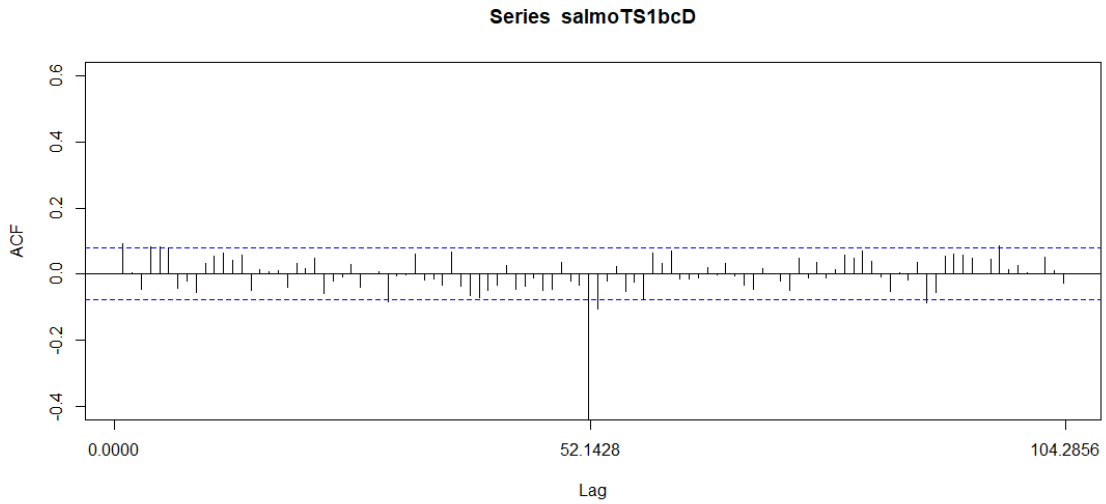

Figura 12. Representación gráfica de la función de autocorrelación (ACF), de la serie temporal de casos semanales de Salmonella esporádicos confirmados por el Departamento de Salud de Minnesota, 2005 – 2017, transformada mediante el método Box-Cox con una lambda = 0,3233895 y con una diferenciación estacional.

Selección de parámetros AR y MA.

Analizando de nuevo la representación gráfica de la ACF (figura 12) y también de la PACF de la serie estacionaria (figura 13) se observó que en ninguna de las partes no estacionales se producen los fenómenos habituales de las series puramente AR caracterizadas por un descenso gradual en el ACF y una caída brusca en la PACF tras unas primeras correlaciones elevadas, ni tampoco las tendencias habituales de las series puramente MA, en las que el PACF desciende gradualmente y la ACF tiene una caída seca tras unas primeras correlaciones intensas (53,54). Debido a la ausencia de alguno de estos patrones que permitiese definir de manera clara los parámetros más probables para la parte no estacional se decidió ajustar modelos con los parámetros AR y MA más habituales, los cuales en cualquier caso pueden concordar con los correlogramas dentro de la ambigüedad que ofrecen en esta ocasión. Por tanto, se valoraron modelos con parámetros en la parte no estacional: (1,0,1), (2,0,1), (3,0,1) y (1,0,2), (1,0,3), (1,0,0), (2,0,0) y (0,0,1), (0,0,2). Para la parte estacional, atendiendo a esa correlación negativa en el retardo 52 y al decaimiento progresivo de las autocorrelaciones en la PACF, se seleccionaron unos parámetros para la parte estacional (0,1,1). Se ajustaron modelos con todas las posibles combinaciones de los parámetros anteriores, con constante y sin constante. Según el AICc se seleccionó para la predicción el modelo con una estructura

(1,0,0) x (0,1,1)52 con constante. En la tabla 1 se puede visualizar el AICc de cada modelo.

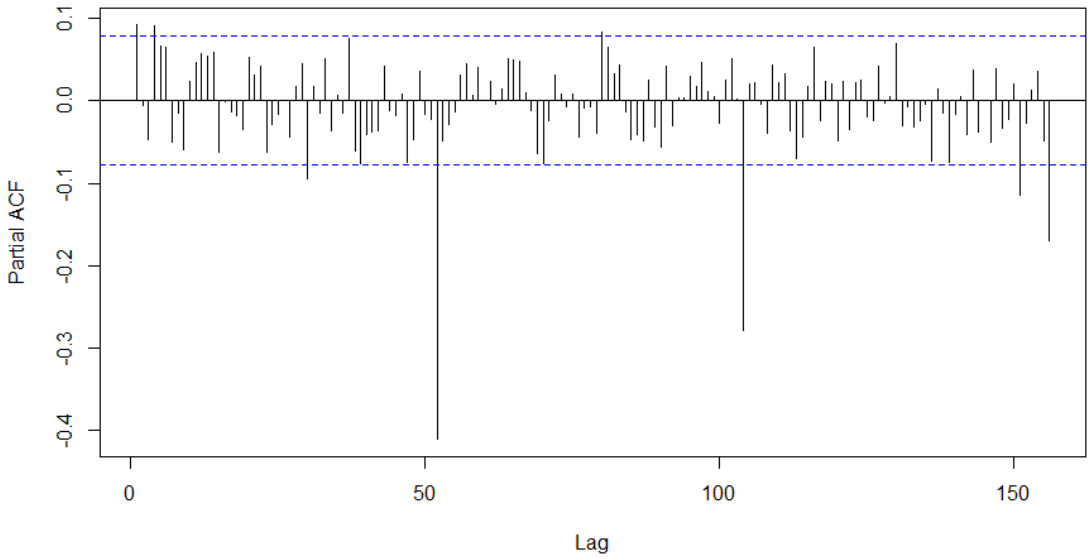

Figura 13. Representación gráfica de la función de autocorrelación parcial (PACF), de la serie temporal de casos semanales de Salmonella esporádicos confirmados por el Departamento de Salud de Minnesota, 2005 – 2017, transformada mediante el método Box-Cox con una lambda = 0,3233895 y con una diferenciación estacional.

Tabla 1. Modelos SARIMA ajustados a partir de la serie temporal transformada mediante Box-Cox (lambda = 0,3233895) de los casos esporádicos semanales de salmonelosis registrados por el Departamento de Salud de Minnesota en el periodo 2005-2017, con su AICc (en negrita el mejor modelo según este criterio.)

Parámetros	Constante (Si,No)	AICc
(3,0,1) x (0,1,1)	Si	1577.115
(3,0,1) x (0,1,1)	No	1602.787
(2,0,1) x (0,1,1)	Si	1574.861
(2,0,1) x (0,1,1)	No	1578.866
(1,0,1) x (0,1,1)	Si	1573.516
(1,0,1) x (0,1,1)	No	1580.791
(1,0,2) x (0,1,1)	Si	1575.069
(1,0,2) x (0,1,1)	No	1579.183
(1,0,3) x (0,1,1)	Si	1577.063
(1,0,3) x (0,1,1)	No	1580.371
(0,0,1) x (0,1,1)	Si	1572.197
(0,0,1) x (0,1,1)	No	1600.234
(0,0,2) x (0,1,1)	Si	1573.504
(0,0,2) x (0,1,1)	No	1599.793
(1,0,0) x (0,1,1)	**Si**	**1571.978**
(1,0,0) x (0,1,1)	No	1599.322
(2,0,0) x (0,1,1)	Si	1573.617
(2,0,0) x (0,1,1)	No	1599.405

Regresión dinámica con errores ARIMA.

Las series temporales de las variables predictoras que se quieren incluir en la regresión dinámica, temperatura media máxima semanal en grados Celsius y precipitaciones (pulgadas) medias por semana, así como la serie analizada de salmonelosis pueden visualizarse en la figura 14.

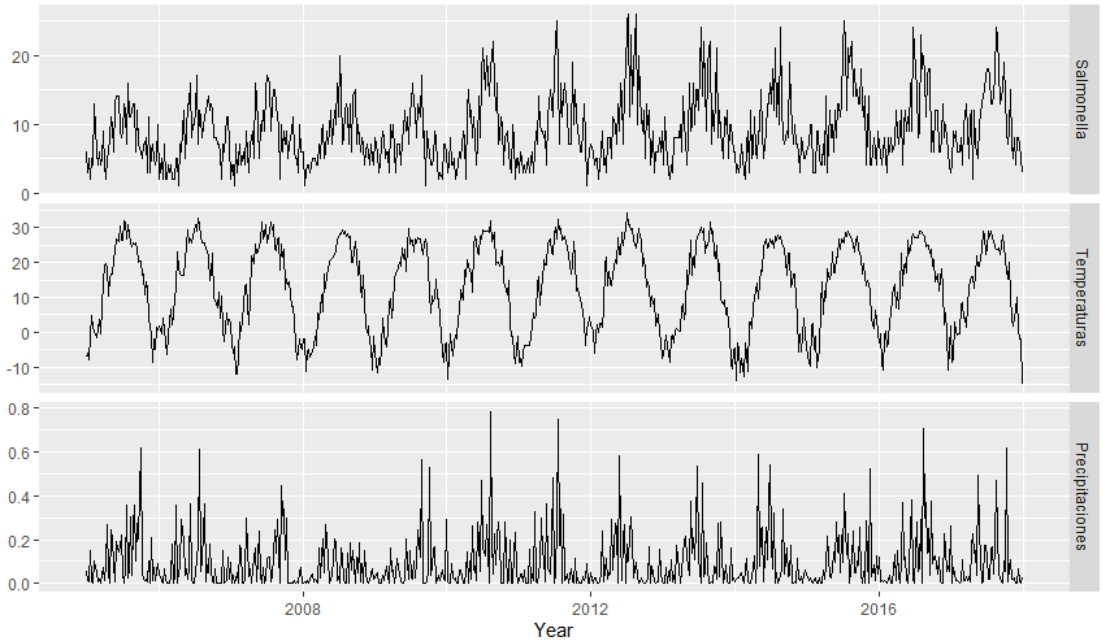

Figura 14. Serie temporal de los casos esporádicos de salmonelosis semanales registrados por el Departamento de Salud de Minnesota (arriba), precipitaciones medias semanales en pulgadas registradas por la estación meteorológica: 215838 (abajo) y temperatura media máxima en grados centígrados registrada por la estación meterológica número: 215838 (centro).

En la figura 15 se aprecia que la ACF de temperatura con una diferenciación estacional (realizada siguiendo las transformaciones hechas en la serie de casos) no deja clara la existencia de estacionariedad al no existir un decaimiento brusco en la significación de las autocorrelaciones, pero atendiendo al test ADF que aceptaba que la estacionariedad existía (p valor < 0,01) y el test KPSS que no rechazaba que esta existiese (p valor > 0,1) se consideró la serie estacionaria.

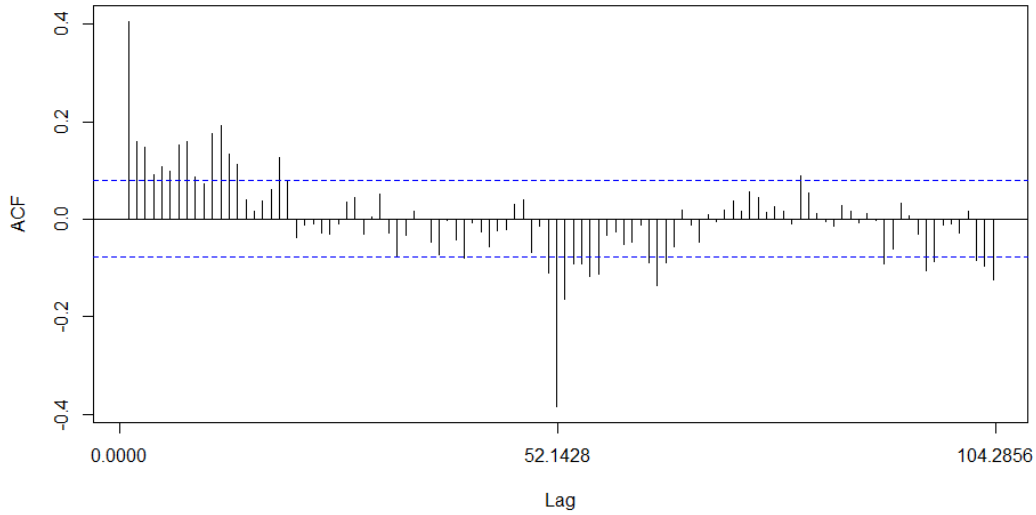

Figura 15. Representación gráfica de la función de autocorrelación (ACF), de la serie temporal de temperaturas medias máximas semanales en grados Celsius, 2005 – 2017, con una diferenciación estacional.

La ACF de la serie de precipitación sí indicó la estacionariedad de la serie como se puede comprobar en la figura 16, con una caída del ACF inmediata ya en el primer retardo. Los tests ADF ($p < 0,01$) y KPSS ($p > 0,1$) confirmaron la existencia de estacionariedad en de esta serie.

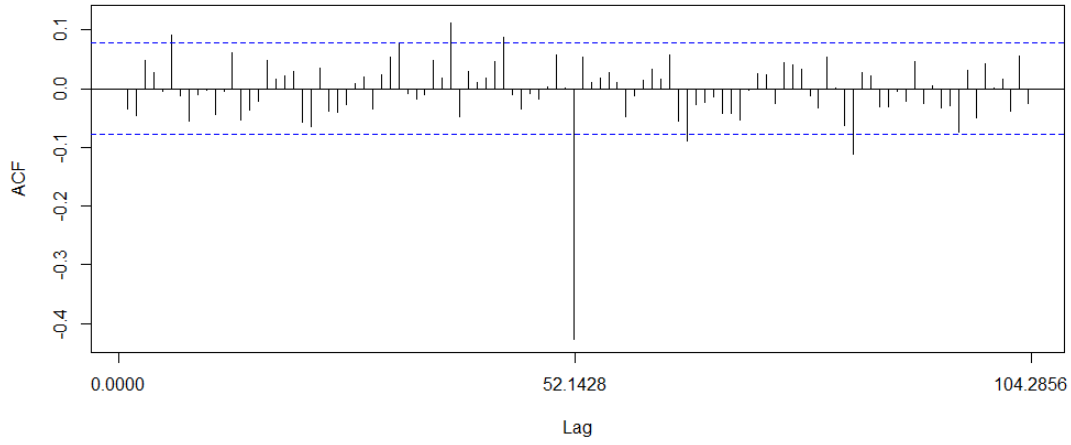

Figura 16. Representación gráfica de la función de autocorrelación (ACF), de la serie temporal de precipitaciones medias semanales en pulgadas, 2005 – 2017, con una diferenciación estacional.

El análisis de regresión lineal preliminar con precipitación y temperatura como variables predictoras para los casos esporádicos de salmonella semanales indicó que, mientras que la temperatura se asoció con el número semanal de casos ($p < 0,0001$), la precipitación no lo hacía ($p = 0,20$).

Por tanto, se fijaron modelos de regresiones dinámicas con la variable temperatura como predictora y con modelos SARIMA para los errores con los parámetros utilizados en el apartado de la modelización del propio SARIMA. El mejor modelo SARIMA para los errores de la regresión fue también el (1,0,0) x (0,1,1)52, resultando su AICc similar al del SARIMA (1571,934 frente a 1571,978) y resultando la variable temperatura no significativa (p = 0,149). Los valores AICc de las regresiones dinámicas ajustadas con los diferentes parámetros SARIMA comparados con los de los modelos SARIMA con constante correspondientes, pueden observarse en la tabla 2.

Al no ser el modelo obtenido a partir de la regresión dinámica mejor que el SARIMA no se consideró para la predicción final.

Tabla 2. Valores de AICc de los diferentes modelos obtenidos a través de las regresiones dinámicas con temperatura como variable predictora con diferentes parámetros SARIMA y AICc obtenido a partir de los SARIMA con los mismos parámetros.

Parámetros	SARIMA/Regresión dinámica	AICc
(3,0,1) x (0,1,1)	SARIMA	1577.115
(3,0,1) x (0,1,1)	Regresión dinámica	1576.886
(2,0,1) x (0,1,1)	SARIMA	1574.861
(2,0,1) x (0,1,1)	Regresión dinámica	1575.204
(1,0,1) x (0,1,1)	SARIMA	1573.516
(1,0,1) x (0,1,1)	Regresión dinámica	1573.744
(1,0,2) x (0,1,1)	SARIMA	1575.069
(1,0,2) x (0,1,1)	Regresión dinámica	1575.019
(1,0,3) x (0,1,1)	SARIMA	1577.063
(1,0,3) x (0,1,1)	Regresión dinámica	1576.994
(0,0,1) x (0,1,1)	SARIMA	1572.197
(0,0,1) x (0,1,1)	Regresión dinámica	1572.145
(0,0,2) x (0,1,1)	SARIMA	1573.504
(0,0,2) x (0,1,1)	Regresión dinámica	1573.505
(1,0,0) x (0,1,1)	**SARIMA**	**1571.978**
(1,0,0) x (0,1,1)	**Regresión dinámica**	**1571.934**
(2,0,0) x (0,1,1)	SARIMA	1573.617
(2,0,0) x (0,1,1)	Regresión dinámica	1573.639

Evaluación de los residuos.

Los residuos del modelo seleccionado (SARIMA (1,0,1) x (0,1,1)52 con constante) se ajustan bastante bien a una distribución normal y se centran en el valor 0 como se puede

observar en el histograma de la figura 17. También mantienen una estructura de ruido blanco, como puede apreciarse en la representación gráfica de la ACF de esa misma figura y como indica el test Ljung-Box, a través del cual se concluye que no se puede rechazar que los residuos se distribuyan de manera independiente (p = 0,491).

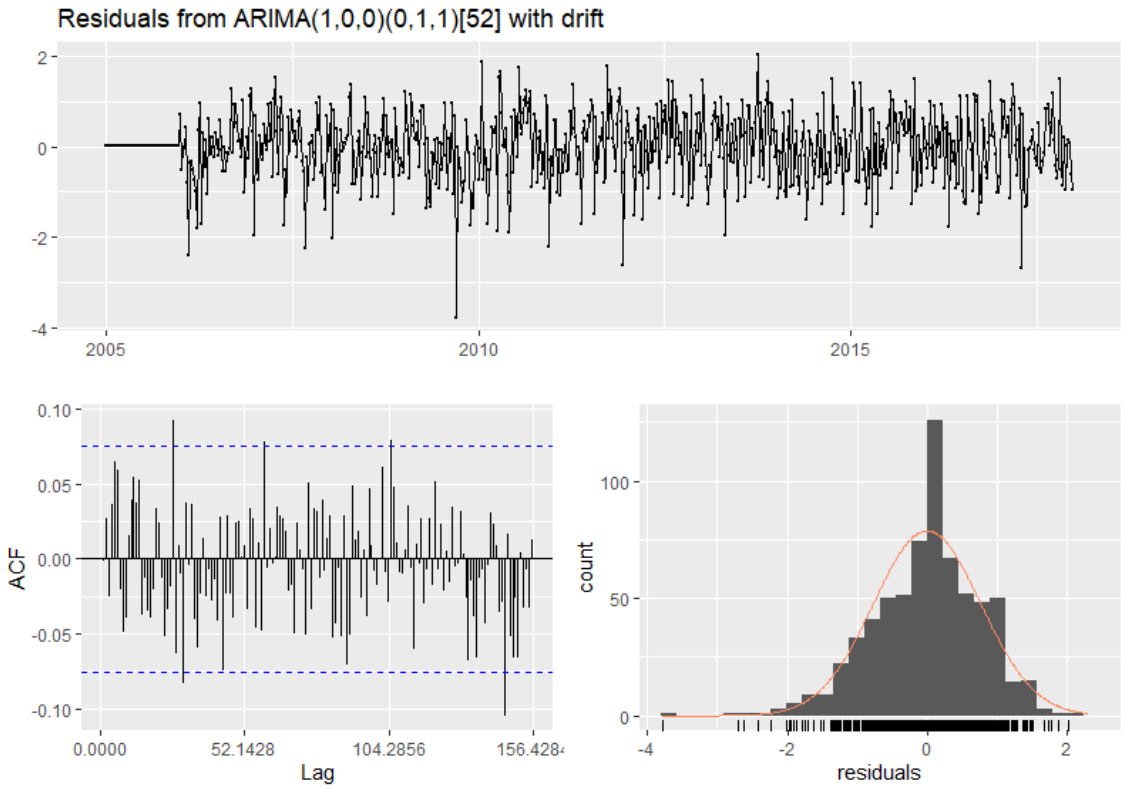

Figura 17. Análisis de los residuos del mejor modelo (SARIMA (1,0,0) x (0,1,1)52 con constante) para la serie temporal transformada mediante Box-Cox (lambda = 0.3233895) de los casos esporádicos semanales de salmonelosis registrados por el Departamento de Salud de Minnesota en el periodo 2005-2017. Serie temporal de los residuos (arriba), representación de la función de autocorrelación (esquina inferior izquierda) e histograma (esquina inferior derecha).

Finalmente, a través del test de McLeod, cuya representación gráfica se puede observar en la figura 18, se concluye que los residuos del modelo no siguen un proceso ARCH, por lo que sus intervalos de predicción se presumen apropiados.

Por tanto, el modelo cumple todos los supuestos para sus residuos y parece apropiado para predecir los casos esporádicos de salmonelosis. En la última parte de este apartado de resultados se comprobó la precisión de sus predicciones, así como su capacidad para la aplicación de la metodología que se pretende desarrollar para la detección de casos de salmonelosis.

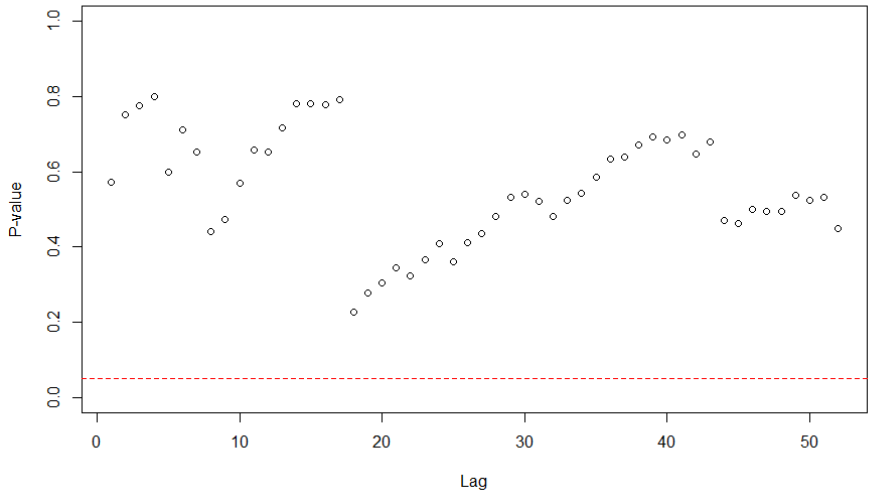

Figura 18. Test de McLeod de los residuos al cuadrado del mejor modelo (SARIMA (1,0,0) x (0,1,1)52 con constante) para la serie temporal transformada mediante Box-Cox (lambda = 0.3233895) de los casos esporádicos semanales de salmonelosis registrados por el Departamento de Salud de Minnesota en el periodo 2005-2017., SARIMA (1,0,0) x (0,1,1)52.

4.3.3 Regresión armónica dinámica

Diferenciaciones

Como ya se observó en la figura 11 es necesario realizar algún tipo de transformación para lograr la estacionariedad de la serie. En la regresión armónica la parte estacional va a ser modelizada a través de los términos de Fournier, por lo tanto, se probó a aplicar una diferenciación a la parte no estacional (conocida como primeras diferencias) para lograr la estacionariedad. Los tests ADF (p valor < 0,01) y KPSS (>0,1) apuntan a la estacionariedad de la serie al igual que la ACF, como puede observarse en la figura 19.

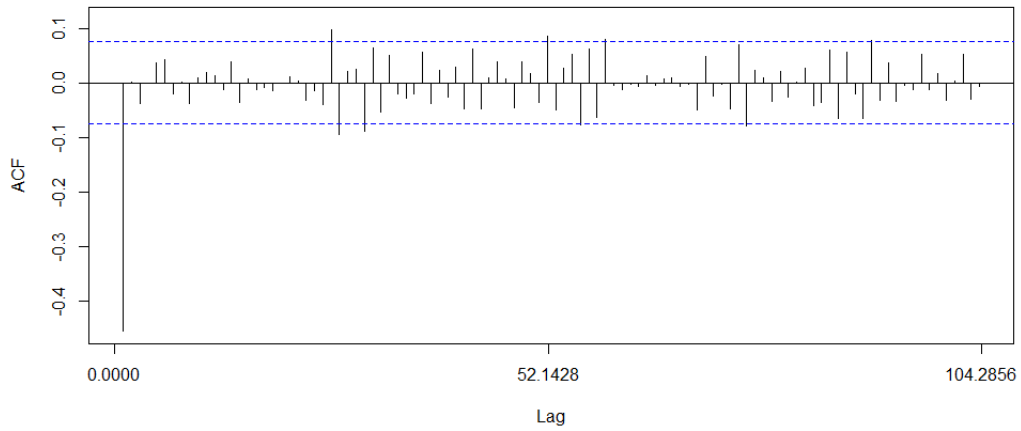

Figura 19. Representación gráfica de la función de autocorrelación (ACF), de la serie temporal de casos semanales de Salmonella esporádicos confirmados por el Departamento de Salud de Minnesota, 2005 – 2017, transformada mediante el método Box-Cox con una lambda = 0.3233895 y con una diferenciación aplicada a su parte no estacional.

Selección de términos de Fourier y parámetros AR y MA

Se seleccionaron los modelos SARIMA para los residuos más habituales en modelos con 10, 9 y 8 pares de senos y cosenos dados los resultados del AICc (tabla 3). Los modelos con los parámetros ARIMA más habituales (52,53) fueron ajustados para los pares de senos y cosenos seleccionados, ajustándose estos con y sin temperatura. El modelo final seleccionado en base a este criterio fue un modelo con 10 términos de Fournier y con un modelo ARIMA (0,1,1) para los errores; además, se observó como la temperatura no mejoraba el valor AICc de los modelos (tabla 4).

Tabla 3. Valores AICc obtenidos con la función auto.arima a partir de los términos de Fourier especificados, en negrita se muestra los términos con los cuales el AICc se homogenizó y cesó su descenso.

Términos de Fournier	AICc
26	1626.019
25	1622.657
24	1619.112
23	1614.981
22	1616.967
21	1613.359
20	1609.679
19	1615.982
18	1613.763
17	1610.156
16	1607.754
15	1605.437
14	1601.695
13	1602.704
12	1598.976
11	1597.633
10	**1594.282**
9	**1594.734**
8	**1593.348**
7	1594.832
6	1595.363

Tabla 4. Valores AICc obtenidos a partir de las combinaciones de términos de Fournier seleccionados y los parámetros ARIMA más habituales (en negrita se resalta la combinación con un menor AICc)

		Parámetros ARIMA								
		(1,1,1)	(2,1,1)	(3,1,1)	(1,1,2)	(1,1,3)	(1,1,0)	(2,1,0)	**(0,1,1)**	(0,1,2)
Terminos de Fournier (k) y temperatura (T)	**k = 10 sin T**	1586.094	1588.069	1590.129	1587.649	1590.106	1833.841	1769.201	**1585.385**	1586.144
	k = 9 sin T	1588.221	1590.311	1592.191	1590.332	1592.044	1838.443	1775.000	1587.728	1588.256
	k = 8 sin T	1586.766	1588.873	1590.621	1588.884	1590.563	1838.040	1775.411	1586.394	1586.795
	k = 10 con T	1586.475	1588.472	1590.510	1588.269	1590.487	1834.450	1770.220	1585.831	1586.526
	k = 9 con T	1588.909	1591.018	1592.853	1839.492	1592.751	1839.492	1776.419	1588.485	1588.944
	k = 8 con T	1587.248	1589.368	1591.068	1589.376	1591.063	1838.807	1776.540	1586.947	1587.275

Residuos del modelo

Los residuos del modelo se ajustan bastante bien a una distribución normal y se centran en el valor 0, como se puede observar en el histograma de la figura 20. También mantienen una estructura de ruido blanco como puede apreciarse en el gráfico de la ACF de esa misma figura y como indica el test Ljung-Box, a través del cual se concluye que no se puede rechazar que los residuos se distribuyan de manera independiente (p = 0,469).

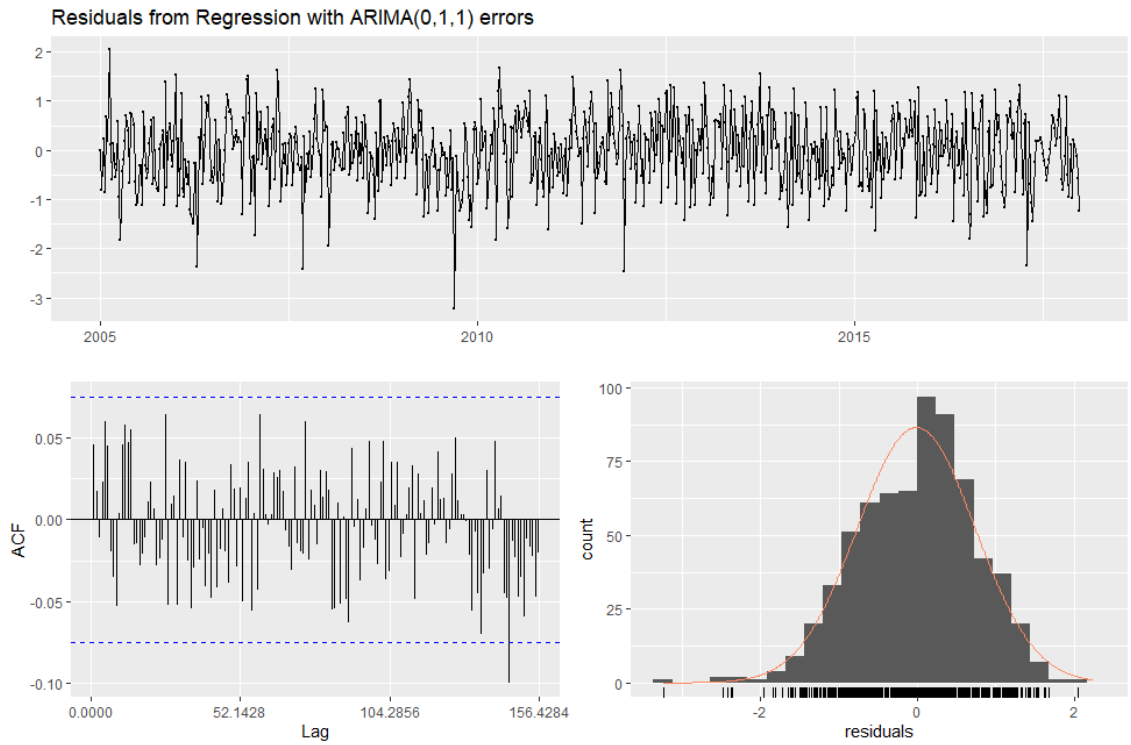

Figura 20. Análisis de los residuos del mejor modelo de regresión armónica dinámica (10 términos de Fournier y ARIMA (0,1,1) para los errores) para la serie temporal transformada mediante Box-Cox (lambda = 0.3233895) de los casos esporádicos semanales de salmonelosis registrados por el Departamento de Salud de Minnesota en el periodo 2005-2017. Serie temporal de los residuos (arriba), representación de la función de autocorrelación (esquina inferior izquierda) e histograma (esquina inferior derecha).

Finalmente, a través del test de McLeod, cuya representación gráfica se puede observar en la figura 21, se concluye que los residuos del modelo no siguen un proceso ARCH, por lo que los intervalos de confianza de las predicciones serán apropiados.

En base a todo lo anterior los residuos cumplen con todos los supuestos y por tanto la regresión dinámica armónica con 10 términos de Fournier y un ARIMA (0,1,1) para los residuos parece apropiada para predecir los casos esporádicos de salmonelosis.

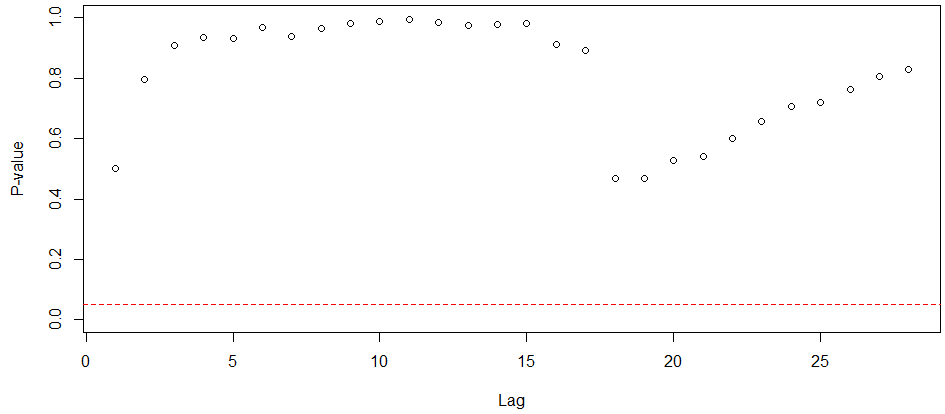

Figura 21. Test de McLeod de los residuos al cuadrado del mejor modelo de regresión dinámica armónica (10 términos de Fournier y ARIMA (0,1,1) para los errores) para la serie temporal transformada mediante Box-Cox (lambda = 0,3233895) de los casos esporádicos semanales de salmonelosis registrados por el Departamento de Salud de Minnesota en el periodo 2005-2017.

4.3.3 NNAR

El modelo NNAR seleccionado por la función nnetar cuando no se incluyó ninguna covariable tenía parámetros (14,1,8), es decir, con 14 retardos no estacionales, 1 retardo estacional como inputs y 8 nodos en la capa oculta. Cuando se introdujeron variables externas los parámetros del modelo NNAR óptimos fueron los mismos (14,1,8) y resultó el mejor modelo el que incluía precipitación y temperatura, con un RMSE de 1,506 frente a un RMSE de 1,853 del modelo sin variables externas. Los modelos con una sola variable mejoraron en RMSE al modelo sin variables externas, siendo el RMSE del modelo con solo lluvia de 1,784 y de 1,717 el de solo temperatura, pero no lo hicieron frente al modelo con ambas variables.

Residuos del modelo.

Los residuos parecen normalmente distribuidos y centrados en el valor 0, señal de que el modelo no sufre un sobreajuste fuerte (figura 22).

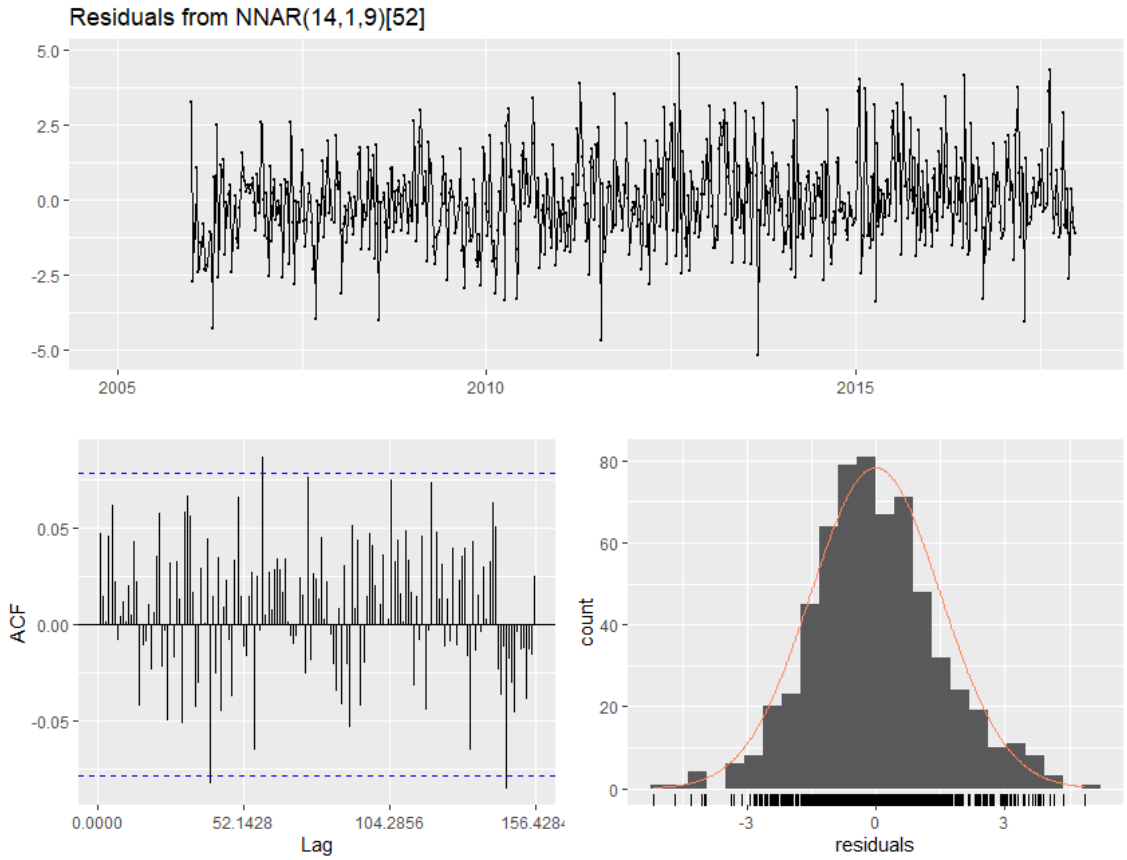

Figura 22. Análisis de los residuos del mejor modelo, NNAR (parámetros (14,1,9)52 con temperatura y precipitación como inputs) para la serie temporal transformada mediante Box-Cox (lambda = 0.3233895) de los casos esporádicos semanales de salmonelosis registrados por el Departamento de Salud de Minnesota en el periodo 2005-2017. Serie temporal de los residuos (arriba), representación de la función de autocorrelación (esquina inferior izquierda) e histograma (esquina inferior derecha).

4.4 Predicción de los casos semanales esporádicos de salmonelosis.

Las predicciones para el año 2018 realizadas a partir de los tres modelos seleccionados, SARIMA (1,0,0) x (0,1,1)52, regresión dinámica armónica con 10 términos de Fournier y con un modelo ARIMA (0,1,1) para los errores y NNAR (14,1,9)52 con temperatura y precipitaciones pueden verse representadas gráficamente en la figura 23.

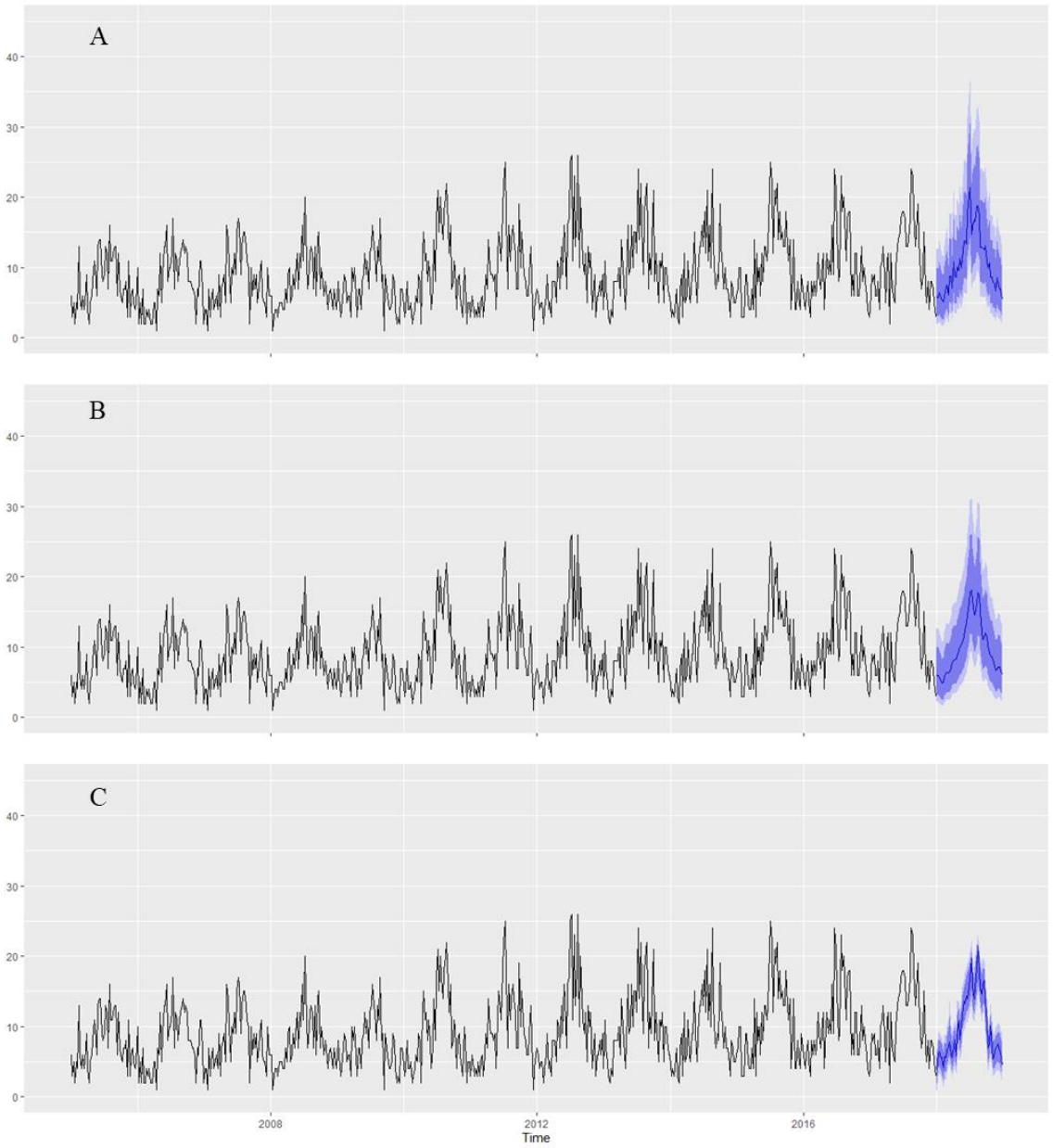

Figura 23. Predicciones de casos esporádicos de salmonelosis para el año 2018 (curva azul) con los modelos SARIMA (1,0,0) x (0,1,1)52 **A,** regresión dinámica armónica con 10 términos de Fournier y con errores ARIMA (0,1,1) **B** y NNAR (14,1,9)52 con temperatura y precipitaciones **C** y sus intervalos de confianza al 80% (azul oscuro) y al 95% (azul grisáceo). En el caso del modelo NNAR los intervalos de la predicción fueron generados mediante simulación. El eje y representa el número de casos de Salmonella esporádicos semanales.

Aunque las predicciones de los tres modelos son de una calidad aceptable en base a su error absoluto medio (MAE) y a la raíz del error cuadrático medio (RMSE) (tabla 5), el mejor modelo según ambos parámetros es la regresión dinámica armónica. En cualquier caso, estas diferencias entre los tres modelos son mínimas y cualquiera de ellos podría

considerarse válido para la predicción de los casos esporádicos semanales de salmonelosis.

Tabla 5. Error absoluto medio (MAE) y raíz del error cuadrático medio (RMSE) para las predicciones de casos semanales de salmonelosis del año 2018 obtenidas a partir del mejor modelo seleccionado para cada técnica utilizada (SARIMA, regresión armónica dinámica y NNAR), (en negrita el mejor resultado obtenido).

	Error absoluto medio **(MAE)**	Raíz del error cuadrático medio **(RMSE)**
SARIMA (1,0,0) x (0,1,1)52	2.904	4.923
Regresión dinámica armónica k = 10, ARIMA (0,1,1)	**2.740**	**4.388**
NNAR (14,1,9)52 con temperatura y lluvia	3.510	4.821

4.5 Detección de brotes

Con base en las predicciones del modelo SARIMA y de acuerdo con el criterio utilizado por defecto, de modo que únicamente es necesario que la predicción puntual sea mayor que el número de casos observados para definir la presencia de un brote, la sensibilidad y especificidad obtenidas son del 96,67% y del 27,27%. Las predicciones puntuales de la regresión dinámica armónica permiten obtener una sensibilidad del 100% y una especificidad del 22,7%, con las del modelo NNAR la sensibilidad y la especificidad resultan en un 96,67% y un 27,27% respectivamente, resultados similares a los obtenidos con el SARIMA (tabla 6).

Las curvas ROC obtenidas a partir de la nueva variable generada "diferencia entre el número de casos totales registrados y la predicción puntual de casos esporádicos de cada semana" se pueden visualizar en la figura 24. La curva ROC de la regresión armónica fue la que ofreció una mayor área bajo la curva, como puede visualizarse en esa misma figura. En cualquiera de los casos, todas las curvas ROC muestran que las diferencias entre los casos totales y los esporádicos son aceptables para la clasificación de brotes siendo todas las áreas superiores a 0,5, así como sus intervalos de confianza.

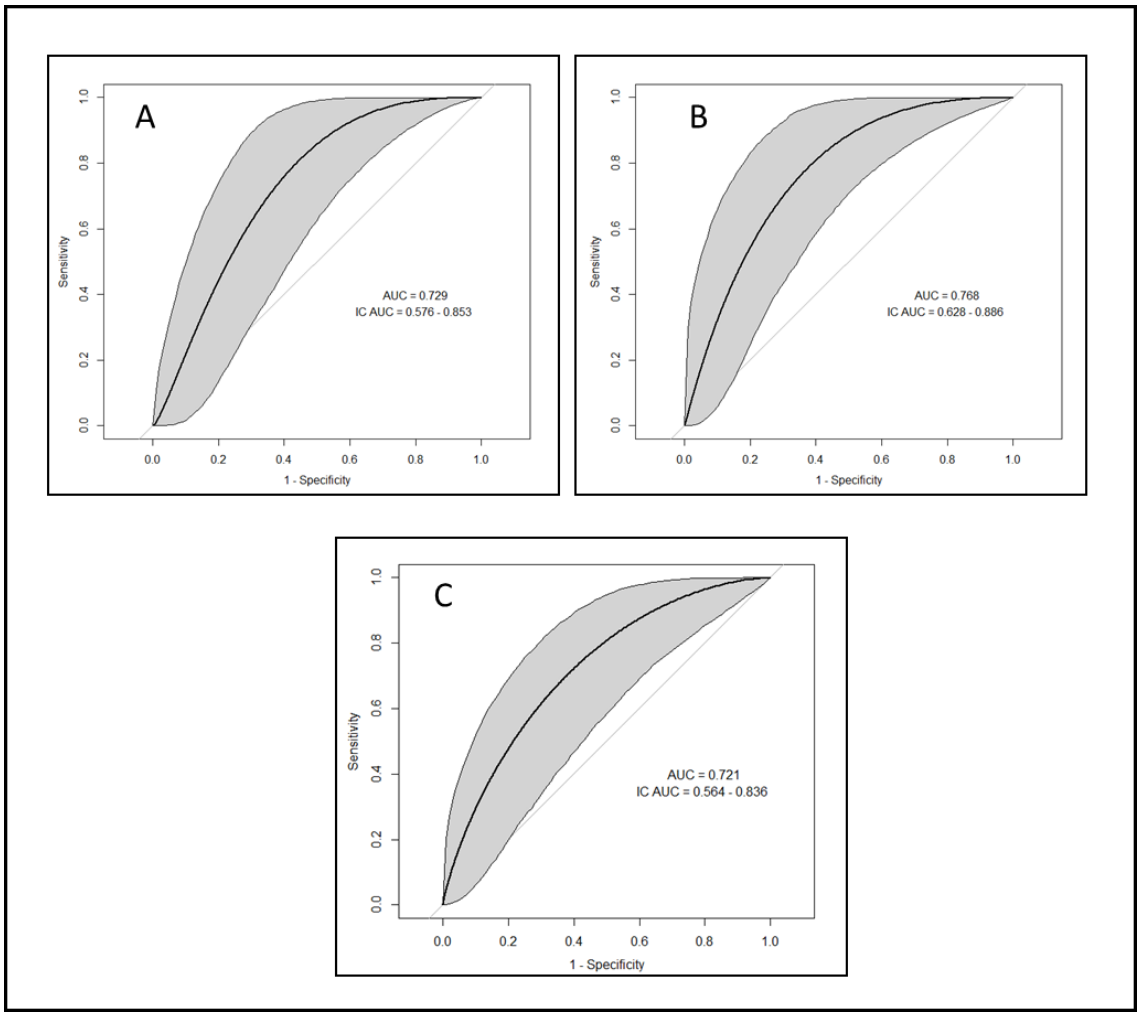

Figura 24. Curvas ROC obtenidas a partir de la variable diferencia entre casos totales registrados y casos esporádicos predichos a partir de los modelos SARIMA (1,0,1) x (0,1,1) 52 [**A**], Regresión armónica dinámica con errores ARIMA (0,1,1) [**B**] y NNAR (14,1,9) 52 [**C**], con la banda de confianza al 95% en gris, los intervalos de confianza del AUC también han sido calculados para una confianza del 95%.

El mejor punto de corte para maximizar la especificidad y la sensibilidad de las diferencias obtenidas a partir del SARIMA es de 2.29 casos totales más que esporádicos predichos (sensibilidad = 83,33%; especificidad =63,63%). En el caso de la regresión armónica el mejor punto de corte está en 3,89 casos totales más que esporádicos predichos (sensibilidad = 76,67%; especificidad = 72,73%), mientras que en el modelo NNAR ese punto de corte es de 1,58 casos totales más que esporádicos predichos (sensibilidad = 93,33%; especificidad = 40,90%). Todos estos resultados pueden visualizarse en la tabla 6. Aunque los resultados obtenidos a partir de la regresión dinámica armónica parecen los mejores, con cualquiera de los otros dos métodos también se obtienen unos resultados que indican una aceptable capacidad para detectar brotes.

Tabla 6. Sensibilidad (Se) y especificidad (Sp) cuando el criterio para definir una semana como con brote activo es que la predicción puntual de casos totales observados sea superior a la predicción puntual de casos esporádicos, es decir, cuando el punto de corte está en cero. Sensibilidad (Se*) y Especificidad (Sp*) cuando el punto de corte (PC) de corte de casos totales observados más que esporádicos predichos fue fijado mediante el criterio de Youden (en gris todo lo relativo a la fijación del nuevo punto de corte).

	Se	Sp	PC	Se*	Sp*
SARIMA (1,0,0) x (0,1,1)52	96.67	27.27	2.29	83.33	63.63
Regresión dinámica armónica k = 10, ARIMA (0,1,1)	100	22.7	3.89	76.67	72.73
NNAR (14,1,9)52 con temperatura y lluvia	96.67	27.27	1.58	93.33	40.90

5. DISCUSIÓN

Los resultados obtenidos en este trabajo indican que tanto los modelos SARIMA, las regresiones dinámicas armónicas como los modelos NNAR pueden ser útiles para predecir los casos de salmonelosis esporádicos futuros, al menos a un año vista. El hecho de que las predicciones de la regresión armónica dinámica superen a las obtenidas a partir del SARIMA concuerda con lo expresado por algunos autores en relación con la mayor idoneidad de las regresiones dinámicas armónicas para la modelización de series temporales con frecuencias altas como la semanal (29,34). Las predicciones obtenidas en este trabajo apuntan en la línea de otras comparaciones desarrolladas entre modelos ARIMA y NNAR, en las cuales no se aprecian grandes diferencias entre las predicciones de ambos modelos o al menos no se puede determinar que un tipo de modelo siempre obtenga mejores resultados que el otro al modelizar distintas series temporales (59), si bien es cierto que hay otros estudios que muestran una superioridad de las predicciones realizadas a partir de modelos NNAR frente a las de los SARIMA (60) . Los modelos NNAR han demostrado ser altamente efectivos para la predicción, por lo que no es descartable que con un proceso de ajuste más exhaustivo que el llevado a cabo en este trabajo se puedan mejorar las predicciones obtenidas a partir de esta técnica. Una de las posibles ampliaciones de este trabajo podría ser la utilización de otras técnicas de *machine learning*, así como la optimización del ajuste del modelo NNAR, para la predicción de los casos esporádicos semanales de salmonelosis.

El hecho de que la temperatura media máxima no ayude a predecir el número de casos de salmonelosis al incluirse tanto en la regresión dinámica con errores SARIMA como en la regresión armónica dinámica podría resultar a priori extraño en base a la literatura vigente (3) y a lo observado en este trabajo, donde se aprecia como el mayor número de casos de salmonelosis se presentan siempre en los periodos de más calor, correspondientes al verano. Sin embargo, hay que tener en cuenta que la función Arima del paquete forecast tiene en cuenta (al ajustar la regresión dinámica con errores ARIMA) la estructura de errores SARIMA a la vez que modeliza la regresión; al explicarse el efecto de la temperatura por el componente estacional del SARIMA se obtiene que la temperatura no ayuda a predecir mejor los casos de salmonelosis. En la regresión armónica dinámica ocurre lo mismo, al explicarse el efecto de la temperatura a través de la inclusión de los términos de Fournier.

La metodología propuesta resulta prometedora para la detección de brotes, ya que las curvas ROC obtenidas a partir de las comparación de los casos totales observados con las predicciones puntuales de los casos semanales esporádicos de salmonelosis obtenidas a partir de los tres tipos de técnicas estudiadas son todas, además de bastante similares, lo suficientemente buenas como para permitir discriminar hasta cierto punto los incrementos de casos sugestivos de la presencia de brotes, como también queda de manifiesto por las sensibilidades y especificidades obtenidas. El método parece válido para complementar a los sistemas de biovigilancia pudiendo facilitar la detección temprana de brotes, especialmente en invierno cuando existe una incertidumbre real sobre la presencia o no de brotes, habiendo en la estación de verano prácticamente permanentemente brotes activos (3), por otra parte se observa como las sensibilidades obtenidas a partir de las predicciones de los modelos NNAR y SARIMA son bastante buenas mientras que las especificidades no son las óptimas, sin embargo esto realmente no es un gran inconveniente puesto que en la gestión sanitaria prima el principio de precaución y la existencia de falsos positivos, a priori, es más aceptable que la de falsos negativos. Para su aplicación práctica sería recomendable reajustar los modelos con las últimas observaciones registradas ya que de otra manera es probable que la precisión de las predicciones se reduzca con el paso del tiempo; en cualquier caso, no sería recomendable predecir a más de un ciclo estacional vista, es decir, a más de un año vista.

Una de las posibles limitaciones de este estudio es que los modelos utilizados se basan en distribuciones gaussianas mientras que los casos de salmonella se distribuyen mediante una Poisson, al ser datos provenientes de un conteo. Los modelos generalizados autorregresivos de media móvil (GARMA) y su extensión para manejar datos no estacionarios y estacionales, los modelos generalizados estacionales autorregresivos integrados de media móvil (GSARIMA) basados en distribuciones de Poisson o binomiales negativas y ajustados mediante estadística bayesiana (61), podrían ser una opción más apropiada y su aplicación a series temporales de salmonelosis podría ser interesante para una futura ampliación de este trabajo. Otras posibles limitaciones de este trabajo es la no consideración del factor espacial, así como del vinculo epidemiológico, los cuales podrían tenerse en cuenta a través de modelos jerárquicos bayesianos (62).

6. CONCLUSIONES

Los resultados obtenidos en este trabajo permiten obtener las siguientes conclusiones:

- Los modelos SARIMA, NNAR y las regresiones armónicas dinámicas son capaces de modelizar correctamente la distribución temporal de los casos esporádicos semanales de salmonelosis y generar predicciones aceptables al menos a un año vista.
- El método desarrollado centrado en la comparación de las predicciones puntuales semanales de casos esporádicos de salmonelosis con los casos totales observados, es decir, el exceso de casos reales respecto a las predicciones puntuales de casos esporádicos permite detectar semanas en las que hay brotes de salmonelosis de manera activa con una sensibilidad elevada y una especificidad aceptable.

7. BIBLIOGRAFÍA

(1) Salmonella EFSA explains zoonotic diseases. EFSA 2014.

(2) Majowicz SE, Musto J, Scallan E, Angulo FJ, O'Brien SJ, Jones T, et al. The Global Burden of Nontyphoidal Salmonella Gastroenteritis. 2010.

(3) Bad Bug Book, Foodborne Pathogenic

Microorganisms and Natural Toxins. Washington, D.C.: U.S. Food & Drug Administration, Center for Food Safety & Applied Nutrition; 2012.

(4) Ficha técnica: Salmonelosis. ANMAT.

(5) National Salmonella Surveillance Overview. CDC 2011.

(6) European Food Safety Authority, European Centre for Disease Prevention and Control. The European Union One Health 2019 Zoonoses Report. EFSA journal 2021;19(2):e06406-n/a.

(7) Martínez-Avilés M, Garrido-Estepa M, Álvarez J, de la Torre A. Salmonella Surveillance Systems in Swine and Humans in Spain: A Review. Veterinary sciences 2019;6(1):20.

(8) Real Decreto 2210/1995

(9) Protocolo de vigilancia de salmonelosis (Salmonella Spp. distinta de S.Typhi y S.Paratyphi) de la red de vigilancia epidemiológica. Instituto de Salud Carlos III.

(10) Local Public Health Institute of Massachussetts. Types of Disease Surveillance .2021; cited 2021 Oct 10. Available at: http://www.masslocalinstitute.info/diseasesurveillance/diseasesurveillance4.html.

(11) Universidad Internacional de Valencia. Vigilancia epidemiológica en salud pública: definición y tipos. 2018; cited 2021 Oct 10. Available at: https://www.universidadviu.com/es/actualidad/nuestros-expertos/vigilancia-epidemiologica-en-salud-publica-definicion-y-tipos.

(12) Jamison DT, Breman JG, Measham AR, Alleyne G, Claeson M, Evans DB, et al. Disease Control Priorities in Developing Countries, Second Edition.: Washington, DC: World Bank and Oxford University Press; 2006.

(13) CDC. Foodborne Diseases Active Surveillance Network (FoodNet): FoodNet 2015 Surveillance Report (Final Data). CDC 2017.

(14) EFSA. Salmonellosis - Annual Epidemiological Report 2016. 2016.

(15) 1instituto de Salud Carlos, III. Resultados de la Vigilancia Epidemiológica de las enfermedades transmisibles. Informe anual. Años 2017-2018. Ministerio de Ciencia e Innovación 2020.

(16) European Food Safety Authority and European Centre for Disease Prevention and Control (EFSA and ECDC). The European Union One Health 2018 Zoonoses Report. EFSA journal 2019;17(12):e05926-n/a.

(17) Ministerio de Sanidad. CMBD. cited 2021 Oct 10. Available at: https://pestadistico.inteligenciadegestion.mscbs.es/publicoSNS/S/rae-cmbd.

(18) CDC. National Enteric Disease Surveillance: Salmonella Annual Report, 2016. CDC 2018.

(19) Lynch M, Marder EP, Zansky S, Tauxe R, Cronquist AB, Geissler AL, et al. Incidence and Trends of Infections with Pathogens Transmitted Commonly Through Food and the Effect of Increasing Use of Culture-Independent Diagnostic Tests on Surveillance — Foodborne Diseases Active Surveillance Network, 10 U.S. Sites, 2013–2016. MMWR. Morbidity and mortality weekly report 2017 Apr 21,;66(15):397-403.

(20) CDC. Surveillance for Foodborne Disease Outbreaks, U.S., 2016: Annual Report; 2018 ASI 4205-59;CS295288-A. 2018.

(21) ECDC. Salmonellosis - Just the tip of the iceberg. 2015.

(22) Mølbak K, Simonsen J, Jørgensen CS, Krogfelt KA, Falkenhorst G, Ethelberg S, et al. Seroincidence of Human Infections With Nontyphoid Salmonella Compared With Data From Public Health Surveillance and Food Animals in 13 European Countries. Clinical infectious diseases 2014;59(11):1599-1606.

(23) Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson M, Roy SL, et al. Foodborne illness acquired in the United States: major pathogens. Emerging infectious diseases 2011;17(1):7-15.

(24) Organization WH. Foodborne disease outbreaks : guidelines for investigation and control. Geneva: World Health Organization; 2008.

(25) Unkel S, Farrington CP, Garthwaite PH, Robertson C, Andrews N. Statistical methods for the prospective detection of infectious disease outbreaks: a review. Journal of the Royal Statistical Society. Series A, Statistics in society 2012;175(1):49-82.

(26) Chatfield C, Xing H. The Analysis of Time Series An Introduction with R. : CRC Press; 2019.

(27) Ward MP, Iglesias RM, Brookes VJ. Autoregressive Models Applied to Time-Series Data in Veterinary Science. Frontiers in veterinary science 2020;7:604.

(28) Moser EB. Repeated Measures Modeling With PROC MIXED. Louisiana State University.

(29) Hyndman RJ, Athanasopoulos G. Forecasting: Principles and Practice. Second print edition ed. Lexington, Ky: Otexts; 2018.

(30) Wang Y, Xu C, Li Y, Wu W, Gui L, Ren J, et al. An Advanced Data-Driven Hybrid Model of SARIMA-NNNAR for Tuberculosis Incidence Time Series Forecasting in Qinghai Province, China. Infection and drug resistance 2020;13:867-880.

(31) Ramanathan K, Thenmozhi M, George S, Anandan S, Veeraraghavan B, Naumova EN, et al. Assessing Seasonality Variation with Harmonic Regression: Accommodations for Sharp Peaks. International journal of environmental research and public health 2020 Feb 18,;17(4):1318.

(32) Nielson A. Practical time series analysis: prediction with statistics and machine learning. : O'Reilly Media; 2019.

(33) Jebb AT, Tay L, Wang W, Huang Q. Time series analysis for psychological research: examining and forecasting change. Frontiers in psychology 2015;6:727.

(34) Michael Foley. Time Series Analysis. 2021; cited 2021 Oct 10. Available at: https://bookdown.org/mpfoley1973/time-series/.

(35) Álvarez Martín T. Análisis y predicción del mercado inmobiliario en la Comunidad de Madrid. 2018.

(36) Department of natural resources. Daily Data New Hope station (ID: 215838) .2021; cited 2021 Oct 10. Available at: https://www.dnr.state.mn.us/climate/historical/daily-data.html?sid=215838&sname=NEW%20HOPE&sdate=por&edate=por&__cf_chl_captcha_tk__=pmd_u.CJ83xYrts5AGyQzGM_y19h9yPBKoUWPb9SPxcuX9E-1631986900-0-gqNtZGzNA1CjcnBszQlR.

(37) Wilkinson L. ggplot2: Elegant Graphics for Data Analysis by WICKHAM, H. Biometrics 2011 Jun;67(2):678-679.

(38) Hadley Wickham (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29. cited 2021 Oct 10. URL http://www.jstatsoft.org/v40/i01/.

(39) Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.5. https://CRAN.R-project.org/package=dplyr.

(40) Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. cited 2021 Oct 10. URL https://www.jstatsoft.org/v40/i03/.

(41) Adrian Trapletti and Kurt Hornik (2020). tseries: Time Series Analysis and Computational Finance. R package version 0.10-48.

(42) Bernhard Pfaff (2008). VAR, SVAR and SVEC Models: Implementation Within R Package vars. Journal of Statistical Software 27(4). URL http://www.jstatsoft.org/v27/i04/.

(43) Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, Petropoulos F, Razbash S, Wang E, Yasmeen F (2021). _forecast: Forecasting functions for time series and linear models_. R package version 8.15, <URL: https://pkg.robjhyndman.com/forecast/>.

(44) Duncan Temple Lang (2020). RCurl: General Network (HTTP/FTP/...) Client Interface for R. R package version 1.98-1.2. https://CRAN.R-project.org/package=RCurl.

(45) Kung-Sik Chan and Brian Ripley (2020). TSA: Time Series Analysis. R package version 1.3. https://CRAN.R-project.org/package=TSA.

(46) Paulo J. Ribeiro Jr, Peter J. Diggle, Martin Schlather, Roger Bivand and Brian Ripley (2020). geoR: Analysis of Geostatistical Data. R package version 1.8-1. https://CRAN.R-project.org/package=geoR.

(47) John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

(48) Kamil Barton (2020). MuMIn: Multi-Model Inference. R package version 1.43.17. https://CRAN.R-project.org/package=MuMIn.

(49) David Stoffer (2020). astsa: Applied Statistical Time Series Analysis. R package version 1.12. https://CRAN.R-project.org/package=astsa.

(50) Mitchell O'Hara-Wild, Rob Hyndman and Earo Wang (2021). feasts: Feature Extraction and Statistics for Time Series. R package version 0.2.2. https://CRAN.R-project.org/package=feasts.

(51) Carl Anners. Forecasting energy usage in the industrial sector in sweden using sarima and dynamic regressionUppsala universitet, Statistiska institutionen; 2017.

(52) Robert Nau. Summary of rules for identifying ARIMA models. Available at: https://people.duke.edu/~rnau/arimrule.htm, 2021.

(53) Robert Nau. Lecture notes on forecasting. 2014.

(54) Rodríguez J, Ruiz E. A powerful test for conditional heteroscedasticity for financial time series with highly persistent volatilities. 2005.

(55) Thamanukornsri C, Tiensuwan M. Applications of Box-Jenkins (Seasonal ARIMA) and GARCH models to dengue incidence in Thailand. Model assisted statistics and applications 2018;13(2):95-105.

(56) Richard Hardy. What is the null hypothesis of the Mcleod and Li test? 2017; cited 2021 Oct 10. Available at: https://stats.stackexchange.com/questions/317556/what-is-the-null-hypothesis-of-the-mcleod-and-li-test.

(57) Aashiq Reza. An Approach to Make Comparison of ARIMA and NNAR Models For Forecasting Price of Commodities. 2020; cited 2021 Oct 10. Available at: https://towardsdatascience.com/an-approach-to-make-comparison-of-arima-and-nnar-models-for-forecasting-price-of-commodities-f80491aeb400.

(58) López Ratón M, Rodríguez Álvarez MX, Cadarso Suárez CM, Gude Sampedro F. OptimalCutpoints: An R Package for Selecting Optimal Cutpoints in Diagnostic Tests. Journal of Statistical Software 2014 Nov 1,;61(1):1-36.

(59) Maleki A, Nasseri S, Aminabad M, Hadi M. Comparison of ARIMA and NNAR Models for Forecasting Water Treatment Plant's Influent Characteristics. KSCE J Civ Eng 2018 Sep;22(9):3233-3245.

(60) İbrahim Demir, Murat Kirişci. Forecasting COVID-19 disease cases using the SARIMA-NNAR hybrid model. medRxiv 2021 May 16,:159.

(61) Briët OJT, Amerasinghe PH, Vounatsou P. Generalized Seasonal Autoregressive Integrated Moving Average Models for Count Data with Application to Malaria Time Series with Low Case Numbers. PloS one 2013;8(6):e65761.

(62) Spencer SEF, Marshall J, Pirie R, Campbell D, French NP. The detection of spatially localised outbreaks in campylobacteriosis notification data. Spatial and spatio-temporal epidemiology 2011 Sep;2(3):173-183.

ANEXOS

ANEXO I. CÓDIGO DE R

El código de R empleado para los análisis desarrollados en este trabajo se encuentra en un repositorio online accesible a partir del siguiente enlace: https://github.com/izanhoyuela/TIME-SERIES-ANALYSIS-FOR-SALMONELLOSIS-OUTBREAKS-DETECTION