Facultad
de
Ciencias

Modelo de previsión de demanda de transporte logístico de palés

Forecast model for demand for logistical transport of pallets

Trabajo de Fin de Máster
para acceder al

MÁSTER EN CIENCIA DE DATOS

Autora: Miriam Cobo Cano

Director: Diego García Saiz

Codirector: Adolfo Garandal Martín

Junio - 2021

Índice

1. Introducción **1**
- 1.1. Motivación y contexto . . . 1
- 1.2. Propósito . . . 2
- 1.3. Organización del trabajo . . . 2

2. Desarrollo **3**
- 2.1. Proceso de ETL . . . 3
- 2.2. Primeras visualizaciones . . . 4
- 2.3. Definición del problema a abordar . . . 8
- 2.4. Tecnologías . . . 13
  - 2.4.1. Holt-Winters . . . 13
  - 2.4.2. Prophet . . . 13
  - 2.4.3. Modelos autorregresivos . . . 14
  - 2.4.4. STL-ARIMA . . . 14
  - 2.4.5. LSTM . . . 15

3. Resultados **16**
- 3.1. Holt-Winters . . . 16
- 3.2. Prophet . . . 17
- 3.3. Modelos autorregresivos . . . 18
- 3.4. STL-ARIMA . . . 19
- 3.5. LSTM . . . 20

4. Discusión **22**
- 4.1. Evaluación de las predicciones en el sector alimentación . . . 22
- 4.2. Evaluación de las predicciones en el sector bodega vitivinícola . . . 23
- 4.3. Limitaciones de los modelos . . . 23
- 4.4. Validación de los resultados con el cliente . . . 24

5. Conclusiones **26**

Agradecimientos

Me gustaría dar las gracias a mi director, Diego, por dirigirme este trabajo y mostrar constantemente su apoyo e interés. Al mismo tiempo, quiero dar las gracias a mi codirector, Adolfo, y a todos los miembros de LIS Data Solutions por la experiencia tan enriquecedora formando parte del equipo. De forma especial a Javi Otero, por ayudarme siempre, y a mis compañeros de prácticas, Alberto, Alex, Andrés y Luis, con quienes comencé esta aventura.

También quiero agradecer a Lara Lloret su enorme implicación y esfuerzo a lo largo del curso, continuamente dispuesta a echarnos una mano. A la vez, quiero dar las gracias a mis amigos y compañeros del máster, es un placer haberos conocido.

Y sin duda, quiero dar las gracias a las personas que están a mi lado, a mis amigos y mi familia, a mis padres y mi hermano por estar siempre ahí, por todo.

Resumen

La predicción de las necesidades de transporte logístico mediante modelos estadísticos, técnicas de minería de datos y algoritmos de inteligencia artificial se ha convertido en un elemento esencial en la planificación del transporte de mercancías, permitiendo a las empresas anticiparse a los picos de demanda y distribuir los medios disponibles de forma más eficiente. La previsión con antelación de los recursos necesarios para poder satisfacer las necesidades de sus clientes posibilita rentabilizar al máximo los esfuerzos de transporte de las compañías, y supone importantes beneficios económicos, además de incrementar la productividad.

A lo largo de este trabajo se desarrolla un modelo de previsión de demanda de transporte por carretera de palés en España. La predicción se lleva a cabo considerando una provincia concreta, así como dos sectores pertenecientes a actividades diferentes, con el fin de analizar situaciones específicas que podrían generalizarse en estudios posteriores. Se tendrán en cuenta la estacionalidad y la tendencia local, para lo cual se usarán técnicas de análisis de series temporales. Se aplicarán diversos modelos estadísticos y de aprendizaje automático, como suavizado exponencial, Prophet, autorregresión o redes neuronales, y se validarán las previsiones obtenidas, con el objetivo de comparar los resultados de cada método y determinar cuál se adapta mejor al problema.

Por un lado, la intención es ofrecer modelos al cliente que le sirvan en determinados sectores y provincias, mostrando la utilidad de las técnicas de análisis de series temporales que se van a aplicar. Por otro lado, el procedimiento seguido puede ser escalable a otras provincias y sectores en futuros desarrollos del proyecto. Además, las soluciones aportadas presentan aplicaciones en la vida real, dando respuesta a un problema de predicción de demanda de una empresa de transporte logístico. El trabajo realizado constituye un proyecto de ciencia de datos que se ha llevado a cabo de principio a fin en todas sus fases.

Palabras clave: ciencia de datos, predicción de demanda, series temporales, modelos estadísticos, aprendizaje automático.

Abstract

The need for forecasting for logistic transport using data mining techniques, artificial intelligence algorithms and statistical models has become an essential element in freight transport planning. These methods enable companies to anticipate demand peaks and efficiently distribute the available means of transport. The advanced prevision of the needed resources to meet the necessities of their customers allows them to make the most of transport efforts, while it entails significant economic benefits, in addition to increasing productivity for the company.

In this dissertation, a forecast demand model for the transport of pallets by road in Spain is developed. Predictions are carried out in a particular province, as well as two sectors that belong to different activities, in order to analyze specific situations that could be generalized in subsequent studies. Time series analysis techniques will be used, considering seasonality and local trends. For this purpose, various statistical and machine learning models will be applied, such as exponential smoothing, Prophet, autoregression or neural networks. Predictions will be validated as a means to compare the results of each method and determine which one best suits the problem.

On the one hand, the intention is to offer models to the client that work in certain sectors and provinces, showing the usefulness of the time series analysis techniques that will be applied. On the other hand, the procedure followed can be scalable to other provinces and sectors in future project developments. Furthermore, the solutions provided present real-life applications, responding to a demand forecasting problem of a logistics transport company. This work constitutes a data science project which has been developed from beginning to end in all its phases.

Key Words: data science, demand forecast, time series, statistical models, machine learning.

Glosario

- **ETL.** *Extract, Transform, Load.* Proceso de extracción, transformación y carga de los datos.
- **ETS.** *Error/Trend/Seasonality.* Descomposición de una serie temporal en las componentes de error, tendencia y estacionalidad.
- **Holt-Winters.** Variante del modelo de suavizamiento exponencial que puede tener en cuenta la estacionalidad si está presente en los datos.
- ***Lag.*** Retraso temporal.
- **LOESS.** *Locally Estimated Scatterplot Smoothing.* Curva de suavizado de un conjunto de datos a partir de regresiones lineales ponderadas a nivel local.
- **LSTM.** *Long Short Term Memory.* Red neuronal recurrente capaz de manejar dependencias a largo plazo en secuencias muy largas.
- **Modelos autorregresivos, AR.** *AutoRegressive models.* Aquellos en los que la variable a predecir viene dada por una combinación lineal de sus observaciones pasadas.
- **Modelos autorregresivos integrados de medias móviles, ARIMA.** *AutoRegressive Integrated Moving Average.* Modelos estadísticos que incluyen tres procedimientos: autorregresión (AR), integración (I) y correlación entre las observaciones (MA). La generalización SARIMA (*Seasonal AutoRegressive Integrated Moving Average*) incorpora la estacionalidad.
- **Modelos autorregresivos vectoriales, VAR.** *Vector AutoRegressive models.* En ellos, la variable a predecir es una combinación lineal de sus observaciones pasadas (*lags*) y de las observaciones de otras variables.
- **Prophet.** Algoritmo para predecir datos de series temporales desarrollado por Facebook.
- **Serie temporal.** Secuencia de observaciones de una variable ordenadas cronológicamente.
- **STL.** *Seasonal-Trend decomposition using LOESS.* Método de descomposición de series temporales aplicando la técnica de suavizado de LOESS.
- **Variable endógena.** Aquella que viene dada por su relación con otras variables dentro del propio modelo.
- **Variable exógena.** Aquella cuyo valor viene predeterminado por factores externos con capacidad explicativa para predecir una variable endógena del modelo.

1. Introducción

En este capítulo se expone en primer lugar la motivación y el contexto dentro del cual se enmarca el trabajo realizado. A continuación, se presentan los principales objetivos del proyecto y, finalmente, se describe la estructura interna de la memoria.

1.1. Motivación y contexto

El transporte de mercancías por carretera resulta fundamental para garantizar el abastecimiento de productos en una economía globalizada como la actual. Las predicciones de demanda de transporte logístico han adquirido gran importancia, en especial a corto plazo para las compañías que prestan este tipo de servicios, pero también a largo plazo con el fin, por ejemplo, de programar la construcción y el mantenimiento de infraestructuras ya existentes por parte de los gobiernos [1].

En este sentido, las empresas dedicadas al transporte logístico de mercancías, bien sea por carreteras u otros medios, requieren de métodos cada vez más precisos capaces de realizar las predicciones de demanda, con la finalidad de reducir costes y poder diseñar con antelación las rutas más eficientes, lo cual supone un considerable ahorro de combustible y una menor contaminación del medio ambiente, además de importantes beneficios para dichas empresas. La planificación previa del volumen de transporte que van a realizar es crucial para aumentar la rentabilidad de su negocio, mejorar la organización y optimizar la distribución de los recursos disponibles [1].

Las predicciones de demanda a corto plazo están influenciadas por diversos factores de naturaleza más incierta que los pronósticos a largo plazo, los cuales siguen en general las líneas de desarrollo económico a nivel regional y nacional [2]. La demanda de transporte en un futuro cercano no solamente se ve afectada por dichas tendencias globales, sino también por aspectos locales de la compañía en cuestión, tales como su posición de mercado, la estrategia de marketing o las relaciones con sus proveedores y clientes. Estos elementos cambian continuamente, lo que dificulta el incremento de precisión de las predicciones [2].

Por otro lado, la calidad del pronóstico depende de la incertidumbre y la aleatoriedad que se encuentran asociadas a la demanda de mercancías. Dicha demanda puede verse afectada por variables muy diversas, desde el índice de producción industrial (IPI) hasta la meteorología o, en los tiempos actuales, la incidencia de casos de Covid-19.

Al mismo tiempo, el sector al que pertenecen los productos transportados es una variable a tener en cuenta por parte de las compañías, particularmente en tiempos de crisis como los que vivimos, puesto que existen áreas estratégicas dedicadas a cubrir necesidades básicas de las personas, como por ejemplo la alimentación, para la cual resulta indispensable garantizar el correcto abastecimiento.

Dentro de los diversos métodos de predicción de demanda se diferencian dos grandes bloques: los procedimientos cualitativos y los cuantitativos [2]. El presente trabajo se centrará en la predicción de demanda siguiendo técnicas cuantitativas, las cuales engloban una gran variedad de modelos matemáticos con los que se realizan las previsiones de futuro. Dichos modelos son entrenados con datos históricos y, en ocasiones, con otras variables capaces de aportar información adicional significativa. Los procedimientos más comunes incluyen el pronóstico de series temporales a través de métodos de regresión y medias móviles, así como redes neuronales para las tareas más complejas [2]. Por el contrario, los métodos de predicción cualitativos están basados principalmente en la experiencia personal y valoraciones subjetivas [2], por lo que su precisión es limitada y no pueden llevarse a cabo de manera automática.

De acuerdo con los motivos expuestos anteriormente, el desarrollo de modelos de pronóstico de demanda fiables, y la interpretación correcta de sus resultados, tiene importantes aplicaciones reales para las compañías de logística, permitiendo reducir tiempos, costes y optimizar estrategias comerciales; así como aumentar beneficios, productividad y rentabilidad.

1.2. Propósito

El objetivo de este trabajo es llevar a cabo una revisión de los métodos de pronóstico de demanda para series temporales y aplicar dichos modelos a un problema real de predicción de pedidos de palés para una empresa de transporte logístico.

Para ello, nos centraremos en una provincia en concreto y dos sectores diferentes de la propia compañía, con el fin de delimitar el problema al que nos enfrentamos y mostrar los resultados en situaciones específicas. Esto no supone una pérdida de generalidad ya que, como se explicará a lo largo del trabajo, el procedimiento seguido puede ser adaptable a cualquier provincia y sector ajustando los parámetros de los algoritmos en función de las características de la serie temporal considerada, y volviendo a entrenar el modelo con los datos correspondientes.

Se comenzará llevando a cabo el proceso de ETL para extraer y tratar las fuentes de datos. Se realizarán visualizaciones de los datos procesados, con el fin de ver si hay correlaciones y conocer mejor el problema al que nos enfrentamos. Se analizarán los diferentes algoritmos que puedan servir a los objetivos de este trabajo y se seleccionarán aquellos que mejor encajen con los mismos y con los datos disponibles. A continuación, se presentarán los resultados obtenidos y se calculará el error cometido en las predicciones con cada modelo. Por último, se discutirán los resultados finales, se validarán las previsiones con la empresa cliente y se evaluará la fiabilidad de cada algoritmo en las conclusiones del trabajo.

1.3. Organización del trabajo

Este trabajo se estructura a partir de aquí en cinco capítulos, de la forma que se especifica a continuación:

- El capítulo 2 incluye una descripción detallada del tratamiento de las fuentes de datos, las primeras visualizaciones y los modelos que se van a aplicar. En primer lugar, se explica el proceso de ETL llevado a cabo con las tablas de datos del histórico correspondiente a una compañía real. También se introducen los primeros análisis de los datos y se realizan visualizaciones de los mismos. A continuación, se presenta la métrica de validación y se define el alcance del problema con el que se van a entrenar los modelos de predicción. Finalmente, se explican los métodos y algoritmos que se utilizarán para realizar las predicciones de demanda de palés.

- El capítulo 3 presenta los resultados obtenidos después de aplicar los diferentes algoritmos explicados en el capítulo anterior, representados a través de gráficas, con comparaciones entre las predicciones y el conjunto de datos reservado para llevar a cabo la validación.

- El capítulo 4 comprende la discusión de los resultados obtenidos en función de la precisión de cada método, realizando un análisis de la fiabilidad del modelo en cuestión. También se incluye la respuesta del cliente cuando se validaron los resultados en la reunión final de presentación del proyecto.

- El capítulo 5 está dedicado a resumir las tareas realizadas, examinar las principales conclusiones e indicar futuras líneas de trabajo.

A lo largo del texto, todas las figuras en las que no se establezca una autoría de forma explícita son obra de la autora.

2. Desarrollo

En los siguientes apartados se comienza describiendo el proceso de ETL. A continuación, se realizan visualizaciones de los datos para familiarizarse con los mismos. Asimismo, se define el alcance del problema que se va a considerar para entrenar los modelos en el capítulo posterior. Para terminar, se presentan los métodos de predicción de demanda más habituales que se aplicarán a lo largo de este trabajo.

2.1. Proceso de ETL

La ETL del histórico de datos del cliente se realizó en KNIME, un software con el cual se crean workflows visuales a través de nodos que permiten modelizar cada paso de este proceso (veáse [3]). Se disponía de siete tablas en formato CSV correspondientes a cada año desde 2014 hasta 2020-febrero 2021, respectivamente.

En las figuras 2.1 y 2.2 se muestran las capturas del workflow ejecutado, con los diferentes nodos que se utilizaron para leer las tablas y realizar el procesado de los datos que se describe a continuación.

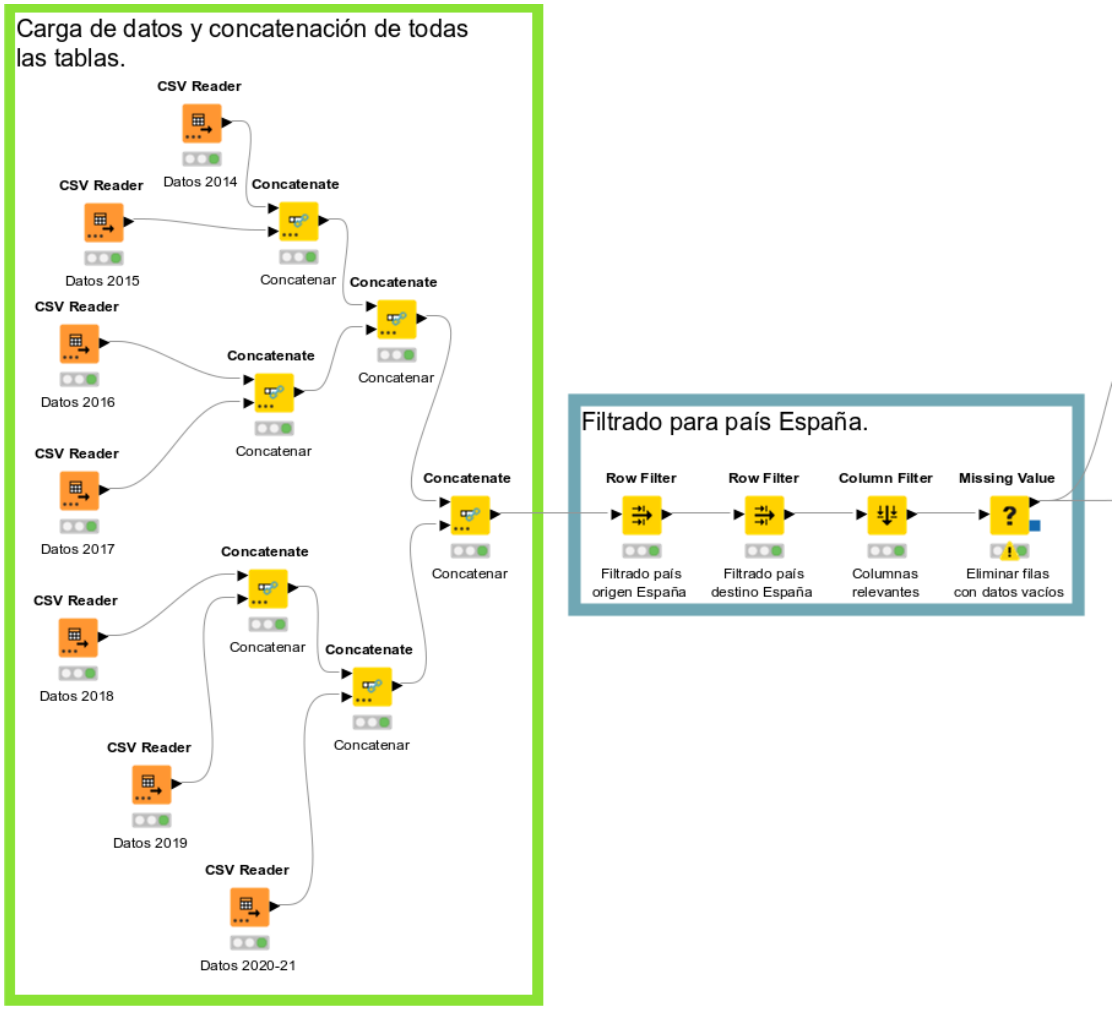

Figura 2.1: Captura del workflow creado en KNIME para realizar la ETL de datos históricos.

En primer lugar, se leyeron y concatenaron todas las tablas, de forma que se juntaron los datos del histórico en una única tabla. Se realizó un filtrado por país, seleccionando solamente aquellos envíos

cuyo origen y destino era España, para lo cual se empleó el nodo *Row Filter*. A continuación, se separó la provincia correspondiente al cliente para el cual se llevó a cabo el transporte, que se obtuvo a partir de los dos primeros dígitos de su código postal, haciendo uso del nodo *String Manipulation*. Este dato del código postal de origen se relacionó con la provincia a través de un diccionario con el nodo *Rule Engine* de KNIME.

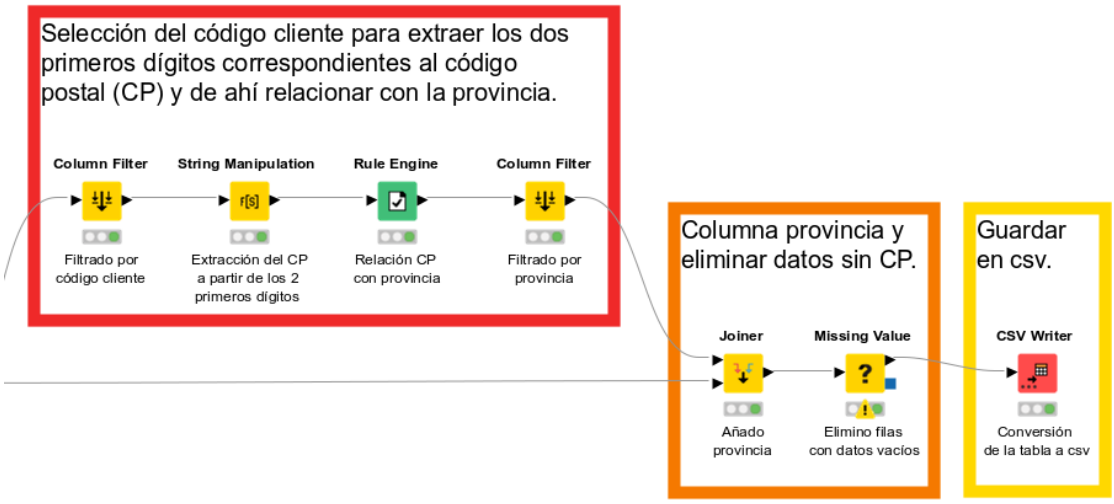

Figura 2.2: Captura de la continuación del workflow creado en KNIME para realizar la ETL de datos históricos.

Posteriormente, se filtraron aquellas columnas con las que íbamos a trabajar empleando un *Column Filter*, atendiendo a los peticiones del cliente a la hora de llevar a cabo las predicciones. En nuestro caso, las columnas de interés eran la provincia, la fecha de salida, el tipo de actividad y el número de palés del pedido. Finalmente, se ignoraron los datos en blanco o nulos, utilizando el nodo *Missing Value*, y se guardó la tabla resultante en un CSV para manipularla después en un *Jupyter Notebook* con *kernel* de *Python*.

2.2. Primeras visualizaciones

El primer paso que se llevó a cabo después de realizar la ETL de los datos históricos de la empresa fue filtrar para el periodo correspondiente a 2015-2019 (ambos inclusive) y por el sector alimentación, puesto que este área es de interés para el cliente al resultar estratégica en momentos de crisis, así que será una de las analizadas cuando se entrenen los modelos. También se considerará el sector bodega vitivinícola, por tratarse de un tipo de actividad donde se esperan observar patrones estacionales en los datos, lo cual posibilitará estudiar una serie temporal con estacionalidad.

Se decidió no considerar los registros a partir del año 2020, debido a que coincide con el periodo de pandemia y los patrones de este año no son extrapolables a los periodos de normalidad. Igualmente, se prescindió del año 2014, porque el volumen de mercancías transportadas por la compañía era considerablemente inferior al de los años posteriores (debe tenerse en cuenta que, conforme transcurre el tiempo, va mejorando el posicionamiento en mercado de la empresa, como se puede apreciar en la tendencia mayoritariamente ascendente de las gráficas que se muestran en la Figura 2.3).

Se agruparon los datos del número de palés transportados por semanas en vez de días, dado que el objetivo es anticipar los recursos necesarios para la provincia correspondiente con una granularidad semanal, de forma que sea posible enviar los camiones estimados con una antelación suficiente. Además, se observó que los datos diarios tenían un número de palés bajo y presentaban grandes fluctuaciones locales, por lo que la agrupación por semanas era lo más razonable en este caso.

De cara a entrenar los modelos, a continuación se mostrarán aquellas provincias que tienen un mayor número de palés transportados, primero, en el sector alimentación y, posteriormente, en la actividad bodega vitivinícola. En esta última, esperamos observar una correlación anual de los datos, asociada a la estacionalidad de este tipo de actividad. Consideraremos que la correlación es significativa si el coeficiente de Pearson es del orden de 0.4 o superior, lo cual se corresponde con un valor moderado o intermedio [4].

Las figuras que se muestran a continuación han sido realizadas con el lenguaje de programación *Python* en el entorno de *Jupyter Notebook*, en concreto, usando las librerías *Numpy*, *Pandas*, *Seaborn*, *Matplotlib*, *Datetime* y *Statsmodels*.

En la Figura 2.3 se recoge la representación gráfica del número de palés transportados en función de la provincia correspondiente al código postal del cliente que realizó el pedido. Se muestran las 9 provincias que presentan un mayor número de pedidos de palés para el sector alimentación.

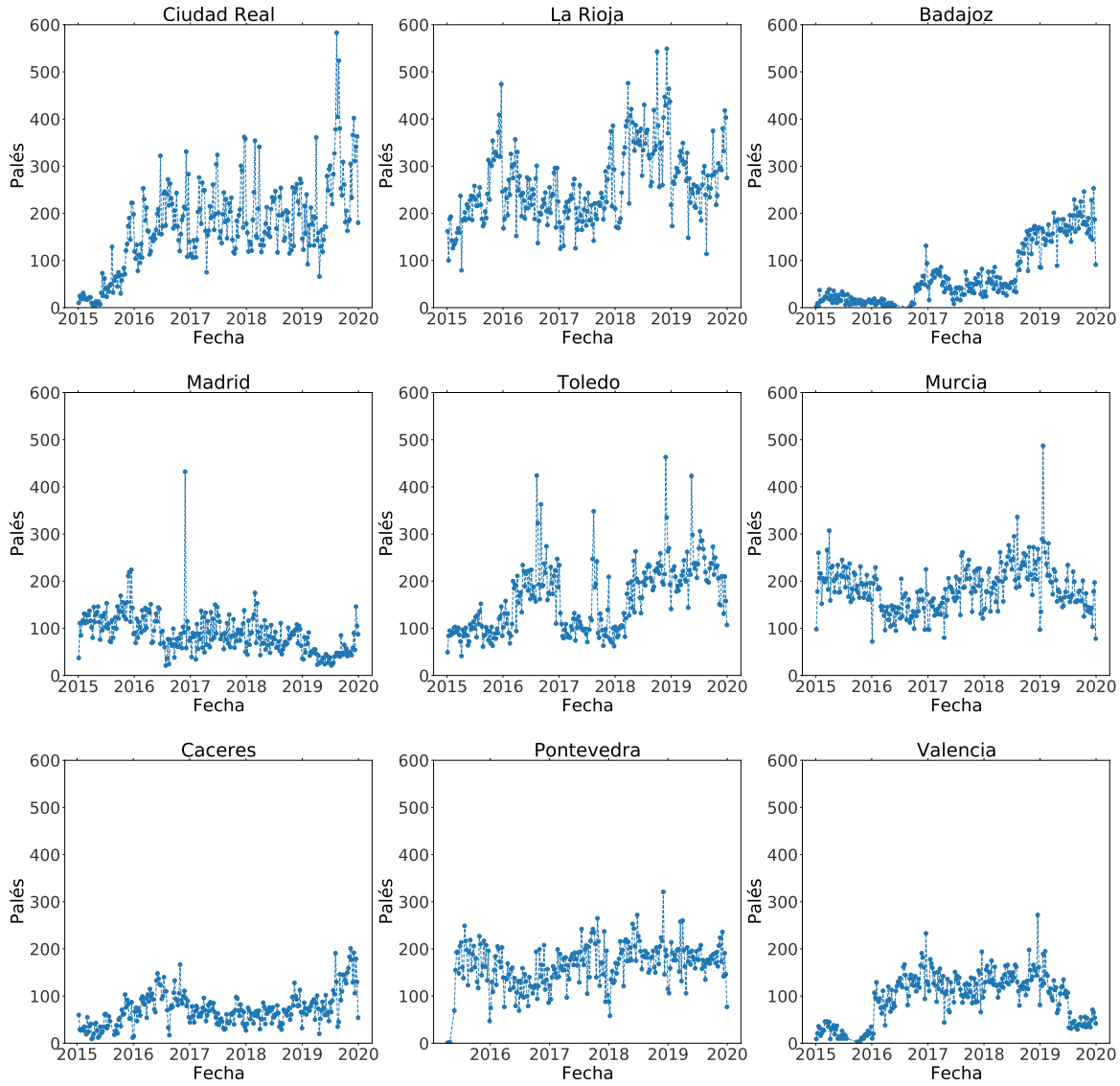

Figura 2.3: Número de palés por provincia entre los años 2015-2019 para el sector alimentación.

Como puede observarse en la Figura 2.3, hay algunas provincias que tienen un número significativamente mayor de palés en promedio, como son, por ejemplo, Ciudad Real, La Rioja o Pontevedra. Además, debe evaluarse si se producen grandes fluctuaciones en el número de palés por semana, lo

cual es un factor a tener en cuenta a la hora de entrenar los modelos de series temporales, especialmente cuando se van a ajustar los parámetros del algoritmo. Ambas circunstancias se valorarán en el momento de extraer conclusiones de las siguientes gráficas, con el objetivo de elegir aquella provincia que posea un número de palés representativo para entrenar con suficientes datos los modelos.

En la Figura 2.4 se representan las autocorrelaciones de Pearson de los datos del número de palés por provincia para el sector alimentación, considerando un retraso pasado (*lag*) de 78 semanas, es decir, un año y medio. Se tiene que para este tipo de actividad no se aprecia una estacionalidad de los datos si nos fijamos, en particular, en aquellas provincias que exhiben un mayor número de palés (Ciudad Real, La Rioja y Pontevedra). Esto concuerda con los rasgos propios del sector de alimentos, que se caracteriza por cubrir una necesidad esencial y, por lo tanto, se consume siempre.

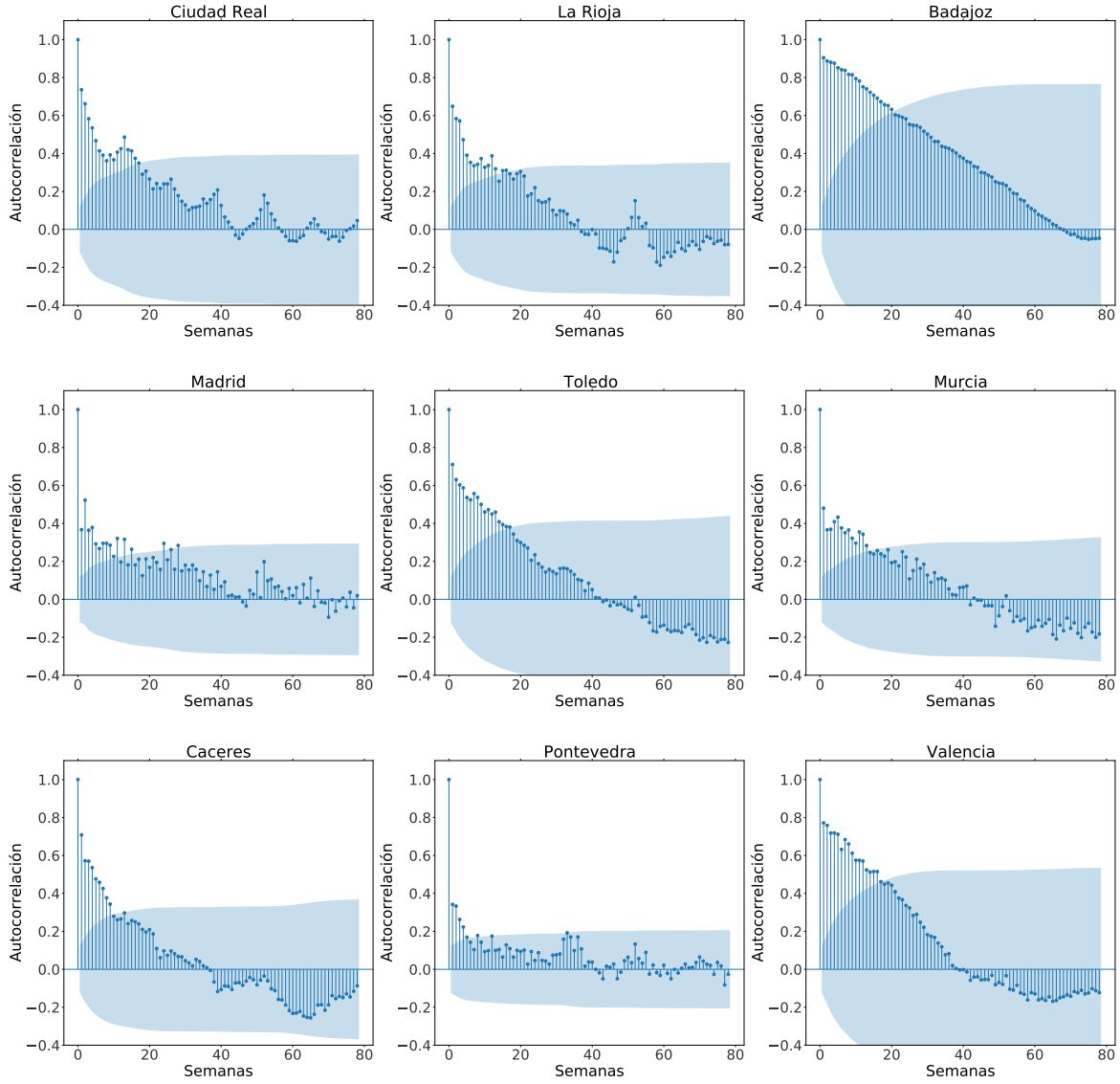

Figura 2.4: Autocorrelaciones del número de palés por provincia para el sector alimentación.

En la Figura 2.5 se recogen las autocorrelaciones de Pearson de los datos del número de palés por provincia para el sector bodega vitivinícola, tomando un *lag* de 78 semanas. Al igual que antes, se muestran las 9 provincias que presentan un mayor número de pedidos en dicho sector. En este caso, puede observarse que existe una estacionalidad anual en los datos de Asturias, La Rioja o Valladolid, por citar algún ejemplo. Esto es coherente con las características del sector, que se rige habitualmente

por una temporalidad de este tipo.

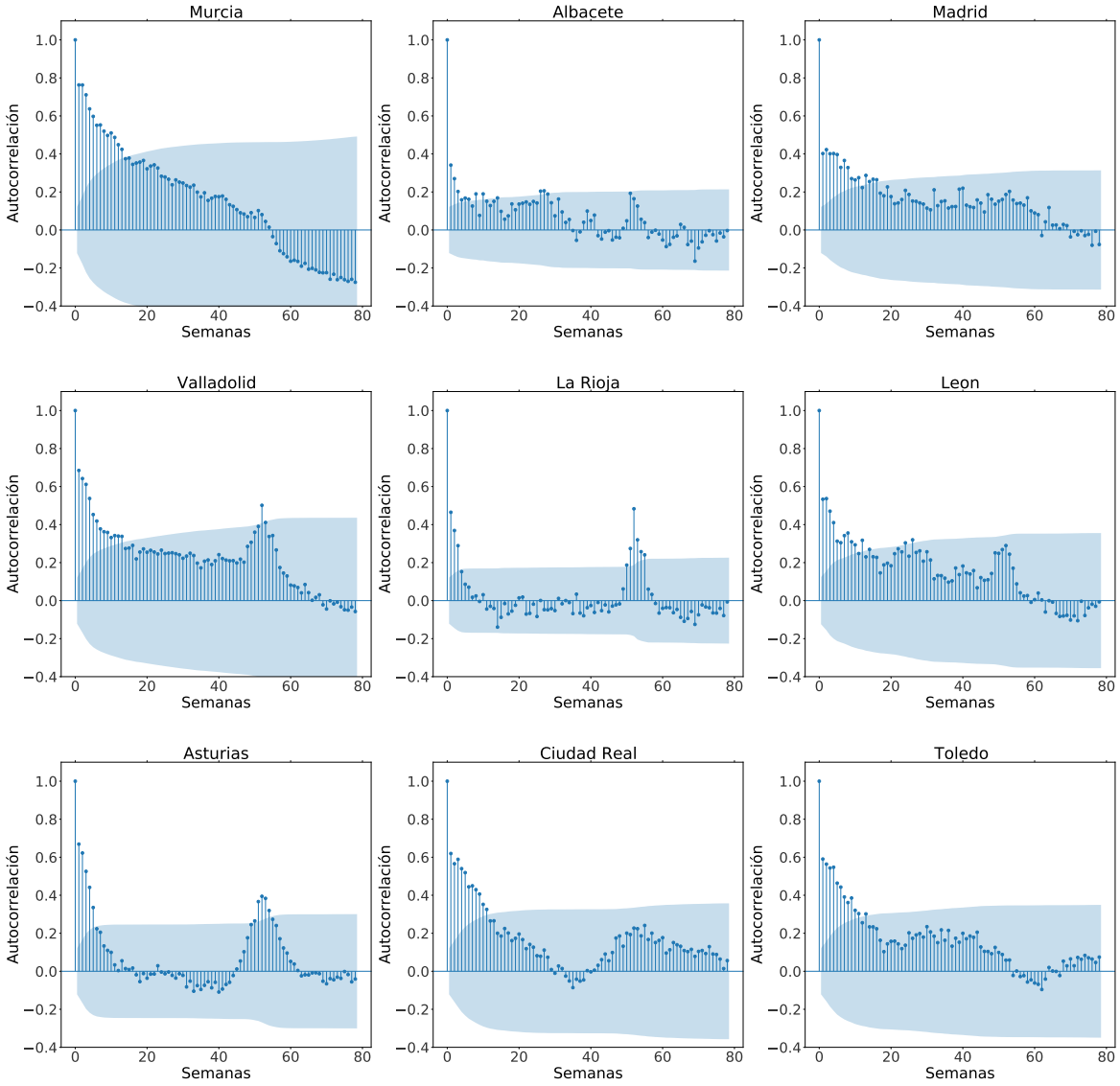

Figura 2.5: Autocorrelaciones del número de palés por provincia para el sector bodega vitivinícola.

Adicionalmente, en la Figura 2.6 se muestra la representación gráfica del número de palés transportados en función de la provincia para el sector bodega vitivinícola. Puede verse que las provincias mencionadas anteriormente (Asturias, La Rioja y Valladolid) tienen un número considerable de palés trasladados.

De acuerdo con lo expresado previamente, se tiene que la provincia de La Rioja muestra un número de pedidos de palés lo suficientemente grande como para que sea viable utilizarla con el propósito de entrenar un algoritmo predictivo. Además, podemos tener en cuenta los dos sectores en los que nos hemos concentrado, por un lado, el de alimentación y, por otro lado, el de bodega vitivinícola, que presenta una correlación en torno a 0.5 para un *lag* de 52 semanas, que puede asociarse con la estacionalidad anual propia de este tipo de actividad. Al mismo tiempo, puede verse que en ambas subdivisiones los datos no muestran por lo general comportamientos extraños, con picos muy extremos o grandes fluctuaciones que dificultarían la implementación de los modelos, por lo que a priori esta provincia parece una elección razonable para nuestro propósito.

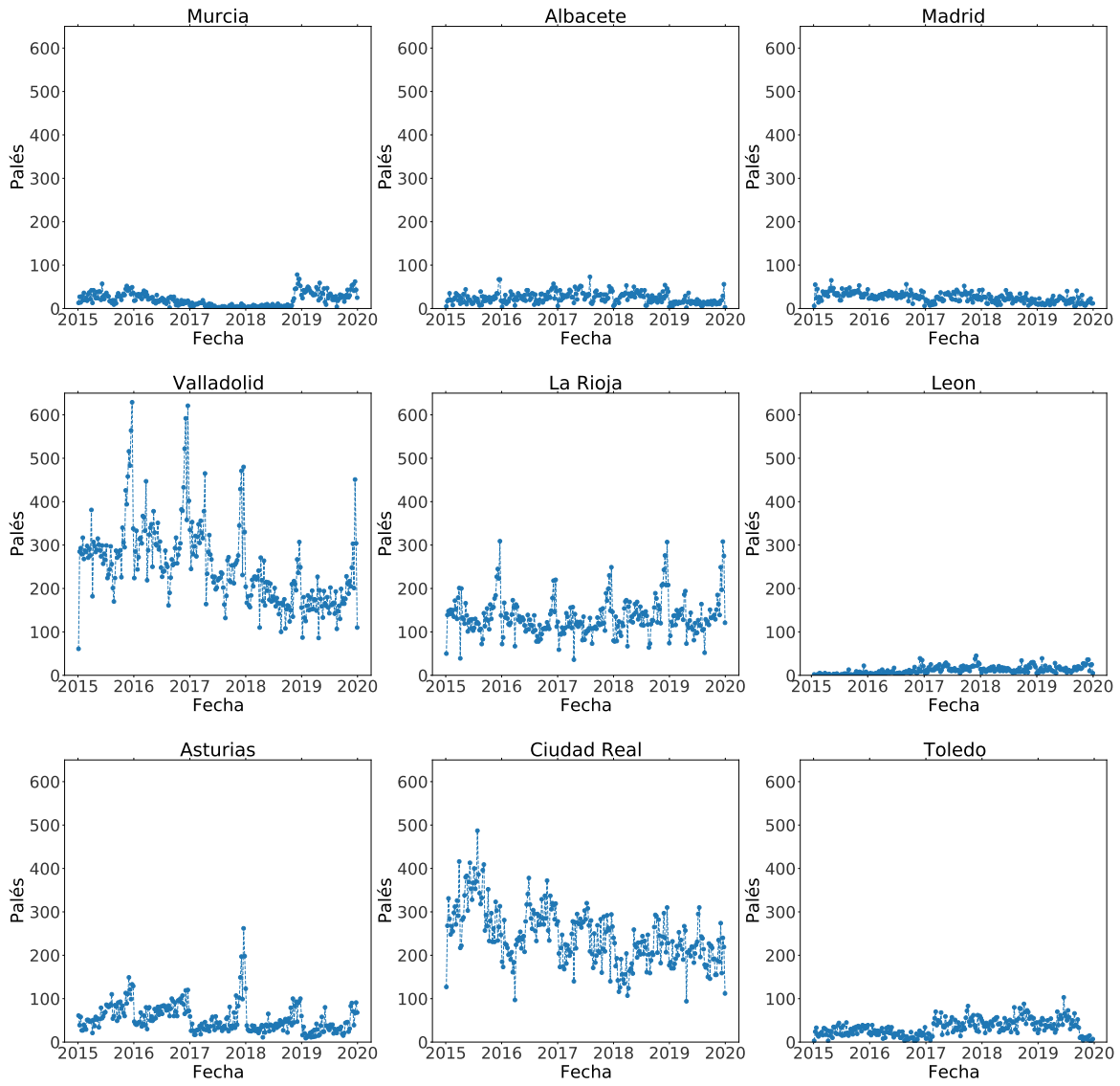

Figura 2.6: Número de palés por provincia entre los años 2015-2019 para el sector bodega vitivinícola.

2.3. Definición del problema a abordar

En la sección anterior se han mostrado diversas representaciones de los datos históricos. Con el objetivo de tratar un problema específico de cara a especializarnos en las predicciones que se llevarán a cabo con diferentes algoritmos, se tomará la provincia de La Rioja como referencia para entrenar nuestros modelos. Como se ha comentado, se considerarán por separado los sectores alimentación y bodega vitivinícola.

La métrica de validación que se empleará será la raíz del error cuadrático medio, RMSE (del inglés, *Root Mean Square Error*), por tratarse de un problema de regresión. El RMSE se define como sigue

$$RMSE = \sqrt{\frac{\sum_{j=1}^{N}(y_j^p - y_j^{obs})^2}{N}} \tag{2.1}$$

donde y^p e y^{obs} son los valores predichos y observados, respectivamente, mientras que N es el número total de datos que se han pronosticado.

A continuación, se presentan diferentes visualizaciones de los datos del sector alimentación en La Rioja, con las cuales se discutirán algunas de las características de la serie temporal que queremos predecir.

En primer lugar, la Figura 2.7 muestra la serie temporal del número de palés para el histórico de datos completo. Como puede observarse, los años 2014 y 2020 presentan patrones distintos al periodo 2015-2019, por esta razón, no serán tenidos en cuenta, de acuerdo con lo mencionado en la sección anterior.

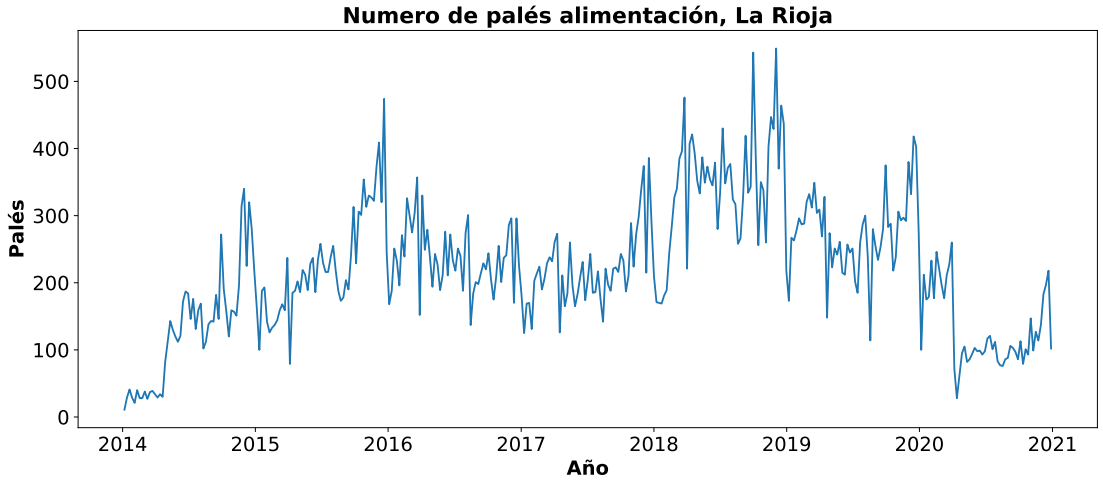

Figura 2.7: Número de palés del sector alimentación en La Rioja en función del tiempo.

Asimismo, en la Figura 2.8 se representan los datos del histórico para el periodo 2015-2019 con una granularidad mensual, en función del año. De nuevo, se deduce que el comportamiento del año 2020 es diferente al resto.

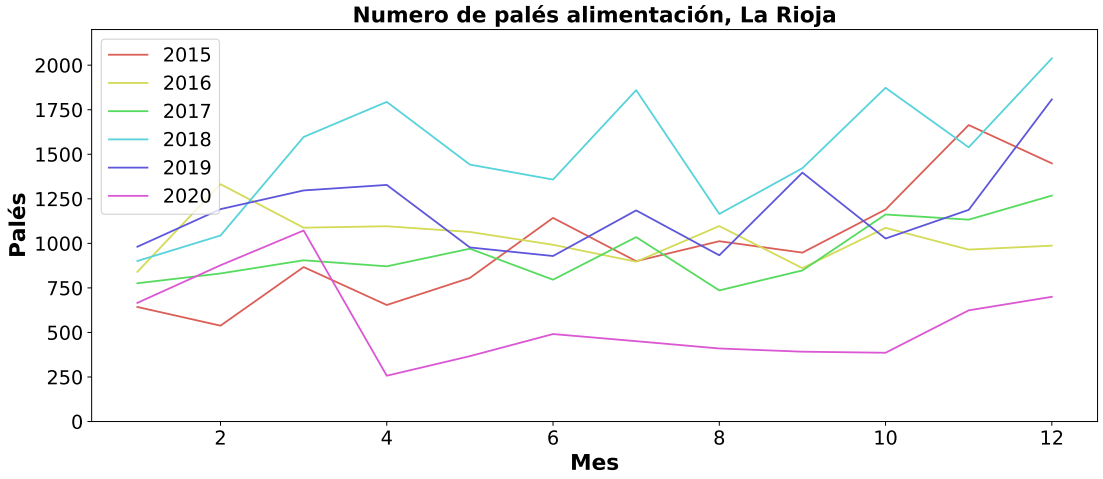

Figura 2.8: Número de palés del sector alimentación en La Rioja en función del mes del año correspondiente.

En la Figura 2.9 se recogen los diagramas de caja correspondientes a la tendencia anual y la estacionalidad mensual, respectivamente, para el periodo 2015-2019. Se observa que los datos presentan estacionalidad si nos fijamos en los diagramas de caja mensuales, ya que se aprecian variaciones en función del mes considerado. El valor de la mediana en los diagramas de caja de tendencia anual está dentro del rango de 880 y 1500 palés. Vemos que los años 2016 y 2019 presentan outliers, es de-

cir, datos poco frecuentes, como consecuencia de picos puntuales de demanda. La mediana de palés en los diagramas de caja de estacionalidad mensual oscila entre 800 y 1500, aproximadamente. En aquellos meses donde la dispersión de los datos es mayor se observa que el máximo y el mínimo, los 'bigotes' del diagrama, están más separados.

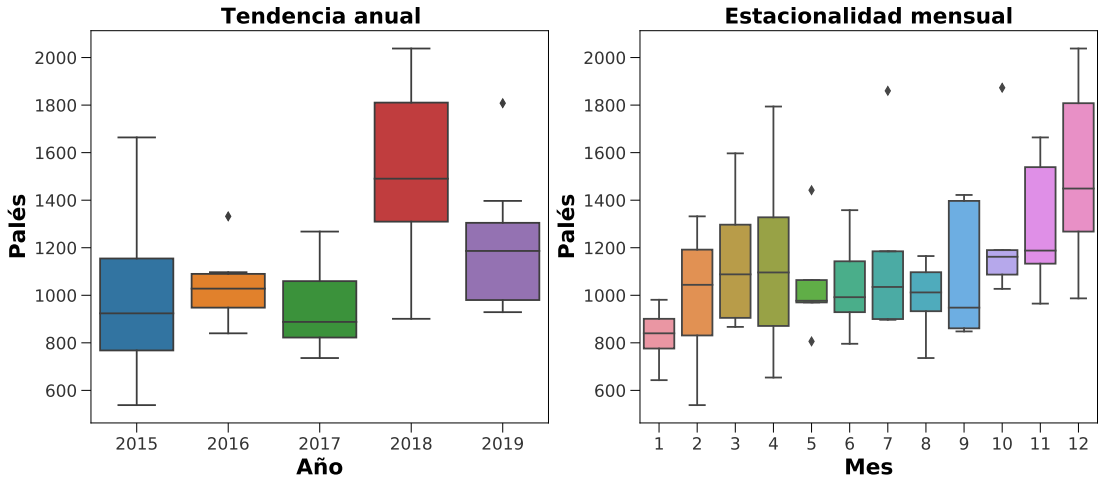

Figura 2.9: Diagramas de caja del número de palés del sector alimentación en La Rioja en función de los años, a la izquierda, y de los meses, a la derecha, para el periodo 2015-2019.

A modo de resumen, en la Figura 2.10 se muestra la descomposición ETS (*Error/Trend/Seasonality* [5]) de la serie temporal en tres componentes: tendencia del ciclo, estacionalidad y error residual. Se ha realizado una descomposición aditiva, puesto que la alteración de las fluctuaciones estacionales no varía de forma significativa con el nivel de la serie temporal. Se ha elegido la descomposición aditiva en lugar de la multiplicativa (no lineal) porque esta última es más apropiada cuando la variación de la componente estacional es proporcional al nivel de la serie temporal [5].

En las figuras 2.11, 2.12 y 2.13 se muestran las mismas representaciones descritas anteriormente para el sector alimentación, pero en este caso para la actividad bodega vitivinícola considerando el periodo 2015-2019.

Como puede verse en la Figura 2.11, se tiene que el año 2020 presenta un comportamiento distinto al resto, por lo que será descartado en el momento de entrenar los modelos. Asimismo, se aprecia que la serie temporal del sector bodega vitivinícola manifiesta de forma evidente patrones estacionales, con un aumento del número de palés en los meses de octubre y noviembre que culmina con un pico de demanda en diciembre, el cual se produce a lo largo de los años pertenecientes al periodo 2015-2019. También se observa en la Figura 2.11 una caída del número de palés en el mes de agosto, coincidiendo con el periodo de vacaciones en verano.

Los diagramas de caja que recogen la tendencia anual de la Figura 2.12 muestran que la mediana del tipo de actividad bodega vitivinícola apenas fluctúa dentro del rango de 500-600 palés. En contraste, si nos fijamos en las variaciones mensuales, puede apreciarse la estacionalidad de los datos de la serie temporal, ya que en este caso la mediana va desde casi 400 hasta un valor ligeramente superior a 900 palés. El mes de diciembre es el que tiene mayor rango de valores, lo cual, como podía observarse también en la Figura 2.11, coincide con el pico de demanda de este sector.

La Figura 2.13 muestra la descomposición ETS de la serie temporal correspondiente al sector bodega vitivinícola. La descomposición realizada es aditiva porque, como ocurría en el sector alimentación, la alteración de las oscilaciones estacionales apenas sufre cambios con el nivel de la serie temporal. Se observa que los datos tienen una importante componente estacional con picos bastante marcados, siendo el patrón de tendencia más débil que en el caso de alimentación, con un rango de unos 30 palés en el tipo de actividad bodega vitivinícola frente a los casi 150 de alimentación (Figura

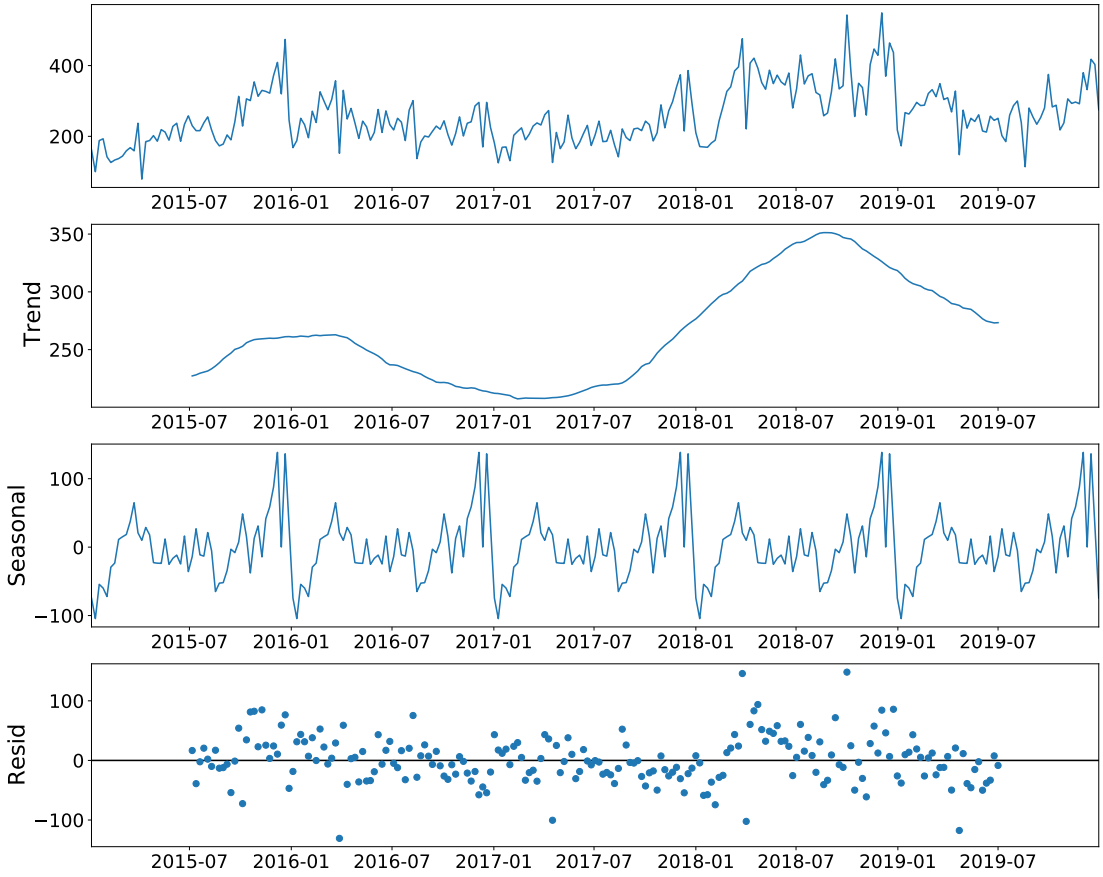

Figura 2.10: Descomposición ETS del número de palés del sector alimentación en La Rioja para el periodo 2015-2019.

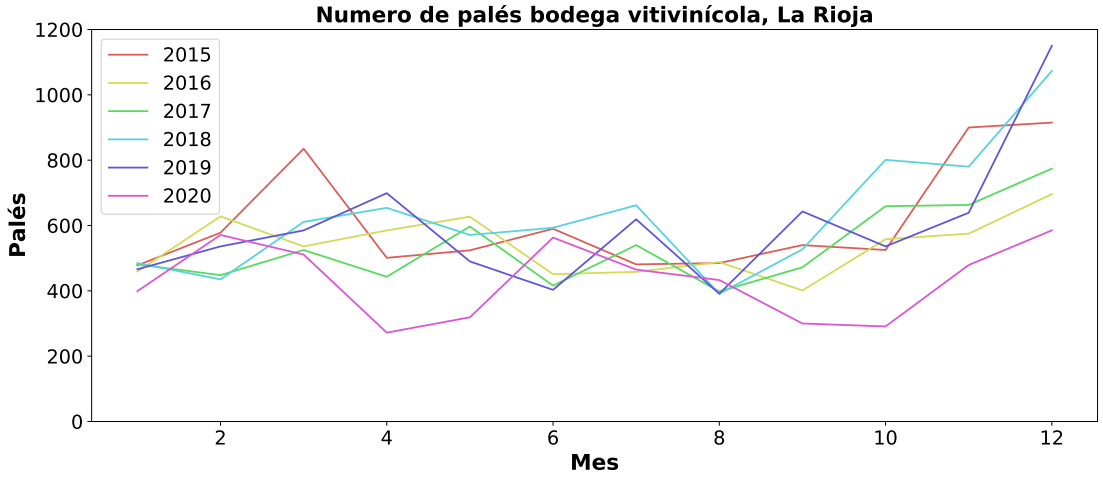

Figura 2.11: Número de palés del sector bodega vitivinícola en La Rioja en función del mes del año correspondiente.

2.10).

Finalmente, en el capítulo 3 los modelos se entrenarán con el histórico desde el año 2015 hasta septiembre de 2019, y se separarán los datos de octubre a diciembre de este último año como conjunto de validación. Se ha decidido hacer predicciones a tres meses porque este periodo posibilita a la

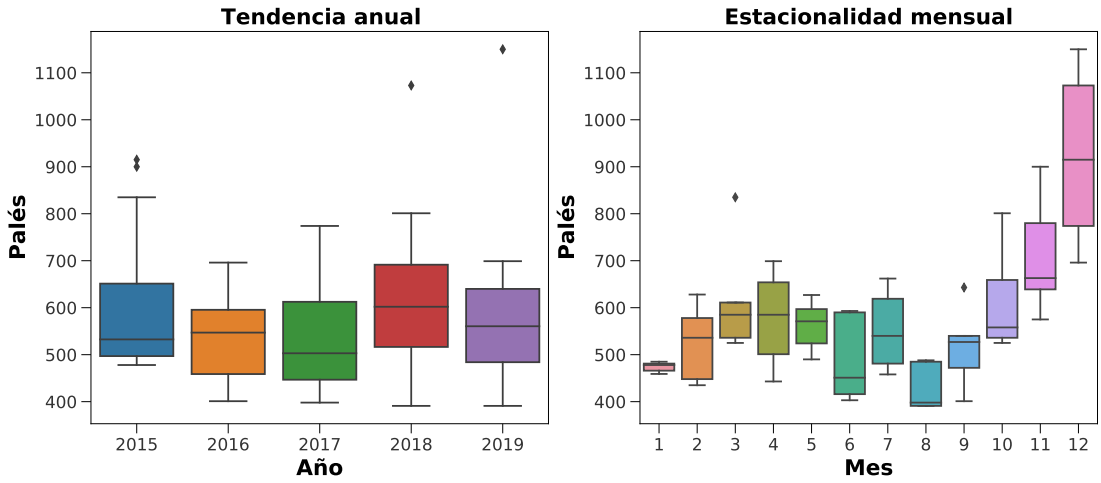

Figura 2.12: Diagramas de caja del número de palés del sector bodega vitivinícola en La Rioja en función de los años, a la izquierda, y de los meses, a la derecha, para el periodo 2015-2019.

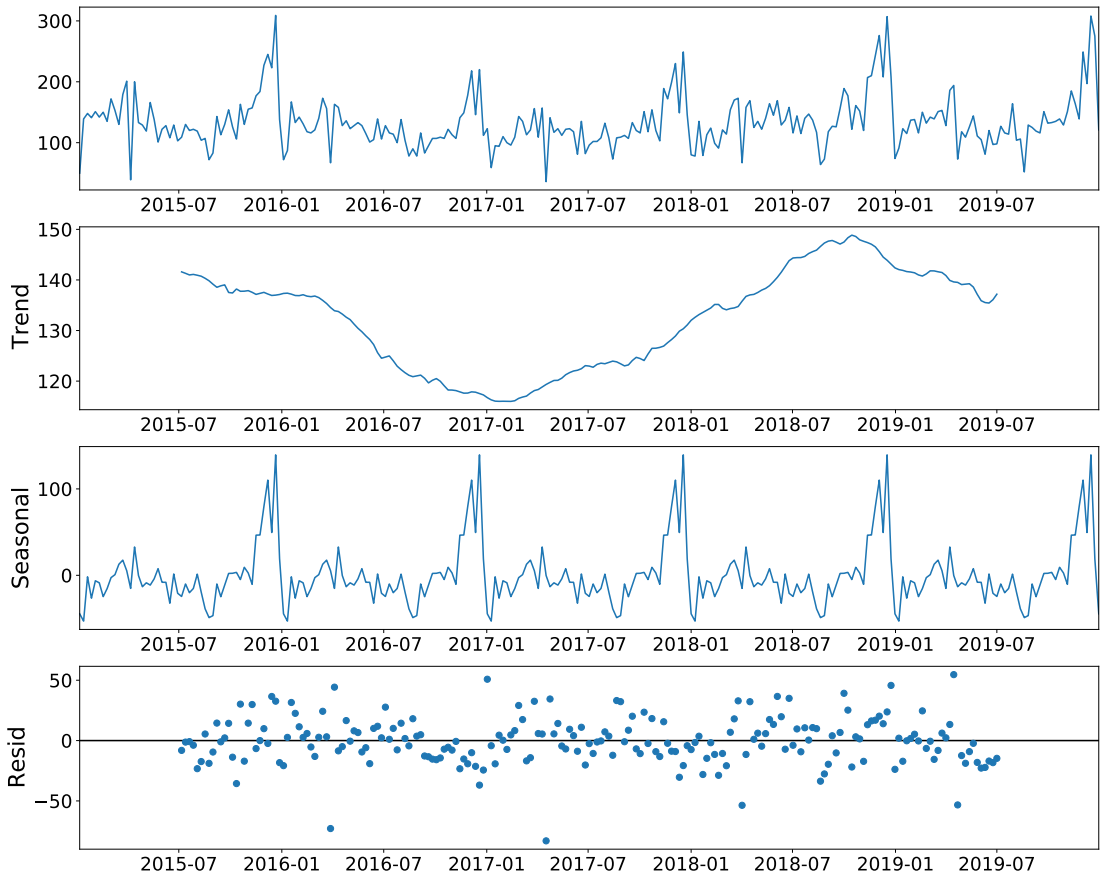

Figura 2.13: Descomposición ETS del número de palés del sector bodega vitivinícola en La Rioja para el periodo 2015-2019.

compañía llevar a cabo su planificación trimestral y no resulta excesivo, ya que cuanto mayor es el rango de las previsiones menor es su precisión. Además, los pronósticos a largo plazo pierden valor para la empresa de cara a organizar su actividad en un futuro cercano, puesto que lo que interesa al cliente es predecir en las próximas semanas. Los resultados de las predicciones con cada modelo se

presentan a lo largo del capítulo 3.

2.4. Tecnologías

La predicción de series temporales consiste en recopilar los valores históricos de la variable a pronosticar, analizarlos y desarrollar un modelo que describa la relación subyacente, con el cual extrapolar la serie temporal a futuro. Este tipo de acercamiento al problema resulta especialmente útil cuando no se conocen a priori otras variables explicativas [6].

A continuación, se describen los algoritmos que se aplicarán para realizar las previsiones de demanda de palés. Se han escogido, por su relevancia y popularidad, los siguientes: Holt-Winters, Prophet, modelos autorregresivos, STL-ARIMA y red neuronal LSTM. Estos modelos se basan en paradigmas diferentes, por lo que permitirán contrastar los resultados obtenidos con el objetivo de determinar cuál de ellos es el más adecuado para las series temporales que vamos a predecir.

2.4.1. Holt-Winters

El modelo Holt-Winters es una variante del suavizamiento exponencial [7], muy popular entre los algoritmos de predicción de demanda. Es un método clásico que suele dar buenos resultados en pronósticos a corto plazo de la serie temporal [7]. Holt (véase [8]) y Winters (véase [9]) extendieron el método de Holt para tener en cuenta la estacionalidad [10].

Los datos se modelan a través de una media, una tendencia y un factor estacional locales que se actualizan mediante un suavizamiento exponencial. El efecto estacional puede ser aditivo o multiplicativo, aunque también puede aplicarse la versión no estacional del algoritmo [7].

La versión del modelo que considera la estacionalidad de forma aditiva supone que los efectos estacionales son de tamaño constante, mientras que la versión multiplicativa asume que los efectos estacionales son proporcionales en tamaño al nivel medio local sin estacionalidad [11].

El método aditivo es preferible cuando las variaciones estacionales son aproximadamente invariables a lo largo de la serie. El procedimiento multiplicativo es recomendable cuando las variaciones estacionales cambian proporcionalmente al nivel de la serie [10].

Antes de entrenar el modelo debe decidirse si se va a considerar la estacionalidad y, en el caso de que así sea, si se tomará la aditiva o la multiplicativa. A los factores estacionales se les debe asignar un valor inicial (si se emplea dicho modelo estacional), y también a la media local y a la tendencia. Así, con cada nueva observación se actualiza tanto la media como la tendencia y los factores estacionales mediante el suavizado exponencial. De este modo, pueden realizarse las predicciones a futuro para un cierto número de pasos [11].

2.4.2. Prophet

Prophet es un método para predecir datos de series temporales basado en un modelo aditivo en el cual las tendencias no lineales se ajustan con una estacionalidad anual, semanal y/o diaria, además de considerar la influencia de los días festivos [12]. Funciona mejor con series temporales que poseen una fuerte componente estacional y varias temporadas de datos históricos [12].

Prophet es un software de código abierto desarrollado por el equipo de Ciencia de Datos de Facebook (véase [13]). El algoritmo puede describirse en líneas generales a través de cuatro componentes [14]:

- Una curva de crecimiento con tendencia lineal o logística que se va actualizando, de forma que Prophet detecta automáticamente los cambios de tendencia seleccionando los puntos de cambio (*changepoints*) de los datos.

- Una componente estacional de tipo anual modelada a través de series de Fourier.

- Una componente estacional semanal que utiliza variables ficticias.
- Una lista de festivos (o fechas relevantes) proporcionada por el usuario.

2.4.3. Modelos autorregresivos

Los modelos autorregresivos (AR, del inglés *AutoRegressive models*) predicen la variable deseada mediante una combinación lineal de valores pasados de dicha variable [10].

Los modelos autorregresivos vectoriales (VAR, del inglés, *Vector AutoRegressive models*) generalizan los métodos autorregresivos univariables al considerar series de tiempo multivariables. Es decir, cada variable es una combinación lineal de los *lags* de ella misma y de los *lags* de otras variables [15].

2.4.4. STL-ARIMA

STL es un método de descomposición de series temporales (del inglés, *Seasonal-Trend decomposition using LOESS*) [10]. Consiste en aplicar de forma secuencial el método de suavizado de LOESS, que es una función de regresión lineal ponderada localmente con la cual se estiman relaciones no lineales, de forma que el resultado es una curva que suaviza la serie temporal. Los datos se descomponen en tres componentes [16]:

- Tendencia: variación de baja frecuencia en los datos junto a cambios de nivel no estacionarios a largo plazo. Puede seguir diferentes patrones, por ejemplo, lineal, exponencial, amortiguado o polinomial [17].
- Estacional: variación de los datos asociada a patrones cíclicos que se repiten en intervalos de tiempo relativamente constantes [17].
- Error o residual: variación de los datos más allá de las componentes estacional y de tendencia, es decir, las fluctuaciones a corto plazo que no son sistemáticas ni predecibles [17].

El algoritmo integrado STL-ARIMA [18] consiste en restar a la serie temporal la estacionalidad estimada con STL y, posteriormente, realizar las previsiones de los datos desestacionalizados utilizando el modelo ARIMA.

A diferencia de los modelos de suavizado exponencial, que están basados en la descripción de la tendencia y estacionalidad de los datos, los procedimientos ARIMA describen autocorrelaciones en los mismos [10].

Los métodos de media móvil (MA, del inglés, *Moving Average*) emplean los errores de pronóstico pasados en un modelo similar a la regresión [10]. Si se combinan las diferencias que transforman la serie temporal en estacionaria con la autorregresión y la media móvil obtenemos el modelo no estacional ARIMA (*AutoRegressive Integrated Moving Average*), que es una generalización del modelo ARMA (*AutoRegressive Moving Average*) [10].

En ARIMA el valor futuro de la variable viene dado por una combinación lineal de varias observaciones pasadas y errores aleatorios [6]. La parte AR hace referencia a la combinación lineal de valores pasados de la propia variable, el término MA se refiere a que el error de la regresión es una combinación lineal de los términos de error residuales del pasado, de forma que transmite la correlación entre las observaciones, es decir, cuánto influyen los valores actuales en los siguientes. Por último, la I indica el número de diferencias necesarias para garantizar que la serie es estacionaria [17] [19, pág. 91] [20]. La generalización SARIMA permite modelar series temporales con variaciones estacionales [17].

La mayor limitación de los modelos ARIMA y sus derivados es que asumen una estructura de correlación lineal entre los datos de la serie temporal, por lo que no son capaces de tener en cuenta patrones no lineales [6].

2.4.5. LSTM

Los algoritmos expuestos en los apartados anteriores requieren un conocimiento previo de la distribución de los datos para construir el modelo de predicción. A continuación se presentan técnicas de aprendizaje automático que, a diferencia de los procedimientos estadísticos, describen propiedades de los datos sin necesidad de conocer en profundidad las características de la serie temporal a priori [17].

Las redes LSTM (del inglés, *Long Short-Term Memory*) son un tipo de redes neuronales recurrentes (RNN, *Recurrent Neural Network*). Las RNN poseen una estructura cíclica y celdas de memoria que permiten guardar temporalmente las entradas anteriores (*inputs*). De este modo, las RNN actúan en secuencias de series temporales entrenando el modelo con un nuevo *input* en cada paso temporal [21].

Las RNN tradicionales presentan una serie de inconvenientes, entre los que cabe destacar que no deben utilizarse arquitecturas muy profundas, ya que se produce el desvanecimiento del gradiente y no son capaces de recordar dependencias a largo plazo [21].

Las redes LSTM se desarrollaron para superar las limitaciones de las RNN, y son capaces de manejar dependencias a largo plazo en secuencias muy largas. Estas redes constan de una serie de celdas de memoria, cada celda está asociada con un estado y múltiples puertas: de entrada, salida y olvido. Estas puertas controlan el flujo de datos entre la capa de entrada y la de salida a través de las celdas de memoria [21].

La puerta de entrada decide si deja pasar el nuevo input a la celda de memoria, la de olvido elimina de esta aquella información que no es importante, y la de salida decide qué información de la celda se envía al resto de la red [22]. Así, al incorporar estas puertas en la arquitectura de la LSTM se da solución al problema de desvanecimiento del gradiente.

3. Resultados

En los siguientes apartados se presentan los resultados obtenidos con cada uno de los algoritmos introducidos en la sección 2.4. Además, se explica cómo se ha llevado a cabo el ajuste de los parámetros de cada uno de los modelos, con el objetivo de mejorar su rendimiento.

3.1. Holt-Winters

El método para implementar el modelo Holt-Winters se importó de la librería *statsmodels.tsa.holtwinters.ExponentialSmoothing* de *Python* [23]. La Figura 3.1 muestra los resultados obtenidos con el modelo Holt-Winters para las predicciones del número de palés en alimentación y bodega vitivinícola frente al valor real en La Rioja.

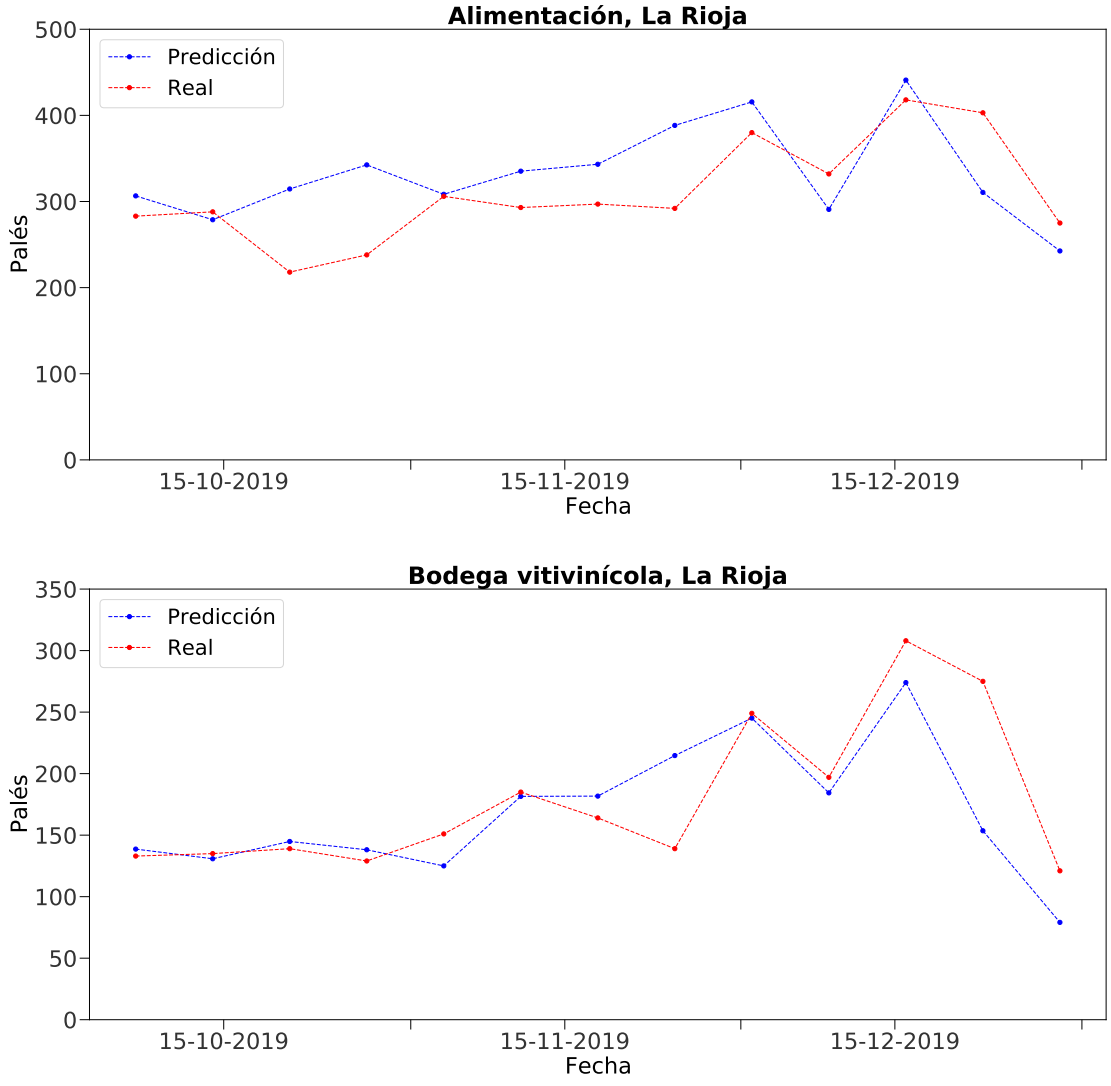

Figura 3.1: Comparación de las predicciones del número de palés en La Rioja con el algoritmo Holt-Winters frente al valor real para los sectores alimentación y bodega vitivinícola.

El algoritmo obtuvo la periodicidad de la serie temporal a partir de la frecuencia semanal de los datos. La tendencia se consideró multiplicativa, dado que de las figuras 2.10 y 2.13 de la sección 2.3 se deduce que este parámetro no sigue una tendencia lineal. La estacionalidad se especificó que era aditiva, puesto que en ambos sectores las variaciones estacionales son aproximadamente constantes a

lo largo de la serie temporal, como se observa en las figuras 2.8 y 2.11. El método ajustaba de forma estimada el valor inicial más conveniente de los parámetros.

3.2. Prophet

Prophet se importó de la librería *fbprophet* y se ajustó con un crecimiento plano, debido a que las series temporales de los sectores alimentación y bodega vitivinícola exhiben más patrones estacionales que cambios bruscos de tendencia [24], como puede verse en las figuras 2.10 y 2.13 de la sección 2.3. También se consideró la estacional anual y la semanal, así como los festivos a nivel nacional en España. La Figura 3.2 muestra las predicciones de Prophet frente a los valores reales para los tipos de actividad alimentación y bodega vitivinícola de La Rioja.

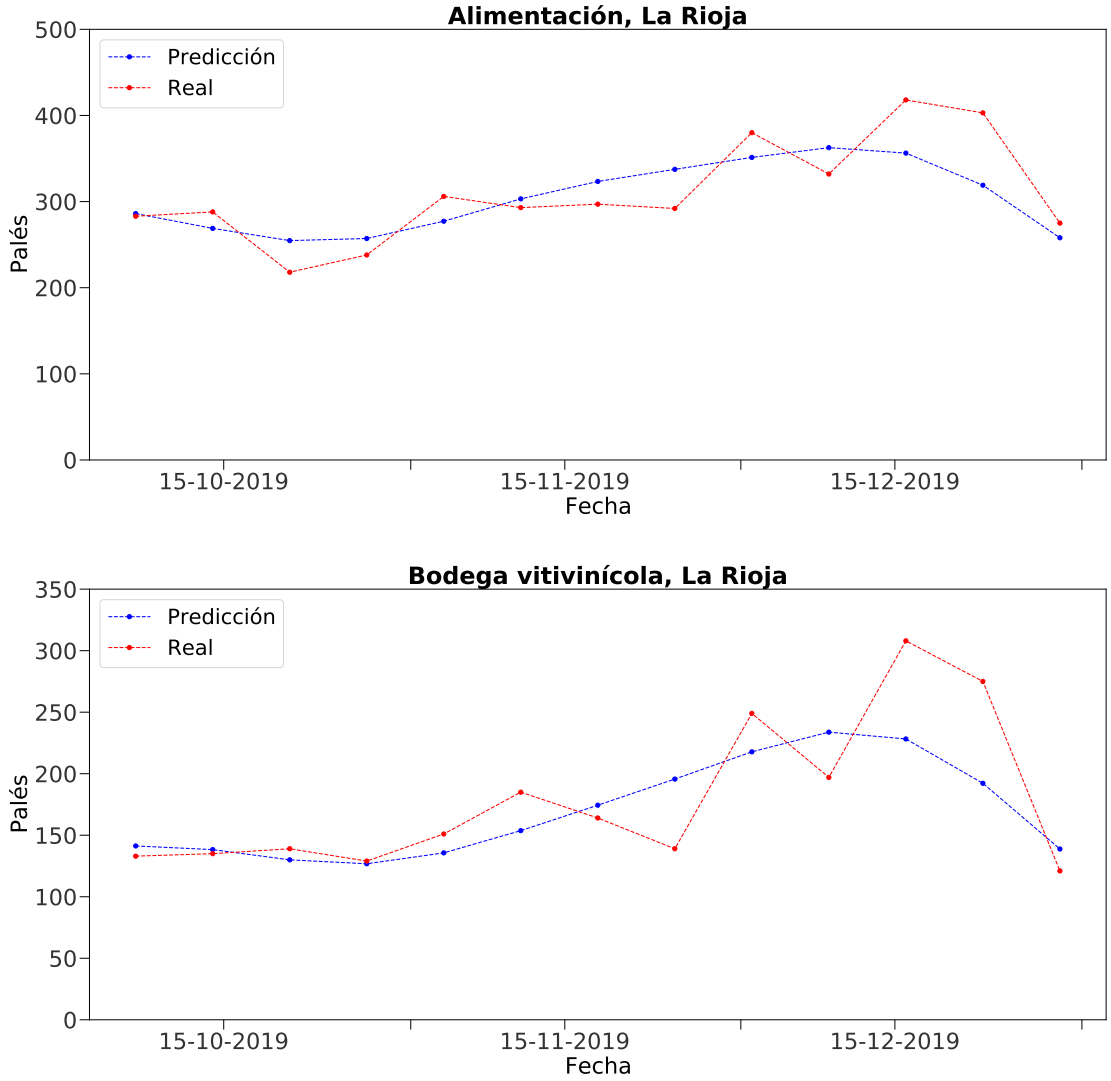

Figura 3.2: Comparación de las predicciones del número de palés en La Rioja con Prophet frente al valor real para los sectores alimentación y bodega vitivinícola.

Se realizaron pruebas incorporando los festivos provinciales de La Rioja, pero se observó que no mejoraban las predicciones del modelo. La razón es que al agrupar los datos con una granularidad semanal, el posible desajuste debido a un festivo regional se compensa con los pedidos del resto de días de la semana. Sin embargo, los festivos nacionales sí se han tenido en cuenta, ya que afectan de forma global a todas las provincias, y los pedidos se transportan a cualquier ubicación dentro de la

península.

Asimismo, con el algoritmo Prophet se realizaron predicciones incorporando una validación cruzada de la serie temporal, utilizando la función *TimeSeriesSplit*, que se importó de *sklearn.model_selection* [25]. Esta validación cruzada es una variación del procedimiento *k-fold* llevada a cabo para pronosticar los valores a futuro de la serie. En este caso, se devuelven los primeros k subconjuntos como *train* y el $(k+1)$ como *test*. Los siguientes subconjuntos de *train* engloban a los anteriores de *test*. De esta forma, se entrena con todo el conjunto de *train*, se predice la primera semana de *test*, se reentrena añadiendo la nueva predicción del paso anterior al conjunto de *train*, se predice un paso más, y así sucesivamente hasta predecir las 13 semanas de *test* correspondientes a las fechas reservadas para validar el modelo. El número de palés obtenido aplicando este método fue muy similar al resultado presentado en la Figura 3.2, por lo que directamente se mostrarán los valores del RMSE en el capítulo 4.

3.3. Modelos autorregresivos

Se aplicaron los métodos AR y VAR, que se importaron de las librerías *statsmodels.ts.ar_model.AutoReg* [26] y *statsmodels.tsa.vector_ar.var_model.VAR* [27], respectivamente.

En AR se ajustó el parámetro tendencia como constante y con el tiempo, esto es, $trend = ct$, indicando también que existía estacionalidad. El *lag* se determinó a partir de la función de autocorrelación parcial, que se muestra en la Figura 3.3. En el caso de alimentación se tomó un *lag* de 12 semanas, es decir, aproximadamente tres meses, mientras que en bodega vitivinícola se consideró un *lag* de 14 semanas, que se corresponde con tres meses y medio. El *lag* se eligió de acuerdo con los resultados de la Figura 3.3.

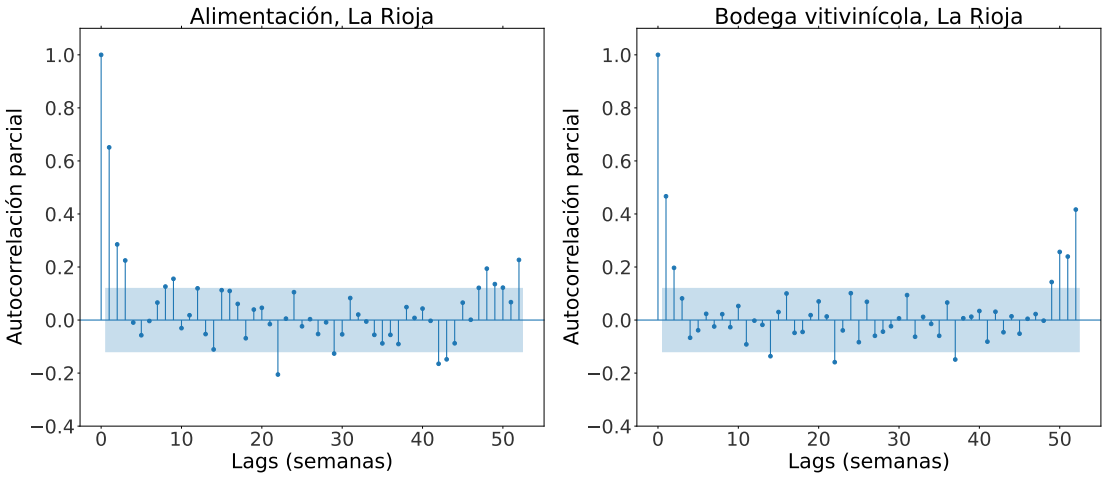

Figura 3.3: Función de autocorrelación parcial de las series temporales de alimentación y bodega vitivinícola en La Rioja.

La Figura 3.4 recoge las predicciones del modelo AR en alimentación y bodega vitivinícola frente a los datos reales en la provincia de La Rioja.

En el caso de VAR, las predicciones obtenidas se descartaron porque eran considerablemente peores que las de AR, con un RMSE más alto, a pesar de ser un modelo que incorporaba como variable endógena los datos históricos del número de palés agrupados de 4 en 4 semanas, además de los datos semanales. El principal motivo encontrado es que, al reagrupar los datos semanales e incluir datos mensuales en el algoritmo, en realidad no se estaba aportando nueva información con respecto al modelo AR. Por esta razón, se omitirán los resultados de VAR.

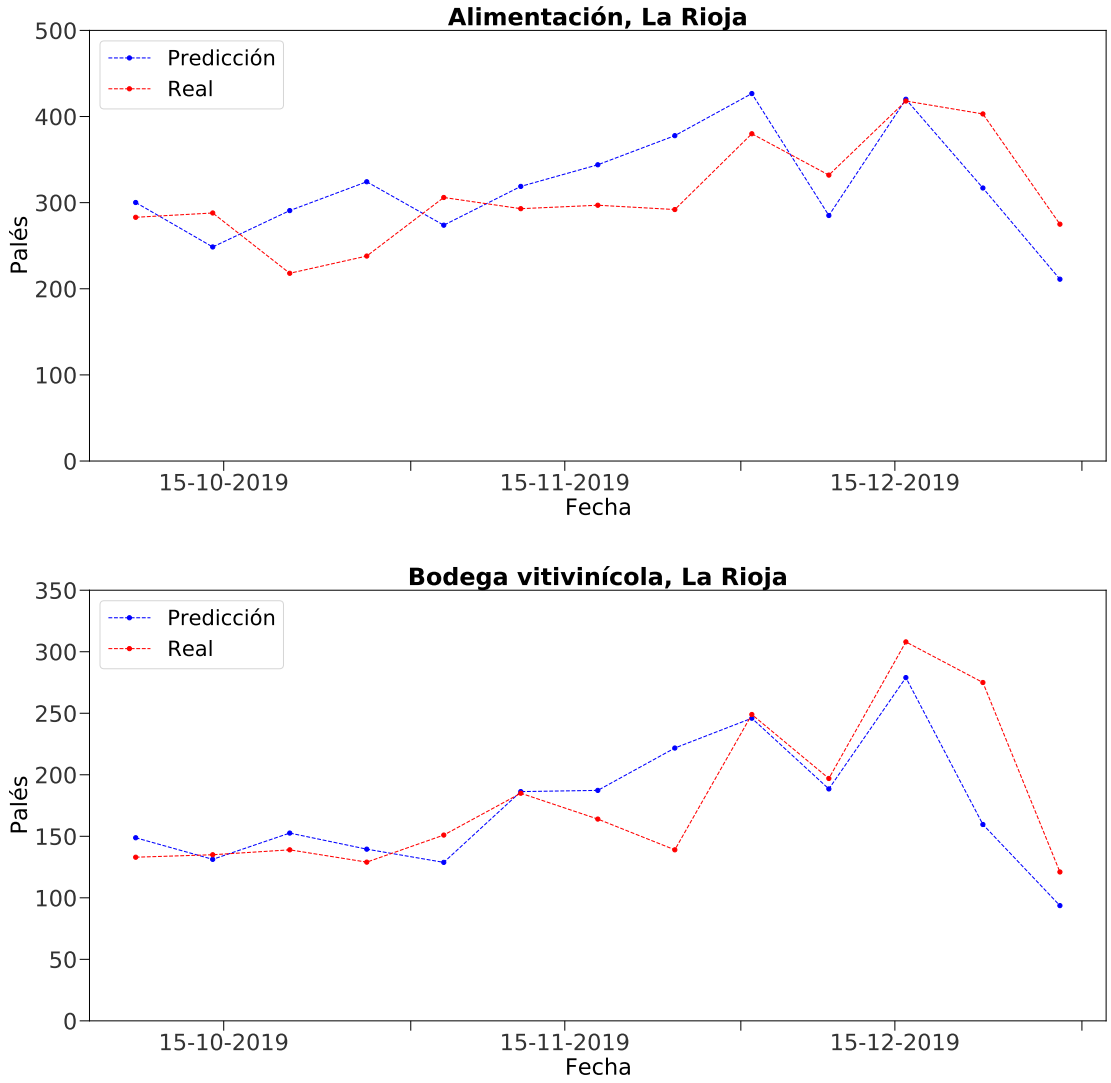

Figura 3.4: Comparación de las predicciones del número de palés en La Rioja con el modelo AR frente al valor real para los sectores alimentación y bodega vitivinícola.

3.4. STL-ARIMA

El método integrado de STL-ARIMA se implementó con STLForecast, importado de la librería *Stats.models.tsa.forecasting.stl* [18], y ARIMA, de *Stats.models.tsa.arima.model* [28].

En *STLForecast* se ajustó el grado de tendencia de LOESS y el grado de estacionalidad a 1, es decir, constante y con tendencia. La periodicidad fue determinada por el algoritmo a partir de la frecuencia semanal de los datos del histórico. Se fijó el parámetro *robust* en *True*, lo que significa que la función de ponderación que se aplica en el suavizado de LOESS es resistente a valores atípicos. La longitud del suavizado estacional debía ser un número entero impar, así que se hizo una búsqueda para encontrar el valor óptimo, para lo cual se probaron todos los enteros impares desde 3 hasta 23. El resultado para ambos sectores fue que el parámetro *seasonal* con valor 9 daba un menor RMSE.

El modelo que se aplicó después de desestacionalizar la serie temporal con STL fue ARIMA, con el cual se realizaron las predicciones del número de palés. ARIMA se ajustó con un grado $(0, 0, 0)$, de acuerdo con las pruebas realizadas para distintos valores de los órdenes de la parte AR, I o MA, respectivamente. En realidad, este grado de ajuste se corresponde con un proceso de ruido blanco [10], lo cual es debido a que la componente con mayor peso en ambas series temporales es la

estacionalidad, que se suprime con STL en este método. La Figura 3.5 muestra los pronósticos del modelo integrado STL-ARIMA frente a los valores reales del número de palés para alimentación y bodega vitivinícola en La Rioja.

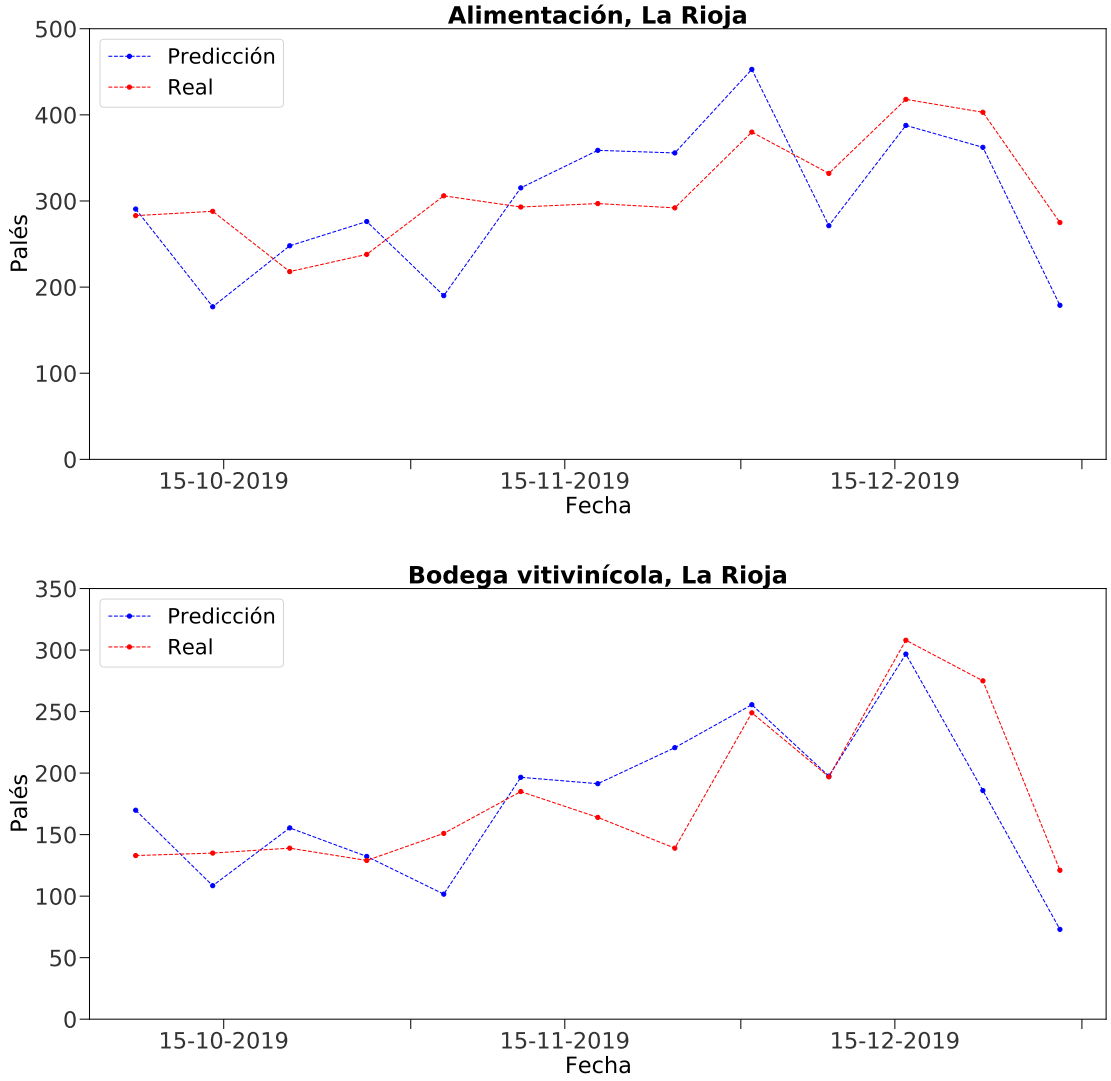

Figura 3.5: Comparación de las predicciones del número de palés en La Rioja con el algoritmo STL integrado con ARIMA frente al valor real para los sectores alimentación y bodega vitivinícola.

3.5. LSTM

La arquitectura de la red neuronal se diseñó en *Python* con *Keras* [29]. En primer lugar, se trató la serie temporal de datos históricos para convertirla en un problema de aprendizaje supervisado, es decir, se extrajo en una tabla el dato del número de palés en el paso $(t-1)$ y los palés en el siguiente paso t, que se corresponden con la variable objetivo que queremos predecir, donde t es el número de palés en una fecha concreta del histórico [30] [31]. El resultado obtenido se normalizó entre 0 y 1 con la función *MinMaxScaler* de *sklearn.preprocessing*.

Se definió un modelo secuencial para cada uno de los tipos de actividad a predecir: alimentación y bodega vitivinícola en La Rioja. Se incorporó una LSTM con 500 neuronas, un *dropout* de 0.2 y, a continuación, una capa densa con función de activación *ReLU*. Se entrenaron 50 épocas, con una

tasa de aprendizaje de 1×10^{-4}, se usó el optimizador *RMSProp*, puesto que era el que mejores resultados daba, y se fijó un tamaño del *mini-batch* de 32. También se separaron un $15\,\%$ de los datos de entrenamiento como subconjunto de validación.

Una vez entrenada la red, se llevaron a cabo las predicciones para los meses de octubre a diciembre de 2019. Los resultados obtenidos se reescalaron con la función *inverse_transform* y se muestran en la Figura 3.6.

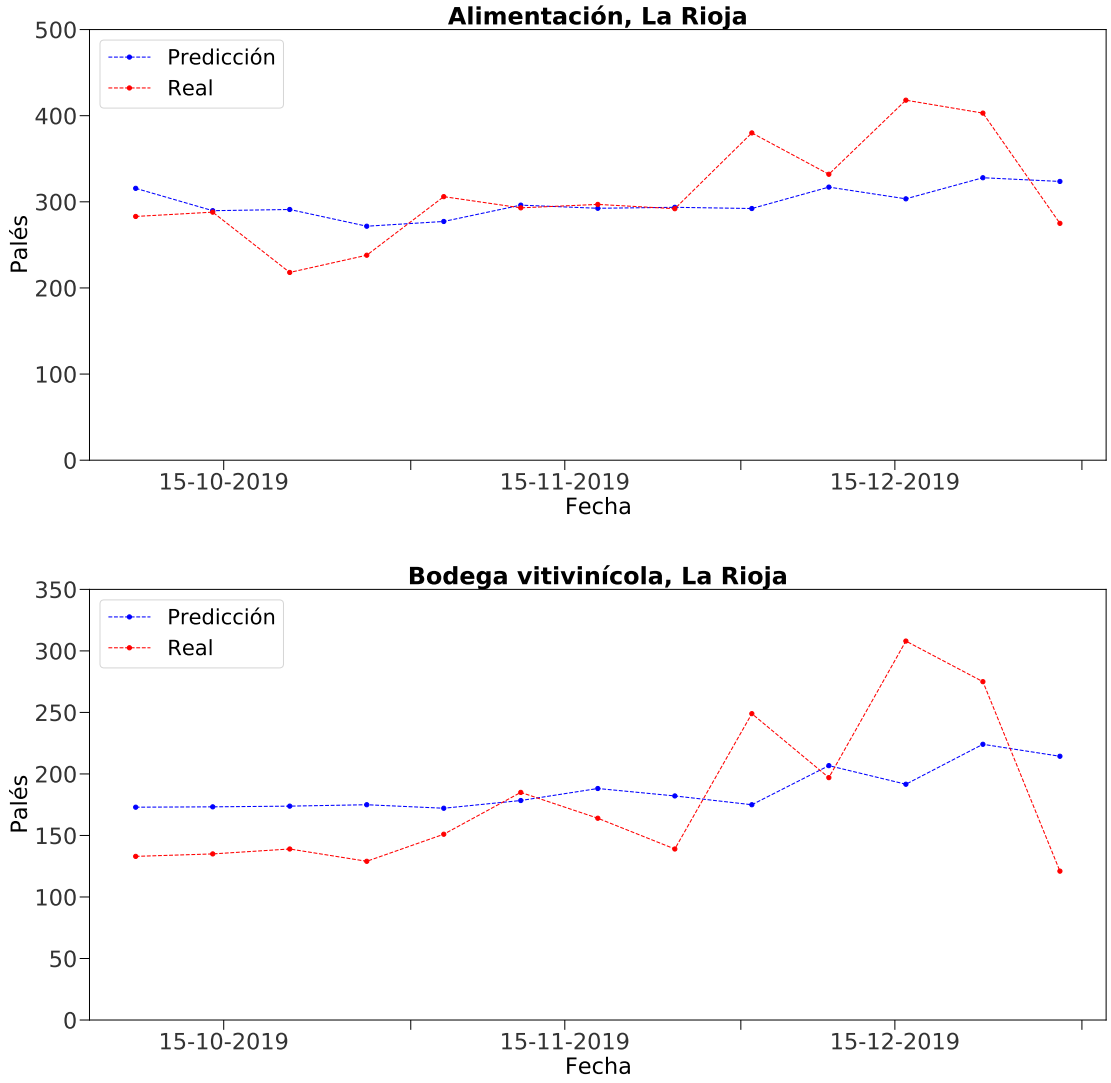

Figura 3.6: Comparación de las predicciones del número de palés en La Rioja obtenidas con la red neuronal LSTM frente al valor real para los sectores alimentación y bodega vitivinícola.

4. Discusión

A continuación se muestran las tablas del RMSE calculado para los periodos correspondientes a un mes (octubre 2019), dos meses (octubre y noviembre 2019) y, finalmente, el trimestre completo para el que se han realizado predicciones (de octubre a diciembre de 2019). Se discutirán por separado los resultados en cada sector analizado. También se explican las limitaciones de los diversos modelos implementados y, finalmente, se incluye la validación de los resultados por parte del cliente.

4.1. Evaluación de las predicciones en el sector alimentación

En este apartado se presenta la discusión de los resultados obtenidos con cada modelo y se analiza su fiabilidad de acuerdo con la métrica de validación elegida, el RMSE, calculado según la ecuación 2.1 del apartado 2.3. Con el objetivo de poner el error de las predicciones en contexto, se dará previamente una descripción de los datos.

El promedio de palés trasladados para el sector alimentación dentro del periodo 2015-2019 en La Rioja es 259, con una desviación estándar de 81, lo cual se corresponde con un error sobre la media del 31 % para el periodo completo. El valor mínimo fue 79 y el máximo 549. En la Tabla 4.1 se recogen los valores del RMSE obtenidos con cada modelo, en función del número de meses considerados.

Modelo	$RMSE_{1\text{mes}}$	$RMSE_{2\text{meses}}$	$RMSE_{3\text{meses}}$
Holt-Winters	72	65	60
Prophet	23	27	38
Prophet con *Time Series Split*	23	27	37
AR ($lag = 12$)	60	57	57
STL-ARIMA	61	68	66
LSTM	43	32	54

Tabla 4.1: Raíz del error cuadrático medio, RMSE, calculado para las predicciones frente a los valores reales del número de palés en el sector alimentación de La Rioja. Se recoge el RMSE en palés para cada modelo, considerando uno, dos y tres meses, respectivamente.

De acuerdo con los resultados de la Tabla 4.1, se tiene que el algoritmo que consiguió un menor RMSE en las predicciones tanto a uno como dos y tres meses fue Prophet. El error del número de palés sobre el valor promedio oscila entre el 9 % en el primer mes, 10 % en los dos primeros meses y 15 % para el trimestre completo. Por lo tanto, la precisión supera el 80 % en todos los casos, así que, según estos resultados, se tiene que en este caso Prophet es el modelo más fiable para hacer las predicciones del número de palés. Como ya se ha mencionado en la sección 3.2, el resultado de los pronósticos con Prophet es muy similar si se incorpora un *Time Series Split*.

Sin embargo, el resto de técnicas no han conseguido tan buenos resultados, con un error en los pronósticos a tres meses que se encuentra en el rango del $20 - 25$ %, aunque sigue siendo menor que el correspondiente a la desviación estándar sobre la media. El siguiente modelo que logra un RMSE más bajo después de Prophet es la red neuronal LSTM. Con la arquitectura de dicha red se consigue un error del 17 % el primer mes, 12 % los dos primeros meses y 21 % en el trimestre íntegro. Las dificultades que se han visto a la hora de ajustar los parámetros de la red LSTM para tratar de reducir el error en el conjunto de *test*, por la mayor complejidad de este modelo, se discutirán por separado en el apartado 4.3, donde se comentarán las limitaciones de los métodos implementados.

4.2. Evaluación de las predicciones en el sector bodega vitivinícola

La media de palés transportados para el sector bodega vitivinícola durante el periodo 2015-2019 en La Rioja es 134, con una desviación estándar de 43, que se corresponde con un error del 32 % sobre el valor promedio en todo el periodo considerado. El mínimo de palés semanal fue 36 y el máximo 309. La Tabla 4.2 recopila los valores del RMSE obtenidos con cada modelo, en función del número de meses considerados.

Modelo	$RMSE_{1\text{mes}}$	$RMSE_{2\text{meses}}$	$RMSE_{3\text{meses}}$
Holt-Winters	6	29	44
Prophet	6	24	40
Prophet con *Time Series Split*	6	24	39
AR ($lag = 14$)	12	32	42
STL-ARIMA	24	39	42
LSTM	40	34	55

Tabla 4.2: Raíz del error cuadrático medio, RMSE, calculado para las predicciones frente a los valores reales del número de palés en el sector bodega vitivinícola de La Rioja. Se recoge el RMSE en palés para cada modelo, considerando uno, dos y tres meses, respectivamente.

Según los valores del RMSE recogidos en la Tabla 4.2 se tiene que, de nuevo, el algoritmo Prophet es el que consigue minimizar el error en las predicciones, siendo este del 4 % cuando se consideran los pronósticos a un mes vista, 18 % en el caso de dos meses y, finalmente, 30 % si se toma el trimestre en su totalidad. Se observa que en este caso el error aumenta considerablemente cuanto mayor es el periodo de tiempo de pronóstico a futuro, a pesar de que no supera el valor correspondiente a la desviación estándar del número de palés sobre la media.

El principal motivo para que el error sea tan alto es que la serie temporal experimenta un pico de demanda a finales de año, lo cual Prophet no predice con suficiente precisión, como puede verse en la Figura 3.2 de la sección 3.2. La considerable dispersión de los datos recogida en el diagrama de caja del mes de diciembre en la Figura 2.12 del apartado 2.3 dificulta que los modelos consigan realizar predicciones con mayor precisión.

El siguiente método que sigue de cerca a Prophet en cuanto a buenos resultados del RMSE es Holt-Winters, que es capaz de capturar correctamente la estacionalidad e incluso pronosticar con bastante precisión los picos de demanda, como se mostró en la Figura 3.1 de la sección 3.1. Las predicciones a tres meses de AR y STL-ARIMA también tienen un error en torno al 31 %, similar al de Holt-Winters y Prophet.

Por último, se tiene que la red LSTM presenta un error muy alto en las previsiones a un mes, aproximadamente del 30 %. Como puede verse en la Figura 3.6 del apartado 3.5, las predicciones de octubre están a primera vista desplazadas verticalmente de forma sistemática en la gráfica. Las complicaciones surgidas con este modelo se desarrollarán en el siguiente apartado.

4.3. Limitaciones de los modelos

Los resultados presentados en los apartados previos muestran que el modelo capaz de predecir el número de palés con un menor RMSE es Prophet para los dos tipos de actividad analizados en La Rioja. En el caso de la serie temporal del sector alimentación, el resto de algoritmos basados en procedimientos estadísticos, es decir, todos los implementados excepto la red neuronal LSTM, no consiguen resultados tan ajustados como Prophet. En dichos métodos puede verse que los pronósticos son mejores cuando la componente estacional es más fuerte, como sucede en el sector bodega vitivinícola, para el cual la precisión a 3 meses es del mismo orden de magnitud en Holt-Winters, Prophet, AR y STL-ARIMA, como se recoge en la Tabla 4.2.

A diferencia de los procedimientos estadísticos, el principal problema afrontado con la red neuronal LSTM ha sido la ausencia de un mayor histórico de datos de la compañía, lo cual ha dificultado el entrenamiento de este modelo y ha supuesto que, a pesar de la mayor complejidad del método, no se haya conseguido un RMSE menor que el obtenido con algoritmos más sencillos. A este respecto, se disponía de 248 datos del histórico para el entrenamiento, de los cuales se separaron un $15\,\%$ en esta fase como conjunto de validación; y se realizaron predicciones para los tres últimos meses de 2019, como se hizo con el resto de modelos, que es el periodo correspondiente a 13 semanas.

La ausencia de una cantidad más significativa de datos históricos dificulta a la red LSTM extraer patrones de dependencia en la serie temporal a largo plazo, como la estacionalidad. Esto es más evidente en el caso del sector bodega vitivinícola ya que, como se vio en la Figura 2.13 de la sección 2.3, el factor estacional tiene mayor peso en esta serie que en la de alimentación, de acuerdo con la Figura 2.10.

En este sentido, una de las ventajas de las redes neuronales con respecto a los modelos estadísticos, en concreto, la posibilidad de ajustarse de forma automática, se convierte en un inconveniente cuando no se tiene suficiente histórico para entrenar, y no es posible introducir la estacionalidad de forma manual en la red. El principal motivo es que la capacidad de LSTM para modelar está limitada por su habilidad para encontrar patrones estacionales en los datos y reproducirlos en el modelo [32].

Se realizaron diferentes pruebas con varias arquitecturas de la red neuronal, con el objetivo de intentar reducir el error de las predicciones. Se probó incluyendo dos capas de LSTM pero, sin embargo, se observó que este diseño no disminuía el RMSE de forma significativa, por lo que finalmente se descartó. Lo mismo sucedía si se incorporaba una tercera capa de LSTM. Esta circunstancia puede atribuirse al overfitting de la red, lo cual es consecuencia de la falta de datos suficientes para el entrenamiento.

Al mismo tiempo, se modificó el número de neuronas por capa pero, de nuevo, se encontró que no mejoraban de forma significativa los resultados. Lo que se detectó fue que al descender el número de neuronas era necesario aumentar las épocas que se entrenaba la red, como cabía esperar, ya que, en esta situación, al ser una arquitectura más sencilla, son necesarios más pasos para entrenar. En cualquier caso, las predicciones seguían teniendo un RMSE muy similar al que se recoge en las tablas 4.1 y 4.2. Se prefirió definir el modelo con un número alto de neuronas en la capa LSTM porque así se observó que la curva de aprendizaje era más estable y presentaba menos fluctuaciones.

Cabe destacar que las redes neuronales son capaces de encontrar eficientemente relaciones no lineales. No obstante, según hemos visto, las series temporales con las que hemos trabajado exhibían mayoritariamente patrones de tipo estacional, en lugar de cambios bruscos de tendencia asociados a efectos no lineales, cuyo resultado habría sido que las variaciones estacionales se modificasen de manera proporcional al nivel de la serie. En esta situación, no está realmente justificada la necesidad de desarrollar un modelo tan sofisticado como una red neuronal.

Una posibilidad para mejorar las previsiones con este procedimiento sería implementar un modelo ARIMA con estacionalidad, es decir, SARIMA (*Seasonal Autorregresive Integrated Moving Average*) [20], que realizase los pronósticos de los términos de tendencia y estacionalidad, mientras que el factor aleatorio podría predecirse con una LSTM [33]. Sin embargo, teniendo en cuenta la calidad de los resultados obtenidos con algoritmos más sencillos, cabe señalar que esta solución sería más indicada para series temporales con patrones menos estables o mayores fluctuaciones, y que incorporaran correlaciones con otras variables.

4.4. Validación de los resultados con el cliente

Se realizó una reunión con la empresa de transporte logístico cliente para la que se llevó a cabo el proyecto. Se presentaron los resultados de las predicciones de Prophet en todas las provincias y tipos de actividad que disponían de datos históricos, puesto que este modelo como se ha visto es el que ha conseguido predicciones más ajustadas.

El cliente aprobó la agrupación semanal de los datos, ya que así podía realizar su planificación del volumen de palés previsto para la semana correspondiente. Además, esta granularidad le posibilitaba ver tendencias y estacionalidad, mientras que, por un lado, a nivel diario el número de palés presentaba grandes fluctuaciones y, por otro lado, a nivel mensual se perdía la capacidad de pronóstico a corto plazo y, al mismo tiempo, se perdía información sobre patrones de estacionalidad. También validó las predicciones si el error sobre la media era inferior al $20\,\%$, lo cual, como se ha explicado en los apartados 4.1 y 4.2, se ha conseguido en las previsiones hasta tres meses vista del sector alimentación, y dos meses en el caso de bodega vitivinícola.

5. Conclusiones

Se ha llevado a cabo un proyecto de predicción de demanda para una empresa de transporte logístico. El proyecto aporta soluciones a un desafío habitual al que se enfrentan este tipo de compañías, que es la predicción de demanda de sus servicios, lo que les permite, por ejemplo, mejorar su rentabilidad y optimizar estrategias comerciales.

Se han aplicado conocimientos adquiridos de forma teórica en el máster, comenzando por la realización del proceso de ETL de las tablas del histórico de datos de la compañía. A continuación, se visualizaron diferentes gráficos con el fin de establecer el alcance del problema y realizar un primer análisis de las series temporales con las que se iban a entrenar posteriormente los modelos.

Se han propuesto distintos algoritmos para realizar las previsiones de demanda de palés, entre los cuales se incluyen procedimientos estadísticos y técnicas de aprendizaje automático, que en su mayoría no se conocían antes de realizar este trabajo. Los modelos se han entrenado con los datos históricos de la provincia de La Rioja en los tipos de actividad alimentación y bodega vitivinícola. El resultado más fiable se ha obtenido con Prophet, que ha conseguido un RMSE más bajo. En las previsiones a un mes vista el error cometido ha sido inferior al $10\,\%$ para ambos sectores, por lo que se ha conseguido una precisión ajustada con este modelo.

Se ha comprobado que las predicciones son más fiables cuanto menor es el rango temporal considerado, ya que conforme aumenta el periodo de pronóstico se incrementa el RMSE, como cabía esperar. Por lo tanto, se tiene que los modelos aplicados tienen mayor precisión en las previsiones a uno o dos meses vista, mientras que en periodos superiores son válidos para realizar estimaciones de demanda basadas en el comportamiento del histórico de datos en el pasado.

Como ya se hizo hincapié en la introducción de esta memoria, debe tenerse en cuenta que las previsiones de demanda se ven muy influenciadas a corto plazo por factores locales relativamente volátiles, y en general difíciles de predecir, que impiden el incremento de precisión en los pronósticos a corto plazo. Además, cuanto menos estables son los patrones que presentan los datos, más complicado resulta hacer las predicciones. A este respecto, las dos series temporales analizadas presentaban características diferentes, dado que en el caso del sector bodega vitivinícola los patrones de estacionalidad eran más marcados que en alimentación.

Asimismo, en el momento de poner en producción los modelos explicados con el objetivo de predecir todas las provincias y tipos de actividad de la compañía debe valorarse, aparte de la fiabilidad del modelo, su eficiencia en términos computacionales. En este sentido, se tiene que los algoritmos que implementan SARIMA o redes neuronales resultan más costosos computacionalmente, y requieren un mayor tiempo de entrenamiento que algoritmos más eficaces como Prophet, por lo que si los errores son similares es preferible optar por el método más eficiente. Teniendo en cuenta que las predicciones con Prophet haciendo *Time Series Split* fueron muy parecidas a las obtenidas sin este procedimiento, lo más conveniente sería usar Prophet para predecir todo el periodo deseado en un único paso.

El trabajo que se ha realizado tiene una importante aplicabilidad en la vida real, ya que da solución a un problema de predicción de demanda de una empresa de transporte logístico, constituyendo un proyecto en sí mismo, llevado a cabo en una consultora logística. El cliente validó los resultados finales en una reunión en la que se mostró satisfecho con las predicciones de Prophet.

En resumen, se ha llevado a cabo un proyecto de *Data Science* de principio a fin, desarrollando las competencias adquiridas en el máster y complementándolas con su aplicación práctica en un problema real, incluyendo también el aprendizaje de nuevos algoritmos y técnicas de visualización de los datos, por ejemplo mediante análisis de correlaciones, siguiendo procedimientos que no se habían impartido en el máster.

Trabajo futuro

Los modelos estudiados se han aplicado en dos sectores de una provincia concreta, La Rioja. En futuros desarrollos del proyecto, el trabajo llevado a cabo sería escalable a cualquier provincia y/o tipo de actividad, siguiendo el mismo procedimiento descrito: visualización de la serie temporal, ajuste de los parámetros de los modelos, realización de las predicciones y validación de los resultados.

Asimismo, podrían implementarse nuevos métodos y algoritmos, como el modelo híbrido SARIMA combinado con LSTM o las máquinas de vectores soporte para regresión, SVR (del inglés, *Support Vector Regression*) [17]. También podrían incorporarse variables exógenas a los algoritmos, para lo cual habría que realizar un estudio previo de correlaciones con el objetivo de encontrar aquellas con mayor capacidad predictiva. Algunas candidatas para este caso serían, por ejemplo, el IPI o el PIB, cuyo histórico puede encontrarse en la página web del INE. Se debe tener en cuenta que los valores a futuro deben ser pronosticados previamente para poder ser introducidas en los modelos como variables exógenas.

Finalmente, el estudio de los principales algoritmos de predicción de demanda que se ha llevado a cabo puede adaptarse a gran variedad de problemas de este tipo, no sólo en el ámbito del transporte logístico, sino también en otras áreas, siguiendo el procedimiento de visualización, análisis y parametrización de los modelos que se ha desarrollado a lo largo de este proyecto.

Bibliografía

[1] T. Tsekeris and C. Tsekeris, "Demand forecasting in transport: Overview and modeling advances," *Economic research-Ekonomska istraživanja*, vol. 24, no. 1, pp. 82–94, 2011.

[2] D. Wang, W. Chen, H. Shi, X. Fang, and W. Luo, "Forecasting inter-urban transport demand for a logistics company: A combined grey–periodic extension model with remnant correction," *Advances in mechanical engineering*, vol. 7, no. 12, p. 168781401562007, 2015.

[3] "KNIME." https://www.knime.com/, fecha de consulta: 04/04/2021.

[4] "SPSS tutorials: Pearson correlation." https://libguides.library.kent.edu/SPSS/PearsonCorr, fecha de consulta: 14/04/2021.

[5] "Time series analysis with theory, plots and code. Part 1." https://towardsdatascience.com/time-series-analysis-with-theory-plots-and-code-part-1-dd3ea417d8c4, fecha de consulta: 07/04/2021.

[6] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, 2003.

[7] C. Chatfield and M. Yar, "Holt-Winters forecasting: Some practical issues," *Journal of the Royal Statistical Society. Series D (The Statistician)*, vol. 37, no. 2, pp. 129–140, 1988.

[8] C. C. Holt, "Forecasting seasonals and trends by exponentially weighted moving averages," *International journal of forecasting*, vol. 20, no. 1, pp. 5–10, 2004.

[9] P. R. Winters, "Forecasting sales by exponentially weighted moving averages," *Management science*, vol. 6, no. 3, pp. 324–342, 1960.

[10] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*. OTexts, disponible en https://otexts.com/fpp2/, 2018.

[11] C. Chatfield, "The Holt-Winters forecasting procedure," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 27, no. 3, pp. 264–279, 1978.

[12] "Prophet. forecasting at scale." https://facebook.github.io/prophet/, fecha de consulta: 06/04/2021.

[13] S. J. Taylor and B. Letham, "Forecasting at scale," *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018.

[14] "Facebook Research. Prophet: forecasting at scale." https://research.fb.com/blog/2017/02/prophet-forecasting-at-scale/, fecha de consulta: 06/04/2021.

[15] "Vector Autoregressive models VAR(p) models." https://online.stat.psu.edu/stat510/lesson/11/11.2, fecha de consulta: 11/04/2021.

[16] R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning, "STL: A seasonal-trend decomposition," *Journal of official statistics*, vol. 6, no. 1, pp. 3–73, 1990.

[17] A. R. S. Parmezan, V. M. Souza, and G. E. Batista, "Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model," *Information sciences*, vol. 484, pp. 302–337, 2019.

[18] "Statsmodels.tsa.forecasting.stl.stlforecast." https://www.statsmodels.org/stable/generated/statsmodels.tsa.forecasting.stl.STLForecast.html, fecha de consulta: 15/04/2021.

[19] G. E. P. Box, *Time series analysis: forecasting and control*. Hoboken, New Jersey: Wiley, 5th ed., 2016.

[20] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "A comparison of ARIMA and LSTM in forecasting time series," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1394–1401, IEEE, 2018.

[21] I. A. Ibrahim and M. J. Hossain, "LSTM neural network model for ultra-short-term distribution zone substation peak demand prediction," pp. 1–5, IEEE, 2020.

[22] A. Yadav, C. Jha, and A. Sharan, "Optimizing LSTM for time series prediction in Indian stock market," *Procedia Computer Science*, vol. 167, pp. 2091–2100, 2020.

[23] "Statsmodels.tsa.holtwinters.exponentialsmoothing." https://www.statsmodels.org/dev/generated/statsmodels.tsa.holtwinters.ExponentialSmoothing.html, fecha de consulta: 11/04/2021.

[24] "Prophet. Flat trend and custom trends." https://facebook.github.io/prophet/docs/additional_topics.html#flat-trend-and-custom-trends, fecha de consulta: 07/04/2021.

[25] "Sklearn.model_selection.timeseriessplit." https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html, fecha de consulta: 19/04/2021.

[26] "Statsmodels.tsa.ar_model.autoreg." https://www.statsmodels.org/stable/generated/statsmodels.tsa.ar_model.AutoReg.html, fecha de consulta: 15/04/2021.

[27] "Statsmodels.tsa.vector_ar.var_model.var." https://www.statsmodels.org/stable/generated/statsmodels.tsa.vector_ar.var_model.VAR.html, fecha de consulta: 15/04/2021.

[28] "Statsmodels.tsa.arima.model.arima." https://www.statsmodels.org/stable/generated/statsmodels.tsa.arima.model.ARIMA.html, fecha de consulta: 15/04/2021.

[29] "Keras." https://keras.io/, fecha de consulta: 19/04/2021.

[30] "Time series forecasting using LSTM." https://www.kaggle.com/gurpreetmohaar/time-series-forecasting-using-lstm, fecha de consulta: 15/04/2021.

[31] "How to convert a time series to a supervised learning problem in Python." https://machinelearningmastery.com/convert-time-series-supervised-learning-problem-python/, fecha de consulta: 19/04/2021.

[32] T.-W. Yoo and I.-S. Oh, "Time series forecasting of agricultural products' sales volumes based on seasonal Long Short-Term Memory," *Applied Sciences*, vol. 10, no. 22, p. 8169, 2020.

[33] Q. Sun, J. Wan, and S. Liu, "Estimation of sea level variability in the China sea and its vicinity using the SARIMA and LSTM models," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 3317–3326, 2020.