
M2.951 · Tipologia i cicle de vida de les dades · Pràctica 1

Taula de continguts

1. Context	2
2. Títol	2
3. Descripció del dataset	2
4. Representació gràfica	3
5. Contingut	3
6. Propietari	4
7. Inspiració	5
8. Limitacions i futurs treballs	5
9. Llicència	6
10. Codi	6
11. Dataset	7
12. Vídeo	7
13. Bibliografia	8
14. Firmes	9

1. Context

En el context de sequera actual, l'objectiu d'aquesta primera part de la pràctica és la obtenció de dades meteorològiques d'un punt del territori. És per aquest motiu que el lloc web que primerament vam pensar fou el corresponent al Servei Meteorològic de Catalunya (Meteocat), empresa pública adscrita al Departament d'Acció Climàtica, Alimentació i Agenda Rural de la Generalitat de Catalunya (Servei Meteorològic de Catalunya, 2023), ja que entre les seves funcions s'inclou (Servei Meteorològic de Catalunya, 2022):

[...]

- d) Tractar, explotar i divulgar les dades procedents dels equipaments meteorològics.
- e) Explotar i gestionar la base documental provinent del Servei de Meteorologia del Departament de Medi Ambient.

[...]

L'adreça del lloc web és <https://www.meteo.cat/>

2. Títol

El títol del *dataset* és:

Dades meteorològiques de l'estació automàtica Guixers – Valls (18/04/2013-17/04/2023).

3. Descripció del dataset

El dataset inclou les dades meteorològiques (resums diaris) dels darrers 10 anys registrades a l'estació automàtica Guixers – Valls, ubicada al municipi de Guixers a la comarca del Solsonès. En les imatges inferiors es pot observar la ubicació de l'estació.



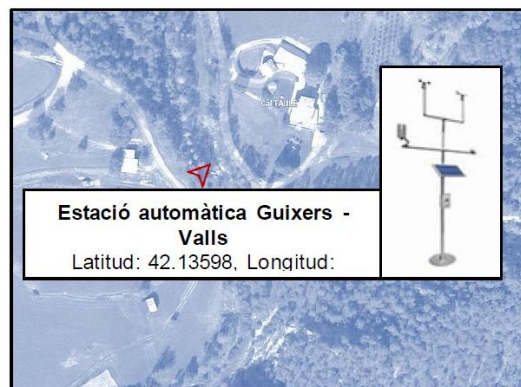
Vista de l'estació i la seva ubicació en el territori

Les dades que proporciona l'estació i que actualment es poden descarregar del lloc web del Meteocat són un resum diari consistent en:

- Temperatura mitjana.
- Temperatura màxima.

- Temperatura mínima.
- Humitat relativa mitjana
- Precipitació acumulada.

4. Representació gràfica



Dia (dd.mm.aaaa)

⌚ Hora	$t_{mitjana}$	$t_{màxima}$	t_{min}	$h_{relativa}$	PPT
00:00	$t_{mitj}^{00:00}$	$t_{max}^{00:00}$	$t_{min}^{00:00}$	$h_{rel}^{00:00}$	$ppt^{00:00}$
...					
hh:mm	$t_{mitj}^{hh:mm}$	$t_{max}^{hh:mm}$	$t_{min}^{hh:mm}$	$h_{rel}^{hh:mm}$	$ppt^{hh:mm}$
...					
23:30	$t_{mitj}^{23:30}$	$t_{max}^{23:30}$	$t_{min}^{23:30}$	$h_{rel}^{23:30}$	$ppt^{23:30}$

Dataset

Dades meteorològiques (de precipitació) de l'estació automàtica Guixers – Valls (18/04/201-17/04/2013)

Dia	Temp. mitjana	Temp. màxima	Temp. mínima	Mitjana humitats relatives	Precipitació acumulada
18.04.2013	$t_{mitj}^{18.04.2013}$	$t_{max}^{18.04.2013}$	$t_{min}^{18.04.2013}$	$h_{rel}^{18.04.2013}$	<i>precipitació^{18.04.2013}</i>
...					
dd.mm.aaaa	$t_{mitj}^{dd.mm.aaaa}$	$t_{max}^{dd.mm.aaaa}$	$t_{min}^{dd.mm.aaaa}$	$h_{rel}^{dd.mm.aaaa}$	<i>precipitació^{dd.mm.aaaa}</i>
...					
17.04.2023	$t_{mitj}^{17.04.2023}$	$t_{max}^{17.04.2023}$	$t_{min}^{17.04.2023}$	$h_{rel}^{17.04.2023}$	<i>precipitació^{17.04.2023}</i>

5. Contingut

El dataset s'inicia el 18 d'abril de 2013 (ara bé l'estació comença a estar operativa al 2015) i acaba el 17 d'abril de 2023 (10 anys) i inclou les següents variables:

- **Dia:** variable quantitativa tipus data, és a dir, dia, mes i any (format dd.mm.aaaa).
- **Temperatura mitjana:** variable quantitativa contínua. Indica la mitjana de les temperatures registrades durant el dia a l'estació (actualment 48 registres disponibles, un cada 30 minuts). Unitat: grau centígrad.
- **Temperatura màxima:** variable quantitativa contínua. Indica la temperatura màxima registrada a l'estació (actualment 48 registres disponibles, un cada 30 minuts). Unitat: grau centígrad.

- **Temperatura mínima:** variable quantitativa contínua. Indica la temperatura mínima registrada a l'estació (actualment 48 registres disponibles, un cada 30 minuts). Unitat: grau centígrad.
- **Humitat relativa mitjana:** variable quantitativa contínua. Indica la mitjana de les humitats relatives registrades durant el dia a l'estació (actualment 48 registres, un cada 30 minuts). Unitat: %.
- **Precipitació:** variable quantitativa contínua. Indica la precipitació acumulada durant el dia a l'estació (la suma dels, actualment, 48 registres, un cada 30 minuts). Unitat: mm.

6. Propietari

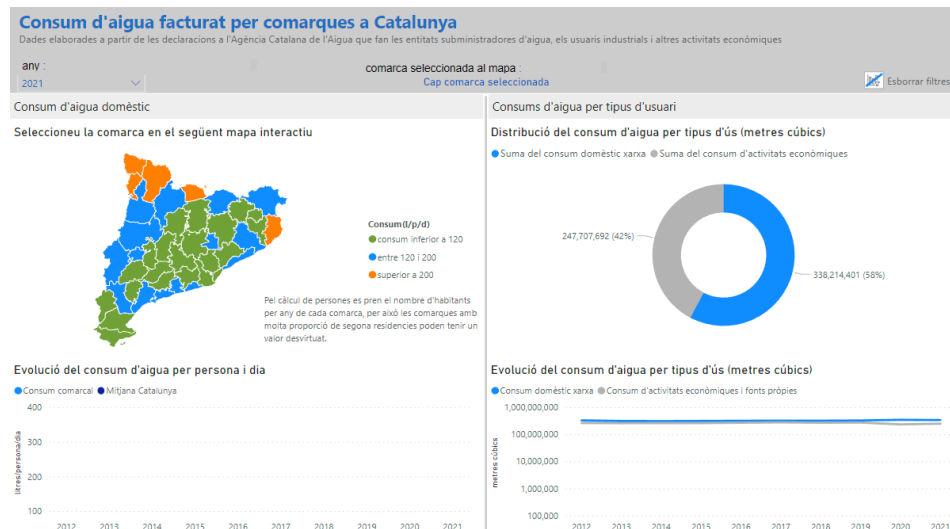
Segons la informació proporcionada per la pàgina web de Meteocat, l'Agència Catalana de l'Aigua (ACA) és la propietària de les dades meteorològiques que provenen de les estacions automàtiques de la xarxa de Meteocat. L'ACA és una agència pública adscrita al Departament d'Acció Climàtica, Alimentació i Agenda Rural de la Generalitat de Catalunya, i té com a missió la gestió integrada dels recursos hídrics i la prevenció de riscos hidrològics a Catalunya (Servei Meteorològic de Catalunya, 2022).

Per tant, les dades meteorològiques que es recullen a través de les estacions automàtiques de Meteocat són propietat de l'Agència Catalana de l'Aigua i són gestionades i difoses per Meteocat, que és el servei meteorològic de la Generalitat de Catalunya. Això implica que les dades són públiques i es poden consultar i utilitzar per a finalitats de servei públic, com ara la prevenció i gestió de riscos, la planificació urbana i territorial, la gestió dels recursos naturals i molts altres.

En general, en el web scraping si les dades obtingudes s'utilitzen per a ús personal, a la pràctica, n'hi ha cap problema (Lawson, 2015). Tanmateix, com és el cas en que ens trobem en aquesta pràctica, si les dades es tornaran a publicar (Github i Zenodo), llavors el tipus de dades obtingudes amb web scraping és important: en aquest cas no es tracta d'opinions ni dades que puguin tenir drets d'autor, ja que són dades sobre fenòmens naturals obtingudes de manera automàtica per una estació meteorològica (si bé passades per un control de qualitat).

En referència a altres estudi fets i que en part complementarien el que tenim, tenim en el portal **gencat.cat** l'anomenat:

- “Visor de la Sequera”: [Visor de la Sequera \(gencat.cat\)](https://visor.sequera.gencat.cat/). El qual enllaça al lloc web de l'Agència Catalana de l'Aigua:
- “Dades obertes”: [Dades obertes. Agència Catalana de l'Aigua \(gencat.cat\)](https://dadesobertes.gencat.cat/), des del qual es pot accedir a la
 - Visualització interactiva de dades (Agència Catalana de l'Aigua, 2023): fa referència al consum d'aigua i a l'estat dels embassaments, a partir d'un *dashboard* generat en Power BI sobre dades dels darrers 10 anys.



Visualització interactiva de dades (Agència Catalana de l'Aigua, 2023)

- Dades obertes en temps real ([Dades obertes en temps real. Agència Catalana de l'Aigua \(gencat.cat\)](https://gencat.cat)): dades de les darreres mesures als embassaments.

7. Inspiració

En el context de sequera actual ens hem volgut preguntar si el ressò mediàtic que actualment hi ha al respecte es reflexa amb la realitat objectiva que poden proporcionar les dades d'una estació meteorològica, és a dir si en períodes anteriors de menys ressò la situació era igual, pitjor o millor que l'actual.

La variable principal és la precipitació però també volem analitzar altres variables proporcionades per l'estació com són la temperatura i la humitat relatives per poder, entre d'altres poder estudiar el grau de dependència que puguin tenir amb la precipitació i si tenen comportaments similars o no.

Som conscients que el nivell en que afecta a la ciutadania i al territori la sequera no depèn sols de la precipitació, sinó també de les necessitats hídriques que s'han de satisfer i de la quantitats d'aigua utilitzable acumulada en l'inici de manca de precipitacions, però precisament aquesta és una pregunta que ens plantejem: si la sequera és sols una conseqüència de les falta de precipitacions o bé també una conseqüència d'altres factors.

8. Limitacions i futurs treballs

L'àmbit de la present pràctica és el basat en les dades meteorològiques obtingudes en els darrers 10 anys d'una sola estació meteorològica ubicada al Solsonès, fet que limita força la capacitat de donar respostes globals al fenomen de la sequera a Catalunya, així com al fet que no s'han tingut en compte altres variables com poden ser les reserves naturals i artificials d'aigua existents.

Per a futurs treballs proposem l'obtenció de dades de més estacions meteorològiques per tal de poder arribar a uns àmbits majors com podrien ser els comarcals, provincials, tot Catalunya i/o de conques hídriques.

9. Llicència

Essent les dades obtingudes de domini públic i essent la nostra voluntat promoure la difusió i el compartiment del seu contingut sense restriccions, hem optat per a una llicència seleccionada ha sigut la CC0 1.0 Universal: la llicència CC0 1.0 Universal és una llicència de domini públic que permet als creadors de contingut renunciar als seus drets d'autor i altres drets de propietat intel·lectual. En altres paraules, aquesta llicència permet a qualsevol persona fer servir, copiar, modificar i distribuir el contingut sense restriccions o necessitat de permís o pagament al creador. (Creative Commons, 2023). Ara bé, els autors no es fan responsables de possibles danys que puguin succeir.

En la imatge inferior es pot veure un resum de les característiques de la llicència usada:

 <p>importasa/pluja_meteocat_public is licensed under the Creative Commons Zero v1.0 Universal</p> <p>The Creative Commons CC0 Public Domain Dedication waives copyright interest in a work you've created and dedicates it to the world-wide public domain. Use CC0 to opt out of copyright entirely and ensure your work has the widest reach. As with the Unlicense and typical software licenses, CC0 disclaims warranties. CC0 is very similar to the Unlicense.</p>	<p>Permissions</p> <ul style="list-style-type: none"> ✓ Commercial use ✓ Modification ✓ Distribution ✓ Private use 	<p>Limitations</p> <ul style="list-style-type: none"> ✗ Liability ✗ Trademark use ✗ Patent use ✗ Warranty 	<p>Conditions</p>
---	---	--	--------------------------

Llicència CC0 1.0 Universal (Creative Commons, 2023)

10. Codi

El codi s'ha implementat en Python, fent ús de Scrapy, un framework de web scraping de codi obert escrit en Python, que té com a propòsit facilitar la creació d'spiders (aranyes) que rastregen i extreuen dades de pàgines web de manera automatitzada (Scrapy, 2023).

Com a punts més rellevants i dificultats que hem tingut cal destacar:

- S'ha hagut d'indicar que no es seguiran les recomanacions indicades a l'arxiu robots.txt¹ (`ROBOTSTXT_OBEY = False` de `Settings.py`) ja que les pàgines que necessitem rastrejar precisament estan afectades pel mateix. La directriu "Disallow" bloqueja l'accés als URLs indicats:

```
# Block duplicate content
Disallow: /*?
Disallow: /observacions/xema?
Disallow: /observacions/xema/dades?*dia=
Disallow: /prediccio/municipal/*?
```

Part de l'arxiu robots.txt que afecta els URLs objectiu de la nostra pràctica

- Al fer varies vegades `get` de la web desitjada, aquesta bloquejava l'accés (error 400 i 401). Per tal de solucionar-ho, s'ha modificat el valor de `download_delay=3` en `settings`.
- Per un altre costat, fer un `for loop` en la shell de scrapy es bastant complicat i una forma de solucionar-ho ha estat cridant als valors un per un, per exemple: `table=response.xpath('//table')[0]`. El que es va fer posteriorment un `cop` conegut la capçalera de la taula/fila objectiu, ha estat cridant el valors fent referència al text d'aquestes capçaleres.

¹ <https://meteo.cat/robots.txt>

Els autors una vegada conegut els problemes anteriors han pogut implementar el codi de github a través de pycharm, utilitzant la seva terminal. Al afegir els 3 segons d'espera ha provocat que es necessitin unes 4hores per tal de baixar totes les dades

El codi es pot trobar en el següent repositori:

https://github.com/mportasa/pluja_meteocat_public

11. Dataset

Les dades es poden trobar tant a Zenodo com en el repositori de Github:

- Zenodo: Marc Porta Sardà, & Roger Raduà Castaño. (2023). Dades meteorològiques de l'estació automàtica Guixers – Valls (18/04/201-17/04/2013) (Version 1) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.7857084>.
- GitHub: https://github.com/mportasa/pluja_meteocat_public/tree/main/dataset

12. Vídeo

EL vídeo es pot visualitzar i descarregar a:

https://drive.google.com/file/d/16mG9ntADQgWJ3L2gS2i7-vHbeeHUYiqE/view?usp=share_link

13. Bibliografia

Agència Catalana de l'Aigua. (23 / Abril / 2023). *gencat - Agència Catalana de l'Aigua*. Recollit de Agència Catalana de l'Aigua - Consum d'aigua per comarques a Catalunya: <https://aca.gencat.cat/ca/laigua/consulta-de-dades/dades-obertes/visualitzacio-interactiva-dades/Consum-aigua-comarques-catalunya/>

Creative Commons. (22 de Abril de 2023). *Creative Commons: CC0 1.0 Universal (CC0 1.0). Oferiment al Domini Públic*. Obtenido de Creative Commons: <https://creativecommons.org/publicdomain/zero/1.0/deed.ca>.

Lawson, R. (2015). *Web Scraping with Python: Successfully Scrape Data from Any Website with the Power of Python*,. BIRMINGHAM - MUMBAI: Packt Publishing.



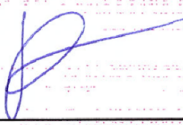

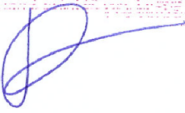
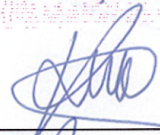
Servei Meteorològic de Catalunya. (21 de 04 de 2022). *Presentació i Funcions: meteo.cat*. Recuperado el 04 de 2023, de meteo.cat: <https://www.meteo.cat/wpweb/sobre-meteocat/funcions/>

Servei Meteorològic de Catalunya. (22 / Març / 2023). *Història: meteo.cat*. Consultat el 04 / 2023, a meteo.cat: <https://www.meteo.cat/wpweb/sobre-meteocat/historia/>

Scrapy tutorial. (22 de Abril de 2023): <https://docs.scrapy.org/en/latest/intro/tutorial.html>

Python Web Scraping & Crawling using Scrapy. (22 de Abril de 2023): https://www.youtube.com/playlist?list=PLhTjy8cBISEqkN-5Ku_kXG4QW33sxQo0t

14. Firmes

Contribucions	Signatura
Investigació prèvia	 
Redacció de les respostes	 
Desenvolupament del codi	 
Participació al vídeo	