Chapter 3
Functional linear regression models for scalar responses on remote sensing data: an application to Oceanography

Nihan Acar-Denizli, Pedro Delicado, Gülay Başarır and Isabel Caballero

Abstract Remote Sensing (RS) data obtained from satellites are a type of spectral data which consist of reflectance values recorded at different wavelengths. This type of data can be considered as a functional data due to the continous structure of the spectrum. The aim of this study is to propose Functional Linear Regression Models (FLRMs) to analyze the turbidity in the coastal zone of Guadalquivir estuary from satellite data. With this aim different types of FLRMs for scalar response have been used to predict the amount of Total Suspended Solids (TSS) on RS data and their results have been compared.

3.1 Introduction

Functional Data Analysis (FDA) concerns with the data sets measured on a continuum such as a dense time interval, space or a spectrum. The data gathered from Remote Sensing (RS) sensors via transmission of electromagnetic energy is also a kind of spectral data. RS data are collected from the earth's surface in terms of reflectance values recorded at different number of wavelengths. They inform us in a fast and economical way about the environment. Therefore, they are used in many fields such as land-use mapping, agriculture, forestry and oceanography to make predictions [2, 3, 5, 11]. In oceanography, RS data are used to estimate ocean characteristic

Nihan Acar-Denizli (✉)
Mimar Sinan Güzel Sanatlar Üniversitesi, Istanbul, Turkey, e-mail: nihan.acar@msgsu.edu.tr

Pedro Delicado
Universitat Politècnica de Catalunya, Barcelona, Spain, e-mail: pedro.delicado@upc.edu

Gülay Başarır
Mimar Sinan Güzel Sanatlar Üniversitesi, Istanbul, Turkey, e-mail: gulay.basarir@msgsu.edu.tr

Isabel Caballero
ICMAN-CSIC, Cadiz, Spain e-mail: isabel.caballero@icman.csic.es

G. Aneiros et al. (eds.), *Functional Statistics and Related Fields*,
Contributions to Statistics, DOI 10.1007/978-3-319-55846-2_3

parameters such as Sea Surface Temperature (SST), Chlorophyll-a content (Chl-a) and Total Suspended Solids (TSS) [2, 3]. Recently, FDA gain importance in analyzing remote sensing sensor data sets [1, 4]. Although, there are many applications of multivariate analysis techniques on RS data in oceanography [2, 3, 10, 11], there are few studies that use FDA approach in this field [9].

The importance of this study is to propose FLRMs alternative to classical statistical methods to predict TSS parameter from RS curves at different time periods. In previous studies, mostly regression models with a combination of different band values or the band values which are most correlated to TSS measurements have been used to predict TSS parameter [2, 3]. FLRMs allows us to use the information recorded at all the bands rather than selecting single band or taking combination of the bands. In this study, the Remote sensing reflectance (Rrs) values, recorded in a spectrum that consists of eight different bands have been considered as functional predictors and TSS content, that are measured from collected in-situ samples have been taken, as scalar response vector. In order to determine the best prediction model several FLRMs for scalar responses have been constructed and their performances have been compared with the performance of classical statistical methods used in the literature. A 10 year data set has been constituted by matching the in-situ measurements with the satellite data recorded between the years 2002-2011. The work exhibits an approach of how to conduct analysis for processing and interpreting large-scale volume of heterogeneous data to improve the present knowledge as an essential piece of the future Big Earth Observation Data monitoring systems.

3.2 Methods

The general form of a functional scalar response model can be expressed by

$$Y = \int_T \chi(t)\beta(t)dt + \varepsilon, \tag{3.1}$$

where Y indicates the scalar response vector, ε is the error term, $\chi(t)$ and $\beta(t)$ define respectively the functional predictor and the parameter function that are defined on a continuous interval T.

To solve this problem, different techniques based on basis functions, eigenfunctions or nonparametric smoothing have been proposed to assess an interpretable estimate of the parameter function [6, 5, 22]. In this study we will focus on two different approaches. The first approach is to use B-Spline basis expansions to define the functional predictor and the model parameter function. The latter approach is based on dimension reduction method functional principal components analysis so that it is named as Functional Principal Components Regression (FPCR). The idea of FPCR is to predict scalar response vector $\mathbf{Y}$ on the functional predictors that are expanded in terms of the eigenfunctions of the empirical covariance operator which form an orthonormal basis in $L^2(T)$ [12, 6].

The main problem here is to determine the number of basis functions or components that will be used to expand data. In this study, Cross Validation (CV) criterion is preferred to choose the optimal number of basis functions. The estimates of the models are found by minimizing the Sum of Square Errors (SSE) as in the classical linear regression problem.

To compare the predictive performance of the mentioned models, an Adjusted version of Mean Error of Prediction (AMEP) based on Leave-One-Out Cross Validation (LOOCV) has been defined by the equation (3.2).

$$\text{AMEP} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_{-i})^2}. \tag{3.2}$$

The term $\bar{y}_{-i}$ in this equation indicates the mean after removing the ith observation from the data set.

3.3 Data

The data set consists of two parts. The in-situ data set is composed of TSS concentrations measured from the collected samples and the satellite data set is composed of the Remote sensing reflectance (Rrs) values recorded by MEdium Resolution Imaging Spectrometer (MERIS), one of the main instruments on board the European Space Agency (ESA)'s Envisat platform between the years 2002-2011. The data set has been constituted by matching the filtered satellite data with the in-situ considering the coordinate and the time that the sample has been collected. The coordinates have been matched considering the exact pixels that the sample is collected and the time difference between in-situ and satellite data has been constrained up to 1.5 hours.

3.3.1 In-Situ Data

The in-situ data consist of the records of TSS concentration which are obtained from the samples collected by the station of Junta de Andalucía and by the cruises of Reserva and Fluctuaciones in the Guadalquivir estuary. The surface samples taken into analysis were collected with a rosette sampler (5 m below water surface) with a distance from coast from 1km to 25 km offshore.

The samples were collected during different time periods. The sampling carried out by Junta de Andalucía covers the period between April 2008 and May 2011 where the samples of Reserva and Fluctuaciones were collected within the periods July 2002 - September 2004 and May 2005 - May 2007 respectively. Each sample is collected by one of the campaigns from a determined coordinate. The coordinate of the station Junta de Andalucía was fixed with the lattitude 36.78° N and longitude 6.37° W where the coordinates of the stations Reserva and Fluctuaciones were

chosen according to the campaign planning. The amount of TSS concentration in each sample has been measured according to the protocols mentioned in [3].

3.3.2 Satellite Data

The study area corresponds to the coastal region of the Gulf of Cadiz in the southwest coast of the Iberian Peninsula ($35.5^\circ - 37.5^\circ$ N latitude and $1^\circ - 10^\circ$ W longitude). The satellite data included within the Region Of Interest (ROI) was downloaded from the Ocean Colour Website (http://oceancolor.gsfc.nasa.gov) in hdf format. SeaDAS image analysis software (SeaWifs Data Analysis System, version 6, http://seadas.gsfc.nasa.gov/) and the interface VMware Workstation 12 Player (https://www.vmware.com/) were used to convert data from hdf format to ascii format. The RS data set consists of Level-2 Remote Sensing Reflectance (Rrs) (sr^{-1}) recorded at eight different wavelengths (413 nm, 443 nm, 490 nm, 510 nm, 560 nm, 620 nm, 665 nm, 681 nm) with 300 m full spatial resolution between the years 2002-2011. The data has been passed through a quality control process corresponding to the L2 flags given in [3] to remove the suspicious and low-quality data points. This filtering process is done by using MATLAB 7.12.0-R2011a software. Considering that the resolution of images is 300 m, the data set consists of 740×3330 pixel images which is equivalent to have 2464200 element vectors for each wavelength.

Statistical validation of satellite-derived products is an essential issue to verify the accuracy provided by the sensor. In this work, data match ups were made by matching the coordinates of Rrs values with the coordinates of the field measurements. Careful consideration of scales is critical when comparing remotely-sensed data with in situ observations, particularly because of the large spatio-temporal heterogeneity of estuarine and coastal water properties influencing those measurements [8]. In this sense, time difference between satellite overpasses and in situ sampling was reduced by a filter of < 1.5 hours from acquisition, thus preventing temporal biases to further evaluate the results of each data set; notwithstanding less number of match-up for validation purposes are available. If we use a wider time window of 4 or 5 hours, we get a major number of match-ups but more variability is encountered with the inconvenience of greater discrepancies between in-situ and RS observations.

3.4 Results

As a result of matching in a time window of 1.5 hour, totally 31 observations are obtained. 5 of them have been excluded from the analysis due to the measurement errors, 6 of them have been removed due to missing values at some wavelengths and 2 of them have not been included into the analysis due to their outlyingness. The analysis have been conducted on 18 observations: 8 observations from Junta

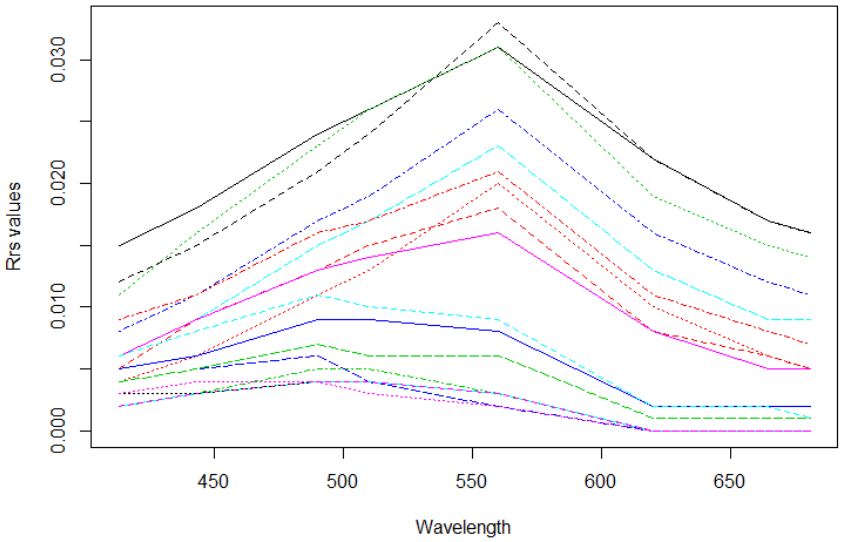

Fig. 3.1: The Rrs curves.

de Andalucía station, 7 observations from Reserva cruise and 3 observations from Fluctuaciones cruise.

The Rrs values at eight different wavelengths have been converted to a functional data object as given in Figure 3.1. Firstly, FLRM with B-Spline approach has been used to predict the TSS parameter. In this case, both the functional predictor and the functional parameter estimate has been smoothed by using B-Spline basis functions. The optimal number of basis functions has been chosen as 5 by the CV criterion [6].

The number of components that will be included in FPCR has been chosen based on their variability. The first 5 components for which the cumulative variance exceeds 95% have been taken into analysis.

The results of FLRMs have been compared with the results of classical approaches in the literature. As offered by [3], exponential regression models have been constructed between TSS and the Rrs values at the band most correlated with the response. The most correlated bands were found as 665 nm (r=0.65) and 681 nm (r=0.66). Therefore, two different single band exponential regression models have been used.

All the functional and exponential regression models were found significant ($p < 0.05$). The coefficient of determination (R^2), Standard Error (Std. Err.) and AMEP values based on LOOCV of the related models are given in Table 3.1.

Regarding R^2 values, FLRMs explain higher amount of variability of the response comparing to exponential regression models.

Table 3.1: R^2, Std. Err. and AMEP values of the models

Models	R^2	Std. Err.	AMEP
FLRM with B-Spline Basis	0.78	0.32	**0.42**
FPCR with the first 5 components	0.82	0.30	0.71
Exponential Regression with 665 nm	0.42	0.46	0.71
Exponential Regression with 681 nm	0.43	0.45	0.70

Among all the models, FLRM using 5 number of B-Spline basis functions has predicted TSS parameter better since the AMEP value of this model is the lowest. Although the AMEP value of FPCR with 5 components is not that low, we see that the predictive performance of the model is as good as the exponential regression models.

3.5 Conclusions

In this study, the performance of FLRMs to predict TSS on RS data has been compared to single band exponential regression models. The data set has been constituted under spatio-temporal filtering by matching exact coordinates in the time window of 1.5 hour difference. Although, the limited number of wavelengths and observations, it is seen that the FLRMs on RS data predict TSS content better than the classical exponential regression models offered in the literature. The best prediction model has been found as FLRM with B-Spline basis approach using 5 basis functions. To conclude, FLRMs estimate the TSS content in Guadalquivir estuary better than other classical approaches that have been used earlier in RS community.

There are several ways to explore in order to improve the prediction ability of the considered models. First, Figure 3.1 suggests the existence of two clusters of curves (well differentiated by Rrs values at wavelengths larger than or equal to 550 nm). These clusters may correspond to clear or low turbid to turbid water conditions in each scene. Then a dummy variable indicating if a day is considered *clear* or *turbid* could be included in the regression models. More observations with different concentrations of TSS will be required to have reliable estimations. Second, the studied observations present (as many environmental data) spatial and temporal dependence (observations from the same boat trip have been taken in close times and close sites). To take into account these dependence in the regression models could lead to more accurate predictions.

Acknowledgements This research was partially supported by the Spanish Ministry of Economy and Competitiveness, and European Regional Development Fund grant MTM2013-43992-R.

References

[1] Besse, P.C., Cardot, H., Faivre, R., Goulard, M.: Statistical modelling of functional data. Appl. Stoch. Model. Bus. **21**(2), 165–173 (2005)

[2] Caballero, I., Morris, Edward P., Ruiz, J., Navarro, G.: Assessment of suspended solids in the Guadalquivir estuary using new DEIMOS-1 medium spatial resolution imagery. Remote Sens. Environ. **146**, 148–158 (2014)

[3] Caballero, I., Morris, Edward P., Ruiz, J., Navarro, G.: The influence of the Guadalquivir River on the spatio-temporal variability of suspended solids and chlorophyll in the Eastern Gulf of Cadiz. Mediterr. Mar. Sci. **15**(4), 721–738 (2014)

[4] Cardot, H., Faivre, R., Goulard, M.: Functional approaches for predicting land use with the temporal evolution of coarse resolution remote sensing data. J. Appl. Stat. **30**(10), 1185–1199 (2003)

[5] Faivre, R., Fischer, A.: Predicting crop reflectances using satellite data observing mixed pixels. J. Agric. Biol. Envir. S. **2**(1), 87–107 (1997)

[6] Febrero-Bande, M., Galeano, P., González-Manteiga, W.: Functional Principal Component Regression and Functional Partial Least-squares Regression: An Overview and a Comparative Study. Int. Stat. Rev. (2015) doi:10.1111/insr.12116

[7] Ferraty, F., Vieu, P.: Nonparametric functional data analysis: theory and practice. Springer, USA (2006)

[8] Fettweis, M.P., Nechad B.: Evaluation of in situ and remote sensing sampling methods for SPM concentrations, Belgian continental shelf (southern North Sea). Ocean Dynam. **61**(2), 157–171 (2011)

[9] Gong, M., Miller, C., Scott, E.: Functional PCA for remotely sensed lake surface water temperature data. Procedia Environ. Sci. **6**, 127–130 (2015)

[10] Nechad, B., Ruddick, KG., Park, Y.: Calibration and validation of a generic multisensor algorithm for mapping of total suspended matter in turbid waters. Remote Sens. Environ. **114**(4), 854–866 (2010)

[11] Nezlin, N., DiGiacomo, Paul M.: Satellite ocean color observations of stormwater runoff plumes along the San Pedro Shelf (southern California) during 1997–2003. Cont. Shelf Res. **25**(14), 1692–1711 (2005)

[12] Preda, C., Saporta, G.: PLS regression on a stochastic process. Comput. Stat. Data An. **48**(4), 149–158 (2005)

[13] Ramsay, J.O., Silverman, B.W.: Functional Data Analysis. Springer, USA (2005)