Projecte final

David Hernández Sánchez
Alumne de DataScience
IT Academy
Barcelona, Catalunya
dissenyador@gmail.com

***Abstract*—En aquest document explicaré com he elaborat diversos models predictius per tal de pronosticar el nivell d'aigua de l'embassament de Darnius-Boadella. Amb aquest objectiu he elaborat diferents models d'autoregressió que fan servir una serie temporal dels últims 22 anys del nivell de l'embassament escollit, així com una altra serie de la precipitació a la mateixa ubicació de l'embassament com a variable exògena. El resultat ens diu que és molt important que les dades de precipitació provinguin d'una ubicació que tingui un impacte important en el nivell de l'embassament, sino no milloraran el model, com ha sigut el nostre cas. En definitiva, que necessitem dades externes amb una forta correlació amb les dades a pronosticar per obtenir una bona predicció, doncs amb un model autoregressiu sense cap variable exògena, serà difícil detectar grans variacions del nivell de l'embassament degut a plujes intenses o periodes de sequera.**

I. Introduction

Tenim una base de dades amb una serie temporal des del 01-01-2000 fins al 13-12-2022 de tots els embassaments de les conques internes de Catalunya, així com un altre serie temporal des del 01-01-2000 fins al 16-12-2022 de la precipitació a l'estació automàtica ubicada a l'embassament de Darnius-Boadella.
Volem predir el nivell de l'embassament de Darnius-Boadella fent servir la serie històrica del nivell de l'embassament així com la precipitació com a variable exògena.
Amb aquest objectiu comprovarem que les series històriques estiguin completes, en cas contrari les completarem. A continuació mirarem la relació existent entre les dues variables que volem fer servir per a la predicció. Posteriorment desenvoluparem diversos models autoregressius fent servir aquestes variables i finalment compararem els resultats de tots els models i extreurem conclusions.

II. State of Art

No he trobat enlloc cap model predictiu sobre el nivell de qualsevol dels embassaments de Catalunya, però si una publicació de l'Agència Catalana de l'Aigua, de títol **Aigua i canvi climàtic** [1] sobre l'afectació a les masses d'aigua de Catalunya pel canvi climàtic, que fa una diagnosi sobre els impactes previstos. En aquest document, encara que no es presenta cap model predictiu, si que es presenten diferents escenaris futurs de les aportacions del riu Muga a l'embassament de Darnius, basats en diferents models climàtics. En tots els casos la previsió es d'una disminució considerable, entre el 16.3% i el 38.4%, en l'aportació del riu Muga a l'embassament de Darnius. Encara que aquestes previsions es focalitzen en periodes molt llunyans, a partir del 2070, així que no es comparable al nostre model, que s'enfocaria en un periode de temps molt més propers, és a dir en els pròxims anys.
Un altre factor a tenir en compte, encara que en molt menor mesura que la precipitació, i que no he introduït al nostre model, és l'augment de la temperatura ambiental que comporta un increment en l'evaporació de l'aigua dels rius i embassaments, disminuint el seu cabdal. En aquesta publicació s'esmenta aquest factor en referència a les previsions de recursos hídrics: *Aquests supòsits encara no incorporen els canvis significatius que es puguin donar en l'evaporació als embassaments (probables creixements del 5%), que tot i tractar-se d'un factor de segon ordre, també seran un element més afegit a escenaris futurs de més escassetat hídrica.*

Per un altre banda, en la memòria tècnica del **Pla especial d'actuació en situació d'alerta i eventual sequera** [2] també elaborat per l'Agència Catalana de l'Aigua, trobem que en l'apartat 2.2.Tendències, i cito textualment, *en l'actualitat es podria estar produint una certa tendència a la baixa en les aportacions fluvials dels nostres rius.[. . .] a nivell pluviomètric només ara es comencen a detectar aquest tipus de tendències, i així ho confirmen els darrers Butlletins Anuals d'Indicadors Climàtics del Servei Meteorològic de Catalunya, que mostren una reducció mitjana de la pluja anual de l'ordre del 1,5% per dècada. [. . .] En definitiva, es tracta d'un fenomen encara en estudi però que en tot cas és preocupant per dos motius: en primer lloc perquè posaria en compromís la representativitat de les sèries històriques en moltes de les anàlisis que les utilitzen suposant la seva invariabilitat; i en segon lloc perquè planteja importants dubtes sobre la gestió dels propers anys si es mantenen aquestes tendències.*

III. Metodologia

A. Preprocessat

a) Dataframes: En aquest projecte hem treballat amb dos dataframes. El dataframe **embassaments.csv** conté dades sobre la quantitat d'aigua als embassaments de les Conques Internes de Catalunya. Conté 8358 registres i es composa de 5 columnes:

Dia: Data de la mesura. Freqüència diària.
Tipus de variable: DateTime (dd/mm/yyyy).

Estació: embassament on s'ha fet la mesura.
Tipus de variable: String.

Nivell absolut (msnm): Nivell de l'aigua a l'embassament.
Tipus de variable: Numérica (float).

Percentatge volum embassat (%): Percentatge de volum d'aigua emmagatzemat respecte la capacitat màxima de l'embassament.
Tipus de variable: Numérica (float).

Volum embassat (hm3): Volum d'aigua emmagatzemat a l'embassament.
Tipus de variable: Numérica (float).

El dataframe **Darnius - Boadella.csv** conté dades sobre la precipitació de l'embassament de Darnius – Boadella. Conté 8345 registres i es composa de dues columnes:

DATA: Data de la mesura. Freqüència diària.
Tipus de variable: DateTime (dd/mm/yyyy).

PPT: Precipitació acumulada diària. Mesurada en mm.
Tipus de variable: Numérica (float).

Del Dataframe **embassaments.csv** hem copiat tota la informació de l'embassament que ens interesa , en aquest cas el Darnius Boadella, en un altre dataframe. Hem llistat la informació d'aquest nou dataframe i hem vist que només conté 8358 registres, així doncs falten dies en la serie histórica, doncs del 01-01-2000 fins al 13-12-2022 hi han 8383 dies. Així que hem completat la serie histórica amb els dies faltants i omplert les les columnes corresponents a aquests dies amb 0, assignant-li al dataframe una freqüencia diària. Després hem passat la columa Dia a format DateTime i l'hem fet servir com a index del dataframe. A continuació hem buscat els possibles NaNs i els que hem trobat els hem substituit per la mitjana del valor anterior i posterior de manera manual, doncs només havia 4 NaNs. Després hem buscat tots els valors que estiguin a 0 mitjançant un bucle for i els hem substituit també per la mitjana del valor anterior i posterior.

Al Dataframe **Darnius - Boadella.csv** hem seguit el mateix procediment inicial que a l'anterior. Hem vist que només conté 8345 registres, així doncs també falten dies en la serie histórica, i l'hem completat amb els dies faltants omplint la columna de precipitació corresponent a aquests dies amb 0, i li hem assignat al dataframe una freqüencia diària. Després hem passat la columa Dia a format DateTime i l'hem fet servir com a index del dataframe. A continuació hem buscat els possibles NaNs però no hem trobat cap. Com que les dades de precipitació contingudes en aquest dataframe no son continues no podem fer servir una mitjana dels valors veïns per omplir els buits com hem fet al dataframe anterior.

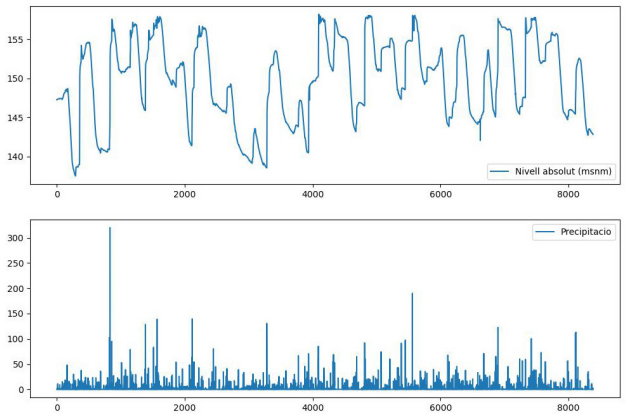
Com que el número d'observacions de la serie temporal de precipitacions té 3 dates més al final que la de la serie temporal de l'embassament, treuem aquestes 3 dates perque ambdues series siguin iguals.

Comprovem que tant la serie de precipitacions com la de l'embassament tenen el mateix número de dades, en el mateix interval de temps, i amb la mateixa freqüencia, i que totes les dades son numériques per poder treballar amb elles.

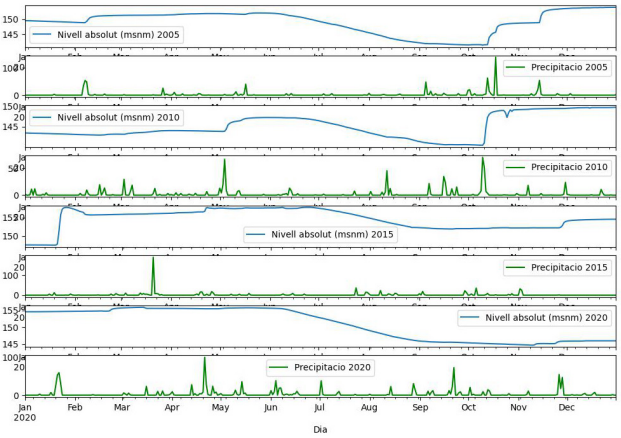
Unim els dos dataframes en un de sol i el guardem per poder fer-lo servir posteriorment.

b) Analisi estadístic: Analitzem estadisticament el Dataframe per veure quina es la mitjana, desviació, així com les diferencies estadístiques entre les diferents columnes. Comprovem que no hi han grans desviacions en general a cap columna. Per una altra banda, fent un cop d'ull als valor minims, màxims i als diferents percentatges(25%,50% i 75%), veiem que els valor es distribueixen de manera força homogenea a totes les columnes.

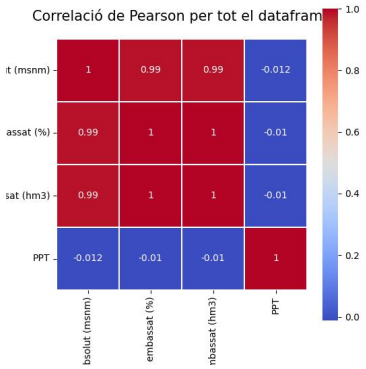
c) Analisi gràfic: Hem elaborat una serie de gràfiques per comprendre la distribució de les dades al llarg de la serie històrica. Primer hem fet una gràfica de tota la serie històrica del nivell absolut de l'embassament i a sota de la precipitació, per veure si trobem alguna relació entre ambdos.



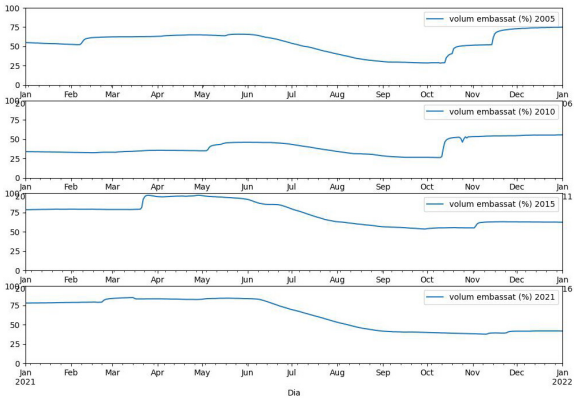
I com que no hem trobat cap hem fet 4 gràfiques amb periodes d'un any al llarg de la serie.



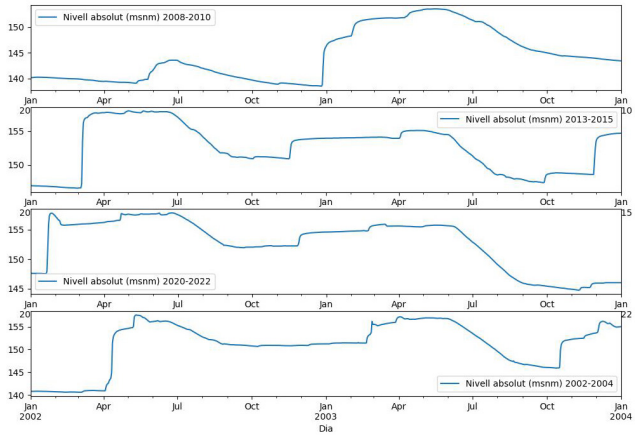
En continuar sense veure gaire relació fem a continuació una gràfica de correlació de Pearson. Veiem que hi ha una correlació negativa molt fluixa entre la precipitació i el nivell de l'embassament.



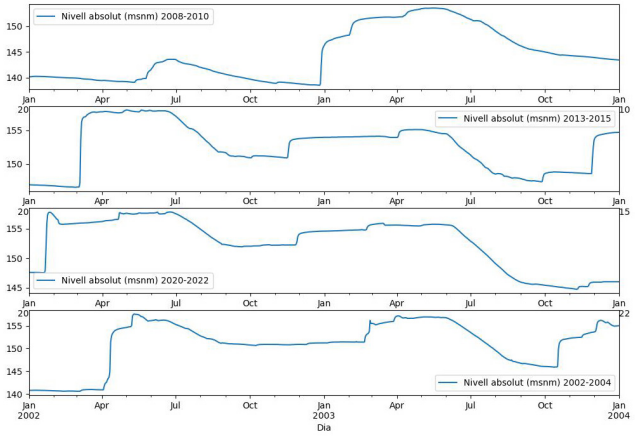
A continuació hem provat de trobar l'estacionalitat del nivell absolut fent gràfiques de diferents periodes d'un any al llarg de la serie històrica.



Com que no hem vist cap estacionalitat en les dades representades, hem graficat un periode de dos anys.

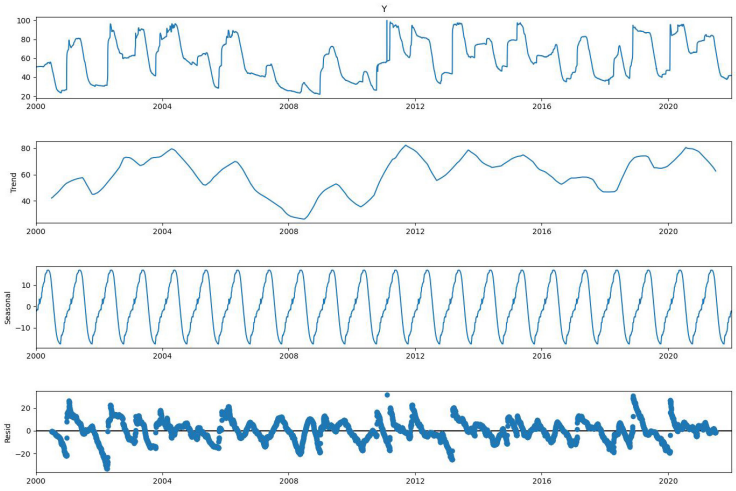


Aquí ja hem pogut percebre una certa estacionalitat, encara que poc marcada, en periode d'uns 8 mesos. Finalment hem graficat un periode de 4 anys només per confirmar si aquesta estacionalitat que hem vist es repeteix any rere any.



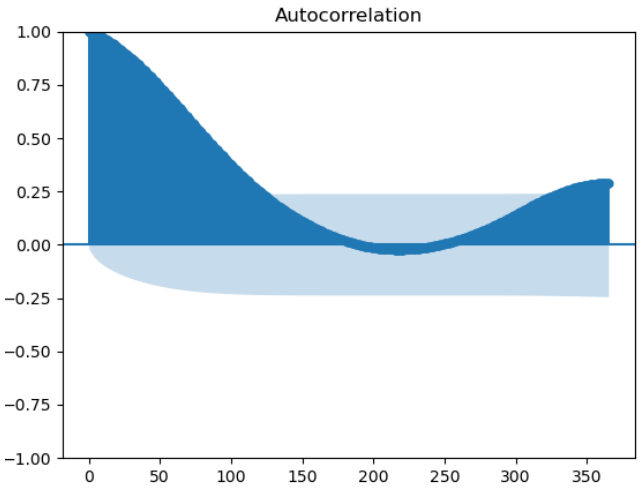
D'aquesta manera hem pogut confirmar que, encara que de manera irregular, hi ha una certa estacionalitat, que es repeteix de manera periodica, en intervals d'uns 8 mesos aproximadament, encara que no de manera molt marcada.

Per fer una analisi més acurada, hem fet una decomposició estacional de tota la serie per veure si la serie té cap tendencia o estacionalitat.

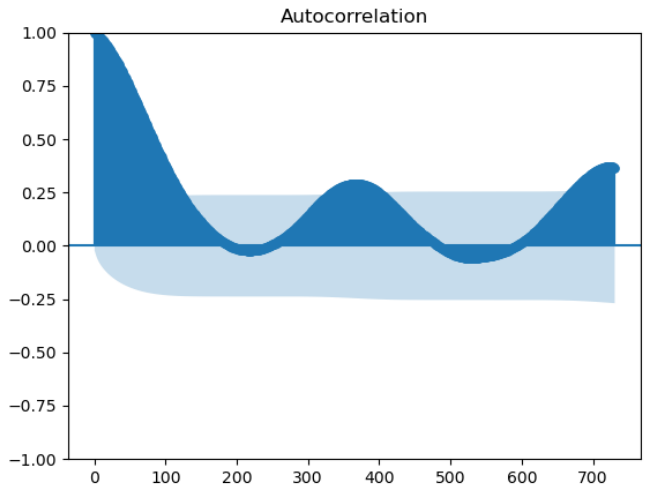


Comprovem que no hi ha cap tendencia, encara que si una estacionalitat gairebé anual, com ja haviem vist en les gràfiques anteriors.

Per finalitzar l'analisi hem fet una gràfica d'autocorrelació (ACF) amb un periode d'un any, per a veure si hi ha autocorrelació de les dades, i de haver-ne en quins intervals es dona. La ACF és útil precissament per veure l'estacionalitat, les tendencies de la serie i altres patrons.



Com que no es veu una estació complerta, es a dir un interval de dades autocorrelacionades, dintre d'aquest periode de temps, hem fet un altre gràfica amb un periode de dos anys per a intentar copsar aquesta estacionalitat.

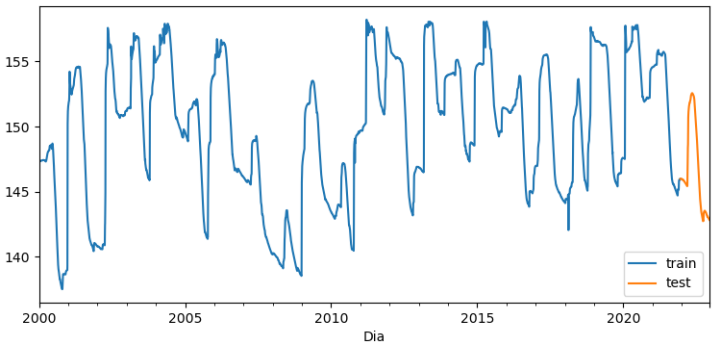


En aquesta ja es veu que la autocorrelació de les dades es dona en cicles de gairebé un any, confirmant el que hem visualitzat en gràfiques previes.

B. Models

En aquest projecte hem fet servir tres models de regressio purs i un d'autoregressió de mitjanes movils que passem a explicar a continuació.

Primer hem separat tot el dataframe en un conjunt d'entrenament i un altre de test, deixant 365 registres per al de test i la resta per l'entrenament.



A continuació hem provat tres models diferents, que passarem a explicar seguidament, primer sense paràmetres, després amb els millors paràmetres que hem buscat amb el mètode grid_search i finalment amb la variable exògena de la precipitació.
Finalment hem provat un model ARIMA trobant els millors paràmetres amb la funció auto_arima i entrenant el model amb ells.

a) Regressius sense paràmetres:

LinearRegression:
Com que aquest és un model purament de regressió, hem fet servir la funció ForecasterAutoreg de la llibreria skforecast per a fer-lo servir com a model autoregressiu.

l'Hem entrenat amb el conjunt d'entrenament sense cap parametre, doncs més endavant buscarem els millors amb el grid search, i posteriorment hem fet una predicció d'un any perque coincideixi amb les dades de test i comparar la predicció envers les dades reals i extreure'n el rendiment.

Hem fet servir com a métriques el Mean Absolute Error (MAE), Mean Squared Error (MSE) i el Coefficient of determination (R2).

Per acabar hem fet una gràfica per veure com encaixen les prediccions en relació a la realitat.

Ridge:

De la mateixa manera que al model anterior, hem fet servir la funció ForecasterAutoreg de la llibreria skforecast per a fer-lo servir com a model autoregressiu.

l'Hem entrenat amb el conjunt d'entrenament, sense cap parametre, i posteriorment hem fet una predicció d'un any perque coincideixi amb les dades de test i comparar la predicció envers les dades reals i extreure'n el rendiment.

Hem fet servir les mateixes métriques: Mean Absolute Error (MAE), Mean Squared Error (MSE) i el Coefficient of determination (R2).

Per acabar hem fet una gràfica per veure com encaixen les prediccions en relació a la realitat.

PassiveAggressiveRegressor:

Igual que als models anterior, hem fet servir la funció ForecasterAutoreg de la llibreria skforecast per a fer-lo servir com a model autoregressiu.

l'Hem entrenat amb el conjunt d'entrenament, sense cap parametre, i posteriorment hem fet una predicció d'un any perque coincideixi amb les dades de test i comparar la predicció envers les dades reals i extreure'n el rendiment.

Hem fet servir les mateixes métriques: Mean Absolute Error (MAE), Mean Squared Error (MSE) i el Coefficient of determination (R2).

Per acabar hem fet una gràfica per veure com encaixen les prediccions en relació a la realitat.

b) Regressius amb els millors paràmetres:

Grid Search: Hem fet una cerca dels millor paràmetres per a cadascun dels 3 models anteriors, mitjançant el mètode grid_search_forecaster de la llibreria skforecast. Un cop fet, hem tornat a entrenat els models i a fer les prediccions, aconseguint un lleuger millor resultat en el model Ridge i PassiveAgresiveRegressor.

c) Regressius amb variable exògena:
Hem tornat a entrenar els tres models amb els millors paràmetres i afegint la variable exògena de la precipitació. El resultat ha sigut que les métriques han restat pràcticament iguals, a excepció del mean absolute error del linear regression que l'ha disminuit lleugerament, mentres que en el cas del model ridge l'ha empitjorat. La resta de les metriques son identiques.

d) ARIMA:
Per aquest model hem fet servir la funcio auto_arima de la llibreria pmdarima per buscar els millor paràmetres.

amb el resultat que ens ha donat l'hem entrenat amb el conjunt d'entrenament i finalment hem fet una predicció d'un any perque coincideixi amb les dades de test i comparar la predicció envers les dades reals i extreure'n el rendiment.

Les metriques per a aquest model son principlament L'AIC.

Per acabar hem fet una gràfica per veure com encaixen les prediccions en relació a la realitat.

Veiem que la predicció, encara després d'haver escollit els millors parametres, no s'ajusta gens a la realitat.

C. Llibreries

Per a tot aquest projecte hem fet servir les següents llibreries:

pandas 1.4.4
matplotlib 3.5.2
seaborn 0.11.2
numpy 1.21.5
scikit-learn 1.0.2
skforecast 0.6.0
statsmodels 0.13.2
pmdarima 2.0.2

IV. Resultats

A. Taules

A continuació presentem cuatre taules amb els resultats de les metriques obtingudes amb els models que hem desenvolupat:

TABLE I
MODELS SENSE PARÀMETRES

Models	*MAE*	*MSE*	*R2*
Linear Regression	1.620418	5.354497	0.737738
Ridge Regression	1.620418	5.624012	0.724537
Passive Agressive	1.620418	5.035074	0.753383

TABLE II
MODELS AMB ELS MILLORS PARÀMETRES

Models	*MAE*	*MSE*	*R2*
Linear Regression	1.620418	5.338832	0.738505
Ridge Regression	1.620418	5.586968	0.726351
Passive Agressive	1.482818	4.820144	0.763910

TABLE III
MODELS AMB ELS MILLORS PARÀMETRES I VARIABLE EXÒGENA

Models	*MAE*	*MSE*	*R2*
Linear Regression	1.615322	5.338832	0.738505
Ridge Regression	1.670257	5.586968	0.726351
Passive Agressive	1.482818	4.820144	0.763910

TABLE IV
MODEL ARIMA

Model	*MSE*	*RMSE*	*R2*
1,1,7	5.354497	5.354497	0.737738

B. Gràfiques

Seguidament mostrem les gràfiques dels 4 models que hem desenvolupat amb la predicció d'un any en verd superposada a les dades reals:

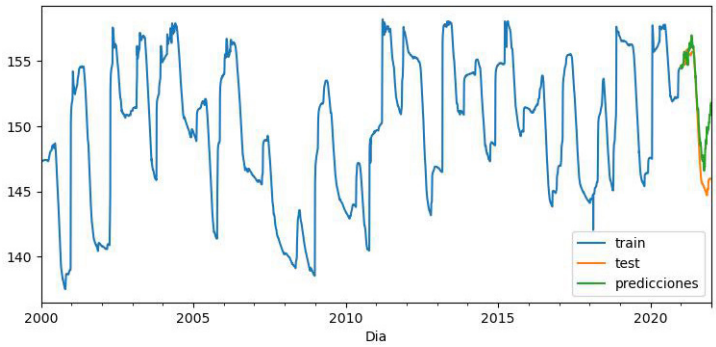


Fig. 1. Linear Regression model.

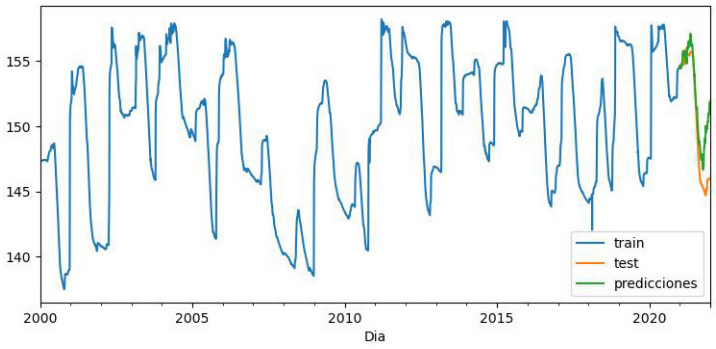


Fig. 2. Ridge model.

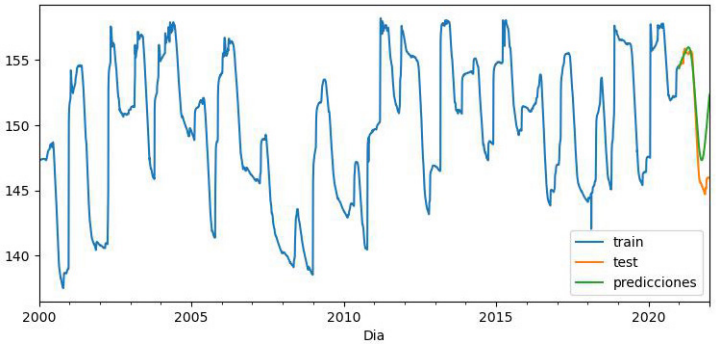


Fig. 3. PassiveAgressiveRegressor model.

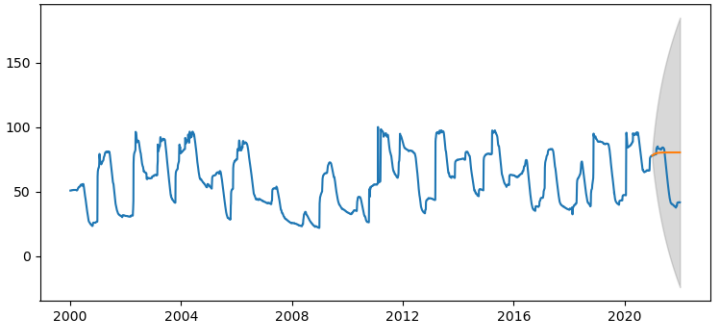


Fig. 4. ARIMA model.

V. Conclusió

De tot aquest projecte podem extreure com a conclusió que en comptes de recollir les dades de l'estació meteorológica situada al mateix embassament hauria d'haver recollit les dades de precipitació de la capçalera del riu Muga, que es el que desemboca a l'embassament de DArnius - Boadella. Doncs serà la pluja recollida en aquesta capçalera la que finalment acabarà omplint l'embassament, i no la precipitació que cau a sobre del mateix pantà. Una vegada em vaig adonar de la gairebé nula correlació entre la precipitació in situ i el nivell d'aigua de l'embassament vaig buscar estacions meteorológiques al llarg de la capçalera de la Muga, però no n'hi ha cap. De fet l'unica estació meteorológica entre la capçalera del riu Muga i l'embassament de Darnius, és la que vaig escollir jo, ubicada al mateix embassament, així doncs no hi ha manera de fer un model predictiu del nivell d'aigua d'aquest embassament fent servir la precipitació com a variable exogena.

Inclús seleccionant un altre embassament de les conques internes tampoc seria possible doncs la majoria dels embassaments es nodreixen de rius que neixen al pirineu, on no hi han gairebé estacions automàtiques, només 7, i cap coincideix amb la capçalera de cap riu.

References

[1] Agència Catalana de l'Aigua, "Aigua i canvi climàtic"

[2] Agència Catalana de l'Aigua, "Pla especial d'actuació en situació d'alerta i eventual sequera"