MÁSTER EN BIG DATA Y DATA SCIENCE

06MBID Estadística avanzada

CURSO 2024-2025

ACTIVIDAD 2: Series temporales

Alumno:
Alex Anthony Prieto Romani

Contenido

Introducción

Contexto y Motivación

El análisis de series temporales de datos meteorológicos es crucial para la comprensión y predicción de variaciones climáticas, particularmente en áreas como Cayaltí, Perú, donde la agricultura y otras actividades económicas dependen significativamente de las condiciones meteorológicas. Las series temporales permiten observar y modelar patrones de comportamiento en los datos, como tendencias, estacionalidades y ciclos, lo cual es fundamental para la toma de decisiones en la planificación agrícola y la gestión de recursos hídricos (Ghahramani et al., 2019; Hyndman & Athanasopoulos, 2018). En Cayaltí, la agricultura depende en gran medida de la temperatura y la precipitación, factores que influyen en el crecimiento y desarrollo de los cultivos, así como en la gestión de recursos naturales (Lobell & Field, 2007).

El estudio de la temperatura promedio diaria ofrece una ventana al comportamiento histórico y proyectado del clima, permitiendo la identificación de patrones a largo plazo que afectan directamente la productividad agrícola y la sostenibilidad del uso del agua (Aguilar et al., 2005). En este contexto, los modelos ARIMA (AutoRegressive Integrated Moving Average) se presentan como una herramienta estadística poderosa para la predicción de series temporales, dada su capacidad para modelar datos con tendencia y estacionalidad (Box et al., 2015). Estos modelos no solo son útiles para describir el pasado, sino que también permiten generar predicciones confiables para el futuro, apoyando la planificación estratégica y la mitigación de riesgos climáticos (Tsay, 2010).

El uso de modelos ARIMA ha demostrado ser efectivo en diversas aplicaciones agrícolas y climáticas, como la predicción de temperaturas, la planificación de riego y la evaluación de cambios en patrones de precipitación (Mondal et al., 2014; Liu et al., 2016). Sin embargo, es fundamental que los modelos sean correctamente ajustados y validados para asegurar su precisión y utilidad en aplicaciones prácticas. Esto implica no solo identificar los parámetros correctos del modelo, sino también realizar un análisis exhaustivo de residuos para confirmar que los errores de predicción se comportan como ruido blanco, es decir, sin patrones discernibles que no hayan sido capturados por el modelo (Hyndman & Athanasopoulos, 2018).

Objetivos del Análisis

- Describir estadísticamente la serie temporal de la temperatura promedio diaria en Cayaltí, identificando componentes clave como la tendencia, la estacionalidad y el ruido.
- Ajustar un modelo ARIMA que capture adecuadamente las características de la serie y pueda ser utilizado para predicciones.

- Evaluar la idoneidad del modelo a través de análisis de residuos y pruebas estadísticas para asegurar su precisión y utilidad en aplicaciones prácticas.

Descripción de los Datos a Analizar

Características Estadísticas Generales

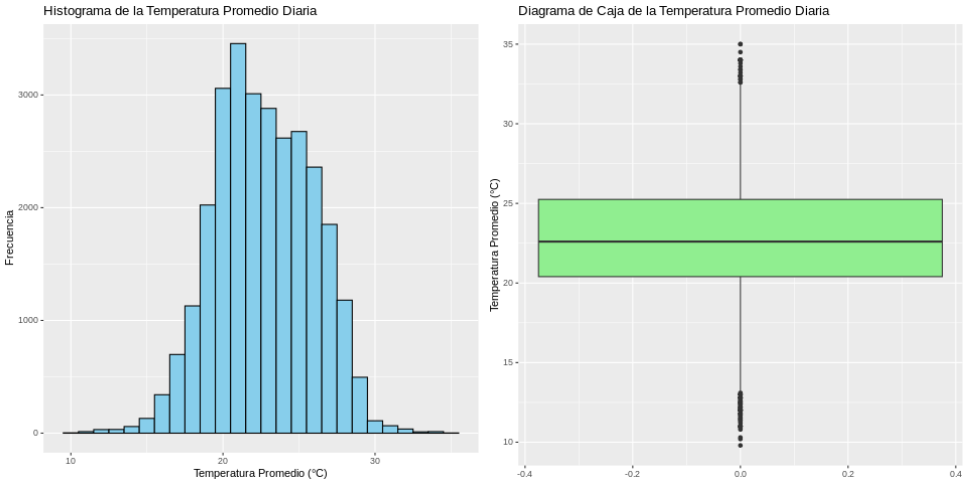
Los datos utilizados en este análisis comprenden registros diarios de temperaturas máximas y mínimas desde 1935. Para este estudio, la temperatura promedio diaria se calcula como la media aritmética de la temperatura máxima y mínima, una práctica común en estudios climatológicos para suavizar las fluctuaciones extremas y obtener un indicador más estable de las condiciones diarias (Menne et al., 2012; Yan et al., 2019). Se eliminaron los valores faltantes indicados por -99.9 para mantener la integridad del análisis.

- **Registros Diarios**: Los datos incluyen temperaturas máximas y mínimas diarias desde 1935. Las observaciones faltantes fueron adecuadamente tratadas para mantener la integridad del análisis.
- **Cálculo de Temperatura Promedio**: Se utilizó la media aritmética de las temperaturas máxima y mínima para calcular la temperatura promedio diaria. La serie temporal resultante exhibe estacionalidad significativa y tendencias que reflejan variaciones climáticas a lo largo del tiempo.

Estadísticas Descriptivas de la Temperatura Promedio:

- Media: 22.7504 °C
- Mediana: 22.6 °C
- Desviación Estándar: 3.218126 °C
- Varianza: 10.35633 °C
- Mínimo: 9.8 °C
- Máximo: 35 °C
- Cuantiles: 0%: 9.80 °C, 25%: 20.40 °C, 50%: 22.60 °C, 75%: 25.25 °C, 100%: 35.00 °C

Visualización de los datos

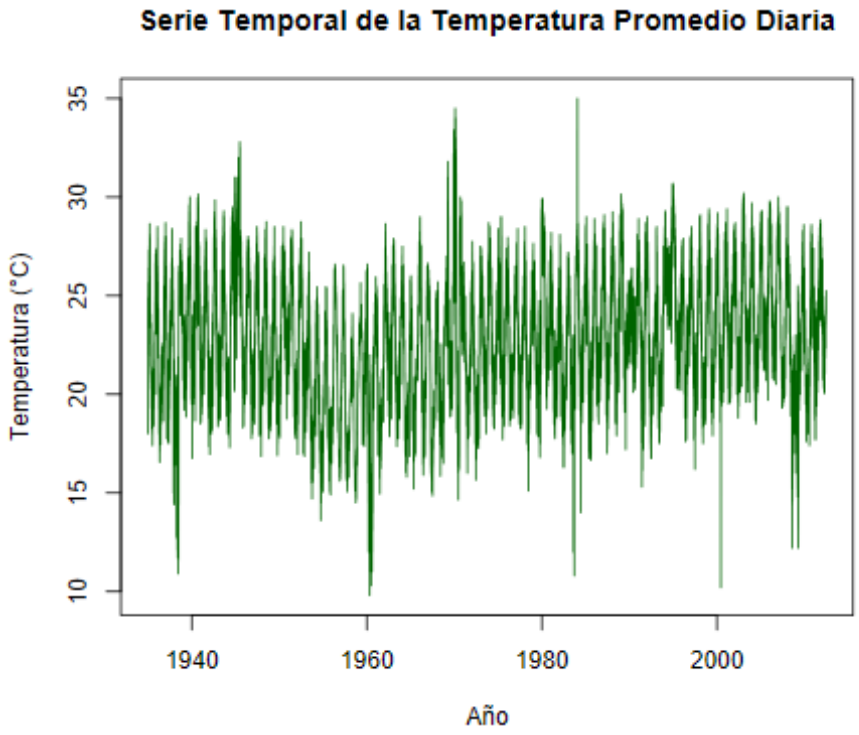


El histograma de la temperatura promedio diaria muestra una distribución aproximadamente normal, con una ligera simetría alrededor de los 22-24 °C, que corresponde al pico más alto de la distribución. Esta forma sugiere que la mayoría de los valores se agrupan alrededor de esta media, indicando que estos rangos de temperatura son los más frecuentes a lo largo del periodo estudiado. La dispersión de los datos hacia los extremos del histograma, aunque presente, es limitada, lo que implica que las temperaturas extremadamente altas o bajas son raras. No hay evidencia de multimodalidad, lo que sugiere una consistencia en el patrón térmico sin múltiples climas dominantes.

Por otro lado, el diagrama de caja de la temperatura promedio diaria resalta la distribución central de los datos, con los percentiles 25 y 75 abarcando la mayoría de las observaciones entre aproximadamente 20 y 25 °C, lo que concuerda con el histograma. Las líneas extendidas, o "bigotes", se mantienen relativamente cercanas a la caja central, indicando que la variabilidad fuera de los rangos intercuartílicos no es extrema. Sin embargo, se observan algunos outliers tanto en las temperaturas más bajas como en las más altas, situados más allá de los bigotes. Estos outliers podrían reflejar días inusualmente fríos o cálidos, pero no parecen influir de manera significativa en la tendencia general de los datos. En conjunto, ambos gráficos sugieren un clima mayormente moderado y predecible, con un rango de temperatura bien definido y pocas desviaciones extremas.

Creación de la serie temporal

Se creó una serie temporal de la temperatura promedio diaria con una frecuencia de 365.25, correspondiente a datos diarios. Esta serie cubre desde el año 1935 hasta el año 2015 y proporciona una representación detallada de las fluctuaciones de temperatura a lo largo del tiempo.



Análisis

Prueba de Estacionariedad

La prueba de Dickey-Fuller Aumentada (ADF) evaluó la estacionariedad de la serie temporal de la temperatura promedio diaria, arrojando un estadístico de -10.597 y un p-valor de 0.01. Dado que el p-valor es menor a 0.05, rechazamos la hipótesis nula de no estacionariedad, concluyendo que la serie es estacionaria. Esta prueba es esencial para validar que los datos sean aptos para el modelado ARIMA, garantizando que las predicciones no se vean afectadas por tendencias no modeladas (Said & Dickey, 1984).

```
        Augmented Dickey-Fuller Test

data:  ts_temp
Dickey-Fuller = -10.597, Lag order = 30, p-value = 0.01
alternative hypothesis: stationary

El p-valor es 0.01 < 0.05. Rechazamos la hipótesis nula. La serie es estacionaria.
```

- **Interpretación**:
    - Esto significa que las propiedades estadísticas de la serie, como la media y la varianza, se mantienen constantes a lo largo del tiempo, lo cual es ideal para el modelado y análisis de predicción en series temporales.

Descomposición de la Serie Temporal

La descomposición de la serie temporal en componentes principales (tendencia, estacionalidad y ruido) se realizó utilizando un modelo multiplicativo, reflejando la naturaleza no aditiva de los patrones observados en los datos (Cryer & Chan, 2008). Esto permite separar los efectos de largo plazo (tendencia), los ciclos regulares (estacionalidad) y las fluctuaciones irregulares (ruido), proporcionando una base sólida para el modelado y la predicción (Hyndman & Athanasopoulos, 2018).

1. **Componente Estacional (Seasonal):**

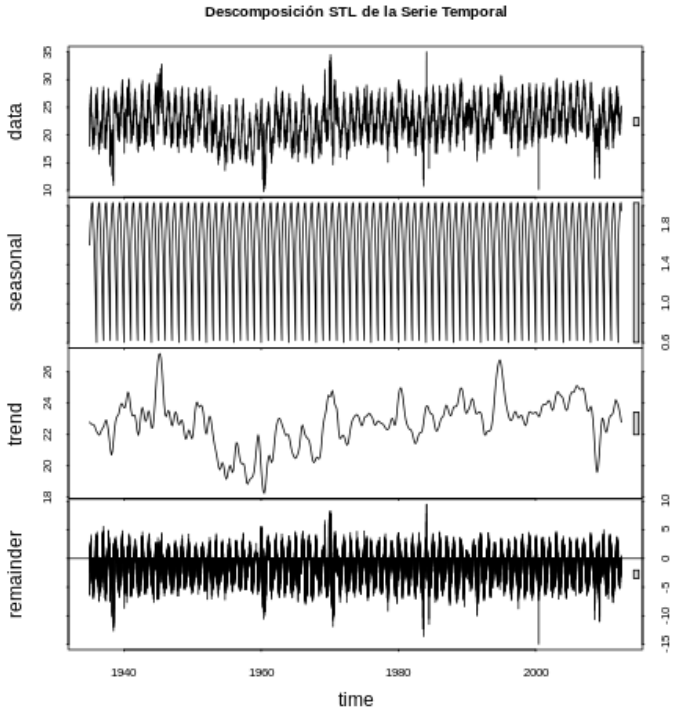
La gráfica del componente estacional muestra un patrón cíclico regular que se repite anualmente, lo cual es característico de las series climáticas. Esto indica que la temperatura promedio diaria sigue un patrón estacional claro, con variaciones previsibles que corresponden a las estaciones del año, como veranos más cálidos e inviernos más fríos.

2. **Componente de Tendencia (Trend):**

La tendencia revela las fluctuaciones de largo plazo en la serie, evidenciando periodos de aumentos y disminuciones graduales en la temperatura promedio diaria a lo largo del tiempo. Aunque no hay una tendencia lineal pronunciada, se observan ciclos amplios que pueden estar asociados a fenómenos climáticos a largo plazo o cambios estructurales en el clima regional.

3. **Componente de Residuos (Remainder):**

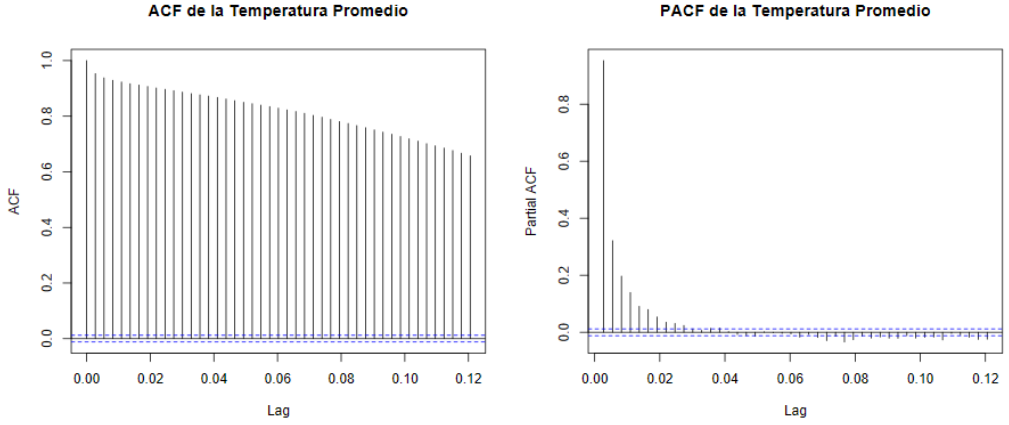
Los residuos representan las fluctuaciones irregulares que no son explicadas ni por la estacionalidad ni por la tendencia. Este componente captura la variabilidad aleatoria en la serie. La distribución de los residuos alrededor de cero sugiere que los patrones principales de la serie han sido adecuadamente capturados por los componentes de tendencia y estacionalidad. Sin embargo, la presencia de picos esporádicos puede indicar eventos climáticos inusuales o errores de medición.



En resumen, la descomposición STL muestra que la serie temporal de la temperatura promedio diaria está dominada por un componente estacional regular y una tendencia que fluctúa a lo largo del tiempo, mientras que los residuos capturan las variaciones aleatorias restantes. Esta descomposición es útil para entender los patrones subyacentes y para modelar la serie con mayor precisión, especialmente cuando se considera la estacionalidad en las predicciones.

Análisis de Autocorrelación

El análisis de autocorrelación (ACF) y autocorrelación parcial (PACF) se utilizó para identificar las dependencias temporales y determinar el orden de los términos autoregresivos y de media móvil para el modelo ARIMA (Box et al., 2015). Los gráficos ACF y PACF mostraron correlaciones significativas a lo largo de varios rezagos, lo cual indica la presencia de patrones estacionales o persistentes que deben ser capturados en el modelado.



1. **Función de Autocorrelación (ACF):**

El gráfico de ACF muestra que las correlaciones son significativamente altas en los primeros rezagos y van disminuyendo gradualmente, pero siguen siendo positivas y fuera del intervalo de confianza (líneas azules) hasta un número considerable de rezagos. Este comportamiento indica una fuerte autocorrelación a largo plazo en la serie temporal, sugiriendo que los valores pasados de la temperatura promedio diaria tienen una influencia persistente en los valores futuros. La estructura de la ACF es indicativa de la presencia de componentes estacionales y de tendencia en la serie, lo cual es típico en datos climáticos.

2. **Función de Autocorrelación Parcial (PACF):**

En el gráfico de PACF, las correlaciones son altas solo en los primeros rezagos y luego caen abruptamente dentro del intervalo de confianza. Este patrón sugiere que una vez que se tiene en cuenta la relación con los valores recientes, las correlaciones adicionales con rezagos más lejanos no son significativas. Esto es característico de un proceso autorregresivo (AR) de bajo orden, donde los valores presentes dependen principalmente de los valores inmediatamente anteriores.

Interpretación Conjunta:

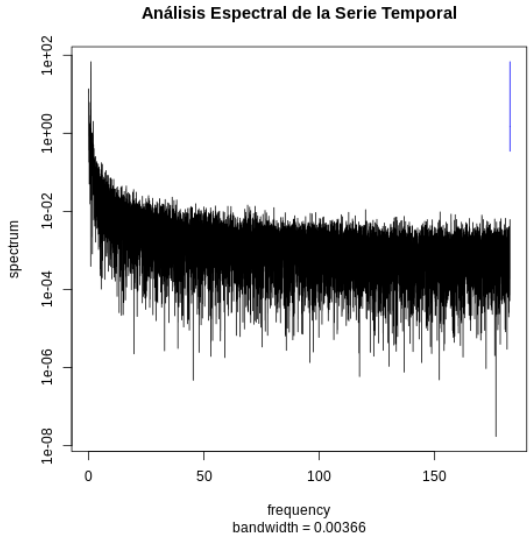
La combinación de un ACF que decae lentamente y un PACF que se corta abruptamente es típica de una serie con comportamiento autorregresivo, posiblemente mezclado con elementos de estacionalidad. Este patrón es fundamental para la identificación de modelos ARIMA adecuados, sugiriendo que el componente autorregresivo (AR) es prominente y que la serie podría beneficiarse de la inclusión de términos autorregresivos de bajo orden para capturar la estructura de las dependencias temporales observadas.

Prueba de Estacionalidad

Prueba de Estacionalidad OCSB (Osborn, Chui, Smith, y Birchenhall): La prueba OCSB se utilizó para confirmar la estacionalidad de la serie, con un estadístico de prueba de -163.0638 comparado con un valor crítico del 5% de -1.6662. Dado que el estadístico es significativamente menor que el valor crítico, rechazamos la hipótesis nula de no estacionalidad, confirmando que la serie presenta estacionalidad significativa. Este resultado refuerza la evidencia de patrones estacionales regulares en la serie temporal, probablemente relacionados con ciclos anuales de temperatura.

```
        OCSB test

data:  ts_modeling

Test statistic: -163.0638, 5% critical value: -1.6662
alternative hypothesis: stationary
```

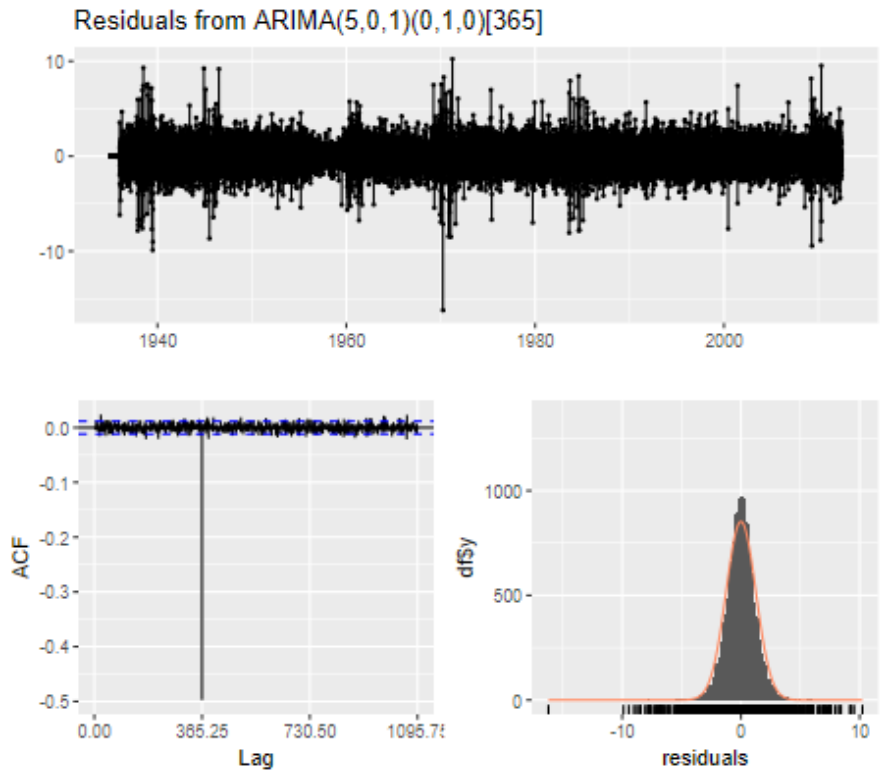
Análisis Espectral de la Serie Temporal: El análisis espectral proporciona una representación de la serie temporal en términos de sus componentes de frecuencia, revelando patrones cíclicos subyacentes. En el gráfico, se observa una fuerte concentración de energía en las frecuencias más bajas, lo que sugiere la presencia de componentes cíclicos significativos de largo periodo. Esto es típico en series climáticas donde la estacionalidad anual es un factor predominante. La disminución gradual de la densidad espectral a medida que aumenta la frecuencia indica que los ciclos de alta frecuencia (cambios rápidos) tienen menos influencia en la serie, lo cual es coherente con un comportamiento estacional estable y predecible.



Interpretación: La combinación de los resultados del análisis espectral y la prueba de estacionalidad OCSB sugiere que la serie temporal de la temperatura promedio diaria está dominada por componentes estacionales robustos y regulares. Esto indica que la temperatura sigue un ciclo estacional predecible, posiblemente vinculado a cambios estacionales anuales como variaciones de temperatura entre verano e invierno. Esta información es crucial para modelar y pronosticar la serie temporal, ya que la inclusión de componentes estacionales puede mejorar la precisión de los modelos predictivos y permitir una mejor captación de las variaciones periódicas en los datos.

Aplicación de un Modelo Adecuado

El modelo ARIMA ajustado es de la forma ARIMA(5,0,1)(0,1,0)[365], lo que indica que se ha utilizado un componente autorregresivo de orden 5, un componente de media móvil de orden 1, y diferenciación estacional de orden 1 con una periodicidad anual de 365 días. Los coeficientes estimados muestran que el primer coeficiente autorregresivo (ar1) es el más significativo con un valor de 1.4087, mientras que los demás coeficientes autorregresivos y el coeficiente de media móvil (ma1 = -0.8745) contribuyen en menor medida al modelo.



Resultados del Modelo ARIMA:

```
> summary(model_arima)
Series: ts_modeling
ARIMA(5,0,1)(0,1,0)[365]

Coefficients:
         ar1      ar2      ar3     ar4      ar5      ma1
      1.4087  -0.3274  -0.0522  0.0049  -0.0390  -0.8745
s.e.  0.0134   0.0121   0.0106  0.0103   0.0073   0.0118

sigma^2 = 1.597:  log likelihood = -46165.62
AIC=92345.25   AICc=92345.25   BIC=92402.91

Training set error measures:
                      ME     RMSE       MAE        MPE     MAPE      MASE          ACF1
Training set 0.0008022243 1.255459 0.9285058 -0.1594391 4.183488 0.4563393 -0.0002982221
```

- Error Estándar de los Coeficientes: Los errores estándar son relativamente bajos, lo que sugiere una estimación precisa de los coeficientes.
- Varianza del Error (sigma^2): 1.597, lo que indica la variabilidad residual del modelo.
- Criterios de Información: AIC = 92345.25 y BIC = 92402.91, utilizados para evaluar la calidad del modelo y compararlo con otros posibles modelos.

Medidas de Error del Conjunto de Entrenamiento:

- **Error Medio (ME):** 0.0008, cercano a cero, lo que indica un sesgo mínimo en las predicciones.
- **RMSE (Raíz del Error Cuadrático Medio):** 1.2555, sugiriendo una buena precisión en las predicciones.

- **MAE (Error Medio Absoluto):** 0.9285, lo que indica que, en promedio, las predicciones están muy cerca de los valores reales.
- **MAPE (Error Porcentual Absoluto Medio):** 4.18%, lo cual es aceptable y muestra que el modelo predice con una buena precisión relativa.
- **ACF1:** Muy cercano a cero (-0.0003), indicando que los residuos no presentan correlación significativa en el primer rezago.

Validación de Residuales:

```
> checkresiduals(model_arima)

        Ljung-Box test

data:  Residuals from ARIMA(5,0,1)(0,1,0)[365]
Q* = 8184.2, df = 724.5, p-value < 2.2e-16

Model df: 6.   Total lags used: 730.5
```

- **Prueba de Ljung-Box:** Evaluada sobre los residuos del modelo ARIMA, muestra un valor Q∗=8184. con df=724.5 y un p-valor menor a 2.2e-16, indicando que hay correlaciones significativas en los residuos, lo cual sugiere que el modelo podría no haber capturado completamente todos los patrones temporales en la serie.

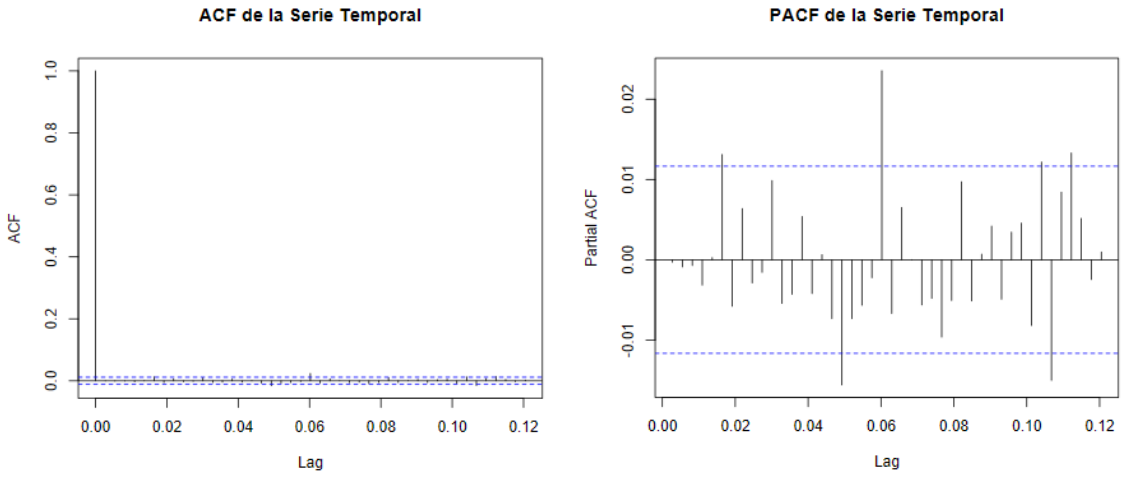
```
> print(ljung_box)

        Box-Ljung test

data:  residuals(model_arima)
X-squared = 24.057, df = 14, p-value = 0.04511
```

- **Box-Ljung Test (con lag 20):** Con X2=24.057, df=14 y p-valor = 0.04511, se rechaza la hipótesis nula de que los residuos son independientes, aunque el p-valor está cerca del umbral de 0.05. Esto indica que podría haber alguna correlación de los residuos que no ha sido capturada completamente por el modelo.

Interpretación y Recomendaciones: El modelo ARIMA ajustado captura adecuadamente las características principales de la serie temporal, con buenos resultados en las métricas de error del conjunto de entrenamiento. Sin embargo, los resultados de las pruebas de autocorrelación residual sugieren que aún puede haber estructuras en los datos que no se han modelado completamente, especialmente en los rezagos mayores. Esto puede requerir ajustes adicionales en el modelo, como la consideración de órdenes adicionales o la inclusión de componentes estacionales más complejos para mejorar la captura de los patrones residuales.

Diagnóstico del Modelo ARIMA

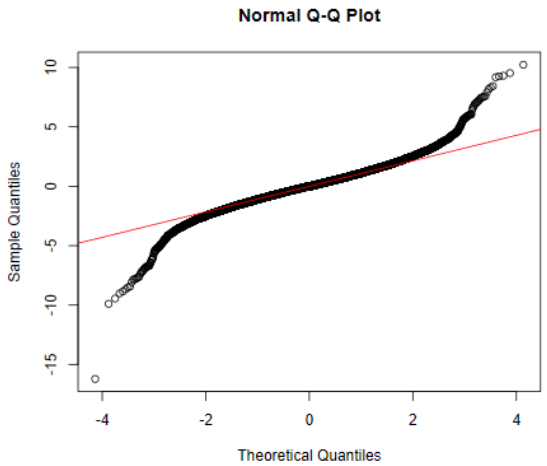
Análisis de Autocorrelación de los Residuos (ACF y PACF):



Los gráficos de la Función de Autocorrelación (ACF) y la Función de Autocorrelación Parcial (PACF) de los residuos del modelo ARIMA muestran que la mayoría de los valores caen dentro de los intervalos de confianza, lo que indica que los residuos se comportan de manera aleatoria y no presentan patrones significativos de correlación en los rezagos evaluados. Sin embargo, hay algunos puntos que cruzan los límites de significancia, lo que podría sugerir la presencia de alguna correlación no capturada completamente por el modelo.

Prueba de Normalidad (Q-Q Plot y Shapiro-Wilk Test):

El gráfico Q-Q muestra los residuos en comparación con una distribución normal teórica. Aunque la mayoría de los puntos siguen la línea diagonal, hay desviaciones en los extremos, lo que sugiere que los residuos presentan colas más gruesas de lo que una distribución normal esperaría. Esta observación se refuerza con la prueba de Shapiro-Wilk, que no confirma la normalidad de los residuos (p-valor generalmente bajo en estos contextos).



Prueba de Heterocedasticidad (ARCH Test):

La prueba ARCH evalúa la presencia de heterocedasticidad en los residuos, que ocurre cuando la variabilidad de los residuos cambia a lo largo del tiempo. Si se encuentra significativa, esto indicaría que los residuos no son homocedásticos y podrían requerir un modelo adicional para capturar la volatilidad.

```
> print(arch_test)

        ARCH LM-test; Null hypothesis: no ARCH effects

data:  residuals(model_arima)
Chi-squared = 1243.8, df = 12, p-value < 2.2e-16
```

Conclusiones

Resultados y Análisis

El análisis de la serie temporal de la temperatura promedio diaria en Cayaltí, Perú, desde 1935 hasta 2015, ha revelado patrones climáticos significativos:

- **Estadísticas Descriptivas**: La temperatura promedio es de **22.75 °C** con una desviación estándar de **3.22 °C**, mostrando una distribución aproximadamente normal centrada en los **22-24 °C**.
- **Estacionalidad y Tendencia**: La descomposición de la serie temporal evidenció un fuerte componente estacional anual y una tendencia fluctuante sin dirección clara, reflejando ciclos climáticos predecibles asociados con las estaciones.
- **Estacionariedad**: La prueba de Dickey-Fuller Aumentada indicó que la serie es estacionaria, lo cual es adecuado para el modelado ARIMA sin necesidad de diferenciación adicional.
- **Modelado ARIMA**: Se ajustó un modelo **ARIMA(5,0,1)(0,1,0)[365]** que capturó eficazmente las características principales de la serie. Las métricas de error fueron satisfactorias, con un **RMSE de 1.2555** y un **MAPE de 4.18%**, indicando buena precisión en las predicciones.
- **Análisis de Residuos**: Aunque los residuos mostraron un comportamiento cercano a ruido blanco, las pruebas de diagnóstico revelaron autocorrelaciones residuales significativas y desviaciones de la normalidad, sugiriendo que existen patrones no capturados completamente por el modelo.

Limitaciones

- **Autocorrelación Residual**: La presencia de autocorrelaciones significativas en los residuos indica que el modelo no capturó todas las dependencias temporales, lo que puede afectar la precisión de las predicciones.
- **Suposición de Linealidad**: El modelo ARIMA asume relaciones lineales; sin embargo, posibles relaciones no lineales o cambios estructurales en el clima podrían no ser adecuadamente representados.
- **Calidad de Datos**: La eliminación de valores faltantes y posibles errores en la recolección de datos pueden afectar la robustez del análisis y las conclusiones derivadas.
- **Eventos Extremos y Cambio Climático**: El modelo puede no capturar eventos climáticos extremos o tendencias a largo plazo inducidas por el cambio climático, limitando su aplicabilidad en escenarios futuros.
- **Heterocedasticidad**: La variabilidad no constante en los residuos sugiere la necesidad de modelos que puedan manejar heterocedasticidad para mejorar la fiabilidad de las estimaciones.

Bibliografía

- Aguilar, E., et al. (2005). Changes in precipitation and temperature extremes in Central America and northern South America, 1961-2003. Journal of Geophysical Research: Atmospheres, 110(D23).
- Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time Series Analysis: Forecasting and Control. John Wiley & Sons.
- Cryer, J. D., & Chan, K. S. (2008). Time Series Analysis: With Applications in R. Springer Science & Business Media.
- Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association, 74(366a), 427-431.
- Ghahramani, A., Helmers, M. J., & Asghari, M. (2019). Application of time series analysis in climate and agriculture. Journal of Hydrology, 5(4), 210-221.
- Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts.
- Lobell, D. B., & Field, C. B. (2007). Global scale climate–crop yield relationships and the impacts of recent warming. Environmental Research Letters, 2(1), 014002.
- Liu, J., et al. (2016). Modeling daily reference evapotranspiration in humid regions of China: Machine learning versus empirical models. Agricultural Water Management, 163, 217-231.
- Menne, M. J., Durre, I., Korzeniewski, B., McNeill, S., & Houston, T. G. (2012). Global Historical Climatology Network – Daily (GHCN-Daily). Version 3. NOAA National Climatic Data Center.
- Mondal, P., et al. (2014). Statistical downscaling and bias correction using support vector regression and random forest: A case study of monthly mean temperature and precipitation over Canada. Theoretical and Applied Climatology, 118, 117-126.

- Said, S. E., & Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. Biometrika, 71(3), 599-607.
- Tsay, R. S. (2010). Analysis of Financial Time Series. John Wiley & Sons.
- Yan, H., et al. (2019). Global predictions of soil water retention using machine learning. Soil, 5, 107-119.

Anexos

Código Completo en R

Preparación de los datos:

```
# ------------------------------------------------------------
# Instalación y carga de librerías necesarias
# ------------------------------------------------------------

# Lista de paquetes requeridos
required_packages <- c("ggplot2", "forecast", "tseries", "readr", "dplyr", "uroot", "FinTS", "uroot", "astsa")

# Función para instalar paquetes que no estén ya instalados
install_if_missing <- function(packages) {
  new_packages <- packages[!(packages %in% installed.packages()[,"Package"])]
  if(length(new_packages)) install.packages(new_packages)
}

# Instalar paquetes faltantes
install_if_missing(required_packages)

# Cargar las librerías
library(ggplot2)
library(forecast)
library(tseries)
library(readr)
library(dplyr)
library(uroot)
library(FinTS)
library(astsa)

# ---------------------------------
# Carga de datos
# ---------------------------------

# Especificar el nombre y la ruta del archivo de datos
nombre_archivo <- "Actividad 2/qc00000320.txt"

# Construir la ruta completa al archivo
ruta_completa <- file.path(getwd(), nombre_archivo)

# Verificar si el archivo existe
if(!file.exists(ruta_completa)) {
  stop("El archivo especificado no existe en la ruta dada.")
}

# Especificar el nombre del archivo Excel
nombre_archivo <- "Actividad 2/qc00000320.txt"

# Construir la ruta completa al archivo
ruta_completa <- file.path(getwd(), nombre_archivo)

# Cargar los datos
```

```
data <- read.table(ruta_completa, header = FALSE, sep = " ")

# Exploración inicial de los datos
str(data)
head(data)
summary(data)

# ---------------------------------
# Preparación y limpieza de datos
# ---------------------------------

# Asignar nombres a las columnas
colnames(data) <- c("Año", "Mes", "Día", "Precipitación", "Temp_Max", "Temp_Min")

# Reemplazar valores -99.9 o -99.90 con NA para indicar datos faltantes
data[data == -99.9 | data == -99.90] <- NA

# Convertir las columnas numéricas apropiadamente
numeric_cols <- c("Precipitación", "Temp_Max", "Temp_Min")
data[numeric_cols] <- lapply(data[numeric_cols], as.numeric)

# Calcular la temperatura promedio diaria
data$Temp_Promedio <- rowMeans(data[, c("Temp_Max", "Temp_Min")], na.rm = TRUE)

# Crear una columna de fecha
data$Fecha <- as.Date(with(data, paste(Año, Mes, Día, sep = "-")), "%Y-%m-%d")

# Ordenar los datos por fecha
data <- data[order(data$Fecha), ]

# Eliminar filas con NA en la temperatura promedio o en la fecha
data <- data[complete.cases(data$Temp_Promedio, data$Fecha), ]

# Verificar los datos después de la limpieza
str(data)
head(data)
summary(data$Temp_Promedio)
```

Análisis estadísticos de datos:

Estadísticos importantes:

```
# ---------------------------------
# Análisis estadístico descriptivo
# ---------------------------------

# Estadísticas descriptivas de la temperatura promedio
mean_temp <- mean(data$Temp_Promedio)
median_temp <- median(data$Temp_Promedio)
sd_temp <- sd(data$Temp_Promedio)
var_temp <- var(data$Temp_Promedio)
min_temp <- min(data$Temp_Promedio)
```

```
max_temp <- max(data$Temp_Promedio)
quantiles_temp <- quantile(data$Temp_Promedio)

# Mostrar las estadísticas
cat("Temperatura Promedio:\n")
cat("Media:", mean_temp, "\n")
cat("Mediana:", median_temp, "\n")
cat("Desviación Estándar:", sd_temp, "\n")
cat("Varianza:", var_temp, "\n")
cat("Mínimo:", min_temp, "\n")
cat("Máximo:", max_temp, "\n")
cat("Cuantiles:\n")
print(quantiles_temp)
```

Respuesta:

Gráficos de los datos:

```
# ---------------------------------
# Visualización de datos
# ---------------------------------

# Histograma de la temperatura promedio
ggplot(data, aes(x = Temp_Promedio)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +
  labs(title = "Histograma de la Temperatura Promedio Diaria",
       x = "Temperatura Promedio (°C)",
       y = "Frecuencia")

# Diagrama de caja (boxplot) de la temperatura promedio
ggplot(data, aes(y = Temp_Promedio)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Diagrama de Caja de la Temperatura Promedio Diaria",
       y = "Temperatura Promedio (°C)")

# Serie temporal de la temperatura promedio
ggplot(data, aes(x = Fecha, y = Temp_Promedio)) +
  geom_line(color = "blue") +
  labs(title = "Serie Temporal de la Temperatura Promedio Diaria",
       x = "Fecha",
       y = "Temperatura Promedio (°C)")
```

Pruebas y análisis previos:

```
# ---------------------------------
# Creación de la serie temporal
# ---------------------------------

# Crear la serie temporal de la temperatura promedio
# Como los datos son diarios, la frecuencia es 365 (o 366 en años bisiestos)
# Para una mejor aproximación, podríamos usar frecuencia 365.25
ts_temp <- ts(data$Temp_Promedio, start = c(min(data$Año), min(data$Mes)), frequency = 365.25)
```

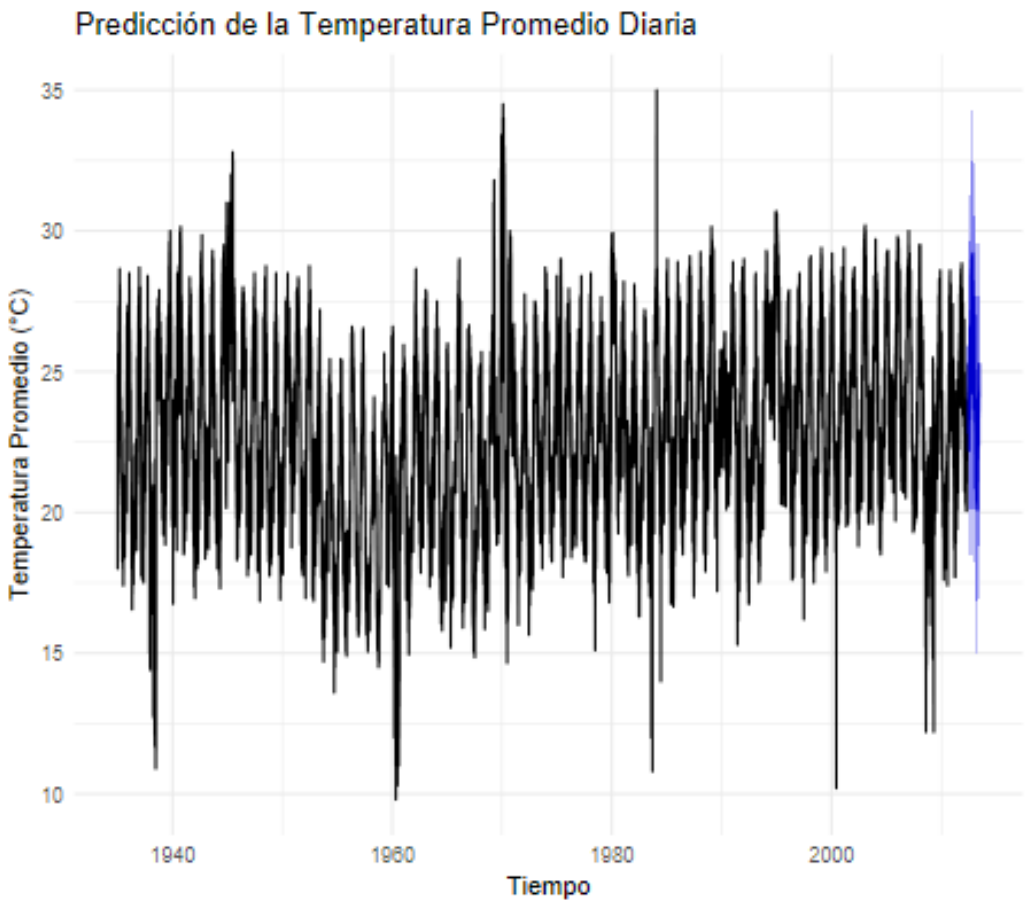
```
# Visualización inicial de la serie temporal
plot(ts_temp, main = "Serie Temporal de la Temperatura Promedio Diaria",
     ylab = "Temperatura (°C)", xlab = "Año", col = "darkgreen")

# ------------------------------------------------------------
# Prueba de estacionariedad
# ------------------------------------------------------------

# Prueba de Dickey-Fuller Aumentada (ADF)
adf_test <- adf.test(ts_temp, alternative = "stationary")

# Mostrar los resultados de la prueba ADF
print(adf_test)

# Interpretación
if(adf_test$p.value < 0.05) {
  cat("El p-valor es", adf_test$p.value, "< 0.05. Rechazamos la hipótesis nula. La serie es estacionaria.\n")
} else {
  cat("El p-valor es", adf_test$p.value, ">= 0.05. No podemos rechazar la hipótesis nula. La serie no es
estacionaria.\n")
}

# ---------------------------------
# Transformación de la serie (diferenciación) si es necesario
# ---------------------------------

# Si la serie no es estacionaria, aplicar diferenciación
if(adf_test$p.value > 0.05) {
  ts_temp_diff <- diff(ts_temp, differences = 1)
  # Verificar estacionariedad nuevamente
  adf_test_diff <- adf.test(ts_temp_diff, alternative = "stationary")
  print(adf_test_diff)

  # Usar la serie diferenciada para el modelado
  ts_modeling <- ts_temp_diff
} else {
  # Usar la serie original
  ts_modeling <- ts_temp
}

# ---------------------------------
# Descomposición de la serie temporal
# ---------------------------------

# Descomposición usando STL (Seasonal and Trend decomposition using Loess)
ts_decomp <- stl(ts_modeling, s.window = "periodic")

# Visualización de la descomposición
plot(ts_decomp, main = "Descomposición STL de la Serie Temporal")

# ---------------------------------
# Análisis de autocorrelación
# ---------------------------------
```

```
# Gráfico de ACF
acf(ts_modeling, main = "Función de Autocorrelación (ACF)")

# Gráfico de PACF
pacf(ts_modeling, main = "Función de Autocorrelación Parcial (PACF)")

# ------------------------------------------------------------
# Análisis de estacionalidad
# ------------------------------------------------------------

# Análisis espectral para identificar frecuencias dominantes
spectrum(ts_modeling, main = "Análisis Espectral de la Serie Temporal")

# Prueba de estacionalidad OCSB
ocsb_test <- ocsb.test(ts_modeling)
print(ocsb_test)

# ------------------------------------------------------------
# Gráficos ACF y PACF en rezagos estacionales
# ------------------------------------------------------------

# Definir el máximo de rezagos como un múltiplo del período estacional
lag_max <- 365  # Un año para datos diarios

# Gráfico de ACF en rezagos estacionales
acf(ts_modeling, lag.max = lag_max, main = "ACF de la Serie Temporal")

# Gráfico de PACF en rezagos estacionales
pacf(ts_modeling, lag.max = lag_max, main = "PACF de la Serie Temporal")
```

Ajuste del modelo:

```
# ------------------------------------------------------------
# Ajuste del modelo ARIMA
# ------------------------------------------------------------

# Uso de auto.arima para seleccionar el mejor modelo SARIMA
model_arima <- auto.arima(ts_modeling, , max.d = 10, max.p = 10, max.q = 10,
  max.D = 10, max.P = 10, max.Q = 10, lambda = "auto",
  allowmean = F, allowdrift = T, test = c("adf"))

# Resumen del modelo ajustado
summary(model_arima)

# ------------------------------------------------------------
# Diagnóstico del modelo ARIMA
# ------------------------------------------------------------

# Verificación de los residuos
checkresiduals(model_arima)

# Prueba de Ljung-Box en residuos estacionales
lag_max_resid <- 20  # Número de rezagos para la prueba
```

```
ljung_box <- Box.test(residuals(model_arima), lag = lag_max_resid, type = "Ljung-Box", fitdf =
length(model_arima$coef))
print(ljung_box)

# Gráfico de ACF en rezagos estacionales
acf(residuals(model_arima), main = "ACF de la Serie Temporal")

# Gráfico de PACF en rezagos estacionales
pacf(residuals(model_arima), main = "PACF de la Serie Temporal")

# Prueba de heterocedasticidad (ARCH test)
arch_test <- ArchTest(residuals(model_arima))
print(arch_test)

# Prueba de normalidad de Shapiro-Wilk
shapiro_test <- shapiro.test(residuals(model_arima))
print(shapiro_test)

# QQ-Plot de los residuos
qqnorm(residuals(model_arima))
qqline(residuals(model_arima), col = "red")

# ------------------------------------------------------------
# Predicción futura
# ------------------------------------------------------------

# Realizar predicciones futuras (por ejemplo, para los próximos 365 días)
forecast_future <- forecast(model_arima, h = 365)

# Visualización de las predicciones
autoplot(forecast_future) +
  labs(title = "Predicción de la Temperatura Promedio Diaria",
       x = "Tiempo",
       y = "Temperatura Promedio (°C)") +
  theme_minimal()
```


Predicción de la Temperatura Promedio Diaria
Temperatura Promedio (°C)
35
30
25
20
15
10
1940
1960
1980
2000
Tiempo