Rainfall prediction in the Mondego river basin using CV and DL

Faculdade de Engenharia da Universidade do Porto

Master in Data Science and Engineering

Computer Vision

João Pedro Pêgo, Kareem Kousa, Pedro Gouveia

Objectives | Methodology

Objective

Using weather forecasts/reanalysis datasets, **estimate the rainfall depth in a river basin**.

Methodology

Target: **daily rainfall** for the **Mondego river basin** (Thiessen method).

Dataset available here: https://filesender.fccn.pt/?s=download&token=0fe569f1-69ea-4b7a-bfb1-f68cb004c953

Workplan

1. Compute daily rainfall in the Mondego river basin (target)	ipynb
1.1. Create Thiessen construction for Portuguese basins	[1]
1.2. Identify which Thiessen polygons intersect Mondego river basin	[1]
1.3. Download data from SNIRH for period 1950 – 2021	[1]
1.4. Load data to dataframes.	[2]
1.5. Verify missing values	[2]
1.6. Eliminate weather stations with too many missing values	[2]
1.7. Reconstruct Thiessen polygons with validated weather stations	[2]
1.8. Compute the spatial average of the rainfall in the Mondego RB	[2]

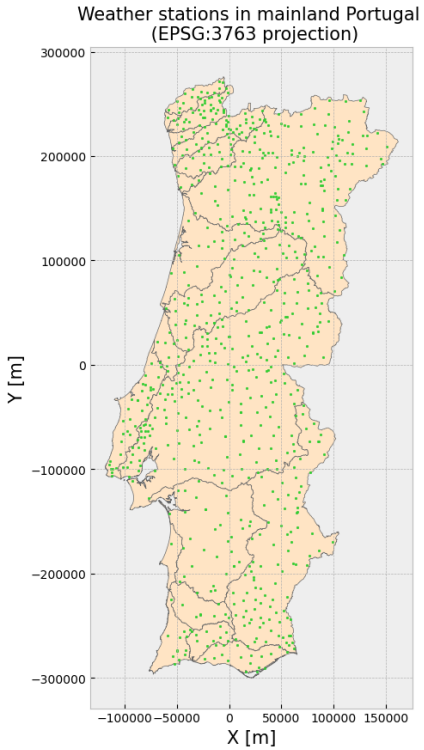
2. Create ECMWF meteorological dataset (tp, tcc, t2m)	ipynb
2.1. Define coordinates (lat/long) for download	[3]
2.2. Download data from ECMWF for period defined in 1.7	[4]
2.3. Convert lat/long to Carthesian coordinates (EPSG:3763)	[3]
2.4. Generate images (rainfall, temperature, total cloud coverage)	[5]

3. Develop CNN for spatial averaged rainfall prediction (VC | TAAC Q2)	ipynb
3.1. Dataset preprocessing	[6]
3.2. Data loader	[7]
3.3. CNN model	[7]
3.4. Training routine	[7]
3.5. Evaluation routine	[7]
3.6. Hyper parameter tunning	[7]
3.7. Testing different image sizes	[7]
3.8. Results evaluation	[7]

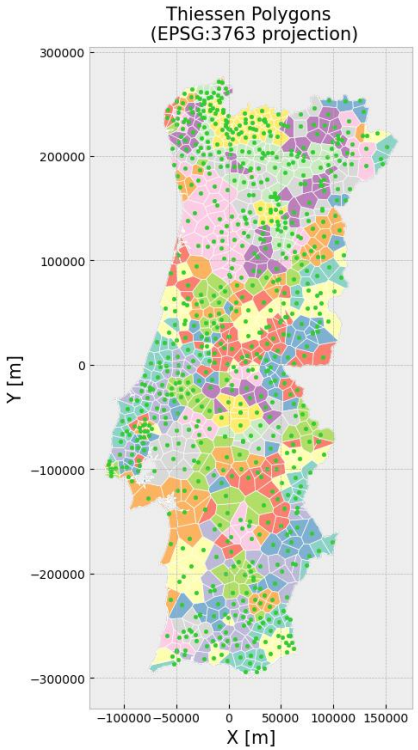
Notebooks	Machine
[1] 20221025_VC_SNIRH_Download	Local
[2] 20221101_VC_SNIRH_Preprocessing	Local
[3] 20221025_VC_ECMWF_Grid_Basin	Local
[4] 20221030_VC-ECMWF_Download	Local
[5] 20221101_VC-ECMWF_ExportImages	Local
[6] 20221110_VC_CNN_dataset	Colab
[7] 20221114_VC_CNN	Colab

Dataset available here: https://filesender.fccn.pt/?s=download&token=a0d86e5b-f854-4133-93a5-1cc924e5f4a1

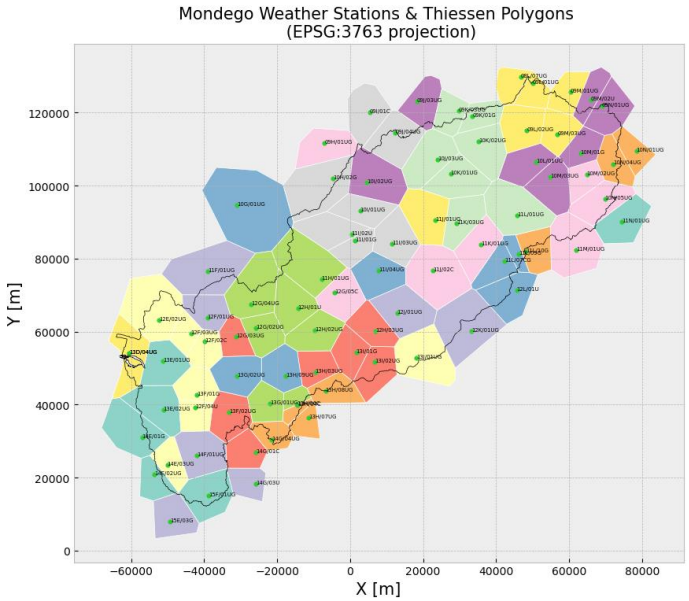
1. Daily rainfall in the Mondego river basin



Voronoi tesselation



Mondego river basin

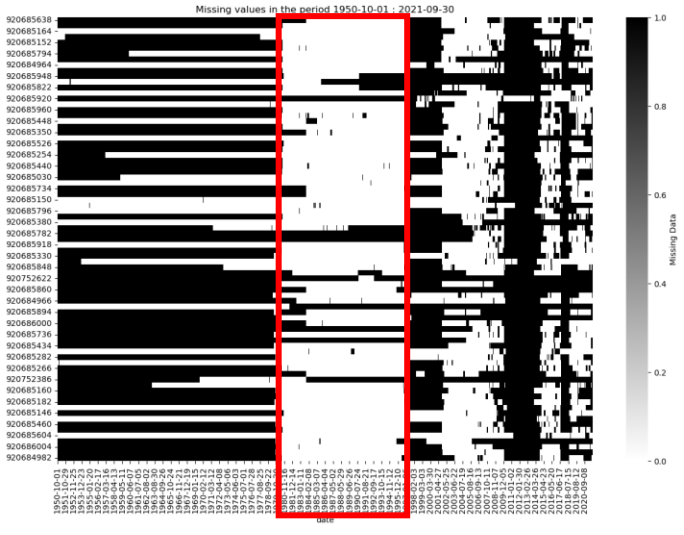



```
# Data sources
# basins boundaries
https://sniambgeoviewer.apambiente.pt/GeoDocs/shpzips/AtAgua_Bacias_bacias_snirh_PC.zip
# weather stations network
https://sniambgeoviewer.apambiente.pt/GeoDocs/shpzips/AtAgua_Meteo_METEO_NET_PC.zip
```

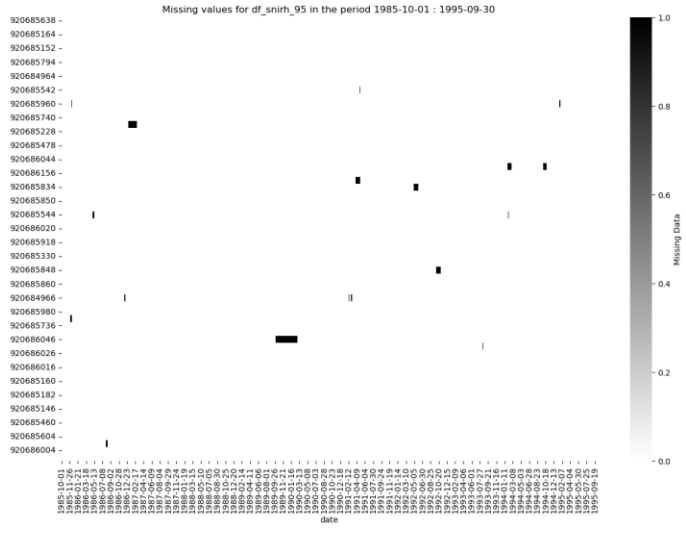
Notebooks:
[1] 20221025_VC_SNIRH_Download

Results:
83 weather stations whose Thiessen polygon intercepts the Mondego river basin

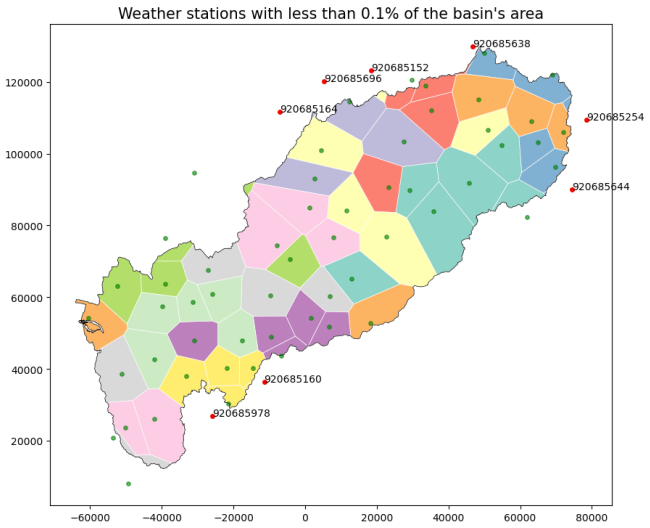
1. Daily rainfall in the Mondego river basin



MV 1950-2021



MV 1985/10-1995/09



Stations with 95% complete records

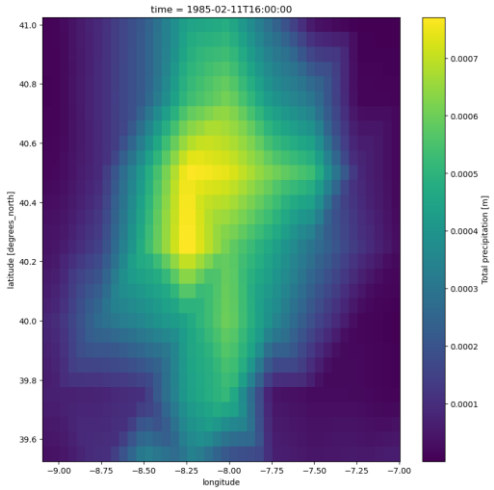
The **spatial average** of days with MV was computed **eliminating the stations with MV** and **recomputing the Thiessen polygons.**

Results:
Period with most complete records: 1985 - 1995
64 weather stations with records >95% of the time
48 weather stations with complete dataset.
Rainfall spatial average over the river basin (target)

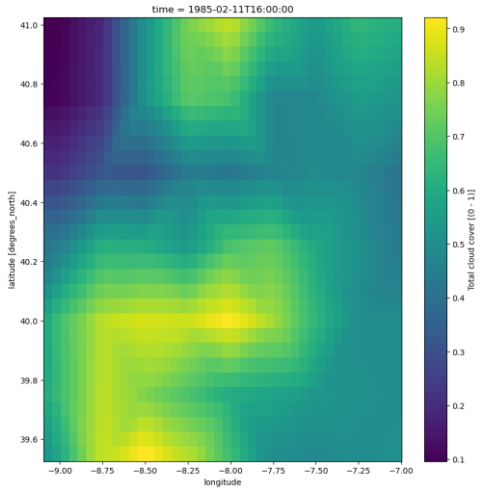
Data sources
Daily rainfall
www.snirh.pt

Notebooks:
[2] 20221101_VC_SNIRH_Preprocessing

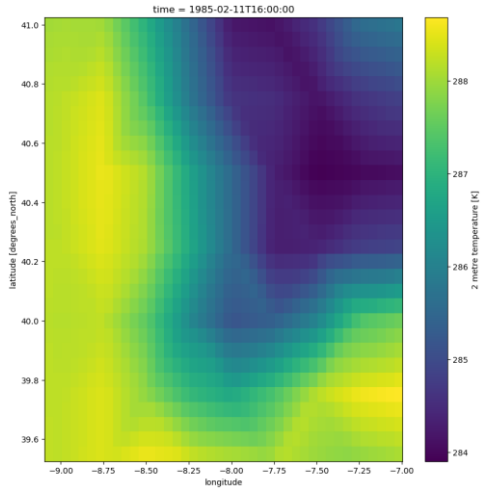
2. ECMWF meteorological dataset (tp, tcc, t2m)



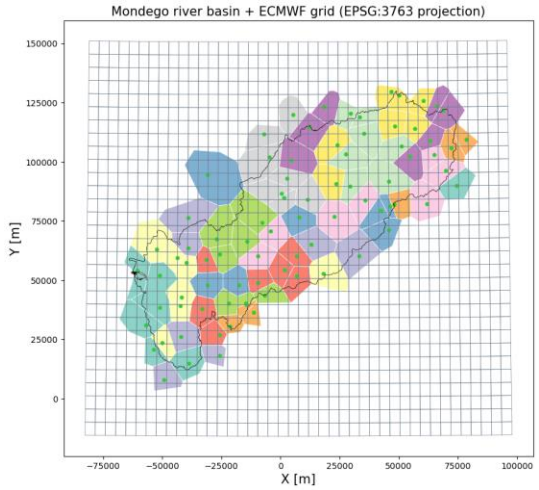
Total precipitation (tp)



Total cloud coverage (tcc)



2m temperature (t2m)



Raster grid (30 x 42)

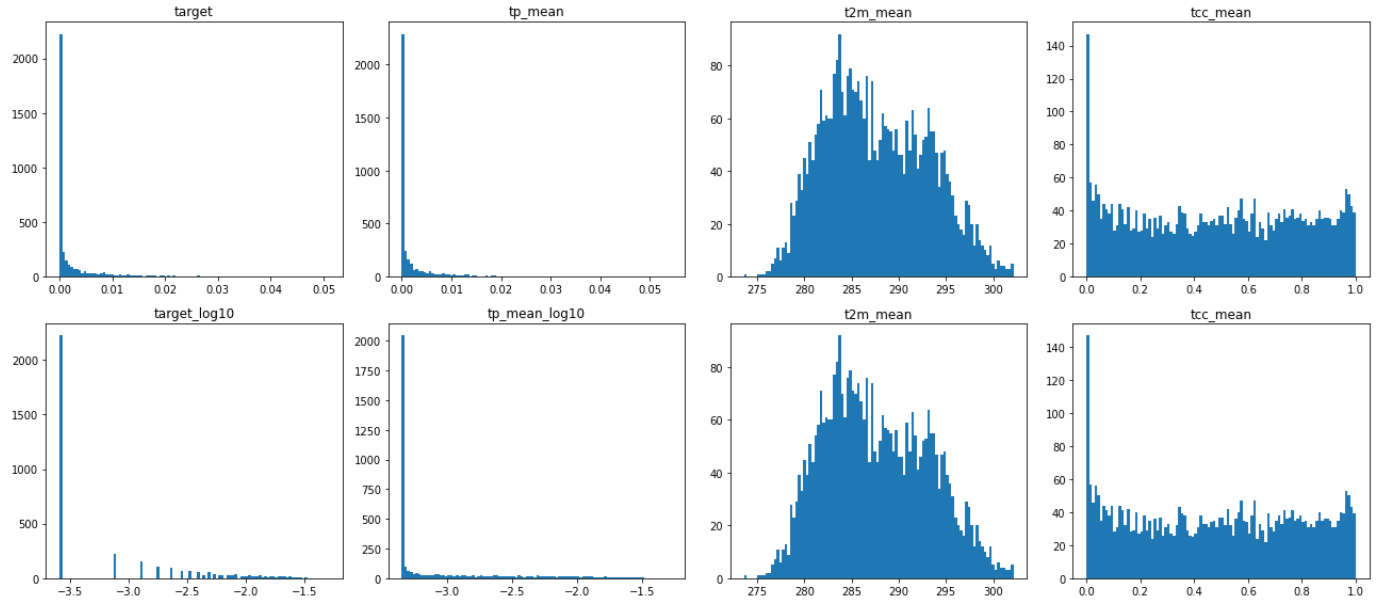
```
# Data sources
# hourly values of total precipitation,
total cloud coverage, 2m temperature
https://cds.climate.copernicus.eu/cdsapp#!
/dataset/reanalysis-era5-land?tab=overview
```

Notebooks:
[3] 20221025_VC_ECMWF_Grid_Basin
[4] 20221030_VC-ECMWF_Download
[5] 20221101_VC-ECMWF_ExportImages
[6] 20221110_VC_CNN_dataset

Results:
ECMWF data on tp, tcc and t2m (features)
1 raster per hour (0h-1h, 1h-2h, …)
Range of lat/long covering the basin

3. CNN dataset

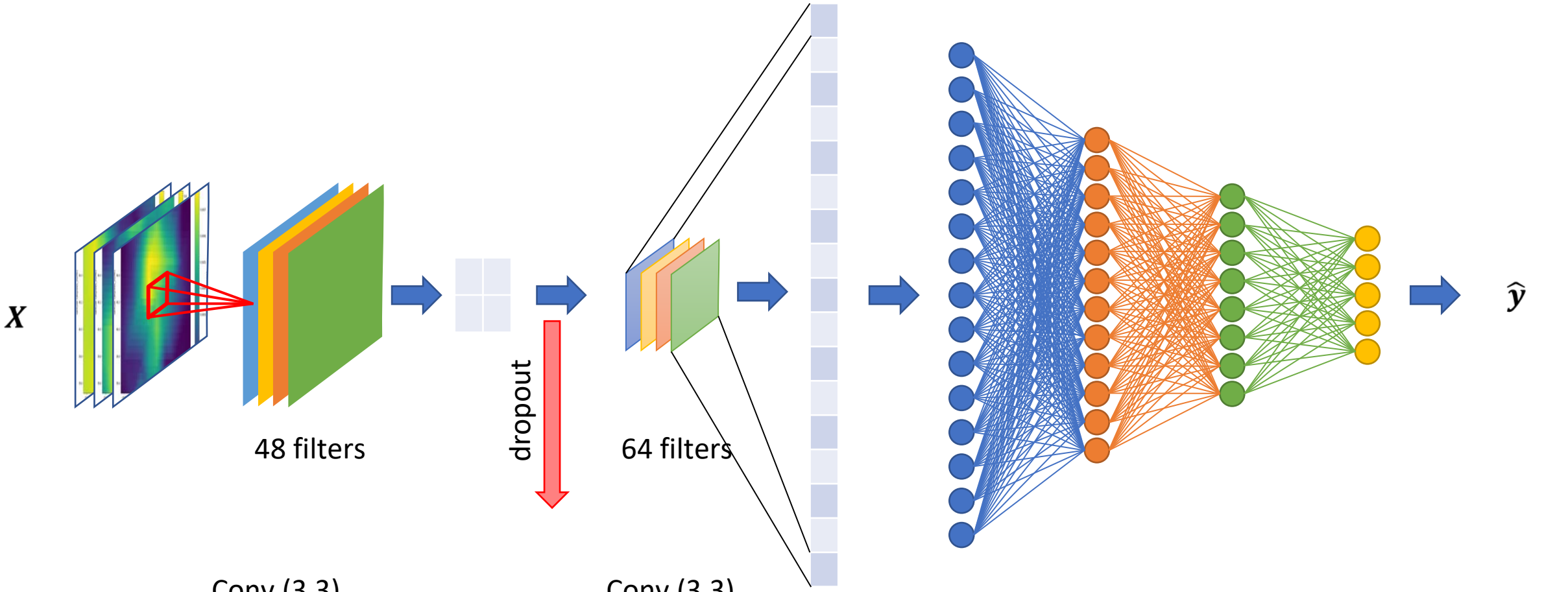
- Composite images
  (tp, tcc, t2m)
- 3652 instances



- Highly imbalanced
- Penalize frequent values
- Log transformation did not produce significant improvements
- MinMax normalization

Notebooks:
[7] 20221114_VC_CNN

3. CNN architecture



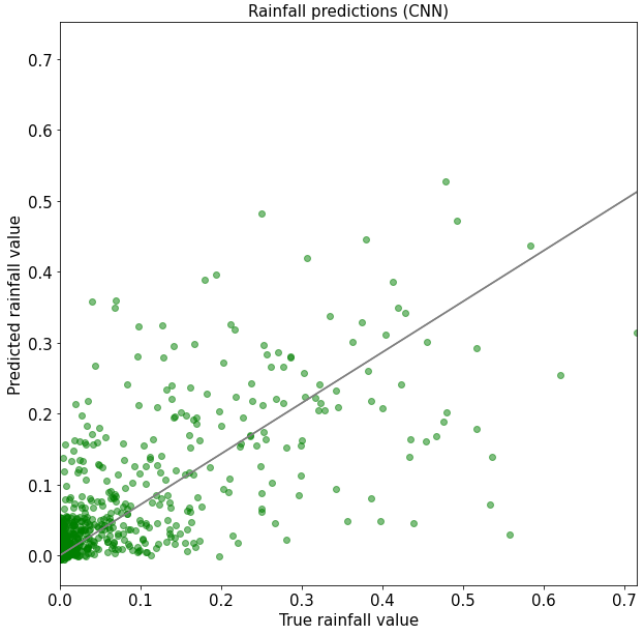
Notebooks:
[7] 20221114_VC_CNN

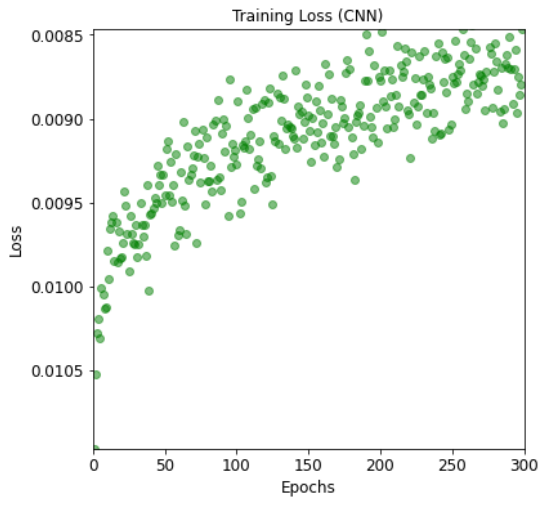
3. CNN results

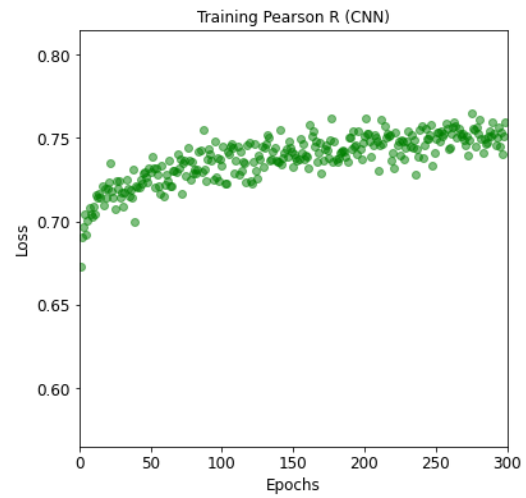
- Best model:
  - Optimizer : Rprop
    - Adam, SGD, NAdam, RMSProp
  - Parameters:
    - batch size - 64
    - epochs – 300
    - lr – 0.001
  - Loss (weighted MSE): 0.007
  - Small values weight: 100
  - Pearson R: 74.7%

Note: Different runs with same parameters produce similar but different results

Notebooks:
[7] 20221114_VC_CNN







Highlights

- Original research
- Dataset created from scratch
- Custom dataloader and loss function (weighted MSE)
- Dataset imbalance compensation (log transform and penalties)
- Successfully trained and tested a CNN for regression problem
- Hyperparameter tunning and different optimizers tested
- Too little sleep!

Future work

- Different CNN architectures
- Effect of image overlap with river basin limits
- Images with time information (e.g., composite tp images [d-3:d+3], hourly images)
- Data augmentation for least represented values of the target
- Other type of features (e.g., cumulative values, humidity, atmospheric pressure, etc.)
- Other datasets (e.g. ERA5-Land hourly data, PERSIANN)
- Write an article!

Rainfall prediction in the Mondego river basin using CV and DL

Faculdade de Engenharia da Universidade do Porto

Master in Data Science and Engineering

Computer Vision

João Pedro Pêgo, Kareem Kousa, Pedro Gouveia