Treball de Fi de Grau

GRAU D'ENGINYERIA INFORMÀTICA

Facultat de Matemàtiques i Informàtica

Universitat de Barcelona

Disseny d'un model basat en tècniques d'aprenentatge automàtic per predir el cabal del riu Ter

---

Autor: **Sergi Ger Roca**

Director: **Dr. Jerónimo Hernández González**

Realitzat a: **Departament de Matemàtiques i Informàtica**

Barcelona, 13 de juny de 2022

Agraïments

Vull agrair, en primer lloc, al meu tutor del treball, el Dr. Jerónimo Hernández González, des del primer moment s'ha bolcat en ajudar-me i guiar-me en tot el que he necessitat. El seu suport i ànims que he pogut sentir durant tota la realització del projecte, només tinc bones paraules per a ell.

Tanmateix, també m'agradaria agrair a la meva família i amics, pel seu amor i suport incondicional i continu.

També vull mencionar el meu agraïment al Juan Jose Villegas, responsable de les dades de les estacions de l'Agència Catalana de l'Aigua. A les diferents administracions públiques per apostar per *OpenData* i cedir les dades per la realització d'aquest projecte.

Vull agrair també al meu amic Albert Vich Gleyal, per la seva ajuda que no va dubtar en aportar quan ell estava també ocupat.

Finalment, a totes aquelles persones que d'una manera o altra m'han ajudat, encara que el seu nom no figuri explícitament en aquestes línies.

Resum

És una obvietat que un dels grans problemes d'aquest segle XXI és el canvi climàtic, les inundacions, sequeres i els temporals violents seran cada cop més habituals. A més a Catalunya, el clima mediterrani és un dels punts més perjudicats d'aquesta crisi global.

Això i la constant i creixent necessitat d'enviar aigua a l'àrea metropolitana de Barcelona a causa del gran consum, fan que la Generalitat hagi de gestionar i regular els cabals de diferents rius de Catalunya.

En el nostre cas ens centrem en el riu Ter, un riu que neix a Ulldeter (Pirineus) i desemboca a l'Estartit (Mediterrani), i que està constantment en el punt de mira per diverses causes. Entre d'altres, riuades i inundacions han acompanyat la història recent del riu. Tot i això, tenim constància que el riu experimenta una gestió per part de les administracions públiques molt meticulosa.

Inspirats pel moviment Data4Good, que agrupa projectes que a partir de dades obertes volen resoldre problemes reals, amb el benefici comú i compromís social com a objectiu final. En aquest treball aportem una anàlisi del riu i una metodologia basada en el *Machine Learning* per predir el seu cabal a 24 hores vista.

Durant aquest projecte s'analitzen en detall i es processen dades obertes de l'Agència Catalana de l'Aigua i del Servei Meteorològic Català referents a la conca del riu Ter. Aquestes dades es creuen i s'utilitzen per entrenar diferents models d'aprenentatge automàtic.

Aquests models intenten predir el cabal del riu Ter en alguns punts estratègics del riu. Estudiant-los extraiem la relació que té una estació meteorològica sobre l'impacte al cabal, localitzant les zones més influents. Les precipitacions tenen un paper directe amb els cabals dels rius, per això és tan important poder utilitzar dades meteorològiques en els models.

Tot això ens permet extreure nova informació valuosa sobre el comportament del riu, que ens ajuda a entendre'l millor per poder fer una gestió més eficient.

Aquest treball posa la primera pedra en l'ús de tècniques d'aprenentatge automàtic com a suport a la gestió del riu Ter. També apunta línies de treball d'interès per continuar millorant aquest tipus de model i exportar-ho a altres rius.

Resumen

Es una obviedad que uno de los grandes problemas de este siglo XXI es el cambio climático, las inundaciones, sequías y temporales violentos serán cada vez más habituales. Además en Cataluña, el clima mediterráneo es uno de los puntos más perjudicados de esta crisis global.

Esto y la constante y creciente necesidad de enviar agua al área metropolitana de Barcelona debido al gran consumo, hacen que la Generalitat tenga que gestionar y regular los caudales de distintos ríos de Catalunya.

En nuestro caso nos centramos en el río Ter, un río que nace en Ulldeter (Pirineos) y desemboca en el Estartit (Mediterráneo), y que está constantemente en el punto de mira por diversas causas. Entre otras, riadas e inundaciones han acompañado a la historia reciente del río. Sin embargo, tenemos constancia de que el río experimenta una gestión por parte de las administraciones públicas muy meticulosa.

Inspirados por el movimiento Data4Good, que agrupa proyectos que a partir de datos abiertos quieren resolver problemas reales, con el beneficio común y compromiso social como objetivo final. En este trabajo aportamos un análisis del río y una metodología basada en el Machine Learning para predecir su caudal a 24 horas vista.

Durante este proyecto se analizan en detalle y se procesan datos abiertos de la Agencia Catalana del Agua y del Servei Meteorològic Català referentes a la cuenca hidrográfica del río Ter. Estos datos se cruzan y se utilizan para entrenar diferentes modelos de aprendizaje automático.

Estos modelos intentan predecir el caudal del río Ter en algunos puntos estratégicos del río. Estudiándolos extraemos la relación que tiene una estación meteorológica sobre el impacto en el caudal, localizando las zonas más influyentes. Las precipitaciones tienen un papel directo con los caudales de los ríos, por eso es tan importante poder utilizar datos meteorológicos en los modelos.

Todo esto nos permite extraer nueva información valiosa sobre el comportamiento del río, que nos ayuda a entenderlo mejor para poder realizar una gestión más eficiente.

Este trabajo pone la primera piedra en el uso de técnicas de aprendizaje automático como soporte a la gestión del río Ter. También apunta líneas de trabajo de interés para seguir mejorando este tipo de modelo y poder exportarlo a otros ríos.

Abstract

It is obvious that one of the biggest problems of this 21st century is climate change, with floods, droughts and violent storms becoming more common. In addition, in Catalonia, the Mediterranean climate is one of the most affected areas of this global crisis.

This and the constant and growing need to send water to the metropolitan area of Barcelona due to high consumption, mean that the Generalitat has to manage and regulate the flows of different rivers in Catalonia.

In our case, we focus on the Ter river, a river that is born in Ulldeter (Pyrenees) and flows into Estartit (Mediterranean), and is constantly in the spotlight for various reasons. Among other things, floods and floods have accompanied the recent history of the river. However, we are aware that the river is undergoing very meticulous management by public administrations.

Inspired by the Data4Good movement, which brings together projects that want to solve real problems from open data, with the common benefit and social commitment as the ultimate goal. In this work we provide an analysis of the river and a methodology based on Machine Learning to predict its flow in 24 hours.

During this project, open data from the Catalan Water Agency and the Catalan Meteorological Service regarding the Ter river basin are analyzed in detail and processed. This data is cross-referenced and used to train different machine learning models.

These models attempt to predict the flow of the river Ter at some strategic points in the river. By studying them we extract the relationship that a weather station has on the impact on the flow, locating the most influential areas. Precipitation plays a direct role with river flows, which is why it is so important to be able to use meteorological data in models.

All this allows us to extract valuable new information about the behavior of the river, which helps us to understand it better, so that, we can manage it more efficiently.

This work lays the foundation stone for the use of machine learning techniques to support the management of the river Ter. It also points out lines of work of interest to continue to improve this type of model and export it to other rivers.

Index

0. Estructura de la memòria 7

1. Introducció 8
- 1.1 El riu 8
- 1.2 Problemàtica 9
  - 1.2.1 La meva solució 10
- 1.3 Motivació 11
- 1.4 Objectius 12
- 1.5 Les dades i estacions 14
- 1.5.1 ACA 15
- 1.5.2 SMC 16

2. Planificació 17
- 2.1 Diagrama de Gantt 18

3. Anàlisis de Dades 20
- 3.1 SMC (Servei Meteorològic Català) 20
  - 3.1.1 Primer análisis de les estacions 21
    - 3.1.1.1 Estacions amb dades desde el 2009 al 2022 22
    - 3.1.1.2 Estacions a completar 24
  - 3.1.2 Generar el DataFrame 25
    - 3.1.2.1 Estacions amb dades desde el 2009 al 2022 25
      - 3.1.2.1.1 Estacions amb diferents freqüències horàries 25
      - 3.1.2.1.2 Estacions KE i M6 26
    - 3.1.2.2 Estacions a completar 28
      - 3.1.2.2.1 DM - WF - XJ Girona 28
      - 3.1.2.2.2 Z4 - ZC Ulldeter 29
      - 3.1.2.2.3 DataFrame a completar final 30
    - 3.1.2.3 DataFrame final 30
- 3.2 ACA (Agència Catalana de l'Aigua) 30
  - 3.2.1 Primer análisis de les estacions 31
  - 3.2.2 Generar el dataFrame de cabal 32
    - 3.2.2.1 Unió de fitxers per cada estació 32
    - 3.2.2.2 Analisis de les dades ajuntades 32
  - 3.2.3 Generar el dataFrame de nivell 33
  - 3.2.4 Tractament de les dades 33
    - 3.2.4.1 Interpolació 34
    - 3.2.4.2 Relacionar el nivell amb el cabal 34
      - 3.2.4.2.1 Valors incomprensibles 35
      - 3.2.4.2.2 S'han fet obres al riu? 35
      - 3.2.4.2.3 KNN Regressor 36
    - 3.2.4.3 Moving average 37
    - 3.2.4.4 Iteracions finals 37

3.2.4.4.1 Primera part 37
3.2.4.4.2 Segona part (Iterative Imputer) 38

4. Disseny dels models 40
4.1 DataFrames 40
4.1.1 Part alta 41
4.1.2 Part baixa 41
4.2 Procediment de validació dels models 42
4.3 Tipus de models utilitzats 43
4.3.1 Random Forest 43
4.3.2 XGBoost 44
4.3.3 Extra Trees 44
4.3.4 Regressió Lineal 44

5. Proves, Avaluació i Resultat 46
5.1 Part alta 46
5.1.1 Random Forest 46
5.1.2 XGBoost 48
5.1.3 Extra Trees 49
5.1.4 Regressió Lineal 50
5.2 Part baixa 50
5.2.1 Random Forest 51
5.2.2 XGBoost 52
5.2.3 Extra Trees 53
5.2.4 Regressió Lineal 54
5.3 Valor RMSE 54

6. Conclusions i treballs futurs 55

7. Referències 58

8. Annexes 60
8.1 Taules 60
8.2 Imatges 62

0. Estructura de la memòria

Aquest treball s'ha dividit en 8 apartats:

1. Introducció: S'introdueix la temàtica i problemàtica que he tractat en aquest treball. Es descriu la motivació i es planteja el que es vol aconseguir amb el treball, els objectius i les dades que utilitzaré.

2. Planificació: S'explica com em vaig planificar aquest projecte en un inici, les temporitzacions de les tasques i es compara com ha acabat sent.

3. Anàlisi de Dades: Es fa el processament de les dades rebudes per part de l'Agència Catalana de l'Aigua i del Servei Meteorològic Català, es descriuen també les tècniques utilitzades per fer-ho possible.

4. Disseny dels models: Es comenta com s'han creat els diferents conjunts de dades que li passaré als models i el procediment que he utilitzat perquè el model retorni dades vàlides.

5. Proves, avaluació i resultat: S'expliquen les diferents proves que s'han fet i quins models s'han analitzat. Es fa una anàlisi i valoració dels resultats obtinguts al realitzar les proves.

6. Conclusions i treballs futurs: Finalització i tancament del treball, enumeració de les diferents conclusions a les que he arribat i descripció de millores o funcionalitats extretes possibles a implementar en un futur.

7. Referències: Enumeració de les diferents fonts que he utilitzat per fer aquest treball.

8. Annexes: Diferent informació extra sobre el desenvolupament del projecte, entre d'altres, taules, imatges, etc.

1. Introducció

En aquest TFG es vol crear diferents models del riu Ter utilitzant l'aprenentatge automàtic i dades referents al riu i la meteorologia. Amb aquests models es pretén predir el cabal en diferents punts del riu per anticipar-se a riuades i inundacions, també per dotar de noves eines i informació a les administracions encarregades de la gestió. En aquest primer apartat es fa una introducció al riu, la seva problemàtica, els objectius, motivacions i les dades utilitzades.

1.1 El riu

El riu Ter neix a Ulldeter a la comarca del Ripollès a uns 2.400 metres d'altitud, al peu d'un antic circ glacial. Transcorre per les comarques del Ripollès, la Selva, Osona, el Gironès i el Baix Empordà on desamboca al mar Mediterrani (Platja de Pals, L'estartit). Amb una longitud de 208 km i una superfície de conca de 3.010 Km2, és, juntament amb el Llobregat, el riu de més recorregut de la xarxa hidrogràfica Pirineus-Mediterrània.

El seu recorregut segueix dos trajectories ben diferents: Nord-Sud (desde Ulldeter a la plana de Vic) i Oest-Est (Plana de Vic fins la desembocadura). La producció hidroelèctrica i l'ús industrial són els principals usos del riu. Actualment hi ha 119 concessions al llarg del Ter i del Freser (Afluent del mateix d'origen pirinenc), on les de Sau, Susqueda i el Pasteral en son les més grans.

També té una gran influència sobre l'agricultura i aporta aigua a altres zones del territori (àrea de BCN). Durant tot el seu transcurs, el Ter pateix una densitat considerable de derivacions de cabal per fins explicats anteriorment.

Després dels tres grans embassaments —Sau (amb un volum de 151,3 hm^3), Susqueda (amb un volum de 233 hm^3) i el de Pasteral (el més petit amb un volum de 2 hm^3)–, els quals podem dir que estan aproximadament a la meitat del transcur del Ter, el cabal del riu es troba plenament regulat. Això vol dir que aquesta segona part del riu (Des de pasteral fins a la desembocadura) no representa les fluctuacions que tenen els cursos fluvials mediterranis que no estan regulats.

A la conca del riu, a part del pantans citats anteriorment hi ha dos embassaments més: el de Colomers i el de la Seva (al riu Gurri).

També, el Ter té una gran quantitat d'afluents durant tota la seva trajectòria, dels quals hi ha rieres i rius més i menys importants: Ritort, Freser, Rigard, riu de Núria, Riera de Vallfogona,

Ges, riu Fornés, Riera de Sorreig, Gurri, Riera de les Gorgues, Riera Major, Riera de Rupit, Riera de l'Om, Brugent, Riera d'Osor, Sot de la Noguerola, Riera de Llémena, Güell, Onyar, Terri, Daró, Riera de Carboners, El Cinyana, Torrent del Gàrrep, Riera de Talamanca, Torrent de Vall-Llobre.

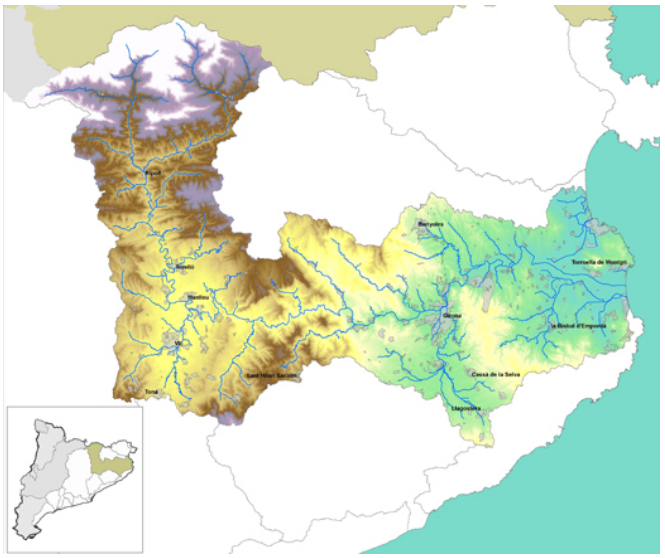

Figura 1: Conca del riu Ter (oriol, https://www.thinglink.com/scene/1185849484622430211,20109)

1.2 Problemàtica

Les problemàtiques i la gestió del riu per part de les administracions públiques han anat força lligades aquest últims 80 anys. Per entendre la que s'aborda en aquest treball, s'ha de tenir una visió amplia de les diferents problemàtiques que ha i està patint el riu.

"La batalla del agua" titulava en portada Los Sitios el 10 d'agost del 1957, on es parlava sobre la reunió que s'havia celebrat al govern civil de Girona per parlar sobre el transvasament d'aigües del riu a Barcelona. Tema que va portar molta controvèrsia (i que en continua portant) però que finalment es va inaugurar el 2 de juliol del 1966.

Les aportacions del Ter a Barcelona estan suposant que un gran tant per cent de l'aigua a l'any vagi cap a l'àrea metropolitana, i com a conseqüència, la reducció del mateix cabal. Des que es va iniciar el transvasament, aquest valor ha sigut molt discutit i en moments molt crític. Com es recull en reportatges i informes com: *La manca de cabal al riu Ter*[1] o la Revista *La llera del ter*[2] entre d'altres.

[1] "LA MANCA DE CABAL AL RIU TER - Museu de la Mediterrània." https://www.museudelamediterrania.cat/pujades/files/La-manca-de-cabal-al-riu-Ter.pdf.
[2] "EL PROBLEMA DE L'AIGUA I EL RIU TER EL ... - Ajuntament de Celrà." https://www.celra.cat/ajuntament/documents/lallera47.pdf.

Per altre banda, els i les habitants de la conca històricament han sigut víctimes de grans aiguats i inundacions. Una de les més rellevant, la de la nit del 17 al 18 d'octubre del 1940 ón el riu es va cobrar només a Torelló: 61 morts, 52 hospitalitzats i 250 edificis van quedar destruïts. Recupero la informació del reportatge del diari NacióDigital: *El gran aiguat del Ter, 75 anys després*[3] i de l'arxiu: *Riuades del riu Ter i Ritort a camprodon*[4] de l'Agència Catalana de l'Aigua on fa un anàlisis de les afectacions que va patir en aquest cas la localitat de Camprodon. No és d'estranyar doncs que una de les rutes temàtiques que es proposen al Ter sigui anomenada: Ruta per les inundacions històriques. Una ruta que permet observar mitjançant de plaques, testimonis i empremtes, les inundacions històriques.

Però no cal traslladar-nos al 1940 per veure que hi han hagut problemes amb les riuades recentment. En el seu pas, el temporal Glòria ens deixava portades com: "El Ter inunda diferents zones de Girona, El riu supera la seva llera després d'una nit de confinament" [5] al diari El País o "El agua del Ter inunda Girona por el temporal Gloria" [6] al diari el Triangle, entre d'altes.

En aquest punt ens preguntem si s'està regulant adequadament el riu Ter? Partim de la idea que el riu Ter són com dos rius: el curs Alt (desde Ulldeter fins al primer embassament, Sau), on la majoria de l'aigua que cau baixa pel riu i no pot ser regulada i el curs mig-baix, (a partir de Pasteral) on està regulat i l'aigua que porta és la que es decideix als pantans més la dels seus afluents a partir d'aquell punt.

Tant excepcional va ser aquell temporal? Quina utilitat tindria una eina predictiva que pugés ajudar a l'Agència Catalana de l'Aigua (empresa responsable de la gestió) a anticipar-se en aquests episodis? Els hi seria útil? A partir d'aquí hom es pot fer moltíssimes preguntes i possibles respostes a aquestes preguntes.

Segurament l'Agència Catalana de l'Aigua ja fa les seves previsions internes, però no sabem com ho fan.

1.2.1 La meva solució

Després de tot això es va creure que podria haver una mancança d'un model que pugés predir el cabal en alguns punts crítics del riu. Un model que pugés creuar dades *OpenData*

---

[3] "FOTOS El gran aiguat del Ter, 75 anys després | Osona.com." 18 d'oct.. 2015, https://www.naciodigital.cat/osona/noticia/48095/fotos-gran-aiguat-ter-75-anys-despres.
[4] "RIUADES DEL RIU TER I RITORT A CAMPRODON - ACA." 20 de des.. 2011, http://aca-web.gencat.cat/sig/fitxes/espais_fluvials/mat/aca_mat_17039_1940b_v1.pdf.
[5] "El Ter inunda diferents zones de Girona | Catalunya - Elpais.cat." 23 de gen.. 2020, https://cat.elpais.com/cat/2020/01/23/catalunya/1579810582_789729.html.
[6] "El agua del Ter inunda Girona por el temporal Gloria - El triangle." 23 de gen.. 2020, https://www.eltriangle.eu/es/2020/01/23/noticia-es-104787/.

no només del propi riu (Agència Catalana de l'Aigua) com el cabal en diferents punts, sinó també dades meteorològiques del moment. Unes dades meteorològiques (Meteocat) que van molt relacionades amb el comportament directe del riu i que juntament amb dades de cabal poden aportar uns resultats molt interessants.

És aquí on apareix la problemàtica que s'intentarà abordar i/o aportar d'alguna manera més informació, per tenir més mecanismes a la hora de gestionar futurs episodis de crescudes i/o inundacions.

Hi ha constància que una solució semblant ja s'ha implementat en altres rius, per exemple en el Ebre on es va fer una Aguathon (Hackathon d'Aigua). Es va proposar crear un model que predís el cabal a 24h, 48h i 72h. Aquest és l'enllaç a la competició: https://www.itainnova.es/blog/eventos/i-hackathon-del-agua-aguathon

La solució que es presenta en aquest projecte està inspirada i comparteix molts dels ideals dels moviment *Data for Social Good / Data Science for Good.* Aquests moviments defensen fer servir *data science* utilitzant *OpenData* amb l'objectiu de contribuir a millores per a la societat.

1.3 Motivació

Una de les motivacions principals que he tingut al elegir i realitzar aquest TFG ha sigut el fet de poder relacionar dos grans mons que realment m'apassionen molt. Per un cantó, el que m'ha acompanyat desde el 2018 quan vaig passar a ser oficialment estudiant de la Universitat de Barcelona. Un món que si el destí ho vol, m'acompanyarà la resta de la vida: l'Enginyeria Informàtica. Per altre banda el de la natura, des de ben petit he sigut membre d'un agrupament escolta a la meva ciutat, on sempre m'han ensenyat a estimar i a cuidar el meu entorn, com a conseqüència les muntanyes i els rius. M'és també molt gratificant, poder veure amb les meves pròpies mans com es poden unir aquests dos grans mons per una causa d'aquesta importància.

Aprendre a fer *Machine Learning* ha sigut una altre de les motivacions que he tingut. El no haver agafat les assignatures corresponents per estudiar les diferents tècniques, s'ha traduït en haver-ho d'aprendre per la realització d'aquest TFG.

Una altra de les motivacions que m'han portat a fer aquest TFG és el fet d'aportar el meu gra de sorra a la problemàtica que envolta el riu Ter. Contribuir en el moviment *Data4Good* i poder dedicar el temps que et proporciona el treball per intentar resoldre o almenys aportar

nova informació. Una nova informació que ens ajudi a anticipar-se al cabal del riu per poder prevenir i actuar acord les diferents situacions que es proposen.

La tecnologia ha avançat molt en els últims anys i està en continu desenvolupament, crec que tenim les eines per poder ajudar d'alguna manera. Les inundacions en un riu regulat tal com el Ter haurien de ser història, o almenys poder avisar als habitants i a les autoritats amb un cert marge d'actuació.

Referent a l'apartat anterior, tenim constància de la viabilitat que té un model d'aquestes característiques. El 2019 es va fer la primera competició d'Intel·ligència Artificial i Big Data que es deia Aguathon organitzada pel Instituto Tecnológico de Aragón. Van proposar modelitzar el comportament del nivell del riu Ebre al seu pas per la ciutat de Saragossa per obtenir prediccions realistes de la seva variació en cada instant. El guanyador d'aquest model va aconseguir un error del 0,15 m en les prediccions de 24h, 0,37m a 48h i 0,47m a 72h uns errors baixos a poques hores i acceptables en tot cas.

La mancança d'un model com aquest a les nostres terres, en aquest cas en un riu com el Ter, ha sigut una altre motivació. A més, la possibilitat de poder creuar dades de cabal del riu amb dades meteorològiques de la zona feia la possibilitat de poder aportar noves variables significatives al model. Unes variables que la gent de l'Aguathon no va tenir i que obre les portes a una possible millora dels models.

L'ús dels model predictors i de la IA en si, està creixent exponencialment en les nostres vides. Cada cop més les empreses i les organitzacions públiques obren les seves dades i/o aposten per aquesta tecnologia ja que donen molt bons resultats. L'OpenData és molt beneficiós per tothom, molts cops les empreses i administracions no poden fer aquest tipus de projecte perquè: o no en saben, no tenen les idees, no tenen els recursos i/o el temps necessari. Això ajuda als investigadors/es a desenvolupar projectes basats amb la filosofia del *Data4Good,* amb un l'objectiu del benefici social.

Perquè no implementar-ho nosaltres també? No es una tecnologia en proves, es consolidada desde fa ja molts anys.

1.4 Objectius

Els objectius d'aquest treball es poden dividir en dos seccions: els objectius generals i els específics. Els objectius generals corresponen a finalitats mes genèriques del projecte, sense tenir un resultat en concret. És a dir, no son quantificables directament ni mesurables. Els específics es refereixen més al que es vol aconseguir concretament i son mesurables.

Com a objectius generals, em vull centrar en investigar i aprendre el màxim possible en tots els àmbits. Desde peculiaritats del riu, estacions, la zona fins la forma d'analitzar/tractar les dades i com funcionen els models de *Machine Learning*.

- Aprendre i consolidar els conceptes bàsics de machine learning per dur a terme el projecte.

Pel que fa referència als objectius més específics:

- Abordar la problemàtica de l'obtenció i el tractament de les dades que necessito per crear el model. Sempre es difícil netejar les dades que les administracions et brinden, ja que acaben passant el que tenen (no sempre tenen el que vols), t'ho passen en el seu format, en diferents formats, amb molts errors, dades no validades, etc …

- Utilitzar diferents tècniques conegudes i algunes que he enginyat que explicaré per realitzar el preprocessament de dades. Aquest tram és molt important ja que és la part més costosa al construir un model de ML i on segurament hi dedicaré més hores. L'última versió de les dades hauran de tenir unes certes qualitats per donar-les com a bones.

- Crear diferents models del riu: si s'escau per la part alta i baixa del riu per separat o un del riu total. En la creació d'aquests models utilitzaré les dades que hauré netejat i preprocessat anteriorment. Poder modelar el riu i obtenir prediccions a x hores en uns punts específics del riu amb un error comparable al que tenia el guanyador de l'Aguathon.

- Aportar/extreure nova informació rellevant sobre el perquè del comportament del riu i quines variables fan que es comporti d'aquesta manera. Tot el que sigui aportar nova informació que beneficiï a la resolució de la problemàtica.

1.5 Les dades i estacions

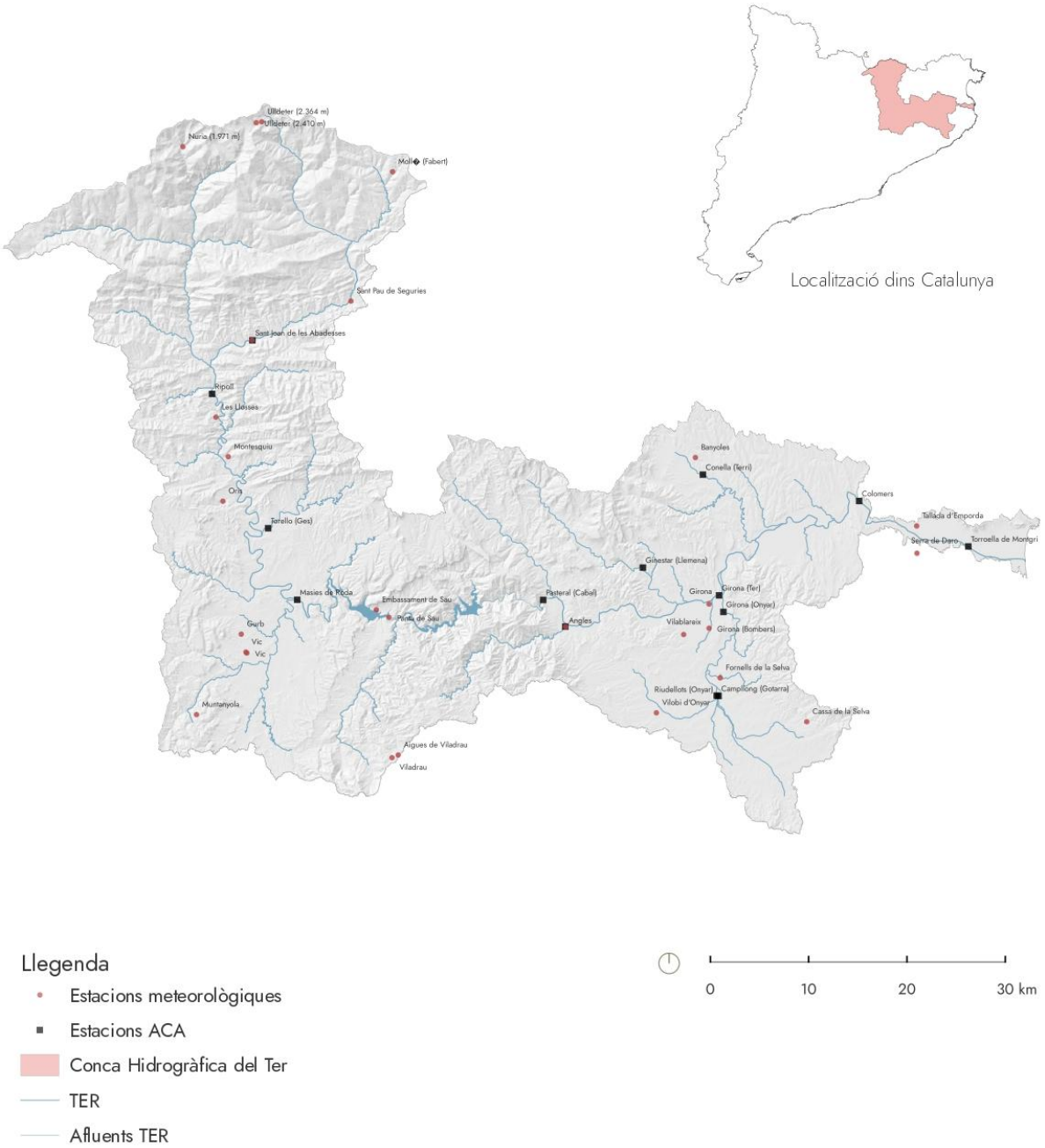

Figura 2: Conca del riu Ter amb les estacions de dades (Albert Vich Gleyal, Estudiant de paisatgisme)

A la Figura 2 podem observar un mapa de la conca hidrogràfica del Ter, amb les diferents estacions d'on he extret les dades per crear el model. En el mateix mapa es poden veure diferenciades les estacions que pertanyen a l'agència catalana de l'aigua i al servei meteorològic català, unes amb un quadrat negre i les altres amb una rodona vermella.

Aquestes no corresponen 100% a les que he utilitzat finalment ja que algunes les he eliminat o tractat.

He rebut i tractat dades d'un total de 41 sèries temporals d'estacions diferents: 13 pertinents a l'Agència Catalana de l'Aigua i 28 al Servei Meteorològic Català. Per fer-se una idea del volum de dades, només l'arxiu que correspon a la estació d'Orís de SMC, té unes 460.945 entrades diferents. Comprenen dades des del 15/11/1995 fins a l'actualitat i amb una freqüència de cada 30'. Per veure un altre exemple, l'arxiu que correspon a l'estació Girona_Ter de l'ACA, té 1.120.757 entrades diferents. Les dades d'aquesta estació són del 2009/01/01 fins 2021/01/01 i amb una freqüència de 5' cada data nova.

Fent càlculs aproximats, la magnitud del volum de dades rebut ha sigut d'unes 24.000.000 líneas de dades en total.

En unes dades del tipus *time series*[7] o sèries temporals, cada valor mesurat li correspon un determinat moment temporal. En aquestes sèries, els valors mesurats estan ordenats cronològicament i, normalment, estan espaiats entre si de manera uniforme (freqüència). Si parlem de les sèries temporals de les diferents estacions que he rebut, moltes de les estacions tenen diferents freqüències i les dates d'inici i final són diferents també.

1.5.1 ACA

L'Agència Catalana de l'Aigua és l'entitat pública que té les competències plenes en el cicle integral de l'aigua en les Conques Internes de Catalunya al marc de la directiva europea de l'aigua.

Les dades de l'ACA les vaig obtenir via la seva bústia de contacte[8] i gràcies a la comunicació amb el Juan José Villegas (responsable de les estacions). A la seva web hi ha dades disponibles però a una freqüència de 24 h, nosaltres volem a 30'.

(Taula 1 als annexes)

Les dades extretes per fer aquesta taula 1 les he obtingut fent recerca per les *aplicacions web interactives*[9] que té la pròpia ACA a la seva pàgina web. En el càs de l'ACA no em van passar una descripció de les dades que em van passar ni la relació entre elles, la majoria d'estacions estaven partides en dos arxius on no totes les variables coincidien.

---

[7] "Sèrie temporal - Viquipèdia, l'enciclopèdia lliure." https://ca.wikipedia.org/wiki/S%C3%A8rie_temporal.
[8] "Contacte - Agència Catalana de l'Aigua - GenCat." https://aca.gencat.cat/ca/contacte/.
[9] "Aplicacions interactives - Agència Catalana de l'Aigua - GenCat." https://aca.gencat.cat/ca/laigua/consulta-de-dades/aplicacions-interactives/.

En l'apartat d'anàlisis aprofundiré més en com he tractat aquestes i explicaré les decisions que he pres cada cop que m'anava trobant en els diferents problemes.

1.5.2 SMC

El Servei Meteorològic de Catalunya és l'empresa pública encarregada de gestionar els sistemes d'observació i predicció meteorològics a Catalunya.

Les dades del SMC les vaig obtenir via el portal de *petició de dades*[10] que tenen actiu a la seva web. També es poden extreure del portar de *dades obertes*[11] de la generalitat de Catalunya, allà estan complertes però s'han de descarregar estació per estació i es una mica lent.

(Taula 2 als annexes)

Les dades extretes per fer aquesta taula 2 les he obtingut del portal de dades obertes de la Generalitat, concretament la taula: *Metadades estacions meteorològiques automàtiques*[12].

Com es pot observar, algunes estacions ja no estan operatives i algunes es van inaugurar vora el 2010.

En l'apartat d'anàlisis aprofundiré més en com he tractat aquestes i explicaré les decisions que he pres cada cop que m'anava trobant en els diferents problemes.

[10] "Petició de fitxer amb dades meteorològiques - Servei Meteorològic ...." https://www.meteo.cat/wpweb/serveis/formularis/peticio-dinformes-i-dades-meteorologiques/peticio-de-dades-meteorologiques/.
[11] "Dades obertes - Servei Meteorològic de Catalunya." https://www.meteo.cat/wpweb/serveis/cataleg-de-serveis/serveis-oberts/dades-obertes/. S'hi ha accedit el dia 31 de maig. 2022.
[12] "Metadades d'estacions meteorològiques automàtiques." https://analisi.transparenciacatalunya.cat/Medi-Ambient/Metadades-estacions-meteorol-giques-autom-tiques/yqwd-vj5e. S'hi ha accedit el dia 31 de maig. 2022.

2. Planificació

En aquest apartat explicaré com m'he organitzat al llarg d'aquest semestre per tal d'anar realitzant el projecte, indicant quines són totes les parts i el temps que l'hi vull dedicar i hi he dedicat a cadascuna. Aquest projecte l'he organitzat en diferents fases, on cadascuna l'he organitzat amb una sèrie de tasques que s'expliquen a continuació.

Les quatre fases en les que es divideix són la Documentació i l'obtenció de dades, el desenvolupament del projecte, la redacció de la memòria i l'entrega del treball.

La primera part es la de la documentació i obtenció de dades, aquestes son les seves tasques:

- Lectura de les diapositives informatives que hi han disponibles al campus virtual per tal de tenir una primera idea de com serà aquest treball i per saber com enfocar-lo i elaborar-lo.

- Documentació i aprenentatge sobre el Machine Learning, amb aquesta recerca s'ha de poder adquirir un ventall de conceptes per tenir una àmplia i crítica visió sobre el tema. Aquesta documentació pot equivaler fer uns cursos de formació.

- Anàlisis de la situació de les diferents estacions necessàries i obtenció de les dades d'aquestes estacions des del 2009 fins al 2022 amb una freqüència de 30'. Aquesta tasca està partida en dos: les dades de l'ACA i les dades del SMC.

La següent fase es la que té en compte el desenvolupament del projecte. Aquesta es divideix en dos altres tasques:

- Anàlisis, triatge i el corresponent tractament dels fitxers de dades per convertir totes les diferents dades rebudes en un dataframe[13] vàlid, per poder-ho passar en un model regresor i tot el que això comporta. Aquesta tasca està partida en tres: el tractament de les dades de SMC, el de les ACA i el de les conjuntes.

- Creació dels diferents models buscant els millors paràmetres per la predicció en punts específics a 12-24h. Comparar diferents models de regressió buscant el que doni millor resultat. Analitzar els resultats obtinguts en els models extraient les variables més importants i extreure'n conclusions.

[13] "Intro to data structures — pandas 1.4.2 documentation." https://pandas.pydata.org/docs/user_guide/dsintro.html.

La penúltima fase és la d'escriure la memòria, fins a aquest punt el TFG ha sigut més com "treball de camp". Aquesta fase és l'encarregada de plasmar al document escrit tot el que he anat fent durant la realització del treball.

Finalment, vaig guardar una última fase per fer una revisió general amb el tutor i fer intercanvi d'opinions i recomanacions.

2.1 Diagrama de Gantt

Un cop definides aquestes tasques principals, he intentat seguir el següent diagrama de Gantt per tal d'organitzar-me durant el semestre. En aquest hi ha tota la feina a fer i el temps que hi havia de dedicar a cadascuna.

Els diagrames de Gantt de les Figures 3 i 4 representen la planificació inicial i la final, respectivament. Les tasques esmentades en aquestes figures es descriuen a la Taula 3.

Descripció	Numero
Documentació TFG	Tasca 1
Obtenció de dades SMC	Tasca 2
Obtenció de dades ACA	Tasca 3
Curs ML	Tasca 4
Tractament de dades SMC	Tasca 5
Tractament de dades ACA	Tasca 6
Tractament de dades SMC i ACA	Tasca 7
Creació dels Models	Tasca 8
Escriure Memòria	Tasca 9
Entregar al Tutor	Tasca 10

Taula 3: Referència de la tasca amb el seu número

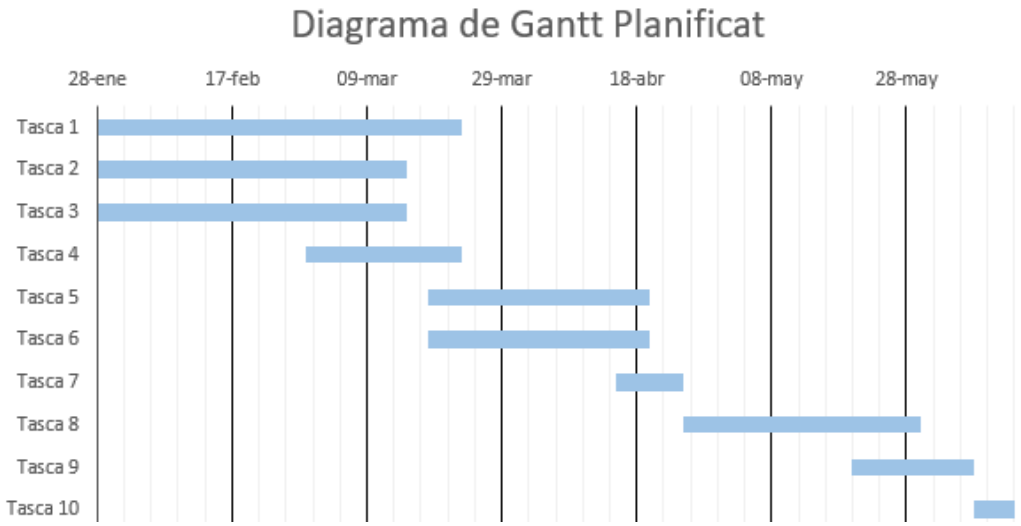

Figura 3: Diagrama de Gantt planificat en l'inici

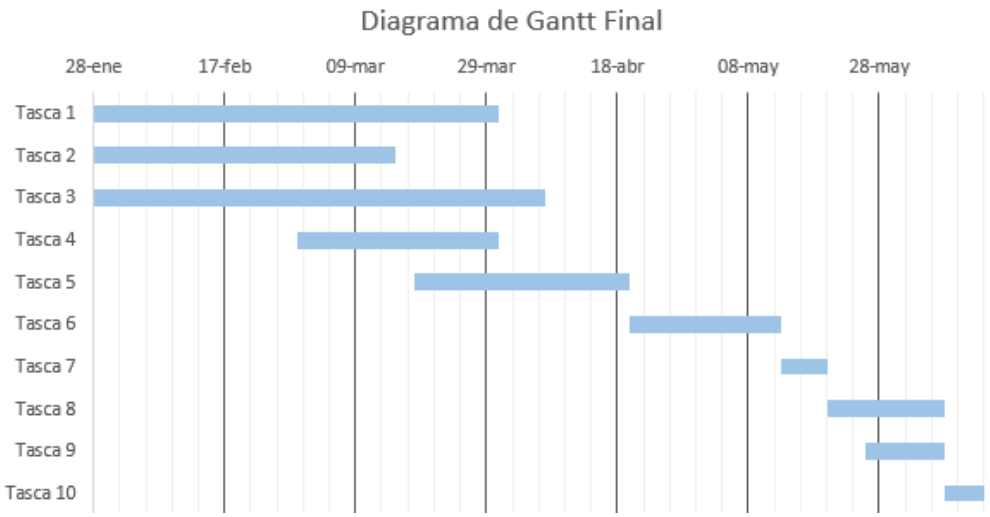

Figura 4: Diagrama de Gantt final

Vaig creure convenient invertir aproximadament el primer mès en la documentació, aprenentatge i obtenció de dades. Va ser una forma d'entrar dins el tema que envolta tot el TFG d'una manera més esglaonada i agafar aquella base conceptual necessària. L'obtenció d'aquestes dades es va allargar una mica més del que havia plantejat, això va fer que la resta de tasques comencessin unes setmanes més tard.

Un cop feta aquesta part, tenia reservat aproximadament un més per fer el tractament de les dades. Aquest plantejament considerava que faria els dos tractaments simultàniament. Però això no ha sigut així, en part a causa del retràs en l'obtenció de les dades de l'ACA, he començat a analitzar primer les del SMC. Sabia que aquesta part del projecte era molt important i costosa però vaig ser una mica més optimista amb el temps el qual he invertit finalment. Puc afirmar que aquesta part m'ha consumit casi la meitat del temps invertit en aquest projecte, s'ha fet una gran feina de processament perquè les dades siguin vàlides per un projecte d'aquest estil.

Seguidament vaig guardar vora un mes per crear els diferents models, fer les proves i extreure conclusions: Tasca 8. La realitat és que com tot ha començat més tard aquesta part també, això m'ha fet retallar coses que hagués volgut fer i que indico a la part de Treballs futurs i conclusions d'aquest document.

Finalment he tingut present un temps per poder escriure la memòria i un període també perquè el meu tutor, el Dr Jerónimo Hernández el pugui revisar.

3. Anàlisis de Dades

En aquest apartat explicaré tot el procés que he seguit i el que he utilitzat per fer l'anàlisis de dades. Ha sigut l'apartat de la planificació que hi he dedicat més temps continuadament. Potser pel fet d'haver sigut la primera vegada que m'enfrentava a un repte d'aquesta magnitud, també acostuma a ser el principal repte en qualsevol aplicació real de l'aprenentatge automàtic.

Com és d'esperar quan es reben les dades s'obtenen en diferents formats com: txt, csv, xlsx, etc.. Cada organisme les té guardades de diferents maneres i tenen diferents potències computacionals, ja que els sistemes de gestió són diferents.

En aquest punt s'han d'agafar aquests documents, pre-seleccionar i extreure les diferents variables que ens interessen. Pot ser que les dades tinguin errors, ja que no estan validades, que hi hagin trossos incomplets i que estiguin en diferents freqüències si parlem de TimeSeries Data. Apart de tot això, podem trobar molts més problemes a les dades que facin més difícil el seu processament. Inicialment les dades que hauré de tractar seran desde el 2009 fins el 2022, vam creure que eren suficients ja que altres models havien funcionat amb aquest volum. A més a més, és el subconjunt en que disposem més dades complertes.

Les dues fonts de dades obertes (*OpenData*) que s'han utilitzat i que s'explicarà el tractament que s'ha fet són: el Servei Meteorològic Català i l'Agència Catalana de l'Aigua.

Donat que les dades es troben en diferent freqüència de mostreig, hem triat unificar-les totes a freqüències d'una mostra cada 30 minuts.

Aquest apartat inclou tot el processament fet des que vaig rebre els diferents fitxers fins que he creat els *dataframes* propis de la part alta i baixa del riu. En tot l'apartat d'anàlisis de dades he utilitzat Jupyter Notebook[14] amb Python 3[15], les versions de cada paquet estan indicades al document *README.txt* del codi.

3.1 SMC (Servei Meteorològic Català)

Pel que fa el SMC em va enviar un email on m'explicava que m'havia de descarregar les dades via WeTransfer, aquest era un fitxer zip on dins hi havia un fitxer excel diferent per cada estació.

[14] "Jupyter Notebook." https://jupyter.org/. S'hi ha accedit el dia 3 de juny. 2022.
[15] "3.10.4 Documentation." https://docs.python.org/. S'hi ha accedit el dia 3 de juny. 2022.

Un cop tenia el zip descarregat, vaig fer una primera inspecció visual obrint un fitxer qualsevol per veure de quina forma estan guardades les dades. Com a ajuda, em van passar unes instruccions de com havia de llegir les dades.:

CC'	DATA	PPT	'SH'
CC	15/11/1995 00:00		SH
CC	15/11/1995 00:30		SH
CC	15/11/1995 01:00		SH
CC	15/11/1995 01:30		SH

Figura 5: Primeres quatre files de dades de l'arxiu que correspon a les dades de l'estació del SMC → CC

L'excel de cada estació el formen 4 columnes i tantes files com registres tinguem d'aquella estació. La primera columna correspon al nom de l'estació, en aquest cas és l'estació CC que si mirem a la taula 2 dels annexes, correspon a l'estació d'Orís sobre Torelló. La següent columna és la de DATA. Aquesta és la que ens informa de quin moment temporal correspon el valor, per això diré que estem tractant amb dades Time Series. Totes les dades els hi correspondrà un moment temporal. La columna PPT hi haurà el valor que ens interessa i que està unit a un "timestamp", en aquest cas els primers valors que dóna aquesta estació son nulls. L'última columna ens diu si el període és amb dades horàries (HO) o siguin semi horàries (SH). A mi m'interessa que siguin SH ja que les necessitaré totes en aquesta freqüència, com hem decidit previament.

Primer de tot citaré les tecnologies utilitzades per fer el processament de les corresponents dades del SMC, aquestes les importo a l'inici del document per tenir-ho organitzat. Aquesta anàlisi es troba al document *SMC_DataAnalysis.ipynb*.

- Numpy[16]
- Pandas[17]
- Matplotlib[18]
- Regex[19]
- Seasonal_decompose[20]
- Administració de fitxers com llistar fitxers d'un directori, etc…

3.1.1 Primer análisis de les estacions

Partint de la gran varietat d'estacions i tenint en compte que hi ha estacions que ja no estan actives o que les van posar en funcionament recentment, les he diferenciat en difernts blocs.

[16] "NumPy." https://numpy.org/.
[17] "pandas - Python Data Analysis Library." https://pandas.pydata.org/.
[18] "Matplotlib — Visualization with Python." https://matplotlib.org/.
[19] "Python RegEx - W3Schools." https://www.w3schools.com/python/python_regex.asp.
[20] "statsmodels.tsa.seasonal.seasonal_decompose." https://www.statsmodels.org/dev/generated/statsmodels.tsa.seasonal.seasonal_decompose.html.

Un primer bloc son les estacions que tinc dades desde el 2009 al 2022 però que s'han d'analitzar. El segon bloc són les que pels motius anteriors tinc diferents períodes, però unint amb dades d'altres estacions en la mateixa situació (pràcticament al costat) es poden completar.

3.1.1.1 Estacions amb dades desde el 2009 al 2022

Els codis respectius de les estacions son: CC, CG, CI, CY, DG, DJ, DN, KE, M6, UB, UN, UO, V3, V4, VN, WS i V5 aquestes estan al directori *dataSMC/*.

El primer pas del análisis és fer una revisió de les dades visualment, per poder veure si a simple vista observem alguna anomalia. Per això, utilitzaré dos tècniques diferents: la funció *seasonal_descompose* i la de plot de *Matplotlib*. *Seasonal_descompose* o la descomposició de sèries temporal implica pensar en una sèrie com una combinació de components de nivell, tendencia, estacionalitat i soroll.

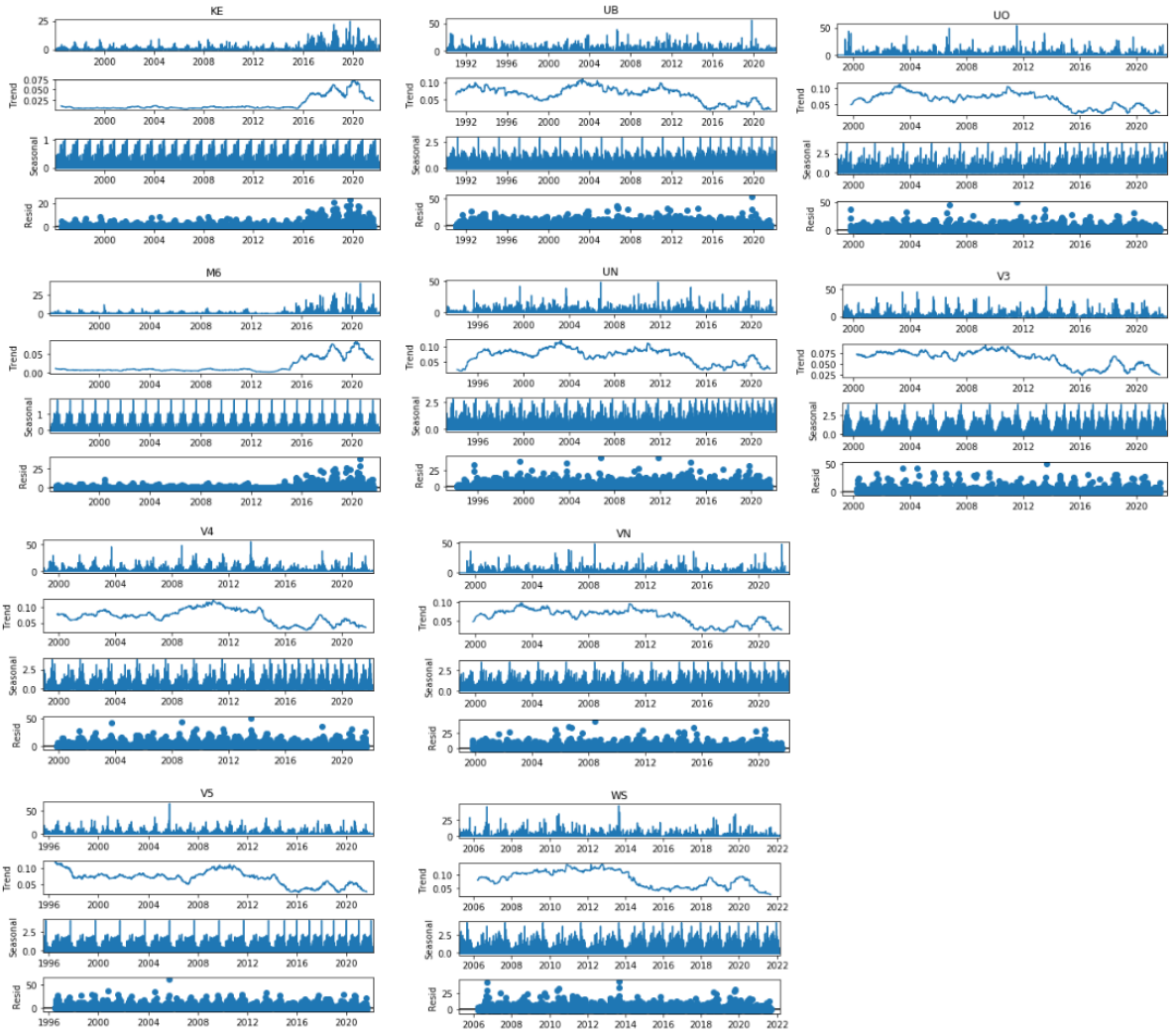

Figura 6: Descomposició temporal de les estacions KE, M6,UB, UN, UO, V3, V4, V5, VN i WS amb dades desde 2009 fins 2022

Les descomposicions de la resta d'estacions es troba a l'Annex 52.

A simple vista de la figura 6 s'observen variacions en el trend i el seasonal de les estacions: UB, UN, UO, V3, V4, V5, VN i WS. Això passa perquè en aquell moment la freqüència de les dades que tinc passa de dades cada 1h a cada 30 min.

També podem afirmar que les estacions KE i M6 tenen algun problema amb les dades, ja que el seu trend i residu comparat amb els altres és diferent. A partir del 2015 aproximadament la mitja de pluges puja considerablement. Això no es normal; intueixo que a partir del 2015 es va gestionar diferent l'estació. A la figura 7 s'observa una mica més ampliat com la mitja de pluges es dispara al 2015 per a aquestes dues estacions.

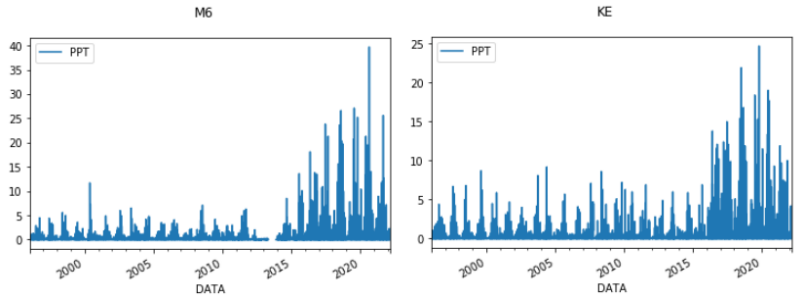

Figura 7: Gràfica de les dades de les estacions M6 i KE

Buscant les estacions KE i M6 per altres fonts (el portal de dades obertes[21]) he trobat que els primers registres disponibles son del 2016. Perquè em passen registres que no estan disponibles a les dades obertes i amb el plus que són erronis? He enviat un correu al SMC avisant de la anomalia.

KE i M6

El SMC m'ha respost:

"... Ens hem de disculpar perquè les dades que us vam enviar d'aquestes dues estacions no eren correctes, així com tampoc les dades de l'estació amb codi KP. L'explicació rau en el fet que aquestes estacions són un cas particular dins la XEMA i el mètode que vam usar per a la confecció dels fitxers no ho va tenir en consideració. Aquestes tres estacions es van integrar a la XEMA durant l'any 2015 (prèviament eren gestionades per l'Agència Catalana de l'Aigua). Per aquest motiu, aquestes estacions només disposen de dades de període SH, a partir de la data de la seva integració a la XEMA en algun moment del 2015, i en lloc de

[21] "Dades obertes - Servei Meteorològic de Catalunya | Meteocat." https://www.meteo.cat/wpweb/serveis/cataleg-de-serveis/serveis-oberts/dades-obertes/.

dades semihoràries com la resta d'estacions de la XEMA disposen de dades de període de 5 min per a tot el període anterior a la data d'integració que us facilitem a continuació …"

En la figura 8 es mostren les dades que em van tornar a passar sobre les estacions KE i M6 quan les gestionava l'ACA fins la seva integració a la XEMA.

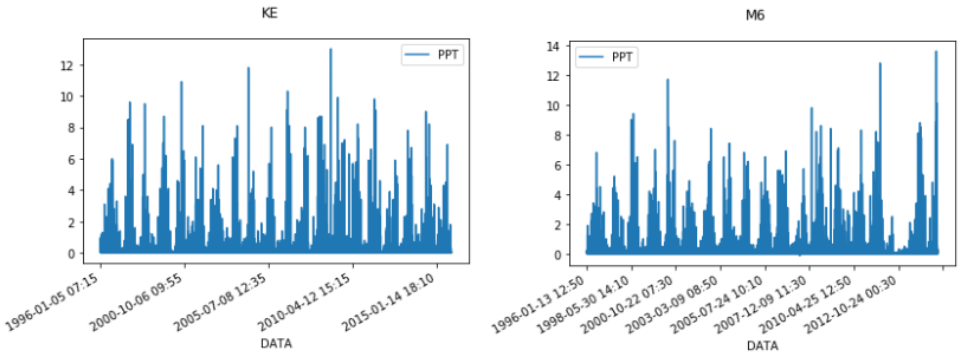

Figura 8: Gràfica de les dades de les estacions M6 i KE tornades a passar

3.1.1.2 Estacions a completar

Hi ha estacions per a les quals només tenim dades durant un subconjunt d'anys, però que les podem combinar per acabar-ne formant una. Els codis respectius de les estacions son: DM, WF, XJ, Z4 i ZC aquestes estan al directori *dataSMC/ATractar/*.

Com he fet amb les estacions que tenia les dades ja "completes" del 2009 al 2022, faig una revisió de les dades visualment, per poder veure si a simple vista observem alguna anomalia. Utilitzo les mateixes tècniques esmentades en l'apartat anterior.

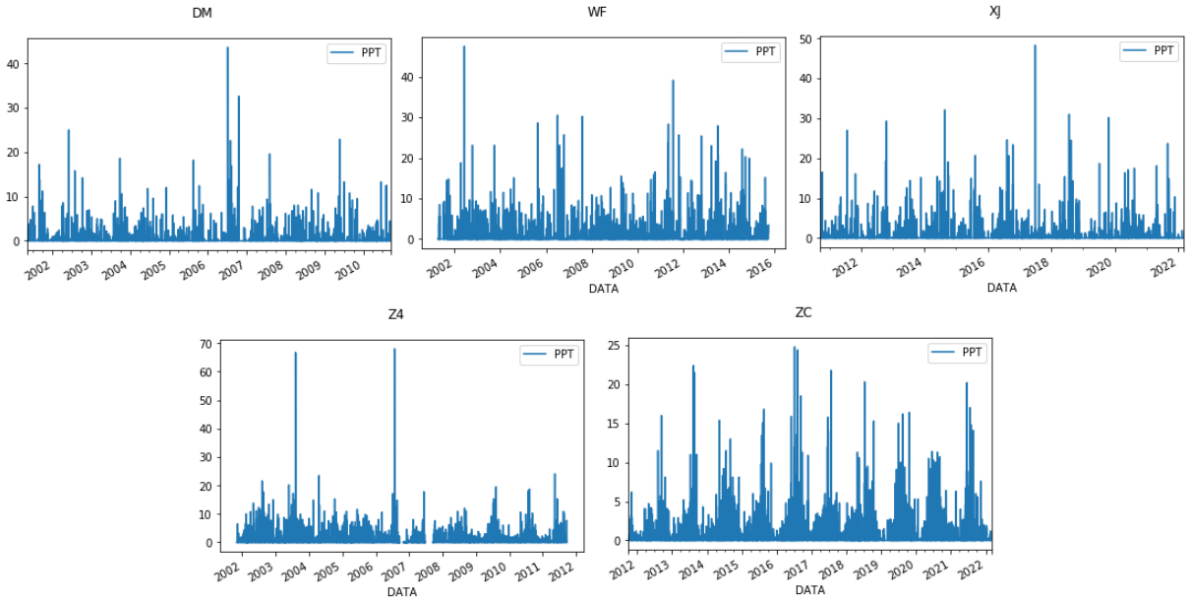

Figura 9: Gràfica de les dades de les estacions DM, WF, XJ, Z4 i ZC

No sembla a simple vista que aquestes tinguin cap anomalia tot i que siguin registres en diferents dates. Hi ha la possibilitat que tinguin dades a diferents freqüències SH i HO.

3.1.2 Generar el DataFrame

L'objectiu d'aquest apartat és modelar les dades de tal manera que em quedi un *dataFrame*, on cada fila serà un moment temporal (cada 30' desde el 2009 al 2022) i cada columna ha de ser una estació meteorològica diferent.

3.1.2.1 Estacions amb dades desde el 2009 al 2022

Creo un primer *dataFrame* amb les dades de les estacions que tinc dades desde el 2009 al 2022 que utilizaré de base un cop vaig tractant els diferents problemes. En aquest *dataFrame* (Figura 10) es pot observar, com ja haviem detectat, que les estacions a partir de la UB tenen almenys alguns primers registres en HO (freqüència horària). Com que aquest *dataFrame* correspon a les dades del camí *dataSMC/* , sabem que les columnes KE i M6 no son correctes.

DATA	CC	CG	CI	CY	DG	DJ	DN	KE	M6	UB	UN	UO	V3	V4	V5	VN	WS
2009-01-01 00:00:00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2009-01-01 00:30:00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2009-01-01 01:00:00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2009-01-01 01:30:00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figura 10: Primeres 4 files del dataFrame que correspon al fitxer: dataSMC/df_dadesComplertesATracatar.csv

3.1.2.1.1 Estacions amb diferents freqüències horàries

En aquest punt recupero el fitxer dataSMC/df_dadesCompletesATractar.csv i selecciono les columnes que tenen les estacions amb freqüència horària: UB, UN, UO, V3, V4, V5, VN i WS.

	DATA	UB	UN	UO	V3	V4	V5	VN	WS
0	2009-01-01 00:00:00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	2009-01-01 00:30:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	2009-01-01 01:00:00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	2009-01-01 01:30:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figura 11: Primeres 4 files del dataFrame extret

En la figura 11 podem observar la selecció feta de les estacions amb una freqüència horària HO que ens interessa passar-ho a freqüencia semi-horària SH.

Per solucionar aquest problema he decidit que quan hi hagi un *null* agafaré la següent dada i la dividiré entre 2 actualitzant el mateix valor. Aquesta decisió va ser presa per la impossibilitat d'accedir a les dades reals i aprofitant que són dades acumulades de pluja. Faig la simplificació d'assumir que el que plou en una hora, ho ha fet la meitat cada 30 min. Tinc en compte que la solució no és 100% realista, però és acceptable. Aquesta solució s'explica gràficament a la figura 12.

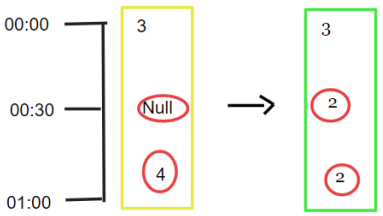

Figura 12: Representació gràfica de la resolució del problema

La funció que fa aquest tractament és la *NaNFrequencyResolver(data)*, on li passem el datagrama per paràmetres i ens el retorna transformat.

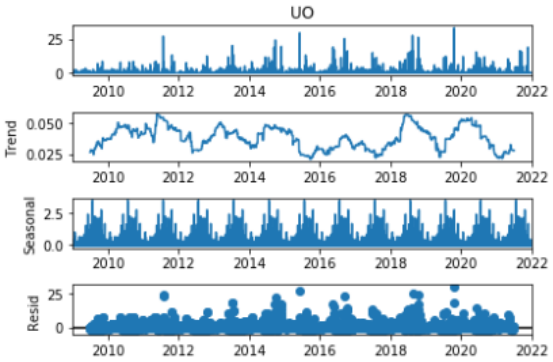

Figura 13: Descomposició temporal de l'estació UO després d'aplicar NANFrequencyResolver()

Un cop aplicat l'algorisme miro si ho ha fet correctament de dos maneres: primer torno a fer la descomposició temporal de totes les estacions fixant-me amb els camps de Seasonal i Trend com es veu en la figura 13 en l'estació UO. Acabo de confirmar que ho ha fet correctament mirant directament en punts localitzats del dataFrame, on sé que ha d'haver algun valor en concret.

Al finalitzar, em guardo l'excel d'aquest *dataFrame* amb el nom de: *dataSMC/fixed/UB-WS_Fixed.xlsx.*

3.1.2.1.2 Estacions KE i M6

De les estacions KE i M6 tinc dos arxius per cada estació, un primer des de la data inici fins a la data d'integració a la XEMA (el que m'han passat a causa de la incidència reportada) i

un altre desde l'inici fins al 2022 però amb dades errònies fins la integració a la XEMA (el que em van passar inicialment). Els fitxers que em van entregar un cop enviada la incidència tenen una freqüència de dades cada 5 minuts.

El primer que faig és canviar la freqüència dels fitxers que son cada 5 min a una freqüència de 30 min. Utilitzo la funció resample de pandas.DataFrame amb la component .sum() al final, això crearà un nou *dataFrame* amb períodes de 30' on aquest valor serà la suma de totes les dades que hi hagin en el període de 30'.

Figura 14: Representació gràfica del canvi de freqüència

A la figura 14 s'observa el comportament descrit anteriorment, com tots els valors desde 18:30 fins 19:00 es sumen i es guarden al nou *dataFrame* amb hora 19:00.

Seguidament, utilitzant la funció *combineTwoDataframesPriorizeLeft*, uneixo el *dataFrame* que li he canviat la freqüència amb les dades rebudes inicialment a partir de la integració a la XEMA.

En aquest punt, torno a imprimir els gràfics de les dades resultats, aquests són els de la figura 15:

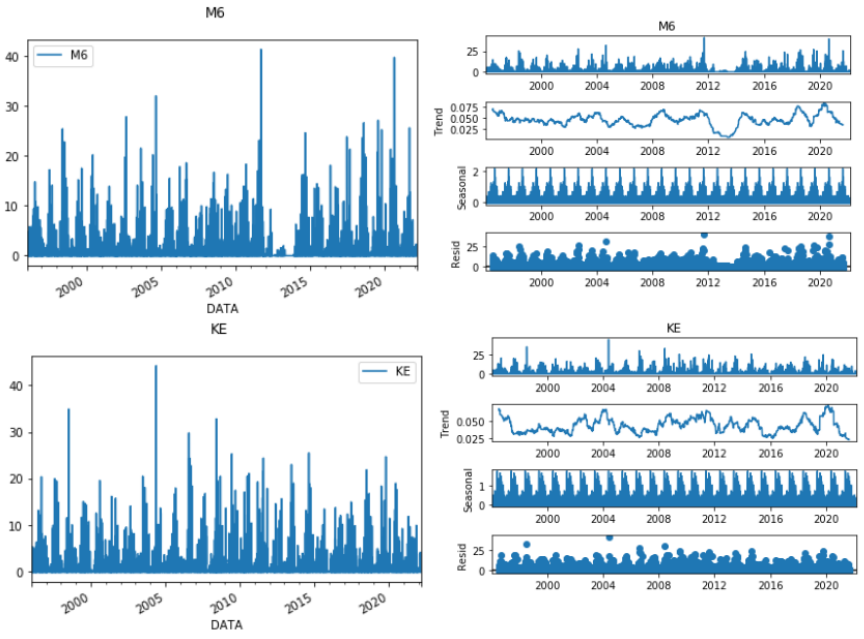

Figura 15: Gràfics de les estacions M6 i KE amb els fitxers ajuntats

Com s'observa a l'estació M6 les dades entre el 2012 al 2014 no semblen del tot correctes, per això no les dono per vàlides i elimino l'estació. En canvi, dono per vàlides les dades de la KE i me les guardo al fitxer *dataSMC/fixed/KE_FixedF.xlsx*.

3.1.2.2 Estacions a completar

Per completar les dades d'aquestes estacions he decidit unir els valors de tres estacions diferents per acabar-ne fent una, les uneixo perquè estan pràcticament al costat una amb l'altre. Algunes d'aquestes com la Z4 eren les estacions antigues que van desmantellar i substituir per les noves: ZC.

Estació1	Estació2	Estació3	Estació final	Lloc
DM	WF	XJ	DM-WF-XJ	Girona
Z4	ZC	DG?	Z4-ZC	Ulldeter

Figura 16: Taula de Correspondències d'estacions finals

3.1.2.2.1 DM - WF - XJ Girona

Primer de tot, la figura 17 mostra les dades de les estacions DM, XJ i WF per separat. Es pot veure que unint les tres es pot convertir en una.

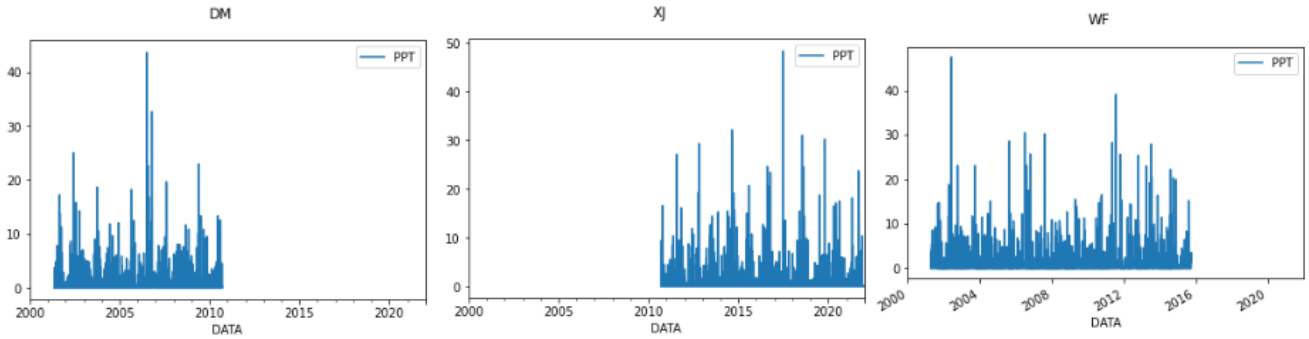

Figura 17: Gràfica de les estacions de Girona: DM, XJ i WF

Per construir el *dataFrame* de Girona, uniré les estacions DM i XJ. Entre l'últim registre temporal de les dades de l'estació DM i del primer de les dades de XJ, hi ha un període d'un dia que no hi han dades. Per aquest dia utilizaré les dades de WF que es una estació de costat de Girona, en aquest cas aquestes són horàries però posaré 0's als 30' que quedin amb NaN. Poso 0's perquè en aquest interval perquè coincideix que no hi han pràcticament precipitacions. El dataFrame resultant li diré DM_WF_XJ → *dataSMC/ATractar/fixed/DM_XJ_WF_Fixed.xlsx*.

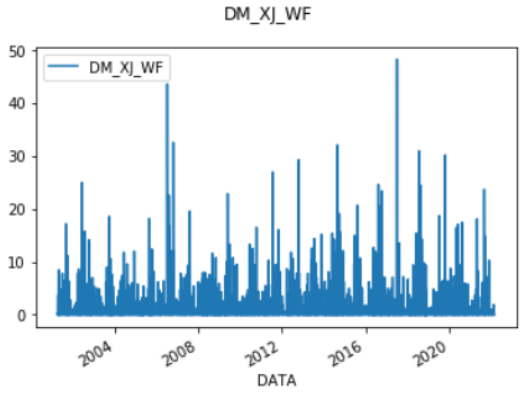

Figura 18: Gràfica de l'estació unida de Girona DM_XJ_WF

3.1.2.2.2 Z4 - ZC Ulldeter

Primer de tot, la figura 19 mostra les dades de les estacions Z4 i ZC per separat. Es pot veure que unint les dos es pot convertir en una.

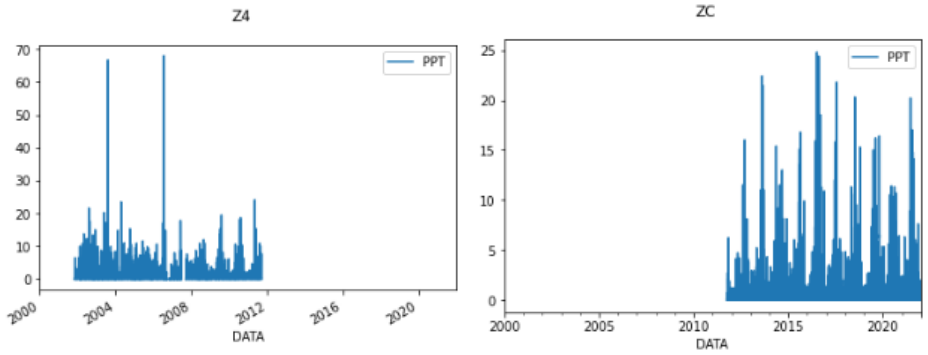

Figura 19: Gràfica de les estacions de Ulldeter: Z4 i ZC

Per construir el *dataFrame* d'Ulldeter, uniré les estacions Z4 i ZC. Ajuntant aquestes dos estacions queden uns dies sense dades, utilitzaré les dades de DG que és l'estació de Núria per omplir aquests dies. Ulldeter es una estació important i les dades de Núria són les que poden ser més representatives de la zona. El dataFrame resultant de la unió li diré Z4-ZC → *dataSMC/ATractar/fixed/Z4_ZC_Fixed.xlsx.*

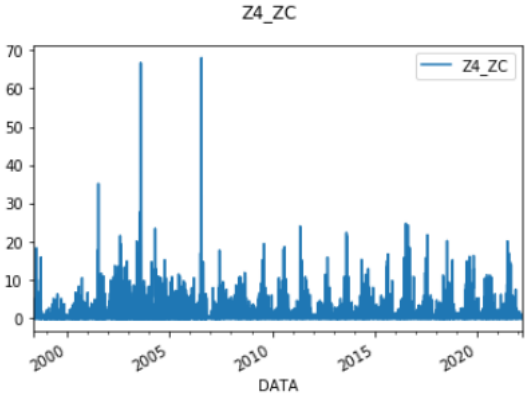

Figura 20: Gràfica de l'estació unida de Ulldeter Z4_ZC

3.1.2.2.3 DataFrame a completar final

Ara que ja tinc les estacions de Girona i Ulldeter, he de sel·leccionar les dades des del 2009 al 2022. Per això faig un *join left* dels dos arxius *DataSMC/ATractar/fixed/DM_XJ_WF_Fixed.xlsx* i *dataSMC/ATractar/fixed/Z4_ZC_Fixed.xlsx* amb un *dataFrame* base que té l'índex temps correcte quedant-me amb els valors d'aquest index (aquest serà el *dataFrame* left). Guardo aquest dataFrame a → *dataSMC/fixed/aTractar_Fixed.xlsx.*

3.1.2.3 DataFrame final

Al final construeixo el dataFrame final del SMC a partir dels diferents arxius que he anat guardant al haver tractat cada problemàtica. La figura 20 és una taula de correspondència dels diferents arxius que he anat tractant, on estan ubicats i quines estacions hi ha.

Arxiu	Ubicació	Estacions
KE_FixedF.xlsx	*dataSMC/fixed/KE_FixedF.xlsx*	KE
UB-WS_Fixed.xlsx	*dataSMC/fixed/UB-WS_Fixed.xlsx*	"UB", "UN", "UO", "V3", "V4", "V5", "VN", "WS"
aTractar_Fixed.xlsx	*dataSMC/fixed/aTractar_Fixed.xlsx*	"DM-WF-XJ Girona", "Z4-ZC Ulldeter"
df_dadesComplertesATracatar.csv	*dataSMC/df_dadesComplertesATracatar.csv*	"CC", "CG", "CI", "CY", "DG", "DJ", "DN"

Figura 21: Taula de referències de les ubicacions i estacions dels diferents arxius

Uneixo utilitzant el join left els diferents arxius/estacions i les guardo com DF_SMC → *finalsDF/DF_SMC.csv.*

DATA	CC	CG	CI	CY	DG	DJ	DN	KE	UB	UN	UO	V3	V4	V5	VN	WS	DM_XJ_WF	Z4_ZC
2009-01-01 00:00:00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2009-01-01 00:30:00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figura 22: Primeres dos files del dataFrame final del SMC → *finalsDF/DF_SMC.csv.*

3.2 ACA (Agència Catalana de l'Aigua)

Referent a l'ACA m'han enviat les dades a partir d'una resposta de la seva bústia de contacte, on m'expliquen que m'he de descarregar les dades via un link al seu servidor. Aquest és un fitxer zip on dins hi ha quatre fitxers txt per cada estació. Dos dels quatre són de cabal i els altres dos de nivell, hi ha dos fitxers de cada perquè un recull les dades fins el 2021 i l'altre, a partir del 2021. Això passa perquè a partir del 2021 les dades son validades a temps de registre, abans ho eren diariament.

El cabal és la quantitat d'aigua que passa en una secció del riu per unitat de temps. En canvi el nivell és l'altura de l'aigua desde el fons del riu fins la superfície.

Un cop tinc el zip descarregat, he fet una primera inspecció visual obrint un fitxer qualsevol per veure de quina forma estan guardades les dades. En aquest cas no em van passar cap document per interpretar les dades, com a conseqüència he hagut d'espavilar-me per trobar que era cada columna.

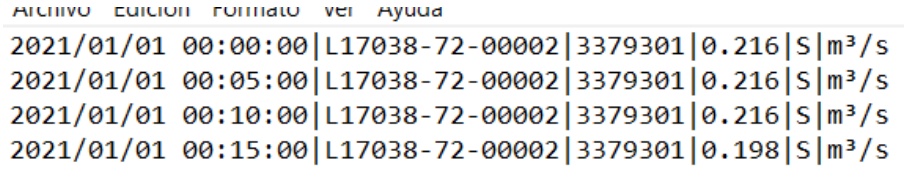

Figura 23: Primeres quatre files de dades de l'estació de l'ACA→ L17038-72-00002

L'arxiu txt de cada estació el formen 6 columnes i tantes files com registres tinguem d'aquella estació fins el 2021 o a partir del 2021. La primera columna correspon al moment temporal que pertany el registre/dada. La segona és el codi de l'estació, en aquest cas el L17038-72-00002 que si mirem a la taula 1 (als annexes), correspon a la riera de Gotarra al municipi de Campllong. La següent columna que hi ha el valor 3379301, és el valor que identifica quin dels dos arxius per cada estació corresponen les dades. La columna 4 hi ha la dada que ens interessa. En aquest càs les unitats en que està la dada són els m^3/s com ens indica la última columna. La S de la columna 5 no he sapigut saber que és, tots els registres i estacions de l'ACA la tenen.

Primer de tot citaré les tecnologies (que no hagi citat en l'anàlisis del SMC) utilitzades per fer el processament de les corresponents dades de l'ACA. Aquestes les importo a l'inici del document per tenir-ho organitzat. Aquest anàlisis fa referència al document *ACA_DataAnalysis.ipynb*.

- Sklearn[22]
- PrettyTable[23]

3.2.1 Primer análisis de les estacions

Primer de tot, com ja he fet amb les estacions del SMC, imprimeixo les gràfiques de tots els fitxers que estan guardats a *dataACA/cabal* per si en puc extreure alguna informació. A la majoria de gràfics s'observa com hi ha valors molt diferents de la resta, anòmals, i fa que sembli que el cabal habitual és 0. La figura 24 n'és un exemple.

[22] "scikit-learn: machine learning in Python — scikit-learn 1.1.1 ...." https://scikit-learn.org/.
[23] "PyPI · The Python Package Index." https://pypi.org/.

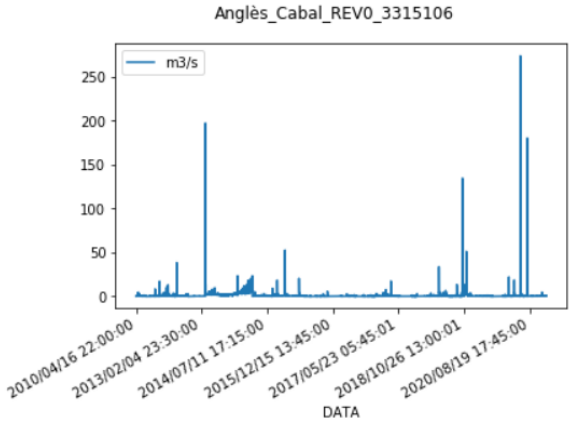

Figura 24: Grafica de les dades de cabal d'Anglès fins el 2021

Pel model final només utilitzaré les dades de cabal però com més endavant s'explicarà, utilitzaré també les de nivell per completar amb mès rigor les de cabal.

3.2.2 Generar el *dataFrame* de cabal

3.2.2.1 Unió de fitxers per cada estació

Com que tinc les dades de cada estació partides en dos arxius, primer he d'unir aquests dos arxius per posar després totes les estacions dins un *dataFrame* final. Els fitxers originals estan guardats a: *dataACA/cabal/*.

Per fer aquesta unió he implementat la funció *combineTwoDataframesPriorizeLeft(dfL,dfR=None)* que a partir de dos arxius els uneix donant prioritat a l'arxiu dfL si hi han dos dades en el mateix punt de temps. La funció retorna un *dataFrame* on hi ha una columna (amb nom de l'estació) que conté les dades de cabal i cada fila es cada moment temporal de la dada. Si només li pases dfL et retorna un dataFrame d'aquest únic arxiu.

	L17038-72-00002
DATA	
2009/01/01 00:00:00	3.892
2009/01/01 00:05:00	3.892
2009/01/01 00:10:00	3.892

Figura 25: Primeres tres files del dataFrame de l'estació L17038-72-00002

3.2.2.2 Analisis de les dades ajuntades

Quan tinc els dos fitxers units miro entre quines dates tinc les dades, per fer-me una idea de quin rang de dates m'han passat realment.

V_llocH	Desde	Fins
F014672	2010/04/16 22:00:00	2021/04/01 00:15:00
L17038-72-00002	2009/01/01 00:00:00	2021/04/01 00:25:00
L17079-72-00004	2009/01/01 00:00:00	2021/07/31 23:55:00
L17079-72-00005	2009/01/01 00:00:00	2022/02/01 23:55:00
L08116-72-00002	2010/07/06 09:35:00	2022/01/01 00:25:00
F001242	2009/01/01 00:00:00	2022/01/01 00:25:00
L17147-72-00005	2010/03/22 14:00:00	2022/01/01 00:25:00
F026458	2009/01/01 00:00:00	2021/04/01 00:25:00
L17167-72-00001	2010/07/01 13:35:00	2021/12/01 23:55:00
L17199-72-00001	2009/01/22 11:15:00	2022/02/01 00:25:00
L17055-72-00002	2009/01/01 00:00:00	2021/01/01 23:50:00
F001243	2009/01/01 00:00:00	2020/12/31 07:30:00
F000005	2017/08/24 19:00:01	2021/01/01 23:45:00
F009891	2009/01/01 02:52:00	2021/01/01 23:45:00

Figura 26: Taula que mostra el rang de temps de les dades

Com es veu a la figura 26, no totes les estacions que m'han passat tinc dades desde el 2009 ni les tinc fins al 2021. Seguidament uneixo els diferents fitxers en un sol dataFrame i ho guardo a *dataACA/Dataframes/DF_noTractat_cabal_ACA.xlsx.* Després recupero aquest fitxer per fer el tractament corresponent.

3.2.3 Generar el *dataFrame* de nivell

En aquest cas repeteixo el procediment dels apartats 3.2.2.1 i 3.2.2.2 però amb les dades de nivell que es guarden a *dataACA/Nivell/*. El *dataFrame* resultant es guarda a *dataACA/Dataframes/DF_noTractat_nivell_ACA.xlsx.*

3.2.4 Tractament de les dades

Al contrari que les dades de meteorologia, aquestes estan bastant incompletes. Fent una primera revisió sobre l'arxiu excel em trobo molts valors de diferents estacions en *null*. Alguns són en llocs a l'atzar, altres són en tot un període de registres consecutius i altres tinc temporades d'anys complets buits.

Una altra diferència amb les dades de pluja, que són TimeSeries però un valor podia ser completament diferent amb el de 30' abans, és que les dades de cabal i nivell mantenen una relació amb les anteriors i posteriors corresponents; és a dir, les sèries són suaus. Això passa perquè són les dades de l'evolució del cabal i nivell en el temps. El sentit comú ens diu que es completament impossible que un registre de cabal passi de 30 a 400 m^3/s en 30'.

En aquest apartat explicaré les tècniques que he aplicat sobre els *dataFrame* a mesura que he anat completant i netejant les dades.

3.2.4.1 Interpolació

“Procediment que, donats els n valors y_1, y_2, ..., y_i, ..., y_n d’una funció y = g(x) en els punts x_1, x_2 ..., x_i, ..., x_n, permet de calcular, aproximadament, els valors de g(x) en punts intermedis als donats.” (Grup enciclopèdia, n.d.)

Com es pot observar a la figura 27, entre el conjunt de les dades hi ha una quantitat de valors nuls que estan rodejats per valors no nuls. Quan el numero de valors nuls seguits entre valors no nuls no és molt alt, l’interpolació es una bona tècnica per omplir aquests espais, ja que els valors nulls es mourien entre els valors que no ho són.

2010-10-12 01:00:00	3,2992	44,4	2970,419	15,88
2010-10-12 01:30:00			2691,883	15,69
2010-10-12 02:00:00	3,2441			
2010-10-12 02:30:00			2724,593	
2010-10-12 03:00:00	3,3557	53,94	3859,186	16,07
				16,07

Figura 27: Exemples de registres on falten dades

Per omplir aquests valors he dissenyat la funció *interpolateOneValue(station,auxiliarDf, pattern).* Aquesta mitjançant el nom d’una estació, un *dataFrame* i el patró de valors nuls que es vol buscar, recorre la columna de valors de l’estació que li passem interpolant en els llocs on detecta el patró de nuls que volem. El patró de nuls que passarem a la funció serà en funció al número de nuls que volem interpolar en aquella iteració com es mostra a la figura 28.

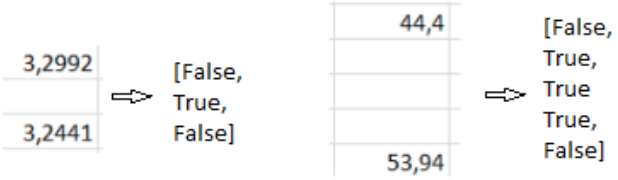

Figura 28: Registres de dades amb els patrons que cal introduir per identificar-los

3.2.4.2 Relacionar el nivell amb el cabal

Les dades rebudes de les diferents estacions de l’ACA no només inclouen dades de cabal sinó també inclouen les de nivell. Perquè no utilitzar aquestes dades de nivell per relacionar-les amb el cabal i completar algun espai que tingui la dada de nivell i no de cabal?

En un riu el cabal té una relació proporcional amb la de nivell, com més aigua més alt està el riu. Això és cert perquè l'espai per on passa el riu no acostuma a canviar i ha de passar més aigua en el mateix espai. Si trobem aquesta relació, que és diferent en cada punt del riu, tenint un valor de nivell podrem predir un de cabal.

Aquesta és una altre tècnica que he utilitzat per completar les dades i que explicaré en aquest apartat.

3.2.4.2.1 Valors incomprensibles

Abans de crear un model que em retorni un valor de cabal segons un de nivell per completar les dades de cabal, relaciono aquests dos valors i analitzo visualment les diferents gràfiques.

Com es pot veure a la figura 29, a on es mostra el cabal *versus* el nivell, hi ha estacions com la L17167-72-00001 que tenen uns valors de cabal molt desproporcionals a la mitjana segons al nivell que marca. Això em fa pensar que l'estació ha enregistrat valors incorrectes durant alguns períodes de temps continus i/o esporàdics.

Utilitzant la funció *movingAverageResolverTotal()* converteixo tots els valors que sobrepassen una xifra en concret a nul, ja que els considero que son incorrectes.

L'algorisme que utilitzo recorre tot el registre i mira si cada valor és més gran o petit que la mitjana més *x* vegades la desviació estàndard de tots els registres. No és ben bé el que fa el moving average, que agafa la desviació estàndard dels seus *y* registres més pròxims.

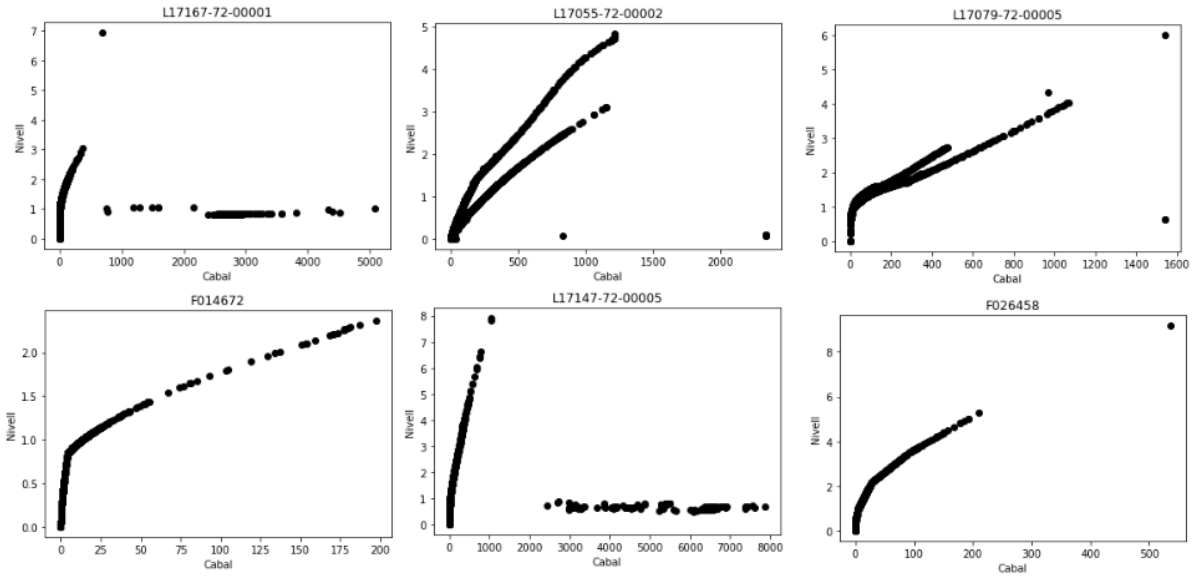

Figura 29: Gràfiques de la relació Cabal - Nivell de 6 estacions de l'ACA abans d'executar *movingAverageResolverTotal*

La resta de gràfiques de la relació Cabal - Nivell estan en els annexes figura 53.

3.2.4.2.2 S'han fet obres al riu?

Un cop trets els valors incomprensibles de cabal per fer la predicció, veig que els valors l'estació L17055-72-00002 semblen seguir dos funcions diferents. Poden haver diferents

circumstàncies que causin aquesta anomalia, una possibilitat es que hagin fet obres al riu i en algun moment la relació cabal - nivell hagi canviat.

Per poder millorar les prediccions de cabal en aquesta estació, tractaré per separat els valors de la funció en blau i en negre com es mostra a la figura 30.

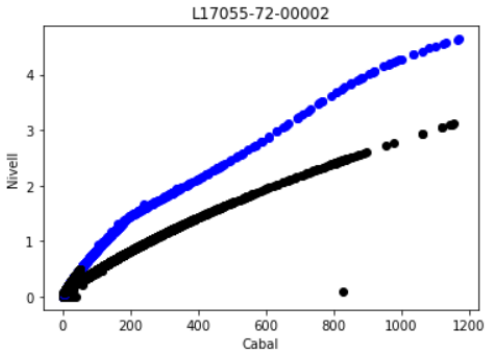

Figura 30: Gràfica de la relació Cabal - Nivell de l'estació L17055-72-00002

Si realment s'han fet obres al riu hi haurà una data la qual els punts canviaran de funció, aquesta serà la data de les obres que haurien modificat el curs del riu causant un canvi em la relació nivell - cabal. La realitat ha sigut diferent, he trobat deu moments en que les dades canvien de funció, per això descarto la hipòtesi d'unes obres al curs del riu. Com que canvia amb una certa freqüència hi hauria la probabilitat que fos una neteja del riu (un dragat).

3.2.4.2.3 KNN Regressor

Després d'haver analitzat les dades, exclòs els registres que tenien valors incomprensibles i trobat les dates per poder gestionar l'estació anterior com a dos estacions diferents, miro quants valors nuls podré omplir. Això es mostra a la figura 31, la reducció de nuls amb aquesta tècnica no és molt gran però en algunes estacions aconseguim erradicar-los quasi al complet.

Estació	Numero de null's	%	Numero de nulls que podem omplir	%Nulls que podem
F014672	118881	52.2	3900	3.3
L17038-72-00002	22574	9.9	190	0.8
L17079-72-00004	8195	3.6	351	4.3
L17079-72-00005	23408	10.3	337	1.4
L08116-72-00002	93244	40.9	26801	28.7
F001242	18277	8.0	38	0.2
L17147-72-00005	25766	11.3	24884	96.6
F026458	22836	10.0	301	1.3
L17167-72-00001	29512	12.9	27672	93.8
L17199-72-00001	11372	5.0	19	0.2
L17055-72-00002	54917	24.1	11347	20.7
F001243	177250	77.8	751	0.4
F000005	215585	94.6	46535	21.6
F009891	167983	73.7	2601	1.5

Figura 31: Taula que mostra el número de nuls podem omplir si aplico el model

A continuació desenvolupo les funcions KNNRegressor i KNNRegressorL17055_72_00002 (Funció explícita per l'estació L17055_72_00002, per tenir en compte les dues funcions dins

la pròpia estació). Aquestes funcions agafen les dades de cabal i nivell de l'estació corresponent i entrenen un KNNRegressor[24] amb les dades completes i amb K=7. Un cop tinc un model entrenat que em relaciona aquestes dos variables, selecciono les files que tinc un valor nul a cabal i no nul al nivell i faig la predicció. Amb aquesta predicció omplo els diferents espais nuls corresponents del dataFrame que retorno.

3.2.4.3 Moving average

Com ja ens van avisar des de l'ACA i com hem vist en apartats anteriors, no totes les dades que tenim en els diferents registres són vàlides. Alguns sensors de cabal i nivell de sobte donen valors molt estranys, també quan utilitzem les tècniques anteriors per introduir valors podem estar introduint valors erronis.

Per netejar les dades he utilitzat la funció *MovingAverageRolling*. Aquesta recorre tot el registre i mira si cada valor és més gran o petit que la mitjana més *x* vegades la desviació estàndard dels seus *y* registres més pròxims. Els valors de *x* i *y* els hi passo per paràmetre a la funció, utilitzo x=20 i y=21.

3.2.4.4 Iteracions finals

Per si soles totes les tècniques anteriors donen resultats prou bons a l'hora d'omplir valors nuls, però el resultat augmentaria si iteres les diferents tècniques una darrere de l'altre uns cops. Això és el que he fet en aquest apartat, dividint en dos les iteracions: una primera part d'iteracions utilitzant les tècniques explicades anteriorment i una segona part on aplico l'algorisme que m'acabarà d'omplir el *dataFrame*.

La primera part és a nivell d'estació: només es fan servir dades de la mateixa estació per a imputar. A la segona part, Iterative Imputer fa servir dades de totes les estacions per a imputar les altres.

3.2.4.4.1 Primera part

En aquesta part faig una funció que aplica les tècniques anteriors un número de vegades determinat. Perquè sigui el màxim d'òptim, ordeno les diferents tècniques en aquest ordre: només en la primera iteració trec els valors incomprensibles com s'explica a l'apartat 3.2.4.2.1, seguidament (ja en totes les iteracions futures) aplico el moving average, interpolo des d'un fins a 5 valors nuls i finalment aplico el *regressor* KNN amb K=7. Tot aquest desenvolupament queda reflectit a la funció *modelIteration.*

[24] "sklearn.neighbors.KNeighborsRegressor." http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html.

En la tècnica del moving average només l'aplico al *dataFrame* de cabal, en canvi la interpolació la faig sobre cabal i nivell.

Per fer-se una idea sobre el temps que tarda la funció en completar-se, li he posat que iteri 4 cops i ho ha fet en 5:39 h. Tarda tant perquè hi ha tècniques que han de recórrer tot el *dataFrame* uns quants cops per poder-se fer.

Estació	1 Numero de null's	1 %	2 Numero de null's	2 %
F014672	129461	56.8	114673	50.3
L17038-72-00002	22754	10.0	22306	9.8
L17079-72-00004	8208	3.6	7791	3.4
L17079-72-00005	23720	10.4	22958	10.1
L08116-72-00002	93371	41.0	66404	29.1
F001242	18468	8.1	18155	8.0
L17147-72-00005	25711	11.3	813	0.4
F026458	23014	10.1	22456	9.9
L17167-72-00001	29528	13.0	1789	0.8
L17199-72-00001	11781	5.2	11148	4.9
L17055-72-00002	55545	24.4	43360	19.0
F001243	177336	77.8	175559	77.0
F000005	215585	94.6	168754	74.0
F009891	173791	76.3	164895	72.4

Figura 32: Millores dels valors nuls després d'aplicar *modelIteration()*

En la figura 32 es mostra la millora del numero de nuls en cada estació després d'haver passat l'algorisme. Tot i que he omplert molts valors nuls hi ha estacions que encara en tenen molts.

A les estacions L08116-72-00002, L17147-72-00005 i L17167-72-00001, he pogut reduir molt el número de nuls. En d'altres com F009891, F000005, F001243 i F014672, el percentatge de nuls encara és molt alt i he decidit eliminar-les de l'estudi. Aquestes corresponen: riu Ges a Torelló, riera Llemana a Ginestar, Cornella Terri a Banyoles i a la riera d'Osor a Anglès.

3.2.4.4.2 Segona part (Iterative Imputer)

En aquest punt ja he omplert el màxim de nuls que podia utilitzant les diferents tècniques que he esmentat. Per finalitzar d'omplir la resta de valors he utilitzat un imputador multivariant que estima cada valor a partir de la resta, aquest es diu *IterativeImputer*[25].

L'IterativeImputer modela cada característica amb valors que falten en funció d'altres característiques i utilitza aquesta estimació per la imputació. Ho fa de manera iterada, en

[25] "sklearn.impute.IterativeImputer — scikit-learn 1.1.1 documentation." http://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html.

cada pas una columna de característiques es designa com a sortida y i les altres columnes son tractades com a entrada X. Un regressor s'ajusta a (X,y) per un y conegut i s'utilitza aquest regressor per predir els valors de y que falten. Això ho fa per cada característica de manera iterativa i es repeteix un número de cops fins que es retornen els resultats finals d'aquestes repeticions.

Com que l'imputador utilitza la resta de variables per omplir un valor, he partit les estacions que formen part a l'alt Ter en una banda i les del baix per una altre. En el mateix *imputer* he utilitzat l'estimador *KNeighborsRegressor* perquè em dóna millors resultats que el que hi ha per defecte.

A partir d'aquesta etapa ja tinc totes les dades completes.

4. Disseny dels models

Després del processament de dades que he explicat a l'apartat 3, ja tenim les dades del SMC i de l'ACA completes i preparades per crear els diferents datasets pels models. En aquest apartat explico des de la creació dels dataFrame fins la metodologia que he utilitzat a la hora de crear i entrenar els diferents models.

A causa de la falta de temps, ja que he necessitat molta energia per fer el preprocessament de les dades, en aquest apartat no s'ha pogut desenvolupar tot el que es volia fer en un principi. Per això deixo les diferents idees a l'apartat de treballs futurs.

Pel disseny d'aquests models he utilitzat les llibreries que he citat anteriorment en l'apartat 3, així com la xgboost[26] pel model XGBoost i la yellowbrick[27] per extreure la importància de les característiques del model.

4.1 DataFrames

En totes les tècniques de machine learning que utilitzaré, necessito passar les dades X i y al model, les quals les X's seran les variables predictoras i la y la variable a predir. L'estructura de dades d'X és una matriu o una llista amb la forma (número de mostres, número de característiques) i la de y és una matriu amb forma (número de mostres, número d'outputs). En el nostre cas, tindrem un únic output, així que y es pot veure com a un vector de mida (número de mostres).

En aquest cas les X's seran variables de cabal i precipitacions de la part alta o baixa segons quina part del riu estiguem modelant. En canvi, la y serà una estació en concret de l'ACA amb les dades corresponents a 24h més tard, serà l'estació que voldrem predir el cabal i la de la part alta serà diferent a la de la part baixa.

Que la y sigui les dades d'una estació en concret 24h més tard, és una decisió de disseny. Volem ser capaços de predir què passarà a un punt concret del riu d'aquí a 24 hores. Això hauria de donar temps a les administracions a planificar-se en cas d'un potencial problema com una avinguda.

[26] "XGBoost Documentation — xgboost 1.6.1 documentation." https://xgboost.readthedocs.io/. S'hi ha accedit el dia 6 de juny. 2022.
[27] "Model Selection Visualizers — Yellowbrick v1.4 documentation." https://www.scikit-yb.org/en/latest/api/model_selection/index.html. S'hi ha accedit el dia 6 de juny. 2022.

Si en aquest moment agaféssim totes les dades d'una estació de l'ACA i les passéssim com a y, estariem passant aquestes amb els valors del moment actual, no les de 24h abans. Per això he utilitzat la funció *makeDfy,* que és l'encarregada de moure els valors de y perquè en cada index temporal li correspongui la data 24h més tard. Aquesta funció se li pot passar amb quantes hores de marge volem que es modifiqui, podent així crear y's a les hores posteriors que es vulgui.

4.1.1 Part alta

Per la creació de la matriu X i el vector y de la **part alta del riu**, he utilitzat les següents estacions:

- Per les X, les estacions de l'ACA: 'L17147-72-00005', 'L08116-72-00002' i 'L17167-72-00001', més les estacions del SMC: 'CC', 'CG', 'CI', 'CY', 'DG', 'V3', 'V4', 'V5', 'WS', 'Z4_ZC'

- La variable a predir de la part alta és l'estació 'L08116-72-00002' (masies de roda) 24h més tard, aquesta és la última estació de l'ACA que tinc abans dels tres pantans del curs mig del riu.

4.1.2 Part baixa

Per la creació de la matriu X i el vector y de la **part baixa del riu**, he utilitzat les següents estacions:

- Per les X, les estacions de l'ACA: 'L17038-72-00002', 'L17079-72-00004', 'L17079-72-00005', 'F001242', 'F026458', 'L17055-72-00002' més les estacions del SMC: 'DJ', 'DN', 'KE', 'UN', 'UO', 'VN', 'WS', 'DM_XJ_WF'

- La variable a predir de la part baixa és l'estació 'L17055-72-00002' (Colomers) 24h més tard.

A la hora de triar l'estació a predir estava entre Colomers i Girona, però donada la certesa que és una zona pròxima a inundacions en situacions d'avinguda i altres interessos de la zona, m'he acabat de decidir per Colomers. He cregut oportú treure dos estacions que físicament s'ubiquen després de Colomers pel que fa el curs del riu, a nivell de planificació del riu no té sentit agafar dades que realment no contribueixen. Aquestes són: 'L17199-72-00001' (sensor de l'ACA a Torroella de Montgrí) i 'UB' (sensor del SMC a la Tallada d'Empordà)

4.2 Procediment de validació dels models

Perquè els resultats dels diferents models es puguin donar com a vàlids s'han de seguir una serie de tècniques que explicaré a continuació. He seguit el mateix procediment de validació per a tots els diferents models.

Primer de tot he dividit les dades en Train i Test, utilitzant el *TimeSeriesSplit*[28]. Aquest mètode em separarà les dades de tres formes diferents tenint en compte que les dades son *time series* i no es pot fer aleatòriament. Això vol dir que es farà tot el procés varis cops diferents (un amb cada partició de Train - Test diferent), en el nostre cas li diem que ho faci 3. Això ho faig per prevenir el model de overfitting i poder extreure resultats més reals. Aquest mètode és un tipus de validació creuada específic per a *TimeSeries Data*.

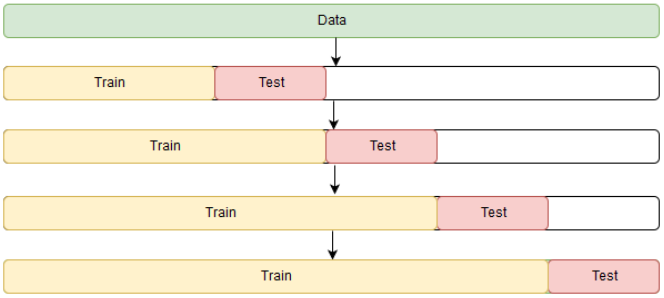

Figura 33: Funcionament de les divisions Train - Test utilitzant TimeSeriesSplit
(https://www.kaggle.com/code/kashnitsky/correct-time-aware-cross-validation-scheme/notebook)

Quan ja tinc les dades de Train i Test separades utilizo el *GridSearchCV*[29], em serveix per gestionar la manera en que faig el *fit* de les dades:

- Automatitzo la selecció dels millors paràmetres del model, a partir d'un diccionari que li passo amb els diferents paràmetres que es poden configurar al model corresponent i els possibles valors que poden tenir aquests paràmetres.

- Aplico la validació creuada sobre les dades de Train passant-li una instància del *TimeSeriesSplit* amb un número de divisions de 5. Això farà que busqui els millors paràmetres però en el conjunt de les 5 divisions Train - Validation.

- Aplico paral·lelització perquè vagi molt més ràpid.

[28] "sklearn.model_selection.TimeSeriesSplit." http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html .

[29] "sklearn.model_selection.GridSearchCV." http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html.

En aquest mateix GridSearchCV li passo el tipus de model de *Machine Learning* que faré servir en cada moment. Tot això fa que per cada partició de dades *Train - Test* inicial, *GridSearchCV* busqui els millors paràmetres del model a partir d'analitzar les dades en 5 subdivisions Train - Validation a partir de les dades de Train inicials.

En cada iteració, em guardo una sèrie d'informació en un *dataFrame* per l'analisis posterior. L'informació és: quin model estic utilitzant, els millors paràmetres del model, l'importància en el model de les diferents variables que li passem, la puntuació del model utilitzant RMSE[30] i el R^2[31] a quantes hores s'ha fet la predicció entre d'altres dades. Aquesta informació es guarda als fitxers *resultsDf/dfmodels/DfResult_AltTer.xlsx i resultsDf/dfmodels/DfResult_BaixTer.xlsx.*

El RMSE: Si $\widehat{y}_i$ es el valor predit de la i-a mostra, i y_i es el valor corresponent i correcte, llavors el RMSE estimat sobre $n_{mostres}$ es definit com a la figura 34:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

Figura 34: Formula RMSE

En canvi l'R^2: Si $\widehat{y}_i$ es el valor predit de la i-a mostra i y_i es el valor corresponent i correcte pel total de les $n_{mostres}$, on $\bar{y} = \frac{1}{n}\Sigma_{i=1}^{n} y_i$ R^2es definit com a la figura 35 :

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Figura 35: Formula R^2

4.3 Tipus de models utilitzats

En aquest projecte, treballarem amb 4 tipus de model d'aprenentatge automàtic diferents: *Random Forest*, *Extra-trees*, *XGBoost* i, com a referència, models de regressió lineal.

4.3.1 Random Forest

Random Forest és una tècnica d'aprenentatge automàtic supervisada basada en arbres de decisió. Això vol dir que es crea un conjunt divers de classificadors (arbres de decisió)

[30] "Raíz del error cuadrático medio - Wikipedia, la enciclopedia libre." https://es.wikipedia.org/wiki/Ra%C3%ADz_del_error_cuadr%C3%A1tico_medio.
[31] "Coefficient of determination - Wikipedia." https://en.wikipedia.org/wiki/Coefficient_of_determination.

introduint l'aleatorietat en la construcció del mateix. La predicció del conjunt es la predicció mitjana dels classificadors individuals.

En el *Random Forest* cada arbre del conjunt es construeix a partir d'una mostra extreta per substitució (mostra *bootstrapping*) del conjunt d'entrenament. També, quan es divideix cada node durant la construcció d'un arbre, la millor divisió es troba entre totes les característiques d'entrada o un subconjunt aleatori (max_features).

El seu principal avantatge és que obté un millor rendiment de generalització. Aquesta millora aconsegueix compensant els errors de les prediccions dels diferents arbres de decisió els quals presenten una gran variància.

4.3.2 XGBoost

XGBoost és una tècnica d'aprenentatge automàtic supervisada basada també en arbres de decisió. El que el diferència del *Random Forest* és que aquesta tècnica crea un arbre a la vegada perquè es tinguin en compte totes les dades relatives a l'arbre de decisió. Amb cada nou arbre, el rendimient del model es calcula i el següent arbre s'apren intentant reduir l'error dels anteriors. Per tant, es considera el gradient dels resultats. Com que es considera aquest gradient, habitualment obté millors resultats que el RF.

4.3.3 Extra Trees

Extra Trees és una tècnica d'aprenentatge automàtic supervisada basada igualment en arbres de decisió. Això vol dir que es crea un conjunt divers de classificadors introduint l'aleatorietat en la construcció del mateix. La predicció del conjunt es la predicció mitjana dels classificadors individuals.

Pel que fa al *Random Forest* (n'és una extensió), en aquest model l'aleatorietat va un pas més enllà en la fòrmula amb que es calculen les divisions dels nodes. El *Random Forest* tria la divisió més òptima mentres el Extra Trees la tria de manera aleatòria.

Comparant-lo amb RF, l'Extra Trees es més ràpid computacionalment. Tot el procediment és el mateix que el de RF però no ha de calcular l'optima divisió.

4.3.4 Regressió Lineal

La regressió lineal és una tècnica de modelatge estadístic que es fa servir per descriure una variable de resposta contínua com una funció lineal d'una o diverses variables predictores. Pot ajudar a comprendre i predir el comportament de sistemes complexos.

Les tècniques de regressió lineal permeten crear un model lineal. Aquest model descriu la relació entre una variable dependent y (resposta) com una funció d'una o diverses variables independents X_i (predictores).

5. Proves, Avaluació i Resultat

En aquesta secció explicaré les diferents proves que he fet, quins models he utilitzat i els resultats d'aquestes proves.

Pel que fa els models que utilitzo són els mateixos per les dos parts del riu: el Random Forest[32], XGBoost, ExtraTrees[33] i finalment la Regressió Lineal[34]. Com que cada model té una implementació diferent, l'objectiu de provar-ne diferents és trobar el que em doni un millor resultat en les prediccions.

Tots els resultats dels diferents models de la part alta i baixa els guardo a *\resultsDf*, en aquesta carpeta tinc en una part els objectes de cada model en forma de *pickle*[35] i per l'altra banda els *dataFrames* on em guardo els resultats.

En els següents apartats mostraré imatges dels resultats de les diferents prediccions, en els annexes hi ha les gràfiques que comparen la predicció amb el valor real per cada model. També hi ha gràfics complementaris que mostren més visualment la importància/pès de cada variable en la predicció que fa el model

5.1 Part alta

Com he explicat anteriorment, el conjunt de dades que utilitzo pel càlcul dels diferents models en la part alta del riu serà sempre el mateix, tal i com s'explica a l'apartat 4.1.1.

5.1.1 Random Forest

El primer tipus de model de la part alta que he provat ha sigut amb el *Random Forest*. A la figura 34 es mostren els resultats obtinguts.

[32] "sklearn.ensemble.RandomForestRegressor." http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html.
[33] "sklearn.ensemble.ExtraTreesRegressor." http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesRegressor.html.
[34] "sklearn.linear_model.LinearRegression." http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html.
[35] "pickle — Python object serialization — Python 3.10.5 documentation." https://docs.python.org/3/library/pickle.html.

	Method	Best_Params	Features	Feature_Importance	Score	RMSE_Score	Prediccio_Hores
0	Random Forest	{'max_depth': 10, 'max_features': 'sqrt', 'n_estimators': 100}	Index(['L17147-72-00005', 'L08116-72-00002', 'L17167-72-00001', 'CC', 'CG', 'CI', 'CY', 'DG', 'V3', 'V4', 'V5', 'WS', 'Z4_ZC'], dtype='object')	[0.18078640797928794, 0.46202659175183536, 0.16503526686908931, 0.008529010393631787, 0.03897723624141299, 0.024112076195114705, 0.008855436056206143, 0.014254663199784499, 0.012523273486239817, 0.014088789041725662, 0.01079136220265008, 0.04547102079270204, 0.01454886579031962]	0.350251	21.771579	1 days 00:00:00
1	Random Forest	{'max_depth': 10, 'max_features': 'sqrt', 'n_estimators': 50}	Index(['L17147-72-00005', 'L08116-72-00002', 'L17167-72-00001', 'CC', 'CG', 'CI', 'CY', 'DG', 'V3', 'V4', 'V5', 'WS', 'Z4_ZC'], dtype='object')	[0.20205816773926552, 0.33125397603075196, 0.16376680313899925, 0.016018784761167786, 0.04746647585366645, 0.042894063728360056, 0.016845329177516365, 0.029787552185481 25, 0.033331653676550034, 0.018190849381093856, 0.015118082674365667, 0.04674366319616543, 0.03652459845661651]	0.176638	13.506175	1 days 00:00:00
2	Random Forest	{'max_depth': 8, 'max_features': 'sqrt', 'n_estimators': 200}	Index(['L17147-72-00005', 'L08116-72-00002', 'L17167-72-00001', 'CC', 'CG', 'CI', 'CY', 'DG', 'V3', 'V4', 'V5', 'WS', 'Z4_ZC'], dtype='object')	[0.21602726931242175, 0.34006943839980286, 0.15948920948836276, 0.011438915307903233, 0.05517050545507773, 0.04006074841071579, 0.014607423053794966, 0.026419827244845 91, 0.02636689722499742, 0.015347549178801703, 0.0131604889135106, 0.04706554308053485, 0.034776184929230555]	0.324907	34.007339	1 days 00:00:00

Figura 36: Resultats de les diferents prediccions del Random forest (Part alta)

Com es pot observar, els paràmetres òptims del model que ha trobat no són constants en les tres proves, el número d'estimadors és completament diferent en totes, però en canvi el màxim de característiques és igual. Això és un indicador de falta d'estabilitat del procés d'aprenentatge.

Pel que fa a la importància de les variables predictores, en les tres proves han trobat com a variables més importants les tres estacions de l'ACA. La que té més importància és la que correspon a l'estació que es vol predir 24h més tard, seguit de les altres dos. Les estacions de meteorologia que les segueixen són les de WS → Viladrau, CG → Molló - Fabert i CI → Sant pau de Segúries.

Com es veu a les primeres tres gràfiques de la figura 44 dels annexes, l'importància relativa d'aquestes estacions del SMC és d'un 15% aproximadament. Si les comparem amb les estacions de l'ACA, tenen poca influència sobre el model.

L'arrel de l'error quadràtic mig és 23,09, un valor molt millorable en un model d'aquestes característiques i l'*estandard deviation* és 8,42 entre aquest tres.

A la figura 44 dels annexes hi ha les gràfiques de les diferents prediccions i de la importància de cada variable d'aquest model

5.1.2 XGBoost

3	XGBoost Regressor	{'max_depth': None, 'max_features': 1, 'n_estimators': 50}	Index(['L17147-72-00005', 'L08116-72-00002', 'L17167-72-00001', 'CC', 'CG', 'CI', 'CY', 'DG', 'V3', 'V4', 'V5', 'WS', 'Z4_ZC'], dtype='object')	[0.13729776, 0.3124329, 0.077890694, 0.008766092, 0.1306806, 0.055532247, 0.043652736, 0.027048096, 0.024549227, 0.057109565, 0.017250586, 0.08486711, 0.022922415]	0.303154	22.546818	1 days 00:00:00
4	XGBoost Regressor	{'max_depth': 25, 'max_features': 1, 'n_estimators': 50}	Index(['L17147-72-00005', 'L08116-72-00002', 'L17167-72-00001', 'CC', 'CG', 'CI', 'CY', 'DG', 'V3', 'V4', 'V5', 'WS', 'Z4_ZC'], dtype='object')	[0.028134793, 0.108188584, 0.033323053, 0.017333046, 0.16420627, 0.096603386, 0.10592702, 0.06614422, 0.09590034, 0.033230066, 0.021250851, 0.12416892, 0.10558949]	-0.108233	15.669402	1 days 00:00:00
5	XGBoost Regressor	{'max_depth': None, 'max_features': 1, 'n_estimators': 50}	Index(['L17147-72-00005', 'L08116-72-00002', 'L17167-72-00001', 'CC', 'CG', 'CI', 'CY', 'DG', 'V3', 'V4', 'V5', 'WS', 'Z4_ZC'], dtype='object')	[0.10767025, 0.26167366, 0.05731047, 0.024346534, 0.11763438, 0.065687984, 0.030311702, 0.03863548, 0.08861965, 0.057064112, 0.031759776, 0.064633414, 0.05465261]	0.309775	34.38635	1 days 00:00:00

Figura 37: Resultats de les diferents prediccions del XGBoost (Part alta)

El model XGBoost ha trobat els paràmetres òptims molt semblants en les tres prediccions, només canvia el valor de la màxima profunditat que arriba el model. Això és bona senyal i vol dir que el procés d'aprenentatge és bastant estable.

La importància de les estacions canvia una mica comparant-lo amb el *Random Forest*, l'estació del SMC → Molló - Fabert passa a estar en les tres estacions més importants per predir el resultat del model. També, l'estació de Viladrau es reafirma com a important juntament amb les tres estacions de cabal.

En aquest model les estacions de pluja agafen la mateixa importància que les de cabal per fer les prediccions. Com es veu a la figura 45 dels annexes, la importància relativa de les estacions de Molló - Fabert (Als pirineus tocant amb França, part nord de la conca del Ter) i Viladrau (Part sud de la conca del Ter) està al mateix nivell o superior a les estacions de cabal.

En aquest model s'està demostrant que la informació de la pluja és important.

L'arrel de l'error quadràtic mig és 24,18, un valor molt semblant al del *Random Forest* i molt millorable en un model d'aquestes característiques. L'*estandard deviation* és 7,73 entre aquest tres.

A la figura 45 dels annexes hi ha les gràfiques de les diferents prediccions i de la importància de cada variable d'aquest model

5.1.3 Extra Trees

6	ExtraTreesRegressor	{'max_depth': 25, 'max_features': 'sqrt', 'n_estimators': 100}	Index(['L17147-72-00005', 'L08116-72-00002', 'L17167-72-00001', 'CC', 'CG', 'CI', 'CY', 'DG', 'V3', 'V4', 'V5', 'WS', 'Z4_ZC'], dtype='object')	[0.15607399151860574, 0.47399261995522907, 0.15437628534452755, 0.009319234088383784, 0.03585497723592673, 0.02387695494546681, 0.010984026261117408, 0.019357666895429476, 0.016029907201233474, 0.01769697943741068, 0.01553882100940632, 0.051394227468656316, 0.015504308638606823]	0.346912	21.827447	1 days 00:00:00
7	ExtraTreesRegressor	{'max_depth': 25, 'max_features': 'sqrt', 'n_estimators': 200}	Index(['L17147-72-00005', 'L08116-72-00002', 'L17167-72-00001', 'CC', 'CG', 'CI', 'CY', 'DG', 'V3', 'V4', 'V5', 'WS', 'Z4_ZC'], dtype='object')	[0.18637364783842686, 0.3222487425925515, 0.14958603744522372, 0.020835486843415556, 0.04535433653737907, 0.04030745986693417, 0.022912719836397706, 0.03707351973811048, 0.034709530367914354, 0.024517014472753805, 0.021524023303988612, 0.05408893555498496, 0.040468545601915595]	0.212275	13.210653	1 days 00:00:00
8	ExtraTreesRegressor	{'max_depth': 25, 'max_features': 'sqrt', 'n_estimators': 200}	Index(['L17147-72-00005', 'L08116-72-00002', 'L17167-72-00001', 'CC', 'CG', 'CI', 'CY', 'DG', 'V3', 'V4', 'V5', 'WS', 'Z4_ZC'], dtype='object')	[0.19001967081291174, 0.3097086232019651, 0.15918234340021806, 0.019787082332750124, 0.046080301333495124, 0.040306778489190294, 0.024815541978308788, 0.03535710308677486, 0.03436205943721931, 0.02429961159931856, 0.02149427941642534, 0.05443572685590247, 0.04015087805552039]	0.328382	33.919703	1 days 00:00:00

Figura 38: Resultats de les diferents prediccions del Extra Trees (Part alta)

Extra Trees com XGBoost ha tornat a trobar valors òptims molt estables, l'únic que ha canviat ha sigut el numero d'estimadors en una predicció.

Pel que fa a la importància de les estacions torna a ser molt semblant al Random Forest, les tres primeres estacions son les de l'ACA sent la pròpia que es vol predir la primera. Les següents tres estacions tornen a ser WS, CG i CI, que es consoliden com a importants, tot i que la Z4_ZC també té força pes en dos dels tres models. Aquesta estació del SMC correspon a Ulldeter, just on neix el riu Ter.

En aquest cas la importància relativa de les estacions del SMC tornen a ser als nivells del *Random Forest,* un 17% d'importància en la predicció. Les estacions de l'ACA tenen una importància d'un 50% i un 100% la pròpia que volem predir.

L'arrel de l'error quadràtic mig és 22,9 el millor valor de moment però molt millorable també i l'*estandard deviation* és 8,49 entre aquest tres.

A la figura 46 dels annexes hi ha les gràfiques de les diferents prediccions i de la importància de cada variable d'aquest model

5.1.4 Regressió Lineal

9	LinearRegression	no avaliable	Index(['L17147-72-00005', 'L08116-72-00002', 'L17167-72-00001', 'CC', 'CG', 'CI', 'CY', 'DG', 'V3', 'V4', 'V5', 'WS', 'Z4_ZC'], dtype='object')	no avaliable	0.268119	23.106667	00
10	LinearRegression	no avaliable	Index(['L17147-72-00005', 'L08116-72-00002', 'L17167-72-00001', 'CC', 'CG', 'CI', 'CY', 'DG', 'V3', 'V4', 'V5', 'WS', 'Z4_ZC'], dtype='object')	no avaliable	0.157734	13.660338	00
11	LinearRegression	no avaliable	Index(['L17147-72-00005', 'L08116-72-00002', 'L17167-72-00001', 'CC', 'CG', 'CI', 'CY', 'DG', 'V3', 'V4', 'V5', 'WS', 'Z4_ZC'], dtype='object')	no avaliable	0.317393	34.196068	00

Figura 39: Resultats de les diferents prediccions de la Regressió Lineal (Part alta)

La diferència de la Regressió lineal amb els models anteriors és la seva senzillesa. En aquest treball, considerem que aquest tipus de model és un *baseline* o referència. S'espera que doni prediccions molt pitjors que les dels altres models. La realitat és que mes o menys el resultat és el mateix. Com a mètode més senzill, no es pot fer servir la mateixa tècnica per mesurar la importància de les variables i tampoc es poden modificar els paràmetres interns.

L'arrel de l'error quadràtic mig és 23,65, un valor acceptable pel que fa a una regressió lineal. Si comparem aquest valor amb la resta de models, no és el model amb pitjor puntuació. És el penúltim per davant del *XGBoost,* tot i que la diferència entre les puntuacions son quasi inexistents. L'*estandard deviation* de les tres puntuacions és de 8,39.

A la figura 47 dels annexes hi han les gràfiques d'una predicció per aquest model.

5.2 Part baixa

El conjunt de dades que utilitzo pel càlcul dels diferents models en la part baixa del riu serà sempre el mateix, tal i com s'explica a l'apartat 4.1.2.

5.2.1 Random Forest

	Method	Best_Params	Features	Feature_Importance	Score	RMSE_Score	Prediccio_Hores
0	Random Forest	{'max_depth': 8, 'max_features': 'sqrt', 'n_estimators': 200}	Index(['L17038-72-00002', 'L17079-72-00004', 'L17079-72-00005', 'F001242', 'F026458', 'L17055-72-00002', 'DJ', 'DN', 'KE', 'UN', 'UO', 'VN', 'WS', 'DM_XJ_WF'], dtype='object')	[0.07332131742887955, 0.1094348875091361, 0.2470332311102765, 0.17381177096317135, 0.05307976132026423, 0.23322296831829073, 0.011807630495869316, 0.01063678029409144, 0.009606477546357907, 0.013142729029663335, 0.010817011799825086, 0.013573341240926795, 0.03557049567987558, 0.00494159726337226]	0.477258	24.974293	1 days 00:00:00
1	Random Forest	{'max_depth': 8, 'max_features': 'sqrt', 'n_estimators': 100}	Index(['L17038-72-00002', 'L17079-72-00004', 'L17079-72-00005', 'F001242', 'F026458', 'L17055-72-00002', 'DJ', 'DN', 'KE', 'UN', 'UO', 'VN', 'WS', 'DM_XJ_WF'], dtype='object')	[0.08524677557708253, 0.11598632298328815, 0.18323974498401424, 0.1942135373940984, 0.03659829061251549, 0.21732611114604083, 0.030387889744812963, 0.016048167507298542, 0.018474851514692192, 0.01614319717997281, 0.015371911486699807, 0.011593158209051078, 0.05014065408037358, 0.00922938758005931]	0.291348	5.752206	1 days 00:00:00
2	Random Forest	{'max_depth': 10, 'max_features': 'sqrt', 'n_estimators': 50}	Index(['L17038-72-00002', 'L17079-72-00004', 'L17079-72-00005', 'F001242', 'F026458', 'L17055-72-00002', 'DJ', 'DN', 'KE', 'UN', 'UO', 'VN', 'WS', 'DM_XJ_WF'], dtype='object')	[0.08794098041124816, 0.12174101374077478, 0.17891128579975726, 0.20505892198558553, 0.049851984042473325, 0.19327357003696247, 0.024092426418034323, 0.015910338600829077, 0.018080666473076122, 0.011945183295723884, 0.014930143401600307, 0.012918120446677707, 0.0519291841013086, 0.013416181245948502]	0.476924	41.50029	1 days 00:00:00

Figura 40: Resultats de les diferents prediccions del Random forest (Part baixa)

Els paràmetres òptims del model que s'han trobat no són constants en les tres proves: el número d'estimadors és completament diferent en totes, però en canvi el màxim de característiques és igual. Aquests resultats són molt semblants al mateix model però a la part alta.

Pel que fa a la importància de les variables predictores, en les tres proves s'ha trobat com a variables més importants les diferents estacions de l'ACA. Les tres més importants són les variables de cabal: L17079-72-00005 → riu Ter a Girona, L17055-72-00002 → riu Ter a Colomers i F001242 → riu Ter a la sortida de Pasteral. La primera estació del SMC amb més importància és la WS → Viladrau, sembla que aquesta estació té importància en el curs alt i baix del riu.

Com es veu a les primeres tres gràfiques de la figura 49 dels annexes, l'importància relativa d'aquesta estació del SMC és d'un 23% aproximadament. Si la comparem amb les estacions de l'ACA, té el mateix pes que l'estació amb menys importància de l'ACA. Això representa una influència sobre el model baixa.

L'arrel de l'error quadràtic mig és 24,07, es veu com aquest valor canvia molt en les tres prediccions diferents, això s'explica al punt 5.3. L'*standard deviation* es 14,60 entre aquest tres.

A la figura 49 dels annexes hi ha les gràfiques de les diferents prediccions i de la importància de cada variable d'aquest model

5.2.2 XGBoost

3	XGBoost Regressor	{'max_depth': None, 'max_features': 1, 'n_estimators': 50}	Index(['L17038-72-00002', 'L17079-72-00004', 'L17079-72-00005', 'F001242', 'F026458', 'L17055-72-00002', 'DJ', 'DN', 'KE', 'UN', 'UO', 'VN', 'WS', 'DM_XJ_WF'], dtype='object')	[0.041147955, 0.028350184, 0.35307893, 0.04635057, 0.102162465, 0.19009347, 0.029886687, 0.013195884, 0.014855327, 0.02411344, 0.01239461, 0.030281143, 0.10668811, 0.0074011562]	0.477991	24.956767	1 days 00:00:00
4	XGBoost Regressor	{'max_depth': None, 'max_features': 1, 'n_estimators': 50}	Index(['L17038-72-00002', 'L17079-72-00004', 'L17079-72-00005', 'F001242', 'F026458', 'L17055-72-00002', 'DJ', 'DN', 'KE', 'UN', 'UO', 'VN', 'WS', 'DM_XJ_WF'], dtype='object')	[0.028389333, 0.02671366, 0.052381046, 0.07960747, 0.039688144, 0.48747867, 0.06882528, 0.006780341, 0.021633081, 0.026082003, 0.012871952, 0.010208263, 0.13060136, 0.008739443]	0.175195	6.205739	1 days 00:00:00
5	XGBoost Regressor	{'max_depth': None, 'max_features': 1, 'n_estimators': 50}	Index(['L17038-72-00002', 'L17079-72-00004', 'L17079-72-00005', 'F001242', 'F026458', 'L17055-72-00002', 'DJ', 'DN', 'KE', 'UN', 'UO', 'VN', 'WS', 'DM_XJ_WF'], dtype='object')	[0.031528514, 0.034335215, 0.06343943, 0.09112481, 0.041752096, 0.4513236, 0.062842004, 0.009211462, 0.03815178, 0.02697609, 0.014046729, 0.00946575, 0.11826197, 0.00754054]	0.394202	44.661432	1 days 00:00:00

Figura 41: Resultats de les diferents prediccions del XGBoost (Part baixa)

Com ja ha passat amb la predicció de la part alta, el model XGBoost arriba a uns valors més estables en trobar els millors paràmetres per la predicció. En aquest cas en els tres conjunts de dades ha trobat que els millors paràmetres per la predicció han sigut els mateixos.

L'estació de Colomers (ACA) torna a ser important per la predicció del model. Colomers és la mateixa estació que volem predir 24h més tard. De la resta d'estacions importants, les de Viladrau (SMC) i de Pasteral (ACA) tornen a sortir prou importants i s'introdueix la del SMC: DJ → a Banyoles i de l'ACA: Riudellots d'Onyar (el riu Onyar).

En aquest model les estacions de pluja tornen a agafar la mateixa importància que les de cabal per fer les prediccions, amb l'excepció que la de Colomers (ACA). Com es veu a la figura 50 dels annexes, l'estació de Viladrau (Part sud de la conca del Ter, SMC) és la segona estació amb la importància relativa més alta, un 25%. La de Banyoles (Part nord de la part baixa del riu) té una importància relativa de 15%. Tot i que no és un nombre molt elevat, té més importància que les altres estacions de l'ACA.

L'arrel de l'error quadràtic mig és 25,27. Hi ha un patró en els mateixos resultats del valor RMSE: el segon conjunt de dades sempre té el valor més petit, l'últim el més gran i el primer un valor intermig (S'explica al apartat 5.3). L'*standard deviation* es 15,70 entre aquest tres.

A la figura 50 dels annexes hi ha les gràfiques de les diferents prediccions i de la importància de cada variable d'aquest model

5.2.3 Extra Trees

6	ExtraTreesRegressor	{'max_depth': 8, 'max_features': 'log2', 'n_estimators': 100}	Index(['L17038-72-00002', 'L17079-72-00004', 'L17079-72-00005', 'F001242', 'F026458', 'L17055-72-00002', 'DJ', 'DN', 'KE', 'UN', 'UO', 'VN', 'WS', 'DM_XJ_WF'], dtype='object')	[0.0740732537592477, 0.08212981180876483, 0.2874272825419713, 0.19220583070590183, 0.029547625560140425, 0.25536092509694985, 0.00826944717506329, 0.006680108982100367, 0.010909678268381316, 0.006983311318302996, 0.00849458114103793, 0.00895071079384179, 0.025590583457340107, 0.00337684939095631 3]	0.419278	26.322891	1 days 00:00:00
7	ExtraTreesRegressor	{'max_depth': 25, 'max_features': 1, 'n_estimators': 200}	Index(['L17038-72-00002', 'L17079-72-00004', 'L17079-72-00005', 'F001242', 'F026458', 'L17055-72-00002', 'DJ', 'DN', 'KE', 'UN', 'UO', 'VN', 'WS', 'DM_XJ_WF'], dtype='object')	[0.06999670589242553, 0.08685952489670373, 0.2106705103268817, 0.16333295413595766, 0.0408436528884432, 0.19892983884002544, 0.03551128063759883, 0.023970881680025802, 0.026790862607231264, 0.02239749007456403, 0.02881626011282192, 0.022586395276697577, 0.04824437522995096, 0.021049267400672398]	0.172645	6.215323	1 days 00:00:00
8	ExtraTreesRegressor	{'max_depth': 25, 'max_features': 'log2', 'n_estimators': 200}	Index(['L17038-72-00002', 'L17079-72-00004', 'L17079-72-00005', 'F001242', 'F026458', 'L17055-72-00002', 'DJ', 'DN', 'KE', 'UN', 'UO', 'VN', 'WS', 'DM_XJ_WF'], dtype='object')	[0.0749323498989874, 0.08561466676697271, 0.193111637225571, 0.17604492038532374, 0.03752015249221338, 0.21259447402701548, 0.03545332281407616, 0.018879899476574508, 0.02487372401803699, 0.019835565709902782, 0.025118641301514814, 0.01968497466882637, 0.057114970662552535, 0.019220700552432093]	0.486602	41.114562	1 days 00:00:00

Figura 42: Resultats de les diferents prediccions del Extra Trees (Part baixa)

Comparant els millors paràmetres amb el mateix model però per la part alta, aquest conjunt no ha aconseguit trobar uns valors tant estables com l'altre.

Els resultats de les tres primeres millor variables, que ha extret Extra Trees, són molt semblants al que ha extret el model Random Forest. En les tres proves, la quarta i cinquena variable més important són les mateixes del ACA: L17079-72-00004 → Girona (riu Onyar) i L17038-72-00002 → Campllong (riera Gotarra).

En aquest cas la importància relativa de les estacions del SMC tornen a ser als nivells del *Random Forest,* un 15% d'importància en la predicció. La resta d'estacions de l'ACA tenen una importàncies de: 100%, 90%, 80%, 40% i 35%.

L'arrel de l'error quadràtic mig és 24,54 i l'*estandard deviation* es 14,30 entre aquest tres.

A la figura 51 dels annexes hi ha les gràfiques de les diferents prediccions i de la importància de cada variable d'aquest model

5.2.4 Regressió Lineal

9	LinearRegression	no avaliable	Index(['L17038-72-00002', 'L17079-72-00004', 'L17079-72-00005', 'F001242', 'F026458', 'L17055-72-00002', 'DJ', 'DN', 'KE', 'UN', 'UO', 'VN', 'WS', 'DM_XJ_WF'], dtype='object')	no avaliable	0.41392	26.444029	1 days 00:00:00
10	LinearRegression	no avaliable	Index(['L17038-72-00002', 'L17079-72-00004', 'L17079-72-00005', 'F001242', 'F026458', 'L17055-72-00002', 'DJ', 'DN', 'KE', 'UN', 'UO', 'VN', 'WS', 'DM_XJ_WF'], dtype='object')	no avaliable	-0.31172	7.825974	1 days 00:00:00
11	LinearRegression	no avaliable	Index(['L17038-72-00002', 'L17079-72-00004', 'L17079-72-00005', 'F001242', 'F026458', 'L17055-72-00002', 'DJ', 'DN', 'KE', 'UN', 'UO', 'VN', 'WS', 'DM_XJ_WF'], dtype='object')	no avaliable	0.424033	43.547936	1 days 00:00:00

Figura 43: Resultats de les diferents prediccions de la Regressió Lineal (Part baixa)

L'arrel de l'error quadràtic mig de la Regressió Lineal a la part baixa del riu és 25,92. Si comparem aquest valor amb la resta de models, és el model amb pitjor puntuació. A pesar de tot la diferència entre les puntuacions de son quasi inexistents, ja que el model que està per sobre té una puntuació de 25,27. L'estandard deviation es 14.58 entre aquest tres.

A la figura 48 dels annexes hi han les gràfiques d'una predicció per aquest model.

5.3 Valor RMSE

Com es pot observar en els valors RMSE dels diferents models de la part baixa i alta del riu, sempre segueixen el mateix patró. La segona iteració sempre té el valor més petit, la tercera el més gran i la primera l'entremig.

Fem un plot de les dades de test que utilitzem en les dues parts del riu per calcular el valor RMSE. Aquest plot és el que està reflectit a les figures 54 i 55 del annex.

S'observa que el patró té relació amb el número i els valors de les avingudes, que tenen els diferents subconjunts de dades de test. Com més avingudes amb uns valors més alts, més alt serà la puntuació RMSE. Això passa perquè els nostres models tenen mancances a l'hora de predir avingudes. Llavors, com més avingudes s'hagin de predir més alt el valor RMSE.

6. Conclusions i treballs futurs

En aquest treball s'ha presentat una aproximació basada en ML per a la predicció del cabal del riu Ter. Barrejant dades històriques del riu i de meteorologia, hem construït models per a dos sectors del riu diferents que ens han aportat nova informació a la problemàtica de les avingudes. Una informació que ajuda a entendre una mica millor el comportament del riu en els dos sectors.

En la realització d'aquest projecte he pogut introduir-me en el món de la Data4Good i el de les dades Obertes. Crec que és un moviment molt potent i interessant, per això animo i dono gràcies a les administracions que de forma desinteressada cedeixen dades per projectes com aquest. Aconseguir les dades molts cops és bastant complicat i requereix de temps, per això s'haurien d'agilitzar els diferents processos.

Una oportunitat convertida en fortalesa d'aquest treball ha sigut treballar en un problema real. Això m'ha fet adonar de la complexitat que té la creació d'un model de *Machine Learning* desde zero en el món real. Sobretot la part més farragosa i menys valorada, el preprocessament de les dades, una feina que s'emporta el 80% dels esforços i que quan l'acabes encara has de començar tota la part de modelatge.

Valoro molt positivament l'oportunitat que m'ha brindat aquest projecte en aprendre des de la base *Machine Learning.* Els conceptes que he adquirit a base de l'autoaprenentatge podrien ser molt comparables al d'una assignatura sencera de la Universitat.

Parlant dels objectius que m'havia plantejat a l'inici del projecte, la gran majoria els he passat satisfactòriament: l'obtenció i preprocessament de dades, la creació dels models i aportar informació rellevant sobre el comportament del riu. També sé que els models no han acabat tenint la precisió que es volia en un principi per un problema *d'underfitting*. *Underfitting* és quan un model li costa trobar la relació entre les X i y generant un error prou gran a l'hora de fer les prediccions.

Tant a la part alta com a la baixa del riu, diverses estacions s'han anat repetint constantment com a importants, en alguns models més i en altres menys. Aquestes puc assegurar que tenen una relació directa amb el cabal del riu en els dos punts que predim:

De la part alta són:

- ACA: les tres estacions que li passem al model: Ripoll, Sant Joan de les Abadesses i Masies de Roda. És d'esperar que el cabal a Masies 24h despés depengui dels diferents cabals del riu.

- SMC: hi ha tres estacions més significatives que les altres: la de Viladrau (Part sud de la part alta del riu), Mollo-Fabert (Part nord tocant a França) i Sant Pau de Segúries (Part nord molt a prop de Camprodon)

De la part baixa són:

- ACA: La majoria d'estacions son importants però les tres que ho són més són: el cabal a Girona (riu Onyar), el cabal a la sortida de Pasteral i la pròpia estació de cabal a Colomers que hem intentat predir.

- SMC: Entre les diferents estacions meteorològiques n'hi ha dos que destaquen per sobre les altes: l'estació que correspon a Viladrau (Inici de la part baixa) i a Banyoles (part intermitja de la part baixa del riu, les precipitacions baixen pel riu Terri)

Un cop finalitzada aquesta part del treball, s'obren una infinitat de portes per la millora dels diferents models. Durant la realització del mateix m'han anat sortint idees sobre possibles millores, que no he pogut dur a terme per falta de temps. Entre d'altres:

- Les prediccions que estem calculant amb els models actuals, tant en el curs alt com en el baix del riu, són a 24 hores. Això vol dir que per predir el cabal actual en els diferents punts, recollim les dades de totes les estacions 24 hores abans. Però realment té sentit predir a 24 hores? Si donéssim la llibertat a l'aprenentatge automàtic perquè triés, per cada estació, amb quant de temps ha de recollir les dades. Podríem millorar considerablement els resultats.

  Perquè? Tarda igual a arribar l'aigua que neix de Ulldeter amb l'aigua que prové d'una riera dels curs mig? Quan tarda l'aigua de la pluja a transformar-se en cabal en cada una de les estacions del SMC?. Tot això es podria trobar amb les dades que disposem.

- Per poder millorar una mica més el problema *underfitting* que tenim al nostre model, es podrien crear noves característiques (*features*) a partir de les existents. Això aumentaria la complexitat del model i com a conseqüència les seves prediccions serien més acurades.

- Una altre possibilitat que podria donar bons resultats podria ser entrenar el model només amb les files de dades que hi hagi temporals. Si el que realment interessa es el comportament del riu en episodis d'avinguda, seria interessant entrenar-ho en aquets episodis. Quan hi ha episodis d'aquest estil les dades de les estacions meteorològiques tenen valors interessants que potser tindrien més pes en el model

final. Hi ha tot un camp del *Machine Learning* que treballa tot això: *anomaly detection*. Es podria aplicar tècniques d'aquesta part del *ML* en aquest problema.

Com es pot observar, aquest treball obre les portes a una problemàtica força rellevant del nostre territori. Nosaltres hem iniciat aquest projecte amb moltes ganes però encara pot tenir molt més recorregut. T'animem a tu lector a agafar el relleu del projecte i sentir-te lliure de preguntar-me qualsevol cosa.

7. Referències

[1] Agència Catalana de l'Aigua. (2011, December 20). *RIUADES DEL RIU TER I RITORT A CAMPRODON - 18 d'octubre de 1940*. ACA. Retrieved May 29, 2022, from http://aca-web.gencat.cat/sig/fitxes/espais_fluvials/mat/aca_mat_17039_1940b_v1.pdf

[2] ASSOCIACIÓ CULTURAL LA LLERA. (2008). El problema de l'aigua i el riu Ter. *El problema de l'aigua i el riu Ter*, -(47), 4-11. https://www.celra.cat/ajuntament/documents/lallera47.pdf

[3] Buisán, L. (2019, May 27). *Un investigador de la Universidad del País Vasco, ganador del "Aguathón"*. Instituto Tecnológico de Aragón. Retrieved May 30, 2022, from https://www.itainnova.es/blog/noticias/un-investigador-de-la-universidad-del-pais-vasco-ganador-del-aguathon/

[4] Carreras, T. (2022, April 19). Comença l'obra del mur del Ter contra les inundacions a Fontajau. *Diari de Girona*. https://www.diaridegirona.cat/girona/2022/04/19/comenca-l-obra-mur-ter-65127023.html

[5] Costa, J. M. (2015, October 18). FOTOS El gran aiguat del Ter, 75 anys després. *NacióDigital*. https://www.naciodigital.cat/osona/noticia/48095/fotos-gran-aiguat-ter-75-anys-despres

[6] elTriangle. (2020, January 23). El agua del Ter inunda Girona por el temporal Gloria. *El triangle*. https://www.eltriangle.eu/es/2020/01/23/noticia-es-104787/

[7] Grup Enciclopèdia. (n.d.). *el Ter | enciclopedia.cat*. Enciclopèdia.cat. Retrieved May 27, 2022, from https://www.enciclopedia.cat/gran-enciclopedia-catalana/el-ter

[8] Grup enciclopèdia. (n.d.). *interpolació | enciclopedia.cat*. Enciclopèdia.cat. Retrieved June 3, 2022, from https://www.enciclopedia.cat/gran-enciclopedia-catalana/interpolacio-3

[9] Martinez, J. (2020, September 19). *Aguathon: mi solución al primer Hackathon del Agua*. IArtificial.net. Retrieved May 29, 2022, from https://www.iartificial.net/aguathon-mi-solucion-al-primer-hackathon-del-agua/#Tecnicas_de_Machine_Learning

[10] MathWorks. (n.d.). *¿Qué es la regresión lineal? - MATLAB & Simulink*. MathWorks. Retrieved June 10, 2022, from https://es.mathworks.com/discovery/linear-regression.html

[11] Petit, A. (2016, July 3). La batalla perduda per l´aigua del riu Ter. *Diari de Girona*. https://www.diaridegirona.cat/dominical/2016/07/03/batalla-perduda-per-l-aigua-49120929.html

[12] Rodríguez, M. (2020, January 23). El Ter inunda diferents zones de Girona | Catalunya | EL PAÍS Catalunya. *Elpais.cat*. https://cat.elpais.com/cat/2020/01/23/catalunya/1579810582_789729.html

[13] Wikipedia. (n.d.). *Ter*. Viquipèdia. Retrieved May 27, 2022, from https://ca.wikipedia.org/wiki/Ter#Hist%C3%B2ria

[14] Breiman, “Random Forests”, Machine Learning, 45(1), 5-32, 2001.

[15] P. Geurts, D. Ernst., and L. Wehenkel, “Extremely randomized trees”, Machine Learning, 63(1), 3-42, 2006.

[16] Stef van Buuren, Karin Groothuis-Oudshoorn (2011). “mice: Multivariate Imputation by Chained Equations in R”. Journal of Statistical Software 45: 1-67.

[17] S. F. Buck, (1960). “A Method of Estimation of Missing Values in Multivariate Data Suitable for use with an Electronic Computer”. Journal of the Royal Statistical Society 22(2): 302-306.

8. Annexes

8.1 Taules

Nom	Codi	Comarca	Riu	Superfície conca drenada	Terme municipal	Posició
Torelló_Ges	F009891	OSONA	RIU GES	88,02 km²	TORELLÓ	42.049547207 2.2646119
Masies de Roda	L08116-72-00002	OSONA	RIU TER	1387,05 km²	MASIES DE RODA, LES	41.983802 2.302718379
Ripoll	L17147-72-00005	RIPOLLÈS	RIU TER	736,91 km²	RIPOLL	42.172192162 2.194229581
Sant Joan de les Abadesses	L17167-72-00001	RIPOLLÈS	RIU TER	301,02 km²	SANT JOAN DE LES ABADESSES	42.221664291 2.243255294
Anglès	F014672	SELVA	RIERA D'OSOR	88,00 km²	ANGLÈS	41.961160899 2.630484738
Ginestar_Llèmena	F000005	GIRONÈS	RIERA DE LLEMANA	77,19 km²	SANT GREGORI	42.015521388 2.725462811
Campllong_Gotarra	L17038-72-00002	GIRONÈS	RIERA GOTARRA	100,43 km	CAMPLLONG	41.897979761 2.818842216
Girona_Onyar	L17079-72-00004	GIRONÈS	RIU ONYAR	321,79 km²	GIRONA	41.975059495 2.824550321
Riudellots_Onyar	F026458	SELVA	RIU ONYAR	117,02 km²	RIUDELLOTS DE LA SELVA	41.898318267 2.816478373
Colomers	L17055-72-00002	BAIX EMPORDÀ	RIU TER	2901,95 km²	COLOMERS	42.076860022 2.991162733
Girona_Ter	L17079-72-00005	GIRONÈS	RIU TER	2257,00 km²	GIRONA	41.990209392 2.819160445
Pasteral_Cabal	F001242	SELVA	RIU TER	1799,32 km²	AMER	41.985432993 2.603218761
Torroella de Montgrí	L17199-72-00001	BAIX EMPORDÀ	RIU TER	2948,48 km²	TORROELLA DE MONTGRÍ	42.035163473 3.125380537
Conellà Terri	F001243	PLA DE L'ESTANY	RIU TERRI	23,78 km²	CORNELLÀ DEL TERRI	42.100848824 2.798988336

Taula 1: Conjunt d'estacions de l'ACA d'on he extret dades

Nom	Codi	Georeferència	Municipi	Comarca	Estat	Alta	Baixa
Serra de Daró	UD	POINT (42.02895 3.06227)	Serra de Daró	Baix Empordà	Desmantellada	06/09/1999 12:00:00 AM	06/01/2013 12:00:00 AM
la Tallada d'Empordà	UB	POINT (42.05398 3.06195)	La Tallada d'Empordà	Baix Empordà	Operativa	05/01/1989 12:00:00 AM	
Girona - Bombers	DM	POINT (41.96012 2.80696)	Girona	Gironès	Desmantellada	05/10/2001 12:00:00 AM	09/15/2010 12:00:00 AM
Vilablareix	WF	POINT (41.95425 2.77578)	Vilablareix	Gironès	Desmantellada	04/11/2001 12:00:00 AM	10/01/2015 12:00:00 AM
Cassà de la Selva	UN	POINT (41.87449 2.92694)	Cassà de la Selva	Gironès	Operativa	03/08/1993 12:00:00 AM	
Fornells de la Selva	UO	POINT (41.91461 2.82069)	Fornells de la Selva	Gironès	Operativa	11/10/1998 12:00:00 AM	
Girona	XJ	POINT (41.98223 2.80686)	Girona	Gironès	Operativa	09/15/2010 12:00:00 AM	
Vic - 1	CX	POINT (41.93582 2.23857)	Vic	Osona	Desmantellada	01/18/1996 12:00:00 AM	01/01/2003 12:00:00 AM
Viladrau - Aigües de Viladrau	V6	POINT (41.84256 2.42636)	Viladrau	Osona	Desmantellada	10/03/1995 12:00:00 AM	03/29/2005 12:00:00 AM
Embassament de Sau	V7	POINT (41.97554 2.39837)	Vilanova de Sau	Osona	Desmantellada	05/20/1997 12:00:00 AM	05/02/2006 12:00:00 AM
Muntanyola	CY	POINT (41.87813 2.17873)	Muntanyola	Osona	Operativa	01/12/1996 12:00:00 AM	
Orís	CC	POINT (42.07398 2.20862)	Orís	Osona	Operativa	11/15/1995 12:00:00 AM	
Viladrau	WS	POINT (41.84008 2.41877)	Viladrau	Osona	Operativa	03/17/2005 12:00:00 AM	
Vic	XO	POINT (41.93497 2.23987)	Vic	Osona	Operativa	12/22/2011 12:00:00 AM	
Montesquiu	V4	POINT (42.11477 2.21483)	Montesquiu	Osona	Operativa	12/02/1998 12:00:00 AM	
Gurb	V3	POINT (41.95224 2.23271)	Gurb	Osona	Operativa	03/31/1999 12:00:00 AM	
Pantà de Sau	KE	POINT (41.96867 2.41404)	Vilanova de Sau	Osona	Operativa	01/05/1996 12:00:00 AM	
Banyoles	DJ	POINT (42.11653 2.78969)	Banyoles	Pla de l'Estany	Operativa	10/11/1999 12:00:00 AM	
les Llosses	CB	POINT (42.15085 2.19914)	Les Llosses	Ripollès	Desmantellada	11/30/1995 12:00:00 AM	06/02/2003 12:00:00 AM
Ulldeter (2.364 m)	Z4	POINT (42.42205 2.2524)	Setcases	Ripollès	Desmantellada	11/08/2000 12:00:00 AM	09/20/2011 12:00:00 AM
Ulldeter (2.410 m)	ZC	POINT (42.42117 2.24565)	Setcases	Ripollès	Operativa	09/28/2011 12:00:00 AM	
Núria (1.971 m)	DG	POINT (42.39848 2.15517)	Queralbs	Ripollès	Operativa	05/15/1998 12:00:00 AM	

Sant Pau de Segúries	CI	POINT (42.25839 2.36429)	Sant Pau de Segúries	Ripollès	Operativa	11/24/1995 12:00:00 AM	
Sant Joan de les Abadesses	M6	POINT (42.22189 2.2427)	Sant Joan de les Abadesses	Ripollès	Operativa	01/13/1996 12:00:00 AM	
Molló - Fabert	CG	POINT (42.37717 2.41456)	Molló	Ripollès	Operativa	06/06/1996 12:00:00 AM	
Vilobí d'Onyar	VN	POINT (41.88244 2.74262)	Vilobí d'Onyar	Selva	Operativa	11/10/1998 12:00:00 AM	
Anglès	DN	POINT (41.96095 2.63108)	Anglès	Selva	Operativa	05/10/2001 12:00:00 AM	
Perafita	V5	POINT (42.03947 2.11993)	Perafita	Osona	Operativa	07/03/1995 12:00:00 AM	

Taula 2: Conjunt d'estacions del SMC d'on he extret dades

8.2 Imatges

Figura 44 : Gràfiques de les diferents prediccions i de la importància de cada variable (Part alta, Random Forest)

Figura 45 : Gràfiques de les diferents prediccions i de la importància de cada variable (Part alta, XGBoost)

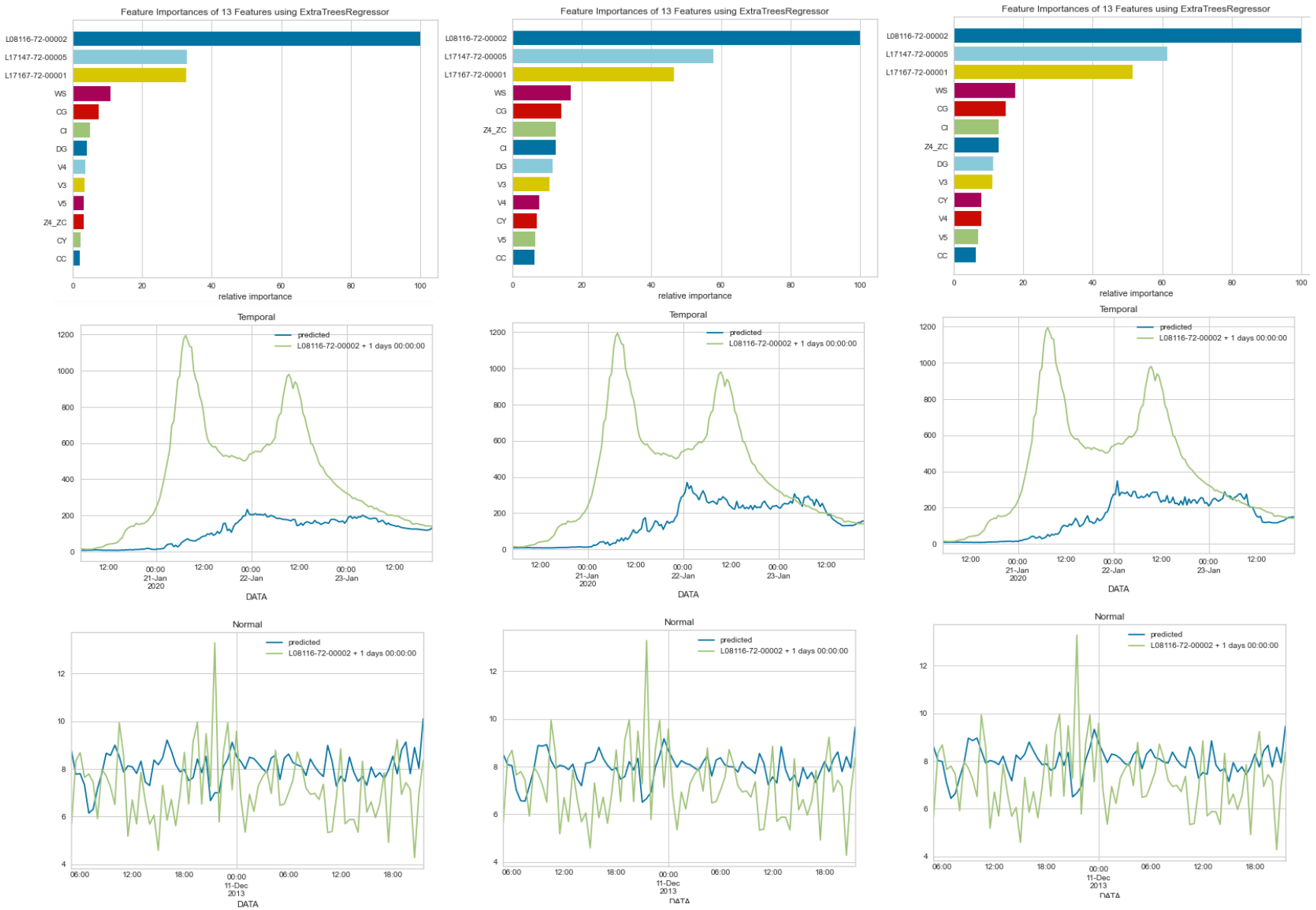

Figura 46 : Gràfiques de les diferents prediccions i de la importància de cada variable (Part alta, Extra Trees)

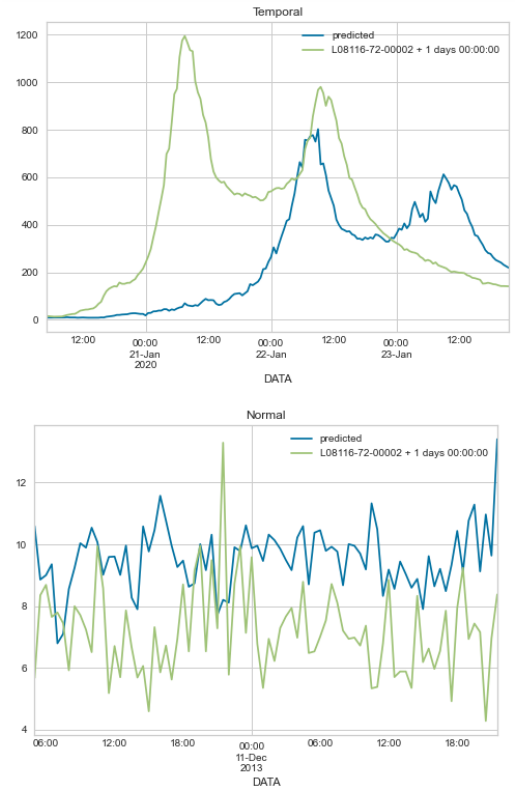

Figura 47 : Gràfiques d'una predicció (Part alta, Regressió Lineal)

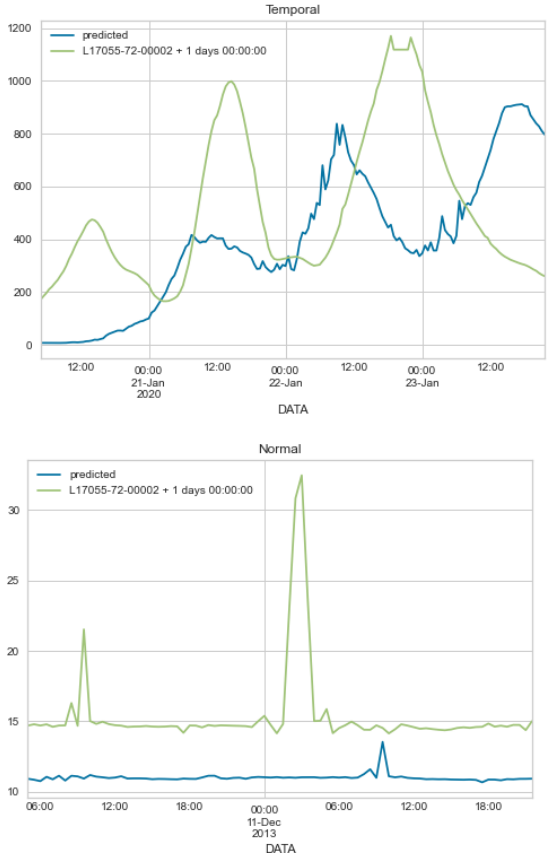

Figura 48 : Gràfiques d'una predicció (Part baixa, Regressió Lineal)

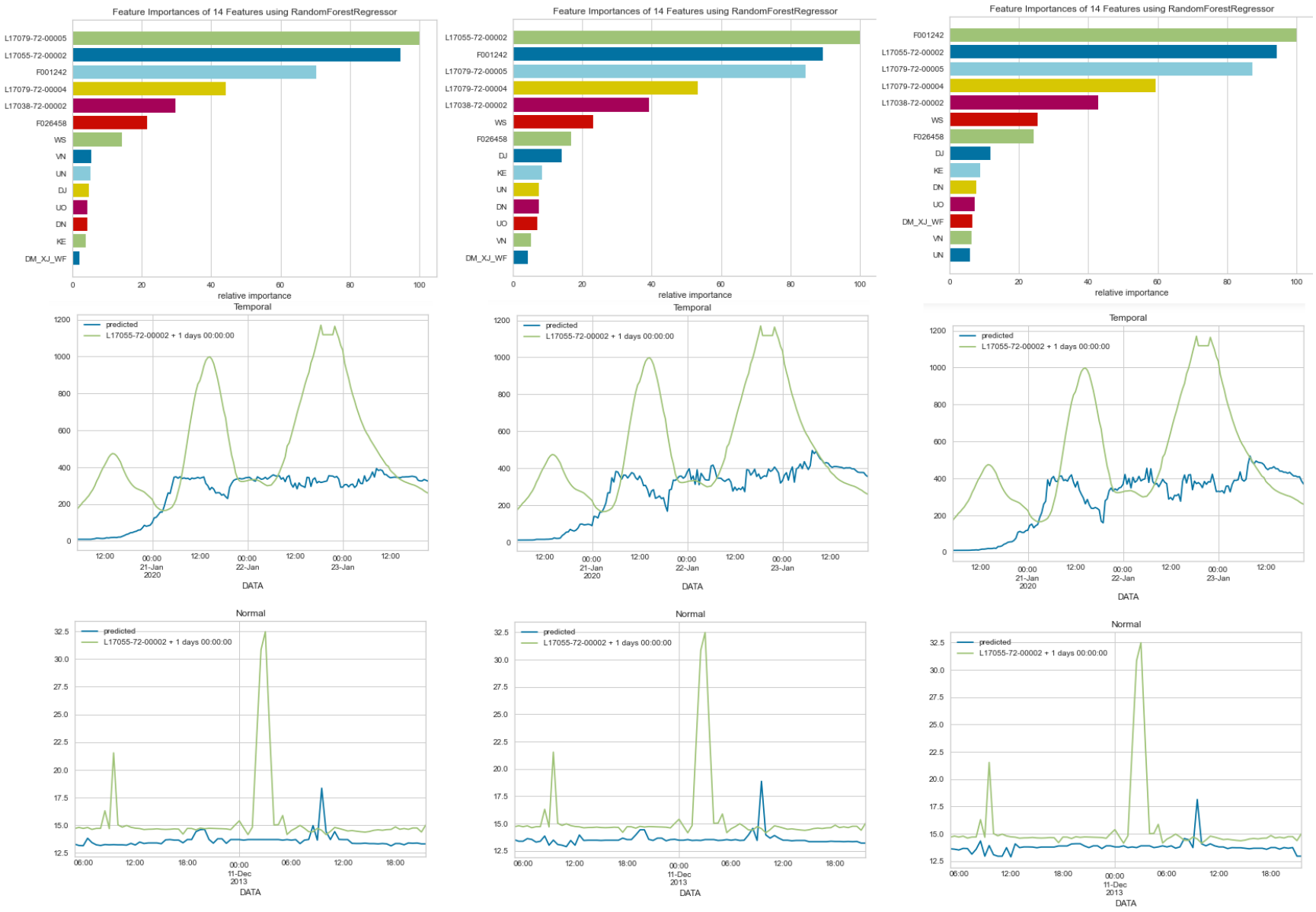

Figura 49 : Gràfiques de les diferents prediccions i de la importància de cada variable (Part baixa, Random Forest)

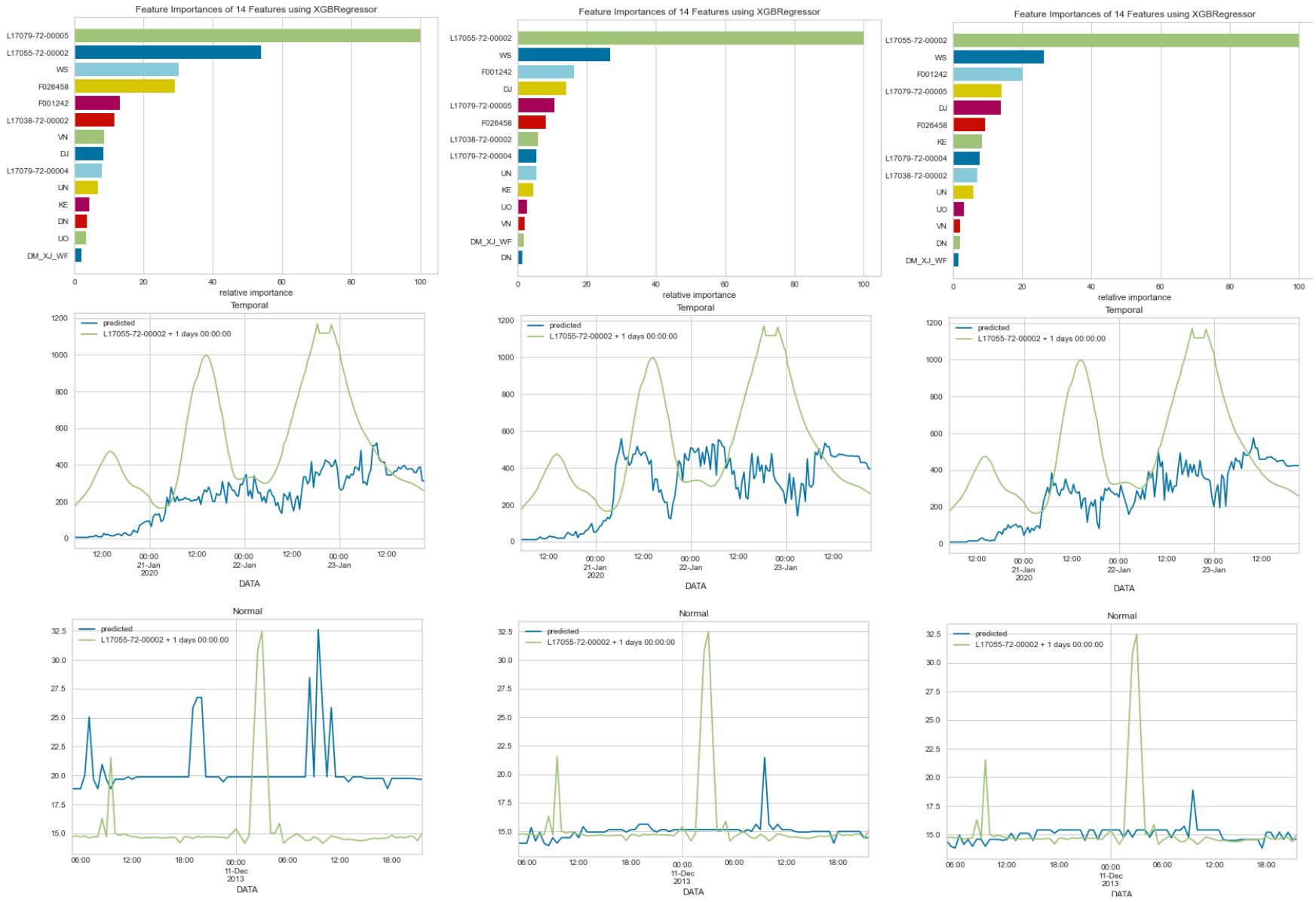

Figura 50 : Gràfiques de les diferents prediccions i de la importància de cada variable (Part baixa, XGBoost)

Figura 51 : Gràfiques de les diferents prediccions i de la importància de cada variable (Part baixa, Extra Trees)

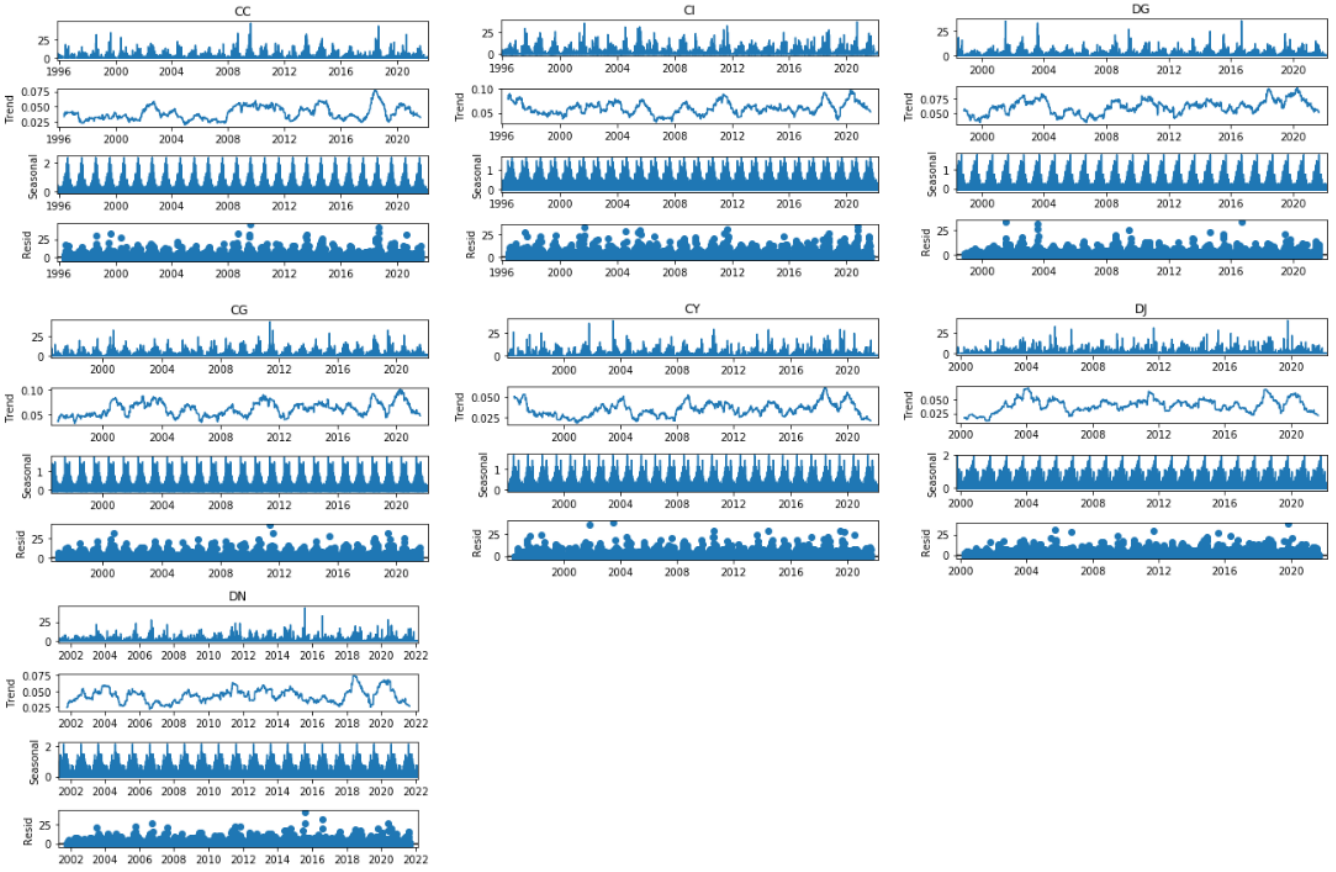

Figura 52: Descomposició temporal de les estacions CC, CI, DG, CG, CY, DJ, DN amb dades desde 2009 fins 2022

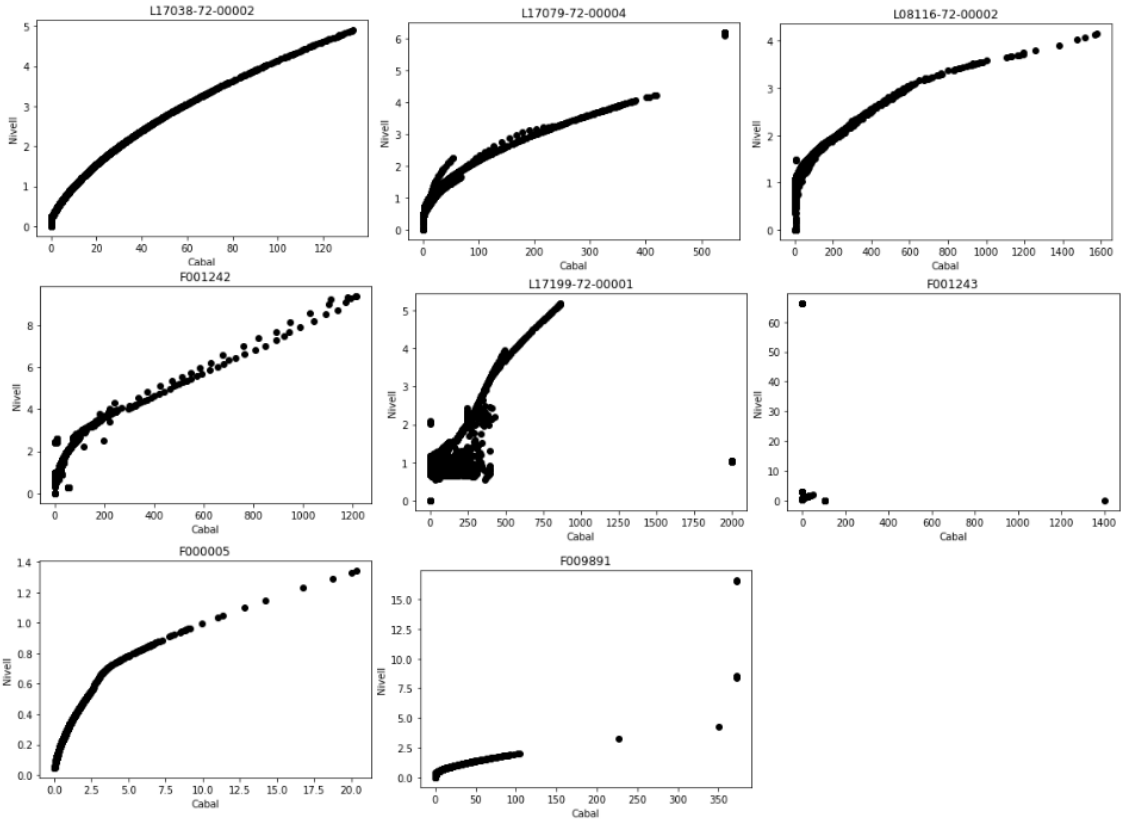

Figura 53: Gràfiques de la relació Cabal - Nivell de la resta d'estacions de l'ACA abans d'executar *movingAverageResolverTotal*

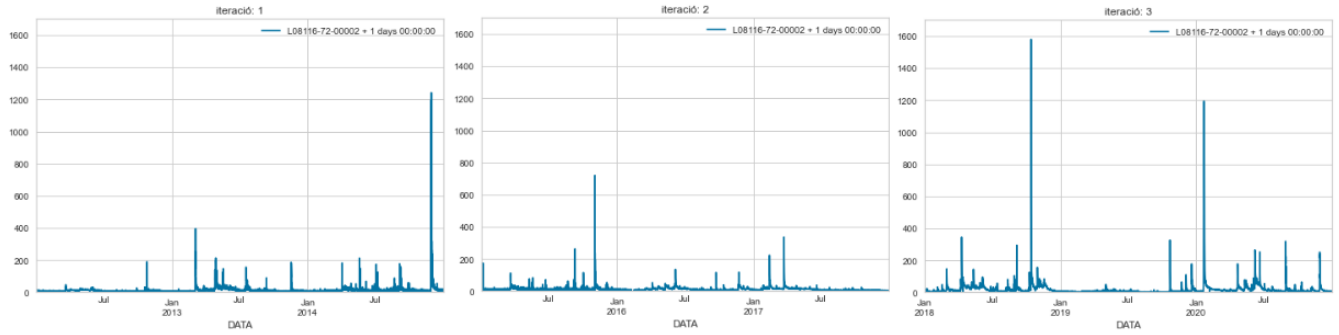

Figura 54: Gràfiques dels valors de test que s'utilitzen per calcular el valor RMSE al alt Ter

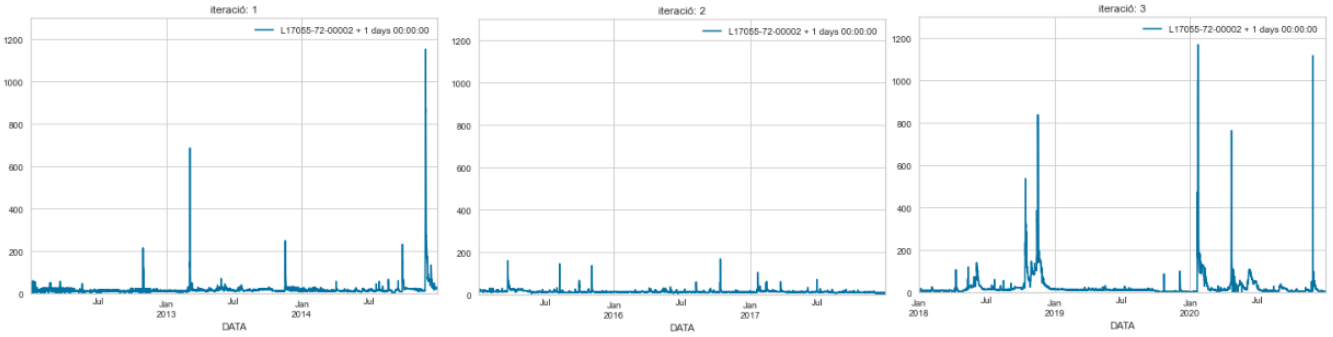

Figura 55: Gràfiques dels valors de test que s'utilitzen per calcular el valor RMSE al baix Ter