

## Treball Final de Màster

**Estudi:** Màster en Ciència de Dades

**Títol:** Detecció de dades anòmales en dades de cabal

**Document:** Memòria

**Alumne:** Joan Saló Grau

**Tutor:** Marc Comas Cufí

**Departament:** Informàtica, Matemàtica Aplicada i Estadística

**Àrea:** Estadística i investigació operativa

**Convocatòria (mes/any):** Setembre 2023

TREBALL FINAL DE MÀSTER

---

DETECCIÓ DE DADES ANÒMALES EN DADES DE  
CABAL

---

AUTOR:

Joan Saló Grau

Girona, Setembre 2023

Màster en Ciència de Dades

TUTOR:

Marc Comas Cufí

# Resum

En aquest treball ens centrarem a detectar les anomalies de les dades majoritàriament causades per un mal funcionament de les estacions d'aforament, que són les estacions que mesuren el cabal d'aigua que circula pels diferents cursos fluvials de les CIC (Conques Internes de Catalunya), a diferència d'altres treballs, on es detecten anomalies relatives a patrons poc freqüents, però no incorrectes, com poden ser episodis de pluja extrema (que causen pics de cabal) o d'extrema sequera (com podria ser el cas actual). D'aquesta manera, el nou conjunt de dades filtrat pot ser utilitzat per entrenar models robustos que ajudin a comprendre els processos hidrològics, i no només això, sinó que també és interessant l'aplicació d'aquest treball pels gestors de les estacions d'aforament, ja que la presència de noves dades anòmales pot indicar un mal funcionament d'aquestes.

Per tal de detectar les anomalies, hem dissenyat dos mètodes: el primer (mètode 1), es basa a modelitzar les dades de cabal utilitzant l'algorisme de modelatge Prophet (dissenyat per Facebook) utilitzant GAM (Model Additiu Generalitzat), i considerar el que estigui fora d'un cert interval de confiança detectat pel mateix algorisme com a anomalia.

El segon mètode (mètode 2) parteix d'un etiquetatge previ de les mostres segons si són anòmales o no. Amb les dades etiquetades, podem modelitzar el cabal de les dades correctes de totes les estacions utilitzant una sola xarxa Long Short-Term Memory (LSTM) però no a partir de les mateixes dades de cabal, sinó a través de variables externes com la precipitació, temperatura, etc., d'aquesta manera, podem arribar a esmenar errors que encara puguem tenir dins les dades. Un cop això, podem entrenar un LSTM Autoencoder, a partir de seqüències de  $n$  observacions i prediccions, on l'última observació és correcta. D'aquesta manera, i basant-se en el supòsit que les prediccions de les mostres correctes seran més exactes que la de les mostres incorrectes, podem dissenyar un classificador que a partir d'una sèrie de prediccions i observacions, ens ajudi a determinar si aquestes observacions són anòmales o no.

A tall de conclusions, els resultats dels algorismes proposats no han sigut satisfactoris, ja que aquests no han sigut capaços de detectar les anomalies, en gran part a causa de la complexitat d'aquestes, les quals no es poden detectar mitjançant mètodes estadístics clàssics, sinó que depenen en gran part del comportament de la sèrie i un conjunt de patrons concrets.

El mètode 1 és el que ha donat pitjors resultats, en gran part a causa de la simplicitat de l'aproximació, i no és capaç de detectar gairebé cap anomalia de forma correcta (menys d'un 1% de les anomalies classificades són correctes). Pel que fa al mètode 2, sí que els resultats han sigut lleugerament positius, i tot i que les prediccions del cabal amb la xarxa LSTM són millorables, i l'error obtingut és elevat (tenim un NSE de 0.2 de mitjana per cada estació), el model entrenat és capaç d'entendre la dinàmica general, tot i que té problemes per predir els pics de cabal. Tot i això, el classificador LSTM Autoencoder no detecta bé les anomalies, però realitzant una

inspecció visual de les prediccions obtingudes, i amb l'ajuda d'un expert, aconseguim detectar algunes anomalies. De totes maneres, caldria continuar treballant en les prediccions provant altres mètodes o integrant més dades per tal de millorar els resultats.

# Continguts

|   |          |
|---|----------|
| Índex de figures                                      | v        |
| Índex de taules                                       | vii      |
| <b>1 Introducció</b>                                  | <b>1</b> |
| <b>2 Literatura relacionada i estat de l'art</b>      | <b>2</b> |
| <b>3 Preliminars</b>                                  | <b>3</b> |
| 1 Dades de cabal . . . . .                            | 3        |
| 2 Sèries temporals . . . . .                          | 7        |
| 3 Prophet . . . . .                                   | 9        |
| 4 Xarxes RNN i LSTM . . . . .                         | 9        |
| 5 LSTM Autoencoder . . . . .                          | 11       |
| 6 Mètriques d'error de regressió . . . . .            | 12       |
| 6.1 NSE . . . . .                                     | 13       |
| 6.2 PBIAS . . . . .                                   | 13       |
| 7 Mètriques d'error de classificació . . . . .        | 13       |
| 8 Optimització d'hiperparàmetres . . . . .            | 14       |
| 8.1 Hiperparàmetres del model Prophet . . . . .       | 15       |
| 8.2 Hiperparàmetres de les xarxes neuronals . . . . . | 15       |
| 9 SWAT+ i dades d'entrada . . . . .                   | 15       |
| 9.1 Dades climàtiques . . . . .                       | 16       |
| 9.2 Model digital d'elevacions . . . . .              | 17       |
| 9.3 Tipus del sòl . . . . .                           | 17       |

|          |  |           |
|----------|--|-----------|
| 9.4      | Usos del sòl . . . . .                 | 18        |
| <b>4</b> | <b>Planificació i metodologia</b>      | <b>21</b> |
| 1        | Metodologia . . . . .                  | 21        |
| 1.1      | Enteniment del negoci . . . . .        | 21        |
| 1.2      | Enteniment de les dades . . . . .      | 21        |
| 1.3      | Preparació de les dades . . . . .      | 21        |
| 1.4      | Modelatge . . . . .                    | 22        |
| 1.5      | Avaluació . . . . .                    | 22        |
| 1.6      | Desplegament . . . . .                 | 22        |
| 2        | Planificació temporal . . . . .        | 22        |
| <b>5</b> | <b>Contribució metodològica</b>        | <b>24</b> |
| 1        | Mètode 1 . . . . .                     | 24        |
| 2        | Mètode 2 . . . . .                     | 25        |
| 2.1      | Entrenament de la xarxa LSTM . . . . . | 25        |
| 2.2      | Detecció d'anomalies . . . . .         | 27        |
| <b>6</b> | <b>Resultats</b>                       | <b>29</b> |
| 1        | Mètode 1 . . . . .                     | 29        |
| 2        | Mètode 2 . . . . .                     | 31        |
| 2.1      | Entrenament de la xarxa LSTM . . . . . | 31        |
| 2.2      | Detecció d'anomalies . . . . .         | 38        |
| <b>7</b> | <b>Conclusions</b>                     | <b>43</b> |

# Índex de figures

|      |  |    |
|------|--|----|
| 3.1  | Filtratge de les estacions d'aforament . . . . .   | 5  |
| 3.2  | Períodes amb cabals basals diferents . . . . .   | 6  |
| 3.3  | Períodes amb cabal constant superior a 0 . . . . .   | 6  |
| 3.4  | Períodes amb hidrograma amb forma d'altiplà . . . . .  | 7  |
| 3.5  | Components d'una sèrie temporal . . . . .  | 8  |
| 3.6  | Operacions internes d'una xarxa RNN (a) i LSTM (b), de [13] . . . . .  | 10 |
| 3.7  | Estat de la cel·la LSTM, extret de [13] . . . . .  | 10 |
| 3.8  | Arquitectura model Autoencoder, extret de [6] . . . . .  | 12 |
| 3.9  | Estacions meteorològiques de Catalunya: a l'esquerra, de precipitacions, i a la dreta, de temperatura . . . . .  | 16 |
| 3.10 | Conques i subconques de les CIC . . . . .  | 17 |
| 3.11 | Tipus de sòl a Catalunya . . . . .   | 19 |
| 3.12 | Usos del sòl de CIC . . . . .  | 20 |
| 4.1  | Diagrama de Gantt de la planificació temporal . . . . .  | 23 |
| 6.1  | Cabal observat (taronja) i predit (blau) de l'estació d'Olvan. En vermell, les anomalies reals no predites (FN), i en lila, les anomalies no observades predites (FP) . . . . .  | 30 |
| 6.2  | Entrenament de la xarxa LSTM . . . . .   | 31 |
| 6.3  | Ajust del model LSTM per als cabals diaris mitjançant la funció objectiu NSE en les diferents estacions d'aforament, representades per un cercle si estan per sobre d'embassaments, o amb un quadrat si estan per sota. A les estacions amb NSE negatiu se'ls hi ha assignat un NSE de 0 . . . . . | 33 |
| 6.4  | Cabal observat (taronja) i predit (blau) per a l'estació de Pont de Molins, amb un NSE de -0.17, utilitzant el model LSTM . . . . .  | 34 |

|      |   |    |
|------|---|----|
| 6.5  | Cabal observat a l'estació de Sant Joan de les Abadesses (en taronja), cabal predit pel model LSTM (blau) i cabal predit pel model LSTM+TL (vermell) . . . . .  | 35 |
| 6.6  | Prediccions amb el model LSTM entrenat amb logNSE (blau), entrenat amb NSE (vermell) contra les observacions (taronja) . . . . .  | 36 |
| 6.7  | Cabal observat (taronja) i predit (blau) per a l'estació de Campllong (riera de La Gotarra), on veiem dos comportaments diferents de les observacions . . . . .   | 37 |
| 6.8  | Entrenament de la xarxa LSTM Autoencoder . . . . .  | 38 |
| 6.9  | Distribució dels errors obtinguts durant la reconstrucció de les mostres utilitzant un LSTM autoencoder . . . . .   | 39 |
| 6.10 | A l'esquerra, observació correcta, i observacions i prediccions corresponents als 13 dies anteriors. A la dreta, el mateix però amb una mostra anòmala. El color taronja representa les observacions, i el blau les prediccions . . . . . | 40 |
| 6.11 | Distribució dels errors obtinguts durant la reconstrucció de les mostres de les estacions amb un $NSE > 0.2$ utilitzant un LSTM autoencoder . . . . .   | 41 |
| 6.12 | A l'esquerra, observació correcta, i observacions i prediccions corresponents als 13 dies anteriors. A la dreta, el mateix però amb una mostra anòmala. El color taronja representa les observacions, i el blau les prediccions . . . . . | 42 |



# Índex de taules

|     |  |    |
|-----|--|----|
| 3.1 | Estacions d'aforament amb incongruències o dades sospitoses . . . . .  | 4  |
| 3.2 | Valors de NSE i PBIAS pel model realitzat amb SWAT . . . . .   | 16 |
| 6.1 | Mitjana dels errors de les prediccions de cabal de cada estació utilitzant Prophet                                       | 29 |
| 6.2 | Mitjana de les estadístiques de les deteccions d'anomalies de cada estació utilitzant Prophet . . . . .                  | 29 |
| 6.3 | Mitjana dels errors de les prediccions de cabal de cada estació utilitzant el model LSTM, separat per conca . . . . .    | 31 |
| 6.4 | Mitjana dels errors de les prediccions de cabal de cada estació utilitzant el model LSTM+TL, separat per conca . . . . . | 32 |
| 6.5 | Mitjana de les estadístiques de les deteccions d'anomalies de cada estació utilitzant un autoencoder . . . . .           | 38 |

# 1. Introducció

En el context de sequera actual que vivim a Catalunya, on des del 2021 estem sotmesos en uns dels períodes de sequera més greus dels darrers anys [2], la disponibilitat dels recursos hídrics és crítica per diferents sectors, entre els quals hi ha l'agricultura, la indústria i l'ús domèstic. En aquest sentit, tot i que només un 3.5% de l'aigua que hi ha a la terra és dolça, i d'aquesta, únicament un 0.006% prové dels rius, aquests són l'origen de l'aigua que utilitza la majoria de gent [7][15]. Per tant, poder predir i modelar el cabal dels rius, que es representa en forma de sèries temporals anomenades hidrogrames, juga un rol vital per tal de ser capaços d'entendre la disponibilitat d'aigua, predir així patrons futurs, i podent prendre decisions per mitigar l'escassetat d'aigua.

Per tal de programar aquests models necessitem dades, que en el nostre cas provenen d'una sèrie d'estacions d'aforament, que són les instal·lacions encarregades de mesurar la quantitat d'aigua que circula pels cursos fluvials. Detectar les anomalies d'aquestes dades és un pas essencial per assegurar la correctesa del model. Aquestes anomalies poden provenir de patrons anòmals, però no necessàriament incorrectes (pluges fortes sobtades, períodes de sequera), o bé perquè aquestes estacions d'aforament han deixat de funcionar correctament, i sí que són dades que s'han de descartar.

Per tant, i a diferència d'altres treballs, en aquest ens centrarem a detectar les anomalies de les dades majoritàriament causades per un mal funcionament de les estacions d'aforament, de manera que el nou conjunt de dades filtrat pot ser utilitzat per entrenar models robustos que ajudin a comprendre els processos hidrològics, i no només això, sinó que també és interessant l'aplicació d'aquest treball pels gestors de les estacions d'aforament, ja que la presència de noves dades anòmales pot indicar un mal funcionament d'aquestes. Per això, hem dissenyat dos mètodes: el primer, d'ara endavant, mètode 1, es basa a modelitzar les dades de cabal utilitzant l'algorisme de modelatge Prophet (dissenyat per Facebook) i considerar el que estigui fora d'un cert interval de confiança com a anomalia. El segon mètode (mètode 2) consisteix a modelitzar el cabal de totes les estacions utilitzant una sola xarxa Long Short-Term Memory (LSTM) però no a partir de les mateixes dades de cabal, sinó a través de variables externes com la precipitació, temperatura, etc., i a continuació, entrenar un LSTM Autoencoder que donat un conjunt de prediccions de la xarxa LSTM i les observacions corresponents, sigui capaç de determinar si la mostra es tracta d'una anomalia o no.

## 2. Literatura relacionada i estat de l'art

La noció de detecció d'anomalies ja ha tingut diverses aplicacions en el camp de la hidrologia. Per tal de detectar anomalies en dades provinents d'hidrogrames, [18] utilitza una *sliding window* i un model ARIMA per tal de fer una predicció sobre les dades. Després, es calcula un interval de confiança sobre aquesta predicció, i tot el que caigui fora d'aquest interval es comprova que realment sigui una anomalia a partir d'aplicar una clusterització amb *K-Means* i processos de Markov. De manera semblant, [20] fa servir un model AR i la distància amb el k-NN per determinar si una mostra és un outlier o no.

A [19] es detecten patrons anòmals primer discretitzant les dades originals en sèries simbòliques, després en construeix una seqüència de símbols a partir d'un PST (*Probabilistic Suffix Tree*), a partir de la qual pot extreure les anomalies.

Encara dins del camp de la hidrologia, però per predir anomalies relatives a la demanda d'aigua per cada hora, [17] proposa un model *encoder-decoder*, que millora els resultats de l'*autoencoder* utilitzat en aquest treball.

Fora d'aquest sector, [11] detecta anomalies relatives al consum elèctric dels edificis, primer creuant diferents dades (consums i temperatures) i creant un model aplicant un *random forest*, per a després amb l'error obtingut entre les observacions i prediccions, detectar els outliers aplicant diferents mètodes clàssics de detecció de valors atípics com els utilitzats al *boxplot* o el mètode 3-sigma.

Amb relació a les prediccions amb Prophet, [4] compara diferents models, entre els quals el de Facebook, per tal de predir el nivell d'aigua subterrània, fent servir les precipitacions com a variable exògena.

Pel que fa a la modelització del cabal dels rius, [8] proposa una aproximació utilitzant xarxes LSTM i estudia com aquesta aprèn les diferents dependències temporals, i com un sol model pot predir el cabal de diferents conques hidrogràfiques, cadascuna amb característiques diferents.

## 3. Preliminars

### 1 Dades de cabal

Pel que fa a les dades de cabal provinents de les estacions d'aforament, disposem d'observacions per a una xarxa de 73 estacions d'aforament pel període comprès entre l'1 de gener del 2001 i el 16 de febrer del 2021. Les sèries de cabal són a resolució diària, però moltes estacions només disposen d'un registre parcial o la qualitat de les observacions és qüestionable. Per aquesta última raó hem realitzat un filtratge previ manual d'aquestes estacions d'aforament, eliminant la totalitat o part de les estacions que presenten incongruències o dades sospitoses, tal com es defineix a la taula 3.1. D'aquesta manera, podrem entrenar els models corresponents utilitzant dades que ja mostren un comportament correcte.

Així, tenint en compte tot el període esmentat i per totes les estacions, hauríem de tenir 536623 registres. D'aquests, 167746 no tenen valor, i 48977 els hem classificat com a incorrectes (42555 corresponen a observacions de les estacions descartades completament i 6422 de les estacions parcialment correctes, seguint la mateixa nomenclatura que la taula 3.1). Per tant, tenim 319900 observacions considerades com a vàlides (el 60% del total).

Adicionalment, hem diferenciat les 64 estacions restants segons si se situen aigües amunt o aigües avall dels embassaments.

Aquesta diferenciació és important pel fet que els embassaments representen un element diferencial en la modelització de la hidrologia, ja que el cabal de les estacions situades per sota d'aquests es veu afectat pel seu buidatge, el qual segueix unes regles poc definides on intervien diferents factors, com el nivell d'aigua d'aquests, l'estat de sequera general o previsió de pluges, entre d'altres.

La distribució espacial de les estacions d'aforament i el seu filtratge es pot veure a la figura 3.1.

Abans que res, cal mencionar que per tots els hidrogrames mostrats en el treball, l'eix de les  $x$  indica el dia de la mostra, l'eix de les  $y$  representa el cabal detectat (en  $m^3/s$ ), i s'utilitza el color blau per representar mostres vàlides, i el vermell per mostres incorrectes.

Per tal d'identificar els comportaments incongruents o sospitosos a què fem referència a la taula 3.1 hem seguit una sèrie de criteris que exposem a continuació. El primer es pot veure a la figura 3.2, on el cabal basal (valor de cabal al qual es tendeix en períodes amb poca precipitació) de les mostres del primer període (fins a 2006) és lleugerament superior respecte el cabal basal de les mostres del següent període (passa d'uns  $5m^3/s$  a 1.5). En aquest cas concret, s'ha considerat com a correcte el segon tram pel fet que és el més extens en el temps. En el cas que els trams

| Estació                             | Acció                               | Raó   |
|-------------------------------------|-------------------------------------|---|
| Anglès (Riera d'Osor)               | Eliminar totalment                  | El comportament hidrològic de tota o gran part de la sèrie és incongruent |
| Campllong (La Gotarra)              |                                     |   |
| Cercs (Riera Peguera)               |                                     |   |
| Cornellà del Terri                  |                                     |   |
| La Bisbal d'Empordà (Daró)          |                                     |   |
| Les Masies de Roda (Gurri)          |                                     |   |
| Sant Quintí de Mediona              |                                     |   |
| Súria                               |                                     |   |
| Castellar del Vallès (Riera d'Osor) |                                     |   |
| Serinyà                             | Eliminar totalment                  | El nombre d'observacions no és representatiu                              |
| Fogars de la Selva (Can Simó)       | Eliminar parcialment                | El comportament hidrològic d'un període és incongruent                    |
| Berga (Olvan)                       |                                     |   |
| Castellbisbal (Assut de Salzers)    |                                     |   |
| QLSup(A) - Martorell (Anoia)        |                                     |   |
| Riudellots de la Selva (Onyar)      |                                     |   |
| Santa Cristina d'Aro                |                                     |   |
| Serra de Daró (Riu Daró)            | Eliminar parcialment                | Hi ha un episodi de cabals alts incongruent.                              |
| Maçanet de Cabrenys (Arnera)        |                                     |   |
| Vilanova de Sau (Riera Major)       | Eliminar parcialment                | Eliminació d'un registre aïllat.  |
| Jorba                               | No eliminar, però estació sospitosa | S'observa un comportament hidrològic sospitós                             |
| Lliçà de Vall                       |                                     |   |
| Tarragona                           |                                     |   |

Taula 3.1: Estacions d'aforament amb incongruències o dades sospitoses

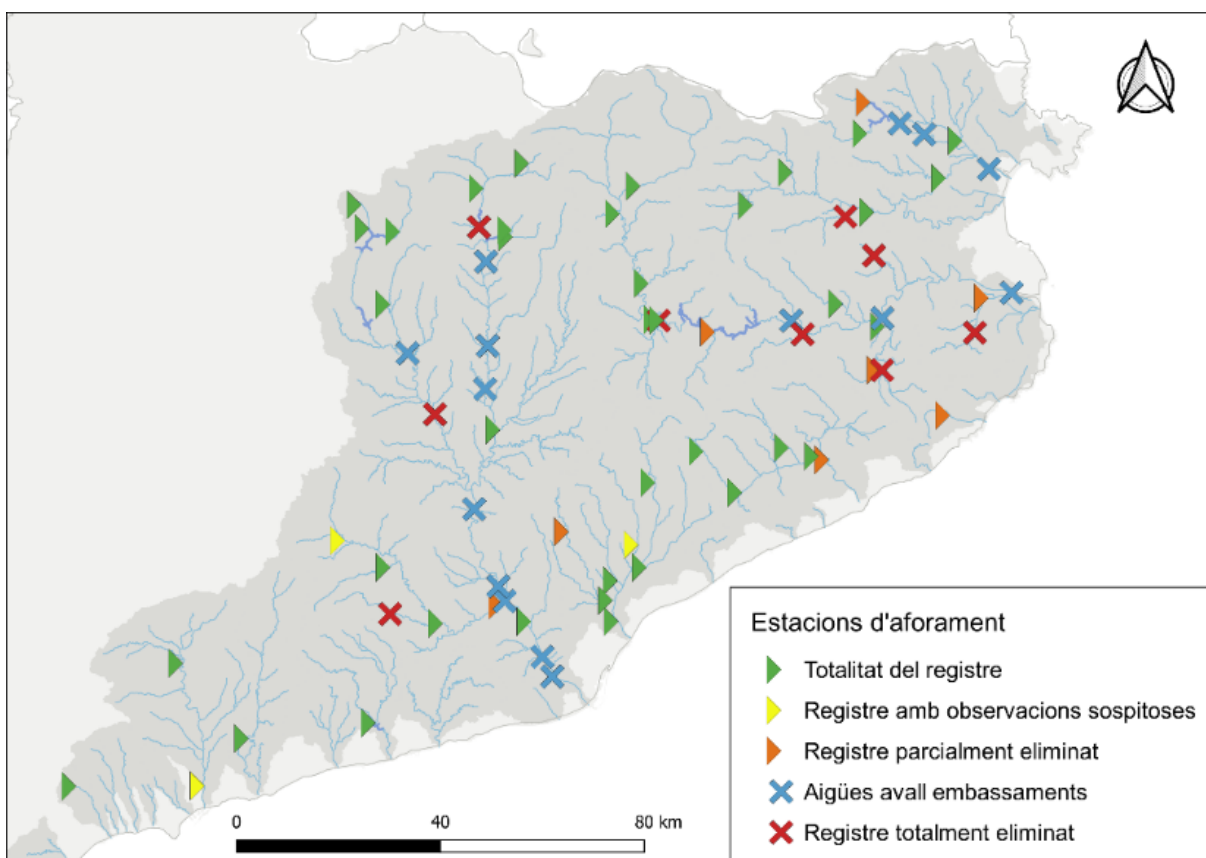


Figura 3.1: Filtratge de les estacions d'aforament

amb comportaments diferents siguin igual d'extens en el temps, hem de descartar totalment les dades, ja que no podem afirmar quin dels dos trams és el correcte.

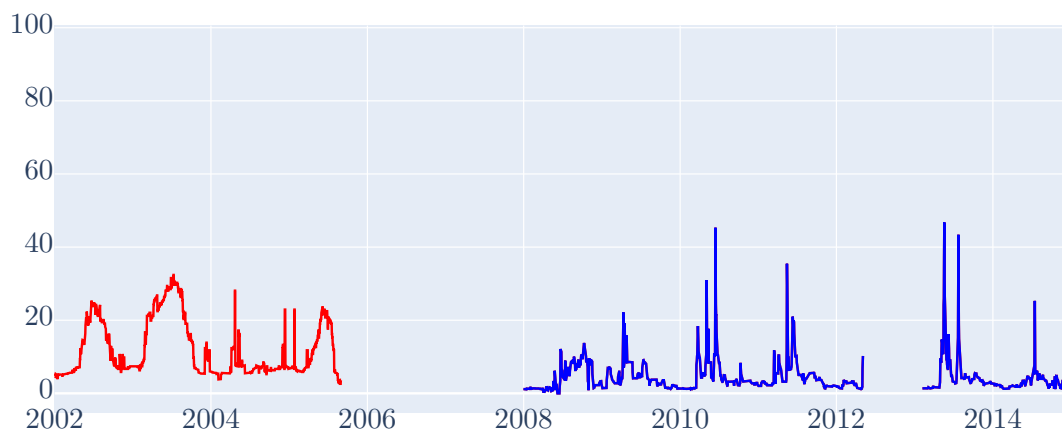


Figura 3.2: Períodes amb cabals basals diferents

També ens hem trobat amb períodes considerablement llargs amb un registre (quasi) constant, que si bé podrien ser certs si aquests fossin valors propers al cabal basal (a causa de precipitacions escasses), tenim casos en què aquests patrons són inusuals, com és el cas concret de la figura 3.3, on durant un període de més d'un any el cabal es manté constant als  $5m^3/s$ .

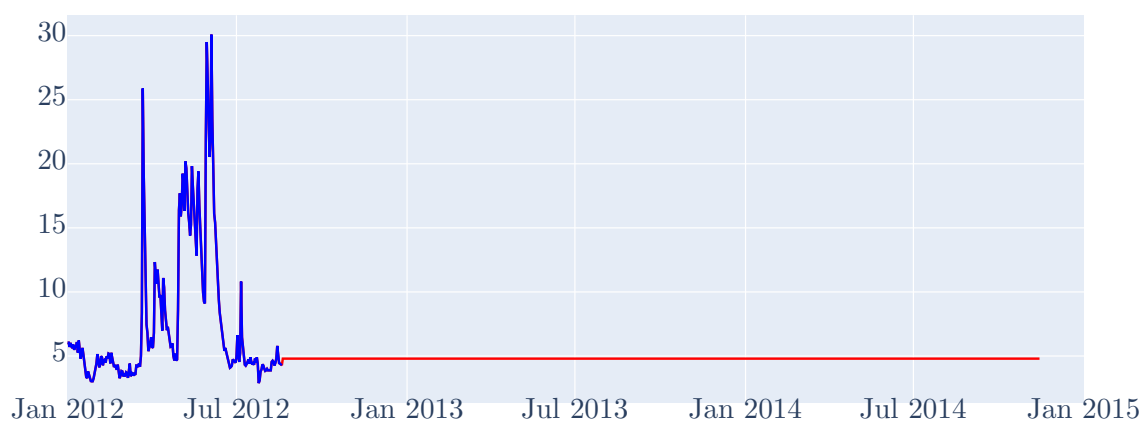


Figura 3.3: Períodes amb cabal constant superior a 0

Cal mencionar també patrons on l'hidrograma segueix una forma semblant a un altiplà, en què el valor de cabal augmenta sobtadament, es queda constant (més o menys) durant un període

de temps, i després torna a disminuir de cop, comportament que es pot veure a la figura 3.4.

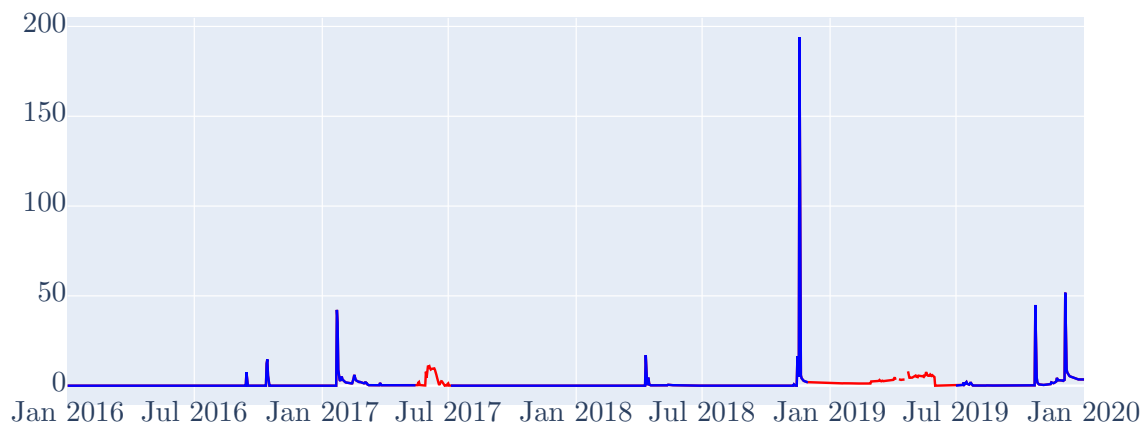


Figura 3.4: Períodes amb hidrograma amb forma d'altiplà

A mode general, s'ha de dir que el filtratge de les mostres no es fa tenint present una sola finestra de temps concreta, sinó que depèn en gran part del comportament general de la sèrie. Per tant, tot i el conjunt de comportaments anòmals definits anteriorment, és complicat establir una sèrie de regles a seguir per tal de determinar si una mostra és incorrecte o no, sinó que en gran part se segueix una revisió manual aplicant criteri expert.

## 2 Sèries temporals

Una sèrie temporal és un conjunt successiu de mesures  $X = x_1, \dots, x_t$ , on  $x_t$  és el valor observat a l'instant de temps  $t$ .

Es pot pensar una sèrie temporal com una suma dels següents components, com també es pot veure a la figura 3.5:

- Tendència: comportament o moviment a llarg termini de la sèrie
- Estacionalitat: cicles o patrons que es van repetint al llarg de la sèrie de forma constant i ordenada
- Soroll: alteracions de sèrie que no segueixen una pauta periòdica ni tendencial reconeixible.



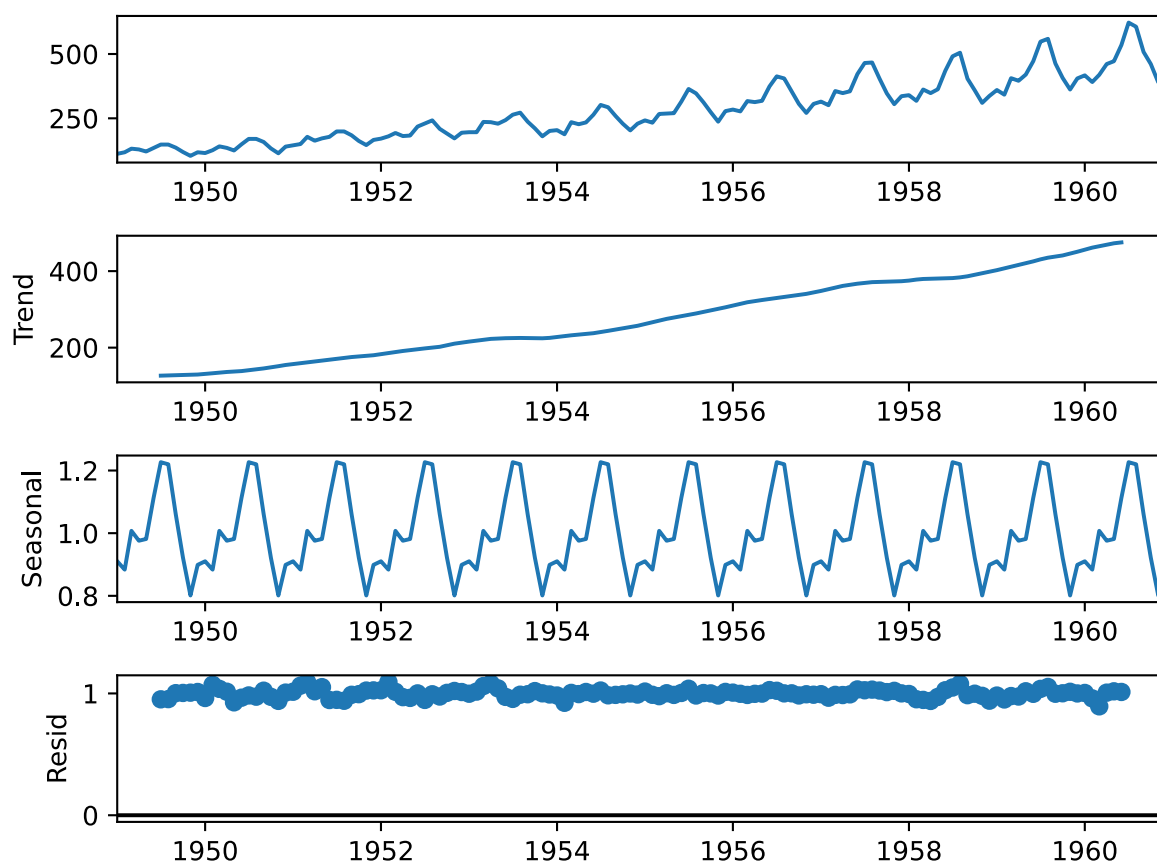


Figura 3.5: Components d'una sèrie temporal

### 3 Prophet

Prophet és un mètode de predicció proposat per Facebook que permet predir dades de sèries temporals basant-se en un model additiu generalitzat (GAM), on les contribucions dels predictors són alguna funció  $f$ , de la forma  $y = \beta_0 + f_1(x_1) + \dots + f_n(x_n) + \epsilon$ , on  $y$  és el terme a predir,  $\beta_0$  és un terme constant i  $\epsilon$  un terme d'error.

Aleshores, Prophet ajusta tendències no lineals tenint en compte l'estacionalitat anual, setmanal i diària, així com sobre l'efecte dels dies festius sobre les dades. A la vegada, és un mètode robust davant la presència de valors nuls i valors atípics, fet que el fa òptim per al nostre problema.

Prophet identifica els canvis de tendència (punts on la línia de tendència canvia) i produeix línies de tendència entre aquests canvis. Per tal de trobar les estacionalitats, calcula una sèrie de Fourier per aproximar una funció periòdica que captura els efectes estacionaris. De forma simple, es pot resumir l'efecte d'aquestes funcions sobre la predicció com es pot veure a l'equació 3.1, on  $y(d)$  és la predicció del model per data  $d$ ,  $g(d)$  és la tendència de la sèrie,  $s(d)$  és la component de l'estacionalitat,  $h(d)$  és l'efecte de les vacances i  $\epsilon_d$  és un terme d'error.

$$y(d) = g(d) + s(d) + h(d) + \epsilon_d \quad (3.1)$$

### 4 Xarxes RNN i LSTM

Una xarxa neuronal recurrent (RNN) és una xarxa neuronal que utilitza dades seqüencials o temporals. Solen tenir un sol estat ocult que permet que la informació passi d'un pas al següent, però pateixen del que es coneix com a esvaïment del gradient, en què la xarxa tendeix a oblidar les dependències a llarg termini. En concret, [5] mostra com una RNN tradicional no pot captar les dependències en seqüències superiors a longitud 10.

En canvi, una xarxa LSTM (*Long Short-Term Memory*) és un tipus de RNN que introdueix una cel·la d'estat (o de memòria), que pot mantenir la informació per un període de temps més llarg. Aquesta cel·la està controlada per tres portes (*gates*): la *forget gate*, la *input gate*, i l'*output gate*.

En el context d'aquest treball, és important aquesta memòria a llarg termini per tal de modelitzar correctament efectes com la caiguda i fosa de neu, els quals tenen una escala temporal superior a altres efectes com la precipitació.

El funcionament i les diferències entre les xarxes RNN i LSTM es poden veure a la figura 3.6.

En una cel·la RNN només existeix un estat intern  $h_t$  (hidden state), que és calculat a cada iteració seguint la següent equació:

$$h_t = g(Wx_t + Uh_{t-1} + b), \quad (3.2)$$

on  $g$  és la funció d'activació,  $W$  i  $U$  són pesos de la matriu de l'estat ocult  $h$  i l'input  $x$ , i  $b$  és el vector de *bias*.

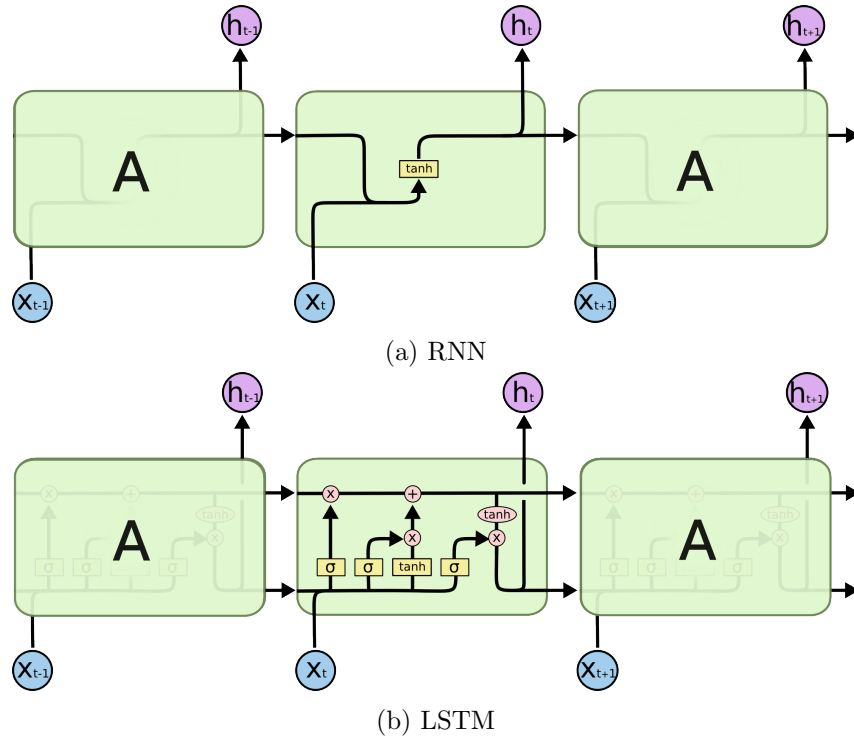


Figura 3.6: Operacions internes d'una xarxa RNN (a) i LSTM (b), de [13]

En canvi, la clau en una cel·la LSTM està en l'estat  $c$ , tal com es pot veure a la figura 3.7.

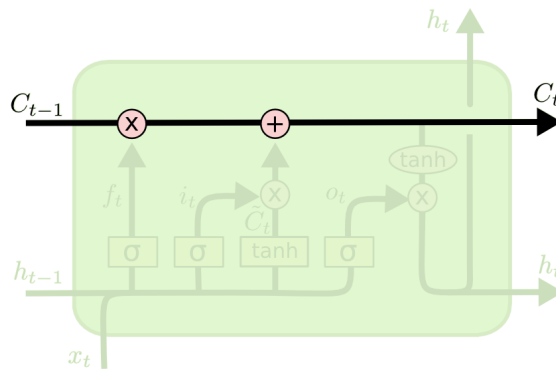


Figura 3.7: Estat de la cel·la LSTM, extret de [13]

A aquest estat li afegim o se li treu informació fent servir les *gates* explicades anteriorment: primer actua la *forget gate*, que controla quina informació s'ha d'eliminar de  $c_{t-1}$ , com es veu a l'equació 3.3:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f), \quad (3.3)$$

on  $f_t$  és el vector resultant,  $\sigma$  és la funció d'activació, i  $W_f$ ,  $U_f$  i  $b_f$  defineixen el conjunt de paràmetres a aprendre per la xarxa, mentre que  $x_t$  és l'input de l'estat de temps actual i  $h_{t-1}$  és l'estat ocult del període de temps anterior.

Pel que fa al següent pas, la *input gate* controla quina informació s'ha d'afegir a la cel·la. Primer calcula un potencial vector d'actualització de l'estat, aplicant:

$$\tilde{c}_t = \tanh(W_{\tilde{c}}x_t + U_{\tilde{c}}h_{t-1} + b_{\tilde{c}}), \quad (3.4)$$

on  $\tanh$  és la funció d'activació tangent hiperbòlica, i  $W_{\tilde{c}}$ ,  $U_{\tilde{c}}$  i  $b_{\tilde{c}}$  són un altre conjunt de paràmetres a aprendre, i a continuació, defineix quina informació de  $\tilde{c}_t$  s'utilitza per actualitzar l'estat:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i), \quad (3.5)$$

on de nou  $W_i$ ,  $U_i$  i  $b_i$  són valors a aprendre per la xarxa. Utilitzant tot això, podem calcular  $c_t$  com:

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}, \quad (3.6)$$

que es pot interpretar com quina informació guardada a  $c_{t-1}$  s'oblida i es guarda (valors de  $f_t$  propers a 0 o a 1 respectivament, i de forma similar,  $i_t$  decideix quina informació emmagatzemada a  $\tilde{c}_t$  s'afegeix).

L'última porta és la *output gate*, que controla quina és la informació de l'estat de cel·la es propaga com a input pel dia següent:

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (3.7)$$

on de nou  $W_o$ ,  $U_o$  i  $b_o$  són valors a aprendre per la xarxa.

Finalment, després de calcular el vector  $o_t$ , es calcula el nou estat ocult  $h_t$  (hidden state), de la següent manera:

$$h_t = \tanh(c_t) \cdot o_t \quad (3.8)$$

Per tant, com es pot veure, és aquesta actualització sobre  $c_t$  que permet captar les dependències temporals a llarg termini [9].

## 5 LSTM Autoencoder

Un autoencoder és un tipus de tècnica no supervisada que consisteix a dissenyar una arquitectura de xarxa neuronal de manera que creem un coll d'ampolla a la xarxa, la qual cosa força una representació comprimida del coneixement de l'entrada original.

Així, partim d'un conjunt d'informació no etiquetat i el tractem com un problema d'aprenentatge supervisat amb la tasca de produir una reconstrucció  $x'$  de l'entrada original  $x$ . Aquesta xarxa és entrenada intentant minimitzar la reconstrucció d'error,  $L(x, x')$  que mesura la diferència entre

l'input original i la reconstrucció pertinent.

Està format pels components encoder i decoder. L'encoder és una xarxa neuronal que produeix el que es coneix com a capa *code* (o bottleneck) a partir de les dades d'entrada. Aquesta capa té un determinat nombre de nodes, on un nombre més petit implica més compressió. Pel que fa al decoder, normalment té una estructura inversa a l'encoder, i parteix de la capa *code* per tal de generar una sortida que és el màxim de similar a les dades d'entrada, com es pot veure a la figura 3.8.

En el nostre cas farem servir un LSTM Autoencoder, on tant com l'encoder com el decoder són xarxes neuronals de tipus LSTM. Partint d'aquesta estructura és possible entrenar un detector d'anomalies. A partir de patrons o seqüències de dades correctes, es pot entrenar aquest model on s'intenta minimitzar una certa funció de pèrdua, i a partir dels errors  $L(x, x')$  obtinguts en aquesta reconstrucció, establir un llindar a partir del qual considerar una mostra com a errònia o correcta. D'aquesta manera, és d'esperar que patrons anòmals tinguin errors grans (superiors al llindar establert) quan el model intenta codificar i reconstruir les dades.

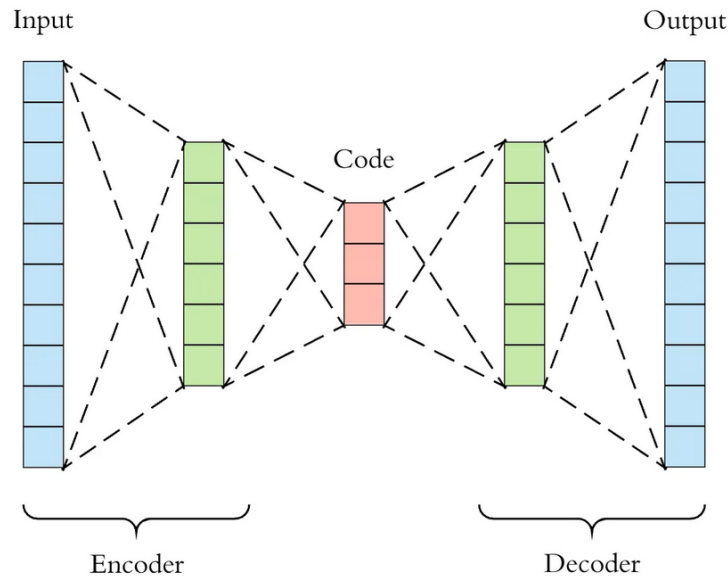


Figura 3.8: Arquitectura model Autoencoder, extret de [6]

Per exemple, suposant  $x = [x_1, \dots, x_n]$ , i una xarxa entrenada utilitzant una funció de pèrdua  $f$ , pots obtenir una reconstrucció  $x' = [x'_1, \dots, x'_n]$ , on pots calcular  $L(x, x')$  utilitzant la mateixa funció  $f$ . Aleshores, reconstruint tots els elements del conjunt de train s'obindrà un conjunt de valors  $[L(x, x')]$ , a partir del qual es decideix quin és el llindar o valor per considerar una mostra com a correcta o incorrecta.

## 6 Mètriques d'error de regressió

Per tal de quantificar l'error de les prediccions del cabal respecte les observacions, s'han fet servir l'NSE i el PBIAS, detallats a continuació, ja que aquestes són les mateixes que s'havien utilitzat per tal d'entrenar un model previ fent servir SWAT+, explicat posteriorment.

## 6.1 NSE

El coeficient d'eficiència del model Nash-Sutcliffe (NSE) és una mètrica utilitzada per avaluar l'eficiència d'un model hidrològic. Compara la similitud entre els valors observats i els simulats pel model, i proporciona un valor entre  $(-\infty, 1]$ , on un valor proper a 1 indica una bona correspondència entre els valors observats i simulats (bona bondat d'ajust), un  $NSE = 0$  indica una eficiència predictiva equivalent a la mitjana de les observacions, mentre que un  $NSE < 0$  indica un poder predictiu del model inferior a la mitjana de les observacions.

La seva definició es pot veure a l'equació 3.9, on  $Q_o^t$  és el cabal observat per al dia  $t$ ,  $Q_m^t$  és el cabal modelat per al dia  $t$ , i  $\bar{Q}_o$  és la mitjana de totes les observacions.

$$NSE = 1 - \frac{\sum_{t=1}^T (Q_o^t - Q_m^t)^2}{\sum_{t=1}^T (Q_o^t - \bar{Q}_o)^2} \quad (3.9)$$

Podríem definir un valor  $NSE > 0.65$  com a molt bo, un  $NSE > 0.5$  com a satisfactori, i un  $NSE < 0.5$  com a insatisfactori [12].

Com es pot veure, aquest és molt semblant al MSE (Mean Squared Error), ja que ambdós són funcions d'errors quadràtiques. El fet que l'NSE estigui fitat entre  $[-\infty, 1]$ , però, el fa més interpretable.

## 6.2 PBIAS

El *Percent bias* (PBIAS) mesura la tendència mitjana dels valors simulats a ser superiors o inferiors als respectius valors observats.

Així, el valor òptim de PBIAS és 0, on valors propers indiquen una bona simulació. Valors positius indiquen valors simulats superiors als observats, mentre que valors negatius indiquen valors predits inferiors a les observacions.

La definició del PBIAS es pot veure a l'equació 3.10, on  $Q_o^t$  indica el cabal observat per al dia  $t$ , i  $Q_m^t$  és el cabal modelat per al dia  $t$ .

$$PBIAS = 100 \cdot \frac{\sum_{t=1}^T (Q_m^t - Q_o^t)}{\sum_{t=1}^T Q_o^t} \quad (3.10)$$

Considerant el valor absolut del PBIAS, podríem dir que un valor inferior al 15% seria un molt bon valor, un valor inferior al 25% seria considerat com a satisfactori, i un valor superior al 25% com a insatisfactori [12].

## 7 Mètriques d'error de classificació

En els problemes de classificació, entre els quals trobem el de detecció d'anomalies, es fa servir diferents mètriques per tal d'avaluar el rendiment de l'algorisme. En concret, hem fet servir les

mètriques de precision, recall i F1, les quals estan definides a continuació.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3.11)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3.12)$$

$$\text{F1} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3.13)$$

On TP fa referència als positius veritables (anomalies correctament detectades), FP al nombre de falsos positius (mostres correctes classificada com a anomalia) i FN fa referència als falsos negatius, és a dir, anomalies classificades com a mostres normals.

En aquest context, Precision indica d'entre totes les anomalies detectades per l'algorisme, el percentatge d'aquests que es corresponen a anomalies reals. Per altra banda, Recall fa referència al nombre d'anomalies detectades respecte del total d'anomalies reals. La puntuació F1 és una mitjana harmònica entre Precision i Recall per tal d'avaluar el rendiment global del mètode.

## 8 Optimització d'hiperparàmetres

Aquest és el problema d'escollir el conjunt d'hiperparàmetres òptim que minimitza l'error obtingut durant l'entrenament del model. Un hiperparàmetre és un paràmetre el valor del qual s'utilitza per controlar el procés d'aprenentatge, i és especificat de forma externa (no s'aprèn durant l'entrenament).

Per intentar trobar aquest conjunt òptim s'utilitzen sistemes d'optimització, entre els quals hem fet servir Grid search i l'optimització bayesiana. Respecte al primer, aquest consisteix en una cerca exhaustiva entre un conjunt de paràmetres especificats manualment. L'algorisme prova sistemàticament totes les possibles combinacions de valors d'hiperparàmetres possibles, de les quals s'escull la millor. Pel que fa a l'optimització bayesiana, aquest és un mètode que construeix un model probabilístic que relaciona els valors dels hiperparàmetres amb una funció objectiu  $f$  (per exemple, l'error obtingut al conjunt de validació), on  $f$  és una funció costosa d'avaluar. Consisteix en, de forma iterativa, anar seleccionant una configuració prometedora d'hiperparàmetres, avaluar-la a  $f$ , i anar actualitzant el model probabilístic utilitzant inferència bayesiana.

Pel que fa als hiperparàmetres a utilitzar i el rang de valors òptim, aquests depenen exclusivament del model entrenat i el problema a resoldre. El valor dels hiperparàmetres és important, ja que normalment valors fora dels rangs correctes fa que el model o bé sigui incapaç d'entendre les especificitats de les dades, o bé que el model aprengui patrons massa complexos i es produeixi sobre ajustament.

S'exposen els hiperparàmetres utilitzats en aquest treball a continuació, mentre que els rangs corresponents es mencionen a 5.

## 8.1 Hiperparàmetres del model Prophet

L'hiperparàmetre de Prophet que més importància té és el de `changepoint_prior_scale`, el qual determina la flexibilitat de la tendència, en concret com pot canviar la tendència entre punts de canvis de tendència. Si és massa petit, la variància de la sèrie que hauria de ser modelada com a canvis de tendència serà identificada com a soroll. Si és massa gran, el model se sobre ajustarà a la sèrie original.

També és important `seasonality_prior_scale`, el qual controla la flexibilitat de l'estacionalitat, és a dir, la capacitat del model a ajustar-se a grans fluctuacions, o bé fer que l'estacionalitat sigui molt petita. En altres paraules, permet que els períodes d'estacionalitat siguin més grans o més petits.

## 8.2 Hiperparàmetres de les xarxes neuronals

Típicament, alguns dels hiperparàmetres que s'optimitzen són la dimensió de l'estat ocult  $h_t$ , així com el dropout, que fixa un percentatge determinat de neurones (escollides aleatòriament) a 0 durant l'entrenament per forçar a la xarxa a aprendre característiques més robustes.

També se sol optimitzar la mida de batch, que el podríem definir com el nombre de mostres que es propaguen per la xarxa. Suposem un conjunt d'entrenament de 100 mostres, amb una mida de batch és de 10, aleshores un epoch (període en què cada mostra d'entrenament és usada per actualitzar els paràmetres del model) consisteix en 10 iteracions (nombre de mostres entre mida de batch), on a cada iteració 10 de les 100 mostres s'escullen aleatòriament (sense repetició, per tant, una mostra s'usa una vegada per epoch) i es fan servir per entrenar el model.

## 9 SWAT+ i dades d'entrada

En aquest treball s'ha partit d'una modelització prèvia de la hidrologia a les CIC (Conques Internes de Catalunya) realitzada per ICRA (Institut Català de Recerca de l'Aigua), en què es feia una predicció del cabal utilitzant SWAT+. SWAT+ és una versió reestructurada i actualitzada de la Soil and Water Assessment Tool (SWAT) que és d'un dels models hidrològics més utilitzats arreu del món, i la seva versió actualitzada ofereix més flexibilitat en termes de representació en l'espai de les interaccions i els processos hidrològics dins una conca, així com una interfície d'usuari més intuïtiva. SWAT+, així com les diferents interfícies utilitzades per configurar, modificar i calibrar el model, són programari lliure.

En aquesta aproximació, i d'ara endavant, model SWAT, es va entrenar un model diferent per cada conca, els resultats dels quals es poden veure a la taula 3.2. El Fluvià presenta un valor de NSE molt bo, mentre que el sistema Besòs-Tordera i el Ter tenen un NSE bo. La Muga, el Sistema Conques del Sud i el Llobregat (sobretot aquests dos últims) en canvi, presenten valors de NSE no acceptables. Respecte al PBIAS, els valors són molt bons excepte per la Muga, el Ter i el Llobregat.

Nosaltres, per al mètode 2, hem utilitzat una versió simplificada de les dades d'entrada utilitzades per la modelització amb SWAT. Per exemple, en el nostre model no hem tingut presents les dades relatives a embassaments, que com ja hem comentat, sol ser un element important al realitzar



| Unitat de modelització | NSE    | PBIAS |
|------------------------|--------|-------|
| Muga                   | 0.130  | -78.5 |
| Fluvià                 | 0.796  | -4.5  |
| Ter                    | 0.551  | 54.1  |
| Besòs-Tordera          | 0.564  | 7.5   |
| Llobregat              | -0.634 | -66.9 |
| Conques del Sud        | -0.142 | -1.9  |

Taula 3.2: Valors de NSE i PBIAS pel model realitzat amb SWAT

les prediccions de cabal.

Així, les dades usades per a l'entrenament del model LSTM són les que expliquem a continuació.

### 9.1 Dades climàtiques

Unes de les variables més importants a l'hora de predir el cabal són probablement les dades climàtiques, en concret la precipitació (que es mesura en mm), i la temperatura mitjana diària (°C), ja que determina l'estat en què cau la precipitació (aigua o neu). Aquestes dades són mesurades per les estacions meteorològiques, la localització de les quals es pot veure a la figura 3.9. En total, i després d'eliminar algunes estacions les quals la qualitat de les seves dades no era òptima, ens hem quedat amb 140 estacions de pluviometria i 104 de temperatura.

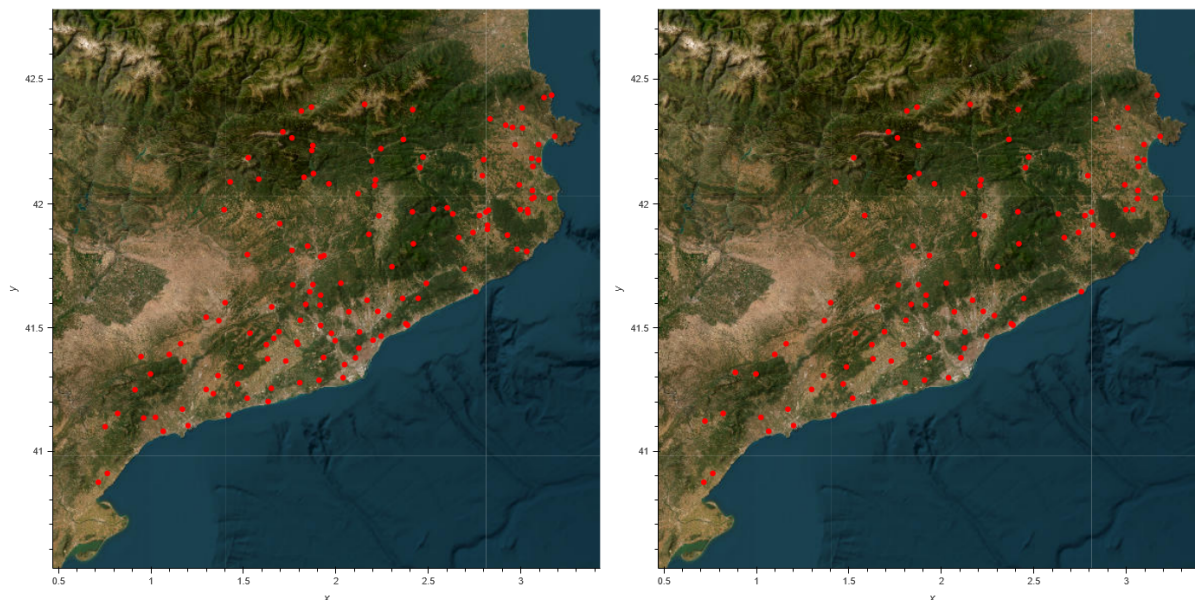


Figura 3.9: Estacions meteorològiques de Catalunya: a l'esquerra, de precipitacions, i a la dreta, de temperatura

## 9.2 Model digital d'elevacions

El model digital d'elevacions (DEM) és una representació digital de la superfície del terreny mitjançant el mostreig de l'altitud del relleu. A partir d'aquest model, es poden delimitar les conques hidrogràfiques, és a dir, regions que drenen aigua cap a una sortida única (un únic punt definit per unes coordenades). Per tant, donat aquest punt podem definir una regió on tota l'aigua que cau en ella (en forma de precipitació o neu) convergeix en aquest punt. D'aquesta regió en podem utilitzar estadístiques com l'elevació mitjana, el pendent mitjà o l'àrea. A la figura 3.10 es pot veure les diferents conques i subconques de les CIC.

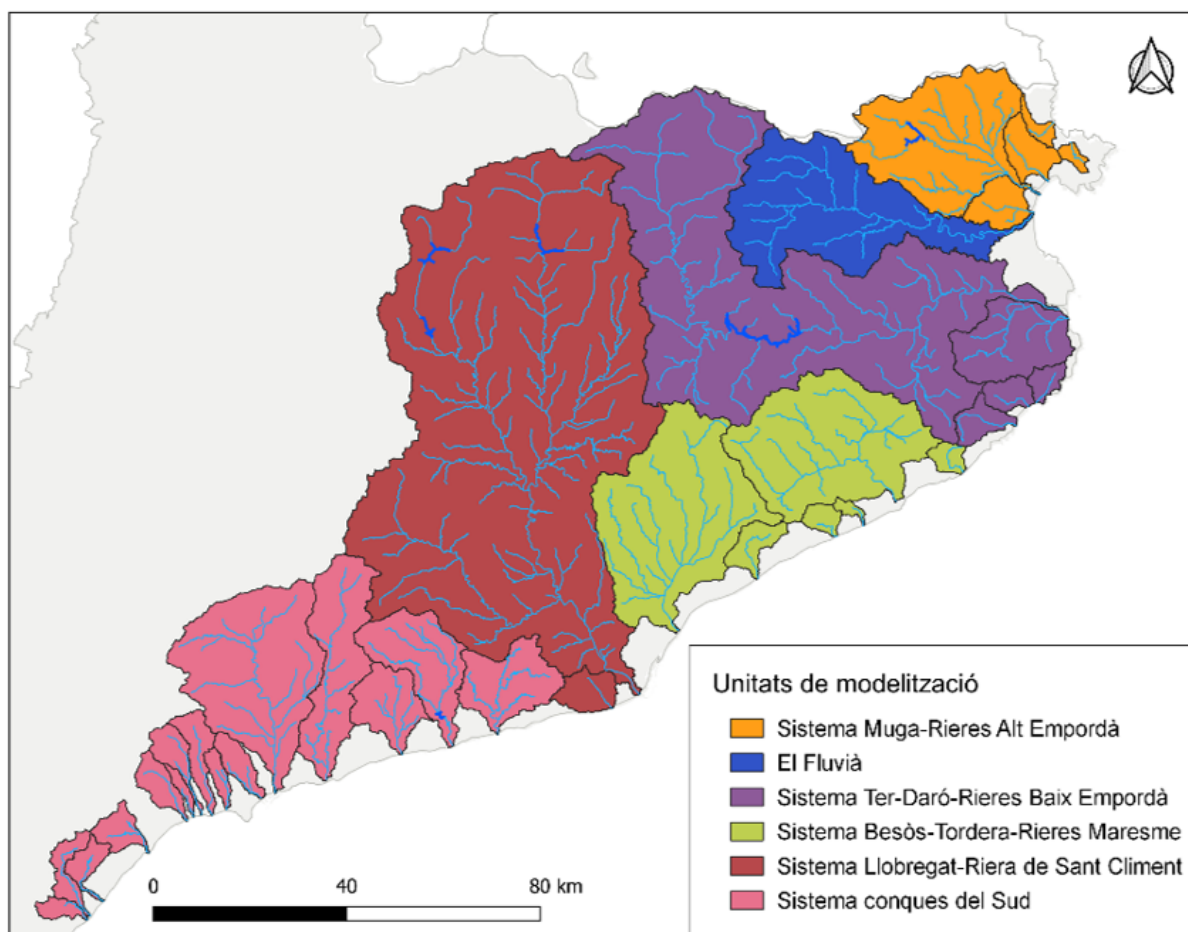


Figura 3.10: Conques i subconques de les CIC

## 9.3 Tipus del sòl

El tipus de sòl és important en la hidrologia, ja que afecta la capacitat d'infiltració de l'aigua, la retenció d'aquesta al sòl, al flux d'aigua subterrània, i a l'escorrentia superficial. Així, el tipus de sòl pot condicionar a com es mou l'aigua i com s'emmagatzema en una conca hidrogràfica. Nosaltres hem utilitzat una versió simplificada de la classificació proposada per l'ICGC (Institut Cartogràfic i Geològic de Catalunya), que consisteix en:

- Sòls desenvolupats a les àrees que presenten un règim de temperatura críic: sòls que ocupen les àrees més fredes del Pirineu, on les temperatures mitjanes del sòl és inferior a 8°C
- Sòls desenvolupats a les àrees que presenten un règim d'humitat údic: sòls que es mantenen humits pràcticament durant tot l'any, i la temperatura mitjana dels quals oscil·la entre els 8 i 15°C.
- Sòls desenvolupats a les àrees que presenten un règim d'humitat ústic: són zones en què els sòls, en condicions naturals, no poden proporcionar tota l'aigua que les plantes necessiten per al seu òptim desenvolupament, tot i que en determinats moments del període de màxim creixement potencial de les plantes, aquests poden disposar d'una certa quantitat d'aigua que els sòls són capaços de retenir i alliberar.
- Sòls desenvolupats a les àrees que presenten un règim d'humitat xèric: zones en què els sòls, en condicions naturals, no poden proporcionar tota l'aigua que les plantes necessiten per al seu òptim desenvolupament, especialment durant el període de màxim creixement.
- Sòls desenvolupats a les àrees que presenten un règim d'humitat arídic: zones en què els sòls es mantenen secs durant llargs períodes de temps, fet que ocasiona una manca important d'aigua per al desenvolupament de les plantes.
- Sòls desenvolupats a les àrees que presenten un règim d'humitat àqüic: aquest règim té lloc en zones que tenen excés d'aigua, i els sòls es caracteritzen per presentar un nivell freàtic a poca profunditat.

Els diferents tipus de sòls de Catalunya es pot veure a la figura 3.11.

#### 9.4 Usos del sòl

L'ús del sòl és el total de les accions, activitats i intervencions realitzades per les persones sobre un tipus de cobertura de sòl, el qual té un efecte important sobre l'aigua. Nosaltres hem considerat únicament agrícola herbaci, agrícola llenyós, bosc, arbusts, zona urbana molt densa, zona urbana poc densa, i zones aquoses. Els diferents usos del sòl considerats sobre les CIC es poden veure a la figura 3.12.

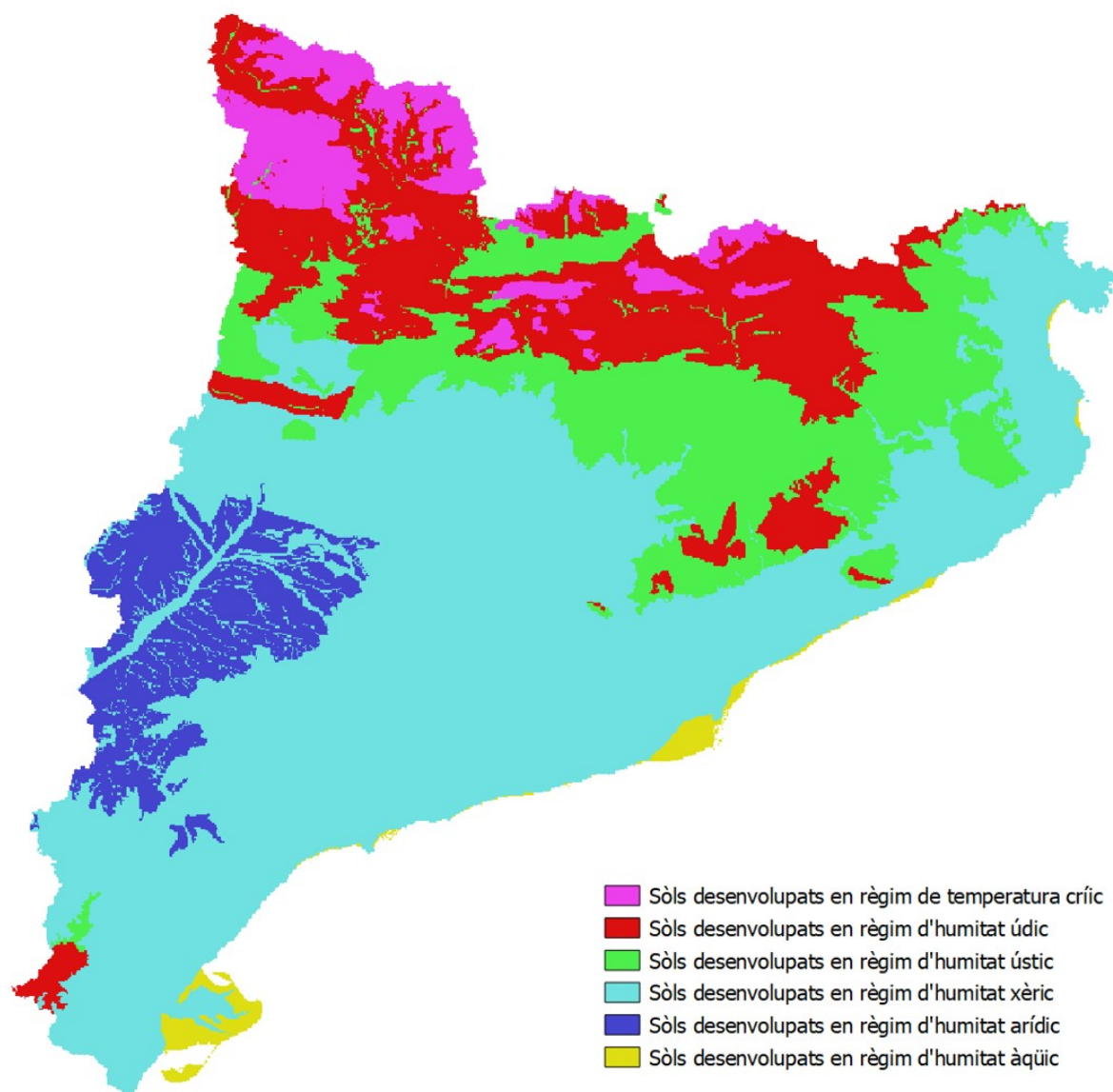


Figura 3.11: Tipus de sòl a Catalunya

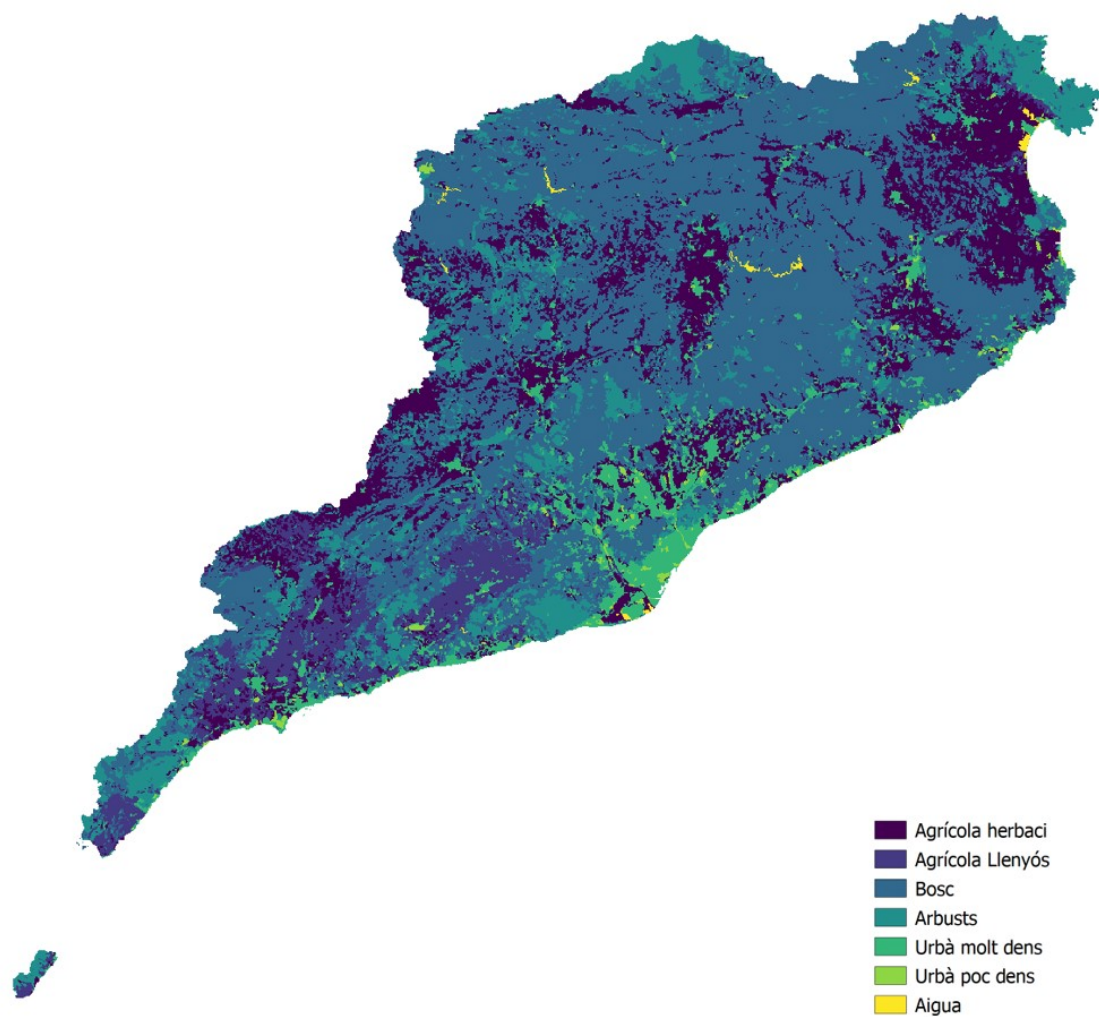


Figura 3.12: Usos del sòl de CIC

## 4. Planificació i metodologia

### 1 Metodologia

La metodologia considerada per al desenvolupament del projecte és CRISP-DM, basada en les sis fases diferents definides a continuació. La seqüència d'aquestes fases no és rígida, sinó que pots anar endavant o enrere repetint fases segons els resultats obtinguts i les necessitats del projecte vagin canviant.

#### 1.1 Enteniment del negoci

Aquesta fase es basa a determinar les necessitats de la indústria o el client, per a continuació definir els objectius del teu projecte en funció dels recursos disponibles, i planificar totes les fases del projecte. En el cas d'aquest treball, la proposta neix a ICRA, amb els objectius explicats a 1, és a dir, millorar la qualitat de les dades de les quals disposaven per tal d'entrenar millors models, i servir com a ajuda als gestors de les estacions d'aforament.

#### 1.2 Enteniment de les dades

Consisteix a recollir totes les dades necessàries per al desenvolupament del projecte, observar-les i explorar-les per poder verificar-ne la qualitat, i si compleixen els requisits necessaris per assolir l'objectiu proposat. L'obtenció i exploració de les dades de les quals hem partit per desenvolupar aquest treball s'ha fet conjuntament amb l'ajuda d'experts a ICRA, amb els quals hem treballat per realitzar l'etiquetatge dels cabals anòmals, i per obtenir i entendre la resta de dades d'entrada dels algorismes (usos i tipus del sòl, etc.).

#### 1.3 Preparació de les dades

Aquesta fase és la que sovint porta més temps, i es basa a determinar quines de les dades obtingudes s'utilitzaran per a la confecció dels models, eliminar-ne (o imputar) valors erronis, definir nous atributs a partir de les dades actuals, formatar les dades en el format correcte, etc. Per realitzar aquesta fase s'han utilitzat, principalment, les llibreries GeoPandas i GDAL (per llegir els fitxers amb informació geogràfica relacionada) i convertir-ho en format tabular, així com pysheds i Xarray per calcular les conques hidrogràfiques i generar totes les estadístiques associades a aquestes, per tal de poder entrenar els models.

## 1.4 Modelatge

En aquest punt, s'escull el model (o models) que poden anar millor segons l'objectiu proposat i les dades obtingudes, s'entrena i s'avalua. En el cas d'aquest projecte, per generar els models s'han utilitzat principalment les llibreries Prophet (per crear el model del Mètode 1), i TensorFlow (per crear i entrenar les xarxes neuronals del mètode 2), juntament amb d'altres com Scikit-learn.

## 1.5 Avaluació

Mentre que en el punt anterior s'avalua el model de forma tècnica, en aquest punt es mira quin és el model que millor compleix les necessitats de la indústria en funció dels resultats obtinguts. També es comprova que tots els passos s'hagin realitzat correctament, i en funció de tot això, determinar els passos següents. Un dels aspectes claus en aquest punt ha estat la visualització de les dades per verificar i entendre els resultats obtinguts, que s'ha realitzat amb HoloViz, una llibreria a alt nivell que engloba l'ús d'altres llibreries relatives a la visualització de dades amb Python.

## 1.6 Desplegament

Aquesta fase depèn bastant dels requisits del projecte, i tant pot anar des de generar una documentació adequada, fins a desplegar el model en un servidor. En general, es podria definir com a la fase que permet accedir al client (o l'usuari final) als resultats del projecte. En el cas d'aquest treball, consisteix en aquesta memòria, escrita amb L<sup>A</sup>T<sub>E</sub>X, juntament amb el repositori de Github clicant aquí o anant al següent enllaç: <https://github.com/zepholus/tfm>.

# 2 Planificació temporal

Si tenim en compte els 15 crèdits que té l'assignatura del TFM i que 1 crèdit ECTS representa 25 hores de feina, a aquest projecte li hauríem d'haver dedicat al voltant de 375 hores. Tenint en compte això, i segons el diagrama de Gantt de planificació temporal (vegeu figura 4.1), podem desglossar la nostra dedicació de la següent manera:

- Tasques de recerca i reunions: 145 hores
- Tasques de programació: 100 hores
- Tasques d'analista: 70 hores
- Redacció de la memòria: 60 hores

Considerant els honoraris de programador i analista a 40€/hora i els d'investigador a 30€/hora, l'import de contractació per dur a terme aquest projecte/treball ascendeix a uns 13.000€. Els impostos estan inclosos en el pressupost, així com totes les despeses i costos indirectes de contractació.

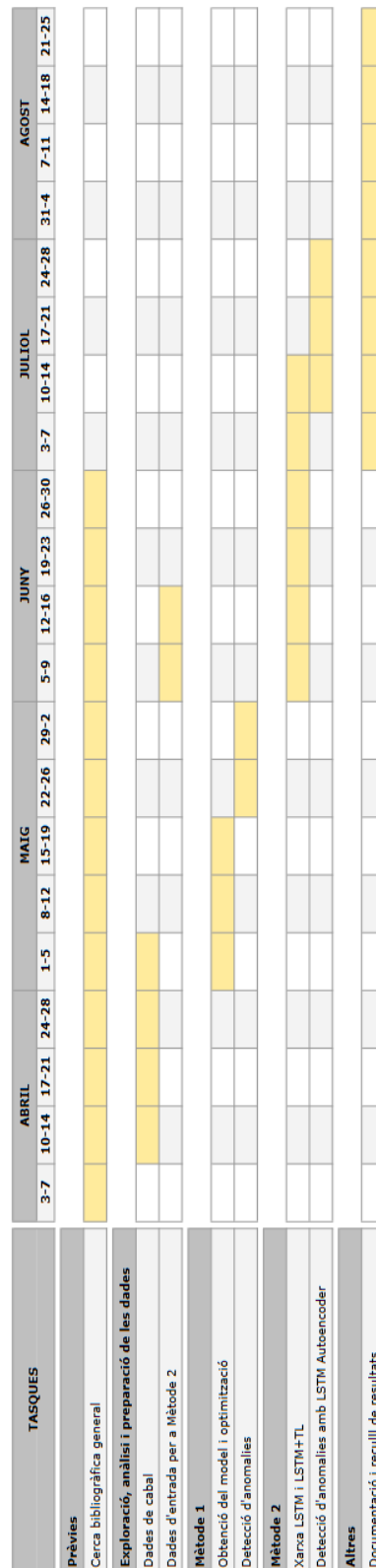


Figura 4.1: Diagrama de Gantt de la planificació temporal



# 5. Contribució metodològica

## 1 Mètode 1

El primer mètode proposa una aproximació no supervisada per tal de detectar les anomalies causades per errors en les estacions d'aforament.

Consisteix a generar un model utilitzant Prophet que s'ajusti bé als valors de cabals observats, no fer una predicció sobre valors futurs. Per això, en aquest cas no aplicarem cap partició de train ni test, sinó que treballarem sobre tot el conjunt de les dades. Com que és una aproximació no supervisada, treballarem sobre les dades en brut, sense filtrar, és a dir, utilitzant tant valors incorrectes com a correctes [1]. Cal destacar, però, que seguint aquesta aproximació, l'NSE i PBIAS obtinguts no són directament comparables amb els altres exposats durant el treball, ja que utilitzant aquest mètode l'algorisme disposa de la totalitat de les dades.

Una de les característiques de la llibreria és que proporciona un output segons un interval de confiança, que per defecte és del 80%. Nosaltres el canviarem al 95% [1], de manera que com a sortida obtindrem la predicció de la mostra i l'interval de confiança d'aquesta.

Aleshores, donada una sèrie temporal, n'apliquem l'algorisme sobre la totalitat d'aquesta, i com a resultat, obtenim una altra sèrie temporal generada per Prophet que s'intenta ajustar a la sèrie original, juntament amb l'interval de confiança de cada mostra.

Per tal d'intentar reduir l'error del model optimitzant els hiperparàmetres, s'ha utilitzat el mètode de Grid Search sobre els hiperparàmetres `changepoint_prior_scale` (amb un rang de [0.001, 0.5]) i `seasonality_prior_scale`, amb un rang de [0.01, 10]. Aquest mètode, hiperparàmetres i rangs seleccionats són els proposats a [14], per tal de permetre al model adaptar-se a la tendència general i evitar el sobre ajustament.

Un cop fet això, podem calcular l'error entre les observacions i prediccions, i la incertesa (diferència entre interval de confiança superior i inferior). Aleshores, si l'error d'una mostra supera el seu valor d'incertesa, considerem aquesta mostra com a anomalia [3].

Tot i la simplicitat d'aquest mètode, la idea és obtenir uns primers resultats que ens serveixin com a punt de partida per al mètode 2.

## 2 Mètode 2

Com ja s'ha comentat, el mètode 2 consisteix primer a modelitzar el cabal de les estacions utilitzant una xarxa LSTM, i després utilitzar aquestes prediccions per entrenar un LSTM Autoencoder que permeti detectar les anomalies. L'objectiu d'utilitzar una sola xarxa neuronal per cada pas, en lloc d'una xarxa per estació, ve fonamentada per diferents motius. El primer, i principal, és que tot i que s'ha realitzat un filtratge previ de les dades i s'ha descartat les principals anomalies, és possible que encara hi hagi dades que siguin errònies. Utilitzant un sol model per totes les conques hidrogràfiques, podem en certa manera arribar a compensar aquests errors, ja que aquestes xarxes són capaces d'extrapolar i aplicar els coneixements apresos d'una conca (on les dades poden ser correctes) a una altra de diferent on les dades puguin ser de menys qualitat [10]. De la mateixa manera, un cop entrenat el model, aquest ens serviria per predir el cabal en conques on no tinguéssim estacions d'aforament (només necessitariem estacions climatològiques), fet que faria molt més aprofitable aquest model. El darrer motiu és que el fet d'utilitzar aquest tipus d'entrenament ens permet entrenar el model amb moltes més dades, de les quals pot haver-hi més patrons, i per tant, la xarxa pot extreure més informació.

Pel que fa a l'LSTM Autoencoder, la intenció és entrenar-lo utilitzant tant les prediccions com les observacions de les estacions d'aforament donat una finestra de 14 dies (longitud definida per criteri expert). L'objectiu és que la xarxa LSTM entengui i compregui bé el comportament del cabal en cada conca, i un cop això, l'autoencoder aprengui els patrons correctes, per exemple:

- Si les prediccions indiquen que el cabal ha d'augmentar (degut a precipitacions), el cabal observat també ho ha de fer
- En cas que no hi hagi precipitacions, el cabal observat ha d'anar tendint cap al cabal basal indicat per les prediccions. Un cop així, s'ha de mantenir més o menys estable, fins que torni a haver-hi precipitacions

Noti's el fet que necessitem realitzar la predicció del cabal, ja que si aquesta és correcta, ens serveix com a pauta o guia per tal de descartar observacions. És a dir, aquesta predicció ens proporciona el context necessari per dictar si un conjunt de mostres són correctes o no.

Tot i que també hàgim pogut entrenar un classificador LSTM, el fet que les dades siguin extremament desbalancejades fa que ens hàgim decantat per l'opció de l'autoencoder.

### 2.1 Entrenament de la xarxa LSTM

Així doncs, per tal d'obtenir i preparar les dades per poder entrenar la xarxa, partim de les coordenades de les estacions d'aforament (i de les observacions de cabal considerades com a correctes), de les quals n'hem extret la conca hidrogràfica associada a partir del DEM. Per cadascuna d'aquestes en calculem la superfície, elevació i pendent mitjà, i els tipus i usos del sòl associats. En relació amb les dades climàtiques, trobem les estacions de temperatura i pluviometria localitzades a la conca en qüestió, i en calculem la mitjana. Així, assumim que la temperatura i precipitació de tota la conca durant un dia concret és la mitjana de valors d'aquest dia de les estacions meteorològiques ubicades dins la conca. Si després de calcular la conca, veiem que no tenim estacions meteorològiques associades, descartem les dades d'aquesta

estació. Per tal d'omplir els valors nuls d'aquestes estacions, aquests s'han substituït a partir de mitjanes de cada mes (i al llarg dels anys).

Donat tot el període en què tenim observacions, s'han agafat les dades corresponents al primer 70% com a conjunt de train, les dades corresponents al següent 15% com a conjunt de validate, i les observacions pertanyent el 15% restant s'han utilitzat com a dades de test. D'aquesta manera, ajudem a reduir el sobreajustament típic d'aquestes xarxes.

En el nostre cas, després d'haver descartat les estacions d'aforament sense estacions meteorològiques associades, així com les dades anòmales, ens hem quedat amb 40 estacions d'aforament disponibles, i 177580 registres del conjunt de train, 34266 del conjunt de validate, i finalment, 33498 del conjunt de test.

Per tal de millorar i agilitzar l'entrenament, les dades s'han estandarditzat restant la mitjana  $\mu$  i dividint per la desviació estàndard  $\sigma$ , on  $\mu$  i  $\sigma$  han sigut calculats a partir de totes les dades de train.

Com ja hem vist abans, les observacions de cabal tenen molts valors buits, a causa de fallades dels sensors. Per tal de solucionar aquest problema, s'han dissenyat dos experiments diferents, amb un model per cada experiment: el model LSTM i el model LSTM + Transfer Learning (LSTM+TL).

- En el model LSTM, els valors buits de les observacions considerades com a correctes s'emplenen utilitzant les seves respectives prediccions del model SWAT, i a continuació es fa optimització d'hiperparàmetres (entrenant sobre el període de train, i avaluant amb el període de validate). Per últim, amb la millor configuració, es reentrena la xarxa utilitzant les dades del període de train i validate.
- El model LSTM+TL està basat en dues fases: primer s'entrena la xarxa LSTM únicament a partir de les prediccions del model SWAT dins del període de train, de manera que la xarxa aprèn els comportaments físics modelats, i després es realitza optimització d'hiperparàmetres (novament, avaluat sobre el període del conjunt de validate). En el segon pas, únicament es reentrena la capa densa (última capa), utilitzant solament les observacions correctes de les estacions d'aforament amb les dades dels conjunts de train i validate, seguint una aproximació similar al Transfer Learning [16].

La manera com hem generat les dades per alimentar els dos models ha sigut la següent: suposant  $n$  dies de finestra de look-back, per cada dia  $d$  i cada estació s'ha construït les parelles  $([X_{d-n}, \dots, X_{d-1}], y_d)$ , on  $X_d$  és el conjunt de variables explicatives per predir el cabal del dia  $d$ , i  $y_d$  és el valor observat de cabal (considerat com observació correcta) per al dia  $d$ .

Pel que fa a l'optimització dels hiperparàmetres (realitzada amb optimització bayesiana), els paràmetres a optimitzar i el rang de valors considerat en els dos experiments ha sigut similar al seguit a [9]. En concret, s'ha optimitzat la dimensió de l'estat ocult (triant un rang entre 64 i 256), i el dropout, amb què s'han provat els valors de 0.0, 0.15, 0.3 i 0.45.

Per tal d'agilitzar l'optimització, s'ha fixat el valor de la mida de batch a 512, ja que normalment valors grans afavoreixen a la convergència de la xarxa. Pel que fa al nombre d'epochs, hem triat un valor de 30 (hem comprovat manualment que era un període suficient), afegint la condició que si durant 4 epochs no es reduïa l'error, es parava l'entrenament (Early Stopping).

També s’han provat dos tipus d’error diferents a minimitzar durant l’entrenament: el Mean Squared Error (MSE), ja que en ser una mesura d’error quadràtica els paràmetres que minimitzen l’MSE solen ser els que maximitzen el NSE [9], i el logMSE, per solucionar alguns dels errors derivats de l’MSE, tal com es veurà a 6. La mida de la finestra de look-back ha sigut de 150 dies (havent provat entre 90, 150, 270 i 365 dies), amb una sola capa LSTM (havent provat entre 1, 2 i 4). Aquests últims valors han estat triats de forma empírica.

Per comparar els resultats d’aquests dos mètodes, els errors NSE i PBIAS s’han calculat únicament a partir de les observacions correctes de les estacions d’aforament durant el període de test.

Després de l’optimització, la millor configuració obtinguda pel model LSTM ha estat de 192 estats ocults i un dropout de 0. Pel que fa al model LSTM+TL, la configuració òptima ha estat de 224 estats ocults, i un dropout de 0.3.

## 2.2 Detecció d’anomalies

Per entrenar l’autoencoder, novament hem partit únicament de les observacions correctes, és a dir, les finestres de  $n$  dies (on, com hem dit abans, considerem que  $n = 14$ ) on el cabal observat de l’últim dia no és anòmal. Per tal de substituir els valors buits hem seguit l’estratègia d’assignar el cabal predit pel model SWAT. Per tant, per cada dia  $d$  on el cabal observat és correcte hem construït la seqüència  $[(X_{d-n}^m, X_{d-n}^o), \dots, (X_d^m, X_d^o)]$  per tal d’entrenar l’autoencoder, on  $X_d^m$  és el cabal modelat i  $X_d^o$  és el cabal observat per al dia  $d$ .

Per tal de normalitzar les dades s’ha seguit una estratègia una mica particular. Com que a nosaltres, donades dues seqüències  $[(X_{a-n}^m, X_{a-n}^o), \dots, (X_a^m, X_a^o)]$  i  $[(X_{b-n}^m, X_{b-n}^o), \dots, (X_b^m, X_b^o)]$ , on  $a \neq b$ , no ens interessa comparar la diferència d’escala entre aquestes, sinó únicament els patrons (és a dir, com es comporta el cabal observat respecte del modelat), el que hem fet ha estat normalitzar les dades de cada seqüència de forma independent. En altres paraules, per cada seqüència agafem els cabals observats i modelats, i escalem els valors entre 1 i 0, de manera que 1 es correspongui al màxim del cabal (d’entre les observacions i les prediccions d’aquella seqüència), i 0 al mínim.

Per la xarxa neuronal de l’encoder, s’han utilitzat dues capes LSTM, la primera on la dimensió de l’estat ocult és de 32, i la segona de 16 (per tant, la mida de la capa *code* és de 16 cel·les), mentre que pel decoder s’ha utilitzat una estructura inversa (la longitud de l’estat ocult de la primera capa LSTM era de 16, i la segona de 32), entrenat sobre el MAE (Mean Absolute Error).

Per tal de fer la partició de train, validation i test hem agafat els mateixos períodes de partició que per entrenar la xarxa LSTM, és a dir, les dades corresponents al primer 70% de la sèrie s’han usat com a train, el següent 15% de validate, i el 15% restant de conjunt de test (així com les dades de totes les estacions descartades, on totes les dades són anomalies). L’autoencoder ha estat entrenat, com ja hem dit, sobre les dades correctes dels conjunts de train i validate, i la detecció d’anomalies s’ha avaluat sobre les dades del conjunt de test.

Un cop fet això, podem calcular una distribució dels errors de la reconstrucció de les dades utilitzant l’autoencoder, i mirar si hi ha alguna diferència entre les distribucions obtingudes en quatre conjunts diferents: el de les dades correctes del període de train i validation, el de les dades incorrectes del període de train i validation, i de forma similar per les dades del conjunt de test (distingint entre correctes i incorrectes). D’aquesta manera, si l’autoencoder és capaç

de detectar les anomalies, esperem que la distribució d'errors de les dades errònies mostri un comportament diferent del de les dades correctes (esperem que les prediccions de les dades correctes siguin millor que el de les dades incorrectes), i puguem establir un llindar a partir del qual considerar quan una mostra és una anomalia.

## 6. Resultats

### 1 Mètode 1

La mitjana dels errors, el valor mínim i el màxim de les prediccions dels cabals desglossat per conca utilitzant Prophet es pot veure a la taula 6.1 (com ja hem explicat, considerant el conjunt de dades complet). Es pot veure com els NSE no són acceptables, tot i que en alguns casos, millor que en el model SWAT (en concret, pel sistema de conques del Sud i per la conca del Llobregat).

| Unitat de modelització | NSE     |         |       |       | PBIAS   |         |        |       |
|------------------------|---------|---------|-------|-------|---------|---------|--------|-------|
|                        | Mitjana | Mediana | Mínim | Màxim | Mitjana | Mediana | Mínim  | Màxim |
| Llobregat              | 0.21    | 0.16    | 0.04  | 0.63  | 0.0006  | 0.001   | -0.02  | 0.02  |
| Muga                   | 0.07    | 0.05    | 0.02  | 0.16  | 0.005   | 0.007   | -0.02  | 0.03  |
| Conques del Sud        | 0.12    | 0.08    | 0.05  | 0.28  | 0.003   | 0.001   | -0.003 | 0.01  |
| Fluvià                 | 0.05    | 0.05    | 0.04  | 0.06  | -0.001  | -0.006  | -0.01  | 0.01  |
| Besòs-Tordera          | 0.11    | 0.06    | 0.03  | 0.47  | -0.003  | -0.003  | -0.01  | 0.02  |
| Ter                    | 0.1     | 0.1     | 0.02  | 0.19  | 0.006   | 0.004   | -0.02  | 0.04  |

Taula 6.1: Mitjana dels errors de les prediccions de cabal de cada estació utilitzant Prophet

Pel que fa als errors de les prediccions de les anomalies, es pot veure a la taula 6.2

| TN  | FP    | FN  | TP    | Precision | Recall | F1   |
|-----|-------|-----|-------|-----------|--------|------|
| 84% | 0.47% | 15% | 0.09% | 0.15      | 0.01   | 0.01 |

Taula 6.2: Mitjana de les estadístiques de les deteccions d'anomalies de cada estació utilitzant Prophet

Podem veure que els resultats no són acceptables. A la figura 6.1 hi ha representat les prediccions dels cabals i de les anomalies per l'estació d'Olvan (Berga), i és un clar exemple de per què els resultats del Mètode 1 no són satisfactoris: tot i que les prediccions del cabal semblen seguir la tendència més general, no és capaç d'identificar els canvis (en gran part lògic, perquè aquí no tenim informació de variables externes, com les precipitacions). On clarament falla, però, és en les prediccions de les anomalies: aquest mètode només detecta valors extrems (per sobre o per sota del que s'esperaria), però no és capaç d'identificar els patrons erronis dins del rang de l'interval de confiança (que són la majoria).

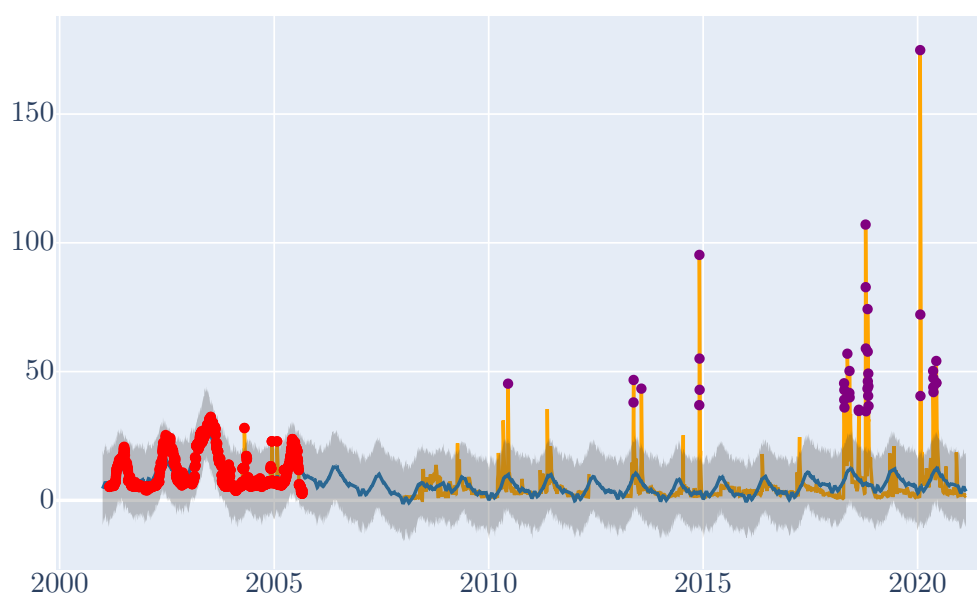


Figura 6.1: Cabal observat (taronja) i predit (blau) de l'estació d'Olvan. En vermell, les anomalies reals no predites (FN), i en lila, les anomalies no observades predites (FP)

## 2 Mètode 2

### 2.1 Entrenament de la xarxa LSTM

#### Model LSTM

El resultat de l'entrenament de la xarxa neuronal pel model LSTM utilitzant com a mètrica d'error a minimitzar el logMSE es pot veure a la figura 6.2. Podem apreciar com l'entrenament no acaba de ser del tot satisfactori, ja que a mesura que anem augmentant els epochs, l'error sobre el conjunt d'entrenament va disminuint, però en canvi, amb pocs epochs l'error comès sobre el conjunt de validació s'estanca, fet que indica un sobreajustament sobre les dades d'entrenament.

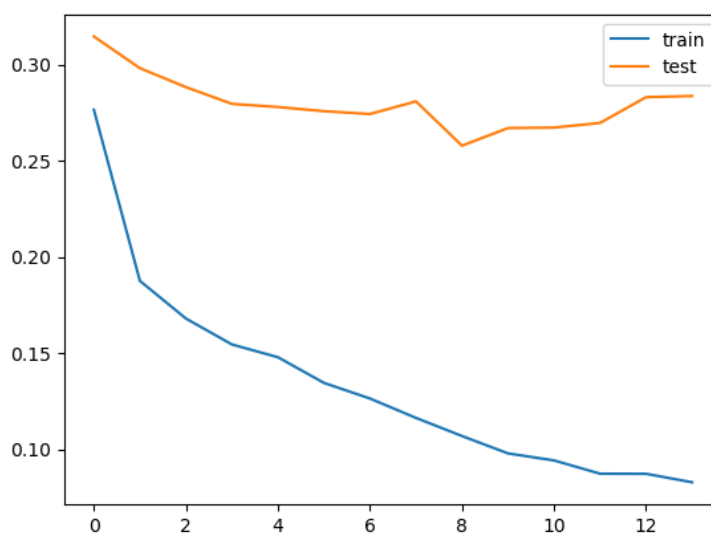


Figura 6.2: Entrenament de la xarxa LSTM

Tot i això, si calculem els errors de les prediccions de cabals sobre les observacions correctes de les estacions d'aforament al conjunt de test, podem apreciar com sense tenir un NSE massa elevat, són millors que amb el Mètode 1, tal com es veu a la taula 6.3.

| Unitat de modelització | NSE     |         |       |       | PBIAS   |         |       |       |
|------------------------|---------|---------|-------|-------|---------|---------|-------|-------|
|                        | Mitjana | Mediana | Màxim | Mínim | Mitjana | Mediana | Màxim | Mínim |
| Llobregat              | 0.08    | 0.12    | 0.45  | -0.85 | -44     | -41     | -24   | -83   |
| Muga                   | -0.03   | 0.009   | 0.21  | -0.39 | -4      | -1      | 40    | -54   |
| Conques del Sud        | 0.3     | 0.26    | 0.5   | 0.19  | -22     | -25     | 3.8   | -42   |
| Fluvià                 | 0.17    | 0.17    | 0.26  | 0.09  | -48     | -48     | -40   | -55   |
| Besòs-Tordera          | 0.22    | 0.21    | 0.27  | 0.18  | -38     | -29     | -22   | -63   |
| Ter                    | -0.86   | 0.09    | 0.4   | -8.9  | -33     | -44     | 31    | -64   |

Taula 6.3: Mitjana dels errors de les prediccions de cabal de cada estació utilitzant el model LSTM, separat per conca

Si mirem la distribució espacial dels resultats obtinguts (figura 6.3) podem observar un patró



clar: les estacions que estan situades a les capçaleres dels rius (i per tant, estan per sobre dels embassaments) són les que tenen un NSE més elevat. Aquest patró és lògic, ja que en ser subconques més petites i per sobre d'embassaments, modelitzar el cabal a través de les precipitacions és més fàcil. En concret, la mediana de l'NSE de les estacions per sobre dels embassaments és de 0.21, mentre la de les que estan per sota és de 0.08. Pel que fa a les estacions amb un NSE inferior a 0 (en concret en tenim 4), cal comentar que tot i que la tendència general pot ser considerada com a bona (en general), en algun fem una predicció d'alguns pics de cabal inexistents (no observats). El detall d'una d'aquestes últimes estacions es pot veure a la figura 6.4. Tot i això, però, en la majoria dels casos el model sí que és capaç de modelar l'augment de cabal en la freqüència temporal que toca, però no en l'escala adequada, és a dir, els pics de cabal modelats sovint són inferiors als observats. De totes maneres, podem afirmar que el model entén la dinàmica general de la hidrologia.

### Model LSTM+TL

Tot i que, com es pot veure a la taula 6.4 en algunes conques l'NSE del model LSTM+TL (utilitzant com a mètrica d'error a minimitzar el logMSE) és més elevat que en el model LSTM, el comportament hidrològic general és més precís en el cas d'aquest últim model, ja que capta millor els pics de cabal (solen ser superiors). El fet que alguns casos l'NSE del model LSTM+TL sigui més elevat és degut al fet que ambdós models a vegades prediuen pics inexistents (o no observats), com ja havíem comentat, i en aquests casos el fet que en el model LSTM els pics predits siguin més elevats té una penalització sobre l'NSE. A la figura 6.7 es pot veure una comparació entre els dos models.

| Unitat de modelització | NSE     |         |       |       | PBIAS   |         |       |       |
|------------------------|---------|---------|-------|-------|---------|---------|-------|-------|
|                        | Mitjana | Mediana | Màxim | Mínim | Mitjana | Mediana | Màxim | Mínim |
| Llobregat              | 0.09    | 0.08    | 0.4   | -0.5  | -45     | -45     | -20   | -75   |
| Muga                   | 0.04    | 0.05    | 0.1   | -0.02 | -42     | -37     | -23   | -70   |
| Conques del Sud        | 0.32    | 0.28    | 0.56  | 0.16  | -28     | -25     | -7    | -56   |
| Fluvià                 | 0.15    | 0.15    | 0.21  | 0.09  | -48     | -48     | -42   | -53   |
| Besòs-Tordera          | 0.23    | 0.2     | 0.3   | 0.18  | -22     | -20     | -7    | -40   |
| Ter                    | -0.15   | 0.11    | 0.3   | -2.5  | -39     | -38     | -12   | -59   |

Taula 6.4: Mitjana dels errors de les prediccions de cabal de cada estació utilitzant el model LSTM+TL, separat per conca

### Entrenament amb MSE o logMSE

Si a l'entrenament de la xarxa neuronal intentem minimitzar l'error MSE en lloc de logMSE (com havíem mostrat fins ara), ens trobem amb la següent situació: tot i que la mediana de l'NSE entre totes les estacions augmenta (en el cas del model LSTM, passem de 0.18 a 0.26), el fet que el càlcul de l'error sigui quadràtic (sense aplicar el logaritme com fa el logMSE) fa que l'entrenament prioritzi la reducció d'errors en cabals alts, i no tant en cabals petits. Per tant, el que acaba passant és que el model capta millor els pics de cabal, però la resta de període de



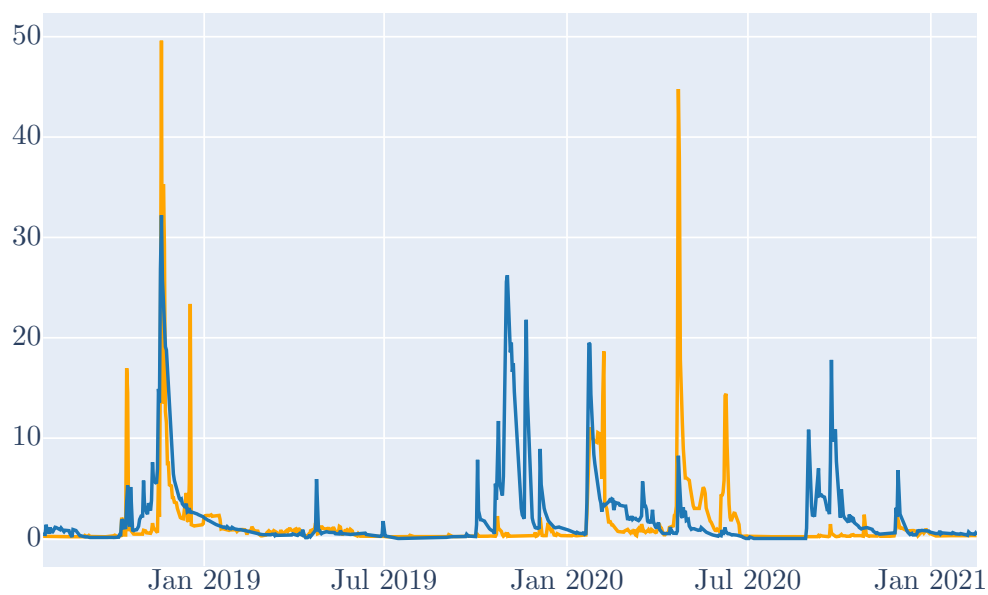


Figura 6.4: Cabal observat (taronja) i predit (blau) per a l'estació de Pont de Molins, amb un NSE de -0.17, utilitzant el model LSTM

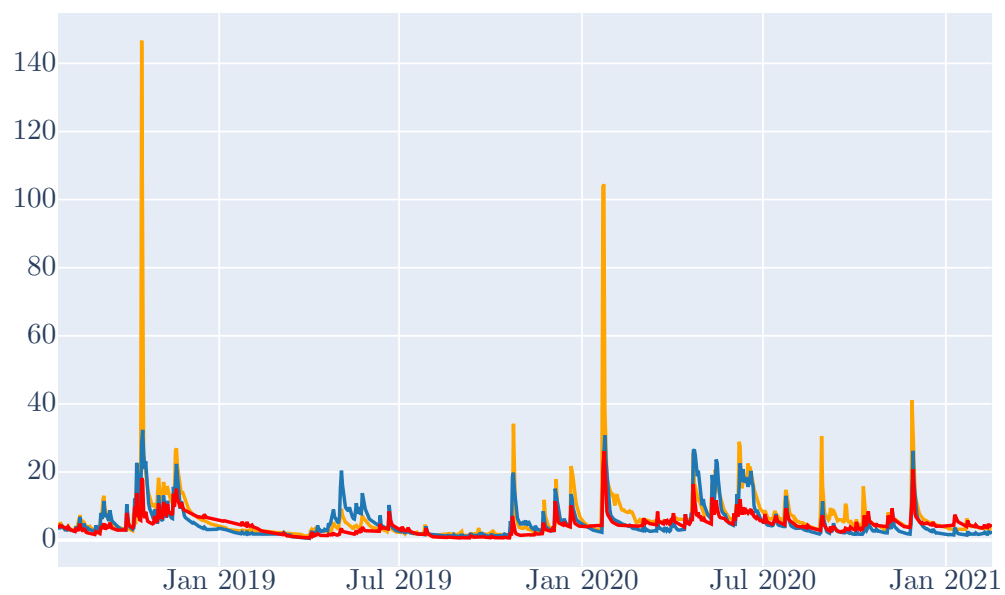


Figura 6.5: Cabal observat a l'estació de Sant Joan de les Abadesses (en taronja), cabal predit pel model LSTM (blau) i cabal predit pel model LSTM+TL (vermell)

temps, on el cabal sol ser baix, sovint fa una predicció de 0, la qual cosa no ens interessa per fer la detecció d'anomalies, com es pot veure a la figura 6.6

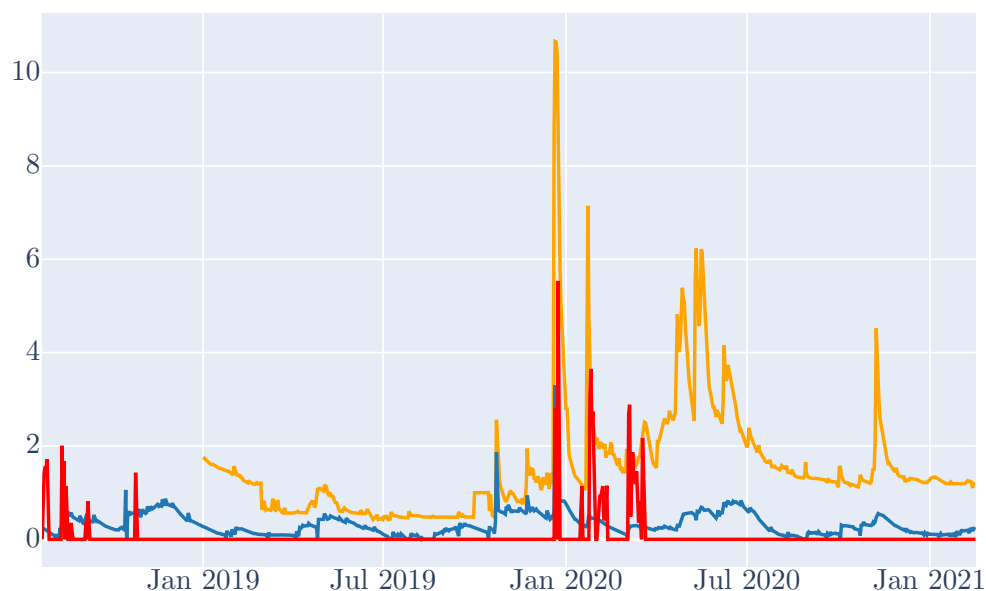


Figura 6.6: Prediccions amb el model LSTM entrenat amb logNSE (blau), entrenat amb NSE (vermell) contra les observacions (taronja)

### Comportament del model LSTM en estacions d'aforament descartades

Si apliquem el model LSTM (el que ha sortit millor dels dos) en conques on s'han descartat la totalitat de les observacions perquè han sigut considerades anomalies, i per tant, el model no ha pogut veure aquestes dades, veiem comportaments diferents.

Per exemple, en el cas de l'estació de Campllong, com que tenim dos comportaments hidrològics lleugerament diferents i de longitud similar, és difícil a simple vista dir quin és el bo i quin és el dolent. En aquest cas, veiem com les prediccions s'ajusten prou bé al primer tram de les observacions, per tant, podríem descartar el segon tram i considerar-les anomalies.

En canvi, en d'altres estacions, les prediccions no són prou bones i no ens serveixen per filtrar les mostres bones de les dolentes.

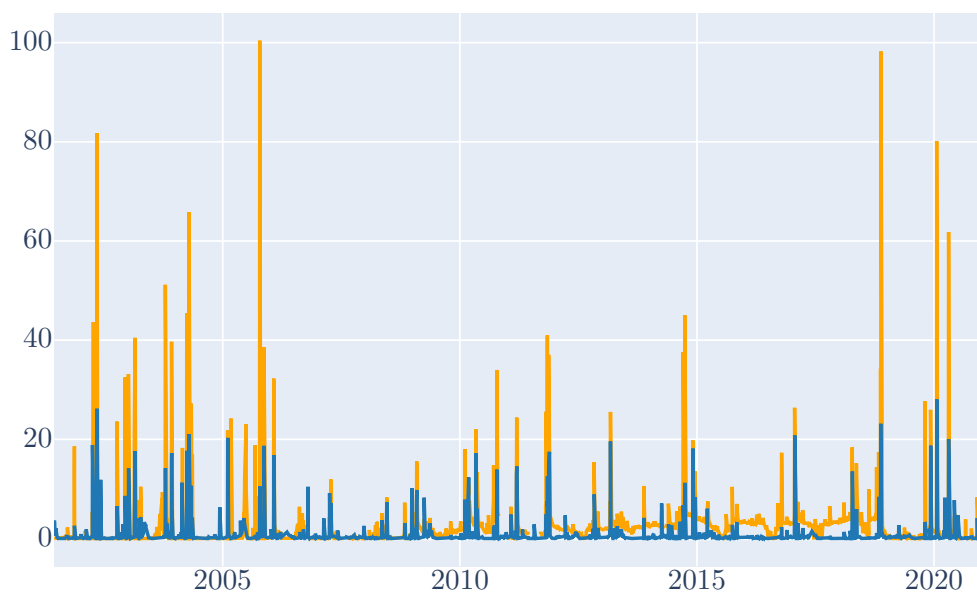


Figura 6.7: Cabal observat (taronja) i predit (blau) per a l'estació de Campllong (riera de La Gotarra), on veiem dos comportaments diferents de les observacions

## 2.2 Detecció d'anomalies

El resultat d'entrenar l'autoencoder a partir de les mostres correctes del període de train es pot veure a la figura 6.8. En concret, l'entrenament sembla ser correcte, assolint on MAE baix.

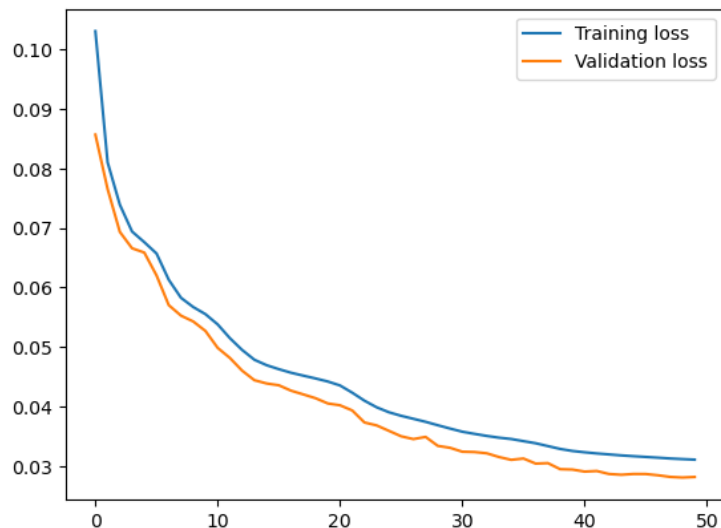


Figura 6.8: Entrenament de la xarxa LSTM Autoencoder

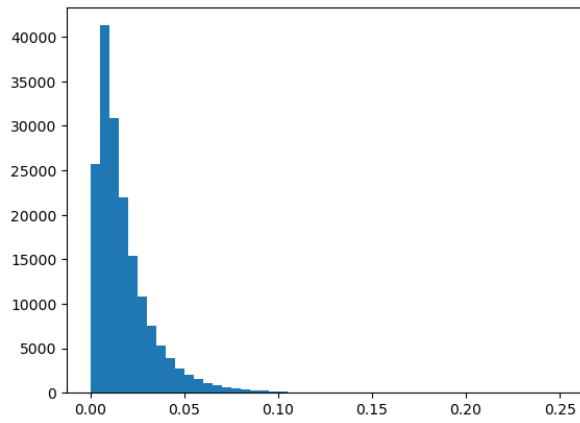
Un cop entrenada la xarxa, podem fer una reconstrucció de les mostres i obtenir els errors comesos en aquestes reconstruccions. Si generem els histogrames d'aquests errors (figura 6.9), hauríem de ser capaços de definir visualment un llindar d'error, ja que les mostres correctes haurien de tenir un error inferior a les mostres incorrectes. Veiem, però, com la distribució és quasi idèntica als histogrames mostrats, cosa que indica que l'autoencoder no és capaç de diferenciar les mostres anòmales. En concret, veiem com la reconstrucció de la majoria de les mostres té un error centrat al voltant de 0.001.

De fet, a la figura 6.10 podem veure com les prediccions generades no tenen per què jugar un factor determinant a l'hora de decidir si una mostra és anòmala o no, ja que aquestes no sempre són prou exactes. De fet, aquesta era la suposició inicial, on assumíem que les mostres vàlides serien més exactes que la de les mostres invàlides.

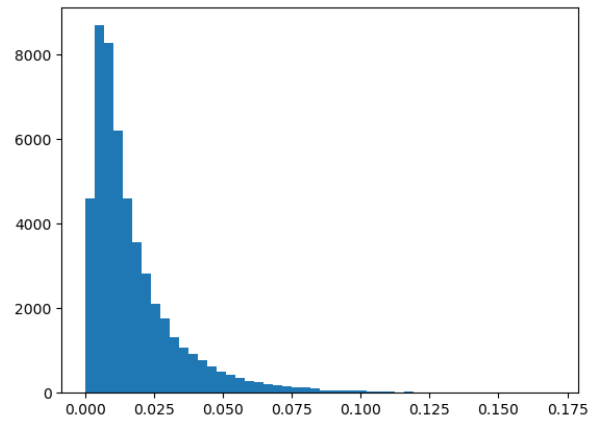
Per tal de donar una puntuació i comparar aquest mètode amb el Mètode 1 (Prophet), triarem com a llindar el percentil 95, que és de 0.04. El rendiment de la detecció d'anomalies amb el llindar establert es pot veure a la taula 6.5, i podem comprovar com els resultats no han sigut satisfactoris, com bé suposàvem anteriorment.

| TN  | FP | FN  | TP   | Precision | Recall | F1   |
|-----|----|-----|------|-----------|--------|------|
| 77% | 5% | 16% | 0.6% | 0.16      | 0.007  | 0.01 |

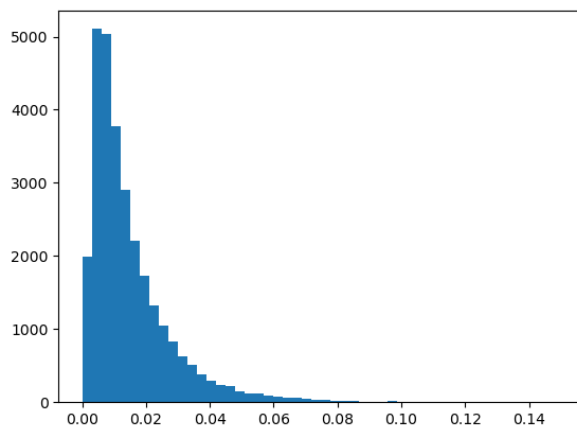
Taula 6.5: Mitjana de les estadístiques de les deteccions d'anomalies de cada estació utilitzant un autoencoder



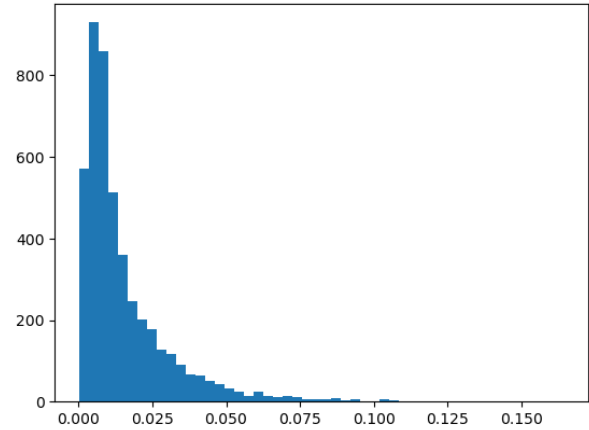
(a) Distribució d'errors en mostres correctes durant el període de train



(b) Distribució d'errors en mostres anòmales durant el període de train



(c) Distribució d'errors en mostres correctes durant el període de test



(d) Distribució d'errors en mostres anòmales durant el període de train i estacions descartades

Figura 6.9: Distribució dels errors obtinguts durant la reconstrucció de les mostres utilitzant un LSTM autoencoder



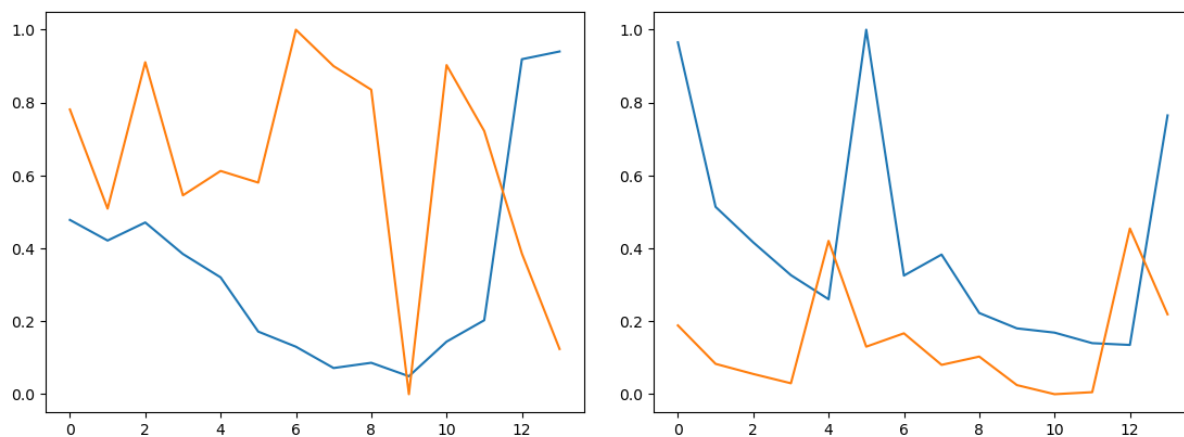
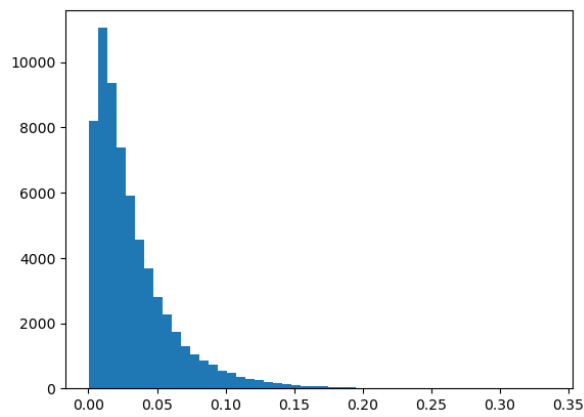


Figura 6.10: A l'esquerra, observació correcta, i observacions i prediccions corresponents als 13 dies anteriors. A la dreta, el mateix però amb una mostra anòmala. El color taronja representa les observacions, i el blau les prediccions

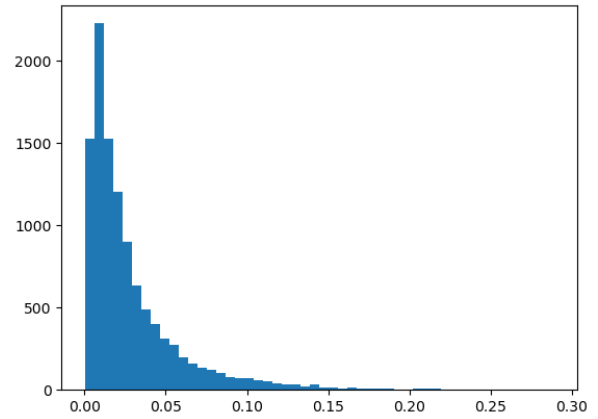
### Detecció d'anomalies amb les estacions amb prediccions més bones

Tenint en compte els resultats anteriors, ens preguntem si aquest mal resultat de la detecció d'anomalies és degut a les males prediccions modelades en algunes estacions, que juntament amb la subtilesa del comportament d'algunes anomalies, pot confondre l'autoencoder. Aleshores, si usem només les dades de les estacions on les prediccions són millors (fem servir les que tenen un NSE superior a 0.2), i avaluem el nou model sobre aquestes mateixes estacions, obtenim la distribució d'errors que es pot veure a la figura 6.11, on novament, podem veure com la xarxa no és capaç de percebre cap diferència entre les mostres anòmales de les normals.

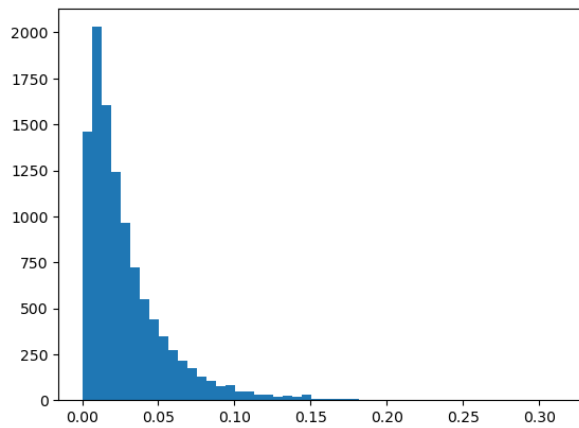
A la figura 6.12 podem tornar a comprovar com, inclús amb les estacions que tenen un NSE superior a 0.2, a escala individual les prediccions de mostres anòmales poden arribar a semblar més correctes que prediccions de mostres normals, fet que impossibilita l'entrenament d'un model utilitzant les prediccions LSTM.



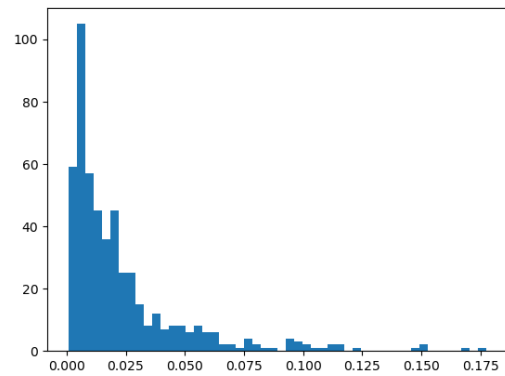
(a) Distribució d'errors en mostres correctes durant el període de train



(b) Distribució d'errors en mostres anòmales durant el període de train



(c) Distribució d'errors en mostres correctes durant el període de test



(d) Distribució d'errors en mostres anòmales durant el període de train i estacions descartades

Figura 6.11: Distribució dels errors obtinguts durant la reconstrucció de les mostres de les estacions amb un  $NSE > 0.2$  utilitzant un LSTM autoencoder

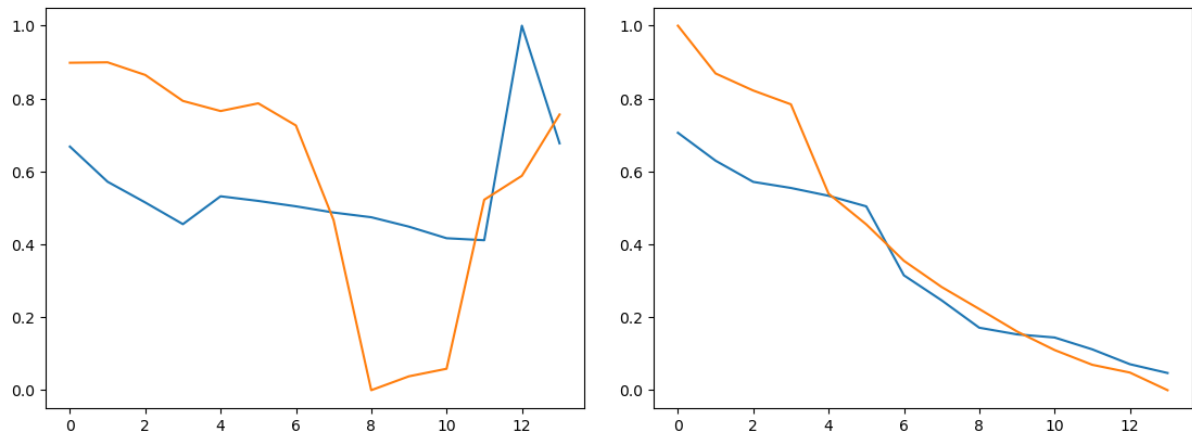


Figura 6.12: A l'esquerra, observació correcta, i observacions i prediccions corresponents als 13 dies anteriors. A la dreta, el mateix però amb una mostra anòmala. El color taronja representa les observacions, i el blau les prediccions

## 7. Conclusions

A tall de conclusions, podríem dir que en general els algorismes proposats no han sigut capaços de detectar les anomalies detectades, en gran part a causa de la complexitat d'aquestes (com hem vist a 3, les anomalies del nostre treball no es poden detectar mitjançant mètodes estadístics clàssics, sinó que depenen en gran part del comportament de la sèrie i un conjunt de patrons concrets). Hem de mencionar la dificultat afegida de treballar en conques de clima mediterrani, caracteritzada per cabals molt baixos i una irregularitat pluviomètrica anual molt marcada, cosa que dificulta les prediccions.

En concret, pels dos algorismes presentats, el mètode 1 és el que ha donat pitjors resultats, ja que, pel tipus d'anomalies que volem detectar, entrenar un model amb Prophet (que no és capaç de comprendre la complexitat de les dades) i traçar un interval de confiança a partir del qual classificar les mostres com a correcta o incorrecta no és una aproximació vàlida.

Pel que fa al mètode 2, hem presentat dos mètodes d'imputació diferents per tal d'entrenar una xarxa LSTM per predir el cabal a partir de variables externes com la precipitació i temperatura. Amb les observacions correctes, juntament amb les prediccions anteriors respectives, podem entrenar un LSTM Autoencoder per a la detecció d'anomalies. Les prediccions de cabal de la xarxa LSTM ha sigut prou correcte (tot i que la mitjana dels NSE de les prediccions està al voltant del 0.2, el comportament d'aquestes és correcte, sobretot en les estacions situades per damunt dels embassaments), i tot i el rendiment inferior als obtinguts amb SWAT+, cal comentar la dificultat que suposa treballar amb SWAT+ respecte d'una xarxa neuronal, ja que aquesta última és molt més ràpida i flexible. Per tant, el model LSTM es pot utilitzar com a alternativa en segons quins casos. En relació amb la detecció d'anomalies, l'LSTM Autoencoder no ha funcionat correctament, ja que donades les prediccions de mostres correctes i incorrectes, aquestes no són prou diferents per poder detectar les anomalies. Això pot ser degut al fet que les prediccions no són prou bones (l'NSE és millorable), per tant, hauríem de continuar treballant en aquest sentit per disminuir els errors, per exemple, utilitzant xarxes neuronals basades en grafs en lloc de xarxes LSTM, i integrar les dades disponibles dels embassaments, modelant totes les dependències espacio-temporals en una sola xarxa neuronal. Tot i això, amb les prediccions de la xarxa LSTM obtingudes, realitzant una comparació visual sobre les observacions, podem ser capaços de detectar algunes anomalies puntuals.

# Bibliografia

- [1] Amy @GrabNGoInfo. *Time Series Anomaly Detection Using Prophet in Python*. URL: <https://medium.com/grabngoinfo/time-series-anomaly-detection-using-prophet-in-python-877d2b7b14b4> (visited on 07/02/2023).
- [2] aca. *Estalvi i eficiència de l'aigua*. URL: <https://aca.gencat.cat/ca/laca/campanyes-i-divulgacio/campanyes/estalvi-i-eficiencia-de-laigua/index.html> (visited on 05/25/2023).
- [3] Raghav Agrawal. *Anomaly Detection Model on Time Series Data in Python using Facebook Prophet*. URL: <https://www.analyticsvidhya.com/blog/2021/12/anomaly-detection-model-using-facebook-prophet/> (visited on 05/29/2023).
- [4] Héctor Aguilera et al. “Towards flexible groundwater-level prediction for adaptive water management: using Facebook’s Prophet forecasting approach.” In: *Hydrological Sciences Journal* 64 (Aug. 2019). DOI: 10.1080/02626667.2019.1651933.
- [5] Y. Bengio, P. Simard, and P. Frasconi. “Learning long-term dependencies with gradient descent is difficult.” In: *IEEE Transactions on Neural Networks* 5.2 (1994), pp. 157–166. DOI: 10.1109/72.279181.
- [6] Arden Dertat. *Applied Deep Learning - Part 3: Autoencoders*. URL: <https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798#3f72/> (visited on 07/19/2023).
- [7] National Geographic. *Surface Water*. URL: <https://education.nationalgeographic.org/resource/surface-water> (visited on 05/25/2023).
- [8] F. Kratzert et al. “Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks.” In: *Hydrology and Earth System Sciences* 22.11 (2018), pp. 6005–6022. DOI: 10.5194/hess-22-6005-2018. URL: <https://hess.copernicus.org/articles/22/6005/2018/>.
- [9] F. Kratzert et al. “Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets.” In: *Hydrology and Earth System Sciences* 23.12 (2019), pp. 5089–5110. DOI: 10.5194/hess-23-5089-2019. URL: <https://hess.copernicus.org/articles/23/5089/2019/>.
- [10] Frederik Kratzert et al. “Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks.” In: *Hydrology and Earth System Sciences* 22 (Nov. 2018), pp. 6005–6022. DOI: 10.5194/hess-22-6005-2018.
- [11] Gustavo Felipe Martin Nascimento et al. “Outlier Detection in Buildings’; Power Consumption Data Using Forecast Error.” In: *Energies* 14.24 (2021). ISSN: 1996-1073. DOI: 10.3390/en14248325. URL: <https://www.mdpi.com/1996-1073/14/24/8325>.

- [12] Daniel Moriasi et al. “Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations.” In: *Transactions of the ASABE* 50 (May 2007). DOI: 10.13031/2013.23153.
- [13] Oinkina. *Understanding LSTM Networks*. URL: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/> (visited on 07/19/2023).
- [14] Prophet. *Diagnostics*. URL: <https://facebook.github.io/prophet/docs/diagnostics.html> (visited on 05/29/2023).
- [15] Water Science School. *Where is Earth’s Water?* URL: <https://www.usgs.gov/special-topics/water-science-school/science/where-earths-water> (visited on 05/25/2023).
- [16] A. Y. Sun et al. “A graph neural network (GNN) approach to basin-scale river network learning: the role of physics-based connectivity and data fusion.” In: *Hydrology and Earth System Sciences* 26.19 (2022), pp. 5163–5184. DOI: 10.5194/hess-26-5163-2022. URL: <https://hess.copernicus.org/articles/26/5163/2022/>.
- [17] Jieru Yan and Tao Tao. “Unsupervised anomaly detection in hourly water demand data using an asymmetric encoder–decoder model.” In: *Journal of Hydrology* 613 (2022), p. 128389. ISSN: 0022-1694. DOI: <https://doi.org/10.1016/j.jhydrol.2022.128389>. URL: <https://www.sciencedirect.com/science/article/pii/S002216942200957X>.
- [18] Feng Ye et al. “Hydrologic Time Series Anomaly Detection Based on Flink.” In: *Mathematical Problems in Engineering* 2020 (May 2020), p. 3187697. ISSN: 1024-123X. DOI: 10.1155/2020/3187697. URL: <https://doi.org/10.1155/2020/3187697>.
- [19] Yufeng Yu et al. “Detecting Pattern Anomalies in Hydrological Time Series with Weighted Probabilistic Suffix Trees.” In: *Water* 12.5 (2020). ISSN: 2073-4441. DOI: 10.3390/w12051464. URL: <https://www.mdpi.com/2073-4441/12/5/1464>.
- [20] Yufeng Yu et al. “Time Series Outlier Detection Based on Sliding Window Prediction.” In: *Mathematical Problems in Engineering* 2014 (Oct. 2014), p. 879736. ISSN: 1024-123X. DOI: 10.1155/2014/879736. URL: <https://doi.org/10.1155/2014/879736>.