Classification et Segmentation de séries temporelles

Journée thématique
"activités Machine Learning à l'IPHC"

Sommaire

- Données
- Méthodes
  - Machine Learning en R
  - Deep Learning en Python et Keras
- Projets
- Expertises
- Annexes
  - Bonnes pratiques
  - Documentation et liens

DataSet WISDM

- Smartphone and Smartwatch Activity and Biometrics Dataset https://www.cis.fordham.edu/wisdm/dataset.php#actitracker
- Type : Raw Time Series Data (Accelerometer X,Y,Z)
- Frequency : 20Hz
- Number of examples: 1,098,207
- Number of attributes: 6
- Class Distribution
  - Walking: 424,400 (38.6%) id:5
  - Jogging: 342,177 (31.2%) id:1
  - Upstairs: 122,869 (11.2%) id:4
  - Downstairs: 100,427 (9.1%) id:0
  - Sitting: 59,939 (5.5%) id:2
  - Standing: 48,395 (4.4%) id:3

Machine Learning en R

- Principe de fonctionnement :
  - Définition d'un dataset Segment * Features = Classe
  - Recherche de la taille des segments
  - Détermination de la classe par segment
  - Calcul des « Features » : magnitude, zerocross, peak2peak, rms, means, stdev, kurtosis, skewness, peak/rms, entropy, autocorelation, percentile, spectral energy, principal frequency, ...
  - Augmentation des données :
    - Ajout de voies (filtres, spectre magnitude, phase,…)
    - Overlap, cropping, flipping
- Documentation : Feature extraction for robust physical activity recognition https://link.springer.com/article/10.1186/s13673-017-0097-2

Machine Learning en R

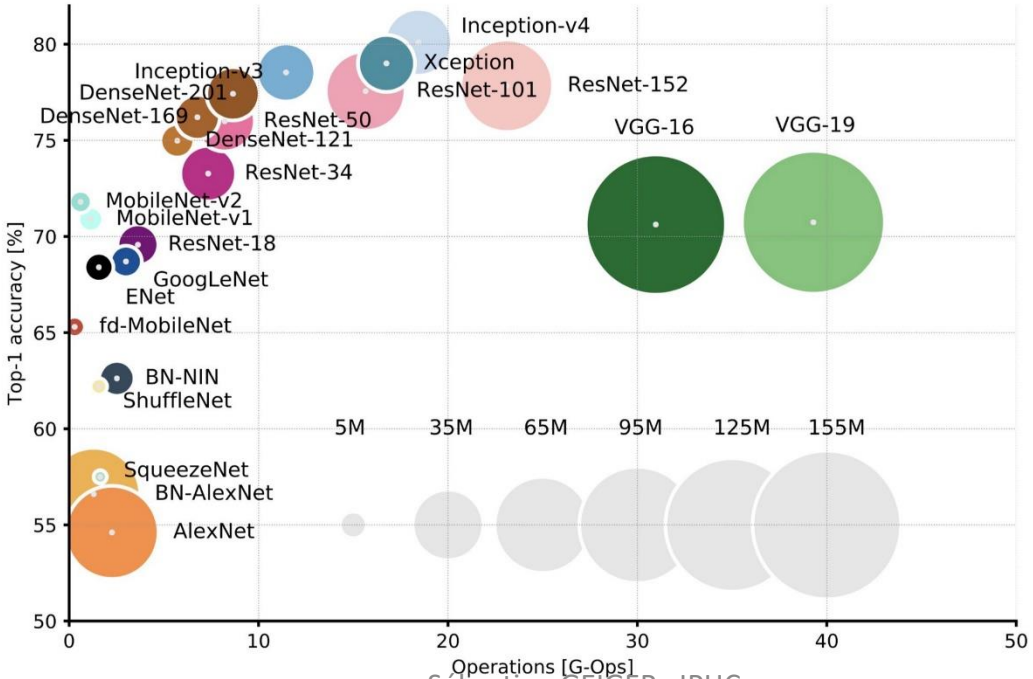
- Taille des segments 200 obs, overlap 40 obs
- Calcul des features : 54710 Obs * 65 variables
- Réduction des dimensions PCA ou Corrélation features : 54710 Obs * 40 variables
- Création jeux d'entrainement 80% et de test 20%
  - train: 43770 Obs * 40 variables
  - test : 1090 Obs * 40 variables
- Evaluation avec différents algorithmes de classification

Machine learning en R

- Résumé :
  - Calcul effectué sur mon laptop sans GPU ;)
  - randomForest :0.9661 (33.930s)
  - xgboost :0.9803 (68.937s)
  - kknn :0.9819 (14.912 s)
- Autres Algos de classification :
  - LightGBM
  - HistGradientBoostingClassifier
- Recherche optimum des Hyper-paramètres
  - Package SuperML (~ scikit-learn's fit,predict,transform)
  - RandomizedSearchCV, GridSearchCV

Deep Learning (Python et Keras)

- Utilisation des réseaux CNN et RNN
- Support 3D, 2D ou 1D (Time Séries)
- Différentes topologies



Principe de fonctionnement

- Détermination des segments et des classes
- Le réseau CNN ou RNN détermine les « features »
- Augmentation des données
- Déterminer les hyper-paramètres
- Utilisation de Colab avec support du GPU
  (Sans GPU 10 minutes, avec GPU 16 secondes)
- Résultat :
  - Resent 1D : 0.93 (30*6s)
  - Conv2D-MelSpectogram : 0.88 (30*16s)
  - LSTM : 0.93 (30*45s)

Projet : rblt-maps

- Projet ANTIDOT de Damien CHEVALLIER et de la thèse de Lorène JEANTET
- Développement d'un outil accessible depuis un navigateur pour des scientifiques spécialisés dans le suivi des tortues marines permettant la visualisation des relations entre données des Bio-loggers et les positions GPS
- Utilisation de modèle(s) pré-entraîné(s) pour identifier automatiquement les comportements à partir des données des Bio-loggers
- Affichage des relations entre comportements et position GPS
- Reconstruction des trajectoires ou plongées en 3D
- Technologie : R, Shiny, Docker, Python, Keras

Projet : Détection d'anomalies dans les métriques de métrologie

- Aide à l'exploitation des métriques de la plate-forme SCIGNE pour déterminer l'état des ressources
- Stockage ElastiqueSearch (base NoSQL)
- Dashboard Grafana, Kibana
- Collecteur Logstash, Prometheus, syslogs, …
- Méthode: détermination des situations exceptionnelles ou des comportements déviants
  - LDOF (Local Distance Outlier Factor), LOF (Local Outlier Factor)
  - Quartiles-based : calcul des quartiles et seuil sur les valeurs min et max
  - SARIMA : stands for Seasonal Auto Regressive Integrated Moving Average
  - Isolation Forest

Expertises

- Outils : Jupyter, Python, Numpy, Pandas, Keras, Scikit-Learn, R, RStudio, Shiny, Colab, C++
- Gestion des données : Alignement des données, filtrages numériques, gestion des formats, TSDB, NoSQL, Segmentations, exploration des Datas
- Infrastructure : Cloud, Docker, Stockage (kubernetes, stockage S3, Ferme GPU du cc)
- Classification / segmentation 2D, la théorie
  - Déterminer un chiffre, chien ou chat ? ;)

Annexes

- Informations complémentaires
- Bonnes pratiques
- Exploration de données

Liens

- Ecole informatique https://gitlab.in2p3.fr/ri3/ecole-info/2020/anf-machine-learning
- scikit-learn https://scikit-learn.org/stable/
- WEKA The workbench for machine learning : https://www.cs.waikato.ac.nz/ml/weka/
- DATA MINING https://www.cs.waikato.ac.nz/ml/weka/book.html
- DataSet public pour l'entraînement :
  - WISDM Smartphone and Smartwatch Activity and Biometrics Dataset Data Set
  - DaLiAc Daily Life Activities https://www.mad.tf.fau.de/research/activitynet/daliac-daily-life-activities/
  - FORTH-TRACE Dataset https://github.com/spl-icsforth/FORTH_TRACE_DATASET
  - UC Irvine Machine Learning Repository https://archive.ics.uci.edu/ml/index.php

Evolution des besoins

- Fourniture de ressources via SCIGNE
- Demande de support
  - R, RStudio, Configuration environnement, Shiny
  - Lecture et visualisation des données
  - Alignement des données bio-loggers, carte GPS
- Nouveaux profils: Data scientist, Data architect
- Gestion des données pour le Big Data
- Outils adaptés à la gestion des Logs ?

Formations

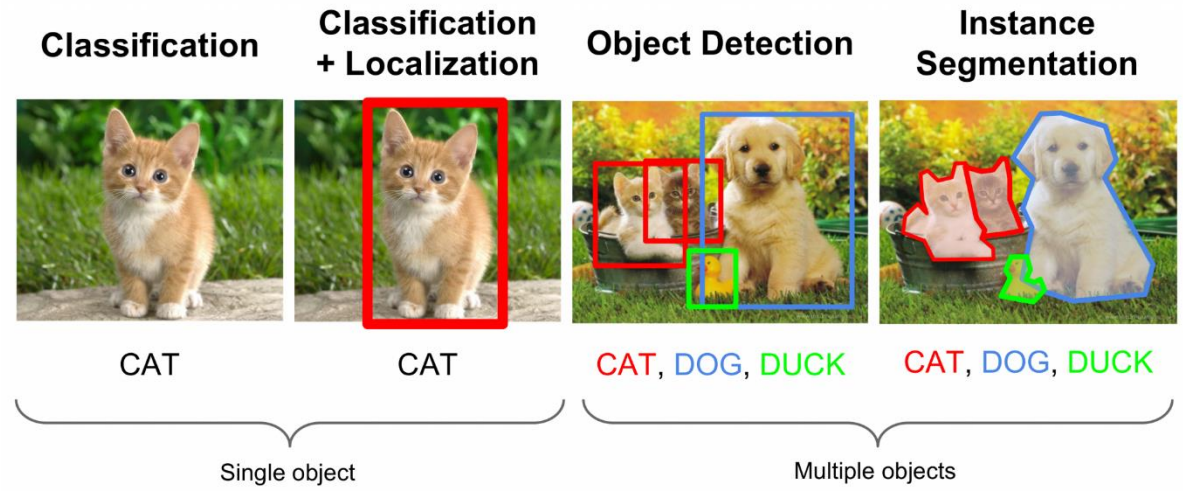
- Ecole informatique: "Concepts et utilisation du Machine Learning pour les informaticiens"
- Fouille de données Master2 Unistra
- Mooc : statistique avec R, Deep Learning, Fondamentaux pour le Big Data, ...
- Livre : L'apprentissage profond avec Python, François Chollet

Bonnes pratiques

- Connaissance des données
- Filtrages et traitements numériques
- Déterminer les relations entre les variables et les classes
- Correction de l'équilibre des classes
- Matrice de confusion
- Recherche des hyper-paramètres
- Classification / Segmentation
- Transfert Learning

Classification / Segmentation

- Des modèles pour chaque usage

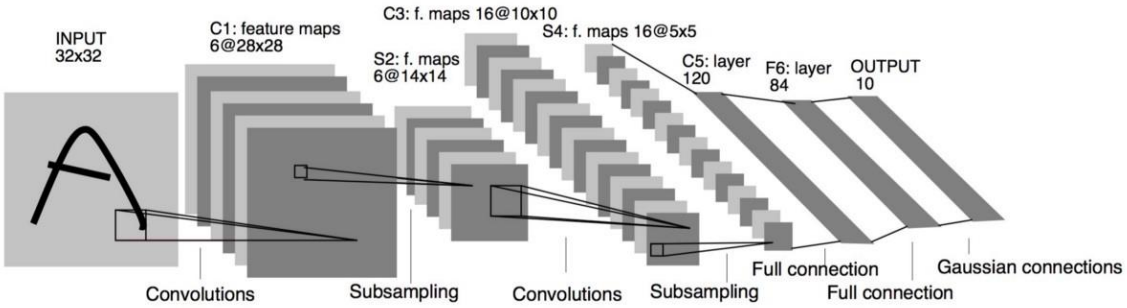


https://medium.com/datadriveninvestor/deep-learning-for-image-segmentation-d10d19131113

- Segmentation: FCN32, Segnet, U-Net, V-Net

Deep Learning (Python et Keras)

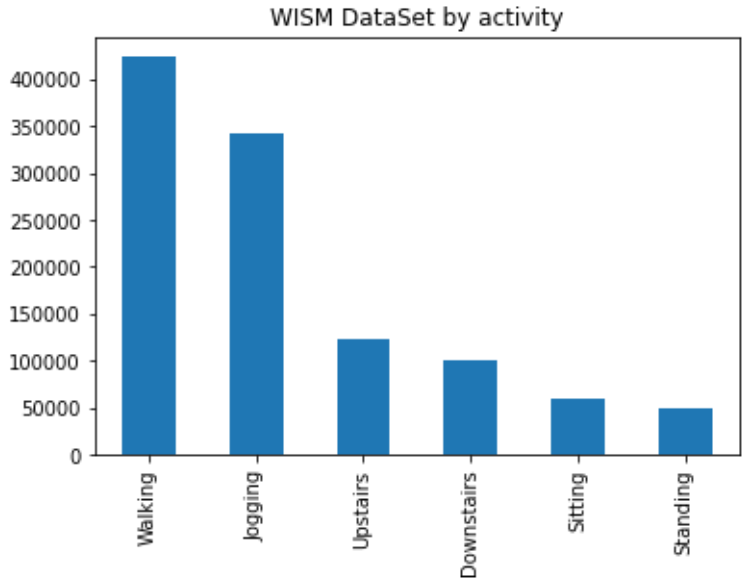
- LeNet5 : Yann LeCun en 1988 ;)



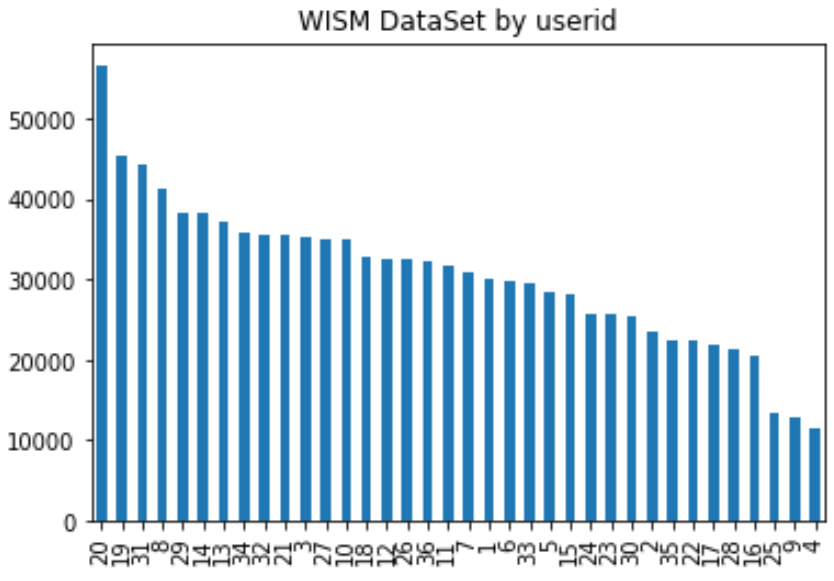
- https://towardsdatascience.com/neural-network-architectures-156e5bad51ba
- Utilisation des réseaux CNN en Deep Learning https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148

DataSet WISDM

- Par activités

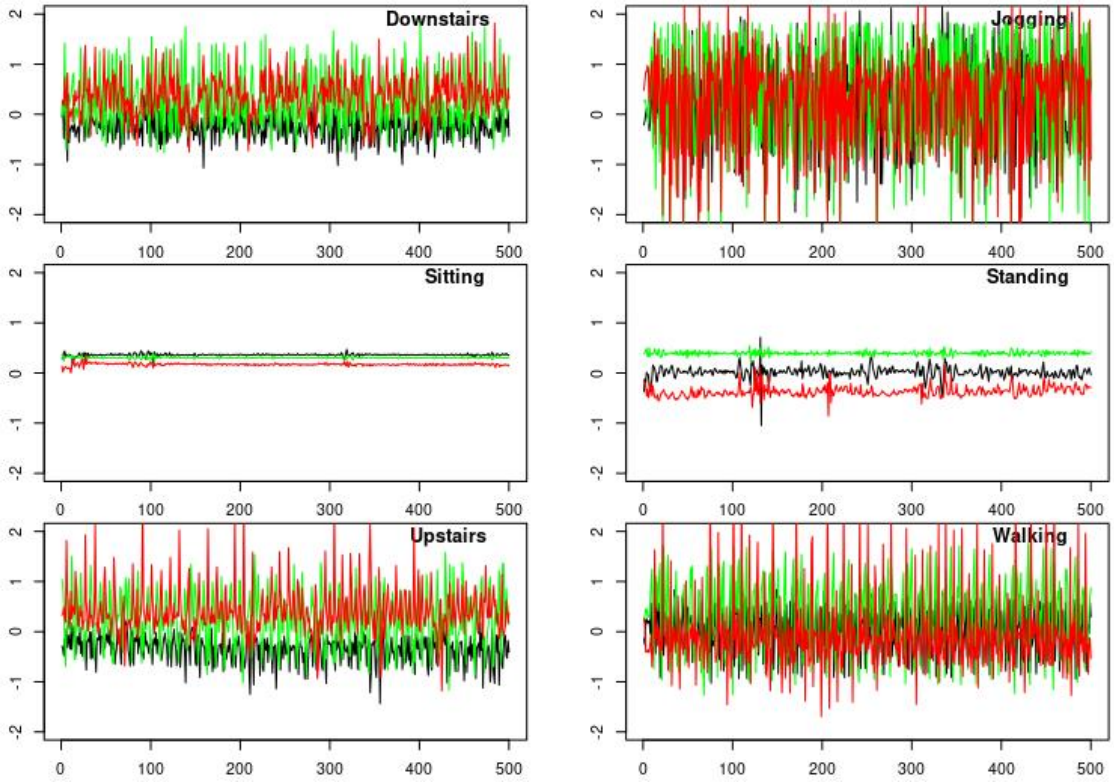


Par utilisateur



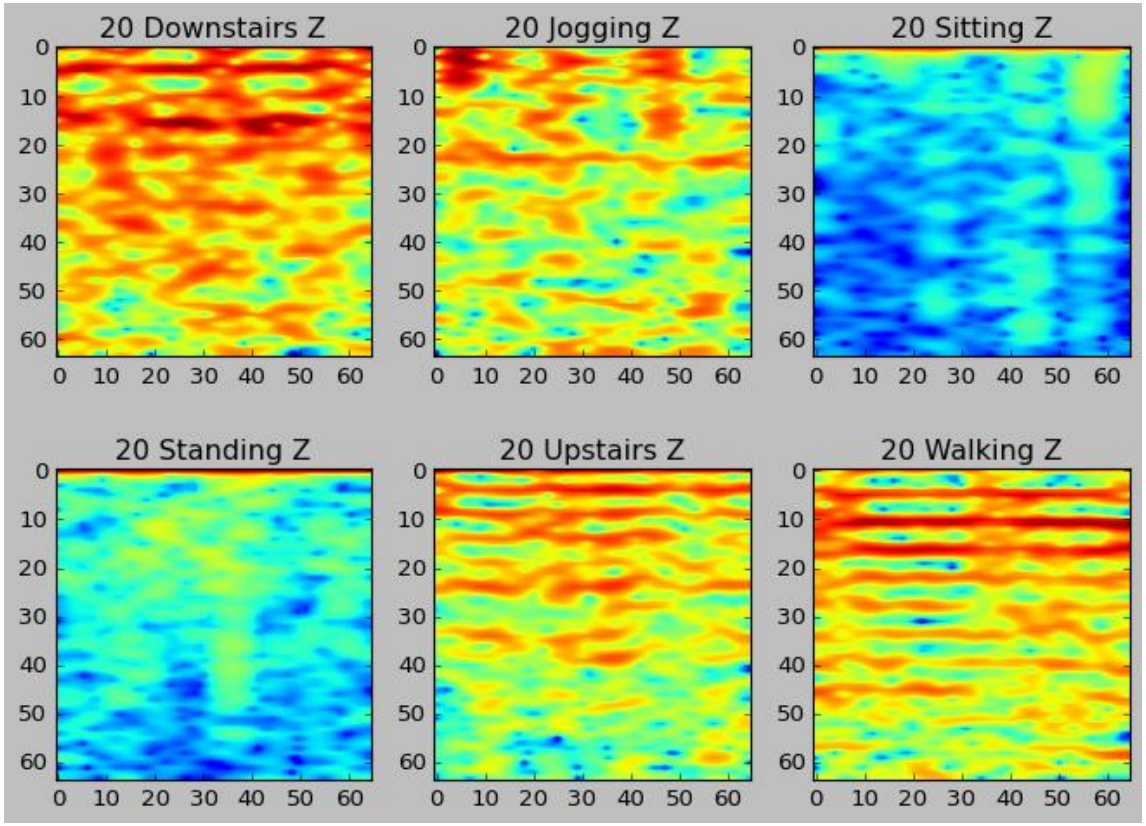
DataSet WISDM x,y,z / activity

- Exemple signaux x,y,z



DataSet WISDM x,y,z / activity

- Exemple de spectrogramme Z par activités



ML- RandomForest

```
nbcore  :  1
fit     :  33.930 s
predict :  0.152  s
Confusion Matrix and Statistics
          Reference
Prediction        0    1    2    3    4    5
             0  893   12    2    1   59   37
             1    4 3385    2    0   10   20
             2    3    0  558   12    1    0
             3    3    2   26  435    1    2
             4   48   21    3    1 1129   26
             5    8   40    0    0   27 4169
Overall Statistics   Accuracy : 0.9661
```

ML-xgboost

```
nbcore  : 6 (support du multicore par défaut)
fit     : 68.937 s
predict : 0.258 s
Confusion Matrix and Statistics
          Reference
Prediction    0    1    2    3    4    5
         0  946    2    1    5   38   12
         1    5 3393    4    1    6   12
         2    1    0  557   14    2    0
         3    1    0    8  456    1    3
         4   50    7    2    0 1164    5
         5    8   13    0    0   15 4208

Overall Statistics : Accuracy : 0.9803
```

ML- kknn

```
nbcore   : 1
fit      : 14.912
predict  : 0.0
Confusion Matrix and Statistics
          Reference
Prediction    0    1    2    3    4    5
         0  964    6    0    1   38    6
         1    5 3394    1    0    1   18
         2    4    2  552   25    3    1
         3    4    0   20  441    1    0
         4   22    5    1    0 1181    9
         5    5   14    0    2    4 4210
Overall Statistics Accuracy : 0.9819
```

DL - Resent 1D

```
Total params          : 512,038
Temps d'apprentissage : 30*6s = 180s
Temps prédiction      : 1s
Accuracy              : 0.93
              precision    recall  f1-score   support

           0       0.83      0.76      0.80       324
           1       1.00      0.96      0.98       992
           2       0.94      0.92      0.93       225
           3       0.90      0.91      0.91       186
           4       0.84      0.80      0.82       365
           5       0.92      0.98      0.95      1200

    accuracy                           0.93      3292
```

DL - Conv2D MelSpectogram

Total params : 67,240,262
Temps d'apprentissage : 30*16s
Temps prédiction : 1s
Accuracy : 0.88

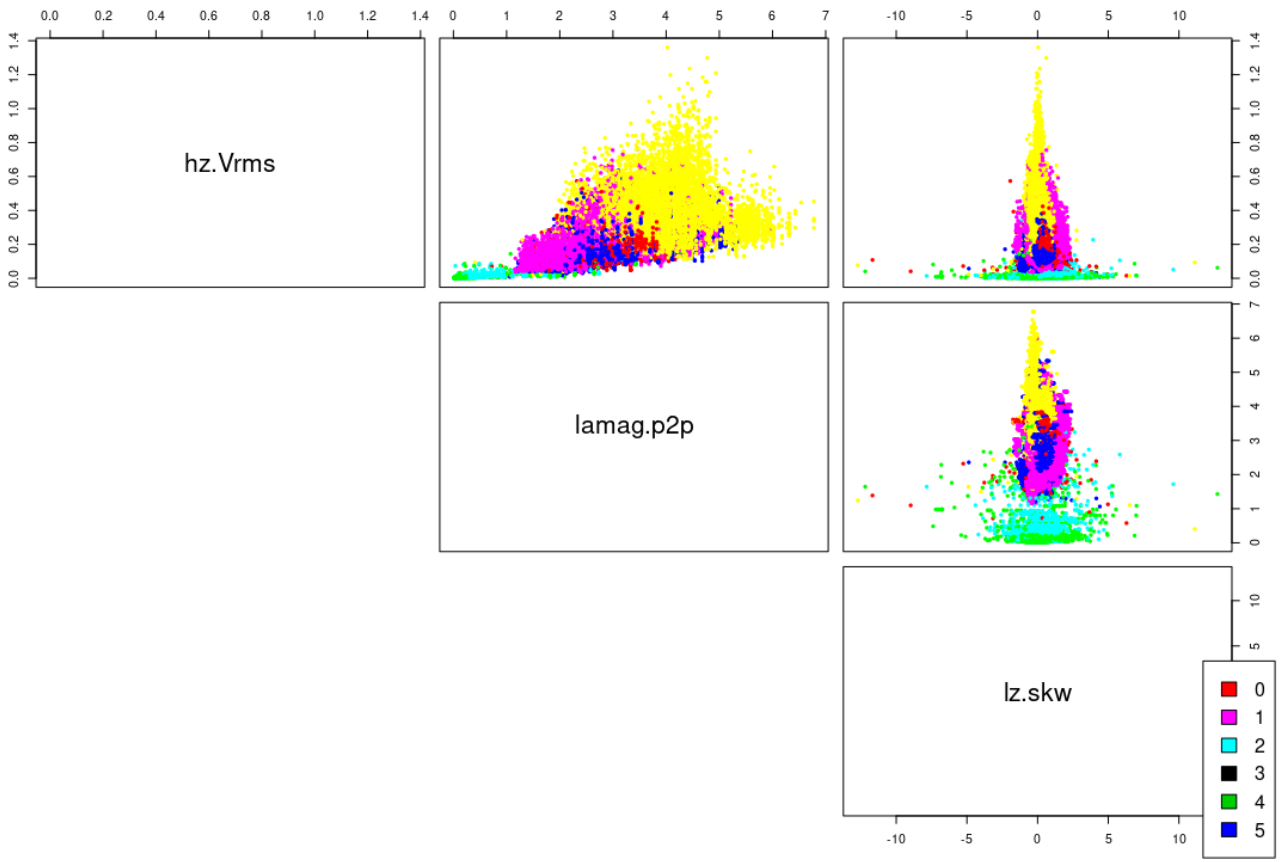
	precision	recall	f1-score	support
0	0.67	0.81	0.73	206
1	0.98	0.91	0.94	620
2	0.95	0.86	0.90	141
3	0.85	0.92	0.89	117
4	0.72	0.59	0.65	223
5	0.91	0.96	0.94	749
accuracy			0.88	2056

DL - LSTM

```
Total params          : 563,078
Temps d'apprentissage : 30*45s
Temps de prédiction   : 1s
Accuracy on test data : 0.93
```

	precision	recall	f1-score	support
0	0.76	0.87	0.81	446
1	1.00	0.92	0.96	1669
2	0.88	0.97	0.92	354
3	0.97	0.80	0.88	252
4	0.79	0.85	0.82	500
5	0.95	0.98	0.97	1771
accuracy			0.93	4992

graphes de corrélations de variables

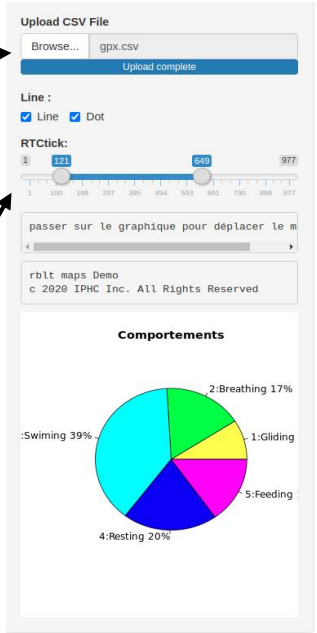


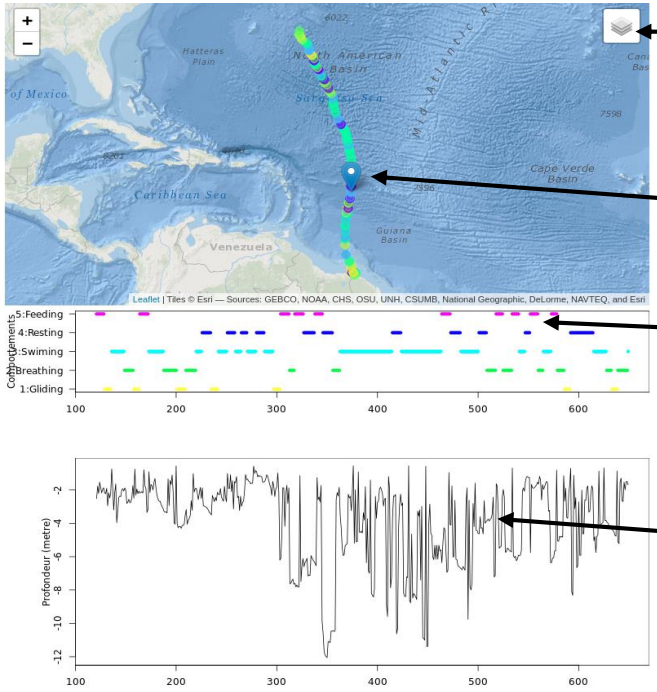
rblt-maps

Développement « Shiny app » et support via SCIGNE la plateforme de CLOUD à l'IPHC

Bouton upload des données

Sélection paramètres d'affichage et périodes de temps





Changement du fond Normal Marin ou Terrain

Point GPS + heure

Comportements

Profondeur