AI data enrichment for drought risk assessment

Paris, 17/11/2022

Guillaume ATTARD[1], David BEHAR[2], Aurélien COULOUMY[3] , Luc GIBAUD[4]

And with the kind contribution of Antoine LABONNE[3] , Ammar MALKI[4], and Thomas ONFROY[3]

1. Ageoce (g.attard@ageoce.com)
2. Dataiku (david.behar@dataiku.com)
3. CCR Group (acouloumy@ccr.fr)
4. Quantmetry (lgibaud@quantmetry.com)

AGEOCE

data iku

CCR

Quantmetry

1. INTRODUCTION

1.1 Context

- **Understanding drought risk** can be hard considering the complexity and the evolutivity of such a risk.

- **AI capabilities** may be helpful to address such issues.

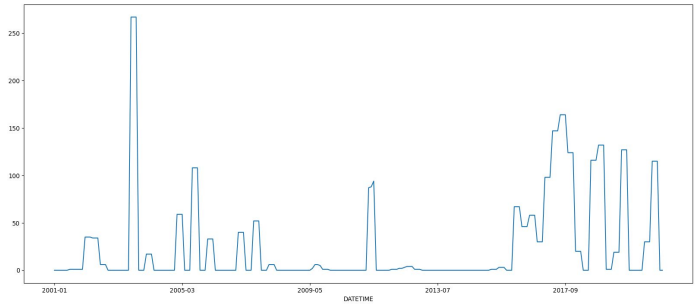


Figure 1: Example of drought nat cat occurrence on 3 french departments through time

- Replacing historical nat cat models by deep learning ones may be a solution but which does not appear yet relevant for business teams.

- Another pragmatic angle is to better understand effects by **getting more features:**

Sources

Time

Types

1.2 Agenda

- Discussion of **3 feature engineering** strategies that have been explored at **CCR Group** this year:

Open source hybrid geo data, and exposition of an API callable at the address using Dataiku to **make the most of public data**

SWI (Soil Wetness Index) time series forecasting to **anticipate future risks analysis** calculations

Computer vision and object detection applied to tree detection to **explore the interest of new data types**

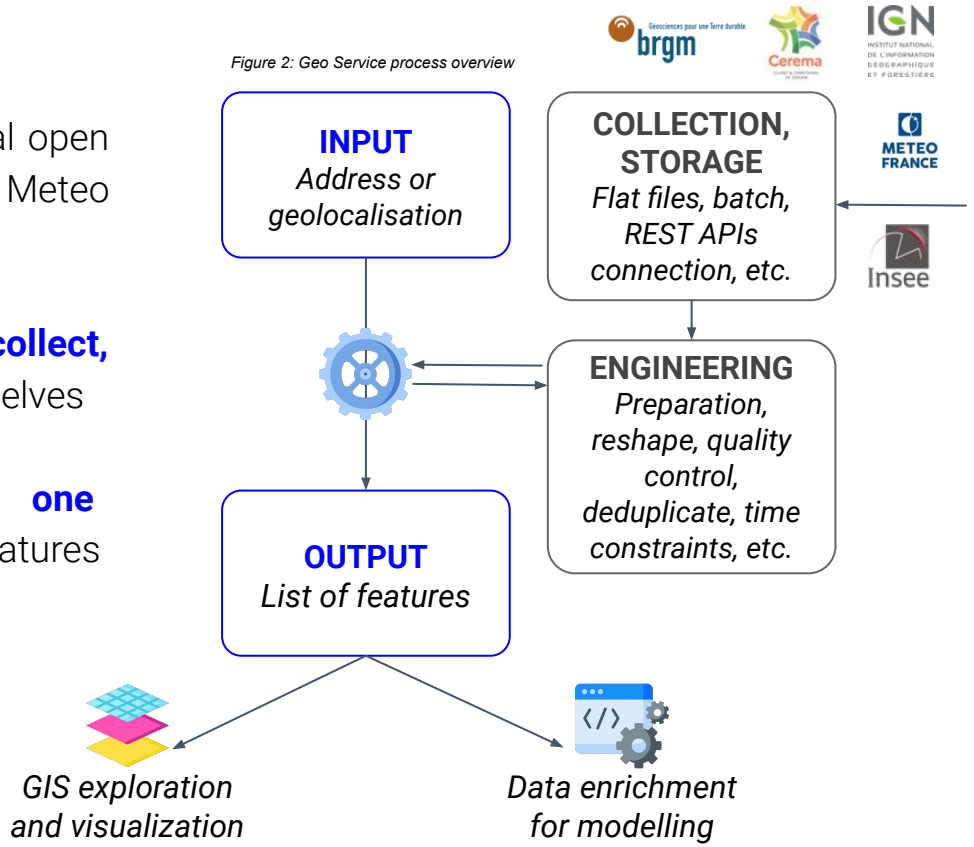
Deep learning proxy models to support nat cat modelling

2. GEODATA WEB SERVICES

2.1 Context and process overview

- There is a **wide panel** of geo or economical open source data in France: BRGM, IGN, Insee, Meteo France, Georisques, etc.
- Most of the open source data are **hard to collect, to maintain and to combine** between themselves
- Goals: provide endpoints to get in **one synchronous call** many standardized features about a precise address (at a precise time).

Figure 2: Geo Service process overview



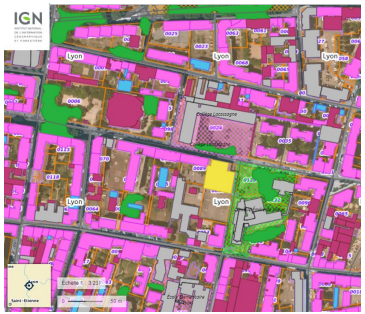
2.2 Data sources

- 4 examples of sources and data explored:

BDTOPO-Bâtiments

https://geoservices.ign.fr/bdtopo

Struct. tabular data in Shape, that describes constructions In France

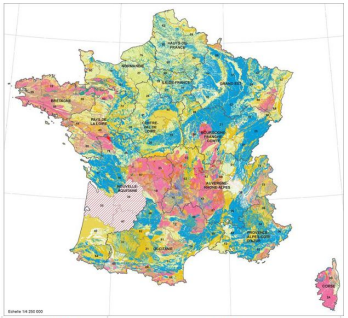


Date construction bâtiment	1990-01-01
Nombre logement bâtiment	28
Nombre étages bâtiment	4
Matériaux murs bâtiment	BETON
Matériaux toiture bâtiment	TUILES
Hauteur bâtiment	10.5
Altitude pied bâtiment	242

INFOTERRE BDCharm-50

https://infoterre.brgm.fr/

Shape tabular description of geological rocks formation

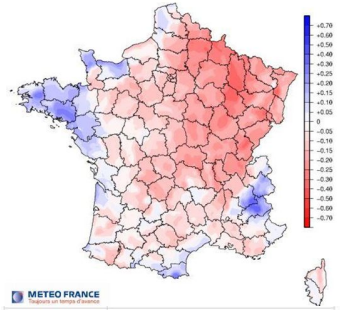


Formations géologiques	Alluvions fluvio-glaciaires indifférenciées du Riss ancien

METEO FRANCE SWI

https://donneespubliques.meteofrance.fr/

Csv monthly data of SWI splitted. in ~10k areas with ext. LambertII coord.

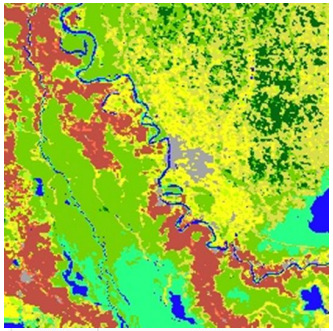


SWI_UNIF_MENS3	1.235

GEE MODIS/006/MCD12Q1

https://developers.google.com/earth-engine/

API of global land cover types at yearly intervals from 6 classif. schemes



Type de surface de l'écosystème	Urban and Built-up Lands: at least 30% impervious surface area including building materials, asphalt and vehicles

Figure 3: Example of data sources

2.3 Solutions

- **Asynch. process** in Dataiku to download, to store and pre process info. Then, synch. call and **response of the core API** for dashboard or other integration.

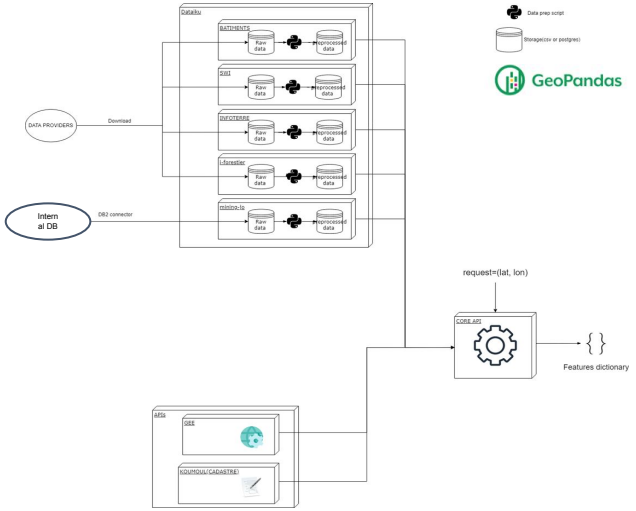


Figure 4: Geo service asynch process on Dataiku

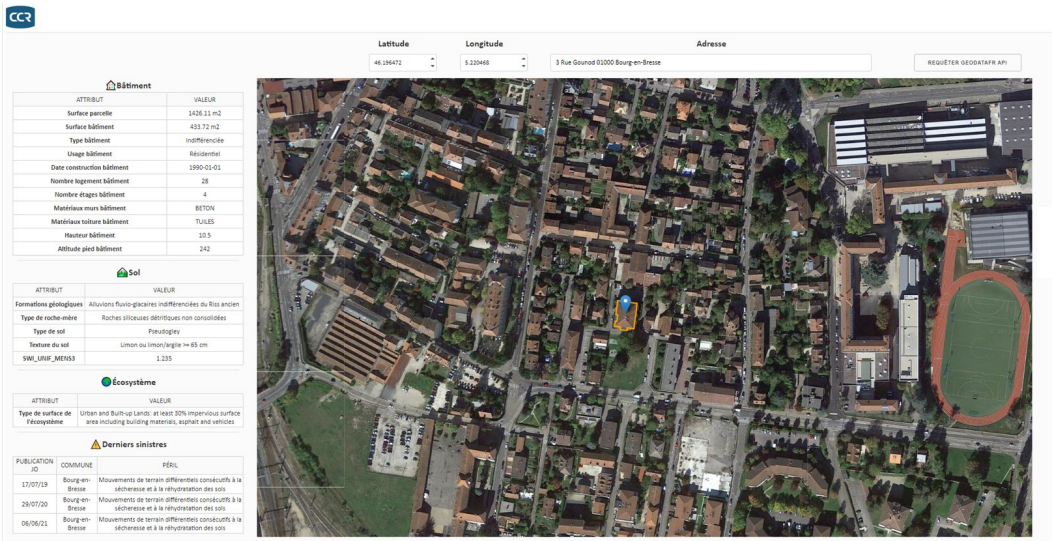


Figure 5: Geo service integration on Dataiku Webapp (CCR credit)

3. SWI TIME SERIES FORECASTING

3.1 Context about SWI

- The SWI is one of the indicators **used** **ns** **occurrence** and severity of drought risk.
- It represents, over a 2m depth, the state of the water reserve in the soil in relation to the useful reserve.
- Meteo France provides such info but sometimes a **time lag** between availability and computations requirements forces to predict it
- Goals: **experiment and develop models** for predicting the SWI incidence;

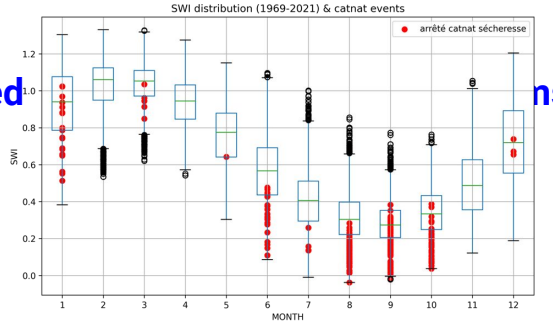


Figure 6: SWI value per area through time

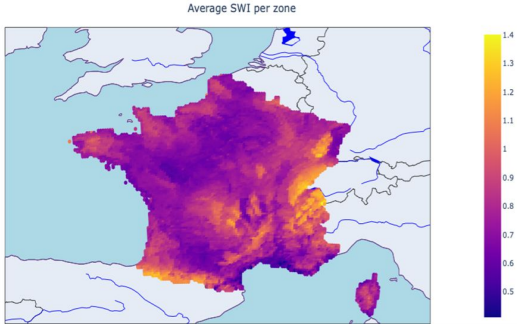


Figure 7: SWI value in France

3.2 Forecast approaches

- **Univariate vs Multivariate**: Time series containing records of a single variable. [1] or containing records of multiple variables. [2]

- For simplicity reason, we have selected only one area, the SWI area 2 (but the modeling can be reproduced easily on all the areas)

- We **decompose data** in :
  - Train set : values from 01/1969-12/2014 (552 values)
  - Test set : values from 01/2015-12/2020 (60 values)

- We have trained many different models to **compare performances** (accuracy, interpretability, complexity).

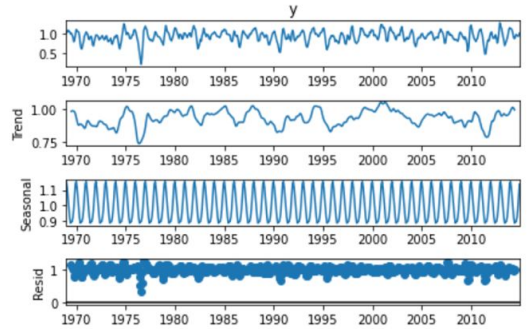


Figure 8: ***Decomposition of the time series***

Statistical models	Machine Learning	Deep Learning
SARIMA	RandomForest + MAPIE *Cyclic encoding*	NeuralProphet
Holt-Winters	RandomForest *Recursive prediction*	TSAI
Prophet	XGBoost *Cyclic encoding*	LSTM *Recursive prediction*
	XGBoost *Recursive prediction*	Bayesian LSTM *Recursive prediction*
		LSTM *Cyclic encoding + Recursive prediction*

PROPHET
scikit learn
MAPIE
PyTorch
Darts

3.3 Model results & discussions

- Using MSE, most models **give similar performance**
- Statistical and deep learning models give the best results
- We also **explore Kmeans Dynamic Time Warping** to provide clustering features of SWI Areas.

	Model	MSE	Confidence Interval Compliance	Confidence Interval Width
Statistical Models	SARIMA	**0.009**	**100%**	0.480
	Holt-Winters	0.014	-	-
	Prophet	**0.009**	**100%**	0.419
Machine Learning	RandomForest + MAPIE *Cyclic encoding*	0.011	98.6%	0.492
	RandomForest *Recursive prediction*	0.012	-	-
	XGBoost *Cyclic encoding*	0.013	-	-
	XGBoost *Recursive prediction*	0.011	-	-
Deep Learning	NeuralProphet	**0.009**	-	-
	TSAI	0.011	-	-
	LSTM *Recursive prediction*	0.014	-	-
	Bayesian LSTM *Recursive prediction*	0.013	91.7%	**0.404**
	LSTM *Cyclic encoding* *Recursive prediction*	**0.009**	-	-

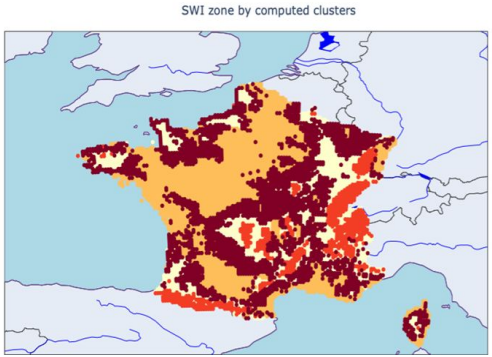


Figure 9: SWI cluster in France

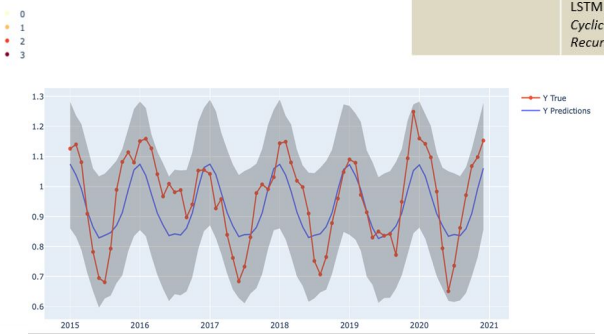


Figure 10: **Predictions and confidence interval (95%) with Prophet**

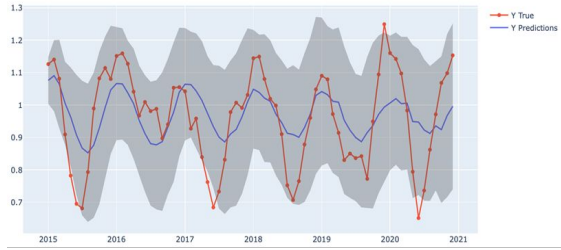


Figure 11: **Predictions and confidence interval with Bayesian LSTM**

4. TREE DETECTION

4.1 Context and data sources

- **Trees around buildings** can increase drought frequency/severity [1, 2].
- **Using datasets** that describe vegetation around buildings (BD ORTHO and BD TOPO) may be helpful
- Goals:
  - Collect and prepare aerial images
  - **Identify trees with Computer Vision**
  - Deduce business tabular features
  - Evaluate importance and scalability

Ranking	Species	Max tree height – H (m)	Max distance for 75% of cases (m)	Min recommended separation in very highly and highly shrinkable clays
1	Oak	16–23	13	1H
2	Poplar	24	15	1H
3	Lime	16–24	8	0.5H
4	Common ash	23	10	0.5H
5	Plane	25–30	7.5	0.5H
6	Willow	15	11	1H
7	Elm	20–25	12	0.5H
8	Hawthorn	10	7	0.5H
9	Maple/ sycamore	17–24	9	0.5H
10	Cherry/plum	8	6	1H
11	Beech	20	9	0.5H
12	Birch	12–14	7	0.5H
13	White beam/ rowan	8–12	7	1H
14	Cypress	18–25	3.5	0.5H

Figure 12: Safe distances between trees buildings (Civilblog.org)

Figure 13: BD ORTHO Sample (IGN)

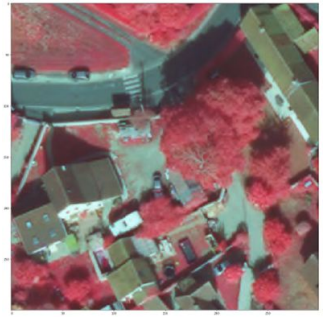


Figure 14: BD ORTHO IRC Sample (IGN)

Figure 15: BD TOPO Sample (IGN)

[1] Satriani, A., Loperte, A., Proto, M., & Bavusi, M. (2010). Building damage caused by tree roots: laboratory experiments of GPR and ERT surveys. *Advances in Geosciences*, *24*, 133-137.
[2]Li, J., & Guo, L. (2017). Field investigation and numerical analysis of residential building damaged by expansive soil movement caused by tree root drying. *Journal of Performance of Constructed Facilities*, *31*(1), D4016003.

4.2 Workflow and annotation

- **DeepForest** is a library for predicting individual tree crowns from RGB imagery [3, 4] that requires to be re-trained to be used on french aerial images.

- Re-train means annotate. To shortcut we **use LIDAR HD for pre-annotation** (which is available for 2 districts: Louhans and Manosque) and the process:

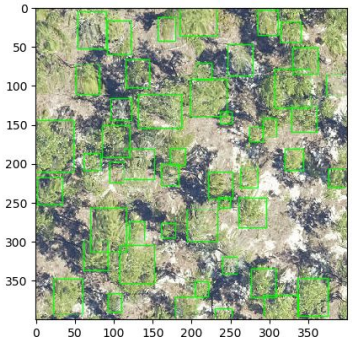


Figure 16: Example of DeepForest Application an a RGB imagery

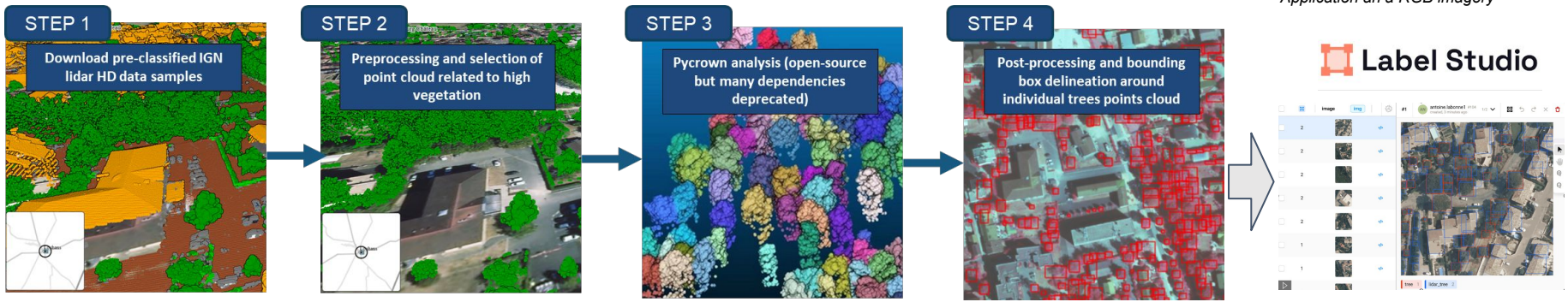


Figure 17: LIDAR pre-annotation process

[3] Weinstein, B.G.; Marconi, S.; Bohlman, S.; Zare, A.; White, E. Individual Tree-Crown Detection in RGB Imagery Using Semi-Supervised Deep Learning Neural Networks. Remote Sens. 2019, 11, 1309
[4] Geographic Generalization in Airborne RGB Deep Learning Tree Detection Ben Weinstein, Sergio Marconi, Stephanie Bohlman, Alina Zare, Ethan P White bioRxiv 790071; doi: https://doi.org/10.1101/790071

4.3 Trainset process and model results

- We have **adapted pre-annotated tiles** to get optimal performances. Final image crop: 100x100px, 500m each

- Data used to create **custom DeepForest** [5]:
  - 100 annotated images from Louhans
  - 50 annotated images from Manosque

- After grid search, **+35% F1-score** on validation set

- We have **tested model on Montigny Le Bretonneux**:
  - F1-score around 70% with 78% recall and 73% precision
  - Optimization of results with softmax threshold at 0.2

Figure 18: Trainset sizing process

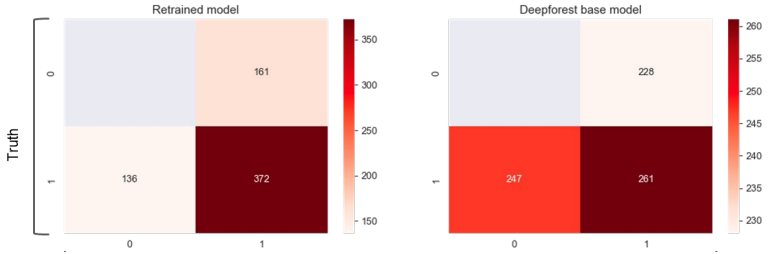


Figure 19: Confusion matrices for base and custom train models

```
F1-Scores on all images:
-----------------------------------------
Model retrained on all images : 0.7080
Model retrained on Louhans images : 0.6817
Model retrained on Manosque images : 0.5675
Model Base DeepForest : 0.5240
```

[5] Onfroy et al., Détection des arbres à partir de données d´imagerie à Très Haute Résolution dans les zones exposées au péril Retrait Gonflement des Argiles In Rapport Scientifique CCR 2022 ; CCR, Paris, France, 2022

4.4 New feature & other results

	lm_trees_count	lm_trees_area	lm_mean_dist	lm_mean_area	lm_density	5m_trees_count	5m_trees_area
count	2585.000000	2585.000000	2585.000000	2585.000000	2585.000000	2585.000000	2585.000000
mean	1.196518	63.164689	6.040312	28.812529	0.274648	3.011605	150.133694
std	1.753167	116.316576	12.588452	35.210272	0.669264	3.218660	195.899878
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	32.480000
50%	1.000000	22.080000	0.000000	19.320000	0.065259	2.000000	95.480000
75%	2.000000	79.200000	7.709191	47.180000	0.287278	4.000000	186.760000
max	17.000000	1217.880000	146.270021	299.040000	9.076617	48.000000	2045.960000

Figure 20: Image to tabular feature

- We have used the new model combined with BD TOPO to **get buildings features related to trees predicted**. We have gathered info in tabular data.

- We have defined for each building surrounded by n trees (distance < 15m), the **TMS score** :

$$TMS = \sum_{k=0}^{n} \frac{S_k}{d_k + 1}$$

Sk: the canopy area of the k-tree
dk: the smaller distance between the k-tree and the building

- Remarks:
  - The score increases with the number of trees around the building,
  - The score increases when the tree-building distance decrease,
  - The score increases when the h-size of the tree increases.

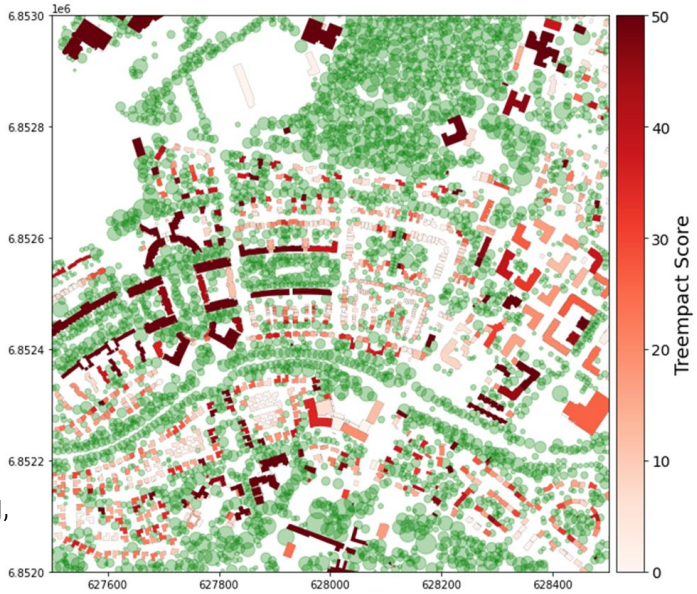


Figure: 21: Calculation of "treempact" scores around the buildings of the area of interest

5. RESULTS & PERSPECTIVES

5.1 Use features enrichment

- All data have **not been yet incorporated** to usual business process of historical modelling.

- Drought events frequency (public statement, district level) are available, so we have created:
  - A time dependent
  - Hypertuned deep learning proxy model
  - That use all data previously introduced

- Dataset:
  - Around 452 initial features (1018)
  - Training time frame 2016/01 to 2019/08
  - 13, 21 and 86 department districts (560)

- LTSM model obtained with grid search

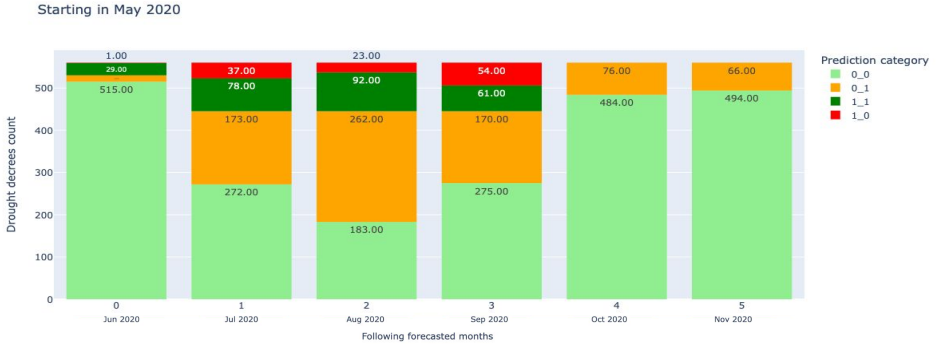


Figure: 22: LSTM results on 2020 test year, categorical representation

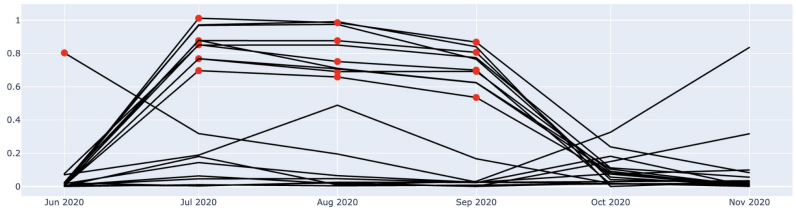


Figure: 23: LSTM nat cat proba inference results on 2020 test year for few districts

5.2 Conclusion & Perspectives

- Use works with historical models, multivariate analysis and **challenge feature importance** (or other metrics)
- Experiment **at scale**: scope (for tree detection 2 departments, Geo service also) and IT.
- More **data to explore** (GEE)

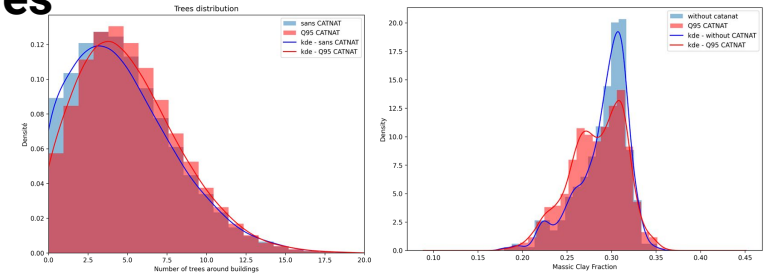


Figure 24: Trees features analysis according nat cat events

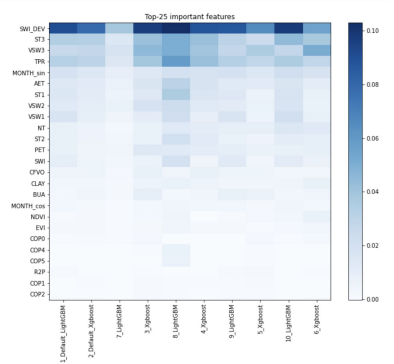


Figure 25: Feature importance tested on basic ML model

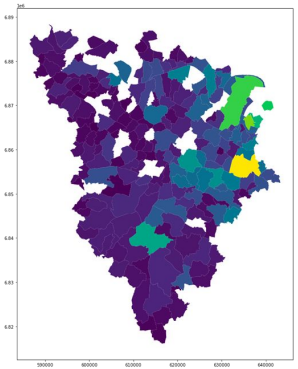


Figure 26: TMS for department 78

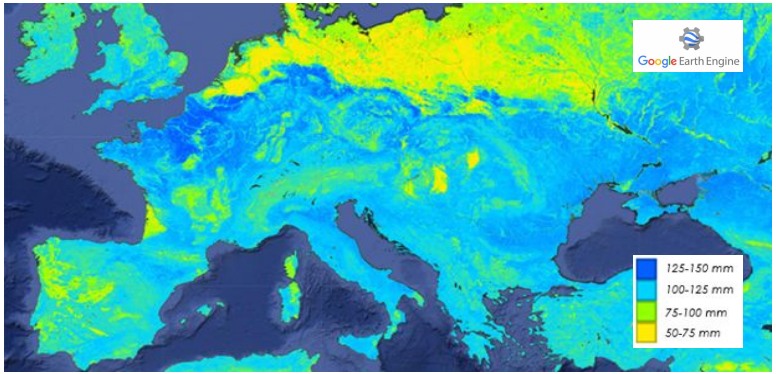


Figure 27: Available Water at Field Capacity calculated with Google Earth Engine (from Attard 2022) [6]

[6] Attard, G. (2022) Implementation of the Thornthwaite-Mather procedure to map groundwater recharge. Google Earth Engine Community tutorials. [link]

Thank you!

Appendix - References

[1] Satriani, A., Loperte, A., Proto, M., & Bavusi, M. (2010). Building damage caused by tree roots: laboratory experiments of GPR and ERT surveys. Advances in Geosciences, 24, 133-137.

[2]Li, J., & Guo, L. (2017). Field investigation and numerical analysis of residential building damaged by expansive soil movement caused by tree root drying. Journal of Performance of Constructed Facilities, 31(1), D4016003.

[3] Weinstein, B.G.; Marconi, S.; Bohlman, S.; Zare, A.; White, E. Individual Tree-Crown Detection in RGB Imagery Using Semi-Supervised Deep Learning Neural Networks. Remote Sens. 2019, 11, 1309

[4] Geographic Generalization in Airborne RGB Deep Learning Tree Detection Ben Weinstein, Sergio Marconi, Stephanie Bohlman, Alina Zare, Ethan P White bioRxiv 790071; doi: https://doi.org/10.1101/790071

[5] Onfroy et al., Détection des arbres à partir de données d´imagerie à Très Haute Résolution dans les zones exposées au péril Retrait Gonflement des Argiles In Rapport Scientifique CCR 2022 ; CCR, Paris, France, 2022

[6] Attard, G. (2022) Implementation of the Thornthwaite-Mather procedure to map groundwater recharge. Google Earth Engine Community tutorials. [link]