GEO ESSENTIAL
ERA-PLANET

Essential Variables workflows for resource efficiency and environmental management

Deliverable 1.6
Data fusion guidelines

Creator	*Department of Mathematics Tullio Levi-Civita, University of Padova. ** Forschungszentrum Jülich, Institute of Bio- and Geosciences: Agrosphere (IBG-3). *** Dipartimento di Salute della Donna e del Bambino, University of Padova. **** Dipartimento di Ingegneria Civile, Edile e Ambientale – ICEA, University of Padova.
Creation date	July 28, 2019
Due date	August 30, 2019
Last revision date	August 30, 2019
Status	Final
Type	Report
Description	This deliverable proposes tools for data fusion
Right	Public
Language	English
Citation	E. Perracchione*, M. Polato*, W. Erb*, F. Piazzon*, F. Marchetti***, F. Aiolli*, B. Bayat**, A. Botto****, S. De Marchi*, S. Kollet**, C. Montzka**, A. Sperduti*, M. Vianello*, M. Putti*, 2019. Data Fusion guidelines.
Grant agreement	GEOEssential Deliverable 1.6., ERA-PLANET No 689443

Table of Contents

TABLE OF CONTENTS 2

INTRODUCTION 2

FUSING FEATURES VIA KERNELS FOR DEBLURRING 3

TESTS FOR DEBLURRING 5

DATA FUSION RPOD APPROACH 7

TESTS FOR RPOD DATA FUSION 8

CONCLUSIONS 9

REFERENCES 9

Introduction

The objective of this deliverable is to provide guidelines on methods of “data fusion” that will form the framework of Task 1.6 entitled “Data Fusion” (DF). This task is intimately related to task 1.4 “Modeling and Processing Services” and uses similar approximation techniques, such as kernel approximation, for the specific purpose of data fusion. Both deliverables aim at providing value-added services to Essential Variables (EVs).

The report addresses a general state of the art with some specific examples for applications in the context of the GeoEssential project.

Information (or data) fusion can be defined as the study of efficient methods for automatically or semi-automatically transforming information from different sources and different points in time into a representation which provides effective support for human or automated decision making (Boström, et al., 2007). In this sense, Reduced Order Methods (ROMs) might be of interest; see e.g. (Azaiez, et al., 2016).

In the context of multi-sensors imagery, data fusion can be thought of as a process of combining images, obtained by sensors of different wavelengths in order to form a composite and more informative image (Jiang, et al., 2009). In this specific report, we focus on novel kernel methods, that fuse more images together. This, as an immediate application, we study algorithms for deblurring images.

More specifically, we drive our attention towards two aspects of data fusion-based modelling:

- Deblurring images (refer to Section Fusing features via kernels for deblurring): combining features from different images to improve the image resolution (De Marchi, et al., 2017) (De Marchi, et al., 2019) (Bozzini, et al., 2015).
- Fusing data via POD (refer to Section Data fusion POD approach): in data-fusion a challenging problem is the one of merging different data and approximate them. For those multivariate problems ROMs might be helpful (Azaiez, et al., 2016) (Azaiez, et al., 2018).

Fusing features via kernels for deblurring

Among several methods available in literature, we here focus on kernel-based method and specifically, we introduce a novel technique in the context of machine learning. The goal of the proposed method is to consider a set of images, then extract different features from them, and finally glue them together for improving the resolution of possibly blurred images.

Let us suppose to have a set of images $I_k, k = 0, \dots, m$, measuring the same variable at different time steps, i.e. sets of function values $f_i^k, i = 1, \dots, n$, all sampled at the set of n points $X = \{x_i \in \mathbb{R}^d, i = 1, \dots, n\}$. Since we take images, for this section we suppose $d = 2$ and we can think of X as the set of pixels. Let us further suppose that for simplicity only one of these images (without any restrictions I_0) is blurred. We here propose a technique, that, given this blurred image, by extracting features from other images $I_k, k = 1, \dots, m$, fuses those features via kernel matrices and enables us to obtain a deblurred approximation of I_0.

We consider kernels $\kappa: \Omega \times \Omega \to \mathbb{R}$ that can be decomposed via the Mercer's Theorem (see e.g. Theorem 2.2., p. 24, in (Fasshauer & McCourt, 2015)) as

$$\kappa(x, y) = \sum_{k \geq 0} \lambda_k \rho_k(x) \rho_k(y), \quad x, y \in \Omega,$$

where the series converges uniformly and absolutely and $\{\rho_k\}_{k \geq 0}$ is a countable set of eigenfunctions (with the associated eigenvalues $\{\lambda_k\}_{k \geq 0}$) of the operator $T: L_2(\Omega) \to L_2(\Omega)$, given by

$$T[f](x) = \int_\Omega \kappa(x, y) f(y) dy.$$

It is worth to note that we can interpret the Mercer series representation in terms of an inner product in the so-called *feature space* F, which is a Hilbert space. Indeed,

$$\kappa(x, y) = \langle \Phi(x), \Phi(y) \rangle_F, \quad x, y \in \Omega,$$

where Φ from Ω to F is known as *feature map*.

Moreover, for the kernels that admit a Mercer expansion (also called valid kernels, following the definition given by (Shawe-Taylor & Cristianini, 2004)), the mapping Φ enables us to obtain the canonical features $\Phi(x) = \kappa(\cdot, x)$. We refer the reader to (Fasshauer & McCourt, 2015) (Shawe-Taylor & Cristianini, 2004) for further details.

We denote by K the kernel matrix constructed via the data set X, i.e. the matrix of entries

$$\mathrm{K}_{ij} = \kappa(x_i, y_j), \quad i, j = 1, \dots n.$$

Since we consider strictly positive definite kernels, such a matrix is positive definite (Fasshauer, 2007).

In approximation theory, radial kernels are the most considered, which are kernels for whom there exists a Radial Basis Function (RBF) $\varphi: [0, \infty) \to \mathbb{R}$ and (possibly) a shape parameter $\gamma > 0$ such that for all $x, y, \in \Omega$

$$\kappa(x, y) = \kappa_\gamma(x, y) = \varphi_\gamma(||x - y||_2^2) := \varphi(r).$$

We now suppose that we are able to extract p features from m images, such as edges. Then, formally it means that we suppose to know a function $\psi: \mathbb{R}^n \to \mathbb{R}^p$, so that the fused kernel becomes

$$\kappa^\Psi(x, y) = k((x, \Psi(x)), (y, \Psi(y))),$$

where κ is a strictly positive definite kernel on $\mathbb{R}^{n+p}$. This coincides with the so-called Variably Scaled Kernels (VSKs), which have been recently studied.

To explain how this method can be used in this framework, let us suppose to have an original blurred image I_0 and related function values f^0. For deblurring and fuse data we can construct an approximation of the form

$$I_0 = \sum_{i=1}^{n} \alpha_i \kappa^\Psi(x, x_i),$$

which, after imposing the interpolation conditions, leads to solving the system:

$$\mathrm{K}^\Psi \alpha = f^0,$$

where $\alpha = (\alpha_1, \dots, \alpha_n)^T$, $f^0 = (f_1^0, \dots, f_n^0)^T$ and

$$\mathrm{K}_{ij}^\Psi = \kappa^\Psi(x_i, y_j), \quad i, j = 1, \dots n.$$

We now focus on an application of such method in the context of the GeoEssential project.

Tests for deblurring

In the numerical experiments that follow, we consider two test images measuring the soil moisture over Europe. Soil moisture is a key variable for hydrology which turns out to be meaningful for many applications, such as modeling climate variability and water resources.

The first images we consider consist of raw data taken by NASA Soil Moisture Active Passive (SMAP) satellite on April 2015 (Entekhabi at al., 2014). Such satellite has been launched on January 31, 2015, and the mission is designed to principally measure soil moisture over the Earth.

The second test images we take consist of simulated soil moisture data obtained via the TERRestrial SYStem Modelling Platform (TerrSysMP) that was developed to simulate the interaction between lateral flow processes in river basins with the lower atmospheric boundary layer (Kollet & Maxwell, 2008) (Shrestha, et al., 2014). The Centre for High-Performance Scientific Computing in terrestrial systems (HPSC TerrSys) is operating TerrSysMP in a forecasting setup over North Rhine-Westphalia and Europe. The model results are made available for the scientific community daily as videos via the YouTube Channel of HPSC TerrSys https://www.youtube.com/channel/UCGio3ckQwasR5a_kJo1GdOw.

We now provide an example. Let us suppose to have a blurred image plotted in Figure 1, (top, left). This image has been reconstructed with the polynomial least squares, as explained in (Perracchione, et al., 2019). The image from which we are able to extract features is plotted in the top right frame. In this case, as augmented features, we extract the edges (see bottom left). Then, such features, are used to construct the *augmented* kernel matrix K^{Ψ}. And finally, by solving the interpolation system, we get the deblurred image plotted in the bottom right frame. In this specific case we took as reference solution the image plotted in the top right frame, so that we can compute the Root Mean Square Error (RMSE) as a post-processing technique. For the blurred image we have RMSE=$1.48\mathrm{E}-02$, while for the deblurred one we obtain RMSE=$8.33\mathrm{E}-03$.

As a second test case, in the same framework, we consider the examples plotted in Figure 2. For the blurred image we have RMSE=$7.12E-03$, while for the deblurred one we obtain RMSE=$3.74E-03$. In both cases, we can affirm that the proposed method, simple from the computational point of view, turns out to be robust.

In what follows, we focus on ROMs and their applications in the context of data fusion.

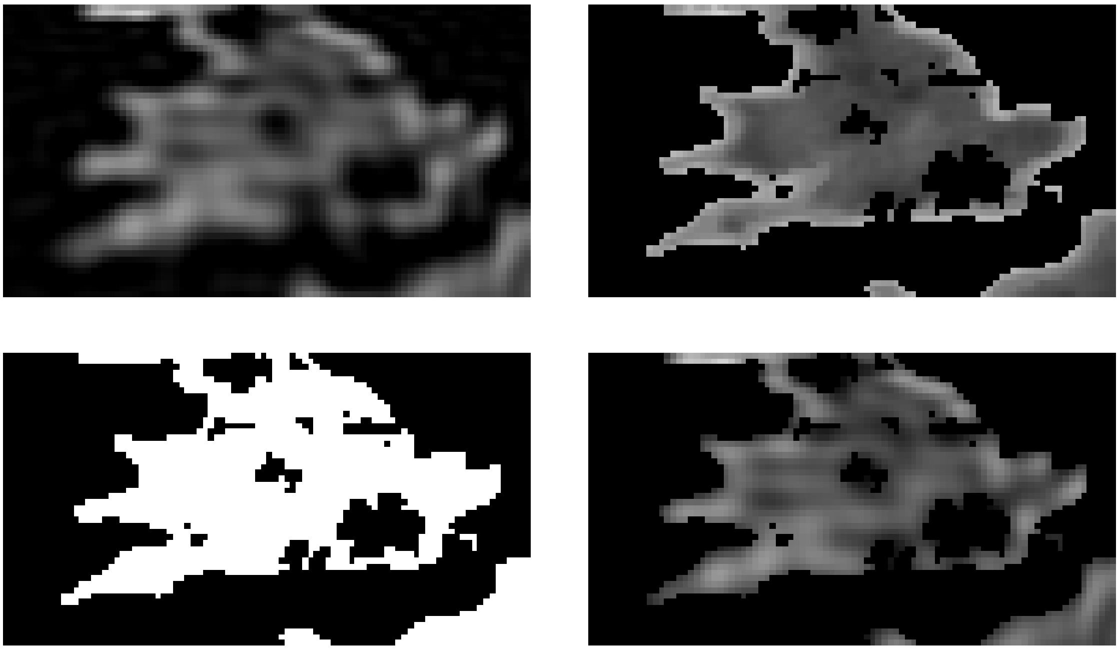

Figure 1: Top: the blurred image (left) and the original image (right). Bottom: the extracted edges (left), i.e. the augmented features, and the deblurred image (right).

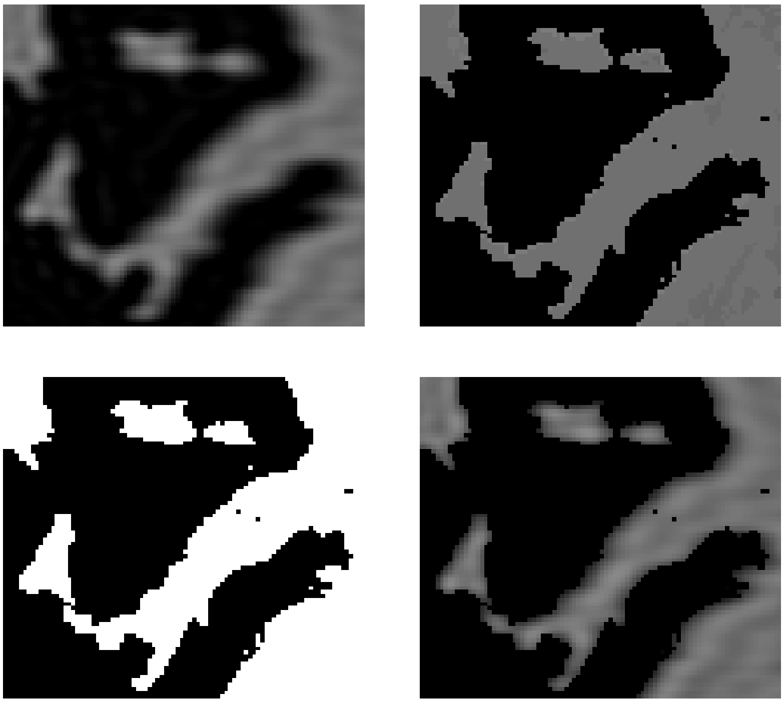

Figure 2: Top: the blurred image (left) and the original image (right). Bottom: the extracted edges (left), i.e. the augmented features, and the deblurred image (right).

Data fusion RPOD approach

The main problem in data-fusion is the one of merging different data and approximate them. Multivariate problems are thus often encountered and in this context ROMs might be of interest. Such methods, in the discrete case, allow to decompose tensors. As a consequence, in approximating tensorial functions, we always reduce to univariate interpolation.

To introduce the Recursive POD (RPOD) for multivariate functions, we take a function f in the Lebesgue space $L^2(X_1 \times X_2 \times \cdots \times X_d)$, where $X_1, \dots, X_d \subset \mathbb{R}$ are bounded domains. Then, we know that f admits the following expansion (Azaiez, et al., 2016) (Azaiez, et al., 2018)

$$f(x_1, x_2, \dots, x_d) = \sum_{i_1 \in N} \sigma_{i_1} v_{i_1}(x_2, \dots, x_d)\, \varphi_{i1}(x_1),$$

where the sum is convergent in $L^2(X_2 \times \cdots \times X_d, L^2(X_1))$, and where $\{\varphi_{i1}\}_{i_1 \in N}$ and $\{v_{i1}\}_{i_1 \in N}$, are two orthonormal sets respectively complete in $L^2(X_1)$ and $L^2(X_2 \times \cdots \times X_d)$.

Continuing in applying recursively the POD, we construct the expansion of

$$v_{i1}, v_{i2}^{(i_1)}, v_{i_{d-2}}^{(i_1 i_2, \dots, i_{d-3})},$$

and we have that the function $f \in L^2(X_1 \times X_2 \times \cdots \times X_d)$, can be written as; refer to Lemma 2.1 in (Azaiez, et al., 2018)

$$\begin{aligned} f &= \sum_{i_1 \in N} \sum_{i_2 \in N} \cdots \sum_{i_{d-1} \in N} \sigma_{i_1} \sigma_{i_2}^{(i_1)} \sigma_{i_{d-1}}^{(i_1 i_2, \dots, i_{d-2})} \varphi_{i_1} \otimes \varphi_{i_2}^{(i_1)} \otimes \cdots \otimes \varphi_{i_{d-1}}^{(i_1 i_2, \dots, i_{d-2})} \otimes v_{i_{d-1}}^{(i_1 i_2, \dots, i_{d-2})} \\ &= \sum_{i_1 \in N} \sigma_{i_1} \sum_{i_2 \in N} \sigma_{i_2}^{(i_1)} \cdots \sum_{i_{d-1} \in N} \sigma_{i_{d-1}}^{(i_1 i_2, \dots, i_{d-2})} \varphi_{i_1} \otimes \varphi_{i_2}^{(i_1)} \otimes \cdots \otimes \varphi_{i_{d-1}}^{(i_1 i_2, \dots, i_{d-2})} \otimes v_{i_{d-1}}^{(i_1 i_2, \dots, i_{d-2})}. \end{aligned}$$

Note that, computationally speaking this can be seen as a SVD in higher dimensions. Therefore, once the tensor is decomposed, we can interpolate via standard kernel-based methods each eigenfunction. This leads to univariate interpolation and hence the stability of standard interpolation is improved. We refer to this method as K-POD and we compare it with the standard kernel based method (K-ST) which work in any dimension but might suffer from instability especially in higher dimensions. With the K-RPOD we take advantage of interpolating univariate functions.

We now provide an example that is meaningful in the context of the GeoEssential project.

Tests for RPOD data fusion

As an application, we consider the problem of scoring the quality of air. In this sense knowing the values of PM_{10} and its relations with other air pollutants and chemical factor is if interest. The data we take as test are available at http://www.blackwellpublishing.com/rss.
Those data are collected from 1994 to 1998 in the months November-February and measures the air pollution over Leeds (U.K.). This is indeed, a typical example of multivariate data analysis. However, due to the high number of input parameters the problem becomes challenging and we here apply our tensor decomposition tool. The 3^4 toy tensor, used in our simulations, is constructed with the state of the art by (De Marchi, et al., 2019).

The data set consists of 532 samples and 4 inputs. We briefly discuss below the inputs and we refer to (Heffernan & Tawn, 2004) for further details.

- O_3. Daily maximum ozone in parts per billion.
- NO_2. Daily maximum nitrogen dioxide in parts per billion.
- NO. Daily maximum nitrogen monoxide in parts per billion.
- SO_2. Daily maximum sulfur dioxide in parts per billion.

The output variable is the value of PM_{10}. Thus, to predict it we need to approximate a function f that depends on the above parameters, i.e.

$$\mathrm{PM}_{10} = f(\mathrm{O}_3, \mathrm{NO}_2, \mathrm{NO}, \mathrm{SO}_2).$$

About the 10% of the data set, specifically, $s = 53$ instances, is used for testing the method. We compare ST-K interpolation with K-RPOD approximation. In Table 1 we report the Relative Maximum Absolute Error (RMAE) and the Relative Root Mean Squares Error (RRMSE) for both methods. The K-RPOD takes advantage of decomposing the problem and turns out to be more accurate than a srandard multivariate interpolation.

Method	K-ST	K-RPOD
RMAE	7.14E+01	1.88E+00
RRMSE	9.23E+00	3.70E-01

Table 1: The results for approximating PM_{10}.

To summarize, such a scheme enables us to fuse data of different types, without knowing which is the underlying function.

Conclusions

In this deliverable, Task 1.6 “Data Fusion” (DF), we gave the main guidelines for information fusion issues. Two novel methods have been briefly reviewed presenting which are pro and cons. Then, we focused on specific experiments related to EVs.

Furthermore, these problems are strictly related to the ones presented in Task 1.4 “Modeling and Processing Services”. Indeed, because of the large amount of available data, instability issues might occur and therefore algorithms such as ROMs should be applied to obtain more robust solutions. As work in progress, we have to fuse the soil moisture data with Earth rain observations in order to get more reliable predictions.

References

Azaiez, M., Ben Belgacem, F. & Rebollo Chacon, T., 2016. Recursive POD expansion for reaction-diffusion equation. *Adv.Model. and Simul. in Eng. Sci.*, p. 3:3.

Azaiez, M., Rebollo, T. C., Perracchione, E. & Vega, J. M., 2018. Recursive POD expansion for advection-diffusion-reaction equation. *Comm. Comput. Physics*, pp. 24: 1556-1578.

Boström, H. et al., 2007. On the Definition of Information Fusion as a Field of Research”. IKI Technical Reports. *IKI Technical Reports*..

Bozzini, M., Lenarduzzi, L., Rossini, M. & Schaback, R., 2015. Interpolation with variably scaled kernels. *IMA J. Numer. Anal.*, pp. 35:199-219..

De Marchi, S., Erb, W. & Marchetti, F., 2017. Spectral filtering for the reduction of the Gibbs phenomenon for polynomial approximation methods on Lissajous curves with applications in MPI. *Dolomites Res. Notes Approx.*, p. 10.

De Marchi, S. et al., 2019. Shape-Driven Interpolation with Discontinuous Kernels: Error Analysis, Edge Extraction and Applications in MPI. *Preprint.*

De Marchi, S., Marchetti, F. & Perracchione, E., 2019. Jumping with Variably Scaled Discontinuous Kernels (VSDKs). *Preprint.*

Entekhabi at al., D., 2014. *SMAP Handbook-Soil Moisture Active Passive.* s.l.:JPL Publication; Pasadena, CA, 2014.

Fasshauer, G. E., 2007. *Meshfree Approximations Methods with Matlab.* s.l.:World Scientific, Singapore.

Fasshauer, G. E. & McCourt, M. J., 2015. *Kernel-based Approximation Methods Using Matlab.* s.l.:World Scientific, Singapore.

Heffernan, J. E. & Tawn, J. A., 2004. A conditional approach for multivariate extreme values. *Journal of the Royal Statistical society B* , pp. 66:497-546..

Jiang, D., Zhuang, D., Huang, Y. & Fu, J., 2009. Advances in Multi-Sensor Data Fusion: Algorithms and Applications. *Sensors*, pp. 10:7771-7784..

Kollet, S. & Maxwell, R. M., 2008. Capturing the influence of groundwater dynamics on land surface processes using an integrated, distributed watershed model. *Water Resour. Res.*, p. 44:W02402.

Perracchione, E. et al., 2019. Modelling and processing services and tools. p. GEOEssential Deliverable 1.3.

Shawe-Taylor, J. & Cristianini, N., 2004. *Kernel Methods for Pattern Analysis.* New York, NY, USA: Cambridge University Press.

Shrestha, P. et al., 2014. A scale-consistent terrestrial systems modeling platform based on COSMO, CLM, and ParFlow. *Mon. Weather Rev.* , pp. 142:3466-3483.