Proyecto Series Temporales

David Cardoner & Arnau Mercader

En el presente documento se pretende ilustrar y realizar un *forecast* a partir de datos recogidos de la empresa **Dexma**, empresa dedicada a la eficiencia y gestión energética. Los datos disponibles estan anonimizados, posiblemente de un cliente del sector alimentario. El rango de estos es el año 2017 y su frecuencia es cuarto horaria (96 observaciones/día). Los datos son de alta frecuencia (*high frequency data*), por lo que la metodología clásica ARIMA no es muy recomendable, ya que el número de parámetros necesarios para controlar el proceso puede llegar a ser excesivo y dar estimaciones inestables e ineficientes. Es por eso, que se implementará primero un análisis de factores (TSFA) y con estos se estimará un modelo vectorial autoregresivo (VAR). Para contrastar el modelo VAR, se comparará con una predicción ponderada de un modelo AR y una descomposición del factor de manera univariante.

Palabras clave: Análisis de factores, *high frequency data*, modelos autoregresivos AR-VAR, variables latentes.

Metodología

El análisis factorial se utiliza como método de análisis de variaciones en la respuesta basada en variables latentes no observadas. Este método esta muy ligado al análisis de componentes principales, pero se asume un error general que de entrada no puede ser explicado por los factores. Al aplicar el modelo de factores, cada uno de ellos se puede interpretar como una serie individual. A la vez, este análisis permite caracterizar relaciones y patrones en los datos estudiando el comportamiento de estos.

Sea y_t un vector m-dimensional de longitud T y n el número de factores o variables latentes ($n << T$) recogidos en el n-vector ϕ. La relación entre y_t y los factores ϕ se asume lineal y se puede caracterizar mediante la expresión:

$$y_t = \alpha_t + L\phi_t + \epsilon_t$$

Donde α_t juega el papel de *intercept* y se puede despreciar si los datos y_t estan centrados. La matriz de loadings L es de dimensión $m\ x\ n$ y se considera invariante en el tiempo, y el error aleatorio ϵ se asume independiente de los factores ϕ. La matriz L es la que permite pasar de la estimación de los factores a la variable respuesta y_t.

El análisis se complementará con la utilización de variables explicativas para intentar mejorar la *performance* del modelo. Se crearán variables binarias para: caracterizar si el día es domingo, días festivos del año 2017 en España y períodos de festividad clásicos (Semana Santa y Navidad conjuntamente). Estos datos se han obtenido de la página: https://www.calendario-365.es/dias-festivos/2017.html (última consulta 06/06) mediante *scrapping*. El Rdata festivos_2017 proporciona esta información además de la variable respuesta.

Validación y determinación del número de factores

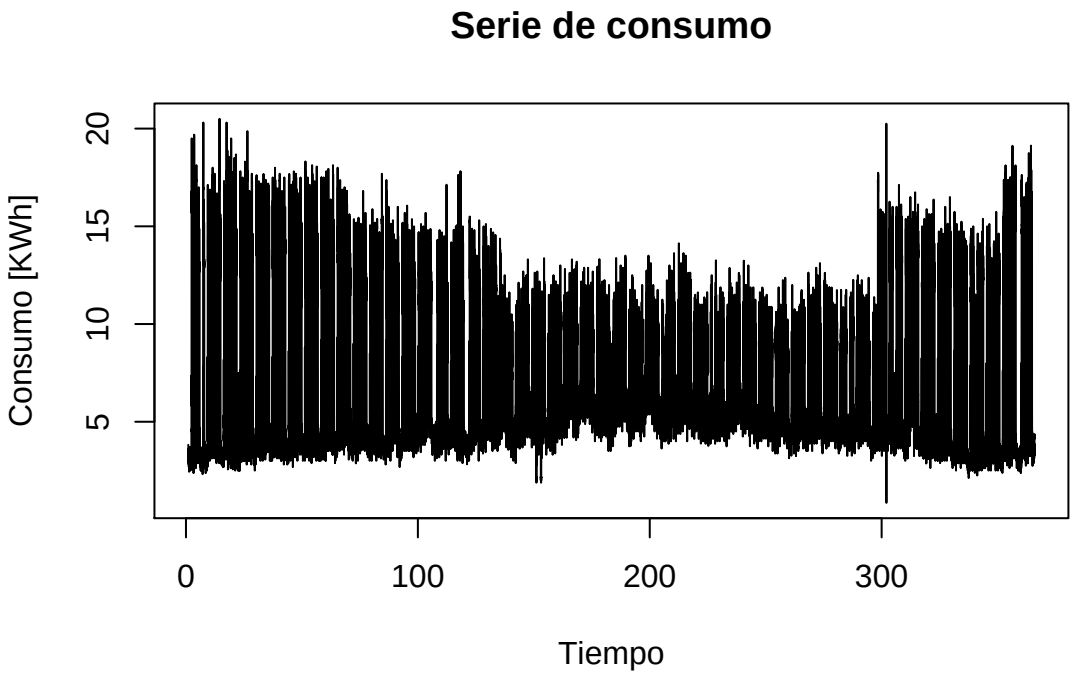
Para determinar el número de factores se contrastarán distintos enfoques: el Test de Bartlett, autovalores asociados a la matriz de correlaciones (dimensión $310 * 96$) de la serie multivariante superiores a la unidad y estadísticos como CFI y RMSEA obtenidos mediante la función *FafitStats* de R. Se reservarán 54 observaciones para validar el poder predictivo del modelo. Como métricas para ver la calidad de estas predicciones se usará RMSE, MAPE y la mediana del APE para intentar corregir el efecto de los domingos (consumo muy inferor) y posibles inestabilidades puntuales en el consumo.

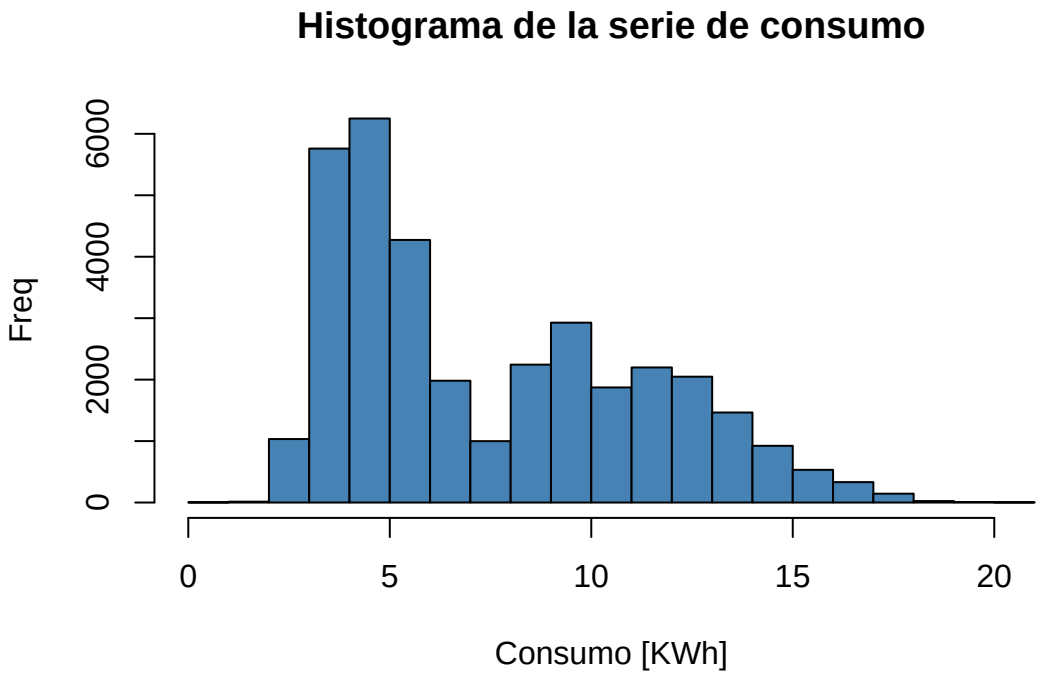
Definición de las métricas:

- RMSE $= \sqrt{n^{-1}\sum_{i=1}^{t}(y_i - \hat{y_i})^2}$
- MAPE $= n^{-1}\sum_{i=1}^{t}\frac{|y_i - \hat{y_i}|}{y_i}$
- median(APE) $= med\frac{|y_i - \hat{y_i}|}{y_i}$

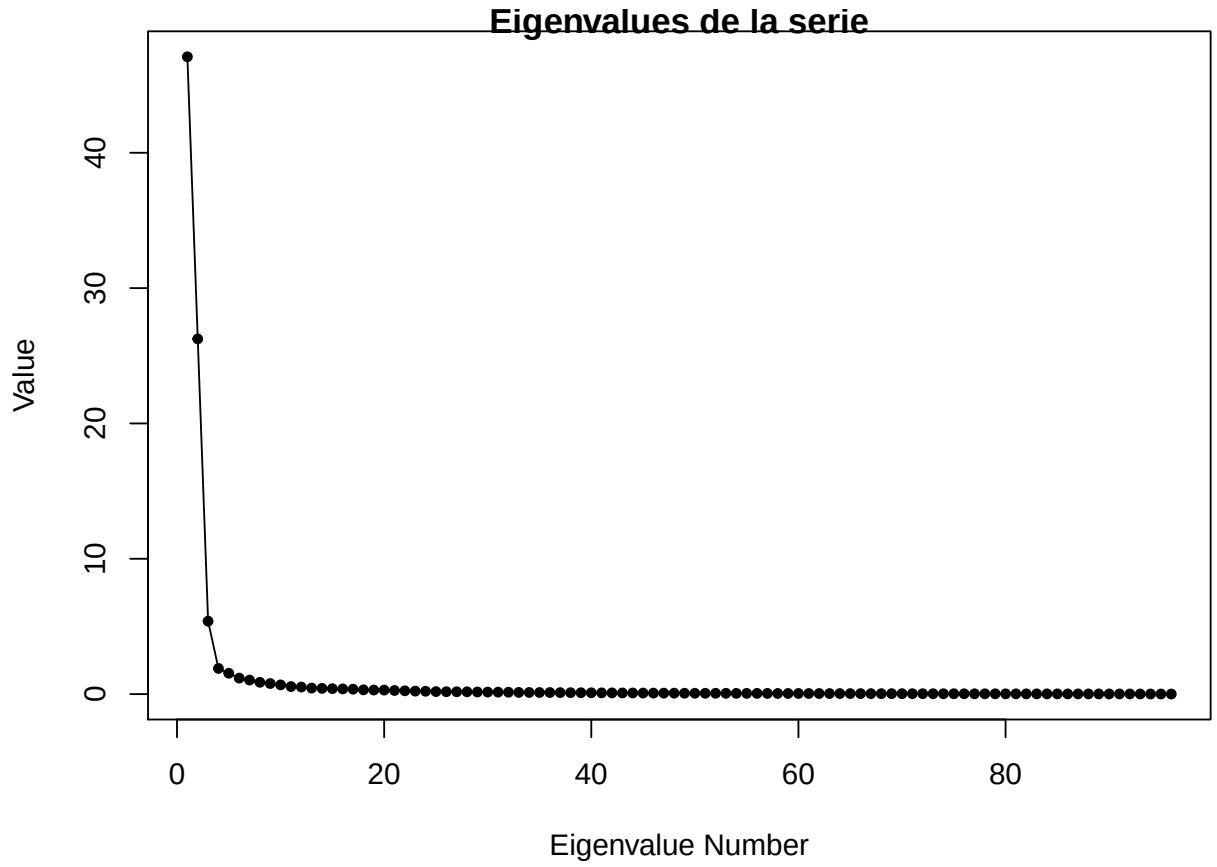
Análisis de los factores

A continuación se muestra un gráfico de la serie a analizar así como un histograma de la distribución conjunta donde se observa bimodalidad.





Seguidamente se muestra la validación basada en los *eigenvalues*. El gráfico ilustra el peso o significación de cada una de las columnas de los datos. Según este criterio se deben escoger como factores aquellos valores superiores a la unidad, en este caso, se deben seleccionar 7 factores.



Según el test de Bartlett que también utiliza como *input* la matriz de correlaciones de los datos, el número de factores resultante debe ser 5.

```
bartlett anderson   lawley
       5       19        5
```

Como última propuesta se usa la función *FAfitStats* para ver la significación de los factores en los estadísticos CFI y RMSEA. El CFI (*comparative fit index*) es un pseudo-R^2, con rango entre 0 y 1 y el RMSEA (*root mean square error of approximation*) mide la pérdida de ajuste por cada grado de libertad, la distribución de referencia se asume χ^2.

Usando esta propuesta, el número máximo de factores será igual o menor a la cota Ledermann, que en este caso es 82.

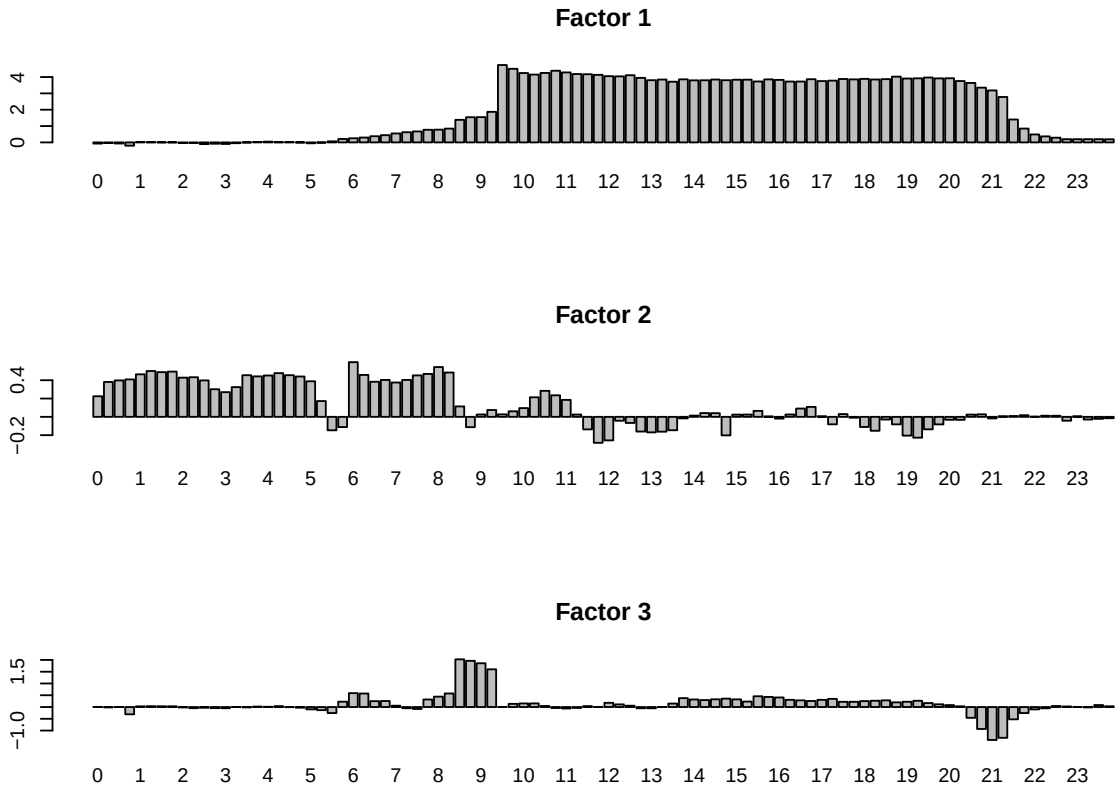
	RMSEA	CFI
0	0.1994223	0.0000000
1	0.1156454	0.6707935
2	0.1065637	0.7264174
3	0.0975812	0.7755312
4	0.0939166	0.7965976
5	0.0919977	0.8091183
6	0.0896762	0.8226653
7	0.0879668	0.8332020

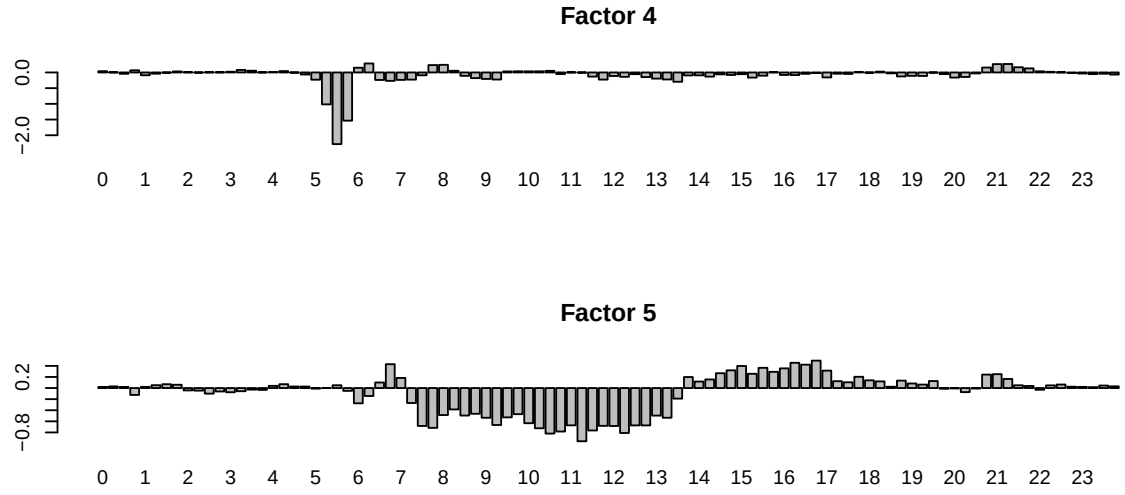
	RMSEA	CFI
8	0.0866468	0.8418550
9	0.0851115	0.8509247
10	0.0835999	0.8595259
11	0.0823979	0.8667561
12	0.0811126	0.8739642
13	0.0802986	0.8794677
14	0.0794539	0.8848796

Analizando todos los indicadores se decide utilizar 5 factores. Está conclusión se hace en base a la interpretabilidad y también analizando el CFI y RMSEA que muestran un estancamiento a partir del factor 5.

Estimación de los factores

Para realizar la estimación de los factores se diferenciarán los datos pero no se aplicará normalización. En los siguientes 5 gráficos se puede ver el peso de los factores en la serie. El factor que más peso tiene es el primero, que podemos interpretar como el consumo durante el tiempo que el establecimiento está abierto de cara al público. El segundo, recoge picos durante la madrugada hasta primera hora de la mañana, quizás necesidades de refrigeración, y los otros tienen un carácter más residual y bastante homogéneo entre los factores 3 y 4, que tienen dos picos, uno por la mañana entre 6 y 10 y otro por la noche sobre las 22.




Factor 4
0.0
−2.0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Factor 5
0.2
−0.8
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

Estimación y forecast

En esta sección, con los valores de los 5 factores se estimarán 2 modelos como ya se ha comentado: VAR y predición basada en la media ponderada de AR y descomposición de la serie univariante para cada factor. Se realizarán 2 variantes, una sin las variables exógenas y otra con ellas, para ver su efecto, además de una comparativa entre ambos enfoques.

Para las dos variantes se usará un modelo VAR con los 5 factores con retardo p=1, y estimando un *intercept*. Los modelos AR univariantes se estimarán mediante la función *auto.arima*, que buscará la mejor combinación con restricciones lag máximo p=7 y considerando el proceso estacionario. Para la descomposición de la serie se usará la función *ets* considerando el error y la tendencia aditiva y no considerando estacionalidad (combinación AAN). Para validar el peso que debemos dar en el forecast a la parte AR y *ets* se utilizará un *grid*, que ponderará de 0.1 en 0.1, los pesos de cada modelo para crear la combinación lineal que obtenga el mejor valor para los 20 días reservados como validación.

Para llevar a cabo todo este proceso se ha escrito una función R, donde en cada iteración se reestima el modelo TSFA con 5 factores para actualizar así los loadings (matriz L) y factores. En el material adjunto, existen 2 animaciones que permiten ver el progreso del algoritmo en el conjunto de test, simulando un poco los outputs que dan librerías recientes como Tensorflow. En la función se guardan las predicciones para poder evaluar las métricas diarias y globales.

Como punto final a destacar, se usará la función 2 veces, una sin considerar variables exógenas y otra añadiendo el código adecuado para poder usar exógenas en nuestros modelos.

Estimación sin variables exógenas

A continuación, se muestran las tablas con el valor de la métrica diaria y global generada para los datos de test.

Table 2: Métricas diarias

	VAR_rmse	Ens_rmse	VAR_mape	Ens_mape	VAR_median_ape	Ens_median_ape
1	2.528	2.500	0.2709	0.2736	0.2166	0.2173
2	2.927	2.804	0.2846	0.2817	0.2070	0.2287
3	3.436	2.235	0.3183	0.2365	0.2726	0.1879
4	4.154	2.404	0.3577	0.2407	0.3282	0.1529
5	3.550	2.280	0.3698	0.2857	0.3210	0.1870
6	4.627	6.144	0.9755	1.2788	0.9374	1.3313
7	2.247	1.756	0.2533	0.2299	0.1832	0.1666
8	2.262	1.667	0.2601	0.2222	0.2154	0.1342
9	2.078	1.713	0.2626	0.2459	0.1973	0.1668
10	2.132	1.563	0.2460	0.2070	0.1883	0.1106
11	1.832	1.602	0.2188	0.2021	0.1517	0.1079
12	2.141	1.636	0.2388	0.2121	0.1681	0.1075
13	4.763	6.285	0.9928	1.2804	1.0562	1.3624
14	2.063	1.768	0.2542	0.2302	0.1752	0.1811
15	1.957	1.570	0.2506	0.2207	0.1847	0.1167
16	1.759	1.580	0.2199	0.2063	0.1553	0.0970
17	1.891	1.503	0.2325	0.2001	0.1861	0.0984
18	2.236	1.716	0.2287	0.2001	0.1548	0.1188
19	2.140	1.547	0.2472	0.2069	0.1867	0.1042
20	4.947	7.136	1.2116	1.6949	1.3653	2.0048
21	1.898	1.662	0.2354	0.2209	0.1402	0.1512
22	1.796	1.693	0.2296	0.2238	0.1642	0.1438
23	1.633	1.550	0.2078	0.2015	0.1536	0.1373
24	1.834	1.541	0.2292	0.2017	0.1639	0.1092
25	1.600	1.467	0.2096	0.1976	0.1499	0.1212
26	1.529	1.324	0.2323	0.2140	0.1608	0.1115
27	6.066	6.376	1.7437	1.8227	1.9507	2.1528
28	1.774	1.587	0.2353	0.2286	0.1247	0.1589
29	1.489	1.595	0.2095	0.2124	0.1363	0.1421
30	5.503	5.399	1.3521	1.3249	0.7791	0.7769
31	2.041	1.618	0.2480	0.2237	0.1505	0.1412
32	1.487	1.638	0.2197	0.2285	0.1202	0.1468
33	1.871	1.909	0.2362	0.2402	0.1751	0.1788
34	5.612	5.870	1.3590	1.4174	1.6267	1.7312
35	1.821	1.843	0.2249	0.2270	0.1294	0.1712
36	1.931	1.901	0.2421	0.2331	0.1729	0.1681
37	1.754	1.771	0.2292	0.2265	0.1605	0.1486
38	1.741	1.877	0.2188	0.2246	0.1547	0.1420
39	1.799	1.758	0.2193	0.2206	0.1434	0.1364
40	1.803	1.690	0.2433	0.2250	0.1754	0.1484
41	6.065	5.915	1.6137	1.5687	1.6581	1.6352
42	2.427	2.817	0.2646	0.2732	0.2189	0.2281
43	2.521	3.565	0.2735	0.3104	0.2070	0.2771
44	3.753	3.564	0.3308	0.3105	0.3083	0.2817
45	3.607	3.528	0.3170	0.3086	0.2904	0.2735
46	3.623	3.374	0.3059	0.2932	0.2879	0.2638

	VAR_rmse	Ens_rmse	VAR_mape	Ens_mape	VAR_median_ape	Ens_median_ape
47	3.836	3.455	0.3190	0.3049	0.2969	0.2634
48	5.644	6.321	1.4552	1.6073	1.6944	1.9307
49	6.334	6.088	1.7056	1.6461	1.5024	1.7268
50	2.858	3.420	0.2780	0.3037	0.2210	0.2855
51	2.499	2.912	0.2575	0.2785	0.2044	0.2459
52	3.210	3.175	0.2943	0.2833	0.2624	0.2372
53	3.864	3.351	0.3094	0.2883	0.2967	0.2520
54	3.843	3.677	0.3180	0.3135	0.2986	0.2829

Table 3: Métricas globales

VAR_rmse	Ens_rmse	VAR_mape	Ens_mape	VAR_median_ape	Ens_median_ape
3.215	3.274	0.4456	0.4548	0.226	0.2058

Estimación con variables exógenas

Tras realizar los modelos sin variables exógenas se adjuntan seguidamente las tablas obtenidas con el mismo formato que antes pero añadiendo variables exógenas en el modelo.

Table 4: Métricas diarias

	VAR_rmse	Ens_rmse	VAR_mape	Ens_mape	VAR_median_ape	Ens_median_ape
1	2.12	1.87	0.245	0.232	0.163	0.142
2	2.47	3.44	0.254	0.339	0.175	0.288
3	3.61	2.48	0.332	0.260	0.296	0.220
4	2.49	1.94	0.251	0.214	0.188	0.109
5	2.49	2.08	0.303	0.271	0.218	0.149
6	2.21	2.37	0.449	0.493	0.413	0.471
7	2.10	1.67	0.242	0.229	0.172	0.146
8	1.61	1.62	0.220	0.222	0.134	0.137
9	1.72	1.70	0.249	0.255	0.173	0.165
10	1.60	1.90	0.213	0.240	0.113	0.175
11	1.71	1.65	0.214	0.211	0.134	0.114
12	1.71	1.71	0.218	0.219	0.114	0.106
13	1.88	1.77	0.407	0.378	0.337	0.278
14	2.14	1.72	0.258	0.230	0.178	0.168
15	1.59	1.62	0.227	0.222	0.126	0.135
16	1.66	1.52	0.215	0.199	0.124	0.110
17	1.57	1.53	0.214	0.201	0.128	0.107
18	1.78	1.74	0.205	0.196	0.142	0.119
19	1.61	1.56	0.213	0.207	0.137	0.110
20	2.16	2.19	0.540	0.541	0.491	0.513
21	1.92	1.59	0.237	0.223	0.145	0.153
22	1.58	1.55	0.225	0.219	0.129	0.127
23	1.47	1.39	0.208	0.196	0.117	0.100
24	1.53	1.49	0.215	0.202	0.133	0.113
25	1.51	1.47	0.211	0.207	0.108	0.099
26	1.44	1.43	0.236	0.228	0.137	0.127
27	2.59	2.01	0.783	0.605	0.754	0.598
28	1.78	1.50	0.233	0.225	0.119	0.137
29	1.52	1.49	0.212	0.207	0.134	0.123
30	5.18	4.60	1.277	1.143	0.791	0.765
31	1.87	1.60	0.233	0.225	0.118	0.131
32	1.75	2.65	0.234	0.285	0.156	0.260
33	1.72	1.67	0.230	0.225	0.148	0.124
34	2.39	1.96	0.584	0.447	0.565	0.388
35	1.89	1.59	0.227	0.211	0.145	0.132
36	1.62	1.66	0.229	0.222	0.129	0.110
37	1.79	1.64	0.229	0.216	0.124	0.127
38	1.58	1.60	0.208	0.202	0.125	0.121
39	1.63	1.55	0.209	0.200	0.107	0.104
40	1.67	1.60	0.231	0.218	0.127	0.111
41	2.41	2.50	0.657	0.686	0.619	0.647
42	2.54	2.44	0.269	0.256	0.241	0.199
43	2.24	2.77	0.261	0.278	0.174	0.213
44	1.85	1.94	0.245	0.254	0.155	0.157
45	1.80	1.77	0.231	0.229	0.148	0.156
46	1.74	1.71	0.207	0.207	0.116	0.101

	VAR_rmse	Ens_rmse	VAR_mape	Ens_mape	VAR_median_ape	Ens_median_ape
47	1.98	1.88	0.236	0.229	0.141	0.140
48	3.98	3.57	1.043	0.939	0.984	0.864
49	6.19	6.28	1.664	1.702	1.485	1.800
50	2.40	3.00	0.252	0.292	0.150	0.248
51	1.60	1.96	0.205	0.228	0.111	0.144
52	1.77	2.42	0.227	0.245	0.143	0.181
53	1.99	2.35	0.228	0.244	0.126	0.169
54	2.00	1.82	0.231	0.221	0.136	0.123

Table 5: Métricas globales

VAR_rmse	Ens_rmse	VAR_mape	Ens_mape	VAR_median_ape	Ens_median_ape
2.26	2.21	0.331	0.32	0.172	0.172

Conclusiones

- Añadiendo variables exógenas al modelo la calidad predictiva mejora moderadamente, debido, por ejemplo al factor domingo, donde el establecimiento esta cerrado y solo se observa un consumo constante debido a la refrigeración del local (por su sector de actividad).
- El enfoque *Ensemble* sin considerar variables auxiliares da más peso a la predicción *ets*. Sin embargo, al usarlas la predicción solo da peso a los AR univariantes ya que proporcionan más información.
- El hecho de utilizar un modelo VAR con un retardo o modelos AR univariantes mezclados con descomposición dan en los dos escenarios analizados resultados similares.
- Podríamos haber enfocado esta predicción como un claro de ejemplo de aplicación de modelos de Machine learning, como XGboost, Random Forest o Redes Neuronales. La idea de reducir la dimensionalidad de los datos mediante factores es rica por su simplez y efectividad, a la vista de los resultados obtenidos. Quizás la parte más compleja es determinar que factores usar y como predecirlos con tal de obtener buenas predicciones. Es en esta última parte donde quizás es bueno combinar una herramienta más clásica o estadística con un enfoque más computacional y moderno como podría ser algun algoritmo de machine learning mencionado arriba.
- Aprender esta metodología ha sido enriquecedor para nosotros y nos ha permitido ver una idea *parecida* al PCA, el cual si habíamos trabajado. Los factores latentes o subyacentes a la serie permiten añadiendo otras variables exógenas como por ejemplo la variable que permite controlar los domingos, generar una predicción bastante ajustada de la serie.
- El hecho de realizar el forecast para los días de navidad genera un handicap adicional provocado básicamente por el aumento de consumo de los hogares. Por ese motivo, se ha añadido la variable festividad que intentará recoger este aumento. Este factor considera como festividades los días de Semana Santa y el periodo de Navidad por igual.
- Como aspectos a mejorar, para seleccionar el peso de la combinación lineal se podrían haber usado técnicas más complejas como k-fold por bloques de tiempo o un rango más amplio de días. La utilización del método *ets* ha sido añadido para poder discutir un método tratado en clase.