A Source Separation Technique for Processing of Thermometric Data From Fiber-Optic DTS Measurements for Water Leakage Identification in Dikes

Amir Ali Khan, Valeriu Vrabie, Jerome I. Mars, Alexandre Girard, Guy d'Urso

A Source Separation Technique for Processing of Thermometric Data From Fiber-Optic DTS Measurements for Water Leakage Identification in Dikes

Amir A. Khan, Valeriu Vrabie, Jérôme I. Mars, Alexandre Girard, and Guy D'Urso

***Abstract*—Distributed temperature sensors (DTSs) show real advantages over conventional temperature sensing technology such as low cost for long-range measurement, durability, stability, insensitivity to external perturbations, etc. They are particularly interesting for long-term health assessment of civil engineering structures such as dikes. In this paper, we address the problem of identification of leakage in dikes based on real thermometric data recorded by DTS. Formulating this task as a source separation problem, we propose a methodology based on Principal Component Analysis (PCA) and Independent Component Analysis (ICA). As the first PCA estimated source extracts an energetic subspace, other PCA sources allow to access the leakages. The energy of a leakage being very low compared to the entire data, a temporal windowing approach guarantees the presence of the leakages on these other PCA sources. However, on these sources, the leakages are not well separated from other factors like drains. An ICA processing, providing independent sources, is thus proposed to achieve better identification of the leakages. The study of different preprocessing steps such as normalization, spatial gradient, and transposition allows to propose a final scheme that represents a first step towards the automation of the leakage identification problem.**

***Index Terms*—Dikes, distributed temperature sensors (DTSs), independent component analysis (ICA), leakage identification, principal component analysis (PCA), source separation, thermometric data.**

I. Introduction

FIBER-OPTIC sensors have long been employed in diverse domains for various applications such as monitoring of civil engineering structures, fault detection and metering in electrical engineering, chemical sensing for environment and pollution control, parameter sensing in oil and gas industry, and early fire detection systems [1]–[7]. In civil engineering, many aging infrastructures may become vulnerable in terms of their stability due to various phenomena like internal erosion, adverse climatic conditions, and other natural phenomena. Amongst these structures, attention is nowadays turned towards dikes for whom it is imperative to detect the possible anomalies such as leakages (significant flow of water) in advance so as to take preventive measures accordingly.

The early conventional methods for detecting anomalies in dikes were based on visual inspections and scheduled investigations performed at the site. The measurements of different parameters such as flow rates, pressure, and deformation form some of the contemporary conventional methods. Recently, some nonconventional methods such as the self-potential method, the resistivity method and the temperature based methods have been used for detection of anomalies [8]–[12]. The self-potential and resistivity methods have been used for internal erosion and leakage detection in many civil engineering structures like dams and dikes. Even if the aforementioned methods provide good solutions for detection of anomalies, the major constraint is that these methods are manual. Moreover, the acquisition setup is not economically viable. On the other hand, the thermometric methods present efficient solutions with the improved possibilities of temperature measurements through the use of fiber-optic temperature sensors. These fiber-optic sensors offer a multitude of advantages such as reduced weight and dimensions, strong immunity to electromagnetic interferences, environmental robustness, scale flexibility for small gauge, long gauge measurements and low cost, etc. [13]. In fact, what sets them apart is the use of low-cost telecommunications grade fiber, which provides them an ability to multiplex large number of sensors along a single fiber thus enhancing their commercial viability. With the limitations of the existing systems in terms of their dependence on tedious and error prone human-based monitoring systems, it is imperative to develop an automatic system which can alert of the possible leakages in advance, while minimizing the false alarms.

The thermometric data acquired at the site through fiber-optic sensors is not directly interpretable in terms of leakage identification and localization. It is, therefore, imperative to process this signal in a way that renders useful information concerning the anomalies. The idea behind the use of temperature signals is that a change of temperature between water of the canal and that of the ground is brought about by a significant flow of water through the structure due to leakages. However, this change of

Manuscript received July 19, 2007; revised October 31, 2007; accepted October 31, 2007. Published July 16, 2008 (projected). The associate editor coordinating the review of this paper and approving it for publication was Dr. Andrea Cusano.

A. A. Khan and J. I. Mars are with GIPSA-Laboratory, Department of Image and Signal Processing (DIS), INP de Grenoble, BP 46, 38402 Saint Martin d'Hères Cedex, France (e-mail: amir-ali.khan@gipsa-lab.inpg.fr; jerome.mars@gipsa-lab.inpg.fr).

V. Vrabie is with Centre de Recherche en STIC (CReSTIC), Université de Reims, 51687 Reims Cedex, France (e-mail: valeriu.vrabie@univ-reims.fr).

A. Girard and G. D'Urso are with Direction Etudes et Recherches, Electricité de France (EDF), 78401 Chatou Cedex, France (e-mail: alexandre.girard@edf.fr; guy.durso@edf.fr).

Digital Object Identifier 10.1109/JSEN.2008.926109

temperature can equally be brought about by other factors such as the seasonal variations, precipitations, existing structures (e.g., drains) etc. The problem of leakage identification can thus be seen as a problem of source separation with the sources being all aforementioned factors which can result in a possible change of temperature. The underlying idea is not only to separate out the response of the near surface, where the acquisitions are made, but also to efficiently identify the leakages in the presence of precipitations and other background effects. The source separation techniques have been successfully employed in domains as diverse as neural networks, biomedical engineering, telecommunications, econometrics, geophysics, image processing, audio signal separation, spatio-temporal data set analysis, etc. [14]–[16]. More recently, they have been employed to analyze fiber sensor signals for the measurement of food color and water monitoring [7]. The aim of this paper is to propose a scheme based on the advanced signal processing techniques so as to uncover the hidden information in the observed temperature signals recorded on an experimental site of Electricité de France (EDF). This experimental site, located in the south of France, near Oraison City, is dedicated to the study of leakage detection in dikes.

Amongst the three major sections of this paper, we start in Section II with a description of the system and the principle of acquisition which is based on distributed temperature sensors (DTSs) [17] using the Raman technique. The representation of the recorded signals and their characteristics are also discussed in this section. The theoretical formulation of the problem is presented in Section III using the advanced techniques of source separation, namely, principal component analysis (PCA) and independent component analysis (ICA). After a brief theoretical description of these techniques, a scheme is proposed to identify the information related to leakages. Section IV presents the application of the proposed scheme to a real data set using different preprocessing techniques like normalization, spatial gradient and transposition of data. A comprehensive discussion on the results, including the choice of various parameters like the number of PCA and ICA sources and the processing window size, also forms the subject of Section IV. This study leads to the selection of a particular scheme for leakage detection. The conclusion follows in Section V, deducing on the results of previous sections and thus elaborating the possibility of identification of leakages in dikes with the application of source separation techniques.

II. Data Set

The sensors used for the acquisition of temperature signals are DTSs based on optical fibers [1], [13], [17]. These sensors provide temperature observations over long distances with high spatial and temperature resolutions. Moreover, their capability to integrate a large number of passive optic sensors within a single low-cost telecommunications grade optical fiber significantly enhances their commercial viability. In most commercial DTSs, the acquisition principle of temperature profiles is based on Raman scattering using optical time-domain reflectometry (OTDR) techniques [18]. The basic setup of temperature sensing based on OTDR is comprised of a pulsed laser coupled to the optical fiber, the sensing element. The emitted photons interact with the molecules of the fiber material. Thermally influenced molecular vibrations in the fiber produce Raman scattering in the form of backscattering of some photons of light. Consequently, this backscattered light carries information about the fiber temperature and can thus be used to obtain the temperature distribution along the fiber [19]. The Raman backscattered light has two components: the Stokes and Anti-Stokes scattering. The principle of temperature sensing lies in the fact that the intensity ratio between Anti-Stokes and Stokes components, $R(T)$, is temperature dependent and can be described by

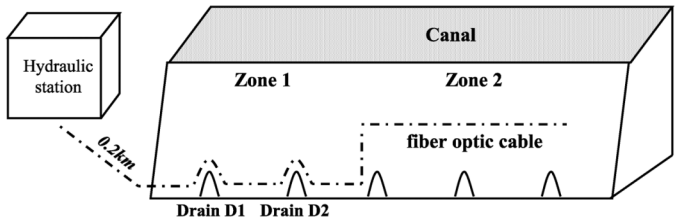

Fig. 1. Schematic representation of temperature monitoring system at site Oraison.

$$R(T) = \left(\frac{\nu_0 + \nu_k}{\nu_0 - \nu_k}\right)^4 \exp\left(-\frac{h\nu_k}{kT}\right) \tag{1}$$

with ν_0 the frequency of the input laser, ν_k the frequency shift of Raman scattering, h the Planck's constant, and k the Boltzmann's constant. Measuring the travel time of probe pulse and the ratio $R(T)$, at the fiber input, allows to obtain the temperature profile along the entire length of the fiber using one-to-one relationship between the spatial resolution and the traveling time.

For studying leakages, EDF proposed installation of a thermometric data monitoring system based on the above principle at an experimental test site located in the south of France (near Oraison City). The aim of this site is to extract the information pertaining to leakages (both natural and controlled) in the dike of a canal. A schematic representation of this data monitoring system is given in Fig. 1. A fiber-optic cable was installed in the abutment at the toe end of the canal so as to intercept the water leakage from the canal. The cable containing 4 optic fibers, of type multimode 50/125, is buried at the downstream toe of the canal at a depth of 1 m. These fibers take measurements in a loop along the entire 2.2 km length of the cable. Two distinct zones, (Zone 1, from approximately 0.2 to 1.25 km and Zone 2, from approximately 1.25 to 2.2 km), corresponding to two different elevation levels, will be exposed with varying intensities to direct sunlight. The cable also circumvents two drains, D1 and D2, situated at 0.561 and 0.858 km, respectively. The temperature data were recorded by a commercial device Sensornet, Sentinel DTS-MR, with a capability of covering up to 8 km range. The temperature resolution of this device is 0.01 °C with 1-m spatial resolution. The basic principle of leakage detection by DTS is that a leakage (significant flow of water) would result in a thermal anomaly and would thus be detected by the fiber-optic cable. The method utilized at the site is the *passive method*, which is a natural measure of temperature. The method takes its

TABLE I
CHARACTERIZATION OF THE STRUCTURES AND LEAKAGES FOR SITE ORAISON

	Leakages			Drains	
	L1	L2	L3	D1	D2
Location (km)	1.562	1.547	1.569	0.561	0.858
Time	10^{th} May(noon)	12^{th} May(noon)	12^{th} May(eve)	-	-
Flowrate (l/min)	5	1	1	-	-

name from the fact that in the absence of any anomaly, the measured temperature is driven by the phenomenon of conduction: the transfer of heat results from the interaction between the temperature of air and that of water present naturally in the ground. The occurrence of leakage results in flow of water which brings along additional heat by the phenomenon of advection [12]. Thus, when advection superposes conduction, thermometry can help in identification of leakages. The identification of leakage thus depends not only on the flow rate but also on the difference in the temperatures between that of ground (which, in turn, depends on the air temperature depending on the depth) and that of water.

The temperature data using DTS are obtained along the entire length of the fiber with a sampling distance of 1 m. In the discussions that follow, this acquisition along the entire length of the fiber will be called a profile. In addition, to monitor the temporal evolution of the anomalies, which in itself is important in any fault detection scheme, the profiles were acquired over three months (April, May, and June). An acquisition period of 2 h, i.e., one profile acquisition every 2 h, was selected to have an adequate resolution. The recorded data set is thus a two-dimensional temperature signal, $y(x,t)$, as a function of displacement along the fiber and time and can be written in matrix format as

$$\mathbf{Y} = \{y(x,t) | 1 \leq x \leq N_x, 1 \leq t \leq N_t\} \tag{2}$$

where N_x and N_t are the number of observation points (also called the temperature sensors) and the total acquisition time, respectively. To test the efficiency of the proposed technique to identify the leakages, three artificial leakages were introduced during different times in the month of May with different flow rates and positions. A description of these leakages along with the localization of some existing drains present in the path of the fiber-optic is given in Table I. In the next section, we are going to briefly discuss the source separation techniques that we have used for this data set along with a description of the proposed scheme.

III. SOURCE SEPARATION TECHNIQUES

A. Data Decomposition

The temperature data acquired by the fiber is affected by various factors like ground response (permeability, physical composition, etc.), natural phenomena (seasonal temperature variations, precipitation), anomalies (leakages), structures of the dike (drains), etc. A source signal as a function of displacement along the fiber is assigned for each factor. Moreover, some factors such as drains or leakages can be modeled by sparse sources, which means that these sources are non-Gaussian. A recorded data set can be considered as a mixture of these sources and we assume that this mixture is linear. Likewise, since the sources originate from noncorrelated phenomena, they are supposed to be independent of each other. The problem can thus be formulated as

$$\mathbf{Y}^T = \mathbf{M}\mathbf{F}^T \tag{3}$$

where $\mathbf{Y} \in \Re^{N_x \times N_t}$ represents the acquired data, $\mathbf{M} \in \Re^{N_t \times p}$ is the mixing matrix, $\mathbf{F} \in \Re^{N_x \times p}$ designates the matrix made up of the p independent sources mentioned above, i.e., a source represents a column of the matrix $\mathbf{F}$, and $.^T$ denotes the matrix transposition. The problem is thus to find out the matrices, $\mathbf{M}$ and $\mathbf{F}$, from the observation matrix, $\mathbf{Y}$, with the only hypothesis that the sources are independent. The identification of each of the above mentioned factors can thus be treated as a source separation problem and in our application the most important factor is leakage. The classical techniques commonly employed are principal component analysis (PCA) and independent component analysis (ICA). We explore briefly these two techniques.

1) Principal Component Analysis (PCA): PCA is widely used in signal processing, statistics, and neural computing with the goal to find out a space in which the desired inherent characteristics of data are represented in a space of the smallest possible dimension [20], [21]. When the principal components are calculated from the covariance matrix of the recorded data $\mathbf{Y}$, an efficient method for their calculation is the singular value decomposition (SVD) [21]:

$$\mathbf{Y}^T = \mathbf{U}_N \Sigma_N \mathbf{V}_N^T = \sum_{j=1}^{N} \sigma_j \mathbf{u}_j \mathbf{v}_j^T \tag{4}$$

where $N = \min(N_x, N_t)$, $\Sigma_N \in \Re^{N \times N}$ is a matrix containing on its diagonal the singular values $\sigma_j \geq 0$ arranged in a descending order and $\mathbf{U}_N \in \Re^{N_t \times N}$ and $\mathbf{V}_N \in \Re^{N_x \times N}$ are orthogonal matrices, containing N left and right singular vectors $\mathbf{u}_j \in \Re^{N_t}$ and $\mathbf{v}_j \in \Re^{N_x}$, respectively. The right singular vectors $\mathbf{v}_j$ are estimators of the sources defined by the above mentioned factors [22]. As these vectors are orthonormal by construction, the estimated sources are decorrelated and normalized. The decorrelation allows to extract Gaussian sources, which is not sufficient in our case because sources associated with some factors such as the drains or the leakages can be modeled by sparse and so non-Gaussian sources. However, it is possible to decompose the initial data into two complementary subspaces, namely, the signal subspace and the noise subspace defined as

$$\mathbf{Y}^T = \mathbf{Y}_{\text{sig}}^T + \mathbf{Y}_{\text{noise}}^T = \sum_{j=1}^{m} \sigma_j \mathbf{u}_j \mathbf{v}_j^T + \sum_{j=m+1}^{N} \sigma_j \mathbf{u}_j \mathbf{v}_j^T. \tag{5}$$

These subspaces are orthogonal, the first one being contained in a space of dimension m and the second one in a space of dimension $N - m$. The critical parameter is the choice of the number "m," the number of singular values retained to construct the signal subspace. This decision is usually based on the observation of the singular values, σ_j, by defining a threshold for keeping the most significant singular values [16]. As we saw above, PCA allows to extract orthogonal sources, but in real practice we cannot ensure that the orthogonality condition holds for the inherent sources. A more realistic technique not driven by the orthogonality condition is based on the independence of the sources and forms the subject of the next section.

2) Independent Component Analysis (ICA): ICA is a blind decomposition of a multichannel data set made up of unknown linear mixtures of unknown source signals based on the assumption that the sources are mutually statistically independent [14], [15], [22]–[24]. The statistical independence of sources means that the cross-cumulants of any order vanish. In general, the third-order cumulants are negligible and are discarded and we use the fourth-order cumulants which are considered suitable for instantaneous mixtures [25]. Considering the noise-free model for ICA

$$\mathbf{Z}^T = \mathbf{A}\mathbf{S}^T \tag{6}$$

where $\mathbf{Z} \in \Re^{N_x \times N_t}$ is an observation matrix, $\mathbf{S} \in \Re^{N_x \times i}$ a source matrix, and $\mathbf{A} \in \Re^{N_t \times i}$ a mixing matrix. The goal of ICA is to estimate the mixing matrix $\mathbf{A}$ and/or the source matrix $\mathbf{S}$ from the observation matrix $\mathbf{Z}$ with the only hypothesis that the sources are independent.

ICA can be usually resolved by a two-step algorithm, consisting of a prewhitening step and a higher order step. The first step is directly carried out by SVD on the raw data to obtain the whitened (decorrelated and normalized) vectors $\mathbf{v}_j$. At this point, a matrix $\mathbf{V}_i$ can be constructed considering i vectors $\mathbf{v}_j$ which can be chosen from those defining one of the two subspaces $\mathbf{Y}_{\text{sig}}$ or $\mathbf{Y}_{\text{noise}}$ in (5). The matrix $\mathbf{Z}$ in (6) thus represents the subspace constructed with the i corresponding vectors $\mathbf{v}_j$ and $\mathbf{u}_j$. The second step then comprises of finding a rotation matrix $\mathbf{B} \in \Re^{i \times i}$, which diagonalizes the tensor of fourth-order cross-cumulants constructed with the columns of $\mathbf{V}_i$. One of the popular algorithms for finding this rotation matrix is the joint approximate diagonalization of eigenmatrices (JADE) algorithm [25]. JADE uses the joint diagonalization of cumulant matrices obtained by unfolding the tensor of fourth-order cross-cumulants. This second step provides independent vectors $\widetilde{\mathbf{v}}_j$ from the decorrelated ones $\mathbf{v}_j$ given by SVD. These independent vectors are the columns of the matrix $\widetilde{\mathbf{V}}_i = \mathbf{V}_i\mathbf{B}$. ICA can then be synthesized as follows [22]:

$$\mathbf{Z}^T = \sum_{j=k+1}^{k+i} \sigma_j \mathbf{u}_j \mathbf{v}_j^T = \sum_{j=k+1}^{k+i} \widetilde{\sigma}_j \widetilde{\mathbf{u}}_j \widetilde{\mathbf{v}}_j^T \tag{7}$$

where the first equality is given by the first step and is written with respect to the SVD decomposition defined previously, while the second equality is given by the second step. The $\widetilde{\sigma}_j$ are called modified singular values [22]. The second step relaxes the orthogonality condition on the vectors $\widetilde{\mathbf{u}}_j$, while imposing a fourth-order independence criterion for the vectors $\widetilde{\mathbf{v}}_j$. This allows to extract non-Gaussian sources thus providing better estimates of the sources defined by the most important factors. The ICA decomposition allows to extract a second signal subspace defined by i_2 sources

$$\mathbf{Z}_{\text{sig}}^T = \sum_{j=k+1}^{k+i_2} \widetilde{\sigma}_j \widetilde{\mathbf{u}}_j \widetilde{\mathbf{v}}_j^T. \tag{8}$$

B. Proposed Scheme

The methodology adopted for the identification of leakages will now be discussed. As a first step of data decomposition in subspaces, PCA is applied. The precipitations and the leakages are ephemeral phenomena in time and time/space domains respectively and are considered as "noise." This, in turn, means that they are not coherent to all the acquired data and are thus not going to be revealed in the first few vectors obtained by PCA. The result of PCA is analyzed in terms of choice of the number of singular values, m, to be kept for constructing the signal subspace $\mathbf{Y}_{\text{sig}}$. The choice of this value is made on the basis of the most significant singular values, which are dependent on the source signals (as we will see in the next section). Following the construction of the signal subspace, the noise or the *PCA residue* subspace can be obtained as the difference between the initial data signal $\mathbf{Y}$, contained in a space of dimension N, and the signal subspace constructed with m PCA sources $\mathbf{Y}_{\text{sig}}$, contained in a space of dimension m. This residue is of dimension $N - m$ and is denoted as $(m+1 : N)$. ICA is then applied on this residue but considering only "i" sources to be estimated, which means that only the decorrelated sources $\mathbf{v}_{m+1}$ to $\mathbf{v}_{m+i+1}$ are considered in the second step of the ICA algorithm and that $k = m$ in (7). This allows a new subspace decomposition giving: 1) a second signal subspace $\mathbf{Z}_{\text{sig}}$ constructed with "i_2" significant ICA sources, and thus contained in a space of dimension i_2 and 2) a noise subspace or *ICA residue* denoted as $(m + i_2 + 1 : N)$, obtained as the difference between the PCA residue and $\mathbf{Z}_{\text{sig}}$. This ICA residue should contain the information uniquely related to the leakages. This seemingly simple technique requires some important decisions like the choice of the number of PCA singular values (m) used to construct the first signal subspace, the number of ICA sources to be estimated (i), and the number of ICA sources to be retained for constructing the second signal subspace (i_2). In addition, the processing can be done on the totality of the data or using a sliding window (in order to consider only short duration analysis). The latter will introduce an additional parameter in form of the sliding window size. The preprocessing steps that can be applied to the data before the application of source separation techniques have to be investigated as well. Summarizing the above discussion, the proposed scheme for the identification of leakages is presented in Fig. 2. In the next section, we will present the results of application of the proposed method on real data preceded by a study of preprocessing methods and followed by a discussion on the results.

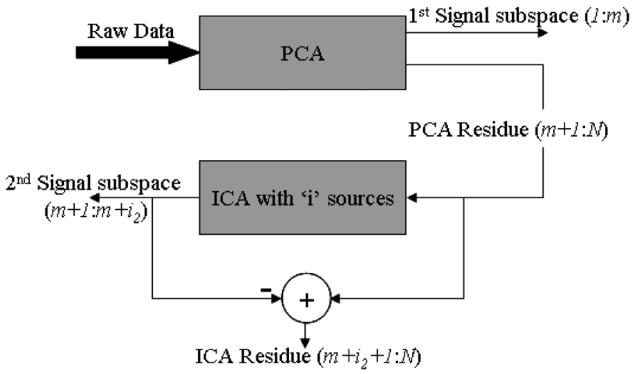

Fig. 2. Proposed scheme based on the temperature data for identification of leakages in dikes using source separation techniques (PCA and ICA).

IV. Results and Discussion

A. Preprocessing

We present next the preprocessing steps that will be used for real data set. Note that these steps are combined as described in the next section.

As a first preprocessing step, the data set is normalized (zero mean and unity variance of each profile, i.e., of each column of the recorded data set $\mathbf{Y}$). This normalization leads to attenuation of the temporal variability of the data (daily and seasonal variations).

The leakages are characterized by high dynamics with a steep slope, whereas other factors are characterized by slow variations (low-frequency contents), either in time or in distance. Taking the gradient (derivative) of the data with respect to time or distance may thus partially remove the slow variations. There are two possibilities for the gradient: the temporal gradient and the spatial gradient. In the temporal gradient, the gradient operator is applied for each sensor along the entire time duration of the signal, whereas in the spatial gradient, the gradient operator is applied for each profile along the entire length of the fiber. The former thus attenuates the effects of existing structures like drains (which exist at all times) and seasonal variations, while the latter attenuates the effects of the ground response and precipitations (which vary slowly along the distance). The fact that the spatial gradient is calculated using relative differences between successive sensors for each profile ensures that the result no more depends on seasonal variations. Note that for this processing, the normalization step is not mandatory. Application of temporal gradient, however, results in the loss of the temporal evolution of the leakages, and thus of useful information. For this purpose, we use only the spatial gradient as a preprocessing step.

Though it may be inapt to state here explicitly as a preprocessing step, yet in certain cases we perform a transposition, after normalization and derivative (provided it is taken), before passing the data on to the source separation process. This means that we consider the data set $\mathbf{Y}$ in (3) instead of $\mathbf{Y}^T$ and that for each factor, the assigned source is a function of time. Processing data in this manner may give a possible characterization of phenomena which exist at all times.

B. Results on Data Set

In this section, we present the results of application of the proposed scheme on a real thermometric data set. The data for the period of three months: April, May, and June is used for analysis purposes and is presented in the Fig. 3(a), with the temperature variations (in °C) expressed by grayscale variations as function of time and distance along the fiber. The normalized data, as discussed in previous subsection, is presented in Fig. 3(b). Note that recordings on 200 m length of the fiber-optic cable corresponding to hydraulic station are not taken into consideration.

1) Processing on the Entire Data: In this section, the results obtained by application of the proposed scheme on three-months data are presented. The normalized version of this data [Fig. 3(b)] is treated to obtain the singular values σ_j, where the first 20 values are represented in Fig. 4(a). The ratio $(\sigma_1/\sum_{j=1}^{N}\sigma_j) = 54.8\%$ is low, meaning that the coherence between the sources $\mathbf{v}_j$ estimated by PCA is not so significant and these sources characterize the recorded noise rather than the useful information. The choice of the parameters "m" and "i" is based on the decrease of these singular values. We started off by setting the threshold at $m = 3$ for constructing the signal subspace using PCA as it marks a change of slope in the singular values graph [16]. Then, $i = 2$ sources are estimated by applying ICA on the PCA residue. Of these two sources, one source each was used to characterize the signal and the noise subspace, thus giving $i_2 = 1$. The sources estimated by PCA and ICA are presented in Figs. 4(b)–(f). The PCA sources do not reveal any discrimination with regards to the leakages (see Table I for details) but the drains, D1 and D2, can be distinguished on the third PCA source at 0.56 and 0.86 km, respectively. We also observe on the ICA sources, $\widetilde{\mathbf{v}}_j$, that ICA does not bring any complementary information concerning the leakages. Note that the sources extracted by PCA are arranged in a decreasing order of energy. However, the energy of a leakage is very low as compared to entire data and thus the leakages remain present in the residue even if we take many singular values for constructing the PCA signal subspace. We tested different values of m, i, and i_2, but the results are quite similar to those obtained previously, which led us to conclude that the estimation of leakages is incomplete. However, the energy of a leakage is relatively large compared to short duration recorded data, so we will now focus our attention on short duration analysis and will show that the processing is better adapted with windowed data.

2) Processing on Windowed Data: We present here the results of processing between the last day of April and the 13th day of May (both inclusive). This windowed section contains the three leakages and the two drains but there are no significant precipitations. The application of the same scheme considering normalized data produces the sources given in Fig. 5 along with the first 20 singular values that result from the PCA. The first PCA source represented in Fig. 5(b) does not contain any information that is pertinent to the leakages. Note that this source is very similar with the one given by considering the three-months data set [see Fig. 4(b)]. This source reveals two temperature zones: Zone 1, from approximately 0.2 to 1.25 km and Zone 2, from 1.25 to 2.2 km. The former has temperature

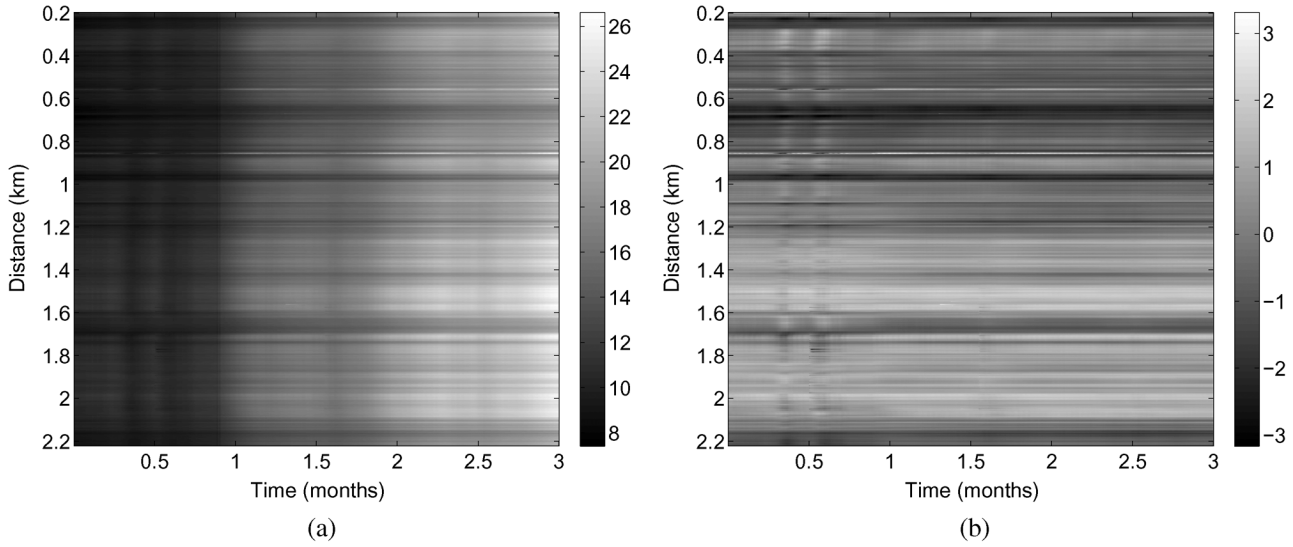

Fig. 3. (a) Raw data along the entire length of the fiber acquired over a period of three months. (b) Normalized data.

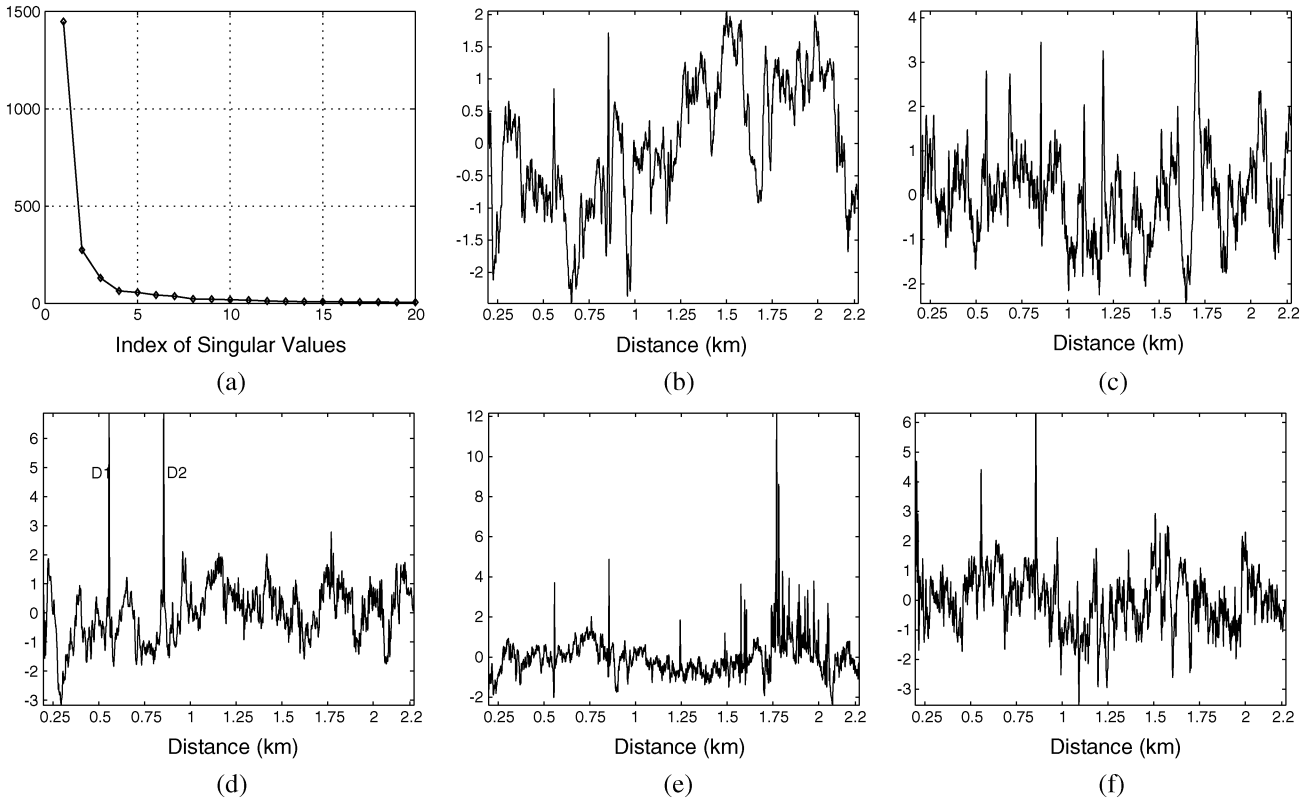

Fig. 4. Singular values and sources for the data of three months of April, May, and June treated in its entirety: (a) first 20 singular values; (b)–(d) first three PCA sources; (e) and (f) first two ICA sources.

values primarily below zero, whereas the latter has values above zero. At the actual Oraison site, these two zones correspond to two different elevation levels, with Zone 1 at lower elevation and Zone 2 at higher elevation (see Fig. 1). These different elevation levels mean that Zone 2 will be more exposed to direct sunlight than Zone 1. Consequently, this first PCA source characterizes the ground response. This, in itself, is an important result as no *a priori* information regarding the elevation levels was incorporated in the separation algorithm, yet we have a ground response characterizing source. The ratio $\sigma_1/\sum_{j=1}^{N}\sigma_j$ is now 80%, which means that the first PCA source extracts a more energetic subspace than in the previous case, allowing to access the leakages by the next PCA sources. Indeed, the other two PCA sources [in Fig. 5(c) and (d)] contain the information linked to the leakage L1 located at 1.562 km (see the arrows). We can thus choose $m = 1$ and $i = 2$. After the ICA step, the first ICA source in Fig. 5(e) contains the information linked to the drains (D1, D2), whereas the second ICA source contains the infor-

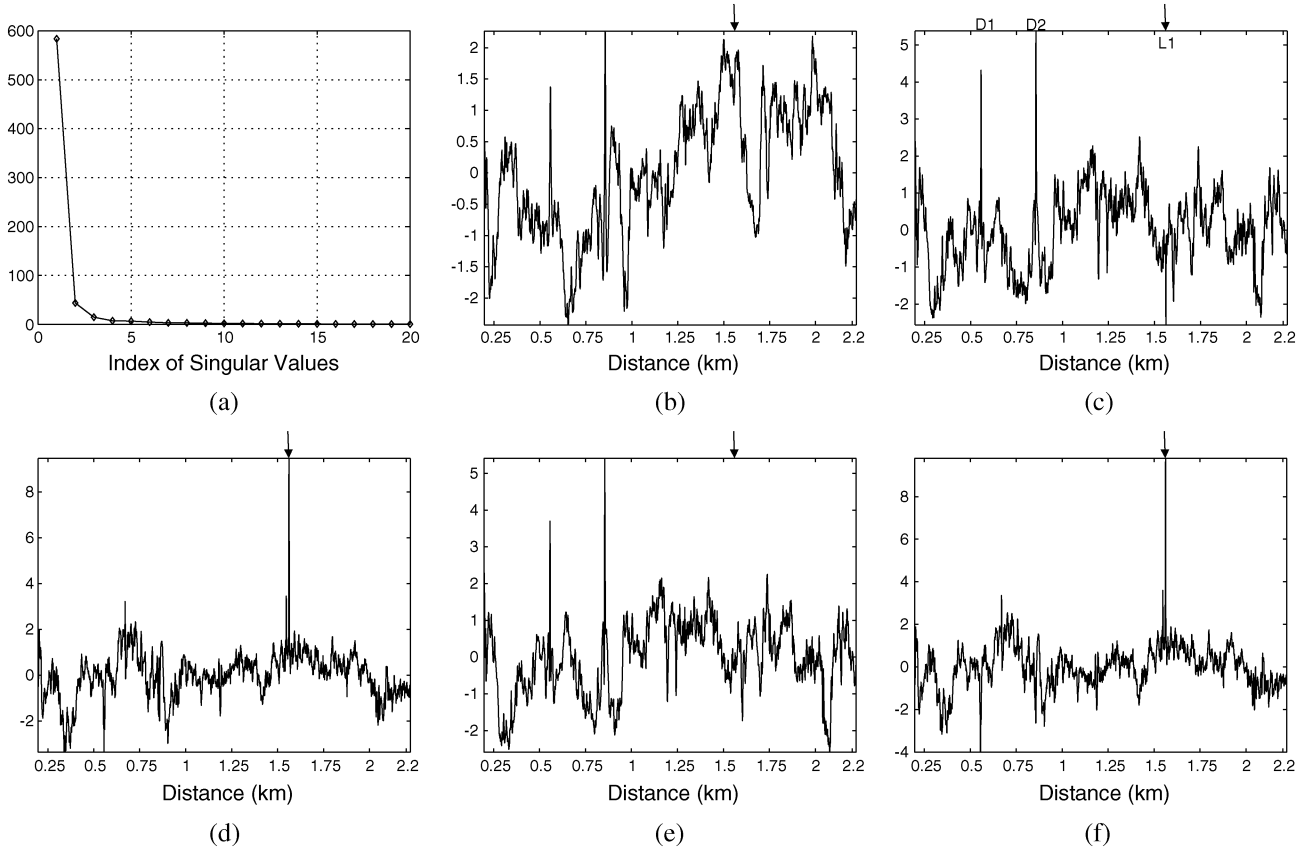

Fig. 5. Source separation for a 14 day windowed data section: (a) first 20 singular values; (b)–(d) first three PCA sources; and (e), (f) first two ICA sources. The leakage L1, whose location is marked by arrows, does not appear in the first ICA source but is present in the second PCA source.

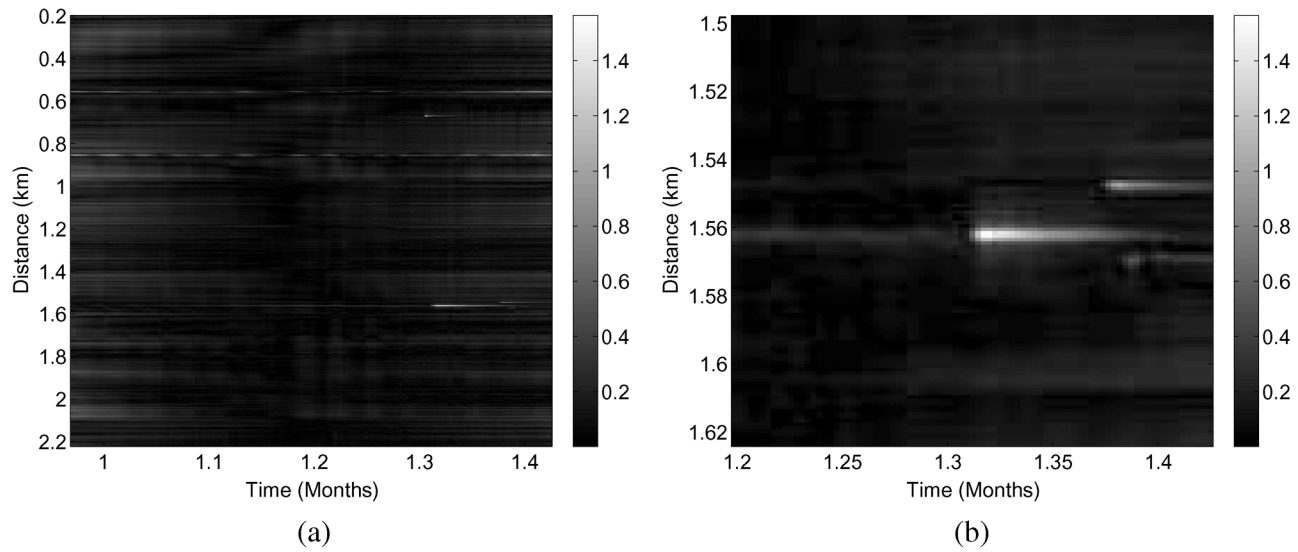

Fig. 6. (a) Envelop of the ICA residue for the windowed data section. (b) Zoom of (a) in the vicinity of leakages.

mation linked significantly to the leakage. We can thus choose $i_2 = 1$ to detect the leakages. The envelope (representing the module) of the ICA residue is shown in Fig. 6, along with its zoomed version in the vicinity of the leakages. It is worthwhile to consider the behavior of PCA and ICA for identification of the leakages. The first PCA source contains no information linked to the leakages so we are quite safe in terms of constructing the first signal subspace with only one PCA source. Meanwhile, the second PCA source contains the leakage and the drains simultaneously. If we compare the second PCA source with the first ICA source in terms of the relative strengths of the leakage and the drains, we observe that the first ICA source contains no information concerning the leakages. Using fourth-order statistics, ICA allows to extract independent factors, which are leakages and drains. Using the first ICA source for constructing the second signal subspace will thus result in a better separation of

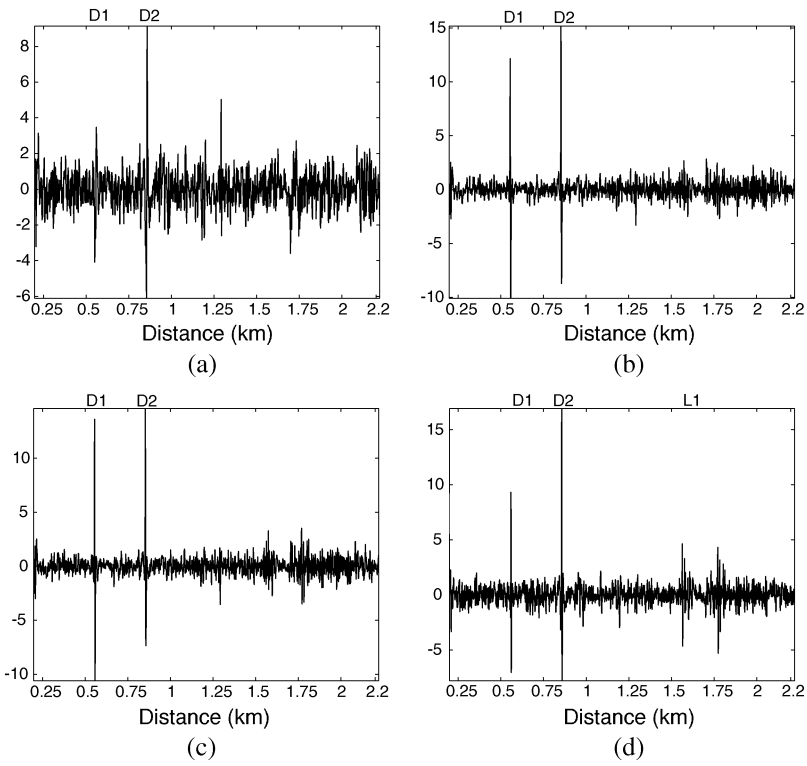

Fig. 7. Sources for the data with spatial gradient taken for a window size of 14 days and an index 121. (a) and (b) First two PCA sources. (c) and (d) First two ICA sources.

drains and leakages. Consequently, when we construct the ICA residue, the drains will be less visual because a part of them has already been retained for constructing the signal subspace with i_2 ICA sources. This justifies the second step of our proposed scheme of performing the ICA on the PCA residue.

The results obtained above highlight the fact that the energy of the identified leakage is very low as compared to the entire data but becomes relatively large when compared to the data of a short duration processing window. Various window sizes were used in order to select a suitable size for processing our data set. Our findings revealed that a window size of 14 days gives the best results. For smaller window sizes, the first ICA source was found to contain a relatively significant information on the leakages. Keeping this first source for the signal subspace construction led to the result that the relative strength of the identified leakages as compared to the drains was low. In addition, for window sizes greater than 14 days, the results were not good and were corrupted with noise artifacts. In order to justify the viability of the proposed scheme on the entire data set, we process the three-months data set with a sliding window of 14 days.

3) Processing in Temporal Sliding Window: As mentioned above, it is important to process the data in windowed sections, so we apply our processing scheme on the entire data using a temporal sliding window of 14 days. We consider an overlapping sliding window method with the results of the processing of each 14 day block placed at the center of the corresponding block. The processing window is then slided in steps of 16 points with each point corresponding to an acquisition of 2 h. This sliding step size smooths the edge effects and was chosen after trials with different sizes. By considering processing on sliding window of 14 days, the proposed scheme extracts the leakages, as shown in Fig. 6. We focus here on two other cases: spatial gradient with and without transposition of data.

Spatial Gradient: The leakages are characterized by high dynamics, whereas other factors such as the ground response have low frequency behavior. Taking the spatial gradient of the data may reduce somewhat the effect of these factors, as well as that of seasonal variations. Following the normalization preprocessing step, the spatial gradient of the data was taken. As we observed in the previous sections, the choice of the parameters m, i and i_2 depends on the singular values, as well as the information contained in the estimated sources. So, we present results with different indices, where an index is defined here as the combination "$m\ i\ i_2$." For example, $m = 1$, $i = 2$, and $i_2 = 1$ constitute the index 121. The estimated PCA and ICA sources are presented in Fig. 7 and the ICA residue with an index 121 is given in Fig. 8. We notice on the first PCA source [see Fig. 7(a)] that effects of the ground response are attenuated as compared to the nongradient case [see Fig. 5(b)]. The first ICA source in Fig. 7(c) contains little information linked to the leakage L1, however, it is highly dominated by the information linked to the drains. The residue obtained is energetic at the position of leakages, as shown in Fig. 8, which gives the best possibility of leakage identification. The residue still contains some information linked to the drains (the horizontal lines in the residue) but the leakages are nevertheless identified. Other parameters have been tested in order to completely eliminate the information linked to the drains. The results obtained with an index 142 and the window length of 14 days are presented in Fig. 9. The results obtained are very much similar in terms of

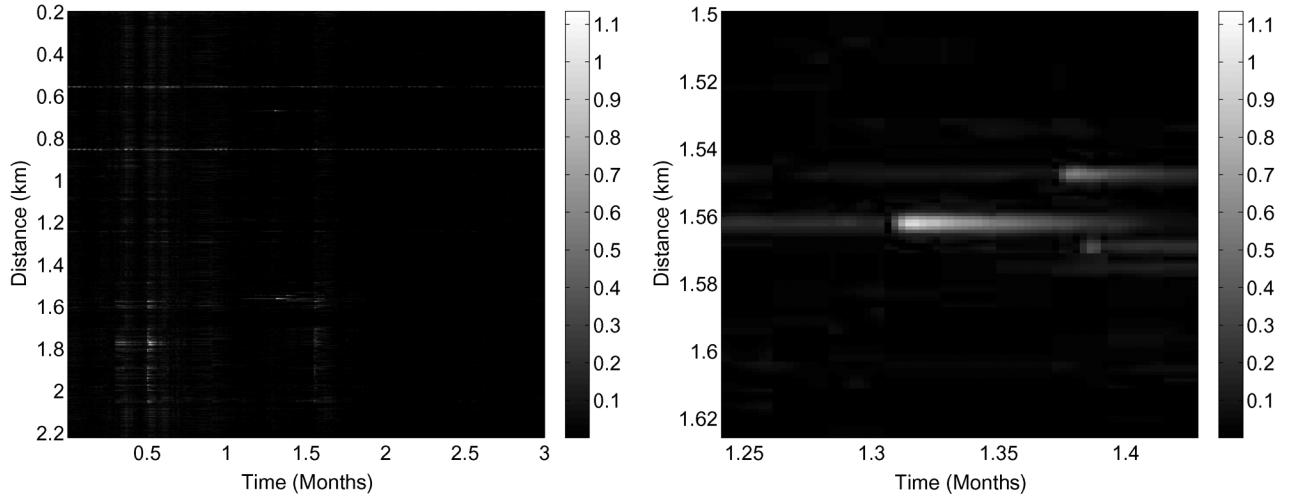

Fig. 8. Residue for the data with spatial gradient taken for a window size of 14 days and an index 121. (a) Envelop of ICA residue. (b) Zoom of (a) in the vicinity of leakages.

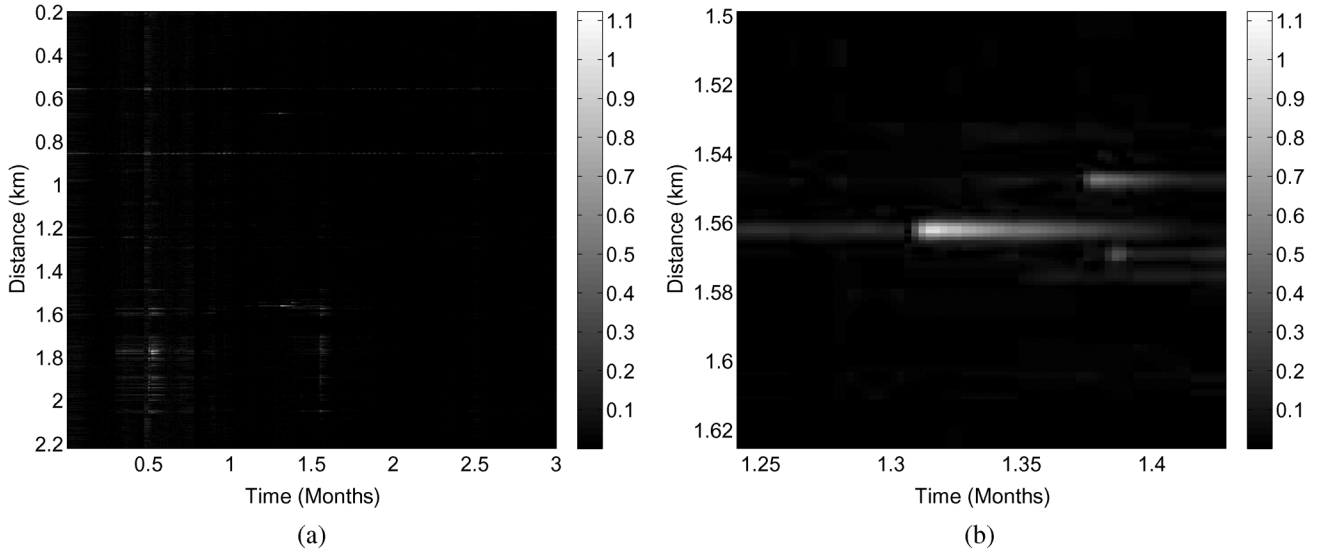

Fig. 9. Residue for the data with spatial gradient taken for a window size of 14 days and an index 142. (a) Envelop of ICA residue. (b) Zoom of (a) in the vicinity of leakages.

the final residue to what were obtained for the index 121. The effects related to precipitation and drains are observable in the residue with their inherent characteristics. Previously, we have also tested different window sizes. However, the best possibility of leakage identification was obtained with an index 121 considering a window size of 14 days.

Spatial Gradient and Transposition of Data: The original data after being normalized is passed through the spatial gradient and then transposed. The idea behind data transposition was to see if we could possibly characterize the existing structures (the drains). This means that we consider the data set $\mathbf{Y}$ in (3) instead of $\mathbf{Y}^T$ and that the extracted sources depend on time. It should, however, be noted that we use the same orientation while plotting the results, i.e., time on horizontal axis and displacement along the fiber on vertical axis. This is done so as to simplify the comparison with the results obtained previously. With a window size of 14 days used for the analysis and an index 121, the ICA residue in Fig. 10 reveals energetic leakages. A close observation of the leakages, however, reveals that the identified leakages are not well localized in distance. For example, the leakage $L1$ originally located at 1.562 km is not identified at the same position in the residue. Instead, two maxima of this leakage occur at 1.561 and 1.563 km, thereby already introducing a localization error. The result with an index 142, using the same window size of 14 days, is given in Fig. 11. It can be clearly observed that there is a loss of temporal evolution of the leakages in addition to the spatial localization error. In order to give a quantitative measure, we compare the energy of the first identified leakage against that of the background in terms of SNR. A window of size 3×6 centered on the leakage $L1$ is first selected. The energy E_L of this leakage is then calculated as the sum of the squares of each element of this window. The energy E_B of the background is then estimated in the same manner but this time centering the window in the background (a

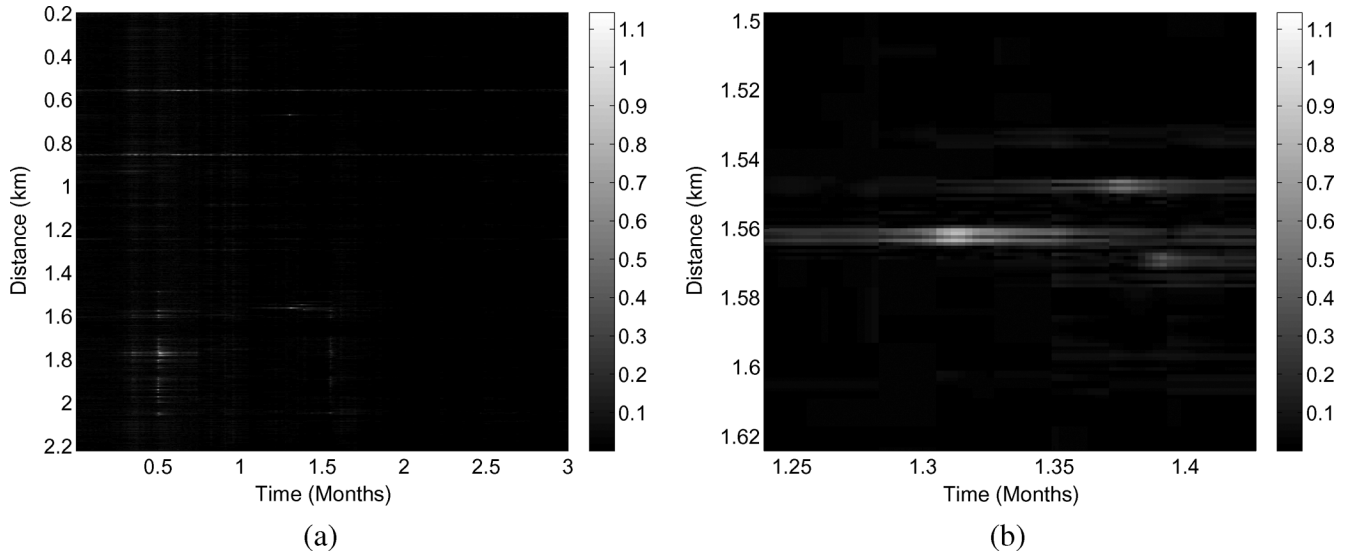

Fig. 10. Residue for the data with spatial gradient taken and transposition applied for a window size of 14 days and an index 121. (a) Envelop of ICA residue. (b) Zoom of (a) in the vicinity of leakages.

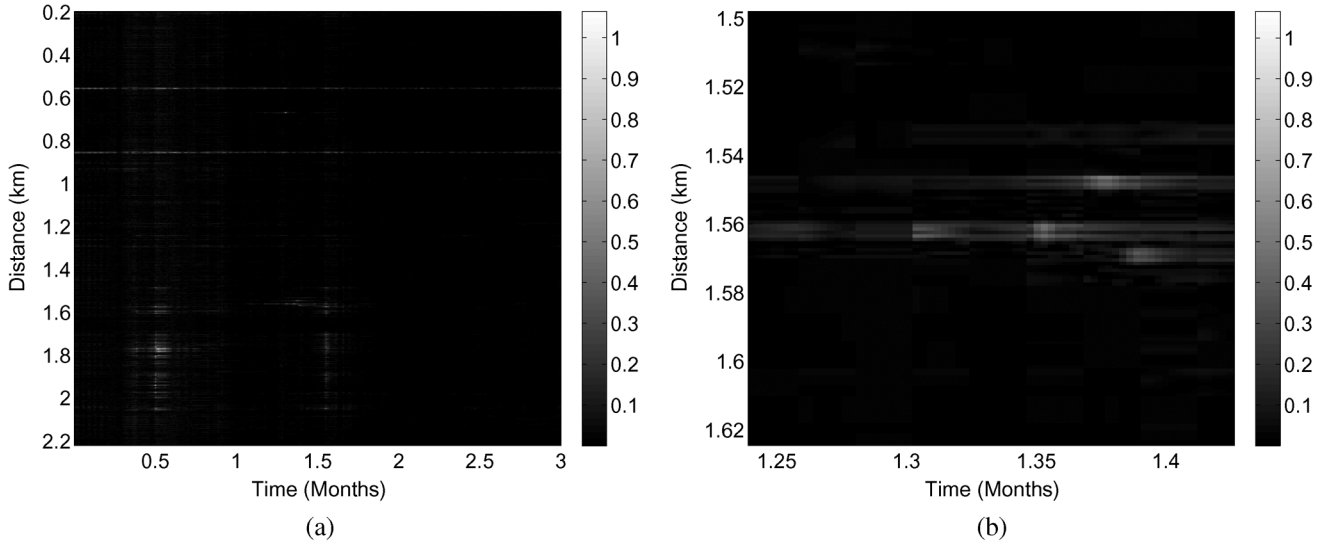

Fig. 11. Residue for the data with spatial gradient taken and transposition applied for a window size of 14 days and an index 142. (a) Envelop of ICA residue. (b) Zoom of (a) in the vicinity of leakages.

region in the leakage vicinity which does not contain any leakages). Note that we have supposed here a constant background, otherwise, the energy E_B must be estimated by computing a mean energy for different window positions. The SNR is then calculated as:

$$\mathrm{SNR} = 10 \log_{10} \left(\frac{E_L}{E_B} \right). \tag{9}$$

It was found out that the SNR reduces from 70 dB for the nontransposed case to about 30 dB for the transposed case. This reduction of SNR coupled with loss of temporal evolution of the leakages and the inability to provide their exact spatial localization render the transposed data analysis less attractive against the nontransposed case.

We also tested the other possible case, i.e., the transposed data without gradient. We found out that the leakages are identified with a relatively low SNR, which might render the eventual detection difficult. In the next subsection, we briefly discuss the observations obtained from these results.

C. Discussion

We have studied the cases with various preprocessing steps, as well as with various window sizes. In addition to that, the choice of the number of sources to be estimated in the two steps of the treatment, i.e., $(m, i, i2)$, was also studied. We found out that for our thermometric data set, a sliding window of 14 days gives the best identification of leakages. It is difficult to make a clear distinction between different window sizes but the leakages identified with a window of 14 days have a higher SNR.

We observed that the analysis with the nontransposed data (i.e., with time on the horizontal axis and the distance on the vertical axis) gives us a better possibility of leakage identification than the transposed case. Moreover, the noise (the undesired effects other than leakages, such as drains and precipitations) that appears in the final residue has lowest energy when we use nontransposed data before processing. The notion of applying the spatial gradient reveals the leakages more clearly, i.e., the identified leakages are relatively more energetic than what we obtain without applying the gradient. Based on the above discussion, we can work out the following scheme for the separation and identification of the leakages.

First of all, the data is normalized thus rendering each column at zero mean and unity variance. Then, the spatial gradient of the data is taken to attenuate the slow variations. This step is followed by the subsequent application of PCA and ICA. We perform this analysis using the sliding window approach, where a window of 14 days is the best in terms of leakage identification.

V. Conclusion

In the present work, we have proposed a method for the identification of leakages in dikes using the temperature data obtained through fiber-optic DTSs. We showed how it is possible to treat leakage identification as a source separation problem. The sources were considered as defining the response of the ground, the known structures in the path of the fiber sensors (drains), the seasonal variations, the precipitations and, of course, the leakages, the last ones being our desired signals. We have shown that with the help of techniques based on data decomposition and source separation, we can identify the leakages. It was shown how certain preprocessing steps, like data normalization and application of spatial gradient, can enhance the possibility of leakage identification. Moreover, we also proved with different analyses that the application of separation techniques in the temporally sliding window gives us a better possibility of leakage identification. The question on the choice of the window size arises for which it was found that a window of 14 days gives the best results. The choice of the number of sources to be estimated by PCA and ICA techniques was addressed by observing the estimated sources as a function of distance along the fiber, as well as the singular values. It was found that for the best results, a single source should be utilized for constructing the signal subspace with PCA. The corresponding residue should then be treated using ICA on two sources and a second signal subspace should be constructed with one ICA source. With the application of some rather simple signal processing techniques, we have shown that it is possible to identify leakages that may exist in dikes. Their flow rate and a development of possible mathematical model for the characterization of leakages along with the application of multidimensional data processing techniques for a better representation of the final residue will form the follow up of this work.

References

[1] A. H. Hartog, "Distributed fiber-optic temperature sensors: Principles and applications," in *Optical Fiber Sensor Technology*, K. T. Grattan and B. T. Meggitt, Eds. New York: Kluwer, 2000, pp. 241–301.

[2] A. H. Hartog, "Progress in distributed fiber-optic temperature sensing," in *Proc. SPIE, Fiber Optic Sensor Technology and Applications, 2001*, M. A. Marcus and B. Culshaw, Eds., 2002, vol. 4578, pp. 43–52, ser. 0277-786X/02.

[3] S. A. Wade, K. T. Grattan, and B. McKinley, "Incorporation of fiber-optic sensors in concrete specimens: Testing and evaluation," *IEEE Sensors J.*, vol. 4, no. 1, pp. 127–134, Feb. 2004.

[4] S. Grosswig, A. Graupner, and E. Hurtig, "Distributed fiber optical temperature sensing technique—A variable tool for monitoring tasks," in *Proc. 8th Int. Symp. Temperature and Thermal Measurements in Industry and Science*, Jun. 19–21, 2001, pp. 9–17.

[5] P. J. Henderson, N. E. Fischer, and D. A. Jackson, "Current metering using fiber grating based interrogation of a conventional current transformer," in *Proc. 12th Int. Conf. Optical Fiber Sensors*, Williamsburg, VA, 1997, pp. 186–189.

[6] A. D. Kersey, "Optical fiber sensors for downwell monitoring applications in the oil and gas industry," in *Proc. 13th Int. Conf. Optical Fiber Sensors*, 1999, pp. 326–331.

[7] E. Lewis, C. Sheridan, M. O'Farrell, D. King, C. Flanagan, W. B. Lyons, and C. Fitzpatrick, "Principal component analysis and artificial neural networks based approach to analyzing optical fiber sensors signals," *Sens. Actuators A*, vol. 136, pp. 28–38, 2007.

[8] A. Rozycki, J. M. Ruiz, and A. Caudra, "Detection and evaluation of horizontal fractures in earth dams using the self potential method," *Eng. Geo.*, vol. 82, no. 3, pp. 145–153, 2005.

[9] P. Sjödahl, "Resistivity investigation and monitoring for detection of internal erosion and anomalous seepage in embankment dams," Ph.D. dissertation, Lund University, Lund, Sweeden, 2006.

[10] B. Vogel, C. Cassens, A. Graupner, and A. Trostel, "Leakage detection systems by using distributed fiber optical temperature measurements," in *Proc. SPIE Smart Structures and Materials 2001: Sensory Phenomena and Measurement Instrumentation for Smart Structures and Materials*, D. I. E. Udd, Ed., 2001, vol. 4328, pp. 23–34, ser. 0277-786X/01.

[11] S. Johansson, "Localization and quantification of water leakage in ageing embankment dams by regular temperature measurements," in *Proc. ICOLD, 17th Congr.*, Vienna, Austria, 1991.

[12] S. Johansson and P. Sjödahl, "Downstream seepage detection using temperature measurements and visual inspection—Monitoring experiences from Røsvatn field test dam and large embankment dams in Sweden," in *Proc. Int. Seminar on Stability and Breaching of Embankment Dams*, Oslo, Norway, Oct. 2004, p. 21.

[13] S. Yin, "Distributed fiber optic sensors," in *Fiber Optic Sensors*, F. T. Yu and S. Yin, Eds. New York: Marcel Dekker Inc., 2000, pp. 201–229.

[14] A. Hyvärinen, "Survey on independent component analysis," *Neural Computing Surveys*, vol. 2, pp. 94–128, 1999.

[15] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.

[16] N. L. Bihan, V. Vrabie, and J. I. Mars, "Multi-dimensional seismic data decomposition by high order SVD and unimodal ICA," in *Signal and Image Processing for Remote sensing*, C. H. Chen, Ed. New York: Taylor and Francis, 2006.

[17] A. Rogers, "Distributed optical fiber sensing," in *Handbook of Optical Fiber Sensing Technology*, J. M. Lopez-Higuera, Ed. New York: Wiley, 2002, pp. 271–308.

[18] A. H. Hartog, "A distributed temperature sensor based on liquid-core optical fibers," *IEEE J. Lightw. Technol.*, vol. 1, no. 3, pp. 498–509, Sep. 1983.

[19] J. P. Dakin, D. Pratt, and G. W. Bibby, "Distributed optical fiber Raman temperature sensor using a semiconductor light source and detector," *Electron. Lett.*, vol. 21, pp. 569–570, 1985.

[20] K. I. Diamantaras and S. Y. Kung, *Principal Component Neural Networks: Theory and Applications*. New York: Wiley, 1996.

[21] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*. New York: Addison-Wesley, 1991.

[22] V. Vrabie, J. I. Mars, and J.-L. Lacoume, "Singular value decomposition by means of independent component analysis," *Signal Process.*, vol. 84, no. 3, pp. 645–652, 2004.

[23] P. Comon, "Independent component analysis, A new concept?," *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.

[24] J. F. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 86, no. 10, pp. 2009–2025, 1998.

[25] J. F. Cardoso and A. Souloumiac, "Blind beamforming for non Gaussian signals," *IEE Proc.-F*, vol. 140, no. 6, pp. 362–370, 1993.

Amir A. Khan received the B.S. degree in electrical engineering from the University of Engineering and Technology, Taxila, Pakistan, in 2002. In 2003, he joined Ghulam Ishaq Khan Institute of Tech. (GIKI), Pakistan, where he received the M.S. degree in electronics engineering, while serving as a Research Assistant. He received the M.S. degree in signal and image processing from the Grenoble Institute of Technology (INP Grenoble), Grenoble, France, in 2005, where he is currently working towards the Ph.D. degree in signal and image processing working with the Grenoble Images Parole Signals Automatics (GIPSA) Laboratory, focusing on source separation techniques and higher order statistics for leakage detection in dikes.

From 2002 to 2003, he worked as a Maintenance Engineer in the Electrical Department at Fauji Fertilizers (FFBL), Karachi, Pakistan.

Valeriu Vrabie received the B.Sc. degree in electronics and telecommunications from the University Politehnica of Bucharest (UPB), Bucharest, Romania, in 1998, the M.Sc. degree in signal processing from the Grenoble Institute of Technology (INP Grenoble), Grenoble, France, in 2000, and the Ph.D. degree in signal processing jointly from the INPG and UPB in 2003.

Since 2004, he has been an Associate Professor at the Centre de Recherche en STIC (CReSTIC), University of Reims, France. His research involves statistical signal processing and source separation methods with applications in geophysics, biomedical engineering, and biometry.

Jérôme I. Mars received the M.S. degree in geophysics from the Joseph Fourier University (UJF), Grenoble, France, in 1986 and the Ph.D. degree in signal processing from the Grenoble Institute of Technology (INP Grenoble), Grenoble, France, in 1988.

From 1989 to 1992, he was a Postdoctoral Research Fellow at the Centre des Phénomènes Aléatoires et Geophysiques, Grenoble. From 1992 to 1995, he served as a Visiting Lecturer and Scientist at the Materials Sciences and Mineral Engineering Department, University of California, Berkeley. He is currently a Professor at INP Grenoble and works in the Signal Processing Department of the Grenoble Images Parole Signals Automatics Laboratory (GIPSA-Lab), France, where he heads the Signal, Image and Physics team. His research interests include seismic and acoustic signal processing, wavefield separation methods, time-frequency and time-scale characterization and applied geophysics.

Dr. Mars is a member of SEG and EAGE.

Alexandre Girard received the Engineering degree from the Ecole Centrale de Paris, Paris, France, in 2000 with a specialization in applied mathematics.

He worked in optimization and robust control for power plants from 2000 to 2005 as a consultant. He is currently a Research Engineer at Electricité de France (EDF) in control and signal processing.

Guy D'Urso received the Engineering degree from ENSIEG, Grenoble, France, in 1993 with a specialization in signal processing.

He is currently a Research Engineer at Electricité de France (EDF) in signal and image processing with an expert status. He is also interested in measurement domain.