A new altimetry data validation approach based on Data Mining and Machine Learning techniques

Eric Jeansou[1]; Romain Bergougnoux[1] , Sophie Le Gac[2]; Charlotte Garcia[2]; Mathilde Cancet[1]; Florence Toublanc[1]; Sylvain Brunato[1]

Corresponding Author: eric.jeansou@noveltis.fr

[1] NOVELTIS, France - [2] CNES, France

Introduction

Data mining techniques allow scientists to extract and evaluate efficiently tendencies from large databases. In that context, the purpose of this study is to explore the potential of *Data Mining* and *Machine Learning* methods to assess the validity of altimetry measurements over ocean and compare their performances with the historical *editing* criteria.

Currently, the detection of spurious data in radar altimetry measurements relies on a legacy data editing method consisting in checking whether the value of some altimetric parameters is outside a validity domain defined by minimum and maximum thresholds. This historical editing method is described in the data user manuals and in the CALVAL reports of altimetric missions. It has been developed and used by the community of experts over the last 20 years.

Our study considers mainly clustering and classification techniques to assess the validity of 1 Hz SLA (Sea Level Anomalies) from 1 cycle of standard JASON-3 GDR data. **Unsupervised** and **supervised** learning techniques have been applied in order to evaluate the capability of such methods in altimetry.

Filtering, standardization, principal component analysis and segmentation are applied to select decisive parameters and to build reliable classifiers.

Finally, measurements validity is determined from their classification in specific groups. Confusion matrices, ROC curves and other performance indicators are produced for validation purpose in order to compare the current "editing" criterion to our results. The first conclusions of our work highlight a correct classification with unsupervised learning models as well as the excellent performances with supervised models.

1. Methodology

Case study

The perimeter of the study was to investigate the performance of **data mining techniques** to discriminate valid and invalid altimetric data from the point of view of the **SLA** (Sea Level Anomaly) over the oceans. Indeed, SLA is the parameter the most commonly used by the altimetric community. The case study was applied to the **Jason-3 mission**.

Approach

This study was carried out according to a typical **data mining project** (Fig. 1). First, an important effort (50-60%) was dedicated to **data selection and preparation**: the good execution of this step is critical for the success of the following steps. Then a **data segmentation phase**, based on unsupervised classification methods, was applied to analyse the statistical properties of the dataset. This was followed by an implementation of a subset of various **supervised methods** and the comparison of their performances according to diverse statistical scores. The fourth step is the model deployment in an operational mode, which was out of the scope of the study. The various steps are not executed in a linear way, but **many iterations** are needed to adjust each steps until the performances are stable. For the sake of clarity, these iterations are detailed.

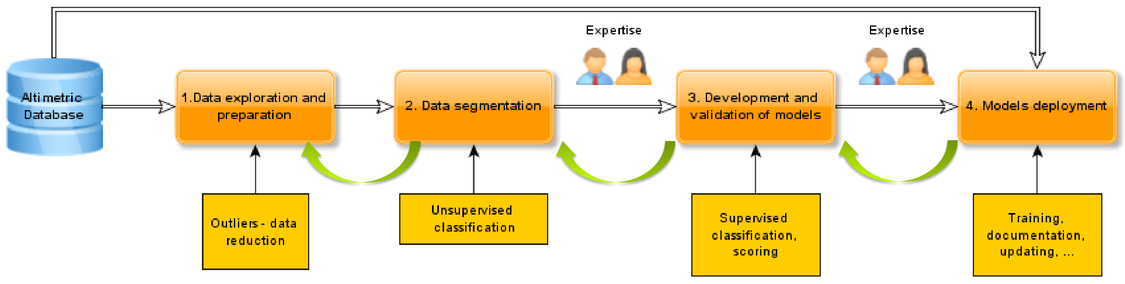


Fig. 1: Data mining methodology

2. Data selection and preparation

Dataset

A dataset was prepared in order to cover a representative subset of Jason-3 observations, covering the variety of **geophysical conditions**. A **"composite" J3 cycle of 254 tracks** was constructed by combining 21 tracks per 10-day repeat cycle sampled every 3 cycles over one year (12 cycles sampled). Doing so, we obtained a composite cycle with tracks regularly sampled both in space and time. Only data over open ocean were kept.

This dataset was visually **analysed by three expert engineers** (Fig 2). With the help of a graphical user interface, they selected individual 1 Hz data in SLA along track plots they suspected to be erroneous and invalid from an oceanographer's point of view. The result was a composite dataset with all 1 Hz data flagged "OK" or "NOK". For some tracks, the classification obtained by the expert was compared with two other data editing criteria : first, with the standard "**CALVAL editing criteria**" (described in altimetric product user manuals and in cycle CALVAL reports); second , with the so-called **20 Hz "iterative flag**" developed in other studies. Though the 20 Hz "iterative flag" is not directly comparable with the "CALVAL flag" and with the newly developed "expert flag", a qualitative comparison permitted to conclude that: 1) the new "expert flag" and the "iterative flag" identify the same data as "OK or "NOK", 2) the "CALVAL flag" is more conservative than the two other flags, i.e. it flags as "NOK" data that are obviously correct.

Data preparation

The dataset was split in a **learning dataset** and a **validation dataset**. The learning subset was used only for the construction of the classification models and the validation set was used to evaluate the performances. Then ANOVAL method was used to select the **10 most discriminating variables** to detect the correctness of the SLA.

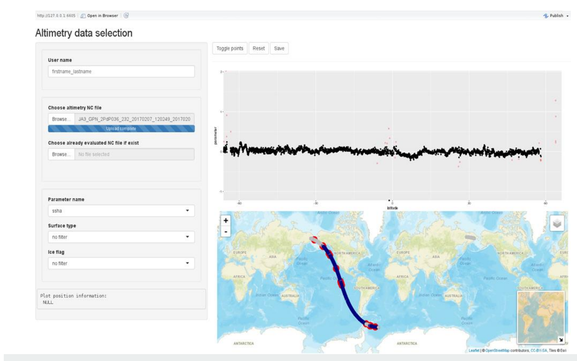


Fig. 2: Expert annotation of the composite cycle

3. Unsupervised approach

Method & Results

We implemented an unsupervised classification method so as to discriminate automatically "OK" from "NOK" SLA, **without a-priori knowledge**. It consist of a mix of Kmeans classification applied to altimetric data grouped by sigma ranges. For each 1 Hz altimetric data record, the ranking of each parameter with respect to the dispersion of the global learning subset (standard deviation – σ) was made: [0-1σ], [1σ-2σ], …[12σ-13σ]. Each 1 Hz altimetric data is represented by a 10-dimension vector containing the following variables: 'iono_corr_alt_ku', 'off_nadir_angle_wf_ku', 'sig0_rms_ku', 'sig0_numval_ku', 'ssha', 'swh_numval_ku', 'swh_rms_ku', 'ssha_carre', 'diff_ssha', 'norm_ssha'. Then each subset was examined by looking at the projection of the multi-dimensional point cloud in the plan consisting of **the two principal components** (after applying a Principal Component Analysis - PCA), Fig. 3. The unsupervised classification highlights clusters of data with a majority of "OK" or "NOK" data, but a significant proportion of data is not well separated (e.g. in the clusters highlighted in light grey and light brown), illustrating the limits of this method. Fig. 4 shows the location of a cluster where 100% of the dots are flagged correctly as "NOK" by the Kmeans method. These dots correspond to isolated coastal data and to data at the border of the Antarctica ice pack.

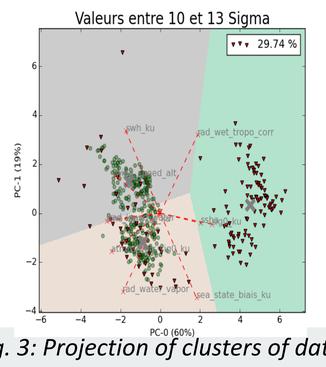


Fig. 3: Projection of clusters of data in the range [10σ-13σ] on the plan containing the 2 largest principal components

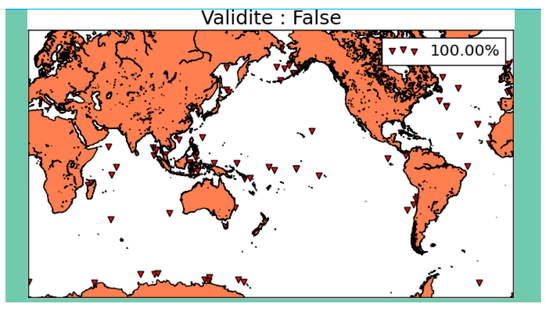


Fig. 4: Location of the clusters highlighted in green in Fig. 3.

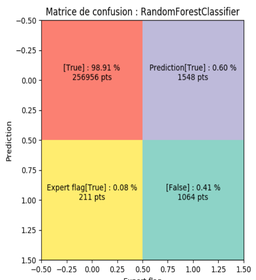
4. Supervised approach

Method & Results

In order to overcome the limitations of the unsupervised approach, a set of supervised methods was tested. The principle was to build models able to predict the class "OK" or "NOK" of the SLA variable. Five methods were tested: **Decision Trees**, **Random Forests**, **Logistic Regression**, Support Vector Machines (**SVM**), and **Naïve Bayes**. The models were trained with the learning dataset and the scores were computed with the validation dataset only. The metrics used to compare the performance of the methods were standard scoring protocols used to evaluate machine learning models: **confusion matrix**, **ROC** curves (Receiving Operating Characteristics), **Precision**, **Recall**, **F-score**.

Results

The best performance is obtained with the Random Forests, followed by the Decision Trees with **99.3%** of data correctly classified (Fig 5., diagonal terms of the confusion matrix), to be compared with the performance of the "CALVAL Flag" (98.3%). Fig. 6 permits to compare the ROC curves for the five methods.



labels	precision	recall	f_score	support
0.0	0.994	0.999	0.997	257167
1.0	0.835	0.407	0.547	2612

Fig. 5: Confusion Matrix and scores for Random Forests

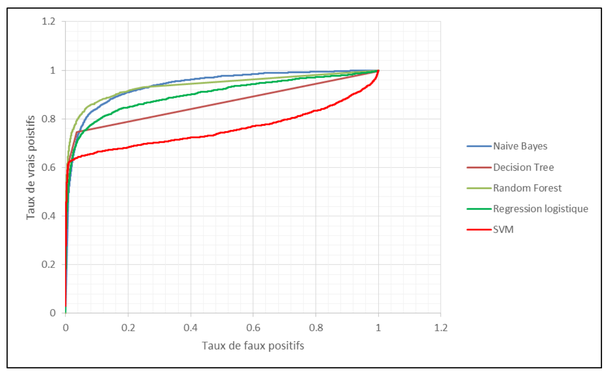


Fig. 6: Superposed ROC curves for the five tested methods

Conclusions & prospects

The study permitted to develop an end-to-end processing prototype, adaptable to further altimetric missions and new machine learning algorithms to be tested. A new **Jason-3 composite cycle** was annotated by altimetry experts and used as a reference dataset completely independent from the standard calval editing criteria. The **unsupervised methods** demonstrated limited performances, but studying them was important to better understand the statistical properties of the multi-dimensional dataset. Amongst the tested **supervised methods**, the best scores were obtained for **Decision Trees and Random Forests**.

In order to further explore the potential of the method, several actions are envisaged: test **new parametrizations** of the tested methods, test **new machine learning algorithms** not considered during the study, process a larger dataset to appreciate the performances in **operational conditions**, apply the methods to **high-resolution altimetric data**, i.e. 20/40 Hz (high rate data are expected to bring more discrimination in evaluating the correctness of altimetric data). The case study considered was focused on the sea level anomaly, but other cases could be taken into account to cover a wider range of applications, such as Hs, wind, backscatter coefficient… Finally, the new EO data should be considered: wide swath altimetry (**SWOT**), other types of microwave sensors (**CFOSAT, SKIM**…) and **optical sensors**.

Acknowledgements: this study was co-funded by the CNES Research & Technology program, grant ref. R-S17/OT-0005-047, R-S18/DU-0002-001.