Trabajo Fin de Máster

Detección de outliers en grandes bases de datos

Sergio Da Vila Davila

Máster en Técnicas Estadísticas

Curso 2019-2020

Propuesta de Trabajo Fin de Máster

Título en galego: Detección de outliers en grandes bases de datos
Título en español: Detección de outliers en grandes bases de datos
English title: Outlier detection in big data
Modalidad: Modalidad B
Autor/a: Sergio Da Vila Davila, Universidad de Santiago de Compostela
Director/a: María José Lombardía Cortiña, Universidade da Coruña
Tutor/a: Esther López Vizcaíno, Instituto Galego de Estadística
Breve resumen del trabajo: La presencia de valores atípicos en un conjunto de datos puede condicionar gravemente las conclusiones que se extraigan de ellos. Por ello, se realiza una comparativa de diferentes métodos de detección de valores atípicos en series temporales a través de un estudio de simulación. Este estudio nos permite observar cuales son los mejores métodos y analizar tanto la presencia de atípicos como el efecto del Covid19 sobre los conjuntos de datos del IGE.
Recomendaciones:
Otras observaciones:

Doña María José Lombardía Cortiña, profesora titular de Universidad del Departamento de Matemáticas da Universidade da Coruña, y doña Esther López Vizcaíno, responsable del Servicio de Difusión e Información del Instituto Galego de Estadística, informan que el Trabajo Fin de Máster titulado

Detección de outliers en grandes bases de datos

fue realizado bajo su dirección por don Sergio Da Vila Davila para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En Santiago de Compostela, a 14 de julio de 2020.

La directora:

Doña María José Lombardía Cortiña

La tutora:

Doña Esther López Vizcaíno

El autor:

Don Sergio Da Vila Davila

Índice general

Resumen

Resumen en español

El objetivo de este trabajo es el de realizar un análisis acerca de la presencia de valores atípicos en los diferentes conjuntos de datos del Instituto Gallego de Estadística. Para ello, realizamos una revisión bibliográfica que nos permite conocer los métodos actuales y el recorrido que se produjo hasta llegar a los mismos. A través de esta revisión comprendemos la idea que se esconde detrás de la detección de valores atípicos en series temporales y somos capaces de proponer nuestros propios métodos.

Para comparar el comportamiento de los diferentes métodos realizamos un estudio de simulación, el cual nos sirve para contrastar el buen comportamiento de nuestras propuestas y para descartar otros métodos.

Finalmente aplicamos los métodos que han reflejado un comportamiento más consistente en el estudio de simulación y analizamos el efecto del Covid19 en el incremento de valores atípicos en los conjuntos de datos. Además, con el objetivo de reducir la incertidumbre asociada al campo de la detección de valores atípicos introducimos una nueva herramienta gráfica.

English abstract

The aim of this project is to carry out an analysis about the presence of outliers in the different data sets of the Galician Institute of Statistics. We carried out a bibliographic review that allows us to know the current methods and the route that took place until we reached them. Through this review we understand the idea behind the detection of outliers in time series and we are able to propose our own methods.

To compare the behavior of the different methods, we carried out a simulation study, which helps us compare the good behavior of our proposals and to rule out other methods.

Finally, we applied the methods that have reflected a more consistent behavior in the simulation study and analyzed the effect of Covid19 on the increase of outliers in the data sets. Furthermore, with the aim of reducing the uncertainty associated with the field of outlier detection, we introduced a new graphical tool.

Capítulo 1

Introducción

El Instituto Galego de Estadística (IGE) es un organismo autónomo de la Xunta de Galicia creado en el año 1988, cuyo objetivo es el de promover el desarrollo del sistema estadístico de la comunidad autónoma mediante servicios de recopilación y difusión de la documentación estadística disponible, desenvolver bases de datos de interés público, analizar las necesidades y la evolución de la demanda de estadísticas y asegurar su difusión. El IGE se erige, por tanto, como una de las principales fuentes de información de datos de carácter socioeconómico de Galicia y de sus divisiones territoriales.

Al igual que el resto de organismos del mismo ámbito, el IGE está experimentando un incremento continuo, tanto en la generación como en el almacenamiento de datos. Es por esto que se hace de vital importancia el incorporar una herramienta que nos permita identificar de un modo automático posibles candidatos a valores atípicos dentro de nuestros conjuntos de datos.

Un dato atípico o *outlier*, en inglés, es una observación anómala en comparación con el resto de datos contenidos en un determinado conjunto de datos. Puede deberse a diversos motivos, desde errores humanos: errores relacionados con la medición, grabación e introducción de los datos; hasta causas como que el dato procede de una población diferente de la que se pretende estudiar.

Existen multitud de definiciones que nos permiten dibujarnos una idea de a qué nos referimos cuando hablamos de valores atípicos:

- "Un dato atípico es una observación que se desvía tanto del resto de observaciones como para crear la sospecha de que fue creada por un mecanismo generador diferente" - Hawkins (1980).
- "Un dato atípico es una observación (o conjunto de observaciones) que son inconsistentes con el resto de datos" - Barnett y Lewis (1996).
- "Un outlier es una observación que se encuentra fuera del patrón general de una distribución"- Moore y McCabe (1999).
- "Los datos atípicos pueden ser, dependiendo de la circunstancia, errores no deseados que pueden afectar negativamente al resultado o valiosas pepitas de información inesperada"- Rousseeuw y Hubert (2011).

Eliminar un dato de una muestra por haberlo considerado atípico puede llevar a perder información relevante debido a una singularidad del mecanismo generador, y a su vez incluir un dato atípico en una

muestra puede confundir los resultados. Ambos casos alteran los análisis posteriores y pueden dirigir a conclusiones incorrectas si se llegara a tomar la decisión equivocada. Por lo tanto, la importancia reside en identificar de manera adecuada qué datos son atípicos y cuáles no.

Esta identificación será de gran utilidad dado que permitirá tanto realizar mejores análisis como poder interpretar sucesos relevantes en el contexto socio-económico gallego.

El problema de la detección de atípicos es un tema ampliamente tratado en la literatura, por lo que nuestro trabajo consistirá en realizar un estudio comparativo de los métodos más comunes en la actualidad y plantear posibles mejoras.

Uno de los criterios a los que le prestaremos mayor atención en este estudio será a la sensibilidad del método, dado que si este es muy sensible e identifica muchos candidatos como atípicos no servirá de gran ayuda. Por otra parte, si apenas señala ningún caso será similar a no aplicar ninguna herramienta de detección.

Para estudiar este comportamiento realizaremos un estudio de simulación en el que, a través de valores atípicos previamente identificados, podremos analizar el comportamiento de cada método. Además, dado que trabajamos con un gran volumen de datos, una característica que se les exigirá a los métodos será una rápida velocidad de ejecución.

La detección de datos atípicos tiene aplicaciones en muchos ámbitos: detección de operaciones fraudulentas en tarjetas de crédito, solicitudes para préstamos de clientes potencialmente morosos, detección de intrusiones en redes de comunicación, monitorización de parámetros de fabricación para detección de producciones defectuosas, anomalías en monitorizaciones médicas, análisis electorales, limpieza de datos, predicción del tiempo o hasta en astronomía, donde un punto anómalo puede implicar el descubrimiento de una nueva estrella.

En este trabajo nos centraremos en datos reales provenientes de la página web del IGE recogidos en siete conjuntos:

- Conjunto 1: Viajeros, noches y estancia media en establecimientos hoteleros y de turismo rural en España, Galicia y sus provincias. Datos mensuales. Los datos proceden del Instituto Nacional de Estadística (INE) de la Encuesta de ocupación hotelera y la Encuesta de ocupación en alojamientos de turismo rural. (`http://www.ige.eu/igebdt/igeapi/datos/3476`)

- Conjunto 2: Población de 16 y más años por sexo, grupos de edad y relación con la actividad económica en Galicia. Datos trimestrales. Los datos proceden de la Encuesta de Población Activa (EPA) elaborada conjuntamente entre el INE y el IGE. (`http://www.ige.eu/igebdt/igeapi/datos/6356`)

- Conjunto 3: Contratos registrados según su modalidad. Datos mensuales. Esta información procede de la Estadística de contratos registrados elaborada por el Servicio Público de Empleo Estatal (SEPE). (`http://www.ige.eu/igebdt/igeapi/datos/308`)

- Conjunto 4: Índice de producción industrial general y por destino económico de los bienes en Galicia (Base 2015). Datos mensuales. La información procede del Índice de Producción Industrial, operación estadística ejecutada por el INE. (`http://www.ige.eu/igebdt/igeapi/datos/9048`)

- Conjunto 5: Transacciones inmobiliarias por régimen y tipo de vivienda. Número, valor total y valor medio. Datos trimestrales. La fuente de esta información es el Ministerio de Fomento y

hace referencia a la compraventa de viviendas elevadas a escritura pública ante notario. (`http://www.ige.eu/igebdt/igeapi/datos/4052`)

- Conjunto 6: Bajas de demandas de empleo según género y duración de la demanda en Galicia y sus provincias. Datos mensuales. La fuente de esta información es el SEPE y hace referencia a las bajas que los servicios de empleo público tuvieron debido a una colocación, no hacer la renovación de la demanda en el periodo establecido o por otras causas. (`http://www.ige.eu/igebdt/igeapi/datos/1243`)

- Conjunto 7: Afiliaciones a la Seguridad Social último día del mes en Galicia y sus provincias. Datos mensuales. La información procede del Ministerio de Seguridad Social y Migraciones y hace referencia a las personas trabajadoras que están en alta en la Seguridad Social. (`http://www.ige.eu/igebdt/igeapi/datos/4885`)

Los motivos por los que se han escogido estos conjuntos obedece a diversas razones. En primer lugar, el más importante, incluir series en las que existe el conocimiento de que se presentan atípicos. A partir de ahí se ha conformado un conjunto de series que permitiesen recoger la mayor amalgama de naturalezas posibles que se presentan en las series que trata el IGE: mercado laboral, turismo, construcción o industria. Serían ámbitos que se recogen a lo largo de la selección de datos realizada. Además, se han incorporado series con distinta frecuencia de obtención, tanto series mensuales como trimestrales.

En este trabajo se recoge una comparativa de métodos, los cuales se pueden englobar bajo dos visiones de cómo abordar la identificación de anomalías en series de tiempo. Por un lado, existen los métodos cuyo mecanismo se fundamenta en modelos de series temporales que realizan un proceso iterativo para estimar el modelo integrando la posible influencia de observaciones atípicas. Por el otro, una visión más reciente, se plantea a través de realizar un proceso de detección en dos partes; en la primera se le aplica a la serie de tiempo un método de descomposición para después aplicar un método de detección de atípicos sobre el residuo.

Dado que ambos enfoques comparten un nexo común, como es la modelización de la serie de tiempo y la localización de atípicos, las podemos tratar conjuntamente en dos capítulos, el Capítulo 2 y el Capítulo 3.

En el Capítulo 2 se estudian las propiedades de modelización de cada método. La importancia de un buen método de modelización de series temporales reside en que cuanto mejor se consiga extraer el mecanismo generador de la serie de tiempo más resaltarán las observaciones atípicas. En el Capítulo 3 se describen los métodos de detección de atípicos. Estos dos capítulos conforman la parte teórica del trabajo.

Para estudiar el comportamiento de los diferentes métodos se plantea un profundo estudio de simulación en el Capítulo 4, en el que se utiliza el Índice de Youden como herramienta discriminatoria acerca de qué métodos ofrecen mejores resultados. En el Capítulo 5 se tratan los datos proporcionados por el IGE a través de los métodos que han presentado un mejor comportamiento en el estudio de simulación. Para finalizar, el Capítulo 6 expone las conclusiones que hemos extraído a lo largo de la realización de este trabajo.

Capítulo 2

Modelización de series temporales

2.1. Conceptos básicos

En este apartado definiremos los elementos principales que rodean a las series de tiempo. Para ello podemos echar mano de infinidad de manuales dado que las mismas son objeto de estudio en multitud de campos, desde las ciencias sociales y económicas hasta ramas que requieren de un bagaje mucho más matemático. En este trabajo haremos uso de Peña (2010) y Woodward, Gray, y Elliott (2017).

En la estadística básica estamos acostumbrados a trabajar con una muestra donde las observaciones $X_1, X_2, ..., X_n$ son variables aleatorias independientes e idénticamente distribuidas. Sin embargo, cuando las muestras se extraen en base a instantes de tiempo es muy probable que las observaciones estén correlacionadas entre si. Este tipo de muestras con dependencia temporal se denominan series de tiempo. Es decir, una serie de tiempo es una secuencia de observaciones ordenadas cronológicamente.

Este tipo de datos se utilizan en múltiples disciplinas, algunos ejemplos podrían incluir desde el ámbito económico como la evolución del Producto Interior Bruto (PIB) o del precio del petróleo, hasta ámbitos como la meteorología midiendo la temperatura de una localidad durante un período de tiempo o el ámbito sanitario siguiendo la evolución de una enfermedad en un paciente.

Uno de los principales inconvenientes a la hora de trabajar con series temporales son los datos atípicos. Estos datos atípicos pueden tener dos naturalezas, errores no deseados o pepitas de información (Rousseeuw y Hubert, 2011). Estas pepitas de información pueden servirnos, por ejemplo, para localizar sucesos importantes como huelgas, medidas de política económica o desastres naturales (Gómez y Taguas, 1995). Al igual que en el resto de campos de la estadística, que en nuestro conjunto de datos se encuentren valores atípicos puede llevarnos a incurrir en errores de especificación. Como se menciona en Chang, Tiao, y Chen (1988) es importante ser capaz de identificar estos sucesos para comprender mejor la estructura subyacente de la serie.

Uno de los modelos más comúnmente usado, y utilizado como punto de partida en el análisis de series temporales, es el modelo ARIMA. Los modelos ARIMA ajustan los valores de la serie en base a las observaciones previas y errores aleatorios con una estructura que le permite incluir tanto componentes cíclicos como estacionales. Siendo $\{X_t\}_t$ una serie de tiempo, un modelo ARIMA sería aquel que admite una representación:

$$\phi(B)(1-B)^d X_t = c + \theta(B)a_t,$$

donde X_t es la observación t de la serie de tiempo, B el operador retardo definido por $BX_t = X_{t-1}$, $\phi(B) = (1-\phi_1(B)-\phi_2 B^2-...-\phi_p B^p)$, $\phi(B) = (1-\theta_1 B-\theta_2 B^2-...-\theta_q B^q)$; c,$\phi_1, \phi_2, ..., \phi_p$,$\theta_1, \theta_2, ..., \theta_p$ son constantes, d es el número de diferencias necesarias para eliminar la tendencia de la serie, a_t es ruido blanco, y p y q los órdenes de los procesos autorregresivos (AR) y de medias moviles (MA).

Con esta representación del modelo ARIMA se puede deducir rápidamente que estos no son más que la mezcla de los dos primeros modelos de series temporales formulados, los modelos AR y los modelos MA. Además, si el proceso que genera la serie de tiempo presenta la posibilidad de estar formado por componentes de largo período temporal, se debe hacer uso de los ARIMA estacional múltiplicativo, ARIMA(p,d,q)x(P,D,Q), cuya representación sería:

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D X_t = c + \theta(B)\Theta(B^s)a_t,$$

En esta nueva formulación, D es el número de diferenciaciones estacionales aplicadas para eliminar la componente estacional, Φ_i y Θ_j las constantes respectivas a la parte estacional para $i = 1, 2, ..., P$ y $j = 1, 2, ..., Q$ con P igual al orden del proceso AR estacional y Q al orden del proceso MA estacional.

Sin embargo, la presencia de atípicos puede llevar a un incorrecto ajuste de los parámetros del modelo debido a que se pueden ver sesgados por el efecto del atípico. La mala estimación de los parámetros nos conduce a diversos errores dado que podemos no diseccionar bien la estructura de la serie como para entender su comportamiento. Además, en el caso de que el atípico se produzca en la última parte de la serie afectaría a las estimaciones futuras incurriendo en un error mayor de predicción.

Por ello surgieron métodos como X-13ARIMA-SEATS o TRAMO-SEATS, que parten de los modelos ARIMA e introducen mejoras como la detección y corrección de atípicos. Estos métodos fueron ampliamente utilizados durante mucho tiempo para la modelización de series temporales. Sin embargo, el desarrollo de la tecnología provocó la creación del Internet de las Cosas (Mahdavinejad et al., 2018), un entorno en el que casi la totalidad de los objetos y personas está conectado a internet dejando huella de su comportamiento. Las características de esta red provocaron un incremento enorme en la creación y obtención de datos, unas características para las cuales no estaban preparados estos métodos diseñados para escalar a conjuntos de tamaño pequeño o mediano.

Es aquí donde surge un nuevo enfoque para la detección de valores atípicos en series de tiempo basado en aplicar un proceso de descomposición y analizar los residuos. Los procesos de descomposición tienen por objetivo diseccionar una serie en tres componentes: tendencia, estacionalidad y residuo, de modo que se pueda establecer un patrón de como se comporta. En este trabajo el objetivo no es buscar el modelo más preciso sino aquel que consiga trazar mejor la estructura subyacente de la serie de forma que resalte los posibles valores atípicos en su componente residual, y es por ello que estos modelos resultan de tanto interés. El primer método basado en esta idea que recibió cierta repercusión fue el presentado por el equipo de Twitter (Hochenbaum, Vallis, y Kejariwal, 2017), el cual fue más tarde ampliado en el paquete `Anomalize` de `R` (Dancho y Vaughan, 2019).

Esta forma de proceder en dos partes, descomposición más detección, se debe a la idea de que visualizar o localizar atípicos en una serie es complicado debido a causas como la estacionalidad o

tendencia de la serie. Aplicarle a la serie un proceso de descomposición permite extraer estas componentes y resaltar las posibles anomalías. A continuación, en la Figura 2.1, mostramos un ejemplo de este procedimiento utilizando uno de los métodos que se emplearán posteriormente en el trabajo, *STL*.

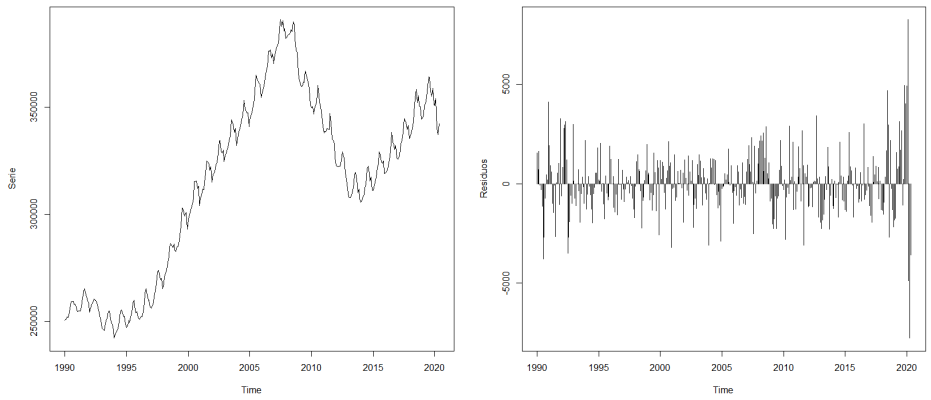

Figura 2.1: A la izquierda la serie referente al Número Total de Afiliaciones a la Seguridad Social el último día de mes en Pontevedra perteneciente al Conjunto 7. A la derecha los residuos obtenidos tras aplicarle la descomposición *STL*.

Como se puede apreciar en la Figura 2.1, observar algún comportamiento extraño en la serie de tiempo es complicado debido a las fluctuaciones que va sufriendo a lo largo de los años. Sin embargo, una vez aplicado el proceso de descomposición y eliminado los efectos de estacionalidad y tendencia podemos ver el efecto de la crisis del Covid-19 apareciendo en los últimos datos de la serie. Esta crisis provoca que los datos reales de esos meses disten mucho de los esperados, lo que se traduce en fuertes incrementos en el tamaño del residuo.

Lo interesante de estos métodos es que los procesos de descomposición de series de tiempo son capaces de analizar multitud de series de características muy diferentes y filtrarlas de tal modo que el problema acabe residiendo en lo que aparentemente es un sencillo problema de detección de valores atípicos en el caso univariante. Mientras que los métodos como *X-13ARIMA-SEATS* o *TRAMO-SEATS* se encuentran limitados en su diseño a trabajar solamente con series con frecuencias mensuales o trimestrales.

Sin embargo, Wilkinson (2017) resalta que el problema de la detección de valores atípicos en el caso univariante es engañosamente sencillo, lo cual puede llevarnos a cometer errores en la detección de atípicos pese a haber realizado una buena descomposición de la serie. Por ello en este trabajo se plantea un capítulo dedicado a analizar las propiedades de diferentes métodos de detección.

Por tanto, en el trabajo coexisten dos mecanismos de detección de atípicos en series temporales. Por un lado, los métodos de ajuste de series temporales que incorporan la detección y corrección de atípicos al modelo de modo automático. Y, por el otro, aquellos métodos de detección de atípicos cuya idea se centra en aplicar un proceso de descomposición más análisis del residuo.

En la Sección 2.2 se describen los procesos de modelización de series temporales utilizados y en el Capítulo 3 se describen los procesos de detección de atípicos.

2.2. Métodos de modelización de series temporales

2.2.1. X-13ARIMA-SEATS

El método *X-13ARIMA-SEATS* es la versión actualizada y mejorada del modelo *X-11* desarrollado por Shiskin, Young, y Musgrave (1967). *X-13ARIMA-SEATS* fue creado por el Censo de los Estados Unidos (US Census Bureau) y forma parte de una familia de modelos de ajuste para series temporales que se han ido desarrollando y están diseñados para trabajar con series económicas estacionales. Una explicación completa del método y de cómo ha de implantarse se puede encontrar en Time Series Research Staff (2017).

Este modelo se basa en la estimación de lo que sus autores denominan modelos regARIMA. Estos son modelos de regresión con errores ARIMA. En concreto, la media de la serie está descrita por una combinación lineal de regresores, y la matriz de covarianzas es la de un proceso ARIMA. Se incluyen regresores para modelar efectos de calendario, como festivos, vacaciones o día de la semana. Además, también se incluyen otro tipo de regresores para modelar el efecto de las anomalías que se puedan presentar. El modo en que se aborda este campo lo trataremos en profundidad en el Capítulo 3, con el resto de métodos de detección.

El proceso de ajuste comienza por un pre-ajuste de la serie a través de eliminar el efecto determinístico mediante un modelo de regresión con errores ARIMA. En el siguiente paso, la serie pre-ajustada se descompone en: tendencia (t), estacionalidad (s) y residuo (i). Esta descomposición puede ser: aditiva ($y = t+s+i$), multiplicativa ($y = t*s*i$), log-aditiva ($log(y) = log(t)+log(s)+log(i)$) o pseudo-aditiva ($y = t*(s+i-1)$). Este paso, en el que se aplica el proceso de descomposición, se basa en el uso del algoritmo *X11*, que descompone la serie a través de filtros lineales. La serie final ajustada debe estar libre de estacionalidad y efectos de calendario.

X-13ARIMA-SEATS está implementado en la librería `RJDemetra` de `R` en la función `x13` y las especificaciones necesarias para llevar a cabo la detección de atípicos se tratan en el Capítulo 3 junto con la explicación del proceso de detección.

2.2.2. TRAMO-SEATS

El método *TRAMO-SEATS* (Gómez y Maravall, 1997) fue diseñado por el Banco de España para el análisis de series de tiempo. Esta herramienta está formada por dos mecanismos. La parte *TRAMO* (Time Series Regression with Arima noise, Missing Observations and Outliers) y la parte *SEATS* (Signal Extraction in ARIMA Time Series).

Como se introduce en Gómez y Taguas (1995), el programa *TRAMO* puede ser utilizado independientemente de *SEATS*, sin embargo *SEATS* ha sido diseñado para trabajar conjuntamente con *TRAMO*. Su finalidad es la de realizar un análisis detallado de series temporales ya que cuenta con mecanismos para realizar tareas de: estimación, predicción, interpolación de modelos de regresión con valores ausentes con errores ARIMA y corrección de valores atípicos. El programa incluye variables de regresión para modelizar los días de calendario que pueden influir en el comportamiento de la serie como puede ser la Pascua y variables de intervención que permiten corregir el efecto de las observaciones atípicas. La idea en la que se basa el modelo es muy similar a la del modelo *X-13ARIMA-SEATS*.

TRAMO elimina de la serie los efectos especiales, identifica y elimina automáticamente los efectos

de varios tipos de atípicos e interpola las observaciones ausentes. También cuenta con un módulo que permite la identificación automática de modelos, herramienta que imitaron los desarrolladores de *X-13ARIMA-SEATS*, y que en su momento le permitió relevar de tareas monótonas a los analistas. Este procedimiento de identificación automática está basado en estimar primero las raíces unitarias y utilizar después el Criterio de Información Bayesiano (*BIC*) para especificar un modelo ARMA a la serie diferenciada.

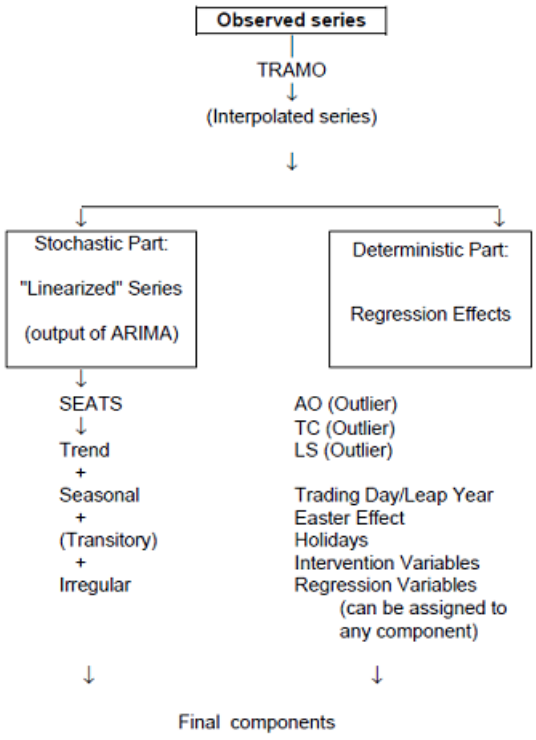

Figura 2.2: Esquema de funcionamiento de TRAMO-SEATS (Gómez y Maravall, 1997).

La Figura 2.2 (Gómez y Maravall, 1997) muestra como se integran ambas partes, *TRAMO* y *SEATS*, para formar en conjunto el mecanismo completo de modelización.

SEATS fue diseñado originalmente para desestacionalizar series temporales. El programa descompone una serie que sigue un modelo ARIMA en varios componentes: tendencia, componente estacional, ciclo y componente irregular. La descomposición puede ser aditiva o multiplicativa. La descomposición parte de la hipótesis de ortogonalidad de los componentes, que a su vez siguen modelos ARIMA. Para identificar los componentes se requiere que, excepto el irregular, estén limpios de ruido blanco. De este modo se maximiza la varianza de este último y, al contrario, la tendencia, el componente estacional y el ciclo son lo más estables posibles.

El modelo tratado por *SEATS* es el de una serie integrada lineal con innovaciones gaussianas. Esta hipótesis puede no ser cierta en múltiples ocasiones, pero siempre es necesario extraer un modelo, es por ello que *SEATS* fue diseñada para ser utilizada de forma conjunta con TRAMO.

TRAMO-SEATS está implementado en la librería `RJDemetra` de `R` en la función `tramoseats` y las

especificaciones necesarias para llevar a cabo la detección de atípicos se tratan en el capítulo 3 junto con la explicación del proceso de detección.

2.2.3. STL

STL es un proceso de filtrado para descomponer series de tiempo en tendencia, estacionalidad y residuo cuyo nombre se debe a sus siglas en inglés, Seasonal-Trend decomposition based on Loess introducido por Cleveland, Cleveland, McRae, y Terpenning (1990). El propósito de este método fue el de desarrollar un método de descomposición de series de tiempo que cumpliese los siguientes requisitos:

1. Diseño sencillo y uso inmediato.
2. Flexibilidad a la hora de especificar las variaciones en la tendencia y estacionalidad.
3. Poder descomponer series de tiempo con valores faltantes.
4. Tendencia y estacionalidad robusta de modo que no se vea distorsionada por datos anómalos.
5. Rápida computación y fácil implementación, incluso para series temporales largas.

El procedimiento de *STL* consiste en una secuencia de operaciones de suavizado realizadas, todas ellas salvo una, por el mismo suavizador: *loess*. El mecanismo consiste en dos bucles, uno interno y otro externo.

Cada vuelta del bucle interno consiste en un suavizado estacional que actualiza dicha componente, seguido de un suavizado de la tendencia que actualiza la componente asociada con la tendencia. Por otra parte, cada vuelta del bucle externo consiste en aplicarle una serie de pesos robustos al bucle interno. Estos pesos se utilizan en la siguiente vuelta del bucle interno para reducir la influencia de los datos anómalos.

La aplicación de la estimación robusta será necesaria cuando exista un conocimiento previo de que los datos tienen un comportamiento no gaussiano que conduce a una variación transitoria extrema, de lo contrario se pueden omitir las iteraciones del bucle externo y *STL* consistiría solamente en el bucle interno.

A diferencia de otros métodos, como *TRAMO-SEATS* o *X-13ARIMA-SEATS*, *STL* puede tratar cualquier tipo de estacionalidad, en vez de estar atado a estacionalidades mensuales o trimestrales. Su componente estacional es capaz de cambiar a lo largo del tiempo, y dicho cambio puede ser controlado por el usuario. Lo mismo sucede con el suavizado de la componente relacionada con la tendencia. Uno de sus inconvenientes es que no opera si en la serie se encuentran valores faltantes.

En `R` encontramos este método bajo la función `stl`, a la cual solo habrá que especificarle un parámetro. Dicho parámetro es el relacionado con el suavizado de la componente estacional e indica al mecanismo el número de observaciones consecutivas a utilizar para estimar los valores de dicha componente. Se podría introducir un valor numérico si tenemos algún indicio sobre el posible desarrollo de la componente estacional, sin embargo, dado que vamos a aplicarlo a un gran número de series el análisis individual sería costoso. La solución utilizada en este trabajo pasa por fijar este parámetro como periódico, lo cual asume que la evolución de la componente estacional es idéntica a lo largo de los años que se desarrolla la serie.

2.2.4. STR

STR es un método de descomposición de series de tiempo presentado por Dokumentov y Hyndman (2015). Sus siglas se deben a su nombre en inglés, Seasonal-Trend decomposition procedure based on Regression, similar a la idea de *STL*, salvo que ahora en vez de basarse en *loess* lo hace en regresión. En concreto, según los autores, *STR* es similar a la regresión *Ridge* (Hoerl y Kennard, 1970), y su versión robusta, *Robust STR* (Dokumentov y Hyndman, 2015), se podría relacionar con una regresión *LASSO* (Tibshirani, 1996).

La idea de su desarrollo se originó debido a que los autores consideran que a pesar de existir varios algoritmos de descomposición de series de tiempo existen muchas características en ellas que dichos algoritmos son incapaces de tratar. Las principales deficiencias que encuentran son:

- Incapacidad para proporcionar un modelo estadístico significativo y simple.
- Incapacidad (o dificultad) para calcular intervalos de confianza.
- Incapacidad para tener en cuenta regresores.
- Incapacidad para tener en cuenta estacionalidad fraccionada.
- Incapacidad para tener en cuenta múltiples estacionalidades.
- Incapacidad para tener en cuenta estacionalidades complejas y regresores que afecten a los datos de un modo estacional.

El objetivo con el que se desarrolló *STR* fue el de corregir estas deficiencias a la par que presentar un método claro, genérico, simple y robusto si fuera necesario. Para solucionar estas deficiencias enfocan la descomposición como un problema a resolver a través de una regresión cuantil o de mínimos cuadrados ordinarios.

De acuerdo a sus desarrolladores, *STR* constituye el método de descomposición de series de tiempo más genérico disponible en el momento de su publicación. Sin embargo, matizan una gran desventaja que presenta el método y es su ineficiencia o la ralentización que sufre el mecanismo a la hora de tratar casos en los que se utilizan varios componentes o predictores estacionales.

Uno de sus desarrolladores, Rob J Hyndman es conocido por sus aportaciones en el campo de las series de tiempo y su posterior implantación en paquetes para el software `R`, como el famoso paquete `forecast`. En este caso también ha colaborado en la implantación de STR con el paquete `stR` (Dokumentov y Hyndman, 2018).

2.2.5. Twitter

En Hochenbaum et al. (2017) se aborda el problema de detección de valores atípicos en series de tiempo introduciendo la idea de descomposición más detección como hemos comentado en la introducción de esta sección. El equipo de *Twitter* plantea en su artículo utilizar en un primer momento *STL* como método de descomposición para obtener los residuos, sin embargo, a la hora de trabajar con este método encuentran una serie de inconvenientes. Estos inconvenientes estarían provocados por ciertas anomalías espurias producidas en el residuo.

Para corregir esta situación optan por utilizar la mediana de la serie de tiempo para representar la tendencia "estable", que después será utilizada para calcular el residuo.

De este modo afirman que reemplazando la tendencia con la mediana se eliminan las anomalías espurias o ilegítimas del residuo. Los mismos desarrolladores publicaron un paquete en la plataforma GitHub (Twitter. Inc., 2015) que nos permitía hacer uso de este mecanismo, sin embargo su mantenimiento cesó a finales de 2015 y el paquete fue eliminado de la plataforma `CRAN` en marzo de 2019. El paquete `Anomalize` (Dancho y Vaughan, 2019) recoge este mecanismo y será el paquete que utilizaremos en este trabajo para utilizar esta modificación del método *STL*, bajo la función `decompose_twitter`.

2.3. Resumen métodos de modelización de series temporales

Método	Ventajas	Contras
X-13ARIMA-SEATS(2017)	Capaz de tratar con valores faltantes. Integra diversos tipos de descomposición. Rápida velocidad de ejecución.	Diseñado para trabajar solo con series mensuales o trimestrales.
TRAMO-SEATS(1997)	Capaz de tratar con valores faltantes. Integra diversos tipos de descomposición. Rápida velocidad de ejecución.	Diseñado para trabajar solo con series mensuales o trimestrales.
STL(1990)	Puede tratar series de cualquier frecuencia. Rápido. Cuenta con versión robusta.	No admite valores faltantes. Carece de flexibilidad cuando el período estacional es largo y se observa mucho ruido en la serie. Aún con su versión robusta, puede verse afectado por la influencia de los atípicos.
STR(2015)	Puede tratar series de cualquier frecuencia. Cuenta con versión robusta. Admite valores faltantes. Sensible a cambios de estacionalidad.	No puede seguir cambios bruscos en la tendencia. Extremadamente lento en comparación al resto de métodos.
Twitter(2017)	Puede tratar series de cualquier frecuencia. Elimina anomalías espúreas en la componente residual. Rápido.	No admite valores faltantes. No cuenta con versión robusta.

Cuadro 2.1: Resumen de los métodos de modelización de series temporales.

Capítulo 3

Detección de valores atípicos

3.1. Conceptos básicos

El problema que supone en la estadística la presencia de valores atípicos apareció muy pronto. En Hawkins (1980) mencionan que este problema pudo haber surgido ya en los siglos XVIII y XIX en encuestas estadísticas que se realizaban. Las primeras decisiones de qué hacer con ellos se basaron en criterios puramente subjetivos, en los que era el propio analista el que decidía descartar una observación si parecía anómala respecto al resto de los datos.

Posteriormente, se comenzaron a plantear posibles soluciones para tratar el problema bajo un criterio. La primera de ellas, fue la presentada por Peirce (1852), basada en lo que se podría entender como un test de razón de verosimilitud, donde todos los valores que superasen el umbral determinado por $c\sigma$ serían rechazados, siendo c una constante a calcular y σ la desviación típica de la muestra. Varios autores realizaron críticas a este método y propusieron su corrección a este criterio, una de ellas fue la de Chauvenet (1963). Esta pasaba por calificar a una observación como atípico si su valor se situaba fuera del intervalo definido por $1/(4n)$ puntos de la distribución Normal, siendo n el tamaño de la muestra.

En ese momento surgen diversas propuestas no muy relevantes que trataron de abordar dicho problema, hasta que surgió la llevada a cabo por Thompson (1935). Su estudio condujo a descubrir la distribución nula de un residuo estudentizado:

$$\frac{x_i - \bar{x}}{s}.$$

donde x_i es la observación sospechosa de ser atípica, $\bar{x}$ la media muestral y s la desviación típica de la muestra. A partir de esto, pudo deducir un procedimiento de detección de atípicos que rechazaría una proporción fija de todos los datos buenos y devolvería una tabla con los valores críticos adecuados.

Esta idea supuso el desarrollo del test de Grubbs (1950) el cual, utilizando la misma notación anterior, basaba su regla en los momentos muestrales de una normal:

$$G = \frac{max_{1 \leq i \leq n} |x_i - \bar{x}|}{s}$$

Grubbs (1950) asoció G con una distribución t con el objetivo de encontrar un máximo y un mínimo

para localizar atípicos. Esta idea fue posteriormente ampliada dando lugar al método *GESD* (Rosner, 1983), el cual se detalla en la siguiente sección.

Posteriormente, se presentó una de las herramientas más comúnmente utilizadas para la identificación de atípicos, el *boxplot*, introducido en Tukey (1977). Sin embargo, el objetivo inicial no fue la detección de atípicos, sino como un método para trazar una idea del intervalo en el que se encuentran la gran parte de nuestros datos. Además, la forma en la que se construye provoca dos grandes inconvenientes. No se puede aplicar a distribuciones sesgadas, y dado que no incluye el tamaño muestral en su formulación tenderá a etiquetar falsos atípicos cuanto más grande sea la muestra.

A la par que se iban presentando estas ideas para la búsqueda de valores atípicos en un conjunto de datos se fueron presentando mecanismos enfocados a las series de tiempo.

Uno de los primeros aportes al tratamiento de datos atípicos en series de tiempo fue propuesto por G. E. P. Box y Tiao (1975). Sin embargo, este método solamente permitía modelar el efecto de la observación atípica si se conocía de antemano el momento de la intervención, por lo que no se podría considerar una herramienta de detección, sino más bien una herramienta de intervención.

Posteriormente se propusieron diversos métodos para abordar el problema de la estimación de modelos ARIMA en series bajo esta circunstancia. El primero de ellos fue el de Abraham y Box (1979), que se basaba en proponer un enfoque bayesiano para resolver el problema, partiendo de ideas similares a las que se habían propuesto en otros contextos por autores como Tukey (1977). Otro método que buscaba resolver también el mismo problema fue el propuesto por Chang y Tiao (1983) cuya idea se centraba en aplicar un proceso iterativo.

Fue en este artículo (Chang y Tiao, 1983) donde se propuso por primera vez un método de detección e identificación de atípicos. Tsay (1986) propuso su versión que se basaba en combinar el proceso iterativo presentado en Chang y Tiao (1983) con la función de autocorrelación muestral extendida desarrollada en Tsay y Tiao (1984), la cual permite eliminar la necesidad de determinar el orden de diferenciación para producir una serie estacionaria que modelar.

En Chen y Liu (1993) se reconocen los numerosos aportes hechos en los anteriores artículos citados pero matiza que algunos problemas seguían estando vigentes:

a. La presencia de atípicos puede resultar en un modelo inapropiado.

b. Incluso si el modelo es el apropiado, los atípicos pueden seguir produciendo sesgo en el parámetro estimado y por tanto, afectar a la detección del atípico.

c. Algunos atípicos no van a ser identificados debido a un problema de enmascaramiento.

Su objetivo era presentar un método que fuese capaz de resolver los problemas b y c, de modo que se pudiese generalizar para aplicar a los cuatro tipos de atípicos en series temporales que se habían ido introduciendo. Los cuatro tipos son:

- IO (atípico innovativo): existe un atípico innovativo en el momento t cuando la innovación en ese punto esté directamente provocada por una cantidad desconocida debido a un suceso imprevisto.

- AO (atípico aditivo): diremos que se produce un atípico aditivo si en el momento t la serie se genera de manera diferente al resto.

- LS (cambio de nivel): existe un cambio de nivel si en el momento t la serie experimenta un incremento de todos sus valores. A diferencia del resto, tiene un efecto permanente sobre la serie una vez aparece.
- TC (cambio temporal): similar al cambio de nivel, solo que el efecto no es permanente y decrece exponencialmente con el tiempo.

X-13ARIMA-SEATS y *TRAMO-SEATS* basan su proceso de detección de atípicos en métodos inspirados en el artículo de Chen y Liu (1993), cuya idea se fundamenta en integrar un mecanismo iterativo diseñado para trabajar con series de tiempo que localiza e integra los atípicos al modelo. Por otra parte, en la siguiente sección introducimos los métodos de detección de atípicos que acompañan a los procesos de descomposición de series de tiempo, estos son: *GESD*, *Isolation Forest* y *HDoutliers*. Estos métodos son desarrollos de los mecanismos e ideas tratados al comienzo de esta sección como el método *Grubbs* (Grubbs, 1950).

3.2. Métodos de detección de valores atípicos

3.2.1. X-13ARIMA-SEATS

El análisis de diagnóstico del modelo regARIMA se realiza a través del análisis de los residuos del modelo estimado. Para el apartado de análisis de valores atípicos, *X-13ARIMA-SEATS* se fundamenta en Chang y Tiao (1983) con extensiones y modificaciones tratadas en Bell (1983) y Otto y Bell (1990), donde los valores atípicos pueden ser de tres de los cuatro tipos antes mencionados: AO, LS y TC.

El enfoque que plantea este mecanismo de detección es similar a la regresión paso a paso, procedimiento de regresión en el que se construye el modelo a través de ir añadiendo, o eliminando, variables predictoras en base a criterios de información.

En este caso las variables candidatas para la regresión son las AO, LS y TC para todos los puntos en los que la detección de atípicos se realiza. Es decir, se calcula el t-estadístico para ver la significatividad de cada tipo de atípico en cada instante temporal, se busca significatividad entre todos estos t-estadísticos y se añade la correspondiente variable de regresión (AO, LS o TC) al modelo con el fin de corregir los efectos que producen estas observaciones atípicas. De este modo se identifica e introduce el efecto de un momento atípico en el modelo. Mientras se produce el proceso de detección de atípicos se utiliza un estimador robusto de la desviación residual estándar, 1.48$*$la mediana del valor absoluto de la desviación residual.

X-13ARIMA-SEATS añade dos variaciones a este método. El método *addone* en el que cada vez que se añade al modelo un atípico se lleva a cabo una re-estimación del modelo, y el método *addall* que re-estima el modelo solo cuando un número de variables atípicas han sido añadidas al modelo.

Este método está implantado en `R` a través de la función `x13` del paquete `RJDemetra`. Dicha función requiere de un parámetro de especificación, `spec`, en que se concreta el modelo que queremos introducir. Este parámetro es el que nos ofrece diversas posibilidades de modelización de la serie, como se recoge en la Figura 3.1, entre ellas los efectos de calendario o los atípicos que presente la serie. Dado que buscamos sucesos atípicos en su sentido más amplio vamos a omitir incluir los efectos de calendario y solamente introducir la detección automática de atípicos. Esto provoca que en nuestro trabajo la

especificación que nos interesa, y con la que hemos trabajado, sea *RSA3*. La opción *RSA3* permite ajustar automáticamente un modelo ARIMA identificando y corrigiendo las posibles observaciones anómalas. El resto de modificaciones que permite introducir el parámetro de especificación de la función `x13` se recogen en la Figura 3.1. Por ejemplo, si estuviésemos interesados en modelizar nuestros datos a través de un modelo ARIMA de líneas aéreas observaríamos en la Figura 3.1 que tendríamos que escoger la especificación *RSA0*. Si ahora quisiéramos añadirle a dicho modelo detección de atípicos tendríamos que cambiar el parámetro de especificación a *RSA1*, y si además también nos interesa modelar los efectos de calendario sobre dicho modelo introduciríamos la especificación *RSA2*. El resto de especificaciones ya trabajan sobre el ajuste de un modelo ARIMA de un modo automático, la más sencilla es la que utilizamos nosotros en este trabajo, *RSA3*, la cual ajusta un modelo ARIMA de forma automática incluyendo la detección de atípicos. Las especificaciones *RSA4* y *RSA5* incorporan, respecto a *RSA3*, modelizar los efectos de calendario. La diferencia entre ellas se encuentra en que *RSA4* incluye en el modelo los efectos de calendario a través de dos variables, que representan los días laborables y los fines de semana, y *RSA5* incluye dichos efectos a través de siete variables, cada día de la semana, y ambas incluyen el efecto de la Pascua en el calendario. Por último se encuentra la especificación *RSAfull*, que incluye el ajuste automático de un modelo ARIMA y de los efectos de calendario e incorpora la detección de atípicos. Todas estas especificaciones son las que se detallan en la Figura 3.1, para profundizar más en ellas se puede consultar el manual de la librería `RJDemetra` (la Tente, Michalek, Palate, y Baeyens, 2020).

Identifier	Log/level detection	Outliers detection	Calendar effects	ARIMA
RSA0	*NA*	*NA*	*NA*	Airline(+mean)
RSA1	automatic	AO/LS/TC	*NA*	Airline(+mean)
RSA2	automatic	AO/LS/TC	2 td vars + Easter	Airline(+mean)
RSA3	automatic	AO/LS/TC	*NA*	automatic
RSA4	automatic	AO/LS/TC	2 td vars + Easter	automatic
RSA5	automatic	AO/LS/TC	7 td vars + Easter	automatic
RSAfull	automatic	AO/LS/TC	automatic	automatic

Figura 3.1: Parámetro de especificaciones a introducir al modelo X-13ARIMA-SEATS.

3.2.2. TRAMO-SEATS

El mecanismo de detección de atípicos que utiliza el programa *TRAMO-SEATS* aborda el problema con la intención de resolver los problemas **b** y **c** introducidos en el inicio del Capítulo 3, dado que entiende que estos no se han resuelto de un modo satisfactorio en Chen y Liu (1993). En Gómez y Taguas (1995) explican que aunque este método funciona de forma bastante satisfactoria su solución presenta una serie de deficiencias:

1. Se estima varias veces por máxima verosimilitud exacta, lo que es costoso.
2. No utiliza residuos exactos.
3. El algoritmo es excesivamente complicado.
4. Las regresiones múltiples no se hacen filtrando los datos y las columnas de la matriz por un filtro exacto, como el filtro de Kalman, sino que se utiliza un filtro condicional.

Por tanto, se presenta un mecanismo que pretende subsanar estas deficiencias a la par que corregir los problemas **b** y **c** mencionados en Chen y Liu (1993). Además, si se utiliza de forma secuencial, junto con el procedimiento de especificación automática del programa, supone un procedimiento alternativo al de Tsay (1986).

Al igual que *X-13ARIMA-SEATS*, *TRAMO-SEATS* está implementado en la librería `RJDemetra` de `R` en la función `tramoseats`. Dicha función también requiere de un parámetro de especificación, `spec`, en el que se concreta el modelo que queremos introducir. La utilidad y mecanismo sobre el que trabaja este parámetro es idéntico al explicado en la Sección 3.2.1 para el método *X-13ARIMA-SEATS* por lo que utilizaremos la especificación *RSA3*. Esta especificación incluye el ajuste automático de un modelo ARIMA corrigiendo las posibles observaciones atípicas.

El resto de posibles especificaciones que se pueden incluir en el modelo se recogen en la Figura 3.2. El significado de esta figura es equivalente al de la Figura 3.1 dado que ambos modelos, *TRAMO-SEATS* y *X-13ARIMA-SEATS*, están implementados en la misma libreria, `RJDemetra`, y comparten el mismo parámetro de especificación en la función que estima sus modelos en `R`.

Identifier	Log/level detection	Outliers detection	Calendar effects	ARIMA
RSA0	*NA*	*NA*	*NA*	Airline(+mean)
RSA1	automatic	AO/LS/TC	*NA*	Airline(+mean)
RSA2	automatic	AO/LS/TC	2 td vars + Easter	Airline(+mean)
RSA3	automatic	AO/LS/TC	*NA*	automatic
RSA4	automatic	AO/LS/TC	2 td vars + Easter	automatic
RSA5	automatic	AO/LS/TC	7 td vars + Easter	automatic
RSAfull	automatic	AO/LS/TC	automatic	automatic

Figura 3.2: Parámetro de especificaciones a introducir al modelo TRAMO-SEATS.

3.2.3. GESD

Introducido por Rosner (1983) como una mejora al por entonces conocido *ESD* (Rosner, 1975), por sus siglas en inglés: Extreme Studentized Deviate. A pesar de que *ESD* contaba con buenas propiedades, ante una amplia alternativa de atípicos, presentaba una serie de defectos que le conducían a etiquetar más candidatos a atípicos que el número apropiado. Esto provocó que se quedase obsoleto en favor de su actualización *GESD*: General Extreme Studentized Deviate.

Los pasos que sigue son los siguientes:

1. Decidir un máximo número de posibles candidatos a atípicos, r. La recomendación es de considerar el 20 % del tamaño muestral.

2. Comenzar las iteraciones, $i = 1$.

3. Calcular $R_i = \frac{max_t|x_i - \bar{x}|}{s}$, siendo x_i una observación sospechosa de ser anómala, $\bar{x}$ la media muestral y s la desviación estándar de la muestra, respectivamente.

4. Eliminar la observación que maximice $x_i - \bar{x}$.

5. Calcular el siguiente valor crítico: $\lambda_i = \frac{(n-i)t_{p,n-i+1}}{\sqrt{(n-1-1+t^2_{p,n-i-1})(n-i+1)}}$;
   donde $t_{p,n-i+1}$ es una distribución t con $n - i - 1$ grados de libertad y $p = 1 - \frac{\alpha}{2(n-i+1}$, siendo α la probabilidad de cometer errores de tipo I.

6. Pasar a la siguiente iteración, repetir los pasos 2 a 5 hasta $i = r$, quedando r definido en el paso 1.

7. El número de atípicos detectados se determina al encontrar el mayor i tal que $R_i > \lambda_i$; donde R_i se define en el paso 2 y λ_i en el paso 5.

GESD es principalmente el test de Grubbs (Grubbs, 1950) aplicado de forma secuencial, sin embargo existen unas pequeñas diferencias que premian su comportamiento. *GESD* realiza ajustes del valor crítico en base al número de atípicos a ser estudiado, algo que el test de Grubbs no hace. Además, si existe efecto de enmascaramiento, aplicar el test de Grubbs de forma secuencial produciría una detención demasiado temprana del algoritmo de detección lo que impediría localizar algunos atípicos.

Los estudios de simulación realizados para estudiar el comportamiento de *GESD* indican que el método es muy preciso cuando contamos con tamaños muestrales $n > 25$, por lo que en nuestro trabajo no tendremos problema. Sin embargo, debido a que es un método iterativo será más costoso computacionalmente a medida que se incrementen los tamaños muestrales, y es algo a tener en cuenta como posible freno al crecimiento futuro del mecanismo de detección.

En `R` podemos encontrar la función `gesd` dentro del paquete `Anomalize` (Dancho y Vaughan, 2019).

3.2.4. Isolation Forest

Isolation Forest, o *iForest*, es el método presentado por Liu, Ting, y Zhou (2009) donde se trata el tema de detección de atípicos desde un punto de vista diferente. Ese punto de vista se basa en dos propiedades relacionadas con la idea de qué es un valor atípico: son la minoría de un conjunto y tienen unos atributos muy diferentes a los de las observaciones normales. Es decir, las anomalías son pocas y diferentes, lo que las hace más propensas a encontrarse aisladas del resto.

Por esta razón los autores presentan un método basado en construir un árbol que aísle cada observación, de tal modo que los valores atípicos se encontrarán en ramas más próximas a la raíz y los valores normales se encontrarán en zonas más profundas del árbol. Es por ello que *iForest* (Liu et al., 2009), se distingue del resto de mecanismos basados en modelos, distancias o densidades.

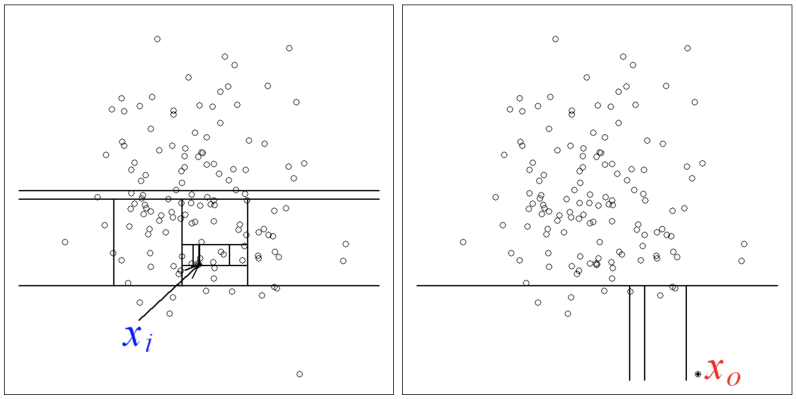

Figura 3.3: Idea sobre la que se sustenta el mecanismo de Isolation Forest.

En la Figura 3.3 se puede observar la idea sobre la que se rige este mecanismo. Un punto normal, x_i, requiere más particiones para ser aislado. Y el opuesto también es cierto, donde un punto anómalo, x_0 requiere menos particiones para ser aislado. Por lo que se trata de calcular la "longitud del camino" a ser aislado, dado que los valores atípicos tendrán caminos más cortos que los de los valores normales.

Para definir como se calcula la "longitud del camino" es necesario definir primero el árbol de aislamiento, *Isolation Tree*. El árbol de aislamiento sería toda la estructura de ramas que trocean el conjunto de datos. La longitud del camino $h(x)$, *path length*, de un punto x se mediría por tanto como el número de ramas a atravesar desde la raíz del árbol al punto.

Es necesario definir una puntuación de cuán anómala es una observación, *score*, para ello se introduce $c(n)$ como la media de $h(x)$, siendo E el operador esperanza, dado el tamaño de la muestra, n. De tal modo que:

- Cuando $E(h(x)) \to c(n), score \to 0.5$;
- Cuando $E(h(x)) \to 0, score \to 1$;
- Cuando $E(h(x)) \to n - 1, score \to 0$.

Haciendo uso de la puntuación anómala que recibe cada punto, *score*, tenemos que:

- Si *score* es cercano a 1 la observación es atípica.
- Si la observación presenta valores de *score* inferiores a 0.5 pueden ser calificadas como observaciones normales.
- Si todas las observaciones devuelven un valor $score \approx 0.5$ entonces la muestra entera carece de atípicos.

En el artículo fijan como atípicos potenciales observaciones con $score \geq 0.6$, en este trabajo somos más estrictos y fijamos el umbral en $score \geq 0.75$. Esta decisión se toma tras probar diferentes umbrales y observar el comportamiento del método en el estudio de simulación y en los datos reales. Otros umbrales que se barajaron y descartaron por los resultados mostrados fueron: $score \geq 0.7$ por ser demasiado laxo y $score \geq 0.8$ por ser demasiado estricto.

En `R` existen dos librerías que implantan dicho método, una es la implantada por la empresa *H20*, la cual ha desarrollado el algoritmo *HDoutliers* utilizado en este trabajo. Y la otra es la librería `solitude` (Srikanth, 2017), la cual hemos usado para el desarrollo de este trabajo.

3.2.5. HDoutliers

HDoutliers (Wilkinson, 2017) es un algoritmo que se puede aplicar en multitud de escenarios posibles, desde el más sencillo que sería el caso univariante hasta más complejos como el multidimensional o el espacial. La idea sobre la que se desarrolla es la de buscar espacios entre los valores ordenados, más que en la de buscar valores extremos. De los métodos que se centran en la idea de buscar valores extremos, como por ejemplo el test de Grubbs y por tanto *GESD*, el autor hace una crítica y es que estos utilizan medidas como la media muestral o la desviación típica las cuales no son robustas frente a valores atípicos.

El concepto anterior se puede explicar de un modo muy sencillo a través de un ejemplo que se recoge en Wilkinson (2017). En los resultados a un examen al que se presentaron 100 alumnos obtenemos una puntuación media de 50 y una desviación estándar de 5. Si entre todas las notas un alumno obtuvo un 100 y la siguiente mejor nota es un 65 se podría sospechar que ese alumno es un genio o un tramposo. Por otra parte, si la nota perfecta se encuentra en el punto más alto de una serie de notas que difieren 5 puntos entre ellas ese alumno ya no sería tan sospechoso. Los test clásicos de valores atípicos no son capaces de diferir entre ambas situaciones.

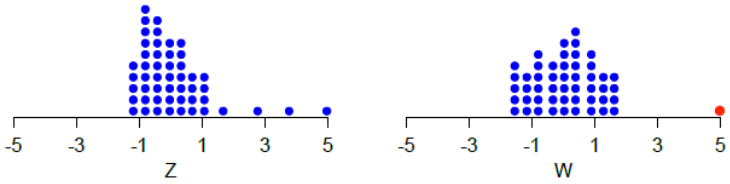

Figura 3.4: El algoritmo de HDoutliers aplicado sobre el ejemplo anterior (Wilkinson, 2017).

En la Figura 3.4 se representa este ejemplo. En el caso de la izquierda estaríamos en la situación en la que no se sospecha que se hayan producido trampas por parte de ningún alumno, y el método *HDoutliers* no detectaría ningún candidato a atípico. A la derecha tendríamos la situación en la que sospecharíamos de que existe un alumno brillante, o tramposo, y el método detectaría una situación atípica, como se señala con el punto rojo.

La solución de buscar espacios permite identificar valores inusuales tanto en el medio de las distribuciones como en los extremos. Para ello *HDoutliers* calcula la distancia al vecino más cercano. Después ajusta una distribución exponencial a la cola superior de las distancias calculadas y calcula el valor $1 - \alpha$, siendo α la probabilidad de cometer un error tipo I, superior de la función de distribución acumulada. Por lo que cada observación que se encuentre significativamente alejada del resto basada en este punto de corte se etiquetará como atípica.

El autor menciona que este método difiere de aquellos que se basan en hacer un ranking de candidatos a atípicos, como podría ser *Local Outlier Factor* (Breunig, Kriegel, Ng, y Sander, 2000), o que los etiquetan bajo un límite arbitrario, como sería el caso de *Isolation Forest* (Liu et al., 2009), dado que pueden conducir a resultados inconsistentes. Para ello lo que hace es asignar una probabilidad a la creencia de que estamos ante un atípico.

En `R` podemos encontrar este método en la librería `HDoutliers` (Fraley y Wilkinson, 2020).

3.3. Resumen métodos de detección de valores atípicos

Método	**Ventajas**	**Contras**
X-13ARIMA-SEATS(2017)	Califica el tipo de atípico. Rápida velocidad de ejecución.	Basado en modelos ARIMA.
TRAMO-SEATS(1997)	Califica el tipo de atípico. Rápida velocidad de ejecución. Idea similar a *X-13ARIMA-SEATS* pero profundizando más en el desarrollo del mecanismo de detección de atípicos .	Basado en modelos ARIMA.
GESD(1983)	Elimina parte de la incertidumbre de que se produzca un efecto de enmascaramiento.	Utiliza medidas como la media o la desviación típica, medidas que se pueden ver distorsionadas por valores atípicos. Es un proceso iterativo, lo que implica mayor coste computacional a medida que se incremente el tamaño de datos.
Isolation Forest(2009)	Escalable a conjuntos de gran dimensión. Diseño sencillo.	El usuario tiene que determinar el umbral de atípico, lo cual puede ser un inconveniente a veces.
HDoutliers(2017)	Identifica valores inusuales tanto en el medio de las distribuciones como en los extremos. Escalable a conjuntos de gran dimensión. Alta velocidad de ejecución.	Tendencia a calificar un exceso de observaciones como atípicas sin serlo.

Cuadro 3.1: Resumen de los métodos de detección de atípicos.

3.4. Métodos de detección de valores atípicos en series temporales

Dado que se han introducido diversos métodos, con diferentes enfoques respecto a la resolución del problema de detección de atípicos en series de tiempo, hemos decidido incluir este apartado al final del contenido principalmente teórico del trabajo para así centrar los métodos que se utilizarán en los siguientes capítulos. Este apartado recoge todos los métodos, los ya existentes y nuevas propuestas realizadas por nosotros, que se utilizarán en este trabajo.

Los métodos existentes son los pertenecientes a las librerías `RJDemetra` (la Tente et al., 2020) y `Anomalize` (Dancho y Vaughan, 2019). La librería `RJDemetra` es la que nos dota de las funciones necesarias para utilizar los modelos *X-13ARIMA-SEATS* y *TRAMO-SEATS*. Mientras que la librería `Anomalize` se ha convertido en uno de los paquetes de referencia en `R` para el análisis de valores atípicos en series de tiempo. A través de ella, podemos utilizar el método tratado en Hochenbaum et al. (2017) basado en una combinación de *Twitter+GESD* y también la otra opción que proponen a través de combinar *STL+GESD*.

A partir de la idea desarrollada por autores como Hochenbaum et al. (2017) o librerías como `Anomalize`, Dancho y Vaughan (2019), proponemos nuevos métodos de detección de atípicos para series temporales. El fundamento bajo el que se sostiene esta propuesta es muy sencillo. Estos mecanismos se fundamentan en aplicarle a la serie de tiempo un proceso de descomposición para extraer la estacionalidad y la tendencia, para posteriormente aplicarle un método de detección de valores atípicos al residuo. Dado que se han propuesto mejoras en ambos campos, tanto en los métodos de descomposición de series temporales como en el de detección de valores atípicos, creemos que su combinación podría suponer una mejora a las ya establecidas. Teniendo en cuenta que se ha introducido un método de descomposición de series temporales, *STR* (Dokumentov y Hyndman, 2015), y dos métodos de detección de valores atípicos, *Isolation Forest* (Liu et al., 2009), y *HDoutliers* (Wilkinson, 2017) lo que se hará será obtener todas las combinaciones posibles a través de mezclar los procesos de descomposición con los métodos de detección, dando lugar a las siguientes combinaciones:

- *STL +iForest*
- *STL +HDoutliers*
- *STL (Robusto) +iForest*
- *STL (Robusto) +HDoutliers*
- *Twitter +iForest*
- *Twitter +HDoutliers*
- *STR + GESD*
- *STR + iForest*
- *STR + HDoutliers*

- *STR (Robusto) + GESD*
- *STR (Robusto) + iForest*
- *STR (Robusto) + HDoutliers*

Estos son los nuevos métodos propuestos en el trabajo, de los cuales estudiaremos su comportamiento en el próximo capítulo a través de un estudio de simulación junto a los métodos ya existentes: *X-13ARIMA-SEATS*, *TRAMO-SEATS*, *STL+GESD* y *Twitter+GESD*.

Existen otros métodos populares en la detección de atípicos para el software `R` los cuales no se han incluido en este trabajo debido a que planteaban peores resultados que los finalmente introducidos, pero que creemos que merecen una mención debido a que pueden ser de interés para alguien que pretenda profundizar en este campo.

El primero de ellos sería el método de detección de atípicos *IQR*, incluido en la librería `Anomalize`, Dancho y Vaughan (2019), el cual se ha descartado debido a que sus combinaciones con *STL* y *Twitter* suponían una menor precisión que las combinaciones con *GESD* (Datacamp, 2018). Los autores lo han incluido en dicha librería debido a que su coste computacional es mucho menor que el de *GESD*, por lo que se puede plantear situaciones en las que sea útil.

El segundo sería la función `tsoutliers` del paquete `forecast` (Hyndman, 2020), la cual hemos descartado por presentar un mecanismo muy similar al que presentaría la combinación *STL+IQR*.

Por último, también se ha descartado la función `tso` del paquete `tsoutliers` (de Lacalle, 2019). El motivo se debe a que su autor enfoca el proceso de detección basándose en las mismas ideas que las desarrolladas en los modelos *X-13ARIMA-SEATS* y *TRAMO-SEATS* pero comenta que este método no está capacitado para trabajar con grandes y heterogéneos conjuntos de series temporales, como *X-13ARIMA-SEATS* o *TRAMO-SEATS*, sino más bien para pequeñas aplicaciones de un modo semi-automático (de Lacalle, 2015).

Capítulo 4

Estudio de simulación

En este capítulo se realiza un estudio de simulación con el objetivo de comparar la eficiencia de los diferentes métodos expuestos en el apartado anterior. En primer lugar, se presenta un resumen del procedimiento seguido para realizar las simulaciones en los diferentes escenarios y el motivo de proponer cada escenario. A continuación, se introduce el Índice de Youden, medida que nos permitirá discernir qué métodos presentan mejor comportamiento en el estudio. Finalmente se presentan los resultados del estudio y las conclusiones extraídas del mismo.

4.1. Escenarios

El procedimiento para obtener las series simuladas ha sido el siguiente:

1. Simular una serie de tiempo sin atípicos

2. A la observación correspondiente al instante t sumarle el valor w, el cual indicará el tamaño del atípico.

3. Si se quiere introducir más de un atípico repetir los pasos 1 y 2.

Se van a proponer ocho escenarios, a continuación pasamos a justificar la elección de cada uno de ellos. Los escenarios 1, 2 y 3 son los planteados en Chen y Liu (1993), artículo utilizado para analizar la eficiencia de los mecanismos de detección que inspiran los métodos aplicados en *X-13ARIMA-SEATS* y *TRAMO-SEATS*. El Escenario 4 plantea el modelo de líneas aéreas, utilizado ampliamente para datos económicos como los que se van a tratar posteriormente. Por último, los escenarios 5 a 8 proponen diversos modelos que han sido generados a partir de seleccionar una serie del IGE aleatoriamente y aplicarle la función `auto.arima` del paquete `forecast` (Hyndman, 2020). A continuación se recogen todos los escenarios planteados en el Cuadro 4.1. Los escenarios 1 a 3 se simulan a través de la función de `R` `arima.sim`, mientras que los escenarios 4 a 8 a través de la función `simulate` del paquete `forecast`.

Escenario	**Modelo**	**Motivo**
Escenario 1	AR(1)	Propuesto en el estudio de simulación de Chen y Liu (1993).
Escenario 2	MA(1)	Propuesto en el estudio de simulación de Chen y Liu (1993).
Escenario 3	ARIMA(0,1,1)	Propuesto en el estudio de simulación de Chen y Liu (1993).
Escenario 4	ARIMA(0,1,1)x(0,1,1)12	Modelos de líneas aéreas, muy utilizado en lo referente a datos económicos.
Escenario 5	ARIMA(1,0,1)x(0,1,2)12	Modelo extraído de aplicar la función `auto.arima` a una serie del IGE escogida aleatoriamente.
Escenario 6	ARIMA(1,1,1)x(1,0,1)12	Modelo extraído de aplicar la función `auto.arima` a una serie del IGE escogida aleatoriamente.
Escenario 7	ARIMA(1,1,2)x(0,1,1)12	Modelo extraído de aplicar la función `auto.arima` a una serie del IGE escogida aleatoriamente.
Escenario 8	ARIMA(0,1,1)x(1,0,0)12	Modelo extraído de aplicar la función `auto.arima` a una serie del IGE escogida aleatoriamente.

Cuadro 4.1: Resumen de los escenarios propuestos para el estudio de simulación.

En cada escenario se estudian cuatro casos:

- Caso 1: Serie con 100 observaciones y un atípico en $t = 40$, el tamaño del atípico es $w = 3.5$.
- Caso 2: Serie con 100 observaciones y un atípico en $t = 40$, el tamaño del atípico es $w = 4.5$.
- Caso 3: Serie con 100 observaciones y un atípico en $t = 2$, el tamaño del atípico es $w = 4$.
- Caso 4: Serie con 300 observaciones y dos atípicos en $t_1 = 40$ y $t_2 = 180$, el tamaño de los atípicos es $w_1 = 4$ y $w_2 = 5$.

En el Escenario 4 hubo que modificar los valores del tamaño del atípico introducido en cada caso debido a que no se obtenían resultados significativos para ningún método. Por lo que finalmente, solo para el Escenario 4, los valores introducidos fueron de: $w = 5.5$ para el Caso 1, $w = 6.5$ para el Caso 2, $w = 6$ para el Caso 1, $w_1 = 6$ y $w_2 = 7$ para el Caso 4.

Los Casos 1 y 2 buscan estudiar la sensibilidad del método de detección dado que en lo único que se diferencian es en el tamaño del atípico. El Caso 3 plantea una situación complicada, al encontrarse el atípico al principio de la serie es más difícil detectar ese valor por dos motivos: que el método no tenga todavía la información suficiente o que el método vea afectada su estimación por este valor tan temprano. El Caso 4 busca estudiar el comportamiento cuando se presenta más de un atípico.

4.2. Índice de Youden

Para comparar los resultados de los diferentes métodos en las simulaciones vamos a utilizar lo que en otros campos, como la investigación médica, se denomina test de diagnóstico. En este tipo de test se compara la realidad con la predicción del mecanismo para analizar su funcionamiento. Estos test de diagnóstico se presentan en lo que se denominan matrices de confusión como la presentada en el Cuadro 4.4. La idea de la matriz de confusión es la de comparar las predicciones realizadas con la realidad, de tal modo que se pueda observar cuando la predicción y la realidad coinciden y cuando no.

		Realidad	
		Negativo(0)	Positivo (1)
Predicción	Negativo(0)	Verdadero Negativo (VN)	Falso Negativo (FN)
	Positivo(1)	Falso Positivo (FP)	Verdadero Positivo (VP)

Cuadro 4.2: Matriz de confusión.

En nuestro trabajo Negativo (0) representaría una observación normal, y Positivo (1) una observación atípica. La diagonal de la matriz está representada por aquellas observaciones en las que la realidad y la predicción coinciden. Los Falsos Negativos (FN) son aquellas observaciones en que el

mecanismo determinó que la observación era normal pero en realidad era atípica. Por otra parte, los Falsos Positivos (FP) suceden cuando el mecanismo dictamina que la observación es atípica pero la realidad es que es una observación normal.

Los Falsos Positivos son también conocidos como Error Tipo I, la probabilidad de aceptar la hipótesis nula siendo esta falsa. Y los Falsos Negativos son conocidos como Error Tipo II, la probabilidad de no rechazar la hipótesis nula siendo esta falsa.

El índice de Youden (Youden, 1950) nació con motivo de poder comparar los resultados de dos o más mecanismos de detección y así tener la capacidad de discernir cual de ellos tiene una mejor capacidad discriminatoria. Es decir, el propósito final del índice es el de resumir la matriz de confusión en un valor que consiga representar el comportamiento del mecanismo de detección, de tal modo que dos mecanismos se puedan comparar directamente.

Este índice considera con la misma importancia tanto los falsos positivos como los falsos negativos. Utilizando otro campo de ejemplo, la medicina, un falso negativo supondría no tratar la enfermedad de un paciente dado que establecemos que está sano, cuando la realidad es que está enfermo. Por otra parte, un falso positivo supondría un coste, como podría ser el monetario financiando el tratamiento o el emocional de un paciente que la realidad es que está sano.

Trasladando ese ejemplo a nuestro trabajo, un falso negativo supondría no calificar como atípica una observación que lo es. Esto provocaría que estaríamos asumiendo como normal un comportamiento en el ámbito socio-económico gallego que no lo fue, estaríamos incurriendo en errores en posteriores análisis o estaríamos obviando sucesos relevantes.

Los falsos positivos también suponen un coste aquí, cada observación que el método detecte como atípica supondrá la intervención de un analista para tratar de entender qué ha sucedido. Si finalmente el analista concluye que la observación no es atípica se habrá desperdiciado una serie de horas y equipo que podrían haber sido invertidos para otro fin. De ambos escenarios surgen dos medidas:

- **Sensibilidad**: $\frac{VP}{VP+FN}$ refleja la proporción de verdaderos positivos sobre el total de positivos.
- **Especifidad**: $\frac{VN}{VN+FP}$ refleja la proporción de verdaderos negativos sobre el total de negativos.

El índice de Youden se construye a partir de ambas medidas del siguiente modo:

$$J = \textbf{Sensibilidad} + \textbf{Especifidad} - 1 = \frac{VP}{VP+FN} + \frac{VN}{VN+FP} - 1$$

El rango en el que se utiliza este índice es de 0 a 1, donde 1 representaría que el mecanismo es capaz de diferenciar perfectamente las observaciones atípicas de las normales y 0 todo lo contrario.

Sin embargo, teóricamente este rango puede variar de -1 a 1 (Shan, G, 2015), los valores menores que cero no se suelen contemplar dado que no tienen una interpretación significativa. En nuestro trabajo sucede esta situación, en la que se presentan mecanismos con valores negativos, y si bien de forma general esto no tiene explicación en otros campos en este estudio si.

Explicándolo a través de lo que podría ser un posible resultado de un mecanismo de detección:

		Realidad	
		Negativo(0)	Positivo(1)
Predicción	Negativo(0)	96	1
	Positivo(1)	3	0

Cuadro 4.3: Ejemplo de posible resultado de un mecanismo de detección en el que el índice de Youden sería negativo.

En el ejemplo del Cuadro 4.2 se analiza una serie de tiempo con 100 observaciones entre las cuales se encuentra un valor atípico. El mecanismo de detección habría detectado 3 posibles atípicos, pero ninguno de ellos es el atípico real. Si calculamos el índice de Youden obtenemos:

$$J = \frac{0}{0+1} + \frac{96}{96+3} - 1 \ = -0.03$$

El resultado es negativo debido a que el mecanismo de detección nos sitúa en una posición peor que no clasificar nada como atípico. Por lo que si 1 refleja que el mecanismo es capaz de diferenciar perfectamente las dos clases y 0 que el mecanismo no nos ayuda en absoluto a diferenciar entre ellas, un valor menor que 0 en nuestro trabajo supondría que el mecanismo sería peor que no clasificar nada, dado que nos infunde un mayor error sobre la realidad.

Un resultado de cero implicaría que el mecanismo no es útil, ya que no es capaz de diferenciar entre lo que es atípico y lo que no. En el Cuadro 4.3 se recoge un posible resultado que conduciría a esta situación. Este ejemplo recoge una situación en la que en una serie de tiempo de 100 observaciones se encuentran 5 valores atípicos, sin embargo, el método no es capaz de detectar ninguna de ellas.

		Realidad	
		Negativo(0)	Positivo(1)
Predicción	Negativo(0)	95	5
	Positivo(1)	0	0

Cuadro 4.4: Ejemplo de posible resultado de un mecanismo de detección en el que el índice de Youden sería cero.

$$J = \frac{0}{0+5} + \frac{95}{95+0} - 1 \ = 0$$

Finalmente, el resultado perfecto implicaría que el mecanismo de detección es capaz de diferenciar de forma precisa aquello que es atípico de lo que no lo es. En el Cuadro 4.4 se recoge un ejemplo de

esta situación. En este ejemplo se plantea una serie de tiempo que cuenta con 100 observaciones de las cuales 5 son atípicas, igual que en el ejemplo anterior, sin embargo, el método ahora es capaz de detectar cada una de ellas sin incurrir en Falsos Positivos ni Falsos Negativos.

		Realidad	
		Negativo(0)	Positivo(1)
Predicción	Negativo(0)	95	0
	Positivo(1)	0	5

Cuadro 4.5: Ejemplo de posible resultado de un mecanismo de detección en el que el índice de Youden sería uno.

$$J = \frac{5}{0+5} + \frac{95}{95+0} - 1 \ = 1$$

Podemos observar como los valores fuera de la diagonal de la matriz, aquellos que representan los Falsos Negativos y Falsos Positivos, son cero. En la diagonal de la matriz se encuentran los resultados, que muestran que el mecanismo ha conseguido identificar de forma perfecta aquellas observaciones que son atípicas. Esta sería la situación ideal, la cual representaría que el mecanismo de detección de atípicos funciona de forma idónea.

4.3. Resultados

A continuación se presentan los resultados para los ocho escenarios introducidos, recogidos en el Cuadro 4.1. Al introducir una gran cantidad de métodos, escenarios y casos, el análisis de los resultados se presenta confuso por lo que se decidió introducir un código de color para resaltar los tres mejores y tres peores resultados en cada caso. Los mejores resultados se asocian al color verde, siendo el verde más intenso el mejor resultado y decayendo gradualmente la intensidad del color según empeora. Por otra parte, los peores resultados están asociados al color rojo, siendo el rojo más intenso el peor resultado y decayendo la intensidad según el resultado mejora.

Se pensó en introducir un gradiente de color por cada caso, de modo que fuese variando de mejor a peor. Finalmente se descartó debido a que su contribución a la comprensión de los resultados era menor que la idea finalmente escogida de resaltar los tres mejores y los tres peores resultados.

Los resultados obtenidos para el índice de Youden se muestran a continuación. A mayores, en el anexo se han incluido tablas referentes a los resultados que han presentado los distintos métodos en base a otras dos medidas que también han resultado de utilidad a la hora de escoger los mecanismos adecuados. La primera de esas medidas es la **sensibilidad**, la cual hemos definido en este apartado, y la otra es el **exceso**, la cual se define como la media de detecciones incorrectas por simulación y cuyos resultados están incluidos en el anexo correspondiente.

	Caso 1	**Caso 2**	**Caso 3**	**Caso 4**
X-13ARIMA-SEATS	0.495	0.853	0.699	0.490
TRAMO-SEATS	0.655	0.926	0.803	0.494
STL+HDoutliers	0.561	0.830	0.562	0.727
STL+GESD	0.265	0.589	0.236	0.531
STL+iForest	0.632	0.882	0.632	0.788
STL(ROB)+HDoutliers	0.550	0.790	0.666	0.707
STL(ROB)+GESD	0.623	0.862	0.732	0.752
STL(ROB)+iForest	0.634	0.873	0.741	0.748
Twitter+HDoutliers	0.462	0.725	0.588	0.616
Twitter+GESD	0.303	0.587	0.426	0.433
Twitter+iForest	0.524	0.797	0.647	0.657
STR+Hdoutliers	0.603	0.833	0.595	0.667
STR+GESD	0.344	0.649	0.292	0.448
STR+iForest	0.670	0.882	0.675	0.719
STR(ROB)+Hdoutliers	0.534	0.803	0.630	0.654
STR(ROB)+GESD	0.379	0.683	0.456	0.557
STR(ROB)+iForest	0.619	0.859	0.695	0.711

Cuadro 4.6: Resultados Escenario 1.Índice de Youden. AR(1).

En el Cuadro 4.6 se recogen los resultados al Escenario 1, escenario en el que se simula un AR(1) igual al realizado por Chen y Liu (1993). Los resultados muestran que en los casos 1, 2 y 3, *TRAMO-SEATS* ejerce un papel superior al resto, buen comportamiento que desaparece en el Caso 4, caso en el que se introducen dos atípicos. Si analizamos conjuntamente los resultados de los cuatro casos podríamos decir que los métodos que reflejan un mejor comportamiento global en el escenario serían *STR+iForest* y *STL(Rob)+iForest*.

	Caso 1	**Caso 2**	**Caso 3**	**Caso 4**
X-13ARIMA-SEATS	0.504	0.828	0.645	0.494
TRAMO-SEATS	0.667	0.913	0.741	0.494
STL+HDoutliers	0.465	0.721	0.473	0.680
STL+GESD	0.217	0.457	0.147	0.428
STL+iForest	0.502	0.781	0.521	0.713
STL(ROB)+HDoutliers	0.457	0.723	0.576	0.642
STL(ROB)+GESD	0.516	0.761	0.651	0.667
STL(ROB)+iForest	0.519	0.798	0.660	0.713
Twitter+HDoutliers	0.457	0.690	0.587	0.640
Twitter+GESD	0.346	0.627	0.480	0.486
Twitter+iForest	0.506	0.774	0.658	0.689
STR+HDoutliers	0.580	0.814	0.566	0.658
STR+GESD	0.297	0.601	0.247	0.449
STR+iForest	0.626	0.866	0.647	0.719
STR(ROB)+HDoutliers	0.514	0.780	0.603	0.644
STR(ROB)+GESD	0.335	0.627	0.405	0.520
STR(ROB)+iForest	0.578	0.836	0.687	0.709

Cuadro 4.7: Resultados Escenario 2. Índice de Youden. MA(1).

En el Cuadro 4.7 se recogen los resultados al Escenario 2, escenario en el que se simula un MA(1) igual al realizado por Chen y Liu (1993). Los resultados que se observan son similares a los extraídos del Escenario 1. *TRAMO-SEATS* es claramente superior en los casos 1, 2 y 3 pero su buen comportamiento desaparece ante el Caso 4. Por otra parte, *STR+iForest* presenta un comportamiento más homogéneo, encontrándose en tres de los cuatro casos entre los tres mejores resultados. Lo mismo sucede con su versión robusta. En esta ocasión la combinación *STL+iForest* solamente se encuentra entre las mejores en el Caso 4.

	Caso 1	**Caso 2**	**Caso 3**	**Caso 4**
X-13ARIMA-SEATS	0.921	0.970	0.926	0.500
TRAMO-SEATS	0.921	0.971	0.921	0.500
STL+HDoutliers	0.306	0.502	0.234	0.435
STL+GESD	0.088	0.229	0.050	0.152
STL+iForest	0.338	0.575	0.279	0.451
STL(ROB)+HDoutliers	0.252	0.426	0.287	0.328
STL(ROB)+GESD	0.287	0.504	0.369	0.316
STL(ROB)+iForest	0.291	0.495	0.329	0.358
Twitter+HDoutliers	0.061	0.109	0.124	0.064
Twitter+GESD	0.015	0.027	0.056	0.025
Twitter+iForest	0.077	0.127	0.167	0.068
STR+HDoutliers	0.400	0.556	0.342	0.436
STR+GESD	0.202	0.365	0.108	0.230
STR+iForest	0.441	0.614	0.385	0.492
STR(ROB)+HDoutliers	0.237	0.391	0.248	0.268
STR(ROB)+GESD	0.199	0.381	0.215	0.254
STR(ROB)+iForest	0.277	0.451	0.279	0.289

Cuadro 4.8: Resultados Escenario 3. Índice de Youden. ARIMA(0,1,1).

En el Cuadro 4.8 se recogen los resultados al Escenario 3, escenario en el que se simula un ARIMA(0,1,1) igual al realizado por Chen y Liu (1993). En este escenario existe una hegemonía de *X-13ARIMA-SEATS* y *TRAMO-SEATS*, los cuales presentan resultados superiores a los de cualquier otro método. La combinación *STR+iForest* es el tercer método que, a pesar de la diferencia con los otros dos, presenta mejores resultados.

	Caso 1	**Caso 2**	**Caso 3**	**Caso 4**
X-13ARIMA-SEATS	0.402	0.592	0.067	0.480
TRAMO-SEATS	0.540	0.705	0.071	0.481
STL+HDoutliers	-0.003	-0.004	0.008	0.000
STL+GESD	-0.002	-0.001	0.002	-0.001
STL+iForest	-0.004	-0.004	0.008	0.001
STL(ROB)+HDoutliers	-0.005	-0.005	0.025	0.000
STL(ROB)+GESD	-0.039	-0.026	0.052	0.002
STL(ROB)+iForest	0.004	-0.002	0.026	0.000
Twitter+HDoutliers	0.000	-0.002	-0.001	0.007
Twitter+GESD	-0.025	-0.019	-0.016	0.003
Twitter+iForest	0.000	-0.001	0.004	0.016
STR+HDoutliers	0.359	0.417	-0.004	0.083
STR+GESD	0.409	0.445	-0.006	0.124
STR+iForest	0.411	0.472	-0.005	0.097
STR(ROB)+HDoutliers	0.014	0.006	0.002	0.039
STR(ROB)+GESD	0.019	0.017	0.008	0.075
STR(ROB)+iForest	0.014	0.005	0.002	0.021

Cuadro 4.9: Resultados Escenario 4. Índice de Youden. ARIMA(0,1,1)x(0,1,1)12.

En el Cuadro 4.9 se recogen los resultados al Escenario 4, escenario en el que se simula el modelo de líneas aéreas (G. Box y Jenkins, 1976). Este modelo es famoso por su uso en series socio-económicas. Por ello, cabría esperar un buen comportamiento de los modelos diseñados para este fin, *X-13ARIMA-SEATS* y *TRAMO-SEATS*. El resto de combinaciones presentan resultados poco consistentes.

	Caso 1	**Caso 2**	**Caso 3**	**Caso 4**
X-13ARIMA-SEATS	0.265	0.628	0.176	0.409
TRAMO-SEATS	0.398	0.747	0.248	0.432
STL+HDoutliers	0.238	0.469	0.281	0.125
STL+GESD	0.091	0.258	0.158	0.024
STL+iForest	0.263	0.531	0.305	0.134
STL(ROB)+HDoutliers	0.265	0.475	0.343	0.097
STL(ROB)+GESD	0.375	0.662	0.452	0.114
STL(ROB)+iForest	0.304	0.545	0.388	0.128
Twitter+HDoutliers	0.243	0.429	0.352	0.103
Twitter+GESD	0.168	0.376	0.309	0.057
Twitter+iForest	0.266	0.479	0.381	0.126
STR+HDoutliers	0.656	0.885	0.231	0.755
STR+GESD	0.424	0.769	0.064	0.567
STR+iForest	0.720	0.936	0.277	0.798
STR(ROB)+HDoutliers	0.189	0.326	0.212	0.523
STR(ROB)+GESD	0.174	0.317	0.192	0.665
STR(ROB)+iForest	0.229	0.364	0.232	0.588

Cuadro 4.10: Resultados Escenario 5. Índice de Youden. ARIMA(1,0,1)x(0,1,2)12.

En el Cuadro 4.10 se recogen los resultados al Escenario 5, escenario en el que se simula un modelo ARIMA (1,0,1)x(0,1,2)12 a partir de aplicarle la función `auto.arima` a una serie del IGE escogida aleatoriamente. En este escenario predomina la descomposición *STR*, ya que son las combinaciones junto a los tres métodos de detección de atípicos, *GESD*, *iForest* y *HDoutliers*, los métodos que presentan mejores resultados. En concreto, la combinación *STR+iForest* supone la combinación más interesante para este escenario.

	Caso 1	**Caso 2**	**Caso 3**	**Caso 4**
X-13ARIMA-SEATS	0.331	0.659	0.354	0.425
TRAMO-SEATS	0.457	0.764	0.477	0.438
STL+HDoutliers	0.513	0.790	0.494	0.624
STL+GESD	0.229	0.558	0.188	0.328
STL+iForest	0.560	0.833	0.567	0.648
STL(ROB)+HDoutliers	0.507	0.776	0.592	0.595
STL(ROB)+GESD	0.528	0.815	0.663	0.560
STL(ROB)+iForest	0.559	0.838	0.661	0.645
Twitter+HDoutliers	0.239	0.416	0.355	0.208
Twitter+GESD	0.101	0.229	0.185	0.100
Twitter+iForest	0.279	0.452	0.383	0.227
STR+HDoutliers	0.520	0.779	0.436	0.636
STR+GESD	0.278	0.584	0.151	0.385
STR+iForest	0.596	0.822	0.494	0.682
STR(ROB)+HDoutliers	0.399	0.666	0.490	0.545
STR(ROB)+GESD	0.274	0.524	0.349	0.404
STR(ROB)+iForest	0.457	0.731	0.568	0.579

Cuadro 4.11: Resultados Escenario 6. Índice de Youden. ARIMA(1,1,1)x(1,0,1)12.

En el Cuadro 4.11 se recogen los resultados al Escenario 6, escenario en el que se simula un modelo ARIMA (1,1,1)x(1,0,1)12 a partir de aplicarle la función `auto.arima` a una serie del IGE escogida aleatoriamente. En este escenario los métodos que muestran unos mejores resultados son *STR+iForest* y *STL+iForest*. También muestra un buen comportamiento el método *STL(Rob)+GESD*, basado en la combinación de la versión robusta de la descomposición *STL* y el método de detección de atípicos *GESD*.

	Caso 1	**Caso 2**	**Caso 3**	**Caso 4**
X-13ARIMA-SEATS	0.332	0.673	0.347	0.422
TRAMO-SEATS	0.426	0.765	0.425	0.453
STL+HDoutliers	0.451	0.693	0.414	0.503
STL+GESD	0.180	0.438	0.142	0.219
STL+iForest	0.509	0.759	0.458	0.527
STL(ROB)+HDoutliers	0.416	0.655	0.468	0.489
STL(ROB)+GESD	0.452	0.702	0.549	0.457
STL(ROB)+iForest	0.484	0.717	0.541	0.525
Twitter+HDoutliers	0.009	0.018	0.037	0.037
Twitter+GESD	-0.024	-0.014	0.011	0.010
Twitter+iForest	0.007	0.016	0.050	0.050
STR+HDoutliers	0.469	0.719	0.369	0.582
STR+GESD	0.228	0.495	0.122	0.329
STR+iForest	0.511	0.774	0.438	0.635
STR(ROB)+HDoutliers	0.282	0.495	0.347	0.330
STR(ROB)+GESD	0.224	0.432	0.301	0.285
STR(ROB)+iForest	0.343	0.557	0.407	0.385

Cuadro 4.12: Resultados Escenario 7. Índice de Youden. ARIMA(1,1,2)x(0,1,1)12.

En el Cuadro 4.12 se recogen los resultados al Escenario 7, escenario en el que se simula un modelo ARIMA (1,1,2)x(0,1,1)12 a partir de aplicarle la función `auto.arima` a una serie del IGE escogida aleatoriamente. En este escenario los métodos que muestran unos mejores resultados vuelven a ser *STL+iForest*, *STR+iForest* y cercano a ellos *STL(Rob)+iForest*. El resto de resultados no muestran ningún método estable para los cuatro casos.

	Caso 1	**Caso 2**	**Caso 3**	**Caso 4**
X-13ARIMA-SEATS	0.298	0.636	0.334	0.434
TRAMO-SEATS	0.419	0.712	0.427	0.481
STL+HDoutliers	0.425	0.697	0.389	0.620
STL+GESD	0.176	0.415	0.108	0.351
STL+iForest	0.480	0.734	0.423	0.662
STL(ROB)+HDoutliers	0.429	0.675	0.483	0.607
STL(ROB)+GESD	0.449	0.718	0.543	0.584
STL(ROB)+iForest	0.486	0.749	0.559	0.675
Twitter+HDoutliers	0.271	0.463	0.378	0.350
Twitter+GESD	0.127	0.286	0.189	0.147
Twitter+iForest	0.300	0.504	0.424	0.356
STR+HDoutliers	0.448	0.692	0.459	0.564
STR+GESD	0.203	0.438	0.167	0.274
STR+iForest	0.492	0.769	0.502	0.600
STR(ROB)+Hdoutliers	0.416	0.680	0.503	0.548
STR(ROB)+GESD	0.260	0.523	0.329	0.366
STR(ROB)+iForest	0.460	0.751	0.573	0.587

Cuadro 4.13: Resultados Escenario 8. Índice de Youden. ARIMA(0,1,1)x(1,0,0)12.

En el Cuadro 4.13 se recogen los resultados al Escenario 8, escenario en el que se simula un modelo ARIMA(0,1,1)x(1,0,0)12 a partir de aplicarle la función `auto.arima` a una serie del IGE escogida aleatoriamente. Los resultados más consistentes para los cuatro casos están formados por las combinaciones de *STR+iForest*, la misma combinación pero con la versión robusta de la descomposición *STR*, *STR(Rob)+iForest*, y la combinación *STL+iForest* y también su versión robusta *STL (Rob)+iForest*.

4.4. Conclusiones

Los resultados de las simulaciones muestran dos realidades. Por un lado nos encontramos con los resultados asociados a los escenarios 1, 2, 3 y 4. En estos escenarios prevalece el comportamiento de *TRAMO-SEATS* por encima de todos los métodos sin lugar a duda, incluso también podríamos incluir a *X-13ARIMA-SEATS* aquí. Cabe recordar que ambos parten de las mismas ideas, pero es *TRAMO-SEATS* el que implanta una serie de mejoras a los métodos de detección de atípicos en los que se inspiran ambos, esto se trata en el Capítulo 3. Por otra parte, en los escenarios 5, 6, 7 y 8 ya no existe tal hegemonía de *TRAMO-SEATS* frente al resto de métodos, de hecho nos encontramos unos resultados bastante más variados.

Entre los métodos de descomposición, (*STL*, *Twitter* y *STR*) el que presenta mejores resultados en las simulaciones es *STR*. En todos los escenarios presenta resultados que lo posicionan entre los tres primeros. Esto nos da una idea de que estamos ante un método flexible, capaz de entender la estructura subyacente de series de tiempo muy diversas y adecuar una respuesta realista del mecanismo generador. Sus buenos resultados son seguidos muy de cerca por la descomposición *STL*. Este es un resultado que concuerda con lo esperado debido a que siendo los dos métodos similares, plantean la misma idea pero la ejecutan de distinta forma, *STR* fue creado con el objetivo de suponer una mejora a *STL*

En lo que respecta a los métodos de detección, *GESD*, *HDoutliers* y *iForest* podemos extraer conclusiones acerca del comportamiento de cada uno:

- *GESD* es un buen método de identificación de atípicos. Obtiene buenos resultados, fácil de aplicar y es sencillo entender su mecanismo, algo muy útil para personas que quieran realizar, o transmitir, un estudio sin ahondar profundamente en materia. Por otra parte, sus buenos resultados en la detección correcta de atípicos se ven eclipsados por su tendencia a clasificar altos valores de observaciones como anómalas incorrectamente. Si estamos realizando un análisis individual esto podría no ser un problema dado que se podría intervenir modificando alguno de sus parámetros y conseguir mejores resultados. Sin embargo, en este trabajo buscamos un método que nos sirva como filtro para analizar multitud de series con la menor intervención posible, por lo que descartamos dicho mecanismo.

- *HDoutliers* supone una mejora a *GESD*. Muestra mejor comportamiento en todas las combinaciones con los métodos de descomposición. *GESD* solo consigue igualar o superar ligeramente sus resultados incurriendo en calificar una gran cantidad de observaciones como atípicas, lo que provoca un incremento en el número de observaciones calificadas erróneamente, esto se puede ver en las tablas de **exceso** introducidas en el anexo. Sin embargo, pese a suponer una mejora, no es el método que presenta mejores resultados.

- *Isolation Forest* o *iForest* es el método de detección más consistente a lo largo de los escenarios y para cada método de descomposición. Consigue un equilibrio entre buenos valores de detección correcta y bajas tasas de atípicos señalados incorrectamente, justo lo que se busca en un método de detección de atípicos. La idea de árbol bajo la que se desarrolla su mecanismo es fácil de comprender; cada observación recibe una puntuación entre 0 y 1 de cuán anómala es, esto permite flexibilizar la decisión del analista con el umbral en el que fija un valor como atípico. En Liu et

al. (2009) califican como observaciones sospechosas de ser atípicas aquellas cuya puntuación sea mayor a 0.6, en este trabajo somos más estrictos y fijamos un umbral superior, 0.75. Además su diseño le permite escalar a conjuntos de gran tamaño y dimensión, (*big data*), por lo que no supone un freno ante situaciones futuras.

Parece interesante también recordar que los mecanismos que emplean los métodos *X-13ARIMA-SEATS* y *TRAMO-SEATS* fueron analizados en estudios que planteaban simulaciones como las que se realizan en los escenarios 1, 2 y 3. Además, también cabe esperar un buen comportamiento que se espera de ellos en un modelo tan común como el de líneas áreas (Escenario 4). Sin embargo, cuando se aplican fuera de estos escenarios en modelos más complejos solamente *TRAMO-SEATS* consigue no quedarse atrás.

La simulación nos permite descartar la descomposición *Twitter*, la cual está quizás más enfocada a datos de alta frecuencia, datos que se producen en frecuencias de tiempo mucho más pequeñas, como minutos, en vez de datos mensuales como los planteados en el estudio de simulación, por lo que refleja peores resultados en escenarios de este ámbito. También nos permite descartar el método *X-13ARIMA-SEATS* dado que supone una idea similar, pero menos desarrollada en lo que a detección de atípicos se refiere, a *TRAMO-SEATS*. Finalmente, el método *GESD* también es descartado debido a sus resultados, en los que muestra un comportamiento inferior frente a *HDoutliers*, método que supone una mejora a la idea sobre la que se basa *GESD*.

Por lo que se concluye que los métodos más adecuados para la continuación del trabajo son: el modelo *TRAMO-SEATS*, las descomposiciones de series temporales *STL* y *STR* y los métodos de detección de atípicos *HDoutliers* y *iForest*, lo que conforman los siguientes mecanismos de detección de atípicos:

- *TRAMO-SEATS*
- *STL+HDoutliers*
- *STL+iForest*
- *STR+HDoutliers*
- *STR+iForest*

Capítulo 5

Aplicación a datos reales

En este capítulo se aplican los métodos que han mostrado unos mejores resultados en el estudio de simulación a las bases de datos del IGE comentadas en Capítulo 1. Estos métodos son: *TRAMO-SEATS*, *STL+HDoutliers*, *STL+iForest*, *STR+HDoutliers* y *STL+iForest*. De este modo, podemos analizar el comportamiento de los métodos con datos reales y extraer nuevas conclusiones que no surgían en el estudio de simulación. Esto se debe a que, si bien el estudio de simulación planteaba una diversidad de escenarios amplia, nunca serán suficientes para recoger la multitud de naturalezas que caracteriza a las series socio-económicas.

El análisis de las datos del IGE se hará en dos niveles. Un primer nivel mostrará el comportamiento global del conjunto mediante cuatro medidas: el total de atípicos detectados por cada método en el conjunto, el máximo número de atípicos que detecta el método en una serie del conjunto, el número de series que no presentan atípicos y la media de atípicos detectados por cada método. El segundo nivel mostrará un análisis más específico, donde se analizará el comportamiento de los métodos respecto a una serie escogida del conjunto.

Para finalizar el capítulo introducimos una herramienta gráfica diseñada con el objetivo de eliminar la mayor incertidumbre posible que acompaña a un campo como el de la detección de valores atípicos y analizamos su comportamiento.

5.1. Análisis Series IGE

Los conjuntos utilizados para este trabajo han pasado primero por un proceso de filtrado en el que se han eliminado todas las series que presentan menos de 24 observaciones o más de 24 ceros entre sus observaciones. De este modo, conseguimos quedarnos con series que recogen la información suficiente para trabajar con ellas.

A continuación se muestra el total de series perteneciente a cada conjunto.

	Conjunto 1	**Conjunto 2**	**Conjunto 3**	**Conjunto 4**	**Conjunto 5**	**Conjunto 6**	**Conjunto 7**
Pre-filtrado	108	72	55	42	162	165	55
Post-filtrado	84	60	50	42	90	165	47

Cuadro 5.1: Número de series de tiempo perteneciente a cada conjunto antes y después del proceso de filtrado.

Además, con el objetivo de capturar la influencia del Covid19 sobre las series de tiempo se procede a dividir el estudio en dos partes. La primera analiza las series con datos hasta 2019, si el conjunto presenta datos de frecuencia mensual esta fecha será diciembre de 2019, por otra parte, si el conjunto presenta datos de frecuencia trimestral esta fecha será el cuarto trimestre de 2019. La segunda analiza las series incluyendo el último dato actualizado, el cual varía dependiendo del conjunto. Los conjuntos con datos trimestrales incluyen hasta el primer trimestre del 2020 mientras que, los conjuntos con datos mensuales, incluyen como última actualización fechas comprendidas entre abril y junio de 2020, por lo que dicha fecha se concretará en el momento que se analiza dicho conjunto. Respecto a la fecha de inicio de las series varía dependiendo del conjunto, por lo que también se especificará en el inicio del análisis del mismo.

Para medir la influencia del Covid lo que se hará será calcular la tasa de variación del total de atípicos detectados entre ambos escenarios, las series hasta 2019 y las series actualizadas al último dato existente de 2020, la cual calcularemos como:

$$\text{Tasa de variación} = \left(\frac{\text{Total atípicos serie completa}}{\text{Total atípicos serie hasta 2019}} - 1 \right) * 100$$

Conjunto 1 (`http://www.ige.eu/igebdt/igeapi/datos/3476`)

Series de datos mensuales que contienen la información de los viajeros, noches y estancia media en establecimientos hoteleros y de turismo rural en España, Galicia y sus provincias. Los datos de este conjunto comienzan en enero de 1999 y finalizan en mayo de 2020.

	TRAMO - SEATS	**STL + HDoutliers**	**STL + iForest**	**STR + HDoutliers**	**STR + iForest**
Atípicos detectados	327	176	111	136	142
Max. número atip. serie	17	11	3	8	4
Series con 0 atípicos	13	20	11	23	3
Media de atípicos detec.	3.89	2.10	1.32	1.62	1.69

Cuadro 5.2: Resultados para el Conjunto 1 con datos hasta diciembre de 2019.

Los resultados del Cuadro 5.2 se producen para un total de 18816 observaciones que muestra el conjunto con datos hasta diciembre de 2019. Ahora vamos a estudiar los resultados de forma que cada serie esté formada por el último dato actualizado y a realizar una comparación entre ambos escenarios. En esta serie, el último dato actualizado es mayo de 2020, lo que conforma un total de 19172 observaciones.

	TRAMO - SEATS	**STL + HDoutliers**	**STL + iForest**	**STR + HDoutliers**	**STR + iForest**
Atípicos detectados	457	243	170	264	204
Max. número atip. serie	17	11	4	15	5
Series con 0 atípicos	1	6	1	10	3
Media de atípicos detec.	5.44	2.89	2.02	3.14	2.43

Cuadro 5.3: Resultados para el Conjunto 1 con datos hasta mayo de 2020.

	TRAMO - SEATS	**STL + HDoutliers**	**STL + iForest**	**STR + HDoutliers**	**STR + iForest**
Hasta 2019	327	176	111	136	142
Hasta mayo 2020	457	243	170	264	204
Tasa de variación	39.76 %	38.07 %	53.15 %	94.12 %	43.66 %

Cuadro 5.4: Total de atípicos detectados por cada método hasta diciembre de 2019 y abril 2020 y tasa de variación que produce el efecto del Covid en el Conjunto 1.

En el Cuadro 5.4 podemos ver como el efecto del Covid sobre el conjunto es muy notorio, algo que cabría esperar dado que estamos tratando series relacionadas con el Turismo en Galicia, uno de los sectores más afectados por la crisis del Covid. El Cuadro 5.3 incorpora datos hasta mayo 2020, lo que supone solamente un crecimiento del 1.89 % del total de datos. Sin embargo, el crecimiento que experimenta el total de atípicos detectado es muy superior, en el caso más suave representa un crecimiento de más del 38 %, llegando a suponer para algún método el 94 %.

A continuación, vamos a analizar un caso concreto de una serie del conjunto. Hemos escogido la serie que hace referencia al número de viajeros residentes en España hospedados en establecimientos de turismo rural en España. El motivo de su elección es que es la serie que provoca un mayor número de atípicos en el Cuadro 5.2, un total de 17 observaciones atípicas señaladas por el método *TRAMO-SEATS*.

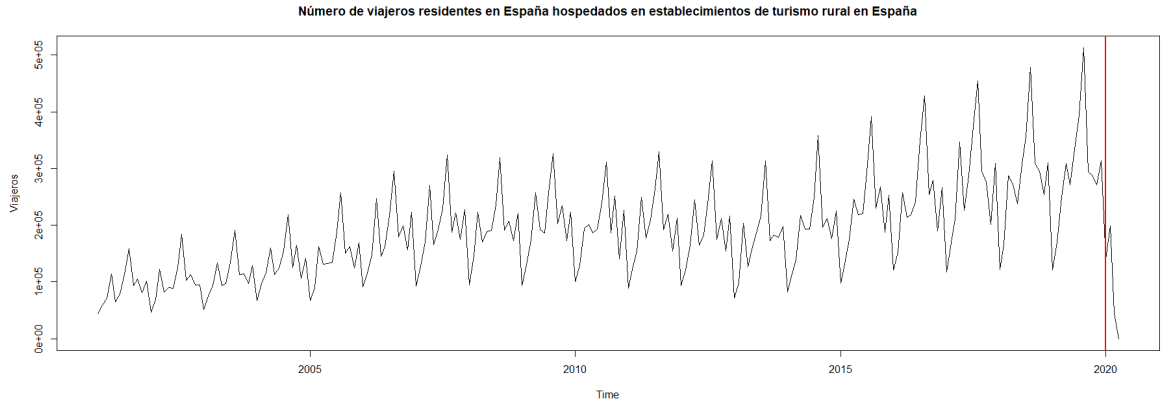

Figura 5.1: Serie de tiempo referente al número de viajeros residentes en España hospedados en establecimientos de turismo rural en España.

La Figura 5.1 recoge la serie de tiempo referente al número de viajeros residentes en España hos-

pedados en establecimientos de turismo rural en España. La línea vertical roja representa el comienzo del año 2020. Observando dicha figura y centrándonos en la serie hasta 2019, nos encontramos con una serie con mucha volatilidad, debido al fuerte carácter estacional que posee un sector como el turismo, y con una ligera tendencia creciente, que se estanca ligeramente en los años posteriores a la crisis económica de 2011, para después continuar en aumento. Sin embargo, no parece que se encuentren valores atípicos en dicha serie. La combinación *STR+HDoutliers* también señala un elevado nivel de atípicos, seis en total, que se reduce con *STR+iForest* a tres posibles atípicos. Finalmente, *STL+HDoutliers* y *STL+iForest* coinciden en señalar un atípico, agosto de 2019.

Si ahora observamos los resultados para la serie completa podemos comprobar que los resultados cambian bruscamente con la introducción de cuatro nuevos datos: enero, febrero, marzo y abril de 2020. Ahora el método *TRAMO-SEATS* señala cuatro fechas, de las que solo dos coinciden con las señaladas anteriormente. Las combinaciones *STL+HDoutliers* y *STL+iForest* mantienen como atípico agosto de 2019 y añaden febrero, marzo y abril 2020 y marzo y abril 2020, respectivamente. Y ahora son las combinaciones *STR+HDoutliers* y *STR+iForest* las que señalan únicamente agosto de 2019.

Es interesante observar el comportamiento de los cinco métodos en esta serie. *TRAMO-SEATS* y las combinaciones basadas en el proceso de descomposición *STR* solían ser las que mejores resultados presentaban, sin embargo, en este ejemplo son los métodos basados en la descomposición *STL*, que si bien presentaron buenos resultados en la simulación tendían a ir por detrás de *TRAMO-SEATS* y los métodos basados en la descomposición *STR*, los que presenta un comportamiento más acorde con lo que sucede en la serie. De hecho, son los métodos basados en la descomposición *STL* los únicos que señalan como atípica la observación de abril 2020, la cual es la única que tiene un valor de 0, y es claramente atípica.

La Figura 5.2 representa los atípicos detectados por el método *STL+HDoutliers* el cual es el método que mejor comportamiento muestra para esta serie.

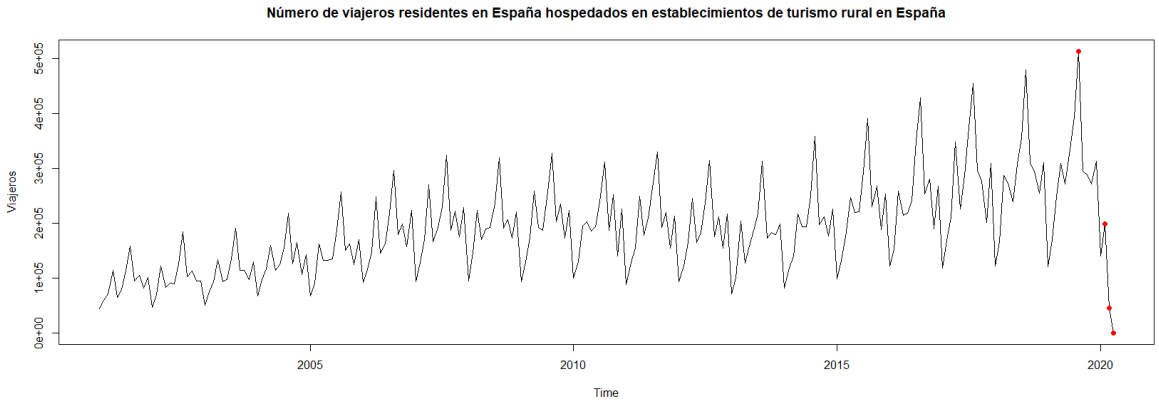

Figura 5.2: Serie de tiempo referente al número de viajeros residentes en España hospedados en establecimientos de turismo rural en España analizada por el método *STL+HDoutliers*.

En la Figura 5.3 se muestran los resultados para la misma serie por el resto de métodos utilizados: *STL+iForest*, *STR+HDoutliers*, *STR+iForest* y *TRAMO-SEATS*.

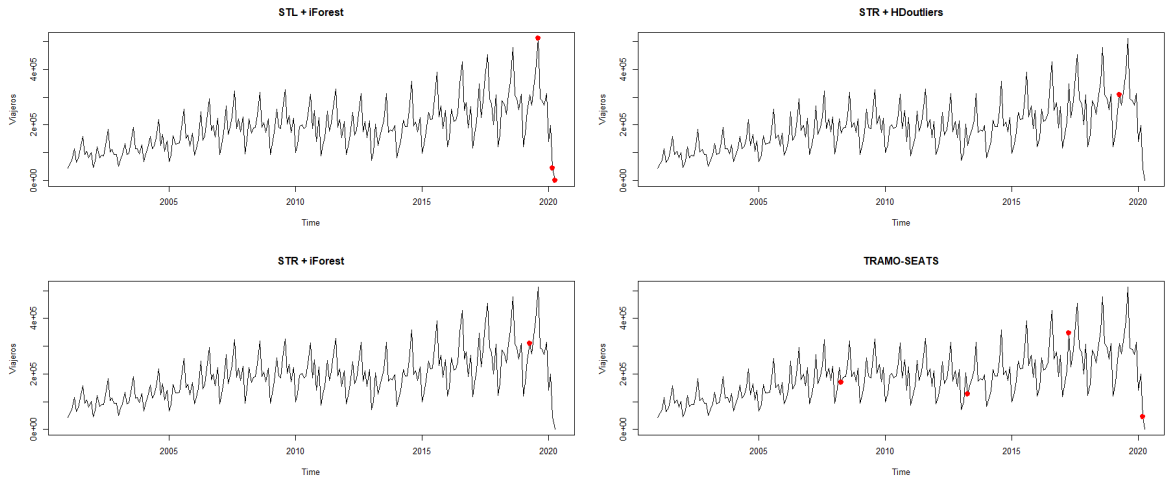

Figura 5.3: Serie de tiempo referente al número de viajeros residentes en España hospedados en establecimientos de turismo rural en España analizada por *STL+iForest*, *STR+HDoutliers*, *STR+iForest* y *TRAMO-SEATS*.

Conjunto 2 (`http://www.ige.eu/igebdt/igeapi/datos/6356`)

Series de datos trimestrales acerca de la población de 16 y más años desglosados por sexo, grupos de edad y relación con la actividad económica en Galicia. Los datos de este conjunto comienzan en el primer trimestre de 1996 y finalizan en el primer trimestre de 2020.

	TRAMO - SEATS	**STL + HDoutliers**	**STL + iForest**	**STR + HDoutliers**	**STR + iForest**
Atípicos detectados	56	70	61	62	66
Max. número atip. serie	5	6	3	5	3
Series con 0 atípicos	28	29	18	26	14
Media de atípicos detec.	0.93	1.17	1.02	1.03	1.10

Cuadro 5.5: Resultados para el Conjunto 2 con datos hasta el cuarto trimestre de 2019.

Los resultados del Cuadro 5.5 se producen para un total de 5760 observaciones que muestra el conjunto con datos hasta el cuarto trimestre de 2019. Ahora vamos a estudiar los resultados de forma que cada serie esté formada por el último dato actualizado y a comparar ambos escenarios. En esta serie, el último dato actualizado es el primer trimestre de 2020, lo que conforma un total de 5820

observaciones.

	TRAMO - SEATS	**STL + HDoutliers**	**STL + iForest**	**STR + HDoutliers**	**STR + iForest**
Atípicos detectados	54	107	62	68	66
Max. número atip. serie	5	45	3	4	3
Series con 0 atípicos	28	30	15	20	12
Media de atípicos detec.	0.90	1.78	1.03	1.13	1.10

Cuadro 5.6: Resultados para el Conjunto 2 con datos hasta el primer trimestre de 2020.

	TRAMO - SEATS	**STL + HDoutliers**	**STL + iForest**	**STR + HDoutliers**	**STR + iForest**
Hasta 2019	56	70	61	62	66
Hasta mayo 2020	54	107	62	68	66
Tasa de variación	-3.57 %	52.86 %	1.64 %	9.68 %	0.00 %

Cuadro 5.7: Total de atípicos detectados por cada método hasta el cuarto trimestre de 2019 y primer trimestre de 2020 y tasa de variación que produce el efecto del Covid en el Conjunto 2.

En el Cuadro 5.7 podemos ver como el efecto del Covid sobre el conjunto es mucho menor que en los datos del Conjunto 1. Esto se debe a la naturaleza trimestral de los datos, por lo que, situando la influencia del Covid en el contexto gallego a finales de marzo de 2020, su impacto es más suave sobre los datos. Sería interesante en el futuro ver que sucede con este conjunto una vez se añadan los datos del segundo trimestre.

A continuación, vamos a analizar un caso concreto de una serie del conjunto. Hemos escogido la serie que recoge los datos de hombres, mayores de 55 años y que forman parte de la población ocupada.

En la Figura 5.4 se muestra la serie de tiempo referente a los hombres, mayores de 55 años, que forma parte de la población ocupada. La recta vertical roja diferencia los dos escenarios que estudiamos, la serie hasta el cuarto trimestre de 2019 y la serie hasta el primer trimestre de 2020-

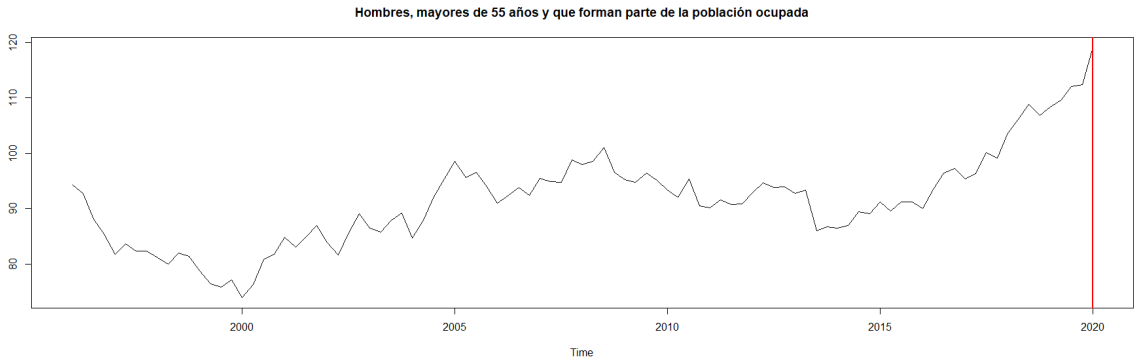

Figura 5.4: Serie de tiempo referente hombres, mayores de 55 años y que forman parte de la población ocupada.

El único dato que las diferencia, el primer trimestre de 2020, sigue con lo que cabría esperar. Y los métodos de detección de atípicos concuerdan con lo que observamos, no existen atípicos ni antes ni después, excepto en un caso. Ahora son los métodos basados en el proceso de descomposición *STL* los que señalan una fecha que parece discrepar con la realidad.

En la serie con valores hasta 2019, son ambas combinaciones, *STL+HDoutliers* y *STL+iForest*, las que señalan como posible atípico julio 2013. Si nos trasladamos a la serie completa, únicamente *STL+iForest* sigue manteniendo ese posible atípico, mientras que *STL+HDoutliers* lo corrige. Esta situación, en las que solo un método señala una fecha, nos lleva a dudar de que realmente esta observación sea atípica.

A continuación mostramos la serie referente a los hombres, mayores de 55 años y que forman parte de la población ocupada analizada por el único método que detecta atípicos para la serie completa: *STL+iForest*.

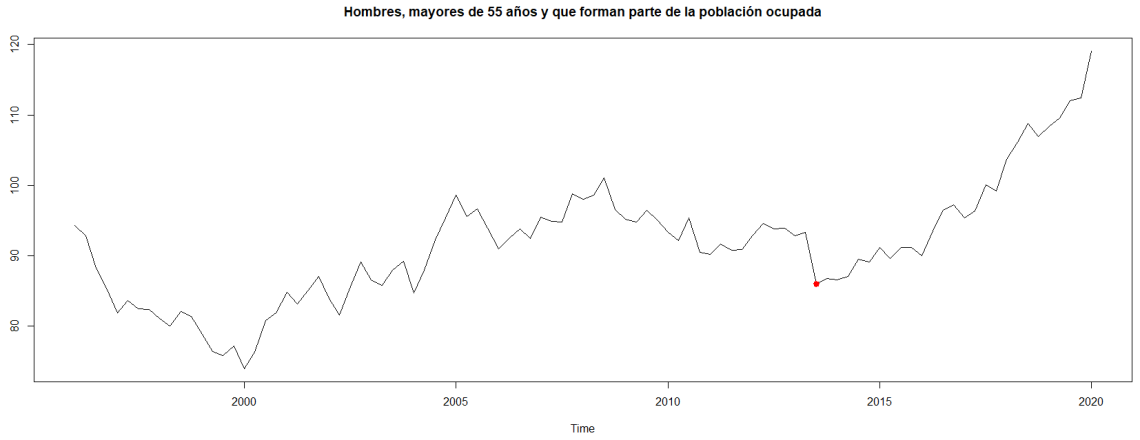

Figura 5.5: Serie de tiempo referente hombres, mayores de 55 años y que forman parte de la población ocupada analizada por *STL+iForest*.

Conjunto 3 (`http://www.ige.eu/igebdt/igeapi/datos/308`)

Series de datos mensuales acerca de contratos registrados según la modalidad del contrato. Los datos de este conjunto comienzan en enero de 1999 y finalizan en mayo de 2020.

	TRAMO - SEATS	**STL + HDoutliers**	**STL + iForest**	**STR + HDoutliers**	**STR + iForest**
Atípicos detectados	193	126	98	146	98
Max. número atip. serie	9	14	5	10	4
Series con 0 atípicos	5	9	1	7	2
Media de atípicos detec.	3.86	2.52	1.96	2.92	1.96

Cuadro 5.8: Resultados para el Conjunto 3 con datos hasta diciembre 2019.

El Cuadro 5.8 recoge los resultados que se producen para un total de 12600 observaciones que muestra el Conjunto 3 con datos hasta diciembre de 2019. Ahora vamos a estudiar los resultados de forma que cada serie esté conformada por el último dato actualizado y a realizar una comparativa entre ambas. En esta serie, el último dato actualizado es mayo de 2020, lo que conforma un total de 12850 observaciones.

	TRAMO - SEATS	**STL + HDoutliers**	**STL + iForest**	**STR + HDoutliers**	**STR + iForest**
Atípicos detectados	242	144	107	153	121
Max. número atip. serie	13	8	5	10	4
Series con 0 atípicos	1	6	0	7	1
Media de atípicos detec.	4.84	2.88	2.14	3.06	2.42

Cuadro 5.9: Resultados para el Conjunto 3 con datos hasta mayo de 2020.

	TRAMO - SEATS	**STL + HDoutliers**	**STL + iForest**	**STR + HDoutliers**	**STR + iForest**
Hasta 2019	193	126	98	146	98
Hasta mayo 2020	242	144	107	153	121
Tasa de variación	25.39 %	14.29 %	9.18 %	4.79 %	23.50 %

Cuadro 5.10: Total de atípicos detectados por cada método hasta diciembre de 2019 y mayo de 2020 y tasa de variación que produce el efecto del Covid en el Conjunto 3.

En el Cuadro 5.10 podemos ver como el efecto del Covid sobre el conjunto es similar al que se presentaba en los datos del Conjunto 1. Este conjunto vuelve a recoger datos de carácter mensual, por lo que la influencia de los meses de 2020 vuelven a reflejar un gran peso en los resultados. Sin embargo, las tasas de variación no son tan grandes como las que se producían en el Conjunto 1. Esto podría deberse a que, mientras el sector turístico se redujo a cero, en el mercado laboral se llevaron a cabo políticas contra la destrucción masiva de empleo. Estos resultados lo que hacen es aflorar que se ha producido un efecto en el conjunto de datos y que podría ser de interés para después desarrollar un estudio en profundidad.

A continuación, vamos a analizar la serie referente a las contrataciones eventuales por circunstancias de la producción para la provincia de Ourense, diferenciando a través de una recta roja vertical los dos escenarios analizados, los datos hasta diciembre 2019 y los datos hasta mayo de 2020.

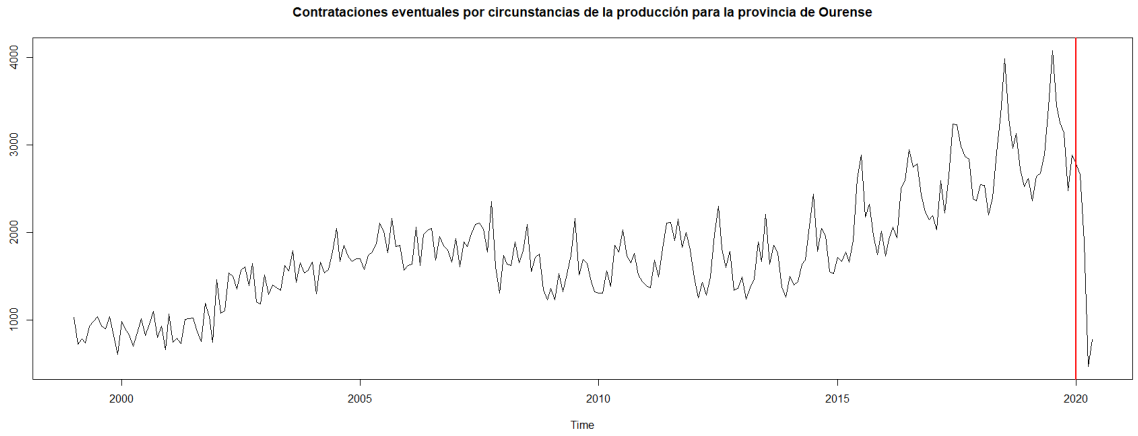

Figura 5.6: Serie de tiempo referente a las contrataciones eventuales por circunstancias de la producción para la provincia de Ourense.

El método *TRAMO-SEATS* señala tres fechas atípicas en ambas ocasiones, en la serie hasta 2019

y hasta mayo de 2020, sin embargo, ninguna de ellas es común. Probablemente esto se deba al comportamiento que podemos apreciar en la Figura 5.6, donde se observa la caída que produce en la serie los meses de marzo, abril y mayo de 2020. La introducción de estos datos reajusta las prioridades sobre qué es un atípico en la serie para *TRAMO-SEATS*, de forma que pasa de identificar octubre 2001, abril 2002 y enero 2012 a señalar como fechas atípicas octubre 2007 y marzo y abril de 2020.

Este suceso también se presenta en el resto de métodos. *STL+HDoutliers* pasa de no identificar ningún dato en la serie hasta 2019 a señalar abril del 2002 y febrero, abril y mayo de 2020. Otro método que experimenta esta situación es *STR+iForest*, que pasa de no calificar ninguna observación atípica a señalar abril y mayo en 2019 y en 2020. *STL+iForest* señala dos y tres atípicos en cada caso; en la serie hasta 2019 señala el mes de julio en 2018 y 2019, mientras que en la serie completa señala los meses de febrero, abril y mayo de 2020. Por último, *STR+HDoutliers* señala como atípico abril del 2002 en la serie acortada, observación que ya no es detectada al incluir la serie completa, donde señala los meses de abril y mayo para los años 2019 y 2020, igual aquí que *STR+iForest*.

Estos resultados conducen hacia un punto común, mientras que en la serie recortada en 2019 no parece haber mucho acuerdo sobre los atípicos, o si existen atípicos, en la serie completa todos los métodos se ponen de acuerdo en que las situaciones de abril y mayo de 2020 son anómalas. Estos resultados se muestran en la Figura 5.7.

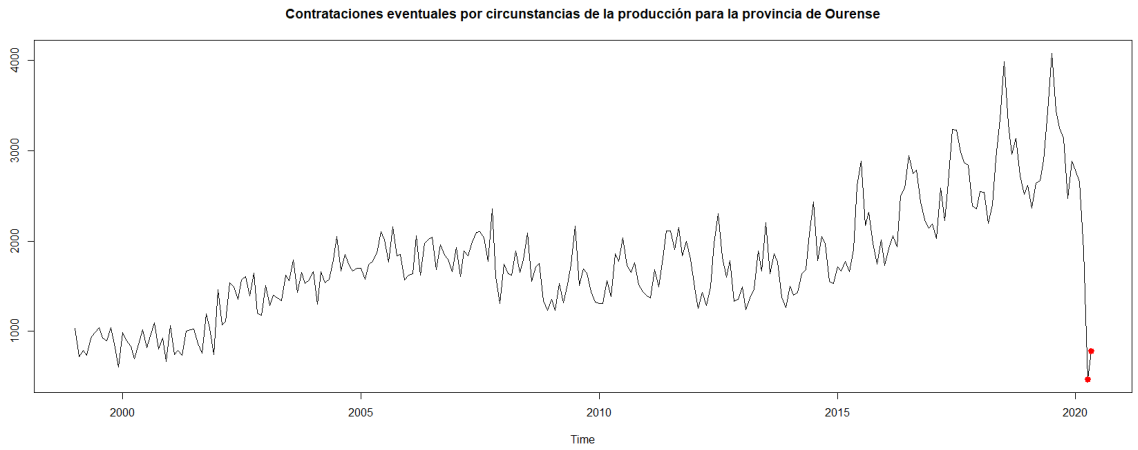

Figura 5.7: Serie de tiempo referente a las contrataciones eventuales por circunstancias de la producción para la provincia de Ourense señalizando las anomalías comunes a los cinco métodos

Conjunto 4 (`http://www.ige.eu/igebdt/igeapi/datos/9048"`)

Series de datos mensuales referidas al Índice de producción industrial general y por destino económico de los bienes en España y Galicia. Los distintos destinos económico son bienes de consumo, duraderos y no duraderos, bienes de equipo, bienes intermedios y energía. Los datos del conjunto comienzan en enero de 2002 y finalizan en mayo de 2020.

	TRAMO - SEATS	**STL + HDoutliers**	**STL + iForest**	**STR + HDoutliers**	**STR + iForest**
Atípicos detectados	288	129	73	212	64
Max. número atip. serie	24	11	3	102	3
Series con 0 atípicos	9	5	3	6	2
Media de atípicos detec.	6.86	3.07	1.74	5.05	1.52

Cuadro 5.11: Resultados para el Conjunto 4 con datos hasta diciembre de 2019.

El Cuadro 5.11 recoge los resultados que se producen para un total de 8904 observaciones que muestra el conjunto con datos hasta diciembre de 2019. Ahora vamos a estudiar los resultados de forma que cada serie esté conformada por el último dato actualizado y a analizar los cambios que se producen entre ambos escenarios. En este conjunto el último dato actualizado es mayo de 2020, lo que conforma un total de 9114 observaciones.

	TRAMO - SEATS	**STL + HDoutliers**	**STL + iForest**	**STR + HDoutliers**	**STR + iForest**
Atípicos detectados	282	92	78	144	79
Max. número atip. serie	24	5	3	50	3
Series con 0 atípicos	4	5	1	2	1
Media de atípicos detec.	6.72	2.19	1.86	3.43	1.89

Cuadro 5.12: Resultados para el Conjunto 4 con datos hasta mayo de 2020.

	TRAMO - SEATS	**STL + HDoutliers**	**STL + iForest**	**STR + HDoutliers**	**STR + iForest**
Hasta 2019	288	129	73	212	64
Hasta mayo 2020	282	92	78	144	79
Tasa de variación	-2.08 %	-28.68 %	6.85 %	-32.08 %	23.44 %

Cuadro 5.13: Total de atípicos detectados por cada método hasta diciembre de 2019 y mayo de 2020 y tasa de variación que produce el efecto del Covid en el Conjunto 4.

En el Cuadro 5.13 podemos ver el efecto del Covid sobre el conjunto, sin embargo, en esta ocasión no podemos ver una dirección clara. Los métodos difieren, y, mientras unos métodos muestran crecimientos en el número de atípicos detectados para el conjunto, otros decrecen y muestran ahora menos atípicos. Sin embargo, esto no implica que el efecto del Covid sobre el conjunto sea menor que por ejemplo en el Conjunto 3. Las variaciones entre las series, las que incluyen datos de 2020 y las que no, son notorias, lo cual implica que el efecto existe.

A continuación, vamos a analizar la serie referente a la producción industrial destinada a bienes intermedios de Galicia diferenciando a través de una recta vertical roja los dos escenarios, los datos hasta diciembre de 2019 y la serie hasta mayo 2020.

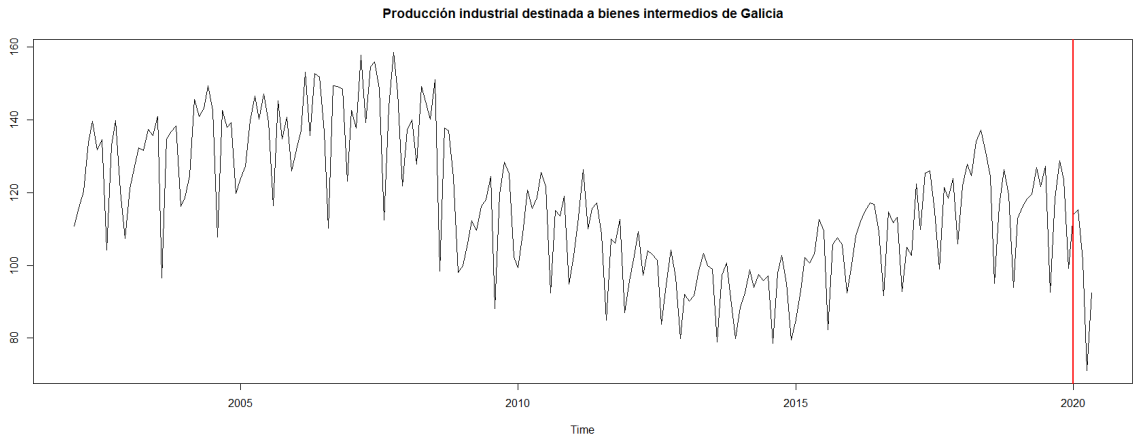

Figura 5.8: Serie de tiempo referente a la producción industrial destinada a bienes intermedios de Galicia.

El comportamiento que muestran los métodos con esta serie nos ayuda a plasmar una idea que desarrollaremos más adelante en el siguiente apartado. Si observamos los datos para la serie recortada

en diciembre de 2019, obtenemos que, salvo *TRAMO-SEATS* que no detecta ningún atípico, los otros cuatro métodos concuerdan en dos fechas: marzo y julio de 2008. Por otra parte, en la serie completa, cinco métodos señalan abril 2020 y cuatro señalan marzo de 2008. Estos dos atípicos, abril de 2020 y marzo de 2008, los representamos en la Figura 5.9.

Analizando la Figura 5.9 el atípico señalado en marzo de 2008 parece dudoso, sin embargo, si profundizamos en los datos observamos que esta fecha rompe una clara tendencia creciente respecto a los meses de marzo de años anteriores.

Este suceso, en el que varios métodos coinciden en identificar una fecha, proveen al analista de una seguridad a la hora de calificar una observación como anómala. Este concepto de seguridad, eliminando la incertidumbre asociada al campo de la detección de valores atípicos, será desarrollado más adelante.

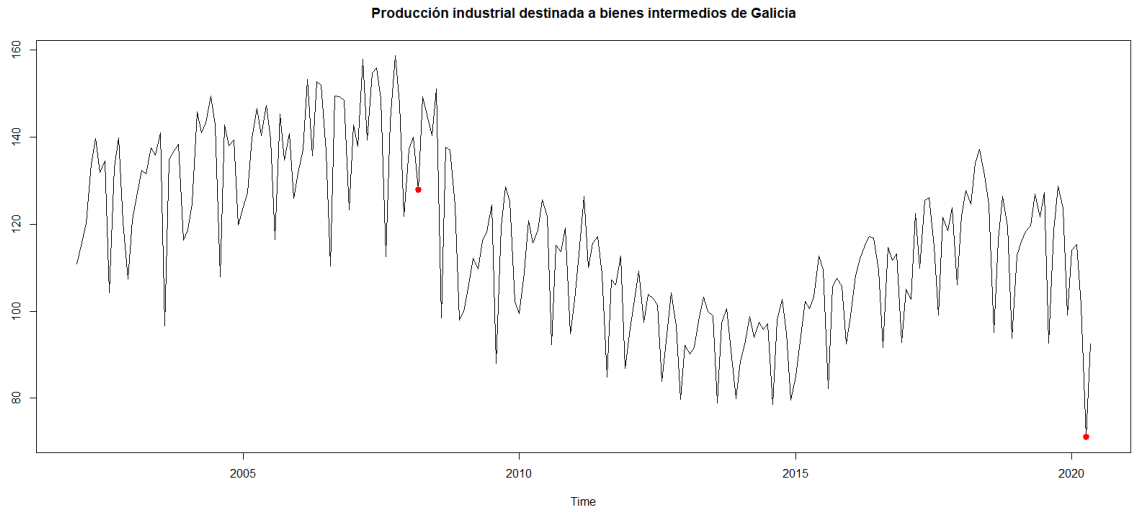

Figura 5.9: Serie de tiempo referente a la producción industrial destinada a bienes intermedios de Galicia con los dos atípicos .

Conjunto 5 (`http://www.ige.eu/igebdt/igeapi/datos/4052`)

Series de datos trimestrales de número de transacciones inmobiliarias por régimen (Libre, Protegida y Total de viviendas) y tipo de vivienda (Nuevas, segunda mano y total de viviendas). Las medidas en las que se toman los datos son por número de viviendas, su valor total (en miles de euros) o su valor medio (en euros). Los datos se presentan a nivel España, Galicia y las provincias gallegas. Los datos se recogen entre el primer trimestre de 2004 hasta el primer trimestre de 2020.

	TRAMO - SEATS	**STL + HDoutliers**	**STL + iForest**	**STR + HDoutliers**	**STR + iForest**
Atípicos detectados	184	122	105	94	97
Max. número atip. serie	7	8	2	6	2
Series con 0 atípicos	13	41	12	47	11
Media de atípicos detec.	2.04	1.36	1.17	1.04	1.07

Cuadro 5.14: Resultados para el Conjunto 5 con datos hasta el cuarto trimestre de 2019.

Los resultados que recoge el Cuadro 5.14 se producen para un total de 5760 observaciones que muestra el Conjunto 5 con datos hasta el cuarto trimestre de 2019. Ahora vamos a estudiar los resultados de forma que cada serie esté conformada por el último dato actualizado y a realizar una comparación entre ambos escenarios. En esta serie, el último dato actualizado es el primer trimestre de 2020, lo que conforma un total de 5850 observaciones.

	TRAMO - SEATS	**STL + HDoutliers**	**STL + iForest**	**STR + HDoutliers**	**STR + iForest**
Atípicos detectados	184	106	108	90	93
Max. número atip. serie	9	8	2	5	3
Series con 0 atípicos	12	42	11	43	13
Media de atípicos detec.	2.04	1.18	1.20	1.00	1.03

Cuadro 5.15: Resultados para el Conjunto 5 con datos hasta el primer trimestre de 2020.

	TRAMO - SEATS	**STL + HDoutliers**	**STL + iForest**	**STR + HDoutliers**	**STR + iForest**
Hasta 2019	184	122	105	94	97
Hasta mayo 2020	184	106	108	90	93
Tasa de variación	0.00 %	-13.11 %	2.86 %	-4.26 %	-4.12 %

Cuadro 5.16: Total de atípicos detectados por cada método hasta cuarto trimestre de 2019 y primer trimestre de 2020 y tasa de variación que produce el efecto del Covid en el Conjunto 5.

En el Cuadro 5.16 podemos ver que el efecto del Covid sobre las series trimestrales vuelve a ser muy pequeño, similar a las conclusiones obtenidas para el Conjunto 2. Si comparamos las tablas 5.14 y 5.15 podemos similitudes en los resultados para las diferentes medidas que se toman. Por ejemplo, si nos fijamos en la fila que mide el número de series que son identificadas con cero atípicos los métodos apenas difieren en sus resultados. Lo mismo sucede con la media de atípicos detectados en cada serie. Estos resultados sostienen la idea de que el Covid repercutió en menor medida en las series trimestrales en el primer trimestre del año, y es de interés realizar este mismo estudio una vez se incorporen los datos del segundo trimestre.

A continuación, vamos a analizar la serie referente al número de viviendas transmitidas de segunda mano en régimen libre para la provincia de Lugo, diferenciando los dos escenarios analizados a través de una recta vertical roja, los datos hasta el cuarto trimestre de 2019 y los datos incluyendo el primer trimestre de 2020.

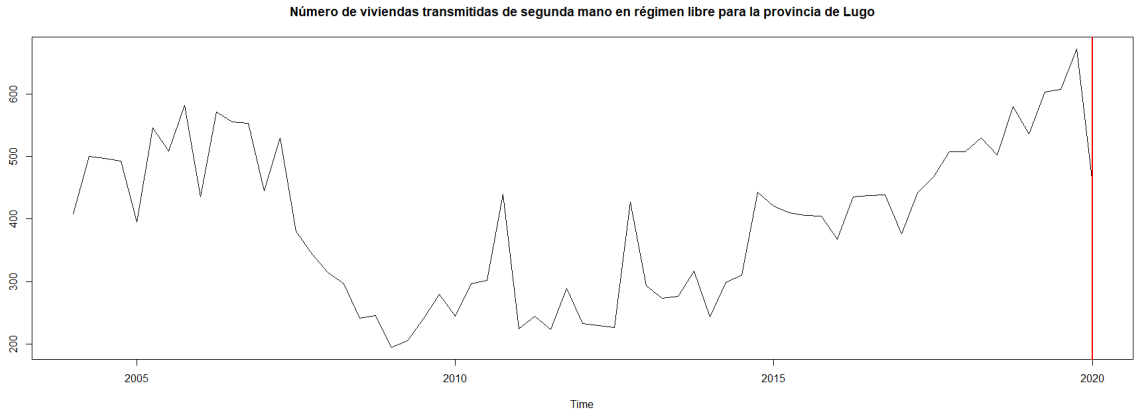

Figura 5.10: Serie de tiempo referente al número de viviendas transmitidas de segunda mano en régimen libre para la provincia de Lugo.

Observando la Figura 5.10, parece que el dato incluido referente al primer trimestre de 2020 supone una caída, una situación anómala, sin embargo, si observamos la evolución histórica de la serie podemos comprobar que solamente en dos ocasiones se han mostrado valores superiores en este primer trimestre. Para ambos escenarios, la serie hasta 2019 y la serie completa, únicamente un método detecta atípicos, *TRAMO-SEATS*, que señala en ambas ocasiones octubre de 2010 y 2012. El resto de métodos concuerdan en que la serie no presenta, en ningún caso, ninguna situación anómala.

La Figura 5.11 muestra las dos fechas que señala como atípicas el método *TRAMO-SEATS*, cuarto trimestre de 2010 y 2012, que si bien reflejan un pico en los espacios temporales en los que se producen, no parecen ser atípicas respecto a la serie completa.

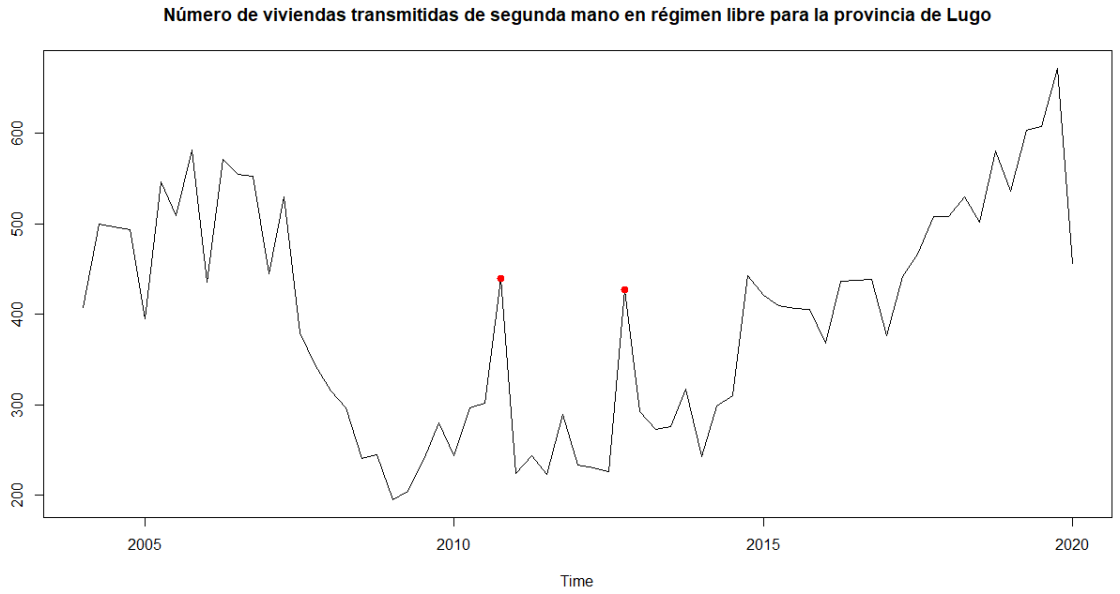

Figura 5.11: Serie de tiempo referente al número de viviendas transmitidas de segunda mano en régimen libre para la provincia de Lugo.

Conjunto 6 (`http://www.ige.eu/igebdt/igeapi/datos/1243`)

Series de datos mensuales referentes a las bajas de demandas de empleo según género y duración de la demanda en Galicia y sus provincias. Los datos de la serie se recogen desde enero de 2009 hasta junio de 2020.

	TRAMO - SEATS	**STL + HDoutliers**	**STL + iForest**	**STR + HDoutliers**	**STR + iForest**
Atípicos detectados	154	271	228	288	228
Max. número atip. serie	7	11	4	7	3
Series con 0 atípicos	85	44	16	40	12
Media de atípicos detec.	0.93	1.64	1.38	1.74	1.38

Cuadro 5.17: Resultados para el Conjunto 6 con datos hasta diciembre de 2019.

Los resultados del Cuadro 5.17 se producen para un total de 21780 observaciones que muestra el conjunto con datos hasta diciembre de 2019. Ahora vamos a estudiar los resultados de forma que cada serie esté conformada por el último dato actualizado y a comparar ambas situaciones. En esta serie, el último dato actualizado es junio de 2020, lo que conforma un total de 22770 observaciones.

	TRAMO - SEATS	**STL + HDoutliers**	**STL + iForest**	**STR + HDoutliers**	**STR + iForest**
Atípicos detectados	484	502	336	504	334
Max. número atip. serie	9	12	4	11	4
Series con 0 atípicos	1	10	3	15	5
Media de atípicos detec.	2.93	3.04	2.04	3.05	2.02

Cuadro 5.18: Resultados para el Conjunto 6 con datos hasta junio de 2020.

	TRAMO - SEATS	**STL + HDoutliers**	**STL + iForest**	**STR + HDoutliers**	**STR + iForest**
Hasta 2019	154	271	228	288	228
Hasta mayo 2020	484	502	336	504	334
Tasa de variación	214.29 %	85.24 %	47.37 %	75.00 %	46.49 %

Cuadro 5.19: Total de atípicos detectados por cada método hasta diciembre de 2019 y junio de 2020 y tasa de variación que produce el efecto del Covid en el Conjunto 6.

En el Cuadro 5.19 podemos ver el efecto que supone la primera mitad del año 2020 en la aparición de valores atípicos en este conjunto. El incremento en el número de datos a analizar no llega a ser del 5 %, pero esa mínima cantidad de datos nuevos dispara las fechas anómalas que se suceden en las series. En el Cuadro 5.18 se puede observar que apenas existen series sin atípicos, independientemente del método que seleccionemos, algo todavía más sorprendente si recordamos que este es el conjunto que posee más series, un total de 165 series.

A continuación analizamos un caso concreto utilizando la serie referente a las bajas de la demanda de empleo en mujeres con una duración de la demanda entre 12 y 18 meses en Ourense. La recta vertical roja indica el inicio del año 2020, sirviendo así para indicar la diferencia entre los dos escenarios analizados en los Cuadros 5.17 y 5.18.

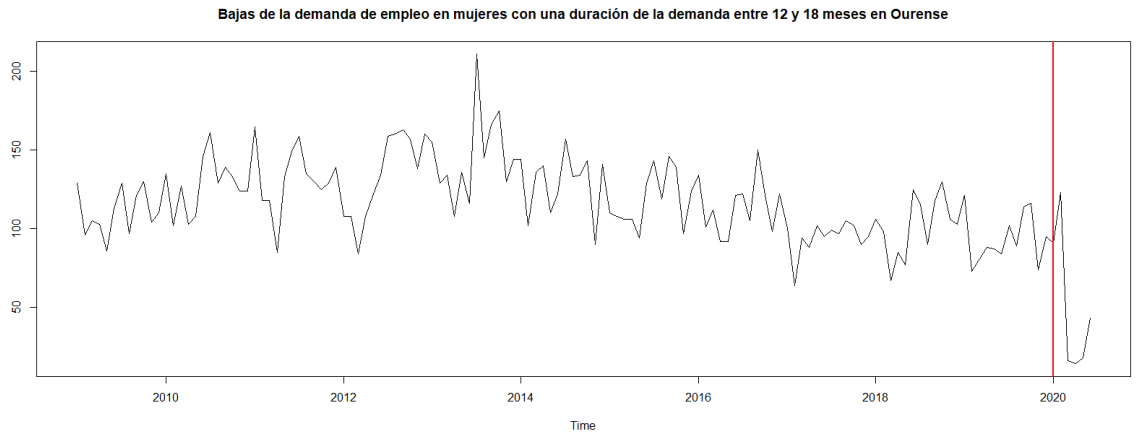

Figura 5.12: Serie de tiempo referente a las bajas de la demanda de empleo en mujeres con una duración de la demanda entre 12 y 18 meses en Ourense.

Esta serie es de interés por diversos motivos. El primero es que posee claros atípicos en ambos

escenarios. En la serie con duración hasta 2019 se puede observar un valor cercano al inicio de 2014 con claro aspecto de ser atípico. La serie con todos los datos muestra el atípico anterior pero también nuevos valores con claro comportamiento discordante al resto de la serie.

Para la serie recortada cinco métodos coinciden en sus resultados, julio de 2013 es atípico, únicamente *TRAMO-SEATS* difiere, el cual no señala ningún atípico en la serie. Si ahora nos trasladamos a la serie completa, *TRAMO-SEATS* señala marzo de 2020 como atípico, mientras que el resto de métodos vuelven a concordar señalando julio de 2013 y febrero de 2020. En la Figura 5.13 se muestra la serie señalizando dichos atípicos: julio de 2013, y febrero y marzo de 2020.

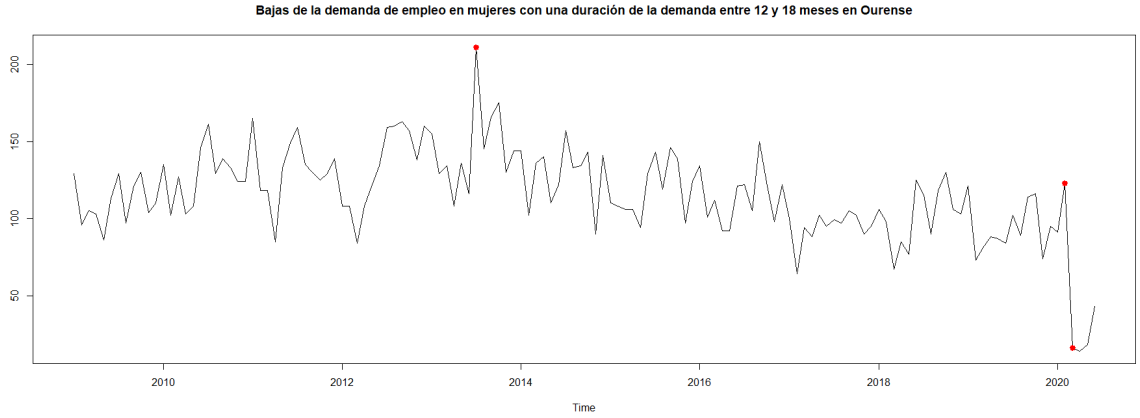

Figura 5.13: Serie de tiempo referente a las bajas de la demanda de empleo en mujeres con una duración de la demanda entre 12 y 18 meses en Ourense con los atípicos identificados.

Si observamos los datos de la serie podemos calificar todas estas fechas como atípicas: julio 2013, febrero 2020 y marzo 2020. Sin embargo, ningún método ha calificado los meses de abril, mayo y junio de 2020, siendo claramente atípicos. Estamos frente a una situación completamente inusual, una situación que se produce en multitud de series del IGE, y es la secuencia de varias observaciones consecutivas muy atípicas, que ejercen un gran peso en la serie.

Este gran peso provoca que los métodos no consigan analizar correctamente este escenario. Los modelos de series temporales como *TRAMO-SEATS* interpretan que se está produciendo un cambio de nivel permanente en la serie, mencionado en el Capítulo 3 como *Level Shift (LS)*, y modelizan la serie como tal. Los procesos de descomposición de series temporales utilizados, *STL* y *STR*, entienden lo que está sucediendo como un cambio en la tendencia, es decir, en su proceso de extracción de la estacionalidad y la tendencia estiman una tendencia decreciente en la serie. Y por último, los métodos de detección de valores atípicos, *HDoutliers* y *iForest*, sufren un grave problema de enmascaramiento. La idea del mecanismo de *iForest* reside en buscar valores aislados, mientras que la de *HDoutliers* se fundamenta en buscar espacios entre los valores ordenados, estas dos ideas son consistentes con la idea de la búsqueda de atípicos, sin embargo, ven disminuida su eficiencia ante una situación tan atípica que permanece durante varias observaciones consecutivas llevando a provocar que observaciones atípicas puedan pasar desapercibidas para el mecanismo.

Conjunto 7 (`http://www.ige.eu/igebdt/igeapi/datos/4885`)

Series de datos mensuales referentes a las afiliaciones a la seguridad social el último día de mes en Galicia y sus provincias por regímenes. Los datos comienzan en enero de 1990 y finalizan en mayo de 2020.

	TRAMO - SEATS	**STL + HDoutliers**	**STL + iForest**	**STR + HDoutliers**	**STR + iForest**
Atípicos detectados	288	96	118	136	121
Max. número atip. serie	24	8	5	10	5
Series con 0 atípicos	4	10	0	4	0
Media de atípicos detec.	6.13	2.04	2.51	2.89	2.57

Cuadro 5.20: Resultados para el Conjunto 7 con datos hasta diciembre de 2019.

El Cuadro 5.20 muestra los resultado que se producen para un total de 16920 observaciones que muestra el conjunto con datos hasta diciembre de 2019. Ahora vamos a estudiar los resultados de forma que cada serie esté conformada por el último dato actualizado y comparar ambas situaciones. En esta serie, el último dato actualizado es junio de 2020, lo que conforma un total de 17155 observaciones.

	TRAMO - SEATS	**STL + HDoutliers**	**STL + iForest**	**STR + HDoutliers**	**STR + iForest**
Atípicos detectados	333	179	128	118	136
Max. número atip. serie	24	14	5	11	6
Series con 0 atípicos	0	6	0	3	0
Media de atípicos detec.	7.09	3.80	2.72	2.51	2.89

Cuadro 5.21: Resultados para el Conjunto 7 con datos hasta mayo de 2020.

	TRAMO - SEATS	**STL + HDoutliers**	**STL + iForest**	**STR + HDoutliers**	**STR + iForest**
Hasta 2019	288	96	118	136	121
Hasta mayo 2020	333	179	128	118	136
Tasa de variación	15.63 %	86.46 %	8.47 %	-13.24 %	12.40 %

Cuadro 5.22: Total de atípicos detectados por cada método hasta diciembre de 2019 y mayo de 2020 y tasa de variación que produce el efecto del Covid en el Conjunto 7.

En el Cuadro 5.22 podemos ver que el efecto que supone la primera mitad del año 2020 en la aparición de valores atípicos en este conjunto. La variación que supone en el número de valores detectados como anómalos es considerable, todavía más si lo comparamos con la variación de nuevos datos introducidos en el análisis, los cuales han aumentado solamente un 1.38 %. De este modo podemos afirmar que la presencia del Covid ha provocado un gran impacto en el incremento del número de observaciones atípicas de los conjuntos que presentan datos de frecuencia mensual.

A continuación analizamos un caso concreto utilizando la serie referente a las afiliaciones a la seguridad social en régimen mar para la provincia de Lugo. La Figura 5.14 muestra la representación gráfica de dicha serie, indicando mediante una recta vertical roja el inicio del 2020.

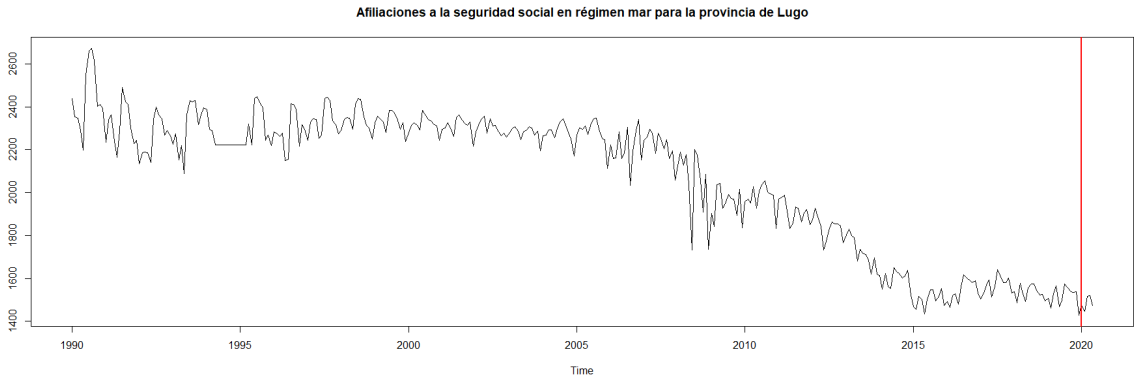

Figura 5.14: Serie de tiempo referente a las afiliaciones a la seguridad social en régimen mar para la provincia de Lugo.

El motivo de seleccionar esta serie reside en una idea que se va desarrollando a lo largo de todo este apartado, y que da lugar a la herramienta que se introduce en el siguiente. Las series socioeconómicas son complejas dado que se ven afectadas por multitud de factores externos que condicionan su desarrollo. Este carácter único de cada serie, más los resultados obtenidos a lo largo del trabajo,

nos conduce a la conclusión de que no existe un método de detección de atípicos con una capacidad de clasificación superior al resto.

Los resultados para esta serie son diversos. Por ejemplo, la combinación *STR+HDoutliers* muestra hasta ocho atípicos para la serie delimitada en el año 2019, sin embargo, al introducir la serie completa esta cifra se reduce solamente a dos. Sucede lo contrario con la combinación *STL+HDoutliers*, que si bien en la serie recortada califica dos observaciones en la completa señala un total de siete. Por su parte, *TRAMO-SEATS* identifica un total de siete atípicos. Sin embargo todos los métodos de detección señalan dos fechas como atípicas, en los dos casos, agosto de 2006 y junio de 2008, como se representa en la Figura 5.15. Esto dota al estudio de confianza, si todos los mecanismos sospechan de dos fechas deberá haber un motivo de peso, algo ha sucedido en la serie para que todos concuerden. Esta idea dio lugar al desarrollo de la herramienta que se introduce en el siguiente apartado.

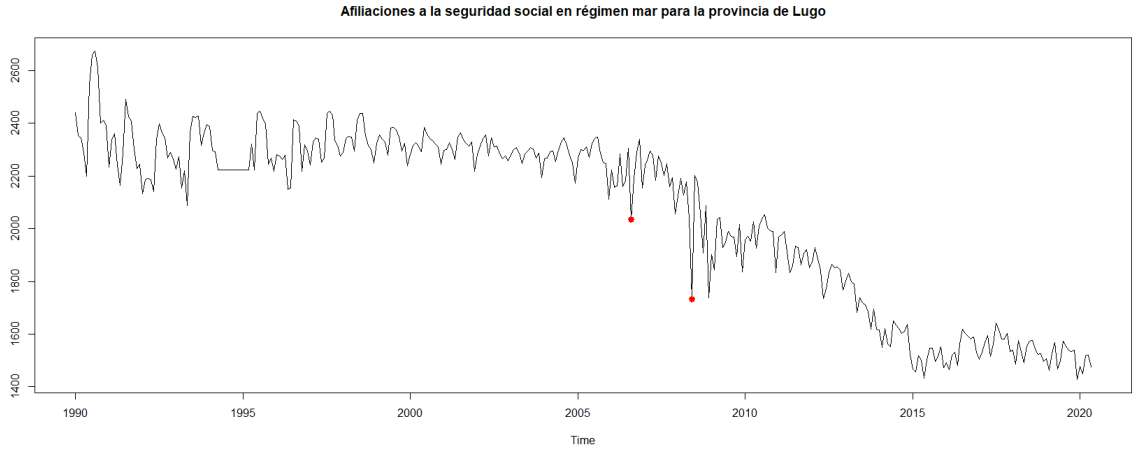

Figura 5.15: Serie de tiempo referente a las afiliaciones a la seguridad social en régimen mar para la provincia de Lugo identificando los atípicos señalados de forma común por los cinco métodos de detección.

Por último, el Cuadro 5.23 recoge el total de atípicos detectados por cada método en los dos escenarios con los que se ha trabajado en este capítulo. El primer escenario incluye datos hasta la última fecha de 2019, si las series son mensuales este dato es el correspondiente a diciembre de 2019, si son series trimestrales el último dato será el cuarto trimestre de 2019. En el Cuadro 5.23 se recoge de modo general para todos los conjuntos bajo el término: "datos hasta 2019". El segundo escenario recoge los datos hasta su última actualización, aquí existe un mayor abanico de fechas como vimos a lo largo de todo este capítulo. Los conjuntos con datos de frecuencia trimestral presentan su último dato en el primer trimestre de 2020, por su parte, los conjuntos de datos mensuales varían su última actualización entre los meses de abril, mayo o junio de 2020. En el Cuadro 5.23 se recoge de modo general para todos ellos bajo el término: "datos hasta 2020".

	TRAMO - SEATS	**STL + HDoutliers**	**STL + iForest**	**STR + HDoutliers**	**STR + iForest**
Tiempo de ejecución (2019)	133.2	2.4	178.2	723.4	880.5
Tiempo de ejecución (2020)	150.3	2.1	204.2	804.3	1076.4
Atípicos detectados (2019)	1490	990	794	1074	816
Atípicos detectados (2020)	2036	1373	989	1341	1033
Tasa de variación	36.64 %	38.69 %	24.56 %	24.86 %	26.59 %

Cuadro 5.23: Tiempo de ejecución (en segundos) y total de atípicos detectados por cada método hasta el último dato de 2019 y hasta última actualización de 2020, y tasa de variación que produce el efecto del Covid en los siete conjuntos de datos.

El Cuadro 5.23 nos permite concluir que el efecto global del Covid ha supuesto un incremento en el número de observaciones atípicas puesto que todos los métodos experimentan una tasa de variación positiva sobre sus cifras de valores atípicos detectados. Las cifras de tasas de variación que se muestran en el Cuadro 5.23 sorprenden todavía más si analizamos el incremento en el número de observaciones analizadas. El escenario que analiza el conjunto de datos hasta 2019 presentan un total de 90540 observaciones, mientras que el escenario que incluye hasta el último dato actualizado recoge un total de 92731. Es decir, el número total de datos analizados se ha incrementado en poco más del 2 %, sin embargo, la tasa de variación más pequeña muestra un incremento del más del 24 % del número de atípicos que se presentan en el conjunto de datos.

Otra medida que incluye el Cuadro 5.23 es el tiempo que tarda en analizar los siete conjuntos de datos cada método, recogida bajo el nombre de "tiempo de ejecución". A través de ella podemos observar que la diferencia entre el tiempo de ejecución es pequeña, pero existe, por lo que muestra que la presencia de atípicos dificulta la modelización de la series temporales. También podemos comprobar como todos los métodos presentan buenas cifras en este medida, los que peores resultados muestran son los métodos que utilizan la descomposición *STR*, la cual ya mencionamos su posible problema de lentitud en el Capítulo 2. Por último, es sorprendente la combinación de *STL+HDoutliers*, que es capaz de analizar las 538 en un tiempo inferior a los tres segundos en ambos casos.

5.2. Análisis Gráfico

El campo de la detección de valores atípicos está caracterizado por una constante incertidumbre. Partiendo del inicio del trabajo, ni siquiera existe una definición exacta de qué es un valor atípico, sino más bien se trazan definiciones en base a ideas de lo que se espera que sea un valor atípico en un conjunto de datos. Es por ello que, dependiendo de los escenarios y las complejidades asociadas a una serie, unos métodos resulten más efectivos que otros. Este es el motivo por el que surgió la idea de crear una herramienta de análisis gráfico para el estudio de series de tiempo.

El objetivo de esta herramienta es la de dotar al analista de un mecanismo sencillo de implantar que le permita trazar una idea de los posibles atípicos con los que podría contar la serie. Para ello se ha creado una función en `R` que solo necesita introducir la serie de tiempo para devolver el gráfico.

Además, dado que esta herramienta está diseñada específicamente para trabajar con series de tiempo, su representación daría lugar a lo que podría denominarse un mapa de calor para zonas atípicas, esto se muestra más adelante mediante ejemplos.

Esta herramienta está compuesta por los métodos que han mostrado mejores resultados en el estudio de simulación. Cada método ocupa una fila, mostrando a lo largo de ella sus atípicos señalados, y cada columna es una fecha detectada por un método. De tal modo que, si una fecha es detectada por todos los métodos mostrará una columna verde y será sospecha clara de que esa fecha es atípica. Por otra parte, si una fecha es detectada por solo un método mostrará solo de color verde la celda asociada a ese método y no será señal tan evidente de situación atípica.

Para la composición del gráfico se utilizan los métodos del apartado anterior, y a mayores se introducen las versiones robustas de los métodos de descomposición de series temporales, *STL* y *STR*. El motivo de su uso en este apartado se debe a que, si bien no mostraron los mejores resultados en el estudio de simulación, ahora pueden resultar de utilidad. Los métodos robustos pueden generar anomalías espurias cuando no existen atípicos debido a que resaltan el tamaño del residuo de una observación provocando su identificación como atípico sin esta serlo. Sin embargo, cuando se producen anomalías que ejercen un gran peso en la serie, como lo sucedido durante el Covid, este tipo de estimaciones son de gran interés, esta situación también se justificará mediante un ejemplo práctico.

El primer ejemplo lo extraemos a través de la serie referente a la estancia media de viajeros de todas las procedencias en establecimientos hoteleros en Galicia, perteneciente al Conjunto 1. En la Figura 5.16 mostramos la representación de la misma, ahora ya mostramos hasta el último dato actualizado, en este caso, mayo de 2020.

Figura 5.16: Estancia media de viajeros de todas las procedencias en establecimientos hoteleros en Galicia.

Esta serie parece contar con dos atípicos, los relacionados con las dos últimas observaciones de la serie, sin embargo, vamos a analizar los resultados del gráfico comparativo.

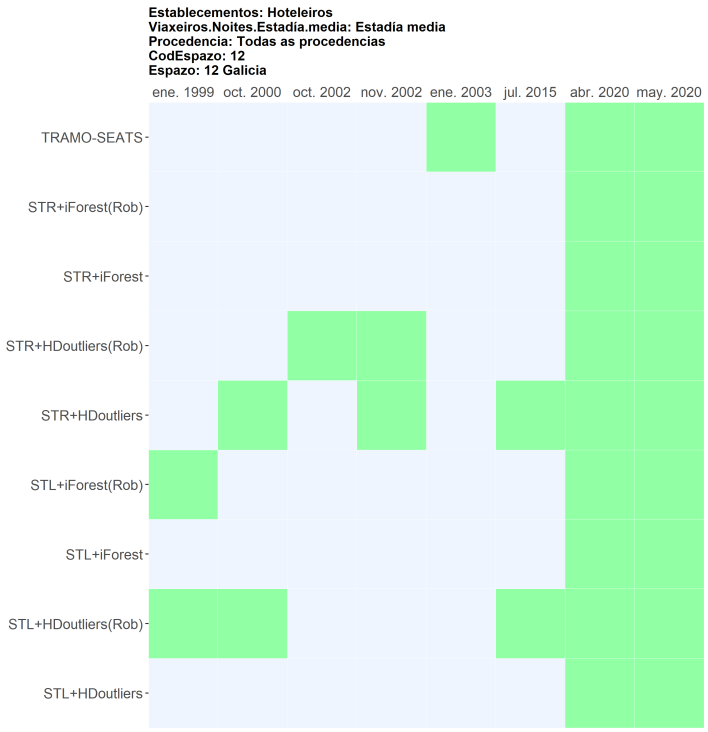

Figura 5.17: Resultados para la serie estancia media de viajeros de todas las procedencias en establecimientos hoteleros en Galicia del Conjunto 1.

La Figura 5.17 muestra el gráfico que hemos desarrollado, el cual vamos a desgranar ahora para entender su utilidad. Si hubiésemos escogido el método *TRAMO-SEATS* hubiésemos señalado tres observaciones atípicas, una de ellas sería enero de 2003. Sin embargo, analizando los datos podemos comprobar que este dato parece normal y estaríamos identificándolo erróneamente como anómalo. De hecho, solo *TRAMO-SEATS* señala esta fecha. Lo mismo sucede con otras fechas como octubre del año 2000, señalada por solo dos métodos. Por otra parte, todos los métodos señalan los meses de abril y mayo de 2020, las dos últimas observaciones de la serie que en la Figura 5.16 mostraban un comportamiento totalmente discordante al resto del conjunto de datos. Es decir, el gráfico permite trazar una idea concisa al analista de qué es anómalo en la serie y que pueden ser falsos positivos.

Observando el gráfico podemos también extraer una conclusión rápida de la posible presencia de valores atípicos en la serie. Se muestran dos comportamientos muy marcados. El primero estaría determinado por la serie hasta el año 2020, en el cual la aparición de atípicos es dispersa, poco concluyente, el analista puede dudar de que estas observaciones sean realmente anómalas. El segundo son los datos de 2020, dos columnas enteras de color verde, una imagen que comunica que en ese intervalo de tiempo está sucediendo algo realmente anómalo.

Vamos a utilizar el gráfico con otra serie, para analizar otra de las características que muestran el interés de analizar una serie de tiempo a través de este gráfico. En la Figura 5.18 se muestra la serie de tiempo perteneciente al Conjunto 3 que recoge la evolución de los contratos iniciales en Galicia. Dicha serie cuenta con tres atípicos claros: marzo, abril y mayo de 2020.

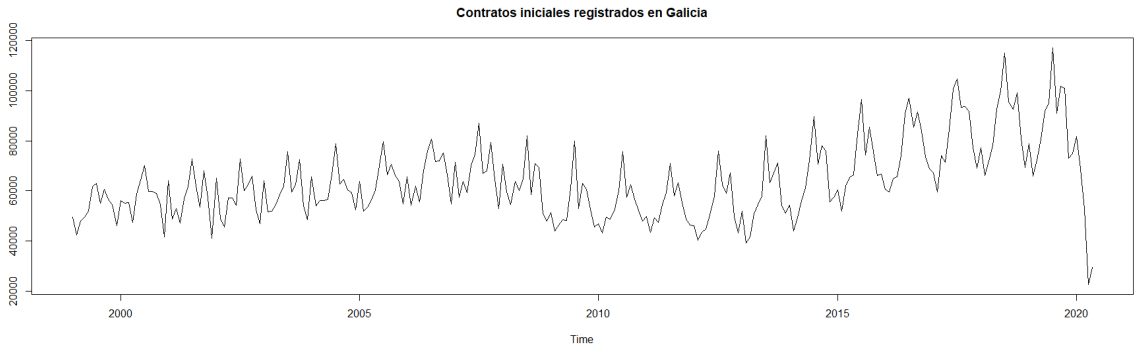

Figura 5.18: Contratos iniciales registrados en Galicia.

Haciendo uso del gráfico comparativo, Figura 5.19, volvemos a obtener un escenario en el que resulta de interés esta herramienta. En este caso, el método compuesto por *STR+HDoutliers* muestra un comportamiento errático, que le conduce a identificar una gran cantidad de atípicos de modo incorrecto. Lo mismo sucede con *TRAMO-SEATS*, que es el único que señala enero de 2001 como atípico, sin existir motivo en los datos para esta calificación.

Por otra parte, este escenario nos permite observar la utilidad de incorporar las versiones robustas. Si bien se equivocan en algún caso, por ejemplo el método *STR+HDoutliers(Rob)* señalando junio de 2002, nos permiten calificar marzo de 2020 como atípico, algo que los métodos que utilizan las versiones no robustas de los métodos de descomposición de series temporales, *STR* y *STR*, no logran. Todas ellas se ponen de acuerdo junto a *TRAMO-SEATS* para la señalización de este mes.

Además, dado que estamos trabajando con series temporales los resultados de la gráfica se muestran en orden cronológico, en vez de por cuán anómala es la observación como se hace en *TRAMO-SEATS* o en *iForest*. Esto nos permite entender la gráfica como un mapa de calor. Al analizar la serie, si existe un intervalo de tiempo en el que se han producido situaciones anómalas, como por ejemplo en este caso los meses relacionados al Covid, se verá reflejado en el gráfico por una serie de columnas con muchas celdas verdes, lo que dará a entender de manera rápida al analista que en ese intervalo de tiempo se han producido situaciones excepcionales. De este modo también conseguimos corregir el problema de enmascaramiento que pueden sufrir los métodos cuando se suceden varias situaciones atípicas consecutivas.

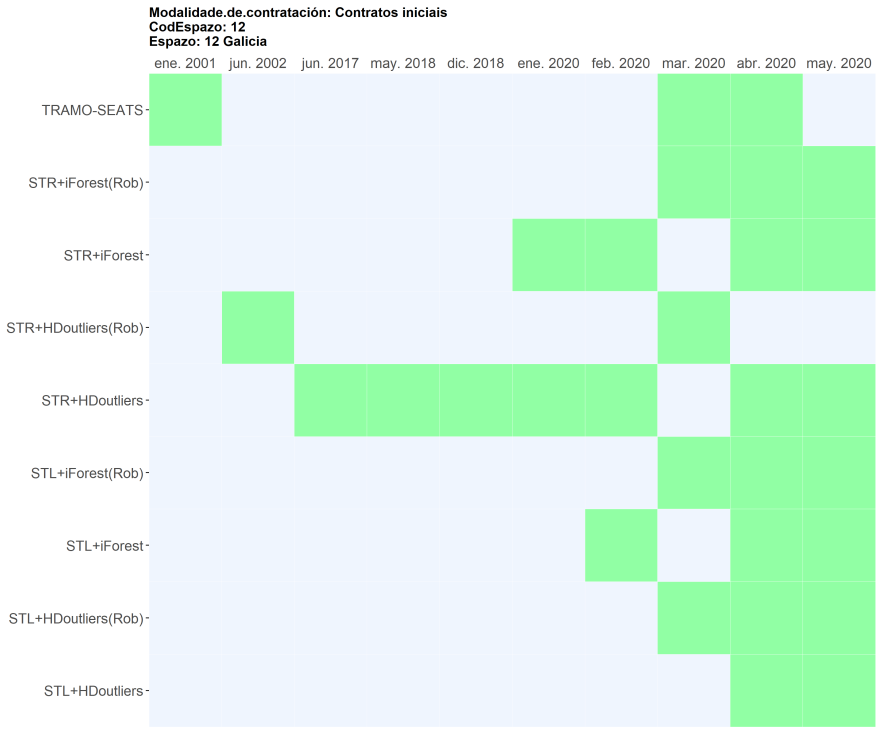

Figura 5.19: Resultados para la serie Contratos iniciales registrados en Galicia del Conjunto 3.

Capítulo 6

Conclusiones

El IGE cuenta con una amplía base de datos que se utiliza para el estudio del contexto socioeconómico en Galicia. Estos datos tienen una base temporal, es decir, son datos referidos a un intervalo de tiempo, en la mayoría de los casos esta periodicidad es anual, trimestral o mensual. La manipulación de los datos puede conllevar a que se introduzcan errores humanos o de medición que provoquen que se incluyan cifras incorrectas a las bases de datos. Estos errores pueden distorsionar las conclusiones de un estudio o conducir hacía resultados alejados de la verdadera realidad. Es por ello que resulta necesario el incorporar una herramienta de detección de valores atípicos al análisis de las bases de datos del IGE.

Para ello se planteó una revisión bibliográfica de los principales métodos de detección de atípicos en series temporales. Esta revisión nos permitió conocer la existencia de un amplio abanico de métodos disponibles con los cuales empezamos a trabajar para analizar su comportamiento. Es aquí donde decidimos descartar una serie de métodos, mencionados al final del Apartado 3.4, debido a que su funcionamiento y sus resultados correspondían a versiones inferiores de otros métodos que finalmente adoptamos. Este primer estudio nos proporcionó los métodos de detección de atípicos en series temporales que supondrían el punto de partida del trabajo: *X-13ARIMA-SEATS*, *TRAMO-SEATS* y la librería de `R` `Anomalize`.

La librería `Anomalize` también nos sirvió de ayuda en el desarrollo de nuestros propios métodos de detección de atípicos para series temporales. La idea sobre la que se fundamenta esta librería es la de aplicar un proceso de descomposición de series de tiempo a una serie para extraer las componentes de la estacionalidad, la tendencia y un residuo, para después aplicar un método de detección de atípicos sobre el residuo. Dado que se han producido mejoras a los métodos de descomposición y detección que se utilizan en la librería `Anomalize` consideramos que una nueva combinación podría reflejar una mejora en los resultados.

Para comparar todos los métodos de detección de atípicos en series temporales, los que sirven de punto de partida del trabajo y los desarrollados por nosotros, decidimos llevar a cabo un ambicioso estudio de simulación. El estudio de simulación recoge diferentes escenarios, con la finalidad de representar la heterogeneidad propia de las series de tiempo referentes a variables sociales y económicas. Los resultados del estudio de simulación confirman nuestra idea acerca de que una nueva combinación puede suponer una mejora en los resultados debido a que varias de nuestras combinaciones reflejan mejores

resultados que las propuestas en la librería `Anomalize`. También nos permite observar que nuestras propuestas compiten en los resultados con un modelo tan desarrollado como *TRAMO-SEATS*. Finalmente, nos permiten descartar otros métodos, que, siendo útiles en otros campos, no parecen resultar idóneos para el nuestro.

Tras el estudio de simulación comenzamos a utilizar los métodos que habían mostrado mejores resultados con los datos proporcionados por el IGE. Fue en este momento cuando nos encontramos con una situación no contemplada en ningún momento, en ningún escenario del estudio de simulación, y en general en ningún momento de lo que podría ser la historia reciente, como fue la crisis del Covid19. Esta crisis provoca un efecto enorme en los datos del contexto gallego como se recoge a lo largo del Capítulo 5, y supone un nuevo escenario en lo que a detección de valores atípicos en series temporales se refiere.

Este nuevo escenario se caracteriza por reflejar durante varios períodos de tiempo consecutivos observaciones enormemente anómalas y los modelos ven empeorado su comportamiento. *TRAMO-SEATS* y los modelos de descomposición de series de tiempo, *STL* y *STR* entienden que lo que se está produciendo es una tendencia decreciente de la serie. Por su parte, los métodos de detección de atípicos, *HDoutliers* y *iForest*, sufren un fenómeno de enmascaramiento en el que son incapaces de señalizar todos los atípicos. Fue aquí donde decidimos crear una nueva herramienta para el análisis.

El objetivo con el que creamos esta herramienta de análisis gráfico fue el de reducir la incertidumbre que supone el campo de la detección de atípicos, de un modo sencillo y directo para el analista que la utilice. Además, su diseño está pensado para el trabajo con series temporales, de modo que si se sucede un período convulso como el actual, la gráfica podrá ser analizada como un mapa de calor, un mapa de calor que transmite visualmente la idea de en qué intervalo de tiempo la serie está experimentando comportamientos atípicos.

Apéndice A

Tablas

A.1. Tablas de sensibilidad

	Caso 1	**Caso 2**	**Caso 3**	**Caso 4**
X-13ARIMA-SEATS	0.496	0.854	0.700	0.826
TRAMO-SEATS	0.657	0.928	0.805	0.851
STL+HDoutliers	0.569	0.838	0.570	0.732
STL+GESD	0.266	0.590	0.236	0.532
STL+iForest	0.636	0.885	0.636	0.790
STL(ROB)+HDoutliers	0.562	0.803	0.678	0.712
STL(ROB)+GESD	0.637	0.876	0.746	0.754
STL(ROB)+iForest	0.640	0.878	0.747	0.786
Twitter+HDoutliers	0.471	0.735	0.597	0.620
Twitter+GESD	0.306	0.590	0.429	0.434
Twitter+iForest	0.529	0.801	0.652	0.660
STR+HDoutliers	0.611	0.841	0.603	0.672
STR+GESD	0.345	0.650	0.292	0.448
STR+iForest	0.674	0.885	0.679	0.721
STR(ROB)+HDoutliers	0.542	0.812	0.638	0.658
STR(ROB)+GESD	0.381	0.685	0.458	0.558
STR(ROB)+iForest	0.623	0.862	0.699	0.714

Cuadro A.1: Resultados Escenario 1.Sensibilidad. AR(1).

	Caso 1	**Caso 2**	**Caso 3**	**Caso 4**
X-13ARIMA-SEATS	0.505	0.829	0.646	0.875
TRAMO-SEATS	0.668	0.914	0.743	0.888
STL+HDoutliers	0.472	0.729	0.481	0.684
STL+GESD	0.217	0.457	0.148	0.428
STL+iForest	0.506	0.784	0.525	0.716
STL(ROB)+HDoutliers	0.469	0.735	0.588	0.648
STL(ROB)+GESD	0.529	0.775	0.665	0.668
STL(ROB)+iForest	0.526	0.803	0.666	0.716
Twitter+HDoutliers	0.465	0.698	0.596	0.646
Twitter+GESD	0.349	0.630	0.483	0.486
Twitter+iForest	0.512	0.778	0.663	0.692
STR+HDoutliers	0.588	0.823	0.573	0.664
STR+GESD	0.298	0.602	0.247	0.449
STR+iForest	0.630	0.869	0.651	0.721
STR(ROB)+HDoutliers	0.521	0.788	0.611	0.649
STR(ROB)+GESD	0.336	0.628	0.406	0.520
STR(ROB)+iForest	0.582	0.839	0.691	0.711

Cuadro A.2: Resultados Escenario 2. Sensibilidad. MA(1).

	Caso 1	**Caso 2**	**Caso 3**	**Caso 4**
X-13ARIMA-SEATS	0.922	0.971	0.927	0.983
TRAMO-SEATS	0.923	0.972	0.923	0.982
STL+HDoutliers	0.312	0.508	0.241	0.440
STL+GESD	0.089	0.230	0.051	0.152
STL+iForest	0.343	0.578	0.284	0.454
STL(ROB)+HDoutliers	0.261	0.435	0.297	0.333
STL(ROB)+GESD	0.311	0.527	0.393	0.322
STL(ROB)+iForest	0.298	0.501	0.336	0.362
Twitter+HDoutliers	0.064	0.112	0.127	0.066
Twitter+GESD	0.028	0.040	0.069	0.032
Twitter+iForest	0.079	0.129	0.169	0.070
STR+HDoutliers	0.407	0.563	0.348	0.440
STR+GESD	0.204	0.367	0.109	0.230
STR+iForest	0.445	0.617	0.389	0.495
STR(ROB)+HDoutliers	0.243	0.397	0.254	0.272
STR(ROB)+GESD	0.209	0.391	0.225	0.259
STR(ROB)+iForest	0.282	0.455	0.284	0.293

Cuadro A.3: Resultados Escenario 3. Sensibilidad. ARIMA(0,1,1).

	Caso 1	**Caso 2**	**Caso 3**	**Caso 4**
X-13ARIMA-SEATS	0.403	0.593	0.068	0.816
TRAMO-SEATS	0.542	0.707	0.072	0.814
STL+HDoutliers	0.005	0.004	0.016	0.004
STL+GESD	0.000	0.001	0.004	0.000
STL+iForest	0.003	0.003	0.015	0.005
STL(ROB)+HDoutliers	0.007	0.007	0.037	0.004
STL(ROB)+GESD	0.009	0.021	0.099	0.029
STL(ROB)+iForest	0.006	0.008	0.036	0.005
Twitter+HDoutliers	0.001	0.000	0.001	0.008
Twitter+GESD	0.006	0.009	0.012	0.023
Twitter+iForest	0.001	0.000	0.005	0.018
STR+HDoutliers	0.370	0.428	0.006	0.088
STR+GESD	0.424	0.462	0.000	0.138
STR+iForest	0.418	0.479	0.004	0.103
STR(ROB)+HDoutliers	0.028	0.019	0.011	0.076
STR(ROB)+GESD	0.064	0.063	0.051	0.140
STR(ROB)+iForest	0.022	0.014	0.010	0.027

Cuadro A.4: Resultados Escenario 4. Sensibilidad. ARIMA(0,1,1)x(0,1,1)12.

	Caso 1	**Caso 2**	**Caso 3**	**Caso 4**
X-13ARIMA-SEATS	0.266	0.629	0.177	0.574
TRAMO-SEATS	0.400	0.750	0.251	0.618
STL+HDoutliers	0.247	0.477	0.291	0.130
STL+GESD	0.093	0.260	0.160	0.024
STL+iForest	0.270	0.536	0.312	0.138
STL(ROB)+HDoutliers	0.278	0.489	0.358	0.102
STL(ROB)+GESD	0.400	0.688	0.478	0.132
STL(ROB)+iForest	0.314	0.553	0.398	0.133
Twitter+HDoutliers	0.254	0.439	0.363	0.108
Twitter+GESD	0.176	0.385	0.317	0.064
Twitter+iForest	0.274	0.486	0.388	0.131
STR+HDoutliers	0.664	0.895	0.240	0.760
STR+GESD	0.425	0.770	0.065	0.567
STR+iForest	0.724	0.939	0.283	0.800
STR(ROB)+HDoutliers	0.197	0.334	0.220	0.529
STR(ROB)+GESD	0.190	0.333	0.207	0.692
STR(ROB)+iForest	0.235	0.369	0.238	0.591

Cuadro A.5: Resultados Escenario 5. Sensibilidad. ARIMA(1,0,1)x(0,1,2)12.

	Caso 1	**Caso 2**	**Caso 3**	**Caso 4**
X-13ARIMA-SEATS	0.333	0.662	0.355	0.618
TRAMO-SEATS	0.460	0.767	0.480	0.656
STL+HDoutliers	0.521	0.798	0.502	0.629
STL+GESD	0.230	0.559	0.188	0.328
STL+iForest	0.565	0.836	0.572	0.651
STL(ROB)+HDoutliers	0.519	0.788	0.605	0.601
STL(ROB)+GESD	0.541	0.828	0.676	0.561
STL(ROB)+iForest	0.566	0.843	0.668	0.648
Twitter+HDoutliers	0.246	0.423	0.362	0.212
Twitter+GESD	0.104	0.232	0.187	0.102
Twitter+iForest	0.284	0.456	0.398	0.230
STR+HDoutliers	0.527	0.788	0.444	0.641
STR+GESD	0.279	0.585	0.152	0.386
STR+iForest	0.600	0.825	0.499	0.685
STR(ROB)+HDoutliers	0.407	0.674	0.497	0.550
STR(ROB)+GESD	0.277	0.527	0.352	0.404
STR(ROB)+iForest	0.462	0.735	0.573	0.582

Cuadro A.6: Resultados Escenario 6. Sensibilidad. ARIMA(1,1,1)x(1,0,1)12.

	Caso 1	**Caso 2**	**Caso 3**	**Caso 4**
X-13ARIMA-SEATS	0.299	0.638	0.335	0.632
TRAMO-SEATS	0.424	0.717	0.432	0.801
STL+HDoutliers	0.432	0.705	0.397	0.626
STL+GESD	0.177	0.416	0.109	0.351
STL+iForest	0.514	0.762	0.463	0.530
STL(ROB)+HDoutliers	0.441	0.686	0.494	0.613
STL(ROB)+GESD	0.461	0.731	0.556	0.584
STL(ROB)+iForest	0.491	0.723	0.548	0.528
Twitter+HDoutliers	0.280	0.471	0.387	0.354
Twitter+GESD	0.129	0.288	0.191	0.147
Twitter+iForest	0.008	0.017	0.051	0.050
STR+HDoutliers	0.456	0.700	0.467	0.568
STR+GESD	0.203	0.439	0.167	0.274
STR+iForest	0.515	0.778	0.443	0.638
STR(ROB)+HDoutliers	0.424	0.687	0.510	0.554
STR(ROB)+GESD	0.262	0.524	0.330	0.366
STR(ROB)+iForest	0.349	0.562	0.413	0.388

Cuadro A.7: Resultados Escenario 7. Sensibilidad. ARIMA(1,1,2)x(0,1,1)12.

	Caso 1	**Caso 2**	**Caso 3**	**Caso 4**
X-13ARIMA-SEATS	0.299	0.638	0.335	0.632
TRAMO-SEATS	0.424	0.717	0.432	0.801
STL+HDoutliers	0.432	0.705	0.397	0.626
STL+GESD	0.177	0.416	0.109	0.351
STL+iForest	0.485	0.738	0.428	0.665
STL(ROB)+HDoutliers	0.441	0.686	0.494	0.613
STL(ROB)+GESD	0.461	0.731	0.556	0.584
STL(ROB)+iForest	0.494	0.755	0.566	0.678
Twitter+HDoutliers	0.280	0.471	0.387	0.354
Twitter+GESD	0.129	0.288	0.191	0.147
Twitter+iForest	0.306	0.509	0.430	0.359
STR+HDoutliers	0.456	0.700	0.467	0.568
STR+GESD	0.203	0.439	0.167	0.274
STR+iForest	0.497	0.772	0.507	0.602
STR(ROB)+HDoutliers	0.424	0.687	0.510	0.554
STR(ROB)+GESD	0.262	0.524	0.330	0.366
STR(ROB)+iForest	0.465	0.755	0.577	0.590

Cuadro A.8: Resultados Escenario 8. Sensibilidad. ARIMA(0,1,1)x(1,0,0)12.

A.2. Tablas de exceso

	Caso 1	Caso 2	Caso 3	Caso 4
X-13ARIMA-SEATS	0.078	0.077	0.066	0.042
TRAMO-SEATS	0.211	0.208	0.222	0.056
STL+HDoutliers	0.788	0.829	0.805	1.534
STL+GESD	0.056	0.055	0.045	0.031
STL+iForest	0.406	0.300	0.404	0.659
STL(ROB)+HDoutliers	1.162	1.242	1.212	1.714
STL(ROB)+GESD	1.411	1.417	1.416	0.461
STL(ROB)+iForest	0.637	0.468	0.575	0.859
Twitter+HDoutliers	0.930	0.955	0.905	1.487
Twitter+GESD	0.257	0.273	0.254	0.069
Twitter+iForest	0.538	0.412	0.490	0.776
STR+HDoutliers	0.759	0.810	0.759	1.464
STR+GESD	0.053	0.062	0.046	0.024
STR+iForest	0.403	0.301	0.381	0.725
STR(ROB)+HDoutliers	0.749	0.842	0.758	1.460
STR(ROB)+GESD	0.199	0.196	0.192	0.099
STR(ROB)+iForest	0.427	0.315	0.413	0.749

Cuadro A.9: Resultados Escenario 1. Exceso. AR(1).

	Caso 1	Caso 2	Caso 3	Caso 4
X-13ARIMA-SEATS	0.054	0.058	0.066	0.058
TRAMO-SEATS	0.144	0.138	0.175	0.061
STL+HDoutliers	0.695	0.772	0.759	1.472
STL+GESD	0.034	0.038	0.050	0.016
STL+iForest	0.443	0.318	0.425	0.706
STL(ROB)+HDoutliers	1.168	1.221	1.180	1.703
STL(ROB)+GESD	1.309	1.349	1.415	0.418
STL(ROB)+iForest	0.682	0.528	0.621	0.877
Twitter+HDoutliers	0.836	0.832	0.911	1.577
Twitter+GESD	0.313	0.329	0.334	0.061
Twitter+iForest	0.552	0.417	0.483	0.755
STR+HDoutliers	0.798	0.885	0.694	1.541
STR+GESD	0.054	0.064	0.044	0.018
STR+iForest	0.407	0.302	0.396	0.724
STR(ROB)+HDoutliers	0.725	0.776	0.745	1.516
STR(ROB)+GESD	0.134	0.143	0.142	0.076
STR(ROB)+iForest	0.435	0.312	0.372	0.727

Cuadro A.10: Resultados Escenario 2. Exceso. MA(1).

	Caso 1	**Caso 2**	**Caso 3**	**Caso 4**
X-13ARIMA-SEATS	0.106	0.060	0.101	0.070
TRAMO-SEATS	0.179	0.123	0.175	0.085
STL+HDoutliers	0.638	0.635	0.686	1.376
STL+GESD	0.052	0.054	0.067	0.020
STL+iForest	0.448	0.346	0.468	0.867
STL(ROB)+HDoutliers	0.932	0.905	1.010	1.604
STL(ROB)+GESD	2.339	2.311	2.338	1.863
STL(ROB)+iForest	0.705	0.573	0.695	1.281
Twitter+HDoutliers	0.311	0.285	0.311	0.743
Twitter+GESD	1.302	1.282	1.331	1.946
Twitter+iForest	0.167	0.153	0.175	0.645
STR+HDoutliers	0.672	0.703	0.623	1.364
STR+GESD	0.180	0.183	0.119	0.050
STR+iForest	0.372	0.297	0.393	0.899
STR(ROB)+HDoutliers	0.598	0.568	0.588	1.225
STR(ROB)+GESD	1.030	1.008	0.999	1.341
STR(ROB)+iForest	0.463	0.383	0.464	1.108

Cuadro A.11: Resultados Escenario 3. Exceso. ARIMA(0,1,1).

	Caso 1	Caso 2	Caso 3	Caso 4
X-13ARIMA-SEATS	0.132	0.120	0.062	0.274
TRAMO-SEATS	0.198	0.180	0.075	0.274
STL+HDoutliers	0.832	0.823	0.837	1.337
STL+GESD	0.231	0.225	0.226	0.165
STL+iForest	0.727	0.710	0.718	1.237
STL(ROB)+HDoutliers	1.197	1.152	1.186	1.230
STL(ROB)+GESD	4.756	4.699	4.689	7.954
STL(ROB)+iForest	0.992	0.993	0.983	1.514
Twitter+HDoutliers	0.144	0.162	0.182	0.376
Twitter+GESD	3.050	2.761	2.813	6.058
Twitter+iForest	0.123	0.128	0.121	0.508
STR+HDoutliers	1.108	1.135	1.027	1.454
STR+GESD	1.469	1.679	0.582	4.430
STR+iForest	0.732	0.674	0.882	1.653
STR(ROB)+HDoutliers	1.382	1.266	0.876	10.915
STR(ROB)+GESD	4.452	4.583	4.289	19.321
STR(ROB)+iForest	0.831	0.845	0.828	1.858

Cuadro A.12: Resultados Escenario 4. Exceso. ARIMA(0,1,1)x(0,1,1)12.

	Caso 1	**Caso 2**	**Caso 3**	**Caso 4**
X-13ARIMA-SEATS	0.109	0.123	0.111	0.069
TRAMO-SEATS	0.239	0.301	0.312	0.139
STL+HDoutliers	0.869	0.762	0.986	1.367
STL+GESD	0.170	0.170	0.198	0.244
STL+iForest	0.667	0.475	0.712	1.291
STL(ROB)+HDoutliers	1.328	1.357	1.499	1.287
STL(ROB)+GESD	2.517	2.530	2.919	5.438
STL(ROB)+iForest	0.992	0.813	0.978	1.444
Twitter+HDoutliers	1.041	0.966	1.065	1.383
Twitter+GESD	0.796	0.854	0.830	2.221
Twitter+iForest	0.808	0.655	0.735	1.412
STR+HDoutliers	0.799	0.979	0.846	1.551
STR+GESD	0.101	0.087	0.083	0.029
STR+iForest	0.383	0.286	0.619	0.669
STR(ROB)+HDoutliers	0.764	0.759	0.755	1.692
STR(ROB)+GESD	1.600	1.591	1.490	8.145
STR(ROB)+iForest	0.603	0.492	0.593	0.894

Cuadro A.13: Resultados Escenario 5. Exceso. ARIMA(1,0,1)x(0,1,2)12.

	Caso 1	Caso 2	Caso 3	Caso 4
X-13ARIMA-SEATS	0.179	0.167	0.141	0.096
TRAMO-SEATS	0.307	0.299	0.297	0.127
STL+HDoutliers	0.792	0.783	0.775	1.598
STL+GESD	0.052	0.050	0.049	0.030
STL+iForest	0.452	0.306	0.459	0.748
STL(ROB)+HDoutliers	1.193	1.207	1.269	1.697
STL(ROB)+GESD	1.290	1.324	1.270	0.437
STL(ROB)+iForest	0.708	0.531	0.670	0.952
Twitter+HDoutliers	0.690	0.698	0.676	1.157
Twitter+GESD	0.266	0.257	0.225	0.523
Twitter+iForest	0.538	0.416	0.459	1.063
STR+HDoutliers	0.740	0.853	0.796	1.429
STR+GESD	0.065	0.077	0.063	0.036
STR+iForest	0.432	0.311	0.493	0.750
STR(ROB)+HDoutliers	0.749	0.799	0.717	1.477
STR(ROB)+GESD	0.317	0.291	0.254	0.167
STR(ROB)+iForest	0.502	0.373	0.463	0.839

Cuadro A.14: Resultados Escenario 6. Exceso. ARIMA(1,1,1)x(1,0,1)12.

	Caso 1	**Caso 2**	**Caso 3**	**Caso 4**
X-13ARIMA-SEATS	0.147	0.138	0.115	0.081
TRAMO-SEATS	0.278	0.285	0.287	0.261
STL+HDoutliers	0.788	0.815	0.776	1.392
STL+GESD	0.045	0.047	0.042	0.031
STL+iForest	0.471	0.337	0.478	0.842
STL(ROB)+HDoutliers	1.122	1.183	1.194	1.785
STL(ROB)+GESD	1.364	1.387	1.340	0.670
STL(ROB)+iForest	0.720	0.559	0.711	1.100
Twitter+HDoutliers	0.125	0.109	0.084	0.092
Twitter+GESD	3.247	3.207	3.185	7.613
Twitter+iForest	0.087	0.066	0.073	0.294
STR+HDoutliers	0.854	0.807	0.773	1.430
STR+GESD	0.069	0.069	0.050	0.036
STR+iForest	0.438	0.357	0.484	0.793
STR(ROB)+HDoutliers	0.765	0.755	0.725	1.370
STR(ROB)+GESD	0.434	0.489	0.510	0.536
STR(ROB)+iForest	0.581	0.471	0.547	1.064

Cuadro A.15: Resultados Escenario 7. Exceso. ARIMA(1,1,2)x(0,1,1)12.

	Caso 1	Caso 2	Caso 3	Caso 4
X-13ARIMA-SEATS	0.090	0.095	0.104	0.056
TRAMO-SEATS	0.487	0.475	0.494	0.446
STL+HDoutliers	0.734	0.814	0.749	1.507
STL+GESD	0.068	0.059	0.052	0.025
STL+iForest	0.496	0.352	0.479	0.747
STL(ROB)+HDoutliers	1.157	1.098	1.083	1.660
STL(ROB)+GESD	1.232	1.281	1.260	0.294
STL(ROB)+iForest	0.764	0.566	0.687	0.918
Twitter+HDoutliers	0.859	0.822	0.860	1.415
Twitter+GESD	0.214	0.209	0.217	0.114
Twitter+iForest	0.606	0.526	0.550	1.031
STR+HDoutliers	0.797	0.792	0.778	1.469
STR+GESD	0.048	0.059	0.042	0.017
STR+iForest	0.461	0.336	0.464	0.782
STR(ROB)+HDoutliers	0.786	0.726	0.699	1.494
STR(ROB)+GESD	0.159	0.146	0.145	0.072
STR(ROB)+iForest	0.520	0.359	0.433	0.830

Cuadro A.16: Resultados Escenario 8. Exceso. ARIMA(0,1,1)x(1,0,0)12.

Bibliografía

Abraham, B., y Box, G. E. P. (1979). Bayesian analysis of some outlier problems in time series. *Biometrika*, *66*(2), 229–236.

Barnett, V., y Lewis, T. (1996). Outliers in statistical data, 3rd edition, (John Wiley & Sons, Chichester), 584 pp., [UK pound]55.00, ISBN 0-471-93094-6. *International Journal of Forecasting*, *12*(1), 175-176.

Bell, W. (1983). A computer program for detecting outliers in time series. *Proceedings of the American Statistical Association, Business and Economic Statistics Section*, 634-639.

Box, G., y Jenkins, G. (1976). Time series analysis, forecasting and control. En (Vol. 134). doi: 10.2307/2344246

Box, G. E. P., y Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association*, *70*(349), 70–79.

Breunig, M., Kriegel, H.-P., Ng, R., y Sander, J. (2000). Lof: Identifying density-based local outliers. En (Vol. 29, p. 93-104). doi: 10.1145/342009.335388

Chang, I., Tiao, G., y Chen, C. (1988). Estimation of time series parameters in the presence of outliers. *Technometrics*, *30*, 193-204. doi: 10.1080/00401706.1988.10488367

Chang, I., y Tiao, G. C. (1983). Estimation of time series parameters in the presence of outliers. *Technical Report 8, University of Chicago, Statistics Research Center*, *30*.

Chauvenet, W. (1963). A manual of spherical and practical astronomy. , *volume 2*, 474-566.

Chen, C., y Liu, L.-M. (1993). Joint estimation of model parameters and outlier effects in time series. *Journal of the American Statistical Association*, *88*(421), 284–297.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., y Terpenning, I. (1990). STL: A seasonal-trend decomposition procedure based on Loess. *Journal of Official Statistics*, *6*.

Dancho, M., y Vaughan, D. (2019). Anomalize: Tidy anomaly detection [Manual de software informático].

Datacamp. (2018). *Detect anomalies with anomalize in r.* `https://www.datacamp.com/community/tutorials/detect-anomalies-anomalize-r`. (Visitado: 2020-06-16)

de Lacalle, J. L. (2015). *Detection of outliers in time series with r.* `https://jalobe.com/tsoutliers/`. (Visitado: 2020-03-16)

de Lacalle, J. L. (2019). Detection of outliers in time series [Manual de software informático]. `https://cran.r-project.org/web/packages/tsoutliers/tsoutliers.pdf`.

Dokumentov, A., y Hyndman, R. J. (2015). STR: A seasonal-trend decomposition procedure based on Regression.

Dokumentov, A., y Hyndman, R. J. (2018). str: Str decomposition [Manual de software informático]. (`https://cran.r-project.org/web/packages/stR/index.html`)

Fraley, C., y Wilkinson, L. (2020). Hdoutliers: Leland wilkinson's algorithm for detecting multidimensionaloutliers [Manual de software informático].

Gómez, V., y Maravall, A. (1997). Programs tramo and seats: instructions for the user. *Mimeo, Banco de España*.

Gómez, V., y Taguas, D. (1995). Detección y corrección automática de outliers con tramo: Una aplicación al ipc de bienes industriales no energéticos.

Grubbs, F. (1950). Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics*, *21*. doi: 10.1214/aoms/1177729885

Hawkins, D. (1980). *Identification of outliers*. London [u.a.]: Chapman and Hall.

Hochenbaum, J., Vallis, O., y Kejariwal, A. (2017). Automatic anomaly detection in the cloud via statistical learning.

Hoerl, A., y Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *8*, 27-51.

Hyndman, R. (2020). Forecasting functions for time series and linear models [Manual de software informático]. `https://cran.r-project.org/web/packages/forecast/forecast.pdf`.

la Tente, A. Q., Michalek, A., Palate, J., y Baeyens, R. (2020). Rjdemetra: Interface to 'jdemetra+' seasonal adjustment software [Manual de software informático].

Liu, F. T., Ting, K., y Zhou, Z.-H. (2009). Isolation forest. En (p. 413 - 422). doi: 10.1109/ICDM.2008.17

Mahdavinejad, M. S., Rezvan, M., Barekatain, M., Adibi, P., Barnaghi, P., y Sheth, A. P. (2018). Machine learning for internet of things data analysis: a survey. *Digital Communications and Networks*, *4*(3), 161 - 175. doi: https://doi.org/10.1016/j.dcan.2017.10.002

Moore, D., y McCabe, G. (1999). *Introduction to the practice of statistics*. W.H. Freeman.

Otto, M., y Bell, W. (1990). Two issues in time series outlier detection using indicator variables. *Proceedings of the American Statistical Association, Business and Economic Statistics Section*, 182-187.

Peirce, B. (1852). Criterion for rejection of doubtful observations. *The Astronomical Journal*, *2*, 161-163. doi: 10.1086/100259

Peña, D. (2010). *Análisis de series temporales*. Alianza Editorial.

Rosner, B. (1975). On the detection of many outliers. *Technometrics*, *17*, 221-227. doi: 10.1080/00401706.1975.10489305

Rosner, B. (1983). Percentagepoints for a generalized esd many-outier procedure. *Technometrics*, *25*, 165-172. doi: 10.1080/00401706.1983.10487848

Rousseeuw, P., y Hubert, M. (2011). Robust statistics for outlier detection. *Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery*, *1*, 73-79. doi: 10.1002/widm.2

Shan, G. (2015). Improved confidence intervals for the youden index. doi: 0.1371/journal.pone.0127272

Shiskin, J., Young, A., y Musgrave, J. (1967). *The x-11 variant of the census method ii seasonal adjustment program*.

Srikanth, K. S. (2017). solitude: An implementation of isolation forest [Manual de software informático]. (R package version 0.2.1 — For new features, see the 'Solitude' file (in the package source))

Thompson, W. (1935). On a criterion for the rejection of observations and the distribution of the ratio of deviation to sample standard deviation. *The Annals of Mathematical Statistics*, *6*. doi: 10.1214/aoms/1177732567

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*, 267-288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Time Series Research Staff. (2017). X-13ARIMA-SEATS Reference Manual. *Center for Statistical Research and Methodology.*.

Tsay, R. S. (1986). Time series model specification in the presence of outliers. *Journal of the American Statistical Association*, *81*(393), 132–141.

Tsay, R. S., y Tiao, G. C. (1984). Consistent estimates of autoregressive parameters and extended sample autocorrelation function for stationary and nonstationary arma models. *Journal of the American Statistical Association*, *79*(385), 84–96.

Tukey, J. W. (1977). *Exploratory data analysis.* Addison-Wesley.

Twitter. Inc. (2015). *Anomalydetection r package.* GitHub. (`https://github.com/twitter/AnomalyDetection`)

Wilkinson, L. (2017). Visualizing big data outliers through distributed aggregation. *IEEE Transactions on Visualization and Computer Graphics*, *PP*, 1-1. doi: 10.1109/TVCG.2017.2744685

Woodward, W. A., Gray, H. L., y Elliott, A. C. (2017). Applied time series analysis with r, second edition. *Journal of Time Series Analysis*, *39*, 618. doi: 10.1111/jtsa.12273

Youden, W. (1950). Youden wjindex for rating diagnostic tests. cancer 3(1): 32-35. *Cancer*, *3*, 32-5. doi: 10.1002/1097-0142(1950)3:13.0.CO;2-3