Statistical learning approaches applied to the calculation of scaling factors for radioactive waste characterization

Biagio Zaffora, Jean-Pierre Chevalier, Francesco La Torre, Catherine Luccioni, Matteo Magistris, Gilbert Saporta

Statistical learning for radioactive waste characterization

Poster · May 2016

DOI: 10.13140/RG.2.1.3956.8240

1 author:

Biagio Zaffora
CERN
14 PUBLICATIONS **15** CITATIONS

SEE PROFILE

Statistical learning for radioactive waste characterization

Biagio Zaffora,[*a] Jean-P. Chevalier,[b] Francesco P. La Torre,[a] Catherine Luccioni,[b] Matteo Magistris,[a] Gilbert Saporta[b]

[a] *CERN, European Organization for Nuclear Research, 1211 Geneva 23, Switzerland*

[b] *CNAM, Conservatoire National des Arts et Métiers, 75003 Paris, 2 Rue Conté, France*

[*] *Corresponding author:* biagio.zaffora@cern.ch

Radiological characterization is needed to dispose of the radioactive waste produced in high energy particle accelerators. We applied statistical learning methods to predict the activity of Difficult-to-Measure radionuclides – which are low-energy X, α- and β-emitters –, to establish criteria for sorting radioactive waste and to quantify prediction errors.

Introduction

DTM Difficult-to-Measure nuclides cannot be easily quantified by non-destructive assay means. Their activity a_{DTM} is often correlated to the concentration of γ-emitters.

CF If a_{DTM} is correlated with the activity a_{KN} of a major γ-emitter - called Key Nuclide *KN* - we can estimate a_{DTM} using a Correlation Factor *CF*:

$$a_{DTM} = CF \times a_{KN}$$

We conducted an extensive numerical experiment to predict the behaviour of CFs.

Simulations

We generated $\sim 2.4 \times 10^6$ CERN activation scenarios. For each scenario:

- We extrapolated the Radionuclide Inventory ⇒ List of produced nuclides
- We identified the Key Nuclide
- We calculated the Correlation Factor for each pair DTM/KN

Statistical learning methods

- We used Decision Trees and Multiple Linear Regression to estimate average CFs and to find possible sorting criteria
- We used Bagging, Random Forests and k-fold Cross Validation to minimize variances and to quantify Prediction Errors

ISO Standard 16966. *Theoretical activation calculation method to evaluate the radioactivity of activated waste generated at nuclear reactors* (2013).

Input Space

The features X considered for the statistical models are:

- Beam Energy ⇒ 6 levels from 160 MeV (Linac 4) up to 7 TeV (LHC)
- Location Inside Tunnel ⇒ 7 levels
- Irradiation Time ⇒ Spaced grid from 0.25 up to 30 years
- Decay Time ⇒ Spaced grid from 1 up to 40 years

Multiple Linear Regression

- We studied the effects of the predictors X_i on CF using linear models:

$$CF = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \epsilon$$

- We used Best Subset Selection. Features are chosen using ⇒ Mallow's C_p coefficient, Bayesian Inference Criterions, Adjusted R^2 or Akaike Information Criterion
- We found that:
  - ⇒ Decay and Irradiation Times are the strongest predictors
  - ⇒ 5 levels of Beam Energy - with the exception of 160 MeV - can be grouped
  - ⇒ The feature Position inside the tunnel plays a minor role when predicting CFs

Prediction Error using k-fold Cross-Validation

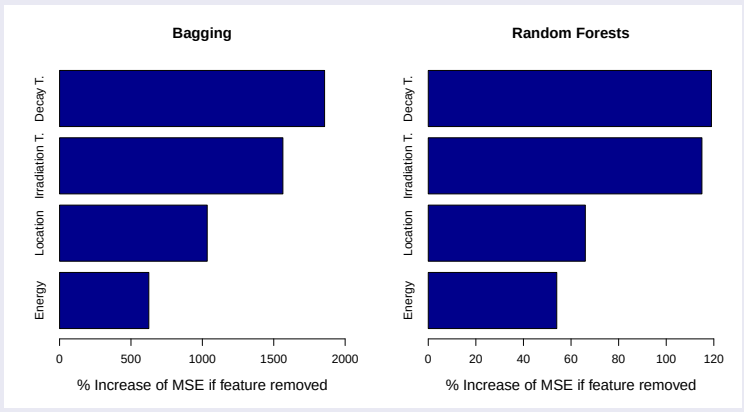
- We used k-fold Cross-Validation $CV_{(k)}$ to estimate Prediction Error
- Scenarios are randomly divided into k groups
- The model is fit on k-1 folds and MSE is calculated in the held-out fold
- The k-fold CV estimate is computed by averaging the calculated MSE:

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} MSE_i$$

Example: The prediction error $CV_{(10)}$ for CF(^{3}H/^{22}Na) in Aluminium 6060, when all predictors are included, is 0.26 ⇒ The average CF is affected by an error $e^{0.26} = 1.3$

C.M. Bishop, *Pattern recognition and machine learning*, Springer (2006).

Figure 2: Variable importance for CF(^{3}H/^{22}Na) in Aluminium 6060



The importance of a variable is estimated as the Percentage Increase of MSE if the variable is removed from the model

Regression Trees

- We used Regression Trees to identify groups of CFs using Binary Splitting
- The feature used to split minimizes the Residual Standard Errors
- Mean CFs are calculated at each node of the tree (red boxes in Figure 1)

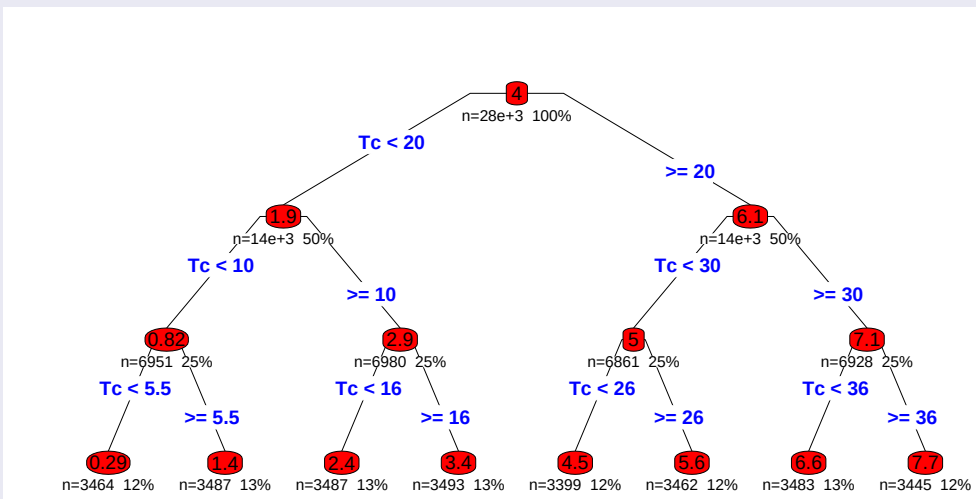
Log-transformation, Training and Test data sets, Prediction Error

- The distribution of calculated CFs is log-normal ⇒ The logarithm of the data is calculated to normalize the observations
- The n data is split into training data set ($m = n/2$) and test data set ($m = n/2$) to build the tree and estimate the error on a new prediction
- The Prediction Error is the square root of the Mean Squared Error of Residuals:

$$MSE = \frac{1}{m} \sum_{i=1}^{m} \left(y_i - \hat{f}(x_i) \right)^2$$

Example: The prediction error of the regression tree in Figure 1 is 0.62 ⇒ The average CF is affected by an error $e^{0.62} = 1.86$.

Figure 1: Regression Tree of CF(^{3}H/^{22}Na) in Aluminium 6060



- T_c ⇒ Decay Time
- n ⇒ Number of scenarios in training set
- 4 ⇒ Logarithm of the Correlation Factor

Variance Reduction using Bagging and Random Forests

Decision trees are often affected by high-variance ⇒ Different training data sets generate very differents outputs.

- We applied 2 known variance reduction techniques to reduce output variability
- We calculated the MSEs and we identified the method that minimizes the error

Bagging

- This method aggregates decision trees obtained by Bootstrapping observations from the training data set
- The prediction error is calculated on the left-out bootstrapped data

Example: The prediction error of the bagged tree of CF(^{3}H/^{22}Na) is 0.0062 ⇒ Bagging reduces the variance of 2 orders of magnitude!

Random Forests

- This technique is similar to bagging but it forces each split to consider only a subset of predictors ⇒ The trees are decorrelated

Example: The prediction error obtained applying random forests for CF(^{3}H/^{22}Na) is 0.61 ⇒ No major improvement from the regression tree.

T. Hastie, R. Tibshirani and J. Friedman. *The elements of statistical learning. Data Mining, Inference, and Prediction*, 2nd Ed., Springer (2009).

Conclusions

To characterize radioactive waste produced in high-energy particle accelerators we must estimate the list of radionuclides produced (α-, β- and γ-emitters) and their activities.

- We estimated the activity of Difficult-to-Measure nuclides using the Correlation Method
- We used Regression Trees and Linear Models to predict Correlation Factors
- We applied Bagging, Random Forests and k-fold Cross-Validation to estimate Prediction Errors
- We identified splitting variables at the nodes of the regression trees as potential sorting criteria for radioactive waste