RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Cerema
CLIMAT & TERRITOIRES DE DEMAIN

Study of the relationship between the composition of road surfaces and their ability to reflect light.

GONIN Alexis

Within team ENDSUM at CEREMA

Supervised **MUZET Valérie**

⁂

5 août 2025

Internship context

- Cerema

- Projet REFLECTIVITY
  - Optimization of public lighting (perception of obstacles on the road)
  - Better understand the photometry of road surfaces and attempt to predict it
    - A complex problem because it depends on internal factors related to the road composition and external factors such as the age of the pavement.

Internship objectives

- Development of a database to facilitate data handling with pre-cleaning and creation of new features.

- Exploratory data analysis to identify links between surface composition and their ability to reflect light.

- Implementation of a predictive method for photometric data, including several clustering possibilities.

Data sources

International Commission on Illumination
Commission Internationale de l'Eclairage
Internationale Beleuchtungskommission

- TC4-50 database, an international database validated by the CIE (International Commission on Illumination),

- COLUROUTE measurements, exclusively French.

Input Data

- Excel files composed of:
- Photometric data:
  - S1: specularity (or gloss).
  - Q0: average luminance coefficient, representing the total amount of reflected light.
  - Qd: average luminance coefficient for diffuse light (daylight).
- Metadata:
  - 27 metadata fields (age, location, color, etc.).
  - 8 additional standardized fields based on the first 27.
  - Almost exclusively qualitative data (strings).
- Associated R tables (matrices).

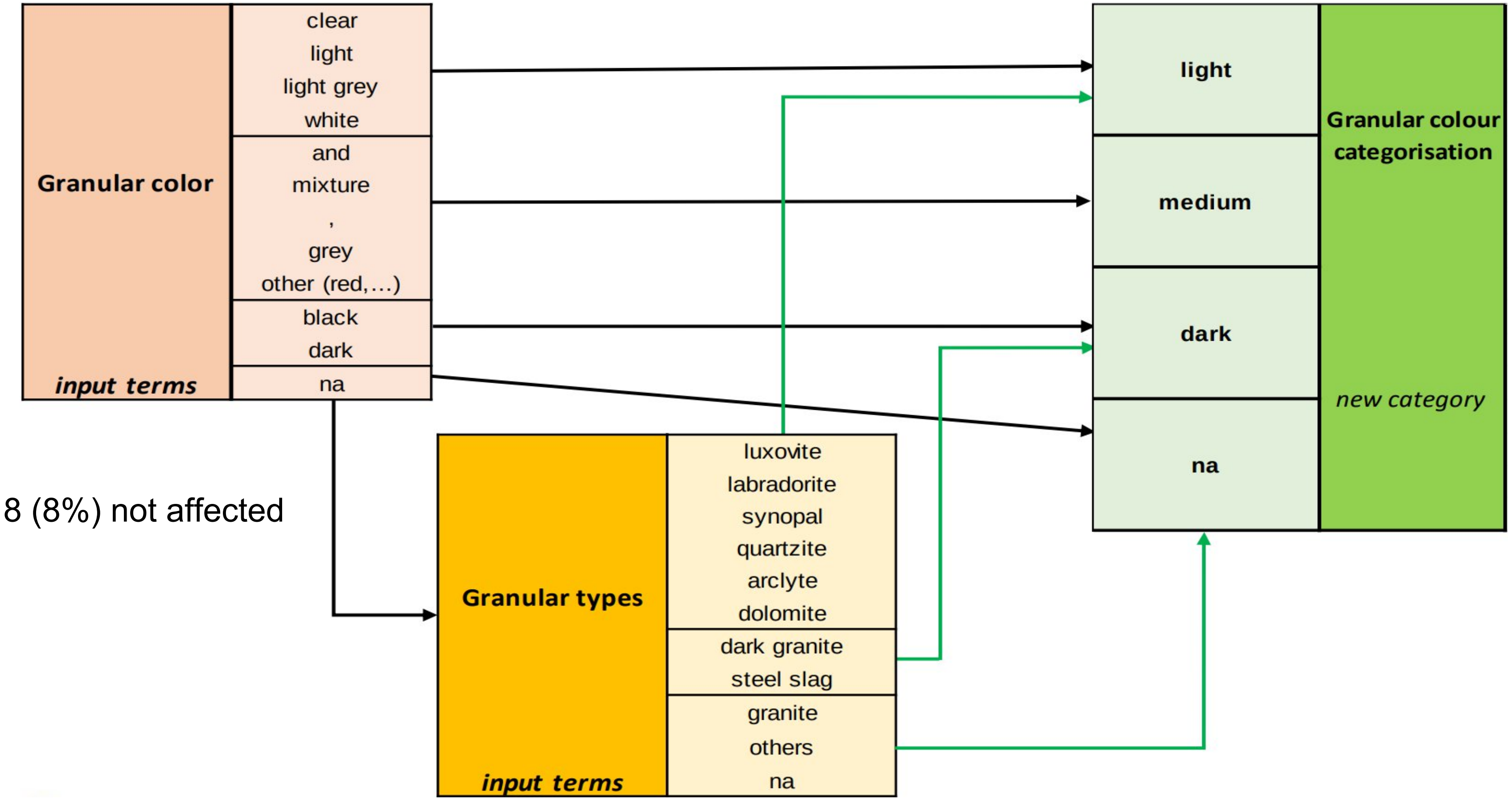
Data cleaning

- Convert to lowercase
- Remove spaces at the beginning and end of strings
- Standardize terms
- To be re-checked with each new entry in the database

sand blasting sandblasting sanded	sand blasted

Adding new features

example for Granular color categorisation



going from 383 (27%) to 118 (8%) not affected

Database Today

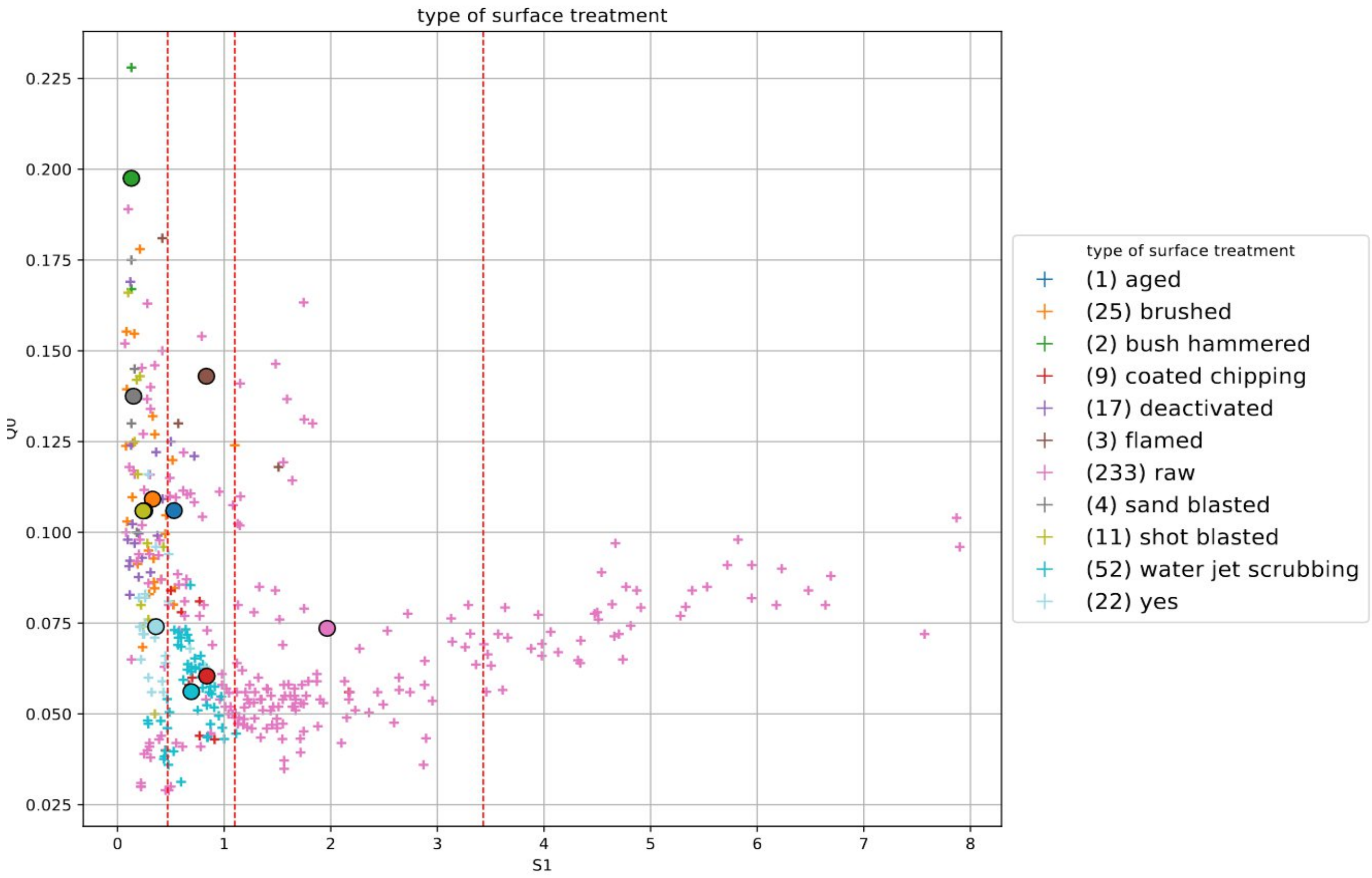
- MONGO DB architecture
- 1397 entries, consisting of 855 TC4-50 entries and 542 field measurements conducted with the COLUROUTE device.
- Memory space used: 11.76 MB
- 2 collections: one for the photometric data and metadata, and one for the associated r tables.
  - First collection: photometric and meta datas
    - 3.01 MB in JSON format
    - 50 unique features
    - Overall data type distribution: str 84.10%, int 6.85%, datetime 0.21%, float 8.85%.
  - Second collection: r tables
    - 4.41 MB in JSON format
    - 3 unique features
    - Overall data type distribution: str 50.00%, list 50.00%.

Development of access and display methods

- Modular and documented Python object oriented architecture
- Queries with conditions on the data and the possibility to select which features to return
- Configurable graphical display functions

2D Graph Display

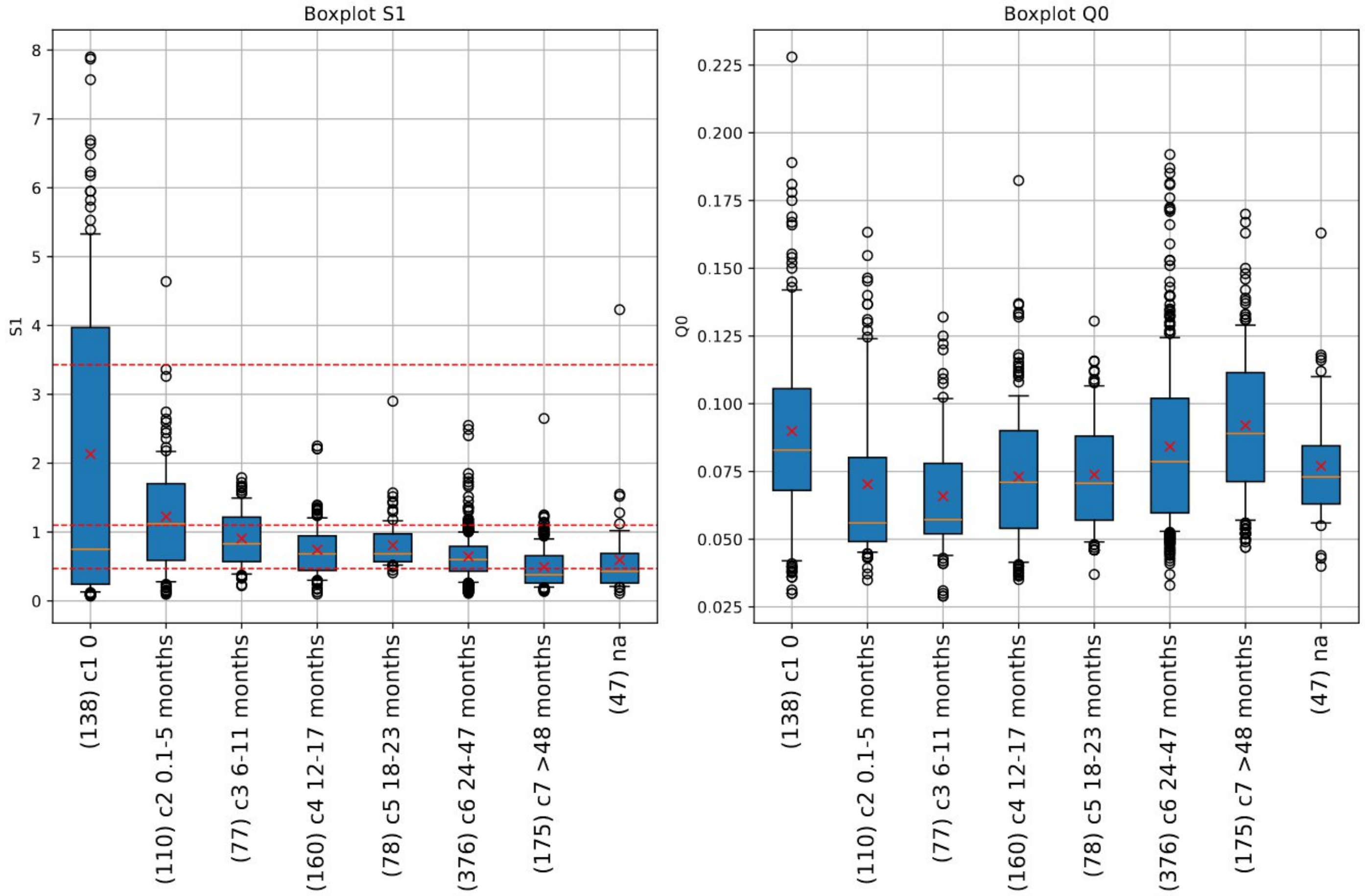
Results conform to litterature



Influence of surface treatments on spécularity

Box plot Display

Results conform to litterature

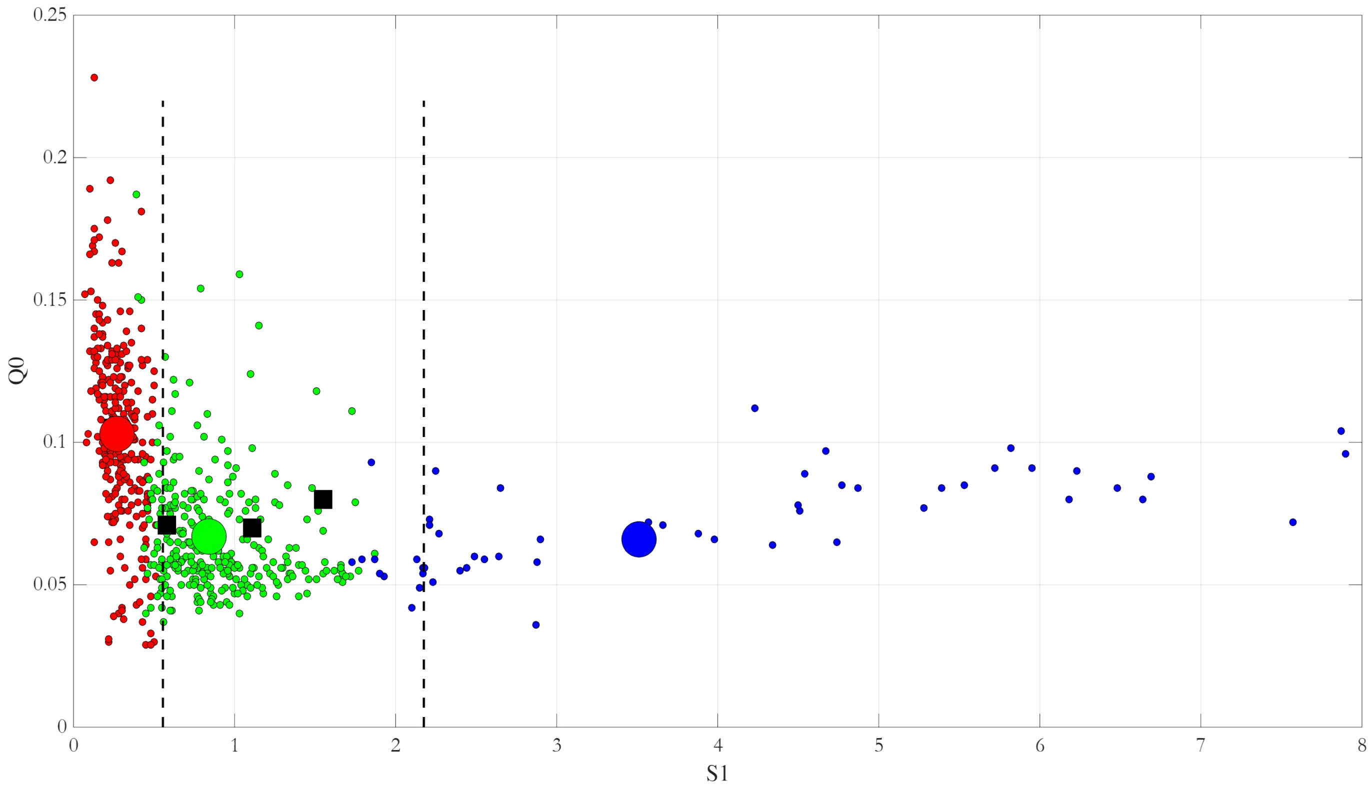


Specularity represented by age in months, with an example of classification threshold in red

Machine Learning

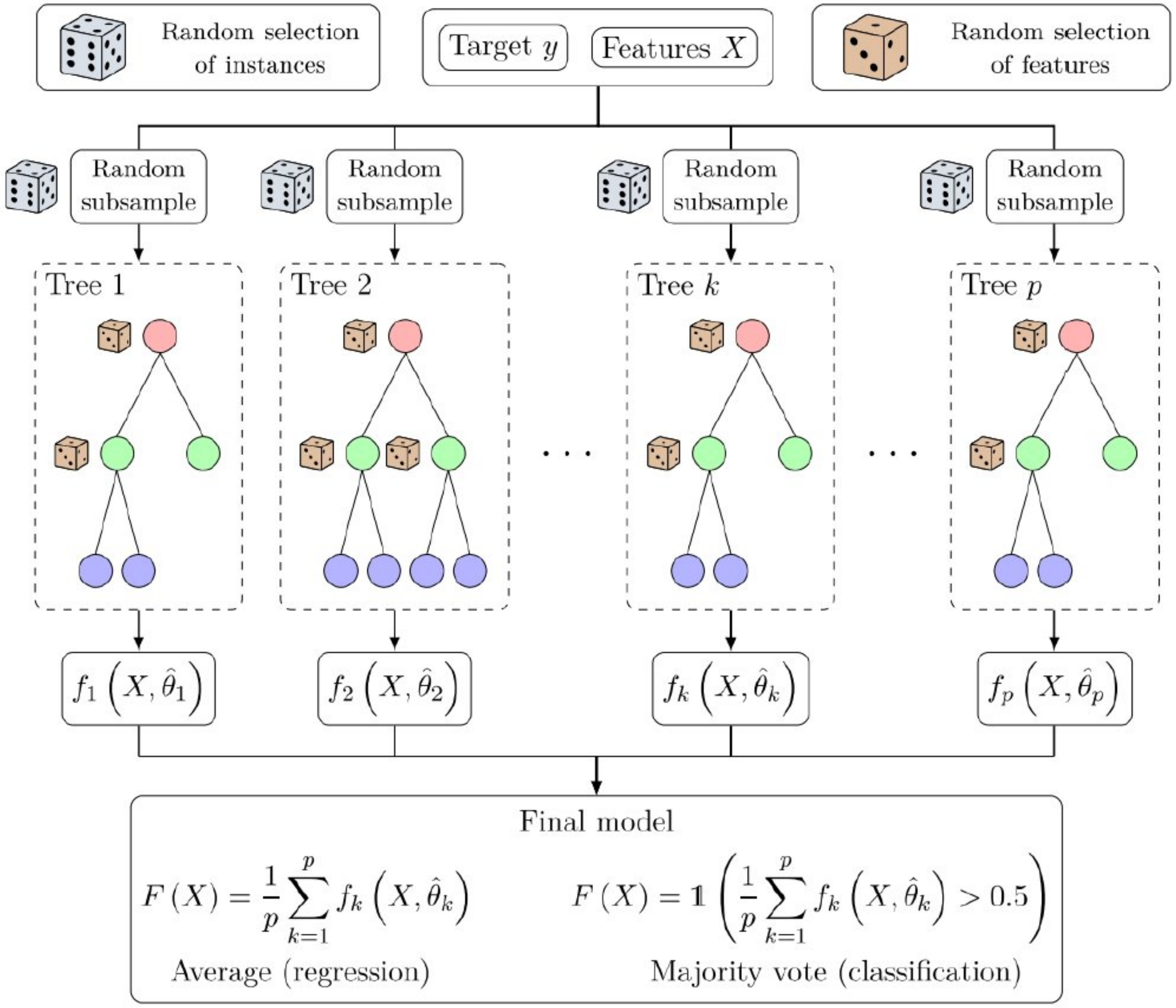
Objectives of machine learning

- Road surfaces are classified according to their specularity S1, which allows typical photometric data to be associated with them.
- Predict the specularity of roads and/or the class to which they belong
- Make the model understandable and interpretable



Model selection

- Decision trees > Deep Learning
- Random Forest > Grandient boosting
- Régressor > classifier for explicability and adaptation
- Out of bag evaluation



Schematic representation of a random forest algorithm

One-Hot encoding

- Make categorical variables usable for learning models.
- Removes the notion of order compared to numerical encoding (cement concrete = 1, surface coating = 2, ...).
- Improves model performance.
- Each variable becomes a binary feature.

general family	general family_cement concrete	general family_surface coating	general family_natural material	general family_bituminous mixture
cement concrete	1	0	0	0
surface coating	0	1	0	0
natural material	0	0	1	0
bituminous mixture	0	0	0	1

Python implementation

- Python library scikit learn :
  - Build and evaluate models
  - One-hot encoding
- Python class specially developed to facilitate usage

Results

Data used for training	Model	Default parameters	Optimised by gridsearch	Out-of-bag evaluation
TC4-50	RandomForestRegressor	0,731	0,765	0,809
	HistGradientBoostingRegressor	0,716	0,716	–
All database	RandomForestRegressor	0,817	0,819	0,825
	HistGradientBoostingRegressor	0,803	0,807	–
TC4-50 ≥24 months	RandomForestRegressor	0,429	0,461	0,445
	HistGradientBoostingRegressor	0,443	0,458	–
All database ≥24 months	RandomForestRegressor	0,377	0,401	0,454
	HistGradientBoostingRegressor	0,422	0,444	–

Comparison of R^2 (coefficient of determination) performance across different datasets and model configurations

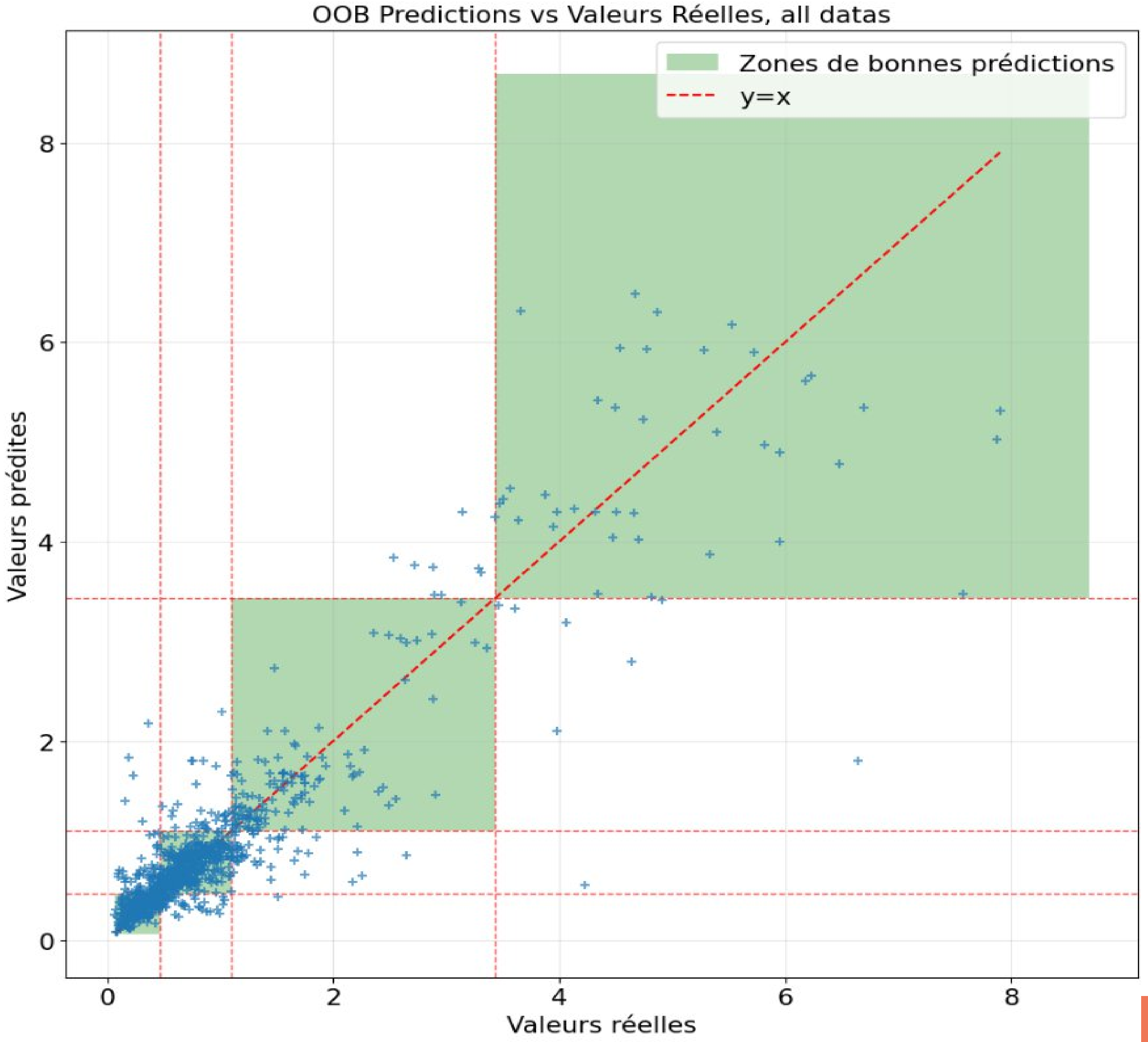
Results

Métrique	**All datas**	**TC4-50 + âge ≥ 24 mois**
R^2	0.825	0.445
Accuracy (3 classes)	0.856	0.801

Performance d'une classification pour 3 classes dans le pire et le meilleur scénario

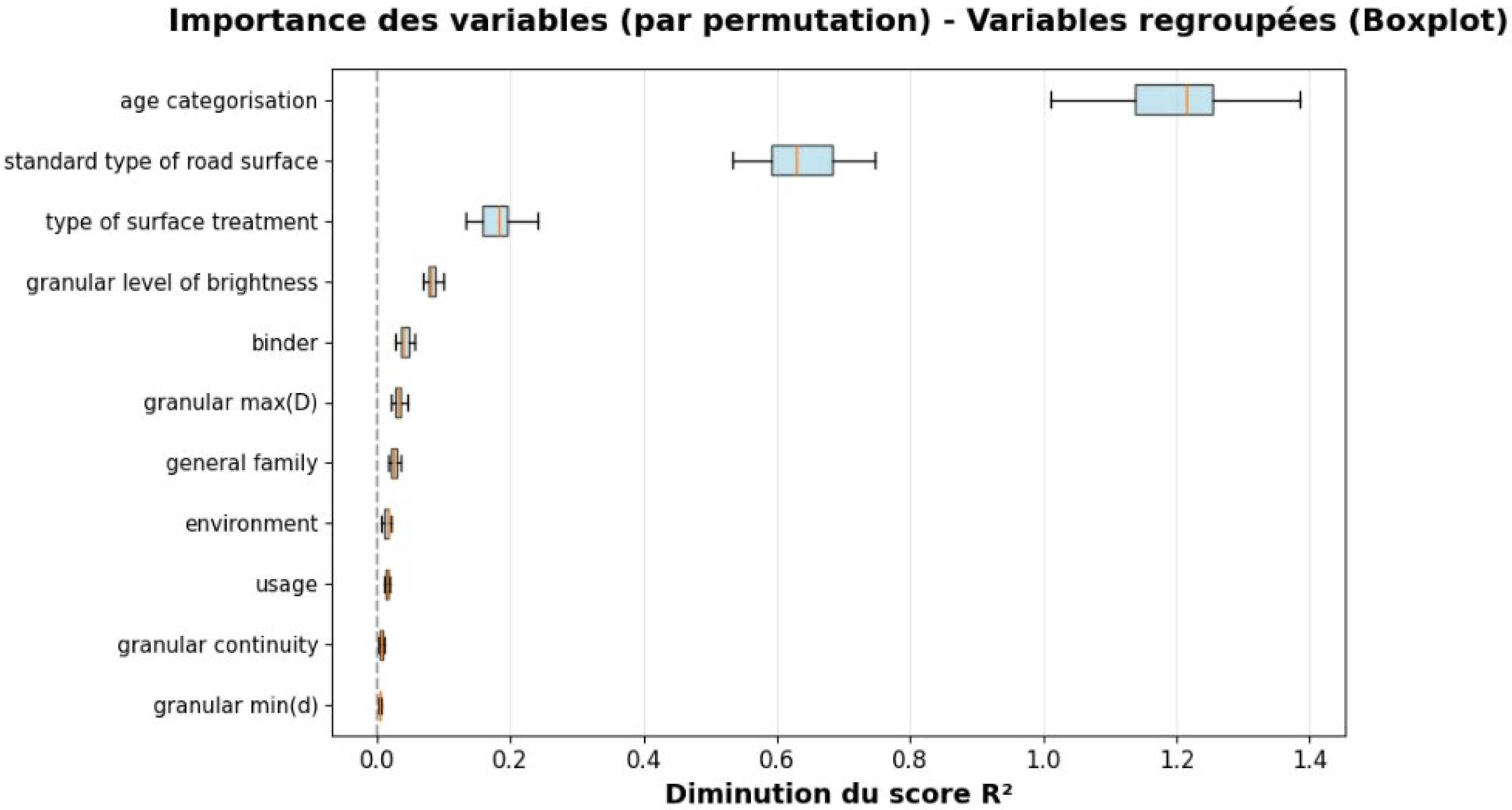
Results

Graphical visualization of the best model performance in the best scenario.



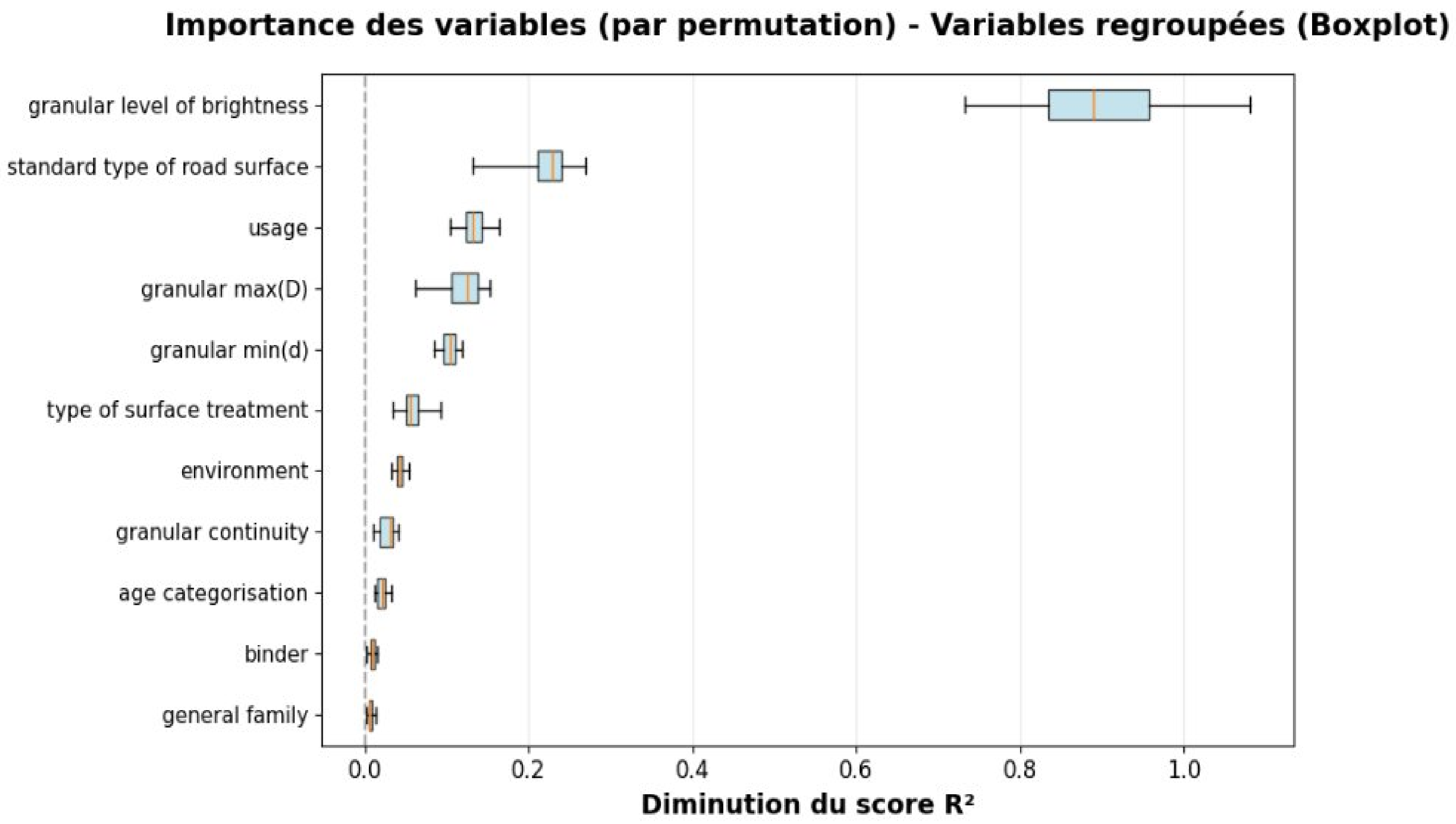
RandomForest interpretation and results explicability

Features importance



Permutation feature importance sorted by decreasing regression score (30 evaluations), Random Forest trained on all datas.

Features importance



Permutation feature importance sorted by decreasing regression score (30 evaluations),
Random Forest trained on roads younger than 24 months

Shap Values : theory

- Based on Shapley values from game theory
- Fairly divides (according to contribution to coalitions) the gains of a game among its participants
- Satisfies 4 axioms:
  - Efficiency: all the gain is distributed among the players
  - Symmetry: if 2 players are interchangeable, they must receive the same gain
  - Null player property: if a player makes no contribution, they receive nothing
  - Additivity: if two games are combined, the total contribution of a player is the sum of their contributions to both games

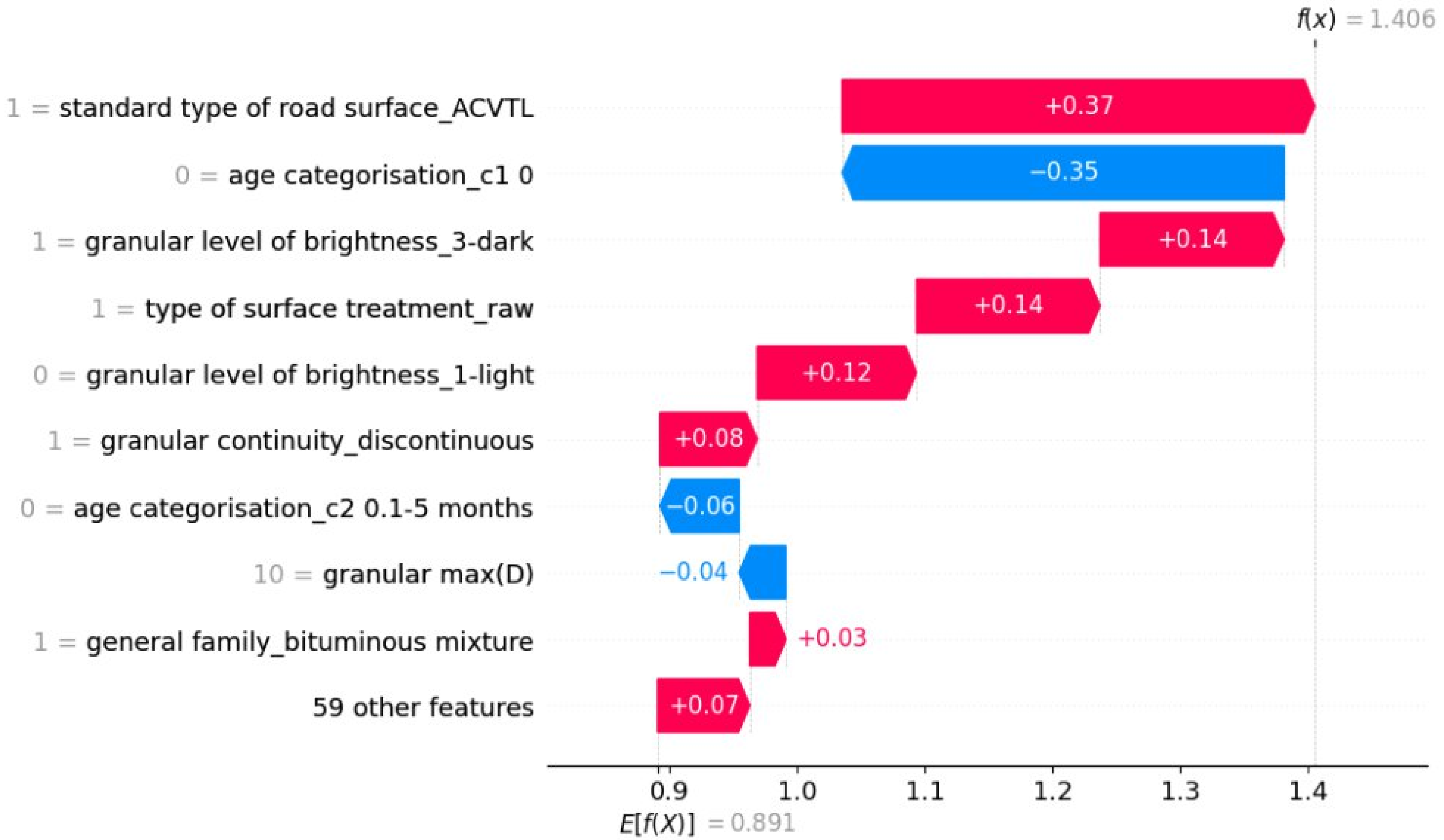
* Lundberg, et al. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat Mach Intell* **2020**, *2* (1), 56–67.

Shap Values applied to machine learning

- Games become predictions and players become features.
- Python library: SHAP
- Associates each variable of each feature for each entry with a “shap value” corresponding to its contribution.
- Allows explanation of a “black box” model.

Shap values

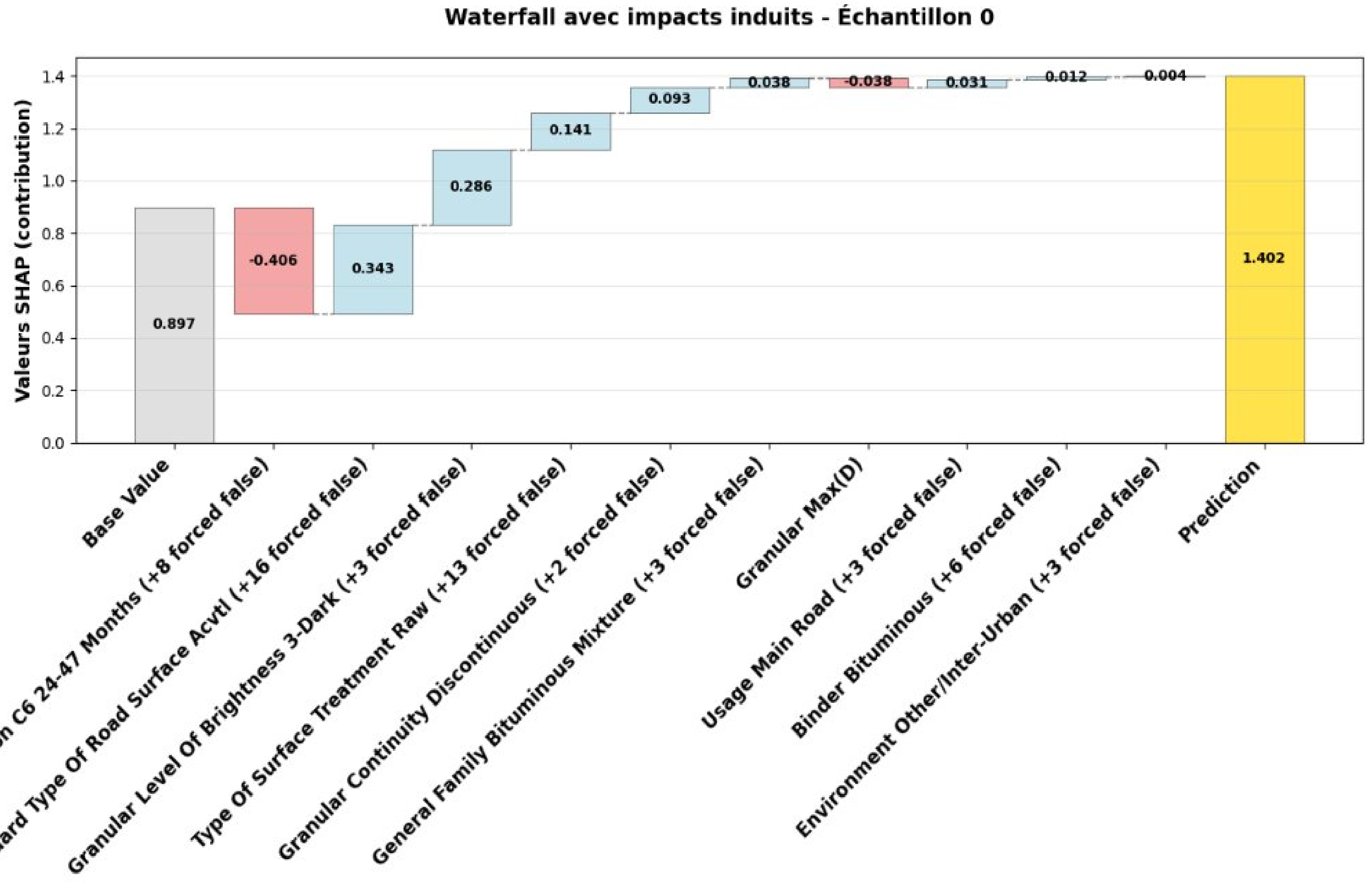
- Waterfall plot
- E[f(x)]+shap values = f(x)
- Too many one-hot features



Visualisation of SHAP values for a single prediction

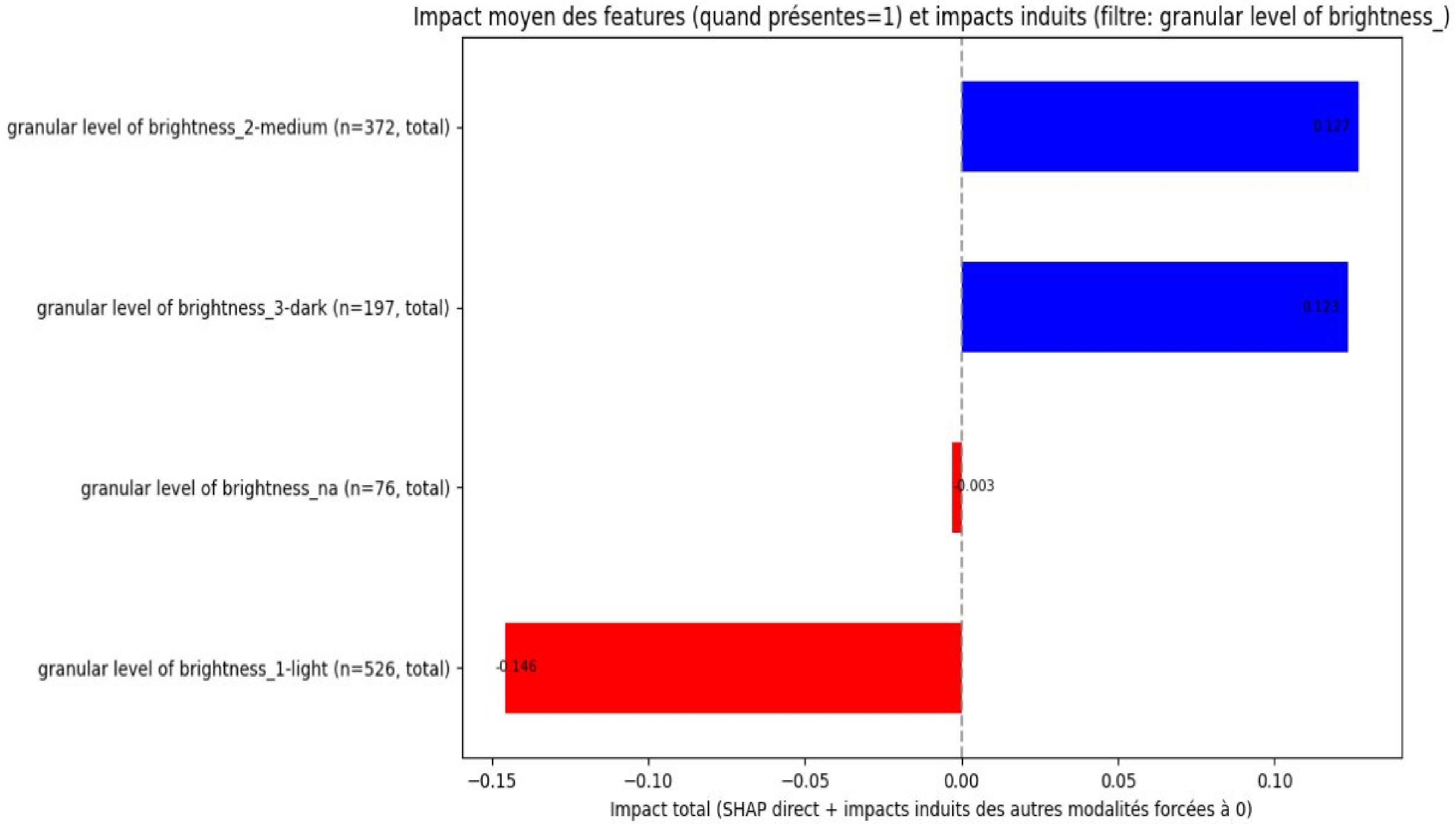
Shap values

- Waterfall plot with cumulative impact by feature



Shap values

- Mean impact of one-hot features



Linear model construction

- With these average feature impact values (taking into account induced impacts), a linear model can be constructed:

$$f(x) = E[f(x)] + \sum_{i=1}^{M} \phi_i \alpha_i$$

- E[f(x)] mean value of the predicted specularity
- ϕ_i average shap values including induced impact for each feature
- α_i is 0 if the feature is 0 and 1 if the feature is 1
- Easy to implement or calculate on the field

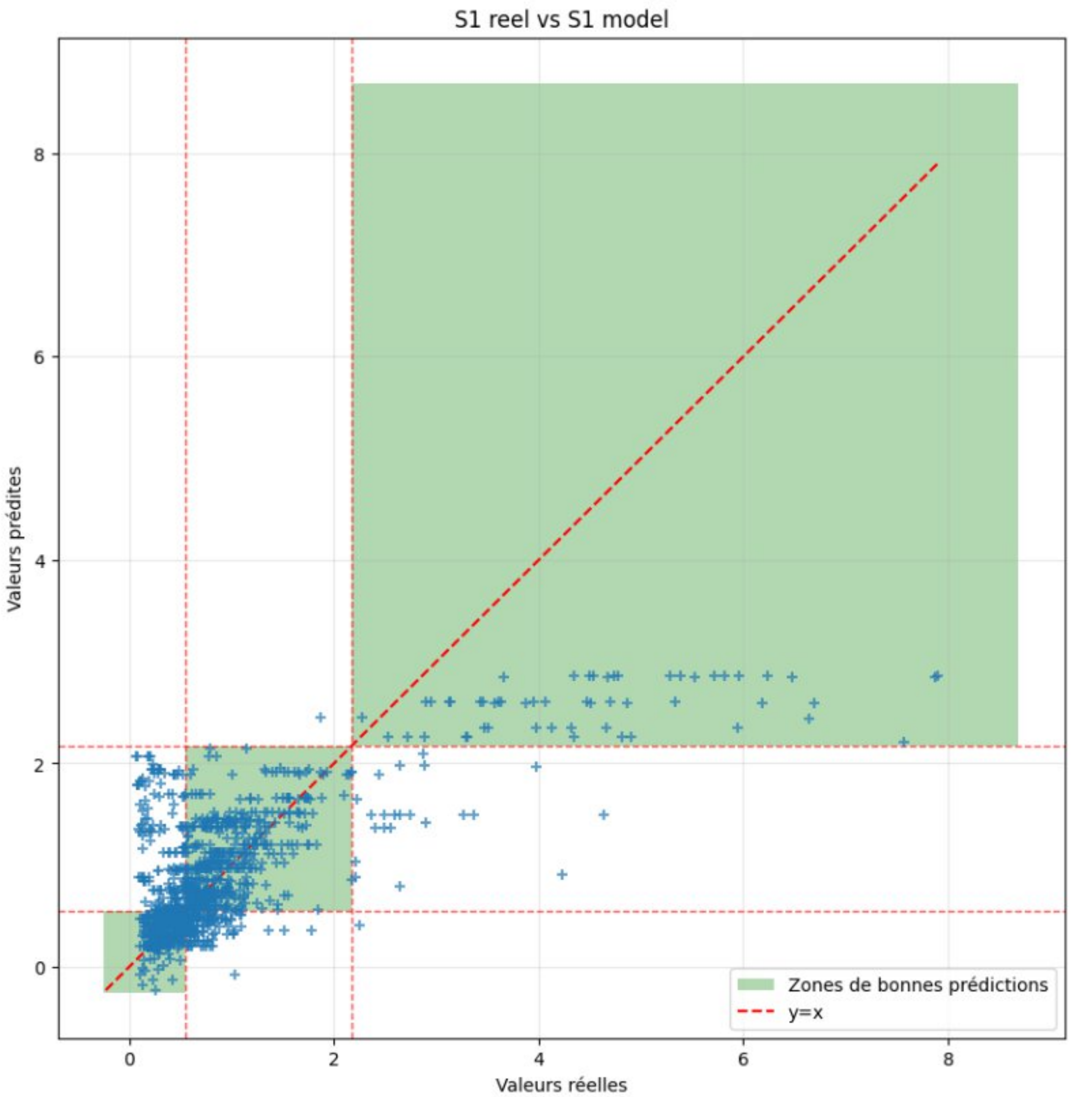
Linear model results

Metric	Optimised RandomForest	Custom linear Model
r^2	0,823	0,475
Accuracy (3 classes)	0,85	0,74

Comparison of performances between Optimised RandomForest and my Linear SHAP values based model.

Linear model results

Graphical visualization of the linear model's performance



Work Transmission

- Containerization VsCode/Docker
- Code on Cerema’s GitLab
- Simple usage through Jupyter notebooks.
- Effort has been made to ensure code reusability for future users.

Conclusion

- MongoDB database from Excel sources
- Visualisation
- Several machine learning models tested, best being random forest
- Interpretation methods

- Usefull tools for future researchs