Estimation of the Contaminant Risk Level of Petroleum Residues Applying FDA Techniques

*Abstract***—In the process of oil extraction, specifically in the refinement and industrialization of hydrocarbons, as is known, multiple wastes are highly polluting for the soil, water and air. In this work, the level of risk of these wastes in affected areas is estimated thanks to the application of statistical models in the field of functional data analysis; And these models have been implemented in a software that allows an early measurement and evaluation, through semiquantitative and quantitative methods, of the level of risk. This measurement is carried out by the human teams of PETROECUADOR in the places close to the affectation. It was used the laser-induced fluorescence technique (LIF) which consists of the spectroscopy that allows to study the structure of the molecules and the detection of the selected species. The data obtained using this technique are used to adjust the following models: Generalized Functional Linear Model (MLFG), which makes it possible to classify the spectrum generated from the application of LIF in two pollution levels: Low and High. A percentage of 99% of correctly classified spectra was obtained. Functional linear regression model with scalar response and functional explanatory variable with the aim of directly estimating the percentage of contamination level. With this model, it was able to explain the 99% (R ˆ 2) of the variability. With these results it is verified that the shape of the laser fluorescence spectrum is highly related to the gasoline content in the sample.**

*Index Terms***—Key Words and Phrases: Quality Control, Laser Induced Fluorescence (LIF), Generalized Linear Functional Model, Linear Regression, Classification.**

I. Introduction

The filtration of oil (or its derivatives), transport and diffusion-dispersion are processes whose study is of vital importance due to the great impact they have on human activity and the environment. The filtration of petroleum in soils causes a level of pollution that is a very complex problem to evaluate, this depends mainly on the following elements: soil type, porosity, hydraulic conductivity, and petroleum properties such as density and viscosity. For this reason, an important task is to determine the state of the system at all times, and as a priority at the initial moment, since it would allow the application of corrective measures in the ecosystem in a more efficient way. Oil is made up of a variety of compounds, some of which produce fluorescence when illuminated with ultraviolet light. The fluorescence of the petroleum depends to a great extent on its chemical composition (Celander, K., Freddricsson, B. Galle, S. and Svanberg, 1988). For this reason there are analytical techniques for the characterization of crude oil, in the parts that the intensity and life time of the fluorescence are related to the chemical composition and density (API) of the oil.

If we combine a source of ultraviolet laser light, a spectrometer and oil, we will have a system to detect the presence of petroleum, such as contaminated lands (O'Neill, RA, Buja-Bijunos, L., Rayner, DM, 1980). Laser light produces fluorescence when there is oil in the earth, which is detected using the spectrometer. As each variety of oil has a characteristic fluorescence spectrum, fluorescence techniques are often used for identification. The data obtained by this technique are used in the first instance to solve a problem of supervised classification and then to carry out a forecast of the level of contamination. Among the statistical techniques that are used to solve a classification problem are: discriminant analysis, logistic regression and cluster analysis, depending on the objective. For example, in R.H. Anderson, D.B. Farrar, S.R. Thoms, (2009) determines the contamination of anthropogenic metals in the soil using the technique of discriminant analysis with clustered chemical concentrations. In the case of the prediction process, Lopez Claudio (2014) applies a linear regression in order to predict the air pollution of carbon dioxide produced by Hawaii's Mauna Loa volcano, in this case time is the independent variable and pollution is the dependent variable. The statistical techniques mentioned are traditional and multivariate techniques, however, in recent times technological change has been able to measure data in a faster and more precise way, and thanks to this evolution, it is possible to work with the functional form of the data. In this work, statistical techniques of functional data analysis (Functional data Analyzes - FDA -) are used. Specifically, a Generalized Functional Linear Model is applied to solve the classification problem and a Linear Regression Model with scalar response and functional explanatory variable to estimate the level of contamination. One of the advantages of using FDA is reducing the influence of noise or observation errors. (Ramsay, J. & Silverman, B. W, 2005).

Classification of functional data is one of the major branches of the FDA (Functional Data Analysis), there are two types of classification that are: supervised and unsupervised. In the case of unsupervised classification, its objective will be to make groups as homogeneous as possible, and at the same time the most distinct among them (Noguerales, 2010); The most common technique is clustering and the method for performing such technique is k-means. For the supervised classification there are different classifiers that will help to classify the base between them: Linear Discriminant, k-NN (nearest neighbor method), Kernel, PLS (generalized linear models), Generalized linear models (Noguerales, 2010). A recent application of a Generalized Linear Functional Model can be found in Miguel

Sample	Level	Total	Sample	Level	Total
GE1	0,3	2	**GE8**	9,1	2
GE2	0,4	2	**GA15**	16,67	2
GE3	0,5	2	**GE9**	16,7	2
GA1	0,5	2	**GE10**	37,5	2
GA3	1,48	2	**GA19**	37,5	2
GE4	1,5	2	**GE11**	50	2
GE5	2,4	2	**GA23**	50	2
GA5	2,44	2	**GE12**	75	2
GE6	3,8	2	**GE13**	83,3	2
GA8	3,85	2	**GA26**	83,33	2
GE7	6,1	2	**GE14**	100	2
GA13	6,1	2	**GA30**	100	2
GA14	9,09	2			

TABLE I
SAMPLES MADE FOR MFLG ADJUSTMENT AND VALIDATION

Flores, Guido Saltos and Sergio Castillo-Paéz (2016), where the types of cancer are classified by DNA information. In this study, the generalized functional linear model was used to classify the contamination level of the hydrocarbon residues, with a variable binary response. This model assumes that the content of the solid element remains constant and that it is indeformable. It is important to note that the model has already been integrated in a software that interacts with the laser spectrometer, built by the team of engineers of the project developed by the National Polytechnic School. On the other hand, the functional linear regression model with scalar response and functional explanatory variable, the model is used as predictor. This class of models has been applied to regional analysis associations as an alternative to standard multiple regression models (Luo, Zhu, Xiong (2012)).

For the development of the models, 25 tests (with two replicates) were carried out, which are differentiated by the percentage (level) of gasoline in the sample (see Table 1). For this work, if a spectrum has a gas percentage less than or equal to 10% it is classified in the low pollution group; Otherwise it is classified in the high pollution group. The model allows to consider any other level of gasoline to discriminate between spectra corresponding to samples with a low or high contamination.

In the second instance, from the sample described in Table 1, the contamination level (percentage) is estimated using the Functional Linear Regression model. Unlike the first model, functional linear regression allows explaining and quantifying the relationship between the scalar response (pollution level) and the functional explanatory variable (spectrum).

II. MATERIALS AND METHODS

The methodology used in this study with functional data is that described in February-Bande and Fuente (2012), which consists of the following stages:

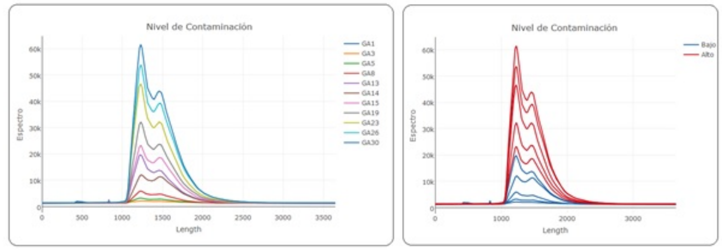

Fig. 1. Spectra by level of contamination

1. Explore and describe the functional data set highlighting its most important characteristics.
2. Explain and model:
2.1 The relationship between a dependent variable and an independent variable using regression models
2.2 Solve the problem of Supervised or Non-Supervised Classification of a set of data regarding some characteristic.
3. Contrast, validation and prediction.

Let T = [a, b] ⊂R. It is generally assumed that there are elements of

$L^2 = \{X : T\text{-})R, such\ that \int'_T \mathrm{E}X^2 < \infty\}(1)$

A. FDA exploratory analysis

The first step is the exploratory analysis of the data to make their characteristics known and know exactly how to manipulate them. The methods of representation are: decrementation and the choice of a reduced basis of functions. One way to represent the functional data is in a nonparametric way. And in most cases this is the best representation. X is defined as the functional variable of interest, spectrum generated through the LIF technique, which takes values in a normalized (or semi-normalized) space F, and is considered as functional data to the results of the 25 tests represented as The set {x_1, x_2, ..., x_n} that come from n functional variables X_1, X_2, ..., X_n identically distributed as X. The functional data are discretized in a total of 3'648 points that are in the range [176,39, 890,62]. These are represanta by the set of points t_j. In Figure 1 we can see in blue color the spectra of low level and in red color the spectros of high level. It can be seen from the figure that the spectra tend to have the same shape however the spectra of High level have a greater amplitude than the low ones and that this is increasing in relation to the percentage of gasoline that is in the sample. This varies from 16.67% to 100% (greater than 10%); On the other hand, the amplitude of the spectra of the low level samples are decreasing according to the percentage of gasoline in the sample. A spectrum is considered with a low level of pollution if the percentage of gasoline is less than 10%, that is from 3% to 9.1%. For the representation of the functional data, a B-spline base was used using the fda.usc package of the statistical software R (February-Bande and Fuente, 2012).

B. Generalized Functional Linear Model (MFLG)

Once the spectra are represented to functional data, a Generalized Linear Functional Model (MFLG) is fitted to estimate the probability that it belongs to one of the two groups. For the adjustment and implementation of the model, the 'fregre.glm' function of the fda.usc package of the statistical software R was used. The MFLG is also known in the literature as Functional Logistic Regression (FLR). The model explains the relationship between Y (binary response) and a functional covariate X (t) with representation based on X (t) and β (t). Π_i is the probability of occurrence of the event Y_i = 1, which in this case corresponds to a high contamination, conditioned to the covariate X_i (t), which is expressed as follows:

$Y_i = \pi_i + \varepsilon_i, where i = 1, ..., n. (2)$

$$\pi_i = P\left[Y = \frac{1}{x_i(t)} : t \in T\right] = \frac{\exp\left\{\int_T X_i(t)\beta(t)dt\right\}}{1 + \exp\left\{\int_T X_i(t)\beta(t)dt\right\}}, \quad \text{donde} \quad i = 1, \dots, n. \qquad (3)$$

Where ε_i are independent errors with zero mean. It is defined as a functional covariate the spectrum denoted by: X = X (t), and as a scalar (binary) response variable the type of pollution denoted by Y (0 = Low pollution, 1 = High pollution). In this case, since the MFLG works with a binary response variable, this model provides a classification rule for the type of contamination (Bayes rule).

C. Linear regression with variable scalar response and functional explanatory variable

For this model the objective will be to understand how a response variable Y being this scalar is related to a vector of variables X∈R ˆ p. Therefore, the regression model is defined as follows (Ramírez John, 2014),

$Y = (X, \beta) + E(4)$

In our case study we will analyze the following model

$$Y = \langle X, \beta \rangle + \epsilon = \frac{1}{\sqrt{T}} \int_T X(t)\beta(t)dt + \epsilon \qquad (5)$$

Where, (. , .) we denote the usual internal product defined in $\mathscr{L}2$ and ε is the random error with mean zero and variance σ2. Specifically, we will perform a non-parametric functional regression model. An alternative to model (5) is

$y_t = r(X_t(t)) +_t (6)$

Where, the unknown real soft function is estimated using the Kernel estimate.

For the regression model it is considered as a scalar response variable the percentage of water presented by gasoline and our functional explanatory variable is the spectrum. To obtain the corresponding estimates, the 'fregre.np' function of the fda.usc package has been used.

D. Validation of the models

For the validation of the statistical models, two samples are used. A sample of training and validation. For each test taken, two replicates of the spectrum were obtained. These are used as follows: one of the replicas is used for the training of the model and the other replica for the validation of the model. In this work, to carry out the validation of the model a confusion matrix is used, with the following structure:

High Low / High True False negatives /Low False positives True negatives)

Where, true positives and true negatives correspond to correctly classified spectra in high and low contamination

	High	Low
High	6	0
Low	0	5

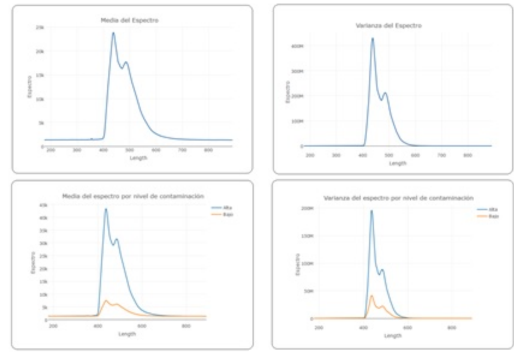

Fig. 2. Functional descriptive measures by level of contamination

respectively, while false negatives and false positives are misclassified spectra by our model; Using the data of the spectra we obtain:

As seen in the previous matrix, the model has an efficiency of 100% On the other hand, for the validity of the functional linear regression model with scalar response and functional explanatory variable, the coefficient of determination R ˆ 2 was used, where a value of 99% was obtained. This implies that the model correctly explains the variability of the data in that percentage.

RESULTS AND DISCUSSION

Prior to the modeling, an exploratory analysis of the data in the functional field has been carried out, ie the mean and functional variance for all data, as well as for the High and Low groups, have been estimated. The way to estimate these descriptive measures is in González-Manteiga and Vieu, 2007.

From Figure 2, it can be seen that there is a clear distinction between the defined groups. This result facilitates and confirms the use of a functional supervised classification model.

In Figure 3, a graph of probabilities resulting from applying the GLFM model is presented. Where it is appreciated that the spectra with a percentage less than 10% are classified as low level; While the rest of the spectra are classified as high level. Therefore, the model correctly classifies 100% of the spectra in each group (Low and High).

Figure 4 shows the level of contamination that the oil sample will present. It is important to mention that several tests were done with different methods trying to find the one that fit the best. It is also observed that the two curves, both the estimation with the model and the actual pollution, coincide in almost every point.

III. CONCLUSIONS

As mentioned in the introduction section, each sample has two replicates, one of which is used for model estimation and the other for its validation. In the case of the sample for the estimation we have that the percentage of correctly classified

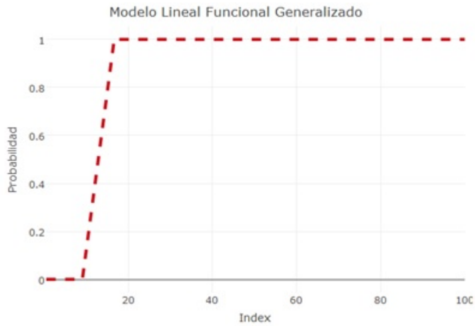

Fig. 3. Estimated probabilities

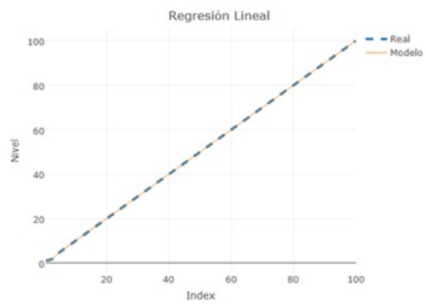

Fig. 4. Actual pollution level vs. estimated by the functional regression model.

spectra in each group (Low and High) is 100%, while for the validation sample the percentage of correctly classified spectra is 99 %. It has been verified that the shape of the laser fluorescence spectrum is highly related to the gasoline content of the sample. Therefore, due to its functional nature, the application of supervised FDA classification techniques provides a reliable solution for the identification of a high or low risk of contamination in potentially affected areas. When applying the functional regression model, we have managed to explain the 99% (R ˆ 2) of the variability, in addition to reach this result has been tested with several models. The corresponding validation tests of the model were also performed, which were statistically significant.

IV. ACKNOWLEDGMENTS

The authors are grateful for the funding provided by the National Polytechnic School for the implementation of the project PII-DM-002-2016: 'Analysis of functional data in statistical quality control'

REFERENCES

[1] Celander, K., Freddricsson, B. Galle, S. y Svanberg. (1988). Investigation of Laser-Induced Fluorescence with application to remote sensing of environmental parameters, Goteborg Institute of Physics Reports GIPR-149.

[2] González-Manteiga, W. y Vieu, P. (2007). Statistics for functional data. Computational Statistics and Data Analysis, 51, 4788-4792.

[3] Li, J., Cuesta-Albertos, J. A., & Liu, R. Y. (2012). DD-classifier: Nonparametric classification procedure based on DD-plot. Journal of the American Statistical Association. Vol. 107, 737-753.

[4] López Miranda Claudio y Cesar Augusto Romero Ramos, (2014). Propuesta de proyecto de estadística: un modelo de regresión lineal simple para pronosticar la concentración de co2 del volcán Mauna Loa. EPISTEMUS, 17:63-69.

[5] [Febrero-Bande, M. and Oviedo de la Fuente, M. (2012). Statistical computing in functional data analysis: The R package fda.usc. Journal of Statistical Software, 51(4):1-28.

[6] Miguel Flores, Guido Saltos and Sergio Castillo-Paéz, (2016), Setting a generalized functional liner model (GFLM) for classification of different types of cancer, Latin American Journal Computing, 3 (2):41-48.

[7] O'Neill, R.A., Buja-Bijunos, L., Rayner, D.M. (1980). Field Performance of laser flourosensor for detection of oil spills. Appl. Opt. 19,863.

[8] Ramsay, J. O. and Silverman, B. W.2005. "Functional Data Analysis", 2nd ed., Springer-Verlag, New York, Pp. 147-325.

[9] Ramírez John. Matemática. (2014), Regresión funcional mediante bases obtenidas por descomposición espectral del operador covarianza, Matemáticas, 12 (2):15-27.

[10] R.H. Anderson, D.B. Farrar, S.R. Thoms, (2009). Application of discriminant analysis with clustered data to determine anthropogenic metals contamination. Elsevier, 408:50-56.

[11] Muñoz Dania, Silva Francisco, Hernández Noslen, Bustamante Talavera. (2014). Functional Data Analysis as an Alternative for the Automatic Biometric Image Recognition: Iris Application. Computación y Sistemas, 18 (1):111-121