

ANÁLISIS AVANZADO Y PREDICCIÓN DE SERIES TEMPORALES APLICADOS AL ESTUDIO DE RETRASOS EN AEROPUERTOS

MÁSTER EN INGENIERÍA DE SISTEMAS DE DECISIÓN

AUTORA: AMAIA SAINZ VILLA
TUTORA: ANA ELIZABETH GARCÍA SIPOLS



Universidad
Rey Juan Carlos

ÍNDICE DE CONTENIDOS

1) INTRODUCCIÓN

2) RECOPIACIÓN, LIMPIEZA Y ESTRUCTURACIÓN DE LOS DATOS

3) TÉCNICAS DE ANÁLISIS DE CONGLOMERADOS

4) TÉCNICAS DE PREDICCIÓN

5) CONCLUSIONES

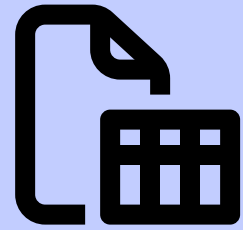


MOTIVACIÓN



- **Saturación aérea y complejidad operativa** creciente.
- **Retrasos** con alto impacto **económico, ambiental** y en el **pasajero**.
- Necesidad de **herramientas predictivas** para anticipar incidencias.
- **Grandes volúmenes de datos** disponibles para el análisis.
- Uso de **modelos estadísticos e inteligencia artificial**.
- Análisis de **50 aeropuertos europeos** con datos reales.
- Aplicación de agrupación y modelos predictivos.
- Evaluación mediante **métricas de error y visualización de resultados**.
- Enfoque práctico para mejorar la **gestión aeroportuaria**.

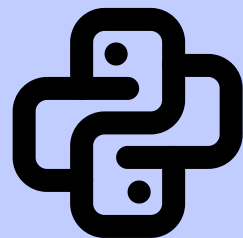
MATERIAL DE TRABAJO



Almacenamiento de datos



Tratamiento, predicción y visualización



Implementación de redes neuronales

ORIGEN DE DATOS

- 50 aeropuertos europeos.
- Registros horarios (07:00–23:00 UTC).
- Periodo analizado: septiembre 2017.
- Dos conjuntos: retrasos medios y tráfico aéreo.
- 17 observaciones diarias por aeropuerto.
- Retrasos: diferencia entre hora real y programada.
- Datos preprocesados y depurados.

Filter Cols: << 1 - 50 >>														
	Date	EGLL	LFPG	EHAM	EDDF	LEMD	LEBL	EDDM	EGKK	LIRF	LFPO	EIDW	LSZH	EKCH
1	2017-09-01 07:00:00	808	194	1053	83	-154	299	170	353	-129	154	-7	240	5
2	2017-09-01 08:00:00	811	426	767	239	138	484	300	519	183	87	90	221	-71
3	2017-09-01 09:00:00	775	323	784	170	483	681	389	515	107	-5	234	417	110
4	2017-09-01 10:00:00	909	356	521	287	351	583	206	510	283	-43	563	795	-65
5	2017-09-01 11:00:00	1109	631	995	183	197	919	-74	796	409	616	147	479	282
6	2017-09-01 12:00:00	1038	653	523	580	448	367	400	1277	2021	447	248	563	-1
7	2017-09-01 13:00:00	1268	644	736	471	596	375	473	1440	2173	360	880	124	-28
8	2017-09-01 14:00:00	1036	586	970	295	571	87	414	1918	934	170	378	379	531

OBTENCIÓN DE SERIES TEMPORALES

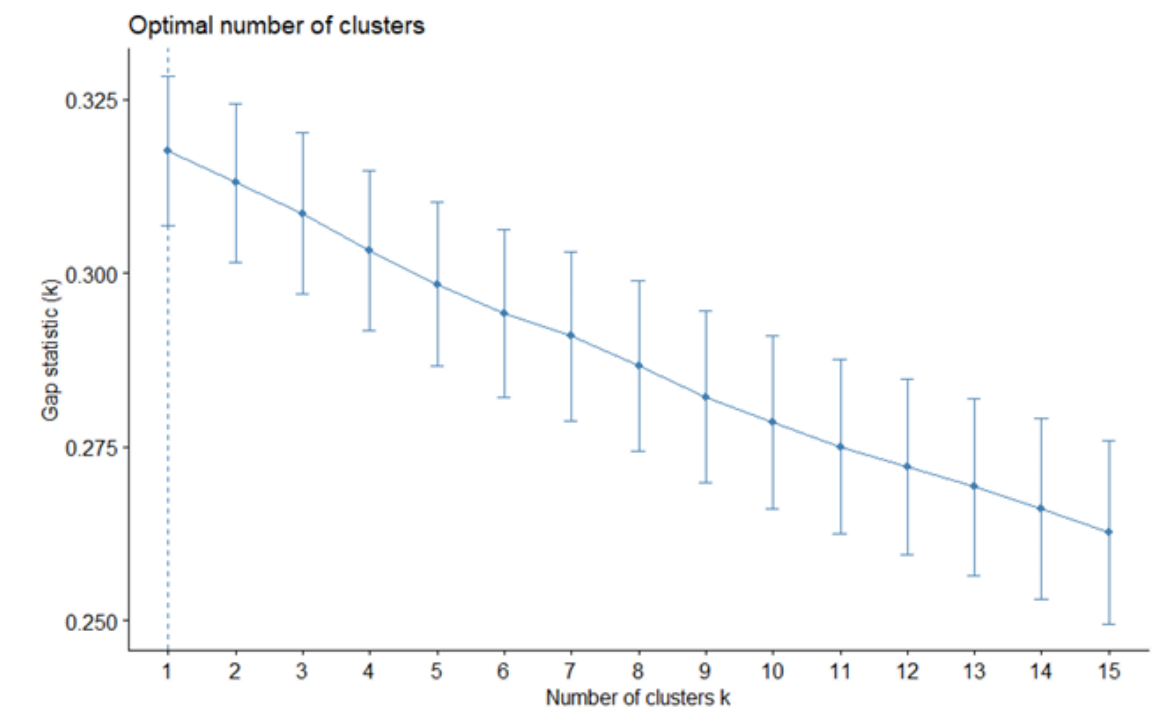
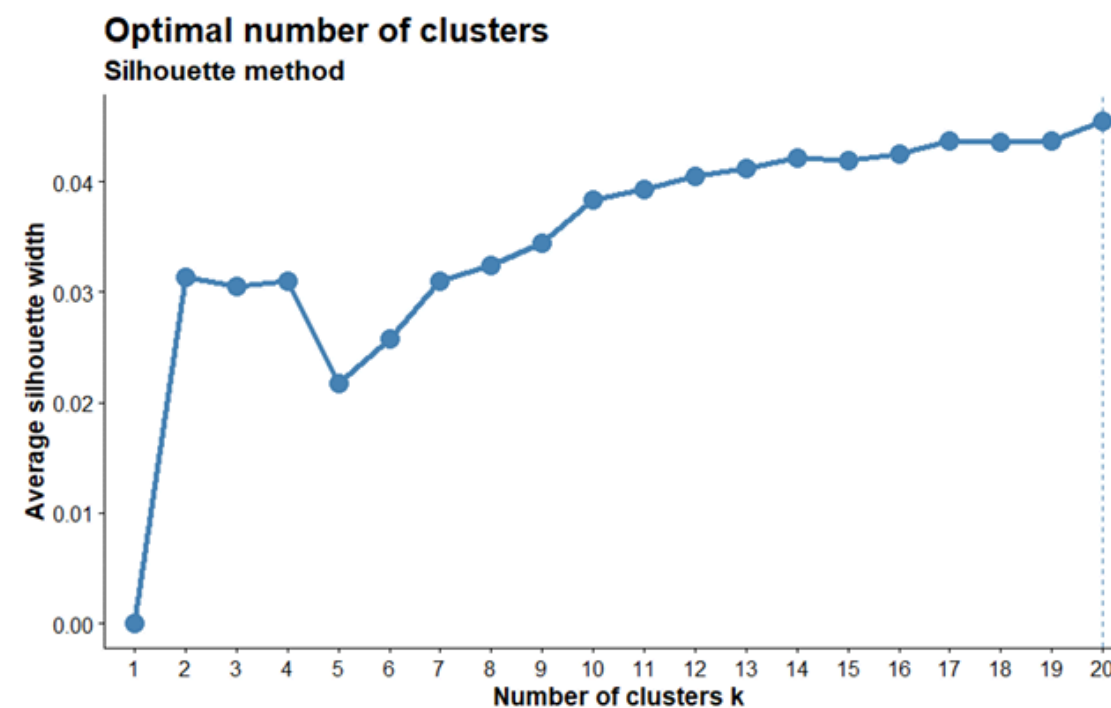
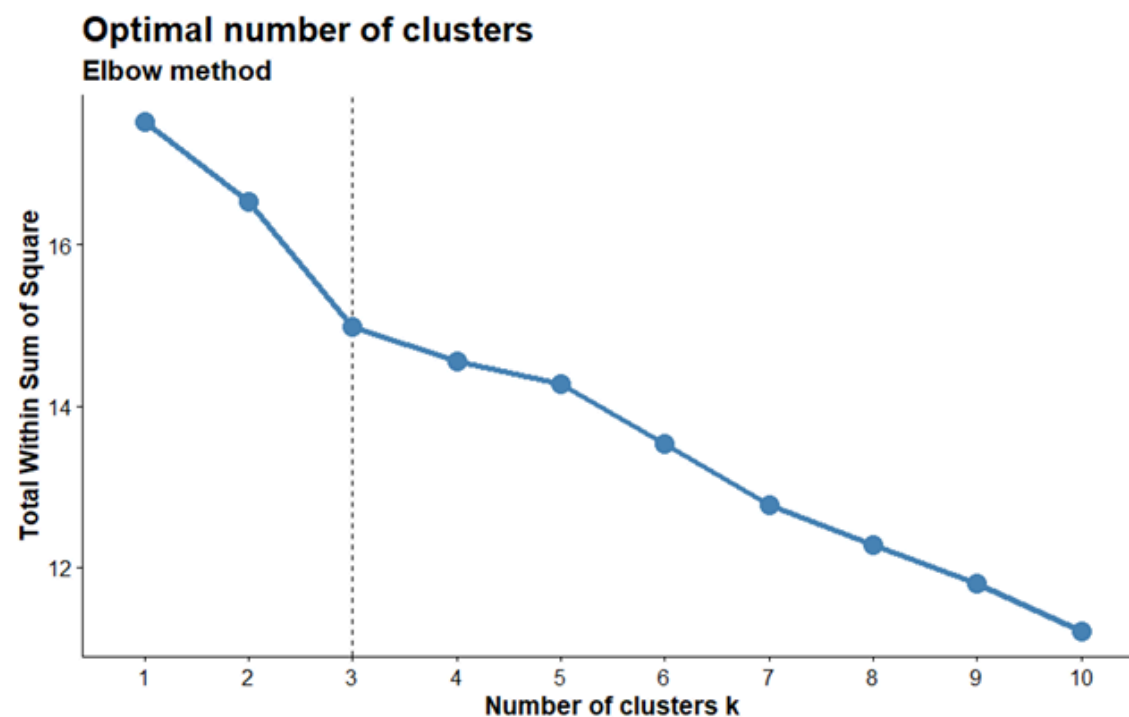
- Visualización inicial para detectar **patrones** y **anomalías**.
- **Gráficos individuales por aeropuerto** para mayor detalle.
- Comportamiento general **estable**, con **picos aislados**.
- **No** se observa **estacionalidad clara**, pero sí posibles **patrones recurrentes**.
- Variabilidad en picos sugiere diferencias en **tráfico** y **respuesta operativa**.
- Se validó que **todas las series** son **variables** (no constantes).
- Datos convertidos a **formato de series temporales** (510 observaciones por serie).

CÁLCULO DE DISTANCIAS

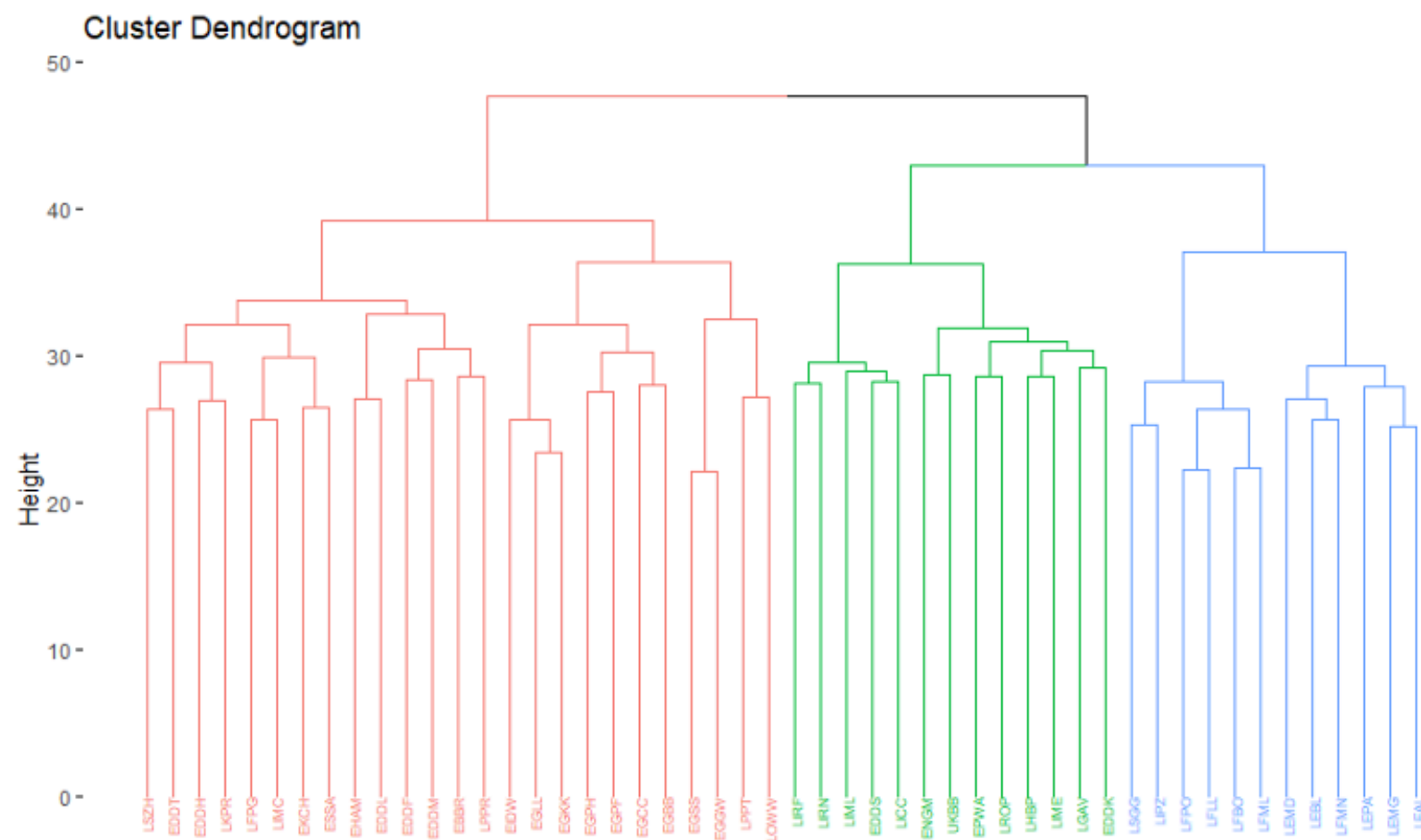
- **Normalización** de las series (media cero, varianza unitaria).
- Selección de la **distancia ACF** (función de autocorrelación) por su capacidad para:
 - Captar **dependencia temporal**.
 - Ser **robusta** a desplazamientos y escalas.
 - Comparar series de **diferente magnitud**.
 - Ser de **fácil interpretación**.
- **Alta similitud entre París y Frankfurt (0,479); mayor disimilitud con Dublín (0,715).**
- Otras métricas como **periodograma** o **euclídea** fueron **descartadas**.
- La métrica ACF permitió una **matriz sólida de similitudes**, base del **clustering jerárquico**.

CREACIÓN DE GRUPOS

- Identificación de **patrones comunes** con **clustering jerárquico aglomerativo**.
- No requieren **definir** el número de **grupos previamente**. 3 métodos empleados:
 - **Método del codo**: analiza la **compactación de los grupos** y detecta el punto donde **añadir más clusters aporta poco valor**.
 - **Coeficiente de silueta**: mide **cohesión interna y separación entre grupos**.
 - **Estadístico Gap**: compara la **dispersión intra-cluster** con la de un agrupamiento aleatorio para ver si los grupos son **estadísticamente significativos**.

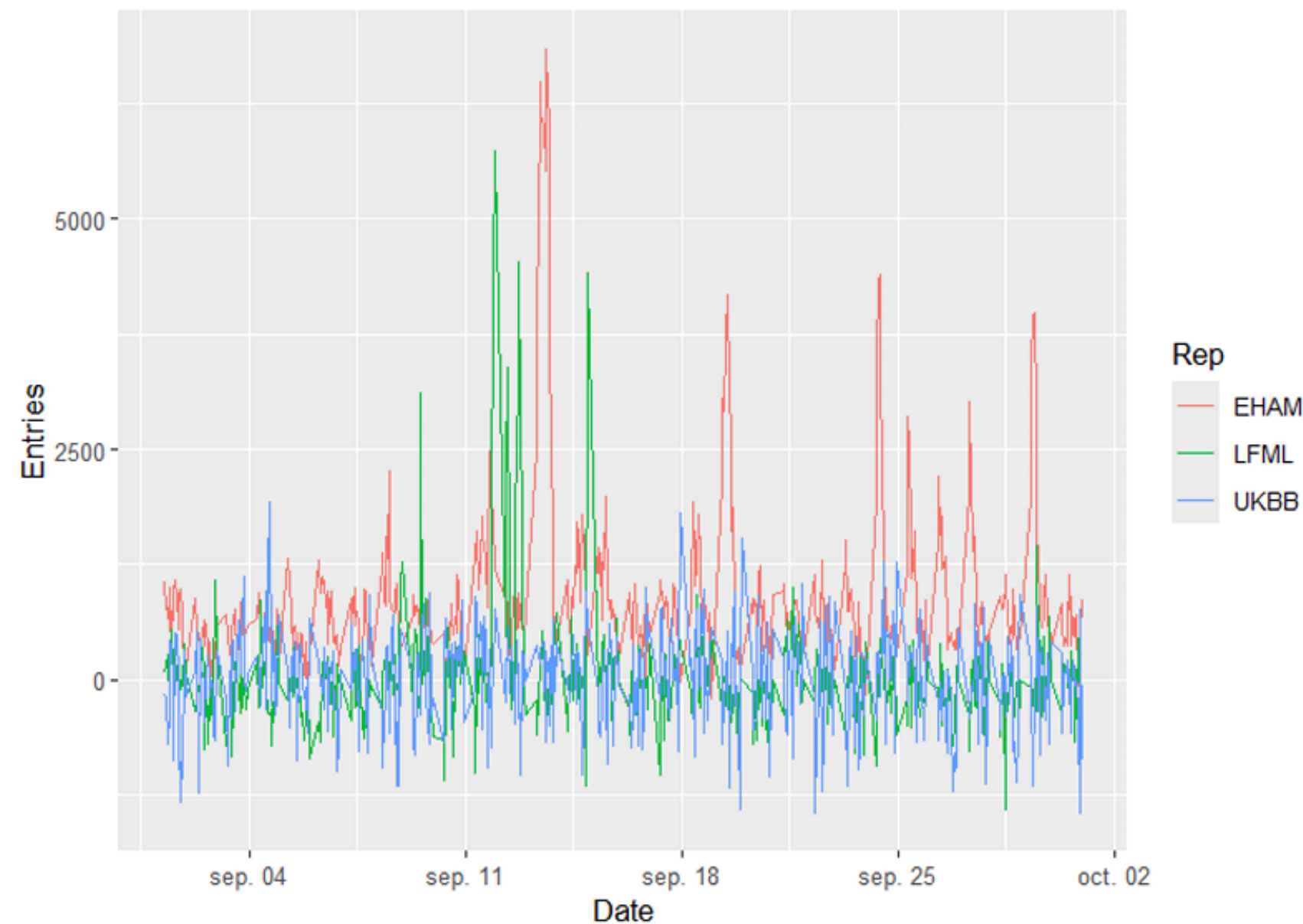


ELECCIÓN DE REPRESENTANTES



- **3 clusters** para equilibrar detalle y claridad interpretativa.
- Las visualizaciones ayudaron a **interpretar la agrupación y validar la coherencia entre series.**
- **Una serie representativa por cluster** (mayor varianza):
 - **EHAM** (grupo 1, 25 aeropuertos)
 - **LFML** (grupo 2, 12 aeropuertos)
 - **UKBB** (grupo 3, 13 aeropuertos)
- Se **descartaron** otros criterios (**media o valores extremos**) por menor representatividad.

INTRODUCCIÓN A LA PREDICCIÓN



- Predicción formulada como regresión de **variable continua**, no clasificación.
- Se usó una **ventana móvil de 7 días** y un **horizonte de predicción de 1 día**.
- **Diferencias claras** entre series:
 - **EHAM y LFML: alta variabilidad y picos coincidentes.**
 - **UKBB: patrón más estable.**

EVALUACIÓN DE MODELOS

Métrica	Penalización	Ventajas	Adecuado para
MSE	Cuadrática	Sensible a errores grandes	Minimizar desviaciones severas
MAE	Lineal	Robusto ante valores extremos	Obtener error promedio realista

MODELOS PREDICTIVOS

- **ARIMA** (clásico): captura **relaciones lineales** y **autocorrelación**, ideal para **series estacionarias**.
- **Holt-Winters** (suavizado exponencial): modela **tendencia** y **estacionalidad**, útil con **patrones repetitivos**.
- **KNN** (vecinos más próximos): basado en **analogía histórica**, **no paramétrico** y **flexible**.
- **ARNN** (red neuronal autorregresiva): captura **relaciones no lineales** entre observaciones pasadas y es capaz de modelar **comportamientos más complejos**.
- **SVR** (máquina de vectores soporte): modela **relaciones no lineales** y es **robusto al ruido**.
- **TBATS** (suavizado avanzado): maneja **múltiples estacionalidades**, para series con **patrones complejos**.
- **LSTM** (red neuronal recurrente): captura **dependencias de largo plazo**, útil para **secuencias complejas**.
- **Combinación de modelos**: **reduce varianza** y **compensa sesgos individuales** mediante:
 - **Media Aritmética**: asigna **pesos iguales** a todos los modelos.
 - **Bates & Granger**: pondera según la **varianza** y la **correlación de errores**.
 - **Mínimos Cuadrados (CLS)**: ajusta pesos según el **desempeño histórico** de cada modelo.

EVALUACIÓN DE MODELOS

- **TBATS** es el mejor modelo para el **Cluster 1**.
- **SVM** es el modelo más **preciso** para los **Clusters 2 y 3**.
- **Modelos tradicionales** como **ARIMA** y **Holt** son aceptables, pero **superados**.
- **KNN** y **ARNN** presentan **mayor error y variabilidad**.

Modelo	Cluster 1	Cluster 2	Cluster 3
ARIMA	758.14	570.58	403.20
HOLT	867.11	884.62	405.48
KNN	973.60	595.91	459.28
ARNN	1096.08	478.59	416.02
SVM	868.03	361.46	364.89
TBATS	622.68	566.59	373.43
LSTM	743.48	562.35	412.86

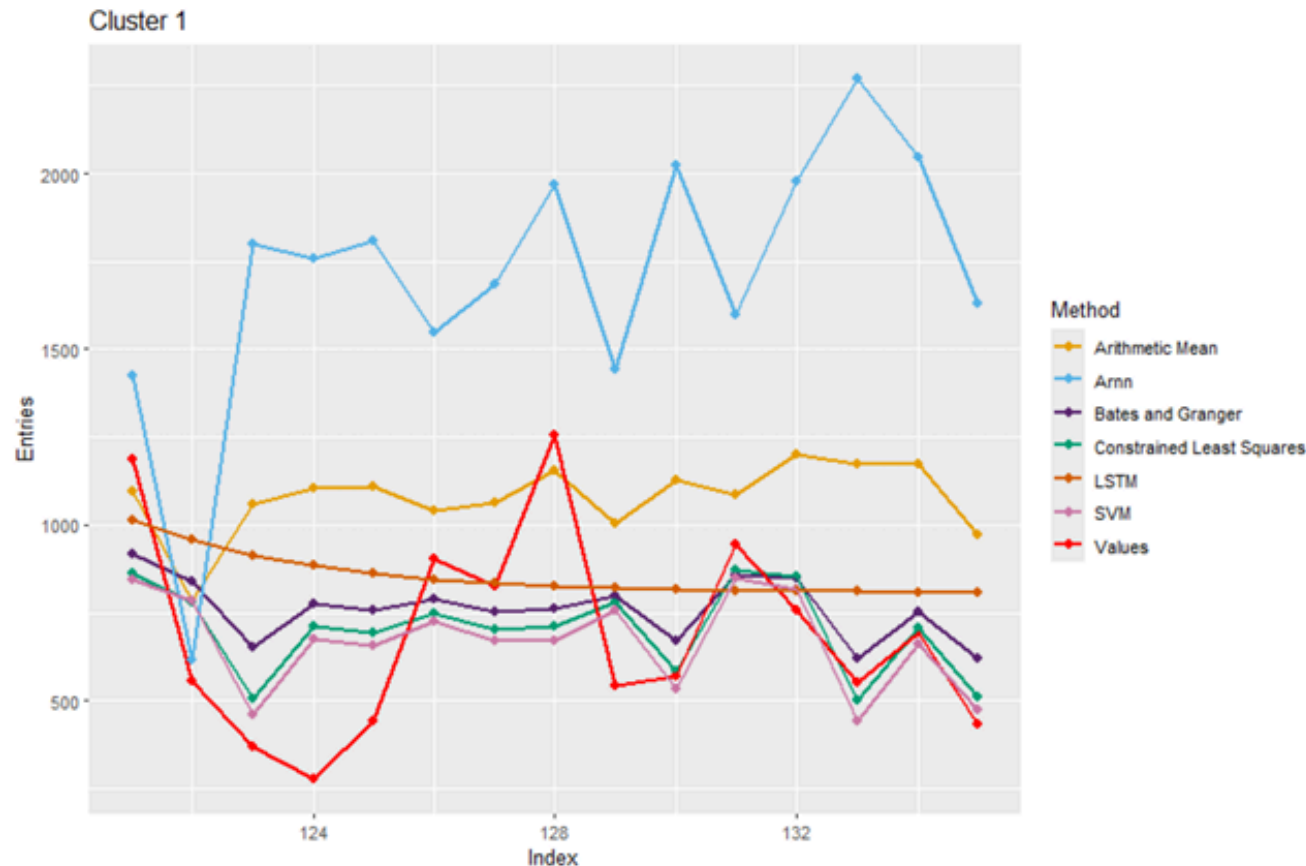
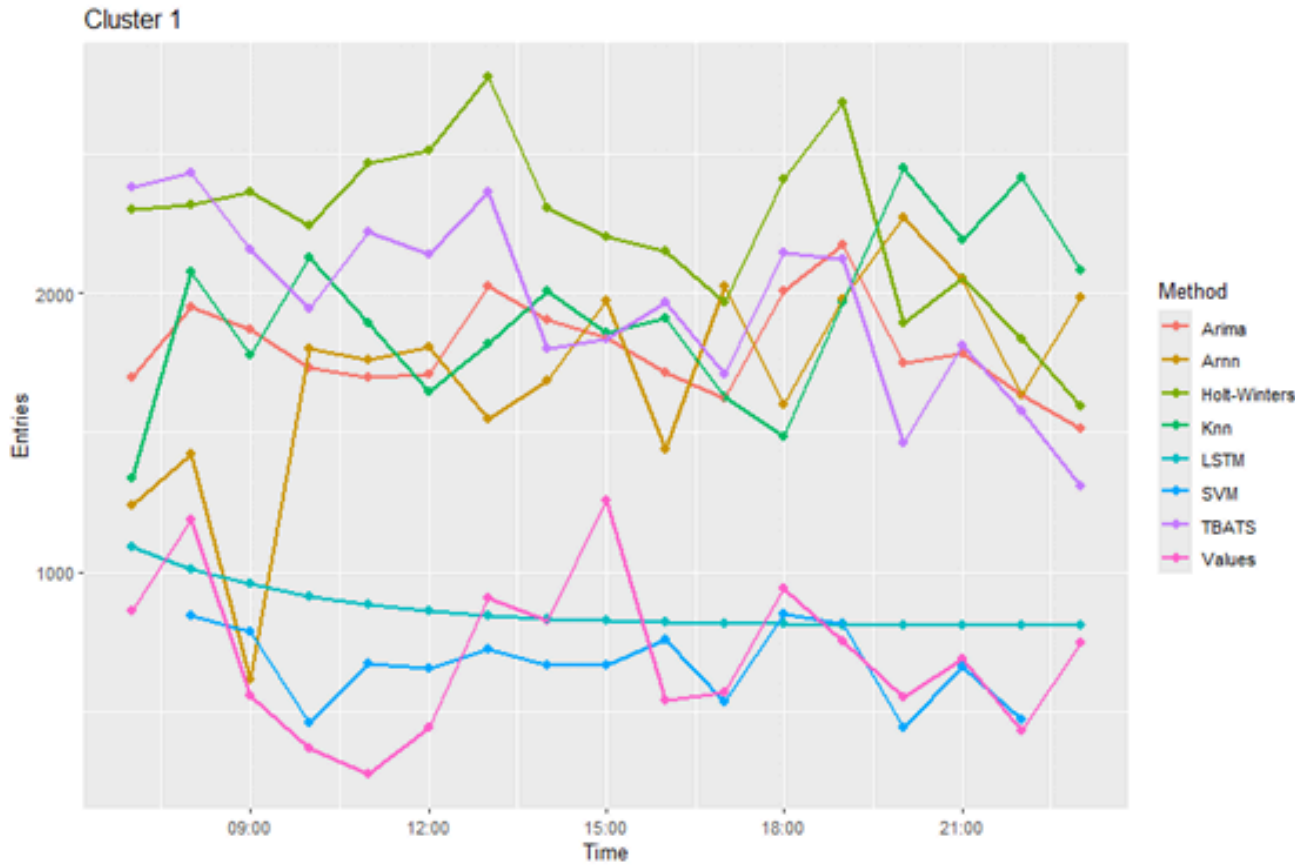
Cuadro 1: MAE promedio de cada modelo por cluster

Modelo	Cluster 1	Cluster 2	Cluster 3
ARIMA	1569217	1150100	266547
HOLT	1700984	5449426	271969
KNN	2608918	916,926	335861
ARNN	2721310	606,506	285287
SVM	11072929	381232	227009
TBATS	1218484	1123774	241919
LSTM	1482774	1225850	273796

Cuadro 2: MSE promedio de cada modelo por cluster

PREDICCIÓN CLUSTER 1

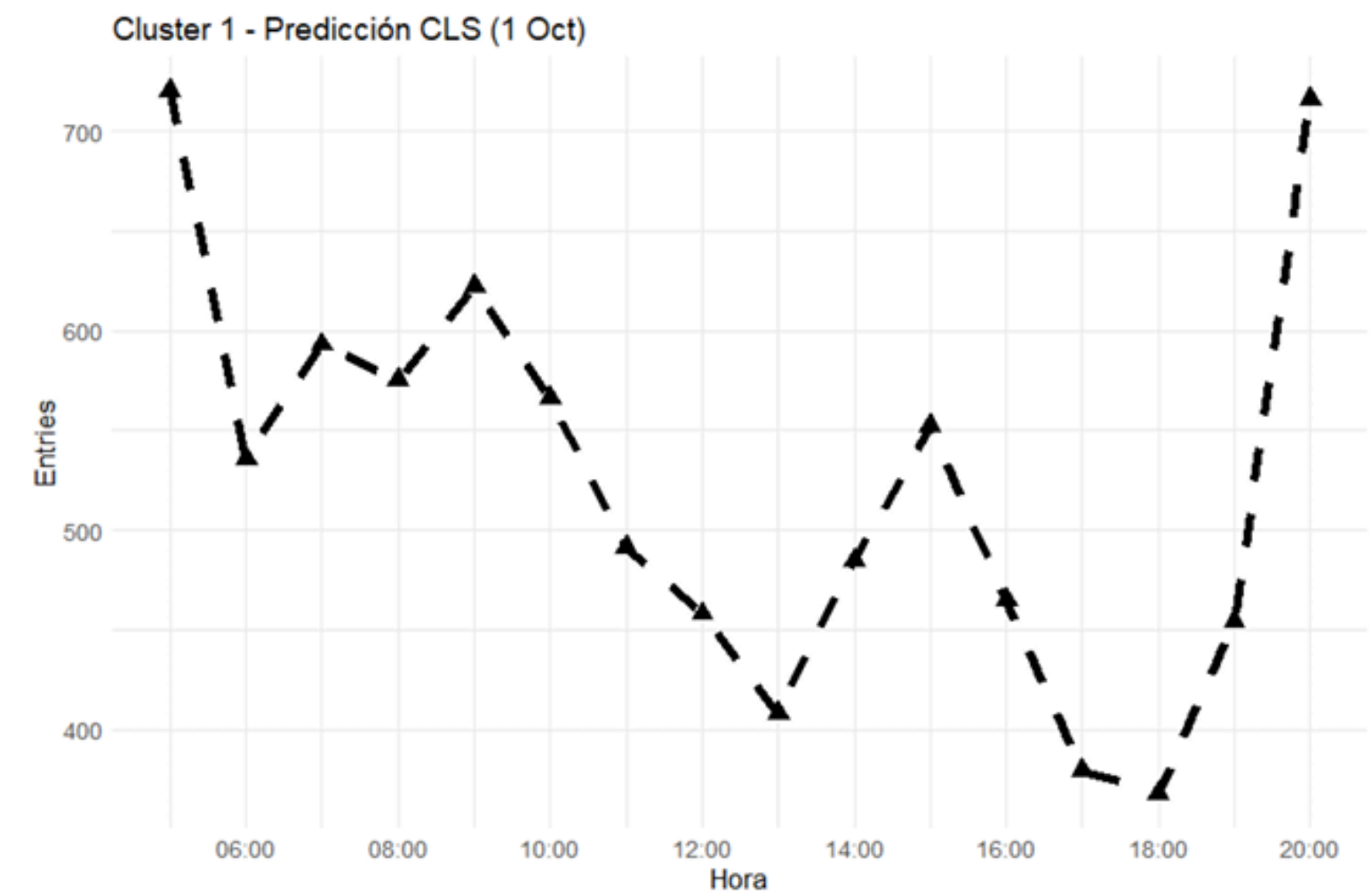
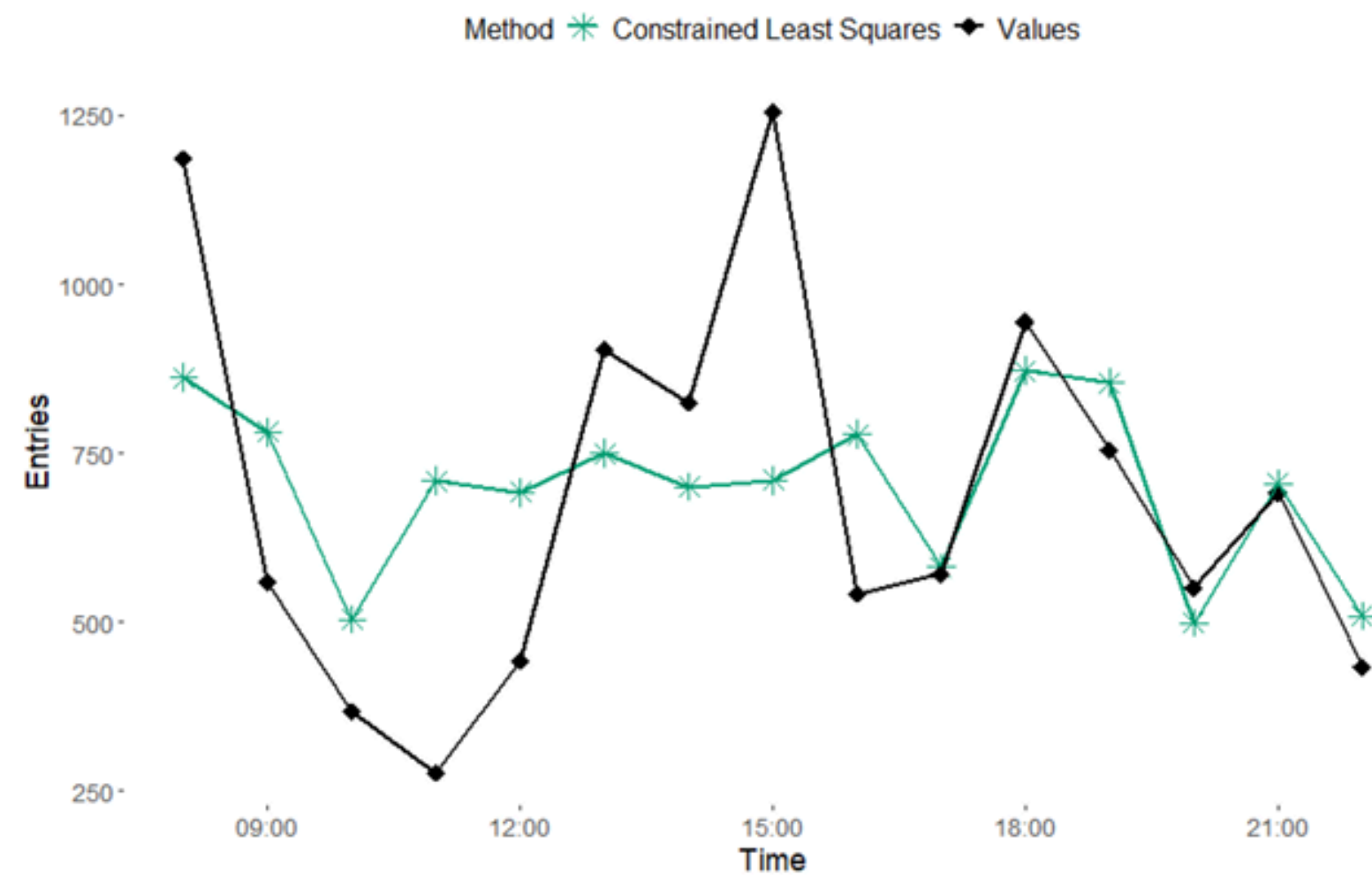
Modelo	MAE	MSE
ARIMA	1100.18	1265292.8
HOLT	1537.41	2497725.5
KNN	1220.53	1678895.4
ARNN	994.82	1210553.6
SVM	185.73	57101.2
TBATS	1261.65	1721901.2
LSTM	258.65	97410.88



Modelo	MAE	MSE
Arithmetic Mean (AM)	415.20	226881.70
Bates and Granger (BG)	212.13	65358.27
Constrained Least Squares (CLS)	183.27	55924.33

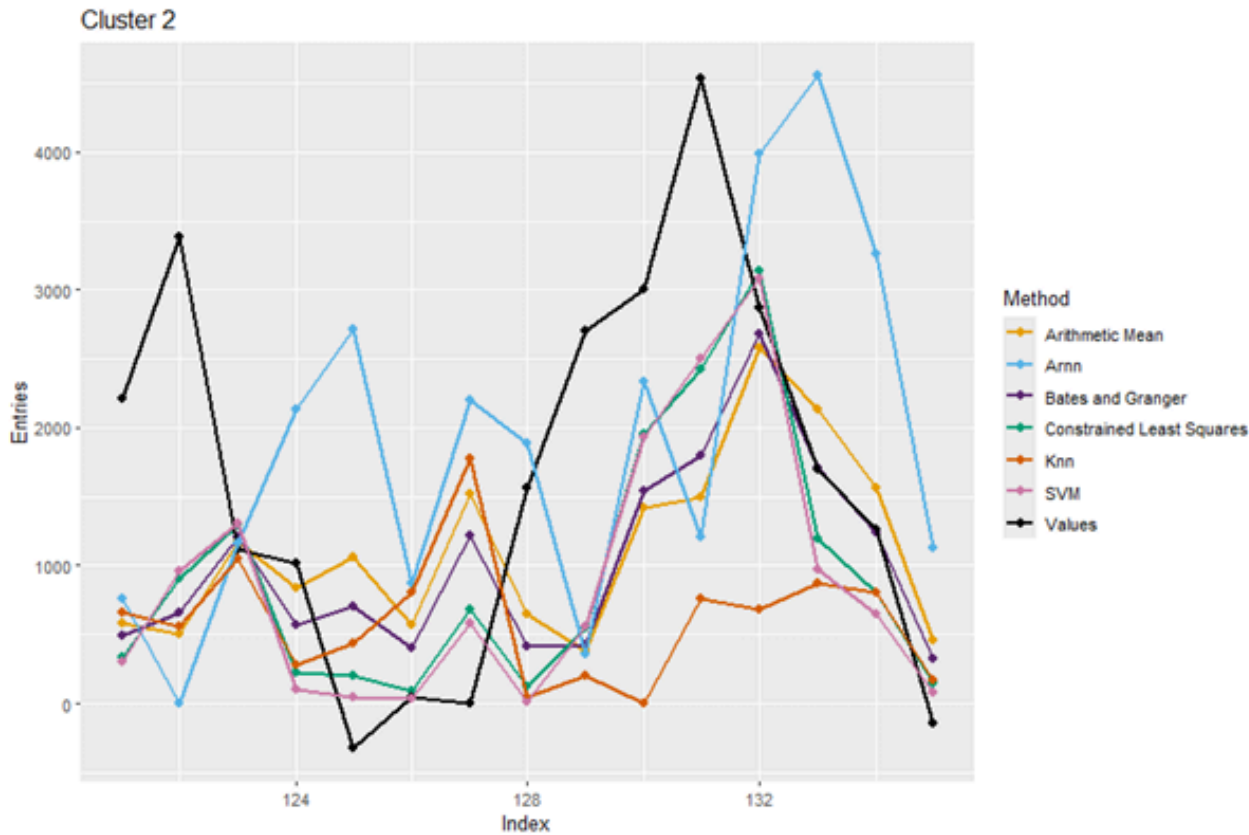
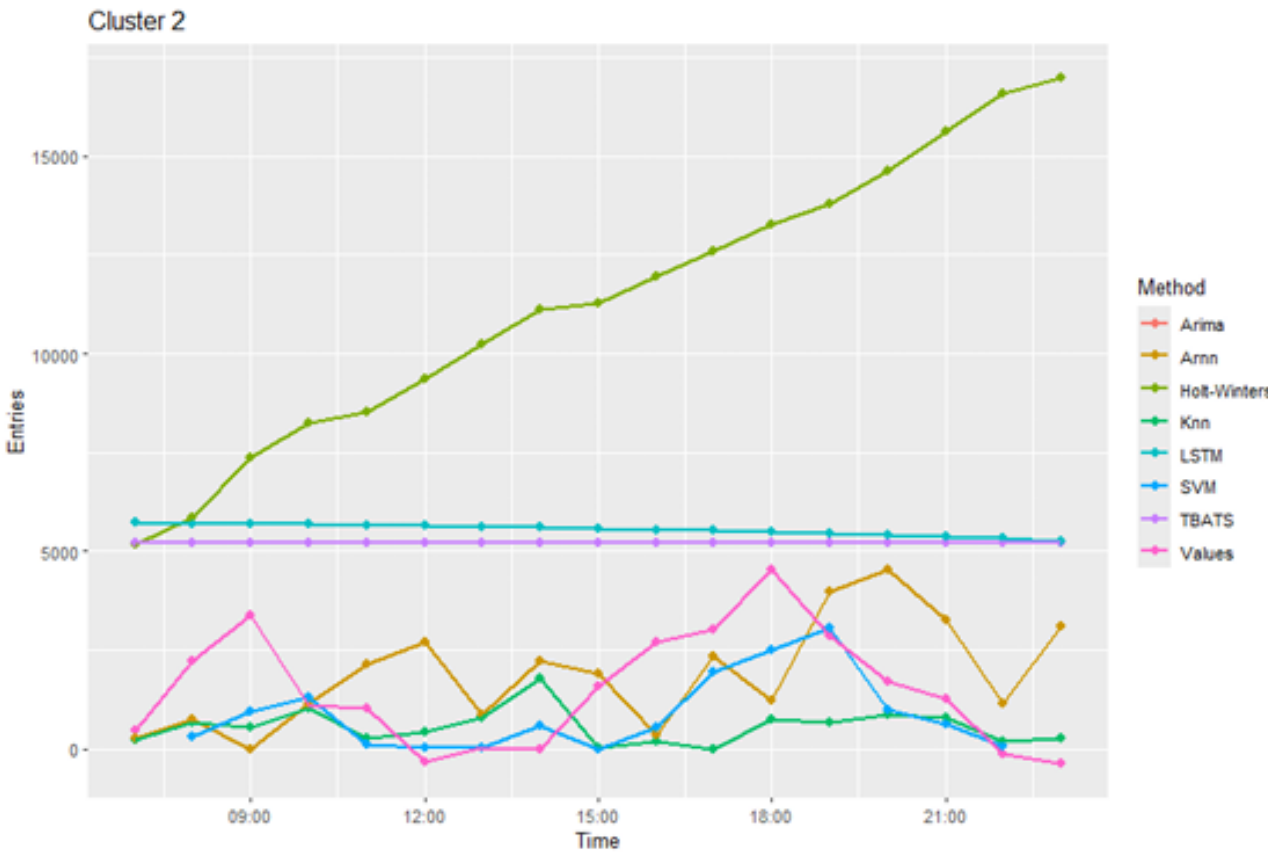
PREDICCIÓN CLUSTER 1

- **Dentro:** el modelo CLS ajusta bien la tendencia general, aunque presenta desviaciones en los picos más pronunciados.
- **Fuera:** coherente sobre datos nuevos, mostrando buena capacidad de generalización.



PREDICCIÓN
CLUSTER 2

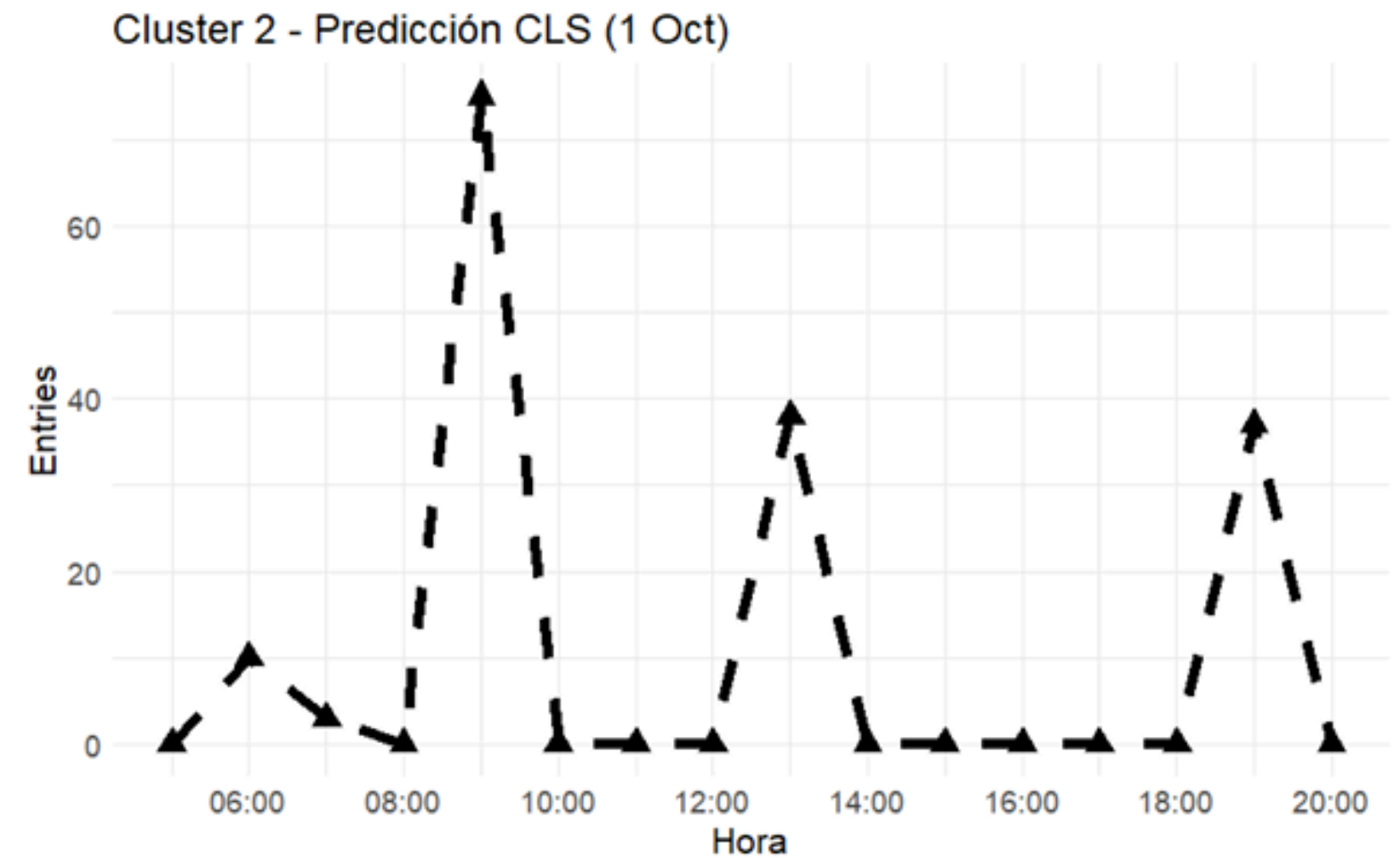
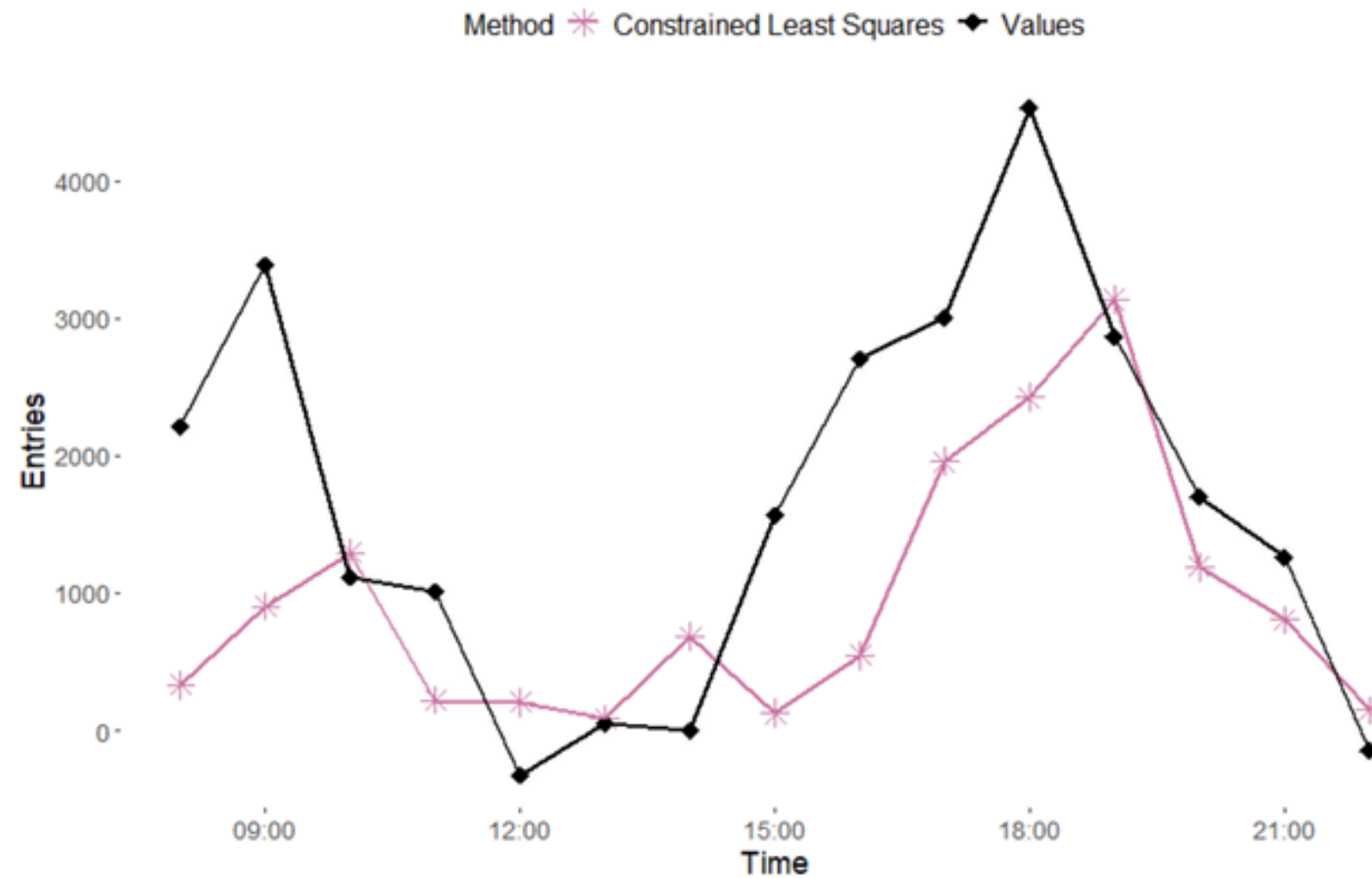
Modelo	MAE	MSE
ARIMA	3758.35	16140704
HOLT	9856.00	111865881
KNN	1407.82	3131977
ARNN	1745.71	4348640
SVM	998.60	1608639
TBATS	3750.35	16080634
LSTM	4078.12	18629458



Modelo	MAE	MSE
Arithmetic Mean (AM)	1178.27	2282163
Bates and Granger (BG)	1059.80	1974388
Constrained Least Squares (CLS)	990.53	1598080

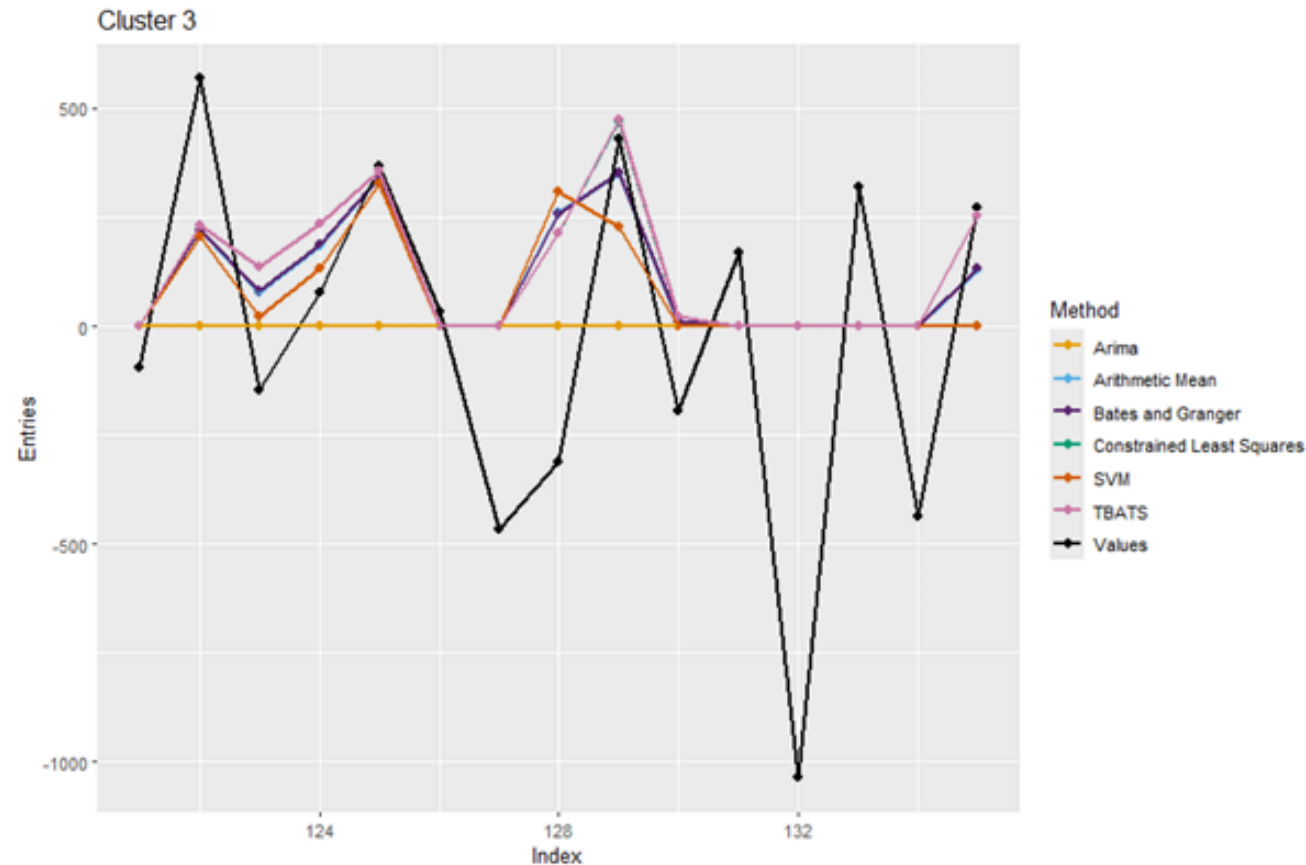
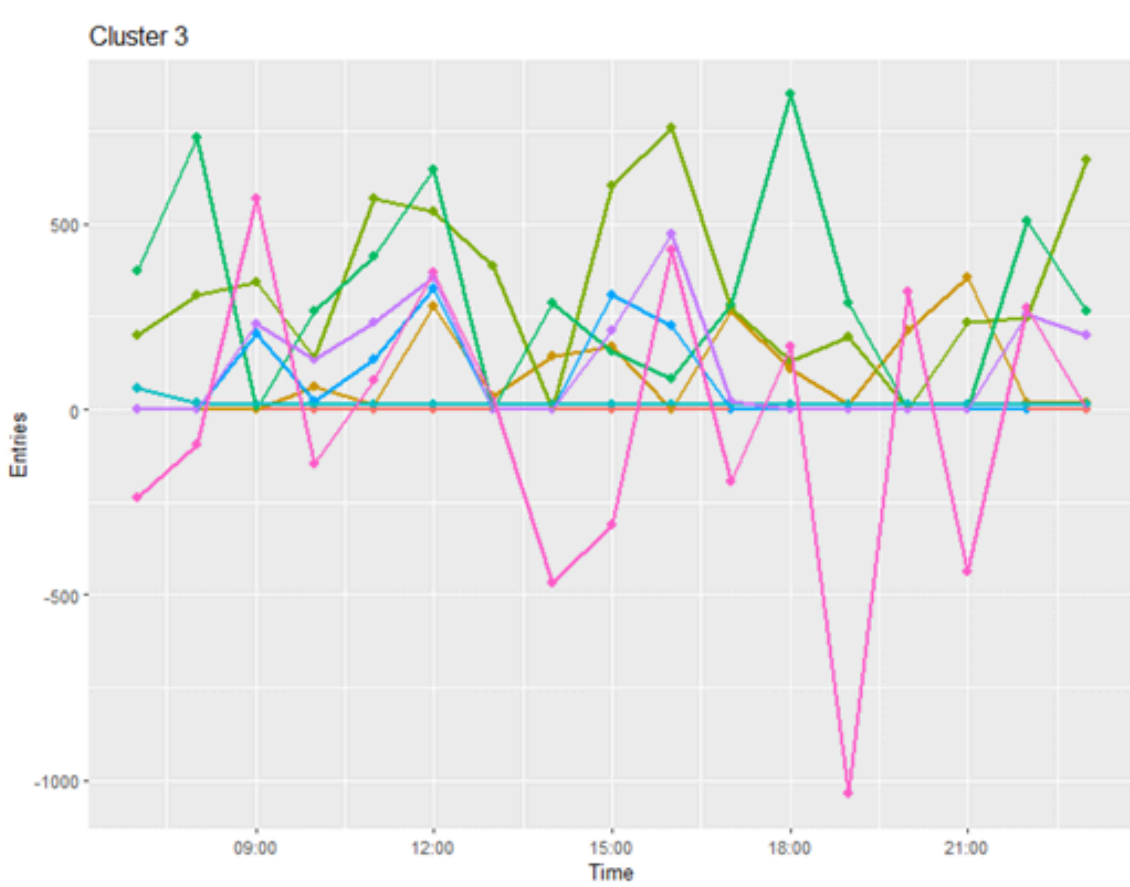
PREDICCIÓN CLUSTER 2

- **Dentro:** el modelo CLS ofrece buenos resultados, aunque presenta **ligeros desplazamientos horarios**.
- **Fuera:** se anticipa una **evolución suave** con varios **picos moderados**, incluyendo un **tramo plano sin entradas** a lo largo del día.



PREDICCIÓN CLUSTER 3

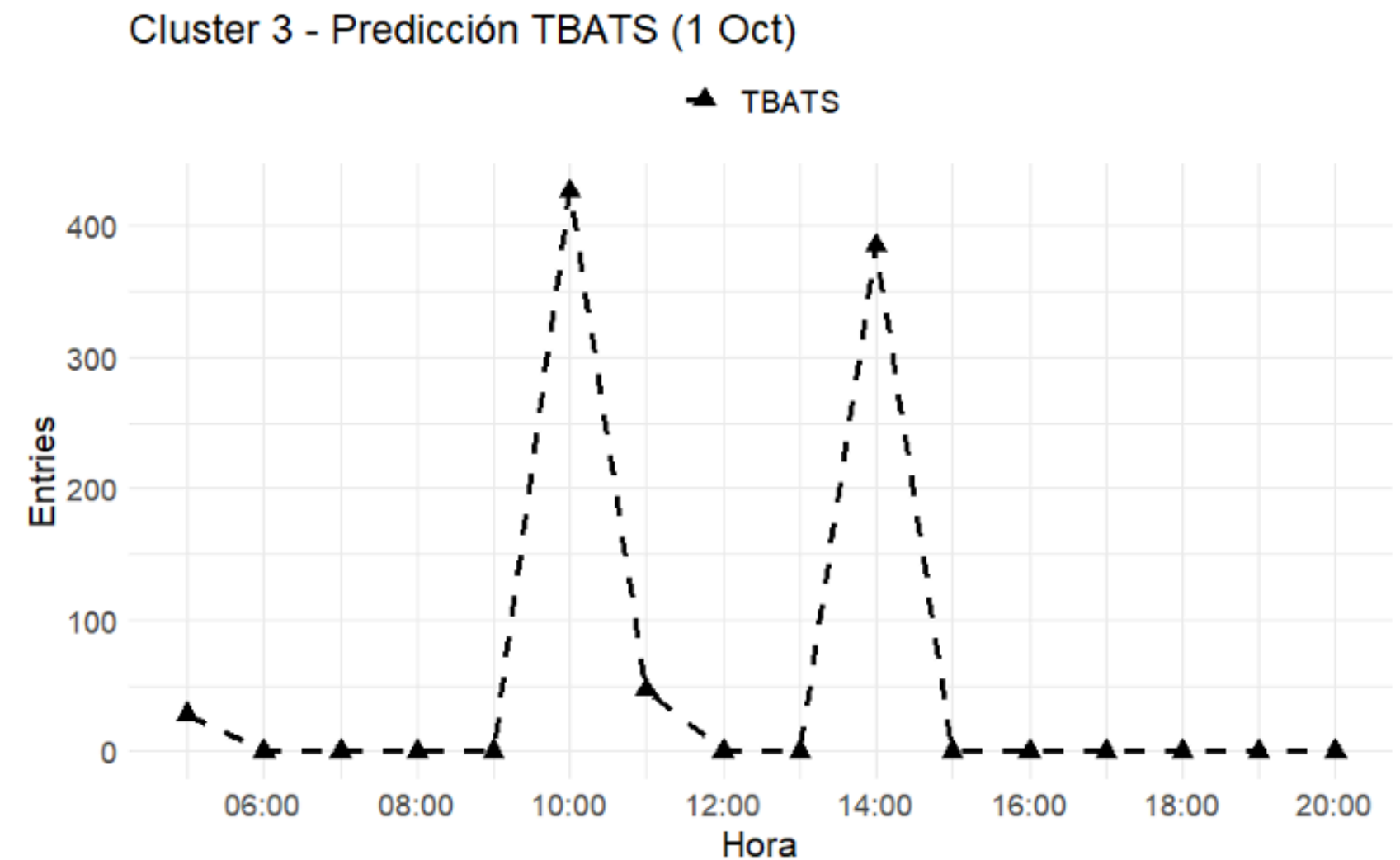
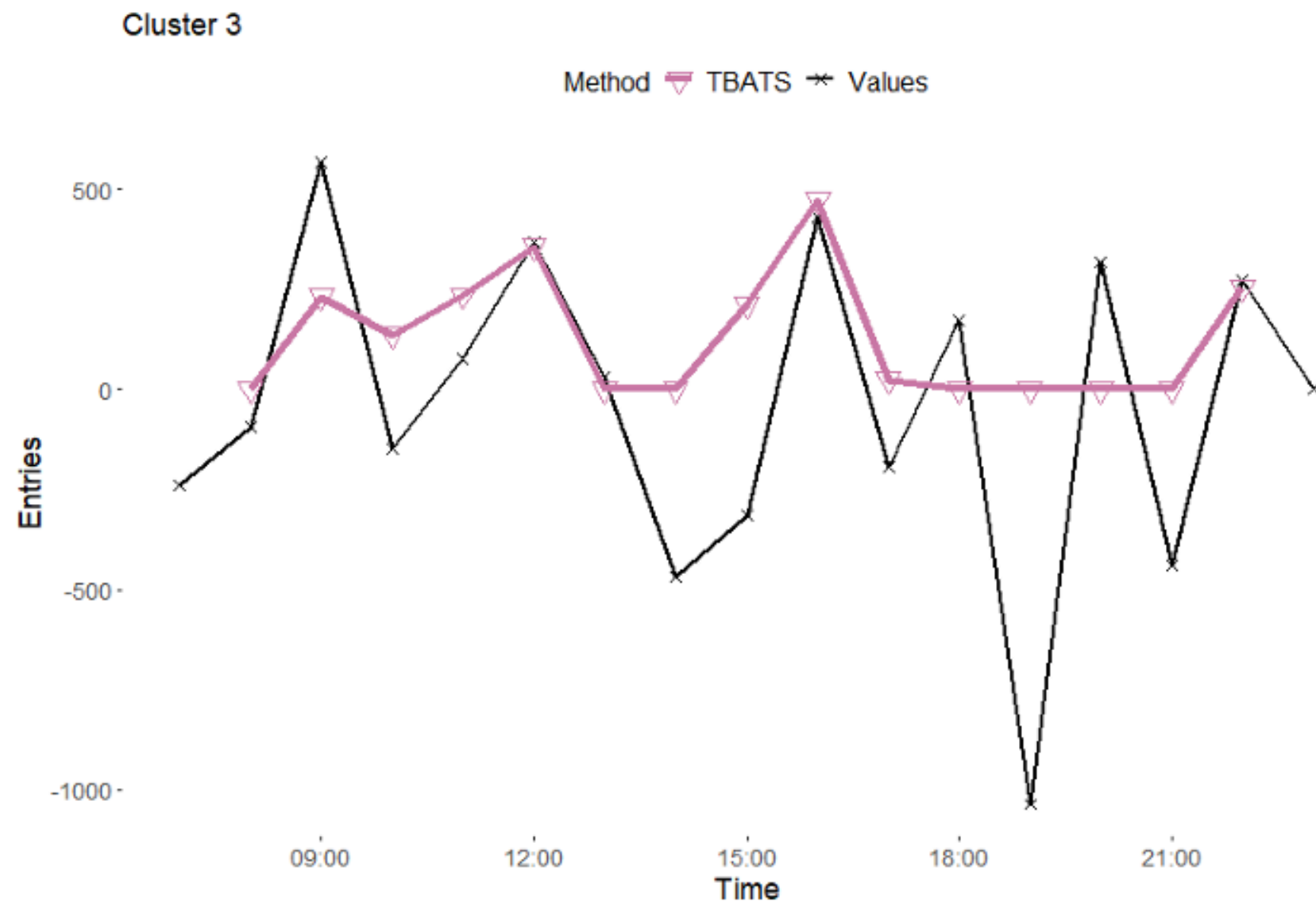
Modelo	MAE	MSE
ARIMA	304.06	150807.2
HOLT	441.41	281934.0
KNN	492.94	324825.2
ARNN	325.65	192073.3
SVM	298.53	155185.5
TBATS	270.18	133387.1
LSTM	307.65	153588.9



Modelo	MAE	MSE
Arithmetic Mean (AM)	284.80	147251.5
Bates and Granger (BG)	284.33	147066.9
Constrained Least Squares (CLS)	276.67	144576.8

PREDICCIÓN CLUSTER 3

- **Dentro:** el modelo LSTM reproduce **razonablemente** la forma general de la serie, aunque **no detecta todos los picos con precisión**.
- **Fuera:** solo se identifican dos picos (a las 9:00 y 13:00).



RESULTADOS GENERALES

- **SVM y TBATS** mostraron **mejor rendimiento** general frente a técnicas tradicionales.
- **Redes neuronales** ofrecieron **buenos con comportamientos complejos**, aunque limitadas por la escasez de datos.
- La **combinación de modelos** mejora la **robustez predictiva**, especialmente en clústeres con alta variabilidad.
- **Predicciones fuera de muestra** demostraron capacidad de **generalización**, incluso con entrenamiento en un solo mes.
- El enfoque por clústeres permite **adaptar modelos según las características del grupo** de aeropuertos, aumentando la precisión.

LÍNEAS DE DESARROLLO FUTURO

- **Ampliar** la serie temporal.
- Explorar **nuevas configuraciones de clustering**, ya que las actuales no garantizan segmentación óptima.
- Aplicar **técnicas de imputación** para manejar datos faltantes o discontinuos sin eliminar muestras útiles.
- **Incorporar variables exógenas** como clima, eventos especiales o incidencias técnicas para enriquecer el modelo.
- **Ampliar el horizonte de predicción más allá de 24 horas** y desarrollar sistemas operativos en tiempo real.

¡GRACIAS!