

GUIA PRACTICA

Selección y Comparación de Modelos en Series de Tiempo

Diagnostico de residuos · AIC / BIC · MAE / RMSE / MAPE · R-cuadrado · Validación cruzada temporal · Ljung-Box · U de Theil · Flujo de decisión

- 1 Diagnostico de residuos — cuatro graficos clave
- 2 Criterios de informacion: AIC y BIC
- 3 Metricas de error: MAE, RMSE, MAPE, MASE
- 4 R-cuadrado y U de Theil
- 5 Validación cruzada temporal
- 6 Pruebas de bondad de ajuste
- 7 Flujo integrado de selección
- 8 Código R de referencia

Introducción

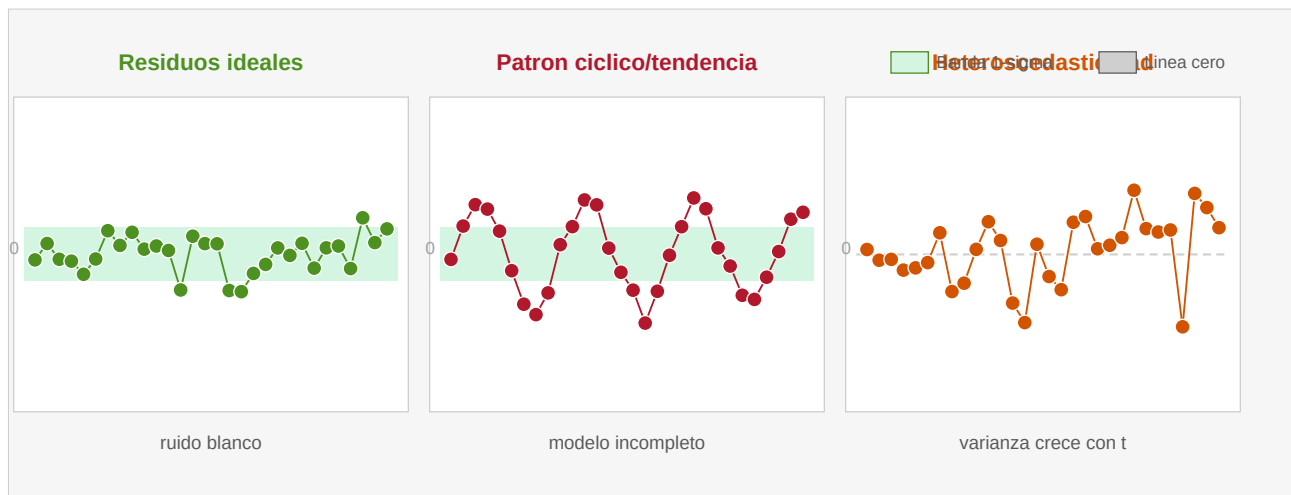
Ajustar un modelo de series de tiempo es solo la mitad del trabajo. La otra mitad consiste en decidir si ese modelo merece confianza y si es mejor o peor que las alternativas disponibles. Esta guía consolida los criterios que un analista experimentado usa en ese proceso: desde el primer vistazo a los residuos hasta la comparación formal por validación cruzada temporal.

Un modelo puede tener AIC bajo y aun así ser inutilizable si sus residuos tienen autocorrelación. El orden correcto: diagnóstico de residuos primero, métricas de predicción después.

1. Diagnóstico de residuos — cuatro gráficos

El residuo es lo que el modelo no explica. Si el modelo es correcto, los residuos deben ser ruido blanco: centrados en cero, sin patrón, varianza constante, no autocorrelacionados. Cuatro gráficos confirman o descartan esto:

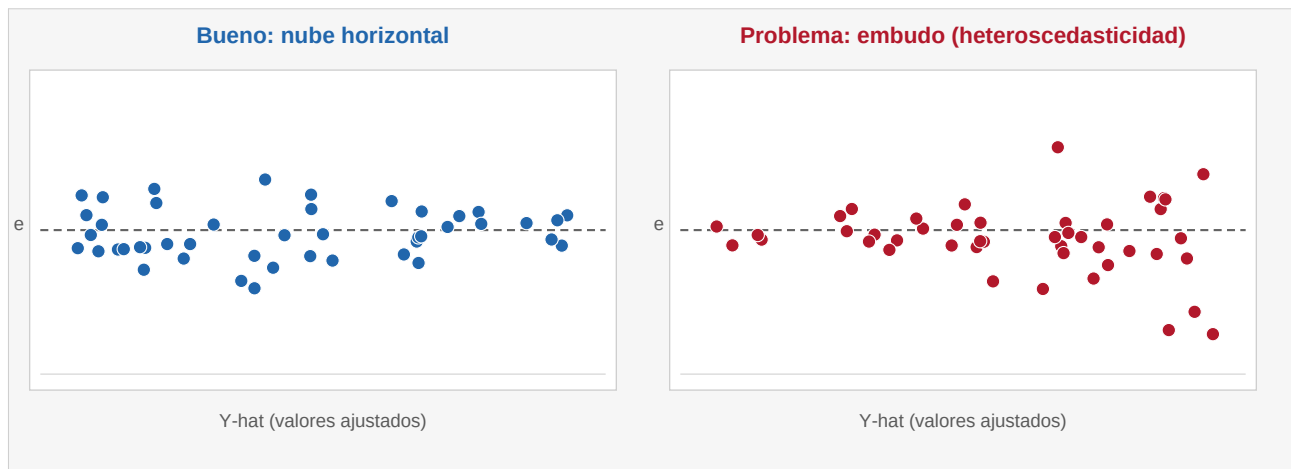
Gráfico 1 — Residuos contra el tiempo



Izq: ruido blanco ideal (banda 1-sigma en verde). Centro: ciclos residuales — falta capturar estacionalidad. Der: embudo — heteroscedasticidad, transformar Y o cambiar modelo.

Gráfico 2 — Residuos contra Y-hat (valores ajustados)

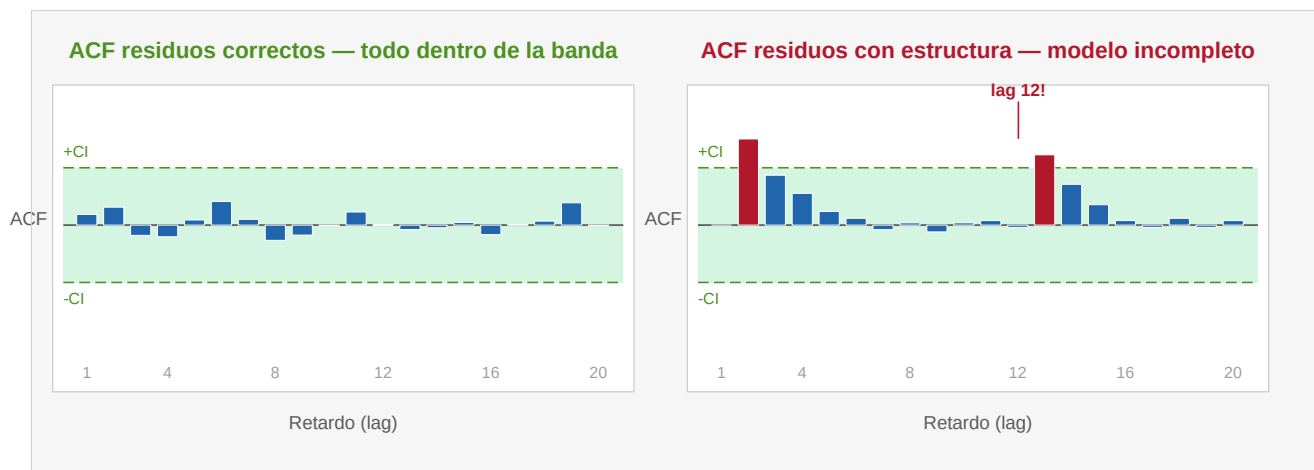
Si la nube de puntos tiene forma de embudo — abre hacia la derecha — la varianza del error depende del nivel de Y. Eso señala que el componente aditivo es inadecuado.



Izquierda: nube horizontal — varianza constante, modelo válido. Derecha: embudo — varianza crece con el nivel, modelo aditivo inadecuado.

Grafico 3 — ACF de los residuos

La función de autocorrelación de los residuos debe tener todas las barras dentro de las bandas de confianza al 95%. Barras significativas en retardos bajos (1, 2, 3) señalan estructura AR/MA no modelada. Barras en lag 12 o 24 indican estacionalidad residual.



Izquierda: ACF ideal — todas las barras dentro de la banda verde (95%). Derecha: ACF con estructura — lags 1-3 y lag 12 significativos (rojo), el modelo necesita términos SAR o SMA adicionales.

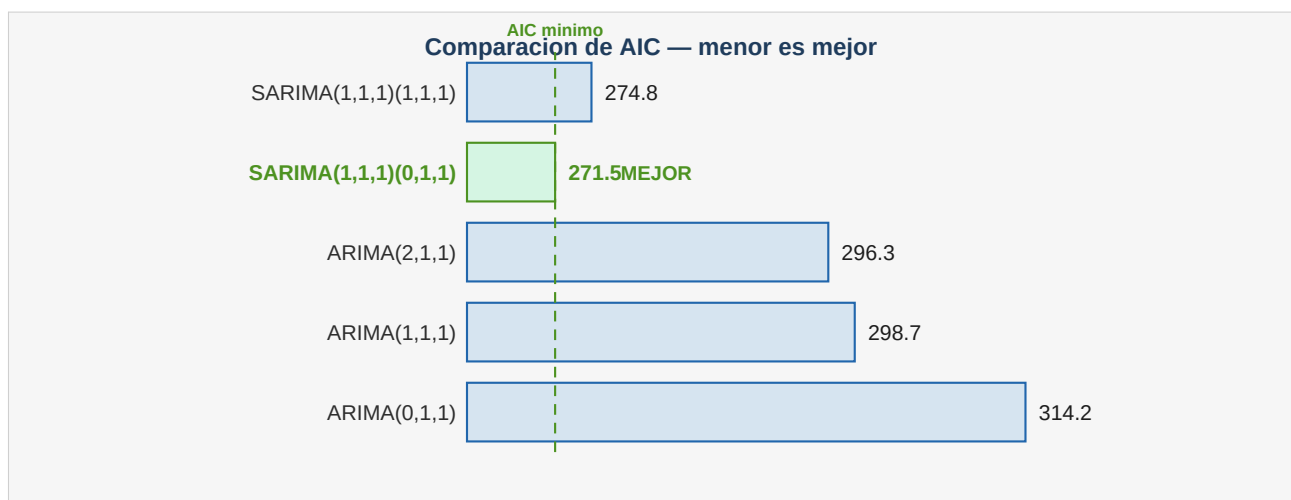
Regla práctica. Si el ACF de los residuos muestra barras significativas, incrementa el orden AR, MA, SAR o SMA según donde estén los lags. No avances a comparar AIC hasta que el ACF este limpio.

2. Criterios de información: AIC y BIC

El AIC y el BIC penalizan la complejidad para evitar el sobreajuste. El mejor modelo entre varios candidatos válidos es el de menor valor.

Criterio	Formula	Penalización extra	Uso principal
AIC	$-2 \log(L) + 2k$	$k = \text{num. parámetros}$	Prefiere predicción
BIC	$-2 \log(L) + k \log(n)$	$n = \text{num. observaciones}$	Prefiere parsimonia

Comparación visual de modelos candidatos



Ejemplo con 5 especificaciones ARIMA/SARIMA. La diferencia entre ARIMA(2,1,1) y SARIMA(1,1,1)(0,1,1) es de 24 puntos: evidencia fuerte a favor del modelo estacional.

Diferencia de AIC	Interpretación
Diferencia < 2	Modelos equivalentes. Elige el más simple.
Diferencia 2-6	Evidencia moderada a favor del menor AIC.
Diferencia > 10	Evidencia fuerte. El ganador es claro.

Atención. AIC y BIC solo son comparables entre modelos con la misma variable de respuesta en la misma escala. Un modelo en $\log(Y)$ y uno en Y tienen AIC en escalas distintas y no se pueden comparar directamente.

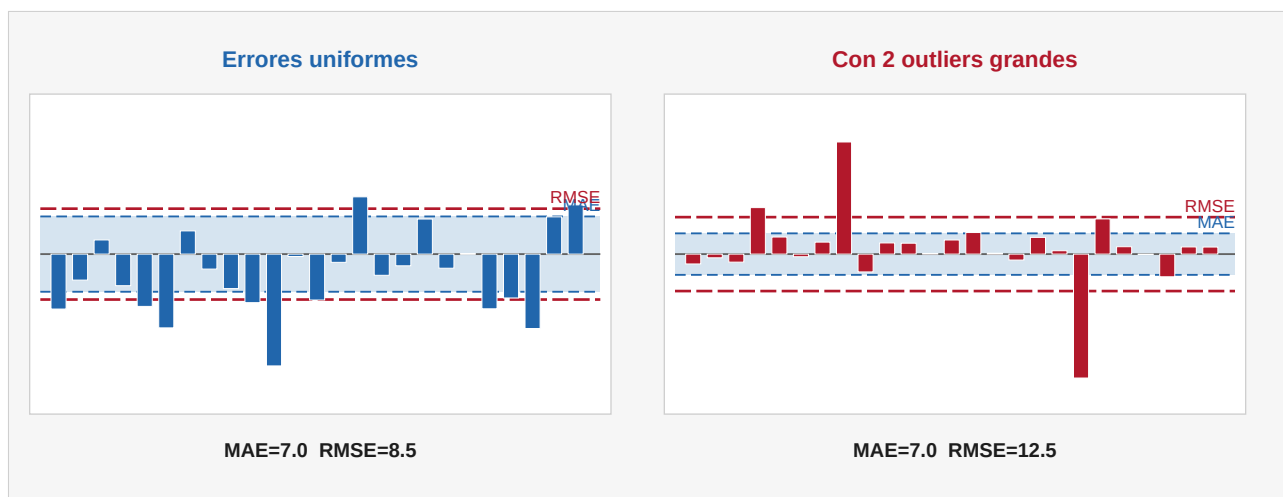
3. Métricas de error en predicción

Se calculan siempre sobre el conjunto de prueba (fuera de muestra). Cada métrica enfatiza algo distinto. Un análisis robusto reporta al menos dos.

Métrica	Cálculo	Fortaleza	Limitación
MAE	Media $ \text{error} $	Interpretable, mismas unidades que Y	No penaliza errores grandes de forma diferencial
RMSE	Raíz de la media del error cuadrático	Penaliza errores grandes; sensible a outliers	Escala depende de Y; no comparable entre series
MAPE	Media $ \text{error}/\text{real} * 100$	Adimensional: compara series de distinta escala	Explota si Y tiene valores cercanos a cero
SMAPE	$2 * e / (Y + Y\text{-hat}) * 100$	Maneja mejor ceros y valores pequeños	Menos intuitivo; simetría artificial
MASE	MAE / MAE del modelo naïve	$MASE < 1$ confirma que supera al naïve	Requiere calcular el naïve como referencia

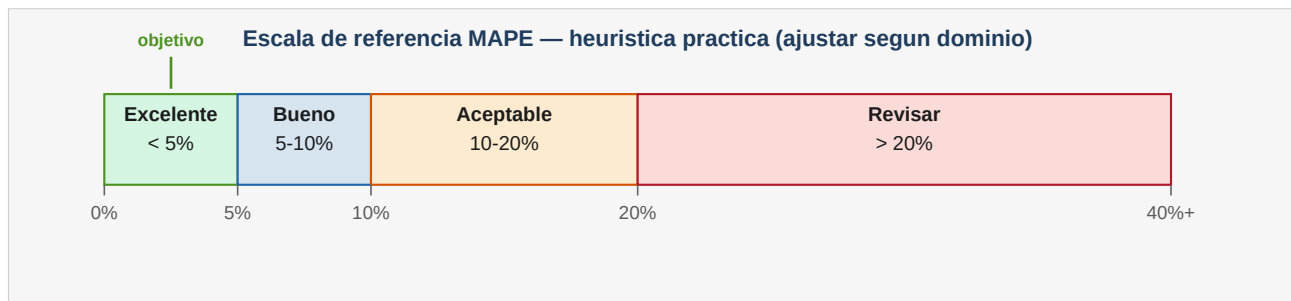
Por que RMSE penaliza mas los errores grandes que MAE

La diferencia entre MAE y RMSE se vuelve crítica cuando hay outliers en los errores de predicción. El siguiente diagrama lo muestra: dos series con similar MAE pero RMSE muy diferente por efecto de dos errores atípicos.



Izquierda: errores uniformes — MAE y RMSE están cercanos. Derecha: con dos outliers grandes (rojo) — el MAE sube poco, pero el RMSE se dispara. La banda azul es MAE; la línea roja es RMSE.

Umbral de referencia para MAPE



Escala orientativa. En series financieras o de alta frecuencia los umbrales son más estrictos; en series de demanda de largo plazo pueden ser más laxos.

Regla práctica. Usa RMSE si los errores grandes tienen consecuencias asimétricas (sobrestock, falla de suministro). Usa MAPE si necesitas comparar el desempeño entre series con escalas distintas. Reporta siempre MASE para confirmar que superas al naïve.

4. R-cuadrado y U de Theil

El R-cuadrado mide la proporción de varianza explicada. En series con tendencia, el R^2 en niveles es casi siempre alto (>0.9) incluso para modelos que solo siguen la tendencia sin capturar nada útil.

Variante	Interpretación práctica
R^2 en niveles con tendencia	Artificialmente alto. Informa poco. Evitar como criterio principal.
R^2 en primeras diferencias	Más informativo: mide si el modelo capta los cambios periodo a periodo.
R^2 negativo	El modelo predice peor que la media de la serie. Problema severo.
R^2 ajustado	Penaliza por número de parámetros. Mejor que R^2 simple en comparación.

La U_2 de Theil: una prueba de valor agregado

Compara el RMSE del modelo con el RMSE del pronóstico naïve (random walk o último valor). $U_2 < 1$ confirma que el modelo supera al naïve. Es la prueba más directa de que el modelo agrega valor.

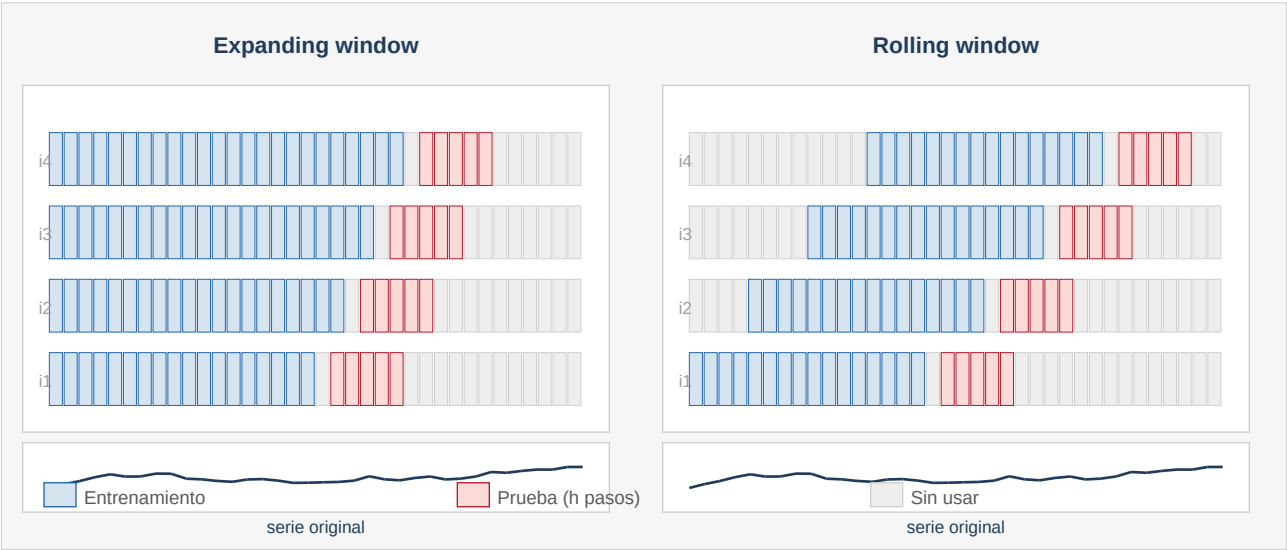
U_2	Significado	Acción
$U_2 < 1$	El modelo supera al pronóstico ingenuo.	Buen resultado.
$U_2 = 1$	Empate con el naïve.	El modelo no agrega valor.
$U_2 > 1$	El modelo es peor que el naïve.	Descartarlo.

Regla práctica. Reporta siempre U_2 junto con MAPE o RMSE. Si $U_2 \geq 1$ en el horizonte de interés, la discusión sobre AIC o bondad de ajuste es irrelevante.

5. Validación cruzada temporal

En series de tiempo NO se hace validación cruzada aleatoria. El orden cronológico es información: el modelo aprende del pasado para predecir el futuro. Mezclar el tiempo destruye esa estructura.

Atención. Nunca uses k-fold aleatorio en series de tiempo. Siempre respeta el orden cronológico.



Esquema	Mecanismo	Cuando usarla
Expanding window	Cada iteración incluye todos los datos previos.	Simula el proceso real de reestimación. Preferida cuando los datos son escasos.
Rolling window	Ventana de tamaño fijo; avanza en el tiempo.	Util cuando el comportamiento reciente es más relevante que el historico lejano.

Regla práctica. Valida siempre en el horizonte h que usaras en la práctica. Reserva al menos 20% de la serie como conjunto de prueba y no menos de dos ciclos estacionales completos.

6. Pruebas de bondad de ajuste

Formalizan los diagnósticos visuales con estadísticos y p-valores. Se aplican sobre los residuos del modelo ya ajustado. Ninguna prueba es definitiva sola: combinalas con los gráficos.

Prueba	Hipotesis nula	Resultado deseado	Nota
Ljung-Box	Residuos no autocorrelacionados	$p > 0.05$	La mas critica. Rechazo invalida el modelo.
Shapiro-Wilk / Jarque-Bera	Residuos normalmente distribuidos	$p > 0.05$	Relevante para intervalos de confianza validos.
Breusch-Pagan	Varianza constante (homoscedasticidad)	$p > 0.05$	Complementa el grafico residuos vs $\hat{Y}$.
ADF sobre residuos	Residuos estacionarios	$p < 0.05$	Confirma que no queda tendencia en los residuos.

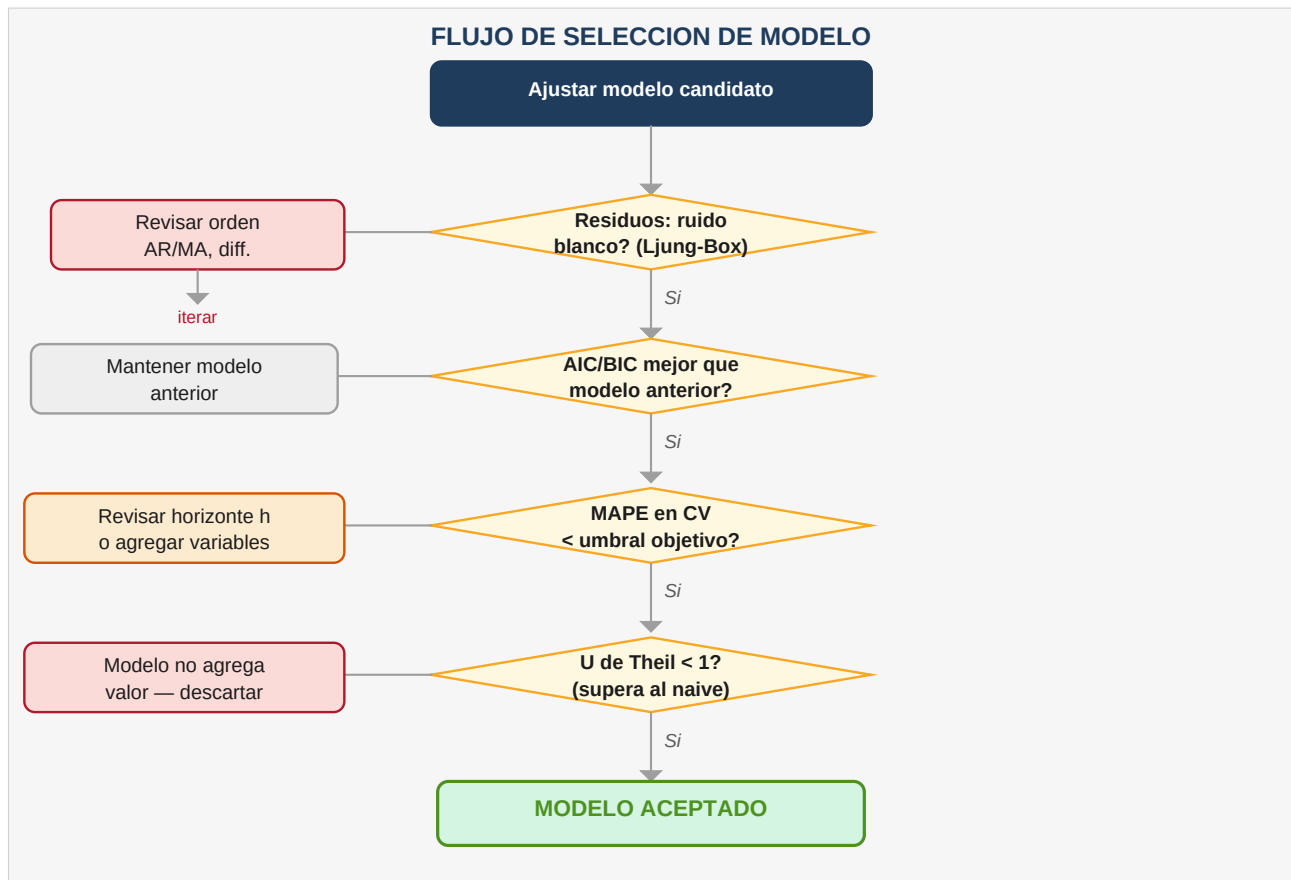
Ljung-Box: cuantos retardos usar

Para series no estacionales: $h = \min(10, T/5)$. Para series estacionales con periodo m : $h = \min(2m, T/5)$. En modelos ARIMA(p,d,q), usa $h > p + q$ para evitar resultado trivial.

Si Ljung-Box rechaza H_0 ($p < 0.05$), incrementa el orden AR o MA, o revisa si la estacionalidad esta completamente capturada. No continues con comparacion de modelos hasta resolver esto.

7. Flujo integrado de selección de modelo

El flujo integra todos los criterios en un proceso de decisión secuencial: primero elimina modelos inválidos (residuos con estructura), luego compara por AIC/BIC entre los válidos, valida el desempeño predictivo en datos no vistos y confirma que supera al naïve.



Cada criterio tiene un rol específico. Ljung-Box garantiza validez estadística. AIC/BIC selecciona entre modelos válidos. MAPE en CV mide utilidad práctica en el horizonte real. U de Theil confirma que el modelo agrega valor sobre el pronóstico ingenuo.

Regla práctica. Si dos modelos pasan todos los filtros y tienen MAPE similar (diferencia < 1 pp), elige el más parsimonioso. La complejidad no justificada por los datos se cobra en sobreajuste.

8. Código R de referencia

Bloques listos para copiar y adaptar. Cubren desde el diagnóstico de residuos hasta la validación cruzada y las pruebas formales de bondad de ajuste.

Código R – diagnóstico de residuos

```
# DIAGNOSTICO DE RESIDUOS
library(forecast); library(tseries); library(lmtest)

fit <- auto.arima(mi_ts)
res <- residuals(fit)

# 1. Residuos vs tiempo
plot(res, main="Residuos vs tiempo"); abline(h=0, lty=2)

# 2. Residuos vs Y-hat
plot(fitted(fit), res, pch=19, cex=0.6,
     xlab="Y-hat", ylab="Residuos")
abline(h=0, lty=2)

# 3. ACF y PACF de residuos
ggAcf(res, lag.max=48)
ggPacf(res, lag.max=48)

# 4. Ljung-Box
h_lb <- min(10, length(res)/5)
Box.test(res, lag=h_lb, type="Ljung-Box",
         fitdf=length(fit$coef))
```

Código R – AIC, BIC, métricas y U de Theil

```
# AIC, BIC Y METRICAS DE PREDICCION

# Criterios de informacion
cat("AIC:", AIC(fit), " BIC:", BIC(fit), "\n")

# Comparar modelos
fit1 <- Arima(mi_ts, order=c(1,1,1), seasonal=c(0,1,1))
fit2 <- Arima(mi_ts, order=c(1,1,1), seasonal=c(1,1,1))
cat("AIC1:", AIC(fit1), " AIC2:", AIC(fit2), "\n")

# Separar train/test
n <- length(mi_ts)
n_test <- round(n * 0.2)
train <- head(mi_ts, n - n_test)
test <- tail(mi_ts, n_test)

# Pronosticar y calcular metricas
fit_tr <- auto.arima(train)
pred <- forecast(fit_tr, h=n_test)$mean
acc <- accuracy(pred, as.vector(test))
cat("MAPE:", round(acc["Test set", "MAPE"], 2), "%\n")
```

```

cat("RMSE:", round(acc["Test set", "RMSE"], 4), "\n")
cat("MAE: ", round(acc["Test set", "MAE"], 4), "\n")
cat("MASE:", round(acc["Test set", "MASE"], 4), "\n")
# U de Theil
naive_pred <- naive(train, h=n_test)$mean
u2 <- accuracy(pred, as.vector(test))["Test set", "RMSE"] /
accuracy(naive_pred, as.vector(test))["Test set", "RMSE"]
cat("U2 de Theil:", round(u2, 3), "\n")

```

Código R – CV temporal y pruebas de bondad de ajuste

```

# VALIDACION CRUZADA TEMPORAL – expanding window
ts_cv <- function(serie, h=6, min_train=36) {
  n <- length(serie); errs <- c()
  for (i in min_train:(n-h)) {
    tr <- window(serie, end=time(serie)[i])
    te <- window(serie, start=time(serie)[i+1],
end=time(serie)[i+h])
    pr <- forecast(auto.arima(tr, stepwise=TRUE), h=h)$mean
    errs <- c(errs,
mean(abs((as.vector(te)-as.vector(pr))/as.vector(te)))*100)
  }
  cat("MAPE CV (h=", h, "):", round(mean(errs), 2), "%\n")
  invisible(errs)
}
cv_err <- ts_cv(mi_ts, h=6)
boxplot(cv_err, main="Distribucion del MAPE por iteracion")

# PRUEBAS DE BONDAD DE AJUSTE
shapiro.test(res)
bptest(res^2 ~ seq_along(res))
adf.test(na.omit(res))

```

La confianza en un modelo se construye cuando pasa los filtros de validez estadística (residuos limpios, Ljung-Box aprobado), supera al naive ($U2 < 1$) y generaliza bien en datos no vistos (MAPE en CV dentro del umbral). Ningún criterio aislado es suficiente.