New analysis methods to push the boundaries of diagnostic techniques in the environmental sciences

2016 JINST 11 C04019

(http://iopscience.iop.org/1748-0221/11/04/C04019)

*J*inst

PUBLISHED BY IOP PUBLISHING FOR SISSA MEDIALAB

RECEIVED: *April 19, 2016*
ACCEPTED: *April 24, 2016*
PUBLISHED: *April 29, 2016*

4TH INTERNATIONAL CONFERENCE FRONTIERS IN DIAGNOSTICS TECHNOLOGIES
30 MARCH 2016 TO 1 APRIL 2016
ROME, ITALY

New analysis methods to push the boundaries of diagnostic techniques in the environmental sciences

M. Lungaroni,[a,1] A. Murari,[b] E. Peluso,[a] M. Gelfusa,[a] A. Malizia,[a] J. Vega,[c] S. Talebzadeh[a] and P. Gaudio[a]

[a] *Associazione EURATOM-ENEA - University of Rome "Tor Vergata", Via del Politecnico 1, 00133 Roma, Italy*

[b] *Consorzio RFX (CNR, ENEA, INFN, Università di Padova, Acciaierie Venete SpA), Corso Stati Uniti 4, 35127 Padova, Italy*

[c] *Asociación EURATOM-CIEMAT para Fusión, Avenida Complutense 40, 28040, Madrid, Spain*

E-mail: michele.lungaroni@uniroma2.it

ABSTRACT: In the last years, new and more sophisticated measurements have been at the basis of the major progress in various disciplines related to the environment, such as remote sensing and thermonuclear fusion. To maximize the effectiveness of the measurements, new data analysis techniques are required. First data processing tasks, such as filtering and fitting, are of primary importance, since they can have a strong influence on the rest of the analysis. Even if Support Vector Regression is a method devised and refined at the end of the 90s, a systematic comparison with more traditional non parametric regression methods has never been reported. In this paper, a series of systematic tests is described, which indicates how SVR is a very competitive method of non-parametric regression that can usefully complement and often outperform more consolidated approaches. The performance of Support Vector Regression as a method of filtering is investigated first, comparing it with the most popular alternative techniques. Then Support Vector Regression is applied to the problem of non-parametric regression to analyse Lidar surveys for the environments measurement of particulate matter due to wildfires. The proposed approach has given very positive results and provides new perspectives to the interpretation of the data.

KEYWORDS: Analysis and statistical methods; Data processing methods; Pattern recognition, cluster finding, calibration and fitting methods

[1]Corresponding author.

doi:10.1088/1748-0221/11/04/C04019

2016 JINST 11 C04019

Contents

1 Introduction

In the last years, the need to reduce particulate emissions has grown substantially, due to both air quality issues and the effects on global and regional radiative forcing and therefore climate change. As a consequence, it has become increasingly important on the one hand to better monitor the environment and, on the other hand, to develop different and non-polluting sources of energy. With regard to the assessment of air quality, the Lidar-Dial techniques are widely recognized as a cost-effective alternative to monitor large regions of the atmosphere. They have been successfully deployed to detect Particulate Matter (PM) and pollutants, emitted by various sources in industrial and city centres and by wild fires in rural areas. With reference to non-polluting energy sources, the research in Nuclear Fusion remains a very active field of activity; to improve the performances of present day reactors, diagnosing plasma instabilities is a very important element. To maximize the effectiveness of the measurements, new data analysis techniques are required. In this paper, the performance of Support Vector Regression (SVR) as a method of filtering is investigated first, comparing it with the most popular alternative techniques. A series of systematic numerical tests indicate that Support Vector Regression provide particularly robust results. Then the signal filtered with SVR are analysed using a nonlinear fitting procedure to obtain information about the properties of the backscattering coefficient. The innovative tools developed allow determining the evolution of the backscattering coefficient versus distance for both the case of concentrated and widespread smoke, providing new perspectives to the application of the technique.

2016 JINST 11 C04019

2 Description of the signals to be analysed

Figure 1. Telescope and laser.

In this paper, different types of Lidar signals will be analysed. The common denominator of these signals is that they are time series, affected by significant levels of noise and therefore quite difficult to interpret. They also present sudden or widespread peaks which correspond to events that have to be identified, if possible in automatic ways.

2.1 The Lidar system

Lidar measurements have become well established laser based techniques for remote sensing of the atmosphere [1]. They are used to probe almost any altitude in the most different conditions, from forests to urban areas. One of the most interesting applications consists of environment surveying of particulate (see [2–5] and [6]). The measurements described in the paper have been performed with the mobile Lidar unit of Industrial Engineering Department, University of Rome “Tor Vergata” [7]. The system consists of an easily transportable compact Lidar system. The transmitter is a Nd:YAG laser that can operate at three wavelengths: 1064, 532 end 355 nm. For this analysis is been used signals by the 1064 nm wavelength because it is better suited to the identification of particulate air pollution.

The signals analysed in this paper have been collected during an extensive experimental campaign, which has been carried out in Calabria, in the south of Italy as shown in figure 2.

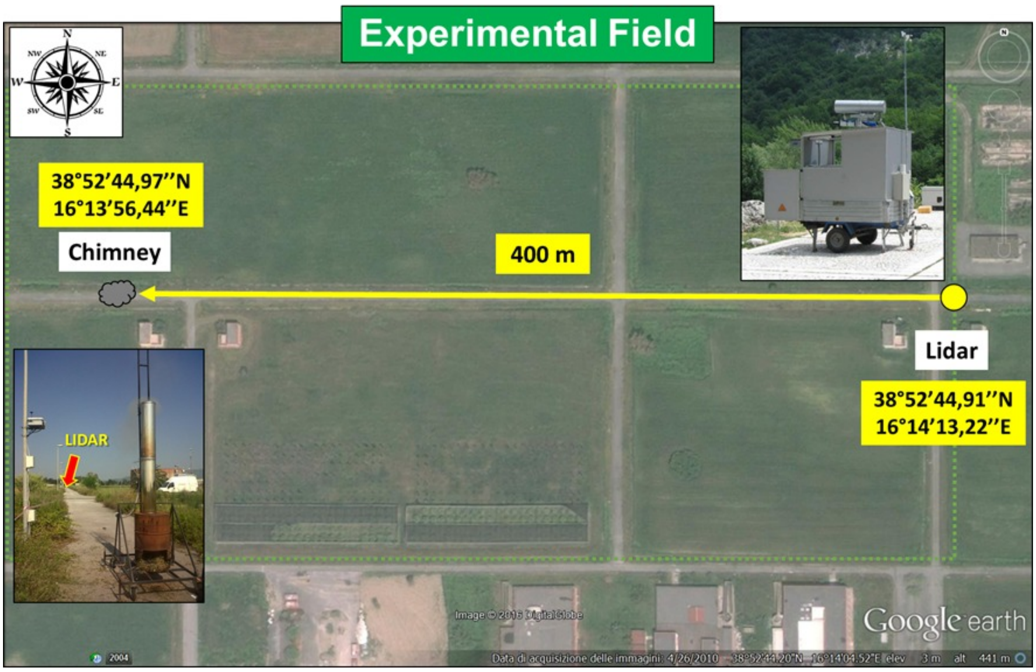

Figure 2. Experimental set up for the measurement campaign in Lamezia Terme (CZ), Calabria, Italy.

2016 JINST 11 C04019

The laser is anchored at the receiver system, a Newtonian telescope as shown in figure 1, and the detector chosen is a Hamamatsu's photomultiplier tube (PMT), R3235 model. These technologies have become relatively standard and therefore they can be procured at reasonable costs. The main characteristics of the mobile unit are reported in table 1.

Table 1. Parameters of Nd:YAG Lidar system [14].

Transmitter:	
Laser	Q-switch Nd:YAG
Pulse time width	8 ns
Divergence angle	5 mrad
Pulse Frequency	10 Hz
Receiver:	
Telescope type	Newtonian
Nominal focal length	1030 mm
Primary mirror diameter	210 mm
Detector	Photomultiplier (PMT)
Photocathode sensibility	0.2 mA/W
Response time	~ 30 ns

The entire apparatus is controlled by a software package, developed by University of Rome "Tor Vergata", written in Labview and Matlab, explicitly developed for this application [8]. The laser activation and the wavelength selection, together with the rotation of the telescope and data acquisition, is controlled by a Labview series of routines. The signal processing algorithms and the visualization of the results have been implemented using Matlab.

The signal processing routines calculate the distance of the fire from the station and also show the fire topographic coordinates.

2.2 Features of widespread smoke and concentrated smoke in Lidar signals

The Lidar technique has been successfully applied to the detection of the smoke plume emitted by wild fires, allowing the reliable survey of large areas, and this because wild fires have become a very serious problem in various parts of the world. The main operational approach envisages the continuous monitoring of the area to be surveyed with a suitable laser. When a significant peak in the backscattered signal is detected, an alarm is triggered. The traditional applications of Lidar systems to atmospheric physics therefore rely on the capability of properly detecting the backscattered peaks of radiation.

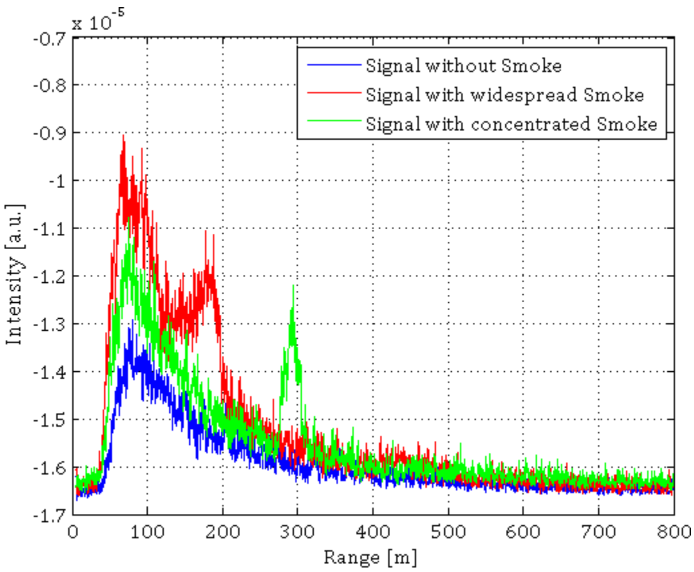

Figure 3. Raw signal by Lidar system, in blue the clear signal without smoke; in red the signal with widespread smoke, in particular between 70 and 200 m; in green the signal with concentrated plume of smoke near 300 m.

More recently, the Lidar technique has been shown to have the potential to provide useful measurements also of widespread smoke, which can be the consequence of strong wind dispersion or non-concentrated sources [9] and [6]. Typical examples of backscattered signals for the alternatives of clear atmosphere, strong smoke plume and widespread smoke are shown in figure 3.

2016 JINST 11 C04019

2.3 Simulating the Lidar backscattered signals

The Lidar raw signals present a decreasing exponential tract, the first increasing part of the signals not having any physical meaning, because it is determined by the intersection between the field of view of the telescope and the scattered laser light cone. So from the first maximum (around 70 m), the real signals decrease exponentially with the distance due to the absorption of the atmosphere, which is considered constant in the following treatment (therefore a constant K_2 is assumed). Since the detected signal consists of backscattered power, a quadratic decrease with respect to the distance is to be included in the treatment. Equation (2.1) is therefore the mathematical model to fit the experimental signals:

$$P\,(R) = \frac{K_1\,(R)}{R^2} exp\,(-K_2 R) \tag{2.1}$$

Where K_1 e K_2 are the parameters of the model and R is the range. In particular K_1includes the effect of the coefficient of backscattering β [3] geometrical and spectral form factor that usually are assume as constant value. In this paper we try to express K_1 as function of distance introducing a new approach to the analysis of the experimental data. Indeed the backscatter coefficient, in presence of an external agent, such as for example smoke, varies substantially. The result, as can be seen from figure 3, is that, when the laser beam interacts with something different from clear air, the backscattering power signal presents a concentrated or a widespread peak, depending on the concentration and on the position of the scattering agent [10]. In the rest of the paper, new data processing techniques are introduced to obtain reliably the spatial evolution of the backscattering coefficient, starting from the raw experimental data.

3 Techniques for Data Processing

This section will discuss the techniques used to process the data so that we can extract from the signals the necessary information. This is achieved by applying the SVR technique for the filtering of the signals. Then a non-linear fitting procedure allows deriving the main quantity of interest, the spatial evolution of the backscattering coefficient.

3.1 Support Vector Regression

Support Vector Regression is an alternative approach to non-parametric regression. Being based on substantially different principles as the most common alternatives, Support Vector Regression presents significant potential advantages for signal processing, which have not been fully explored so far. Similarly to classification problems, a non-linear model is usually required to adequately model data. In the same manner as the non-linear SVC approach, a non-linear mapping can be used to map the data into a high dimensional feature space where linear regression is performed. The kernel approach is again employed to address the curse of dimensionality. The non-linear SVR solution, using an ε-insensitive loss function, is given by

$$\max_{\alpha,\alpha^*} W\,(\alpha,\alpha^*) = \max_{\alpha,\alpha^*} \sum_{i=1}^{l} \alpha_i^*(y_i - \epsilon) - \alpha_i\,(y_i + \epsilon) - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(x_i, x_j) \tag{3.1}$$

with constraints of the equations (3.2) and (3.3):

$$0 \le \alpha\alpha^* \le C, i = 1, \dots, l \tag{3.2}$$

$$\sum_{i=1}^{l} \left(\alpha_i - \alpha_i^*\right) = 0 \tag{3.3}$$

Solving the previous equation with constraints determines the Lagrange multipliers, α_i, α_i^*, and the regression function is given by,

$$f(x) = \sum_{SVs} \left(\bar{\alpha}_i - \bar{\alpha}_i^*\right) K(x_i, x) + \bar{b} \tag{3.4}$$

where

$$\langle w, x \rangle = \sum_{i=1}^{l} \left(\alpha_i - \alpha_i^*\right) K(x_i, x_j) \tag{3.5}$$

$$\bar{b} = -\frac{1}{2} \sum_{i=1}^{l} \left(\alpha_i - \alpha_i^*\right) \left(K\left(x_i, x_j\right) + K\left(x_i, x_j\right)\right) \tag{3.6}$$

As for the case of Support Vector Classification the equality constraint may be dropped if the Kernel contains a bias term, b being accommodated within the Kernel function, and the regression function is given by,

$$f(x) = \sum_{i=1}^{l} \left(\bar{\alpha}_i - \bar{\alpha}_i^*\right) K(x_i, x) \tag{3.7}$$

The optimisation criteria for the other loss functions are similarly obtained by replacing the dot product with a kernel function. The ε-insensitive loss function is attractive because unlike the quadratic and Huber cost functions, where all the data points will be support vectors, the SV solution can be sparse.

In order to investigate the potential of the SVR regression for the analysis of time series, a systematic comparison with traditional non-parametric regression techniques has been undertaken first. The methods used for comparison are: Moving average, Lowess, Loess, Rlowess: robust version of Lowess, Rloess: robust version of Lowess and Savitzky-Golay.

The moving average is obtained by calculating a series of averages of different subsets of the full data set. Given a time series and a fixed subset size, the first element of the moving average is calculated by taking the average of the initial fixed subset of points of the series. Then the subset is changed by a process of "shifting forward"; that is, excluding the first number of the series and including the next number. This operation generates a new subset of numbers, which is averaged. This process is repeated over the entire series.

The acronyms "Lowess" and "Loess" are derived from the term "Locally weighted scatter plot smooth", as both methods are based on locally weighted linear regression to smooth the data. The linear regression is performed over a limited number of points called the span. The smoothing process is therefore local because each smoothed value is determined only by neighbouring data points, the ones within the span. The process is weighted because a regression weight function is used to fit the data points contained within the span. The data point to be smoothed has the largest weight and the most influence on the fit. The more distant the points from the one to be fitted the lower their weight and points outside the span are given zero weight and do not influence on the fit. The two approaches are similar but differ in the model used for the regression: Lowess implements a linear polynomial, while Loess implements a quadratic polynomial.

Rlowess and Rloess are robust versions of Lowess and Loess, to reduce the sensitivity to outliers. Robustness is achieved by an appropriate choice of the weight function. These robust methods include an additional calculation of robust weights, based on MAD, which is resistant to outliers.

The Savitzky-Golay filter can be considered a generalisation of the Loess approach allowing to choose higher order polynomials for the fit. In the applications presented in this paper the degree of the polynomial is always 6 or higher.

Table 2. Root Mean Square Error for the two different plume signal of figure 3: widespread and concentrated.

Method	RMSE widespread plume	RMSE concentrated plume
SVR	0.035931	0.040395
Moving	0.037002	0.042335
Lowess	0.038525	0.045942
Loess	0.036643	0.042956
S-Golay	0.038116	0.041668
Rlowess	0.036705	0.042080
Rloess	0.037445	0.044489

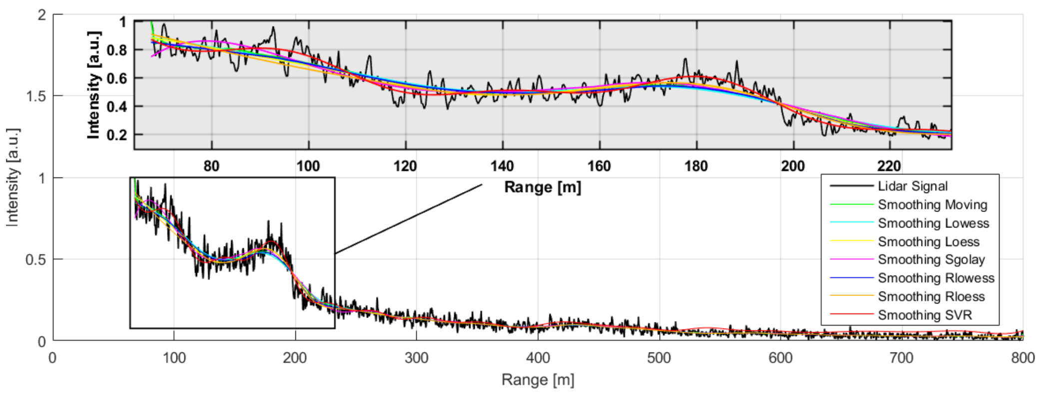

Figure 4. In black the Lidar experimental signal in case of **widespread smoke** is reported; in red the signal smoothed with SVR and in: green, cyan, yellow, magenta, blue and orange the results of the smoothing with the other methods: Moving, Lowess, Loess, S-Golay, Rlowess and Rloess respectively.

In table 2 the Root Mean Square Errors for the various smoothing methods are reported. It can be noted, as show in [11] that SVR reduces the RMSE between Lidar signal and the fit. The results of the various smoothing methods of Lidar experimental signals are shown graphically in figure 4 and figure 5, again for the two experimental curves of figure 3. In figure 4 the case of widespread smoke is reported and, as can be seen more clearly in the zoomed part on top, the improvement of the SVR smoothing is appreciable. The same effect can be seen in figure 5 for a case of concentrated smoke.

3.2 Smoothing and non-linear fitting of Lidar signals

As mentioned in section 2.3, the experimental Lidar signal requires some pre-processing before being fitted. First, the growing part of the signal has to be excluded; then some form of smoothing is required and in this paper this has been achieved with the SVR approach. In general, before smoothing it is advantageous to normalize the signal to its maximum value as shown in figure 4 and figure 5.

2016 JINST 11 C04019

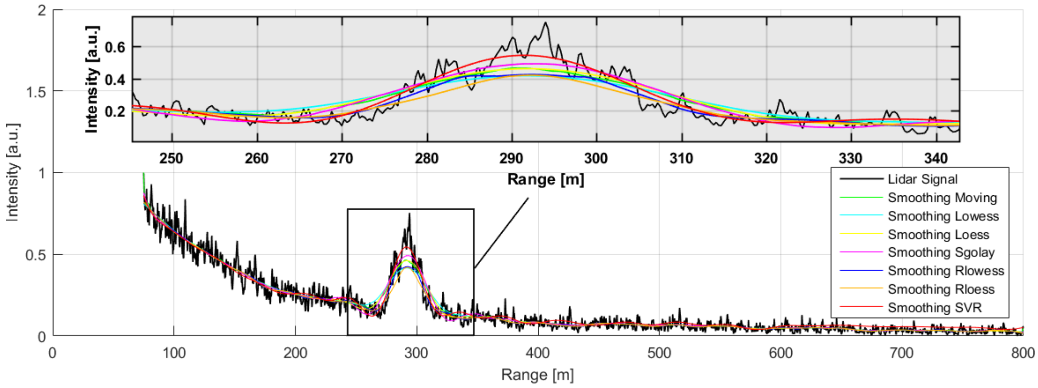

Figure 5. In black the Lidar experimental signal in case of **concentrated plume** is reported; in red the signal smoothed with SVR and in: green, cyan, yellow, magenta, blue and orange the results of the smoothing with the other methods: Moving, Lowess, Loess, S-Golay, Rlowess and Rloess respectively.

After smoothing the signal with the SVR we can re-multiply by the normalization factor to recover the dimensional result. This dimensional smoothed signal is the input to the fitting routine. The fitting routine used a Matlab function that solve the nonlinear least-squares problems:

$$\min_x \|f(x)\|_2^2 = \min_x \left(f_1(x)^2 + f_2(x)^2 + \ldots + f_n(x)^2\right) \tag{3.8}$$

Where into $f(x)$ there are the equation model with the parameters that want to fit, K_1 and K_2. By inserting an appropriate initial condition it is possible to detect, through the minimization of the problem, the best parameters of the nonlinear fit.

The model given by equation (2.1) is the equation to fit to the data. To perform this step a Matlab tool has been developed by the Research Group of University of Rome "Tor Vergata". First of all, the program searches the peaks in the smoothed signal and keeps track of their location, height and width. It then performs the non-linear fit of the smoothed signal with the proposed model, equation (2.1), to obtain the parameters. The algorithm goes to convergence when it finds a $K_1(R)$ which provides a good fit of the smoothed backscattered signal. In this way K_1 as function of the distance is derived which is equivalent to obtaining a backscattered coefficient versus radial position.

4 Tests of the fitting procedure with synthetic but realistic signals

This session will present the analysis to get the K_1 as function of distance from synthetic signals very similar to the experimental data of figure 3 after smoothing with SVR.

4.1 Concentrated smoke signals

This section presents the analysis results conducted on a test signal with two peaks concentrated at distances of 200 and 650 meters. As you can see from the first graph of figure 6, the model equation (2.1) can be fitted to the data quite well. With regard to the two fitting parameters K_2 is constant and equal to 4.44E-3, while K_1 is shown in the bottom graph of figure 6. As can be seen from the figure 6, the value of K_1 for the first peak is lower than the second peak located at 650 meters. This is due to the fact that the acquired signal decreases with the square of the distance and the presence of a small peak in the backscattered light at a very long distance is indicative of a strong variation of the backscatter coefficient at that distance.

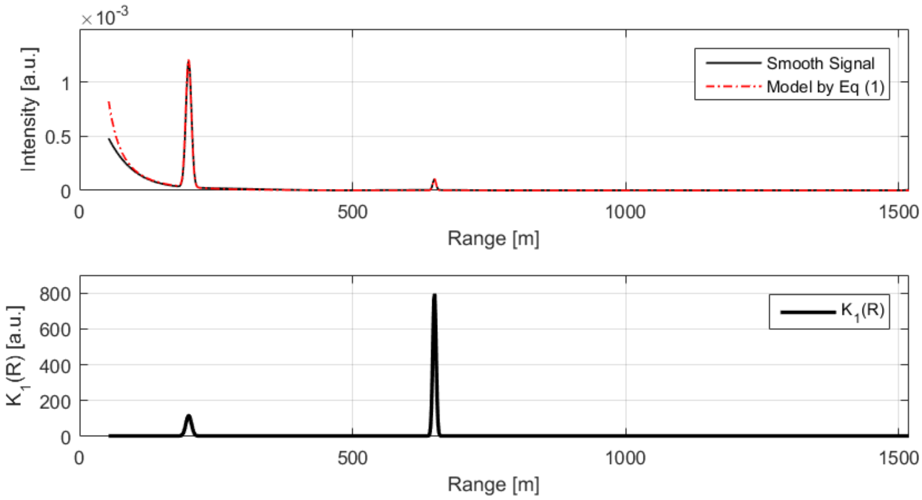

Figure 6. Top graph: in black the smoothed signal with two concentrated peaks and in red the fit with equation (2.1). The bottom graph reports the amplitude of K_1 versus distance.

4.2 Widespread smoke signals

In this case two widespread peaks, at the distance of 400 and 800 meters have been added to the decaying backscattered signal of clear atmosphere. Figure 7 shows again the developed algorithm can properly fit equation (2.1) to the data. With regard to the K parameters K_2 is constant and equal to 2.02E-3, while K_1 is shown in the lower graph of figure 7. Also in this case the distance plays a significant role requiring a much more significant change in the backscattering coefficient for the second peak compared to the first peak localized 400 meters closer.

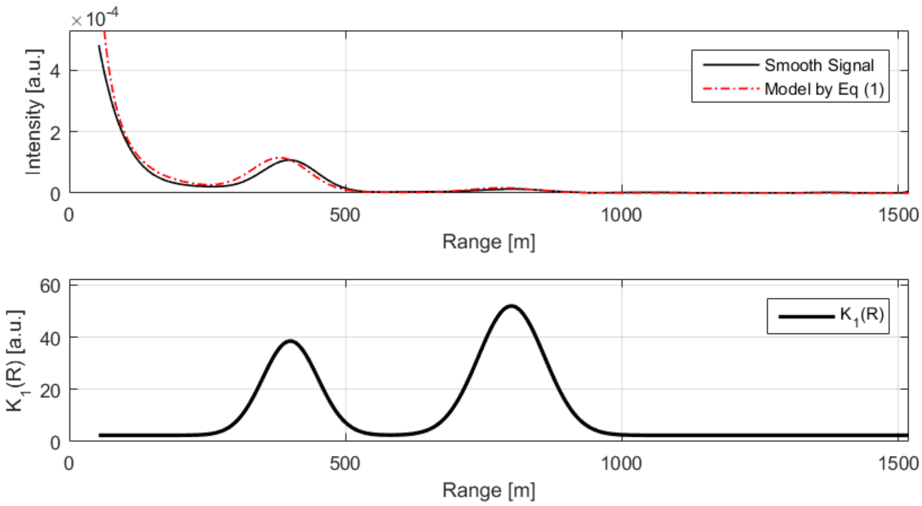

Figure 7. Top graph: in black the smoothed signal with two wide spread peaks and in red the fit with equation (2.1). The bottom graph reports the amplitude of K_1 versus distance.

2016 JINST 11 C04019

5 Conclusions

In this paper a new signal processing approach has been proposed to analyse the backscattered signal of Lidar systems for the monitoring of the environment. The method proposed consists of smoothing the signal with SVR before applying a specific non-linear fitting routine. The model of equation (2.1) allows deriving a profile of the backscattering coefficient, without any assumption about the behaviour of the atmosphere. The first results are very encouraging, a shown from the analysis of signals from both concentrated smoke plumes and wide spread smoke. It is also worth mentioning that the approach, and particularly the smoothing step, is already being considered for implementation by other communities such as Nuclear Fusion [12] and [13].

References

[1] F. Andreucci and M. Arbolino, *A study on forest fire automatic detection system, 2-smoke plume detection performance*, *Nuovo Cim.* **16** (1993) 51.

[2] F. Andreucci and M. Arbolino, *A study on forest fire automatic detection system*, *Nuovo Cim.* **16** (1993) 35.

[3] C. Bellecci et al., *Application of a CO2 dial system for infrared detection of forest fire and reduction of false alarm*, *Appl. Phys.* **B 87** (2007) 373.

[4] C. Bellecci, M. Francucci, P. Gaudio, M. Gelfusa, S. Martellucci and M. Richetta, *Early detection of small forest fire by Dial technique*, *Proc. SPIE* **5976** (2005) 59760C.

[5] P. Gaudio et al., *New frontiers of Forest Fire Protection: A portable Laser System (FfED)*, *WSEAS* **9** (2013) 195.

[6] M. Gelfusa et al., *First attempts at measuring widespread smoke with a mobile Lidar system*, in proceedings of *17th Italian Conference on Photonics Technologies, Fotonica AEIT* (2015).

[7] P. Gaudio et al., *Automatic localization of backscattering events due to particulate in urban areas*, *Proc. SPIE* **9244** (2014) 924413.

[8] C. Bellecci et al., *Reduction of false alarms in forest fire surveillance using water vapour concentration measurements*, *Opt. Laser Technol.* **41** (2009) 374.

[9] M. Gelfusa et al., *UMEL: A new regression tool to identify measurement peaks in LIDAR/DIAL systems for environmental physics applications*, *Rev. Sci. Instrum.* **85** (2014) 063112.

[10] A.B. Utkin, A.V. Lavrov, L. Costa, F. Simões and R. Vilar, *Detection of small forest fire by lidar*, *Appl. Phys.* **B 74** (2002) 77.

[11] M. Gelfusa et al., *Advanced signal processing based on support vector regression for LIDAR applications*, *Proc. SPIE* **9643** (2015) 96430E.

[12] J. Vega, A. Murari, S. González and JET-EFDAContributors., *A universal support vector machines based method for automatic event location in waveforms and video-movies: applications to massive nuclear fusion databases.*, *Rev. Sci. Instrum.* **81** (2010) 023505.

[13] A. Murari et al., *Extensive statistical analysis of ELMs on JET with a carbon wall*, *Plasma Phys. Contr. F.* **56** (2014) 114007.

[14] C. Bellecci et al., *Evolution study of smoke backscattering coefficients in a cell by means*, *Proc. SPIE* **6745** (2007) 67451S.

2016 JINST 11 C04019