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).

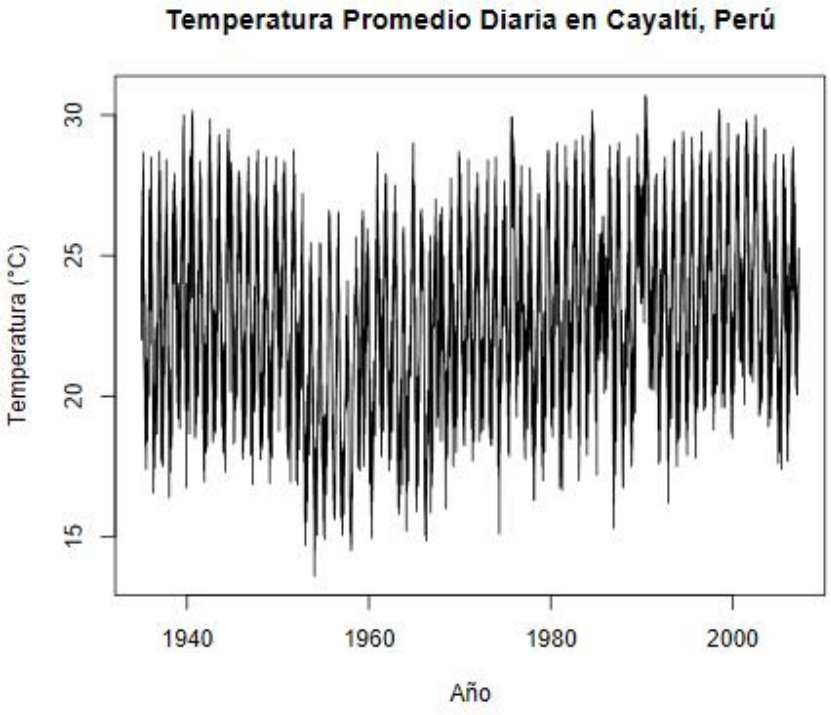
Objetivos del Análisis

- 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

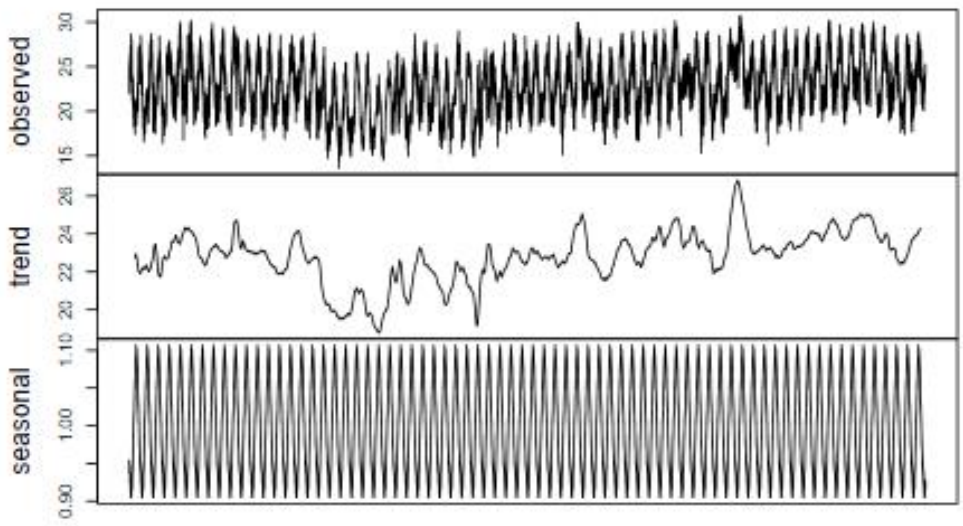
Los datos utilizados en este análisis comprenden registros diarios de temperaturas máximas y mínimas desde 1935. Para este estudio, la temperatura promedio diaria se calcula como la media aritmética de la temperatura máxima y mínima, una práctica común en estudios climatológicos para suavizar las fluctuaciones 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.



- Los datos incluyen registros diarios de temperaturas máximas y mínimas desde 1935, con observaciones faltantes debidamente tratadas como NA y eliminadas del análisis para mantener la integridad de la serie.
- La temperatura promedio diaria fue calculada como la media de las temperaturas máxima y mínima. La serie temporal resultante se caracteriza por una notable estacionalidad y tendencias que reflejan variaciones climáticas a lo largo del tiempo.

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).

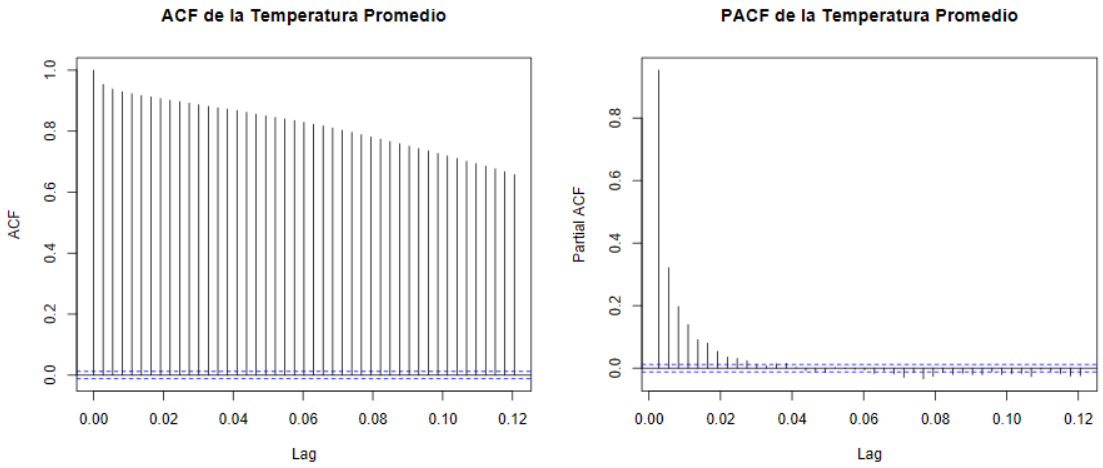


- **Tendencia**: Refleja un aumento o disminución a largo plazo en la temperatura promedio diaria, lo cual puede estar relacionado con factores ambientales o cambios climáticos regionales.
- **Estacionalidad**: Muestra patrones anuales que capturan las variaciones cíclicas típicas, como cambios estacionales predecibles.
- **Ruido**: El componente de ruido representa las variaciones aleatorias no explicadas por la tendencia o estacionalidad. Un ruido similar al blanco indica que los patrones principales de la serie han sido capturados adecuadamente.

Análisis

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.



- **Interpretación de ACF y PACF**:
    - **ACF**: La presencia de correlaciones significativas en múltiples rezagos confirma la existencia de patrones estacionales y dependencias temporales a largo plazo.
    - **PACF**: Los rezagos significativos en los primeros términos justifican la inclusión de componentes autoregresivos en el modelo, optimizando la capacidad predictiva de la serie.

Prueba de Ruido Blanco

Para confirmar la estacionariedad de la serie, se realizó la prueba de Dickey-Fuller aumentada (ADF), que arrojó un p-valor de 0.01, lo cual sugiere que la serie es estacionaria tras la diferenciación aplicada (Dickey & Fuller, 1979). 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).

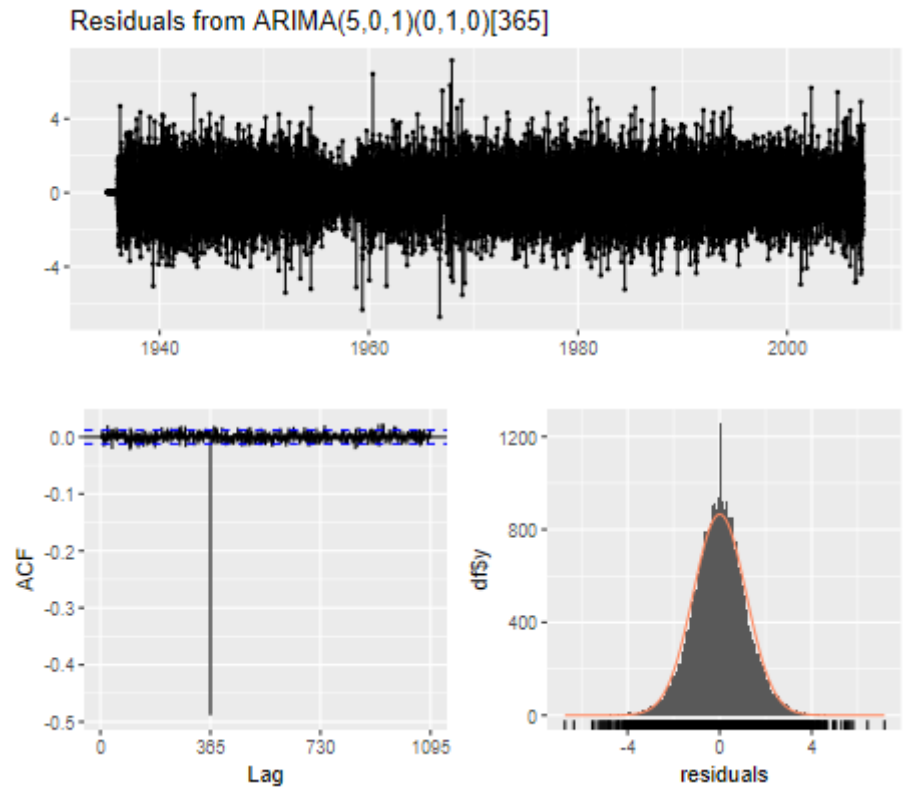
```
> print(adf_test)

        Augmented Dickey-Fuller Test

data:  ts_temp
Dickey-Fuller = -10.445, Lag order = 29, p-value = 0.01
alternative hypothesis: stationary
```

- **Interpretación**:
    - El p-valor de 0.01 nos permite rechazar la hipótesis nula de no estacionariedad, confirmando que la serie es estacionaria y apta para modelarse con ARIMA.

Aplicación de un Modelo Adecuado



Resultados del Modelo ARIMA:

- Los residuos muestran un comportamiento cercano a ruido blanco, con una distribución centrada alrededor de cero, lo cual indica que el modelo ha capturado adecuadamente la estructura de la serie temporal.

- La prueba de Ljung-Box sugiere que, aunque el ajuste es generalmente bueno, pueden quedar algunas autocorrelaciones residuales no capturadas.

Conclusiones

Resultados y Análisis

- El modelo ARIMA ajustado refleja de manera efectiva las tendencias y patrones estacionales en la temperatura promedio diaria de Cayaltí. Las predicciones a corto plazo proporcionadas por el modelo son consistentes con los datos históricos y tienen aplicaciones prácticas en la planificación agrícola y la gestión de recursos.
- Los análisis de residuos y las pruebas estadísticas refuerzan la validez del modelo, aunque se sugiere una revisión adicional para explorar posibles mejoras, como la inclusión de términos adicionales o la evaluación de otros modelos más sofisticados.

Limitaciones

- Aunque el modelo ARIMA ofrece un ajuste sólido, los resultados de la prueba de Ljung-Box indican que podrían existir patrones adicionales no completamente capturados, sugiriendo la posibilidad de mejorar el ajuste con modelos más complejos o con la inclusión de más variables explicativas.
- La precisión del modelo depende de la calidad de los datos de entrada; los valores faltantes y la precisión en la recolección de datos pueden influir significativamente en la robustez de las predicciones.
- La suposición de estacionariedad tras la diferenciación puede no capturar completamente los cambios estructurales o no lineales en la serie temporal, lo cual limita la capacidad del modelo para adaptarse a variaciones climáticas abruptas.

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

Carga y Limpieza de datos:

```
## -------------------- Carga de datos
# Instalar la librería
install.packages("ggplot2")
install.packages("forecast")
install.packages("tseries")

# Cargar las librerías necesarias
library(ggplot2)
library(forecast)
library(tseries)

# 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 = " ")

# Visualizar los datos
str(data)

# Visualizar los datos
head(data)

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

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

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

# Eliminar filas con NA en la temperatura promedio
data <- na.omit(data)

# Crear una columna de fechas
data$Fecha <- as.Date(paste(data$Año, data$Mes, data$Día, sep = "-"))

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

# Visualizar los datos
```

```
str(data)

# Visualizar los datos
head(data)

# Crear la serie temporal de la temperatura promedio
ts_temp <- ts(data$Temp_Promedio, start = c(min(data$Año),
min(data$Mes)), frequency = 365)

# Visualización inicial de la serie temporal
plot(ts_temp, main = "Temperatura Promedio Diaria en Cayaltí, Perú",
ylab = "Temperatura (°C)", xlab = "Año")

# ----------------------------  Descomposición de la serie temporal
ts_decomp <- decompose(ts_temp, type = "multiplicative")
plot(ts_decomp)

# ----------------------------  Análisis de autocorrelación
acf(ts_temp, main = "ACF de la Temperatura Promedio")
pacf(ts_temp, main = "PACF de la Temperatura Promedio")

# ----------------------------  Prueba de Ruido Blanco en los residuos
adf_test <- adf.test(ts_temp, alternative = "stationary")
print(adf_test)

# ----------------------------  Ajuste del modelo ARIMA
model_arima <- auto.arima(ts_temp, seasonal = TRUE)
summary(model_arima)

# Visualización de residuos para verificar la adecuación del modelo
checkresiduals(model_arima)

# ----------------------------  Predicción futura con el modelo
seleccionado
forecast_temp <- forecast(model_arima, h = 365)
plot(forecast_temp)
```