OPTIMAL TRANSPORT FOR DATA FUSION IN REMOTE SENSING

Nicolas Courty[1], Rémi Flamary[2], Devis Tuia[3], Thomas Corpetti[4]

[1]IRISA/Université de Bretagne-Sud, Vannes, France
[2] Lagrange/CNRS/UNS/OCA, Nice, France
[3] MMRS/University of Zurich, Zurich, Switzerland
[4] LETG/CNRS, Rennes, France

ABSTRACT

One of the main objective of data fusion is the integration of several acquisition of the same physical object, in order to build a new consistent representation that embeds all the information from the different modalities. In this paper, we propose the use of optimal transport theory as a powerful mean of establishing correspondences between the modalities. After reviewing important properties and computational aspects, we showcase its application to three remote sensing fusion problems: domain adaptation, time series averaging and change detection in LIDAR data.

Index Terms— Optimal transport, domain adaptation, time series analysis, change detection, LIDAR

1. INTRODUCTION

Data fusion deals with the integration of multiple data sources into a single coherent and consistent representation that can be used for several purposes. In the case of remotely sensed data, these sources can be expressed through different acquisitions [1, 2]: spatial, spectral and temporal data describing the same physical object or phenomenon. When possible, those different modalities have to be matched in a common mathematical and numerical representation. In this work, we advocate the use of optimal transport (OT) to perform this matching. Here, the different modalities are expressed as distributions in their respective spaces, and OT then seeks for an optimal coupling, *i.e.* a way of transporting one distribution onto another. This optimality is relative to a cost of transportation from one space to the other, which can model specific interactions between the modalities.

OT was firstly proposed to solve mass transportation problems in the $19th$ century and reappeared in the middle of the 20th century in the work of Kantorovitch [3], and found recently surprising new developments of several fundamental problems [4]. It was applied in a wide panel of fields, including among others image analysis and processing [5, 6], computer graphics [7], or machine learning [8].

D. Tuia acknowledge the Swiss National Science Foundation for financial support, through the grant PP00P2-150593

We first begin by a short introduction to OT and the associated computational numerical methods (Section 2). We then describe three different applications of this framework in remote sensing data fusion problems (Section 3).

2. A BRIEF INTRODUCTION TO OPTIMAL TRANSPORT

We aim at matching two distributions μ_1 and μ_2 that are related to the same object of interest (but not forcedly registered). Let $\mathcal{X}$ be a space that embeds both distributions. For example, in the case of two images at different resolutions in a resolution enhancement problem, $\mathcal{X}$ is the spatial coordinates domain over which the two images were registered. Let c be a cost function $\mathcal{X}^2 \to \mathbb{R}^+$. This cost can be related to a metric d over $\mathcal{X}$, such as a Euclidean distance.

Wasserstein distance. Noting $\Pi(\mu_1, \mu_2)$ be the space of probability distributions over $\mathcal{X}^2$ with prescribed marginals μ_1 and μ_2, the p-Wasserstein distance between μ_1 and μ_2 is defined as:

$$W_p(\mu_1, \mu_2) = \left(\inf_{\pi \in \Pi(\mu_1, \mu_2)} \int_{\mathcal{X}^2} c(\mathbf{x}_1, \mathbf{x}_2)^p d\pi(\mathbf{x}_1, \mathbf{x}_2) \right)^{\frac{1}{p}} \quad (1)$$

In the remainder, we will consider the Wasserstein distance of order 1, simply noted $W(\mu_1, \mu_2)$. This distance, also known as the Earth Mover Distance in computer vision community [5], allows to compute a distance on the metric space $\mathcal{P}(\mathcal{X})$. Remarkably, this minimization problem admits a minimizer π_0 [4], which is called an optimal transportation plan, and can be intuitively understood as a probabilistic coupling between μ_1 and μ_2 which minimizes the expected distance between the coupled elements.

Discrete case. Whenever μ_1 and μ_2 are available as empirical distributions (probability masses over Diracs located in $\mathcal{X}$), they belong to probability simplices Σ^{n_1} and Σ^{n_2} of dimension n_1 and n_2. Consider the previous examples of two images at different resolutions. Those images are considered as empirical probability density functions, defined by a set of Diracs located at the pixels spatial positions $\mathbf{X}_1$ and $\mathbf{X}_2$, with a vector of probability masses $\mathbf{m_1}$ and $\mathbf{m_2}$. Here,

$\mu_1 = \mathbf{m_1}^T\mathbf{X}_1$ and $\mu_2 = \mathbf{m_2}^T\mathbf{X}_2$. This probability mass per pixel can be defined as the intensity value of the pixel divided by the sum of all intensities in the image, so that $|\mathbf{m_1}|_1 = |\mathbf{m_2}|_1 = 1$. $\Pi(\mu_1, \mu_2)$ is then defined as a set of matrices of size $n_1 \times n_2$ with constrained marginals :

$$\mathcal{P} = \left\{\gamma \in (\mathbb{R}^+)^{\mathbf{n_1}\times\mathbf{n_2}} | \, \gamma\mathbf{1_{n_2}} = \mu_\mathbf{1}, \gamma^\mathbf{T}\mathbf{1_{n_1}} = \mu_\mathbf{2}\right\} \quad (2)$$

where $\mathbf{1}_l$ is a l-dimensional vector of ones. The Wasserstein metric becomes:

$$W(\mu_1, \mu_2) = \min_{\gamma\in\mathcal{P}} \langle\gamma, \mathbf{C}\rangle_F \quad (3)$$

where $\langle ., . \rangle_F$ is the Frobenius dot product and $\mathbf{C} \geq 0$ is a cost matrix of size $n_1 \times n_2$ which gathers all the costs for transporting the Diracs of μ_1 to Diracs of μ_2. This problem can be solved by linear programming, with combinatorial algorithms such as the simplex methods and its network variants (transport simplex, network simplex, etc.) [9], but is is to be noted that the number of variables in this problem scales with the product of the number of bins in the discrete representations of μ_1 and μ_2. This makes the original problem computationally costly.

Displacement interpolation. OT can be used not only to compute distance but also to interpolate between the distributions. Once the optimal transport matrix $\boldsymbol{\gamma}_0$ has been found, we can transform the source elements $\mathbf{X}_1$ in an interpolated version $\hat{\mathbf{X}}_1$:

$$\hat{\mathbf{X}}_1 = \text{diag}((\boldsymbol{\gamma}_0\mathbf{1}_{n_1})^{-1})\boldsymbol{\gamma}_0\mathbf{X}_2. \quad (4)$$

This way, the new expression of the support of the μ_1 distribution is expressed with barycentric coordinates found in $\boldsymbol{\gamma}_0$.

Regularized optimal transport. In many situations, we assume a smooth transportation plan between distributions. In these cases, we can enforce some kind of regularization to help obtaining a better transportation by including additional prior in the optimization problem, such as Laplacian [10]), entropy [11], or class-regularization [12] on the transport matrix . When using a regularization, the optimal transport optimization problem can be reformulated as

$$\boldsymbol{\gamma}_0 = \underset{\boldsymbol{\gamma}\in\mathcal{P}}{\arg\min} \quad \langle\boldsymbol{\gamma}, \mathbf{C}\rangle_F + \lambda R(\boldsymbol{\gamma}), \quad (5)$$

where $\lambda \geq 0$ is a regularization parameter and $R(\cdot)$ is a regularization term encoding prior information on the transportation matrix. The formulation in [11] also has the advantage of being computationally efficient since it can be solved using a multiplicative algorithm, which reduces the computational cost from a quadratic algorithm to a linear scaling.

Wasserstein barycenters. The space of probability measures over $\mathbf{X}$ equipped with the Wasserstein metric defines a complete metric space, the Wasserstein space. The geometry of a Wasserstein space embeds implicitly knowledge about the problem through the particular cost function c. Hence, geometrical concepts such as means or barycenters are now at hand and can be used for several purposes. Notably, one can now try to find a probability measure $\mu \in \Sigma^n$ (with n possibly different from n_1 and n_2) of several distributions $\{\mu_k\}_{k=1...K}$ in the Wasserstein sense. It is defined as:

$$\mu^\star = \underset{\mu\in\Sigma_N}{\arg\min} \sum_{k=1}^{K} w_k W(\mu, \mu_k), \quad (6)$$

w_k being user defined weights (barycentric coordinates in the Wasserstein space simplex, $w_k = 1/n \; \forall k$ defining the mean). Here again, efficient algorithms have been proposed to tackle this complex optimization problem [13, 14].

3. APPLICATIONS TO DATA FUSION

We showcase in this section three examples for the use of OT in fusion problems: domain adaptation, time series analysis and change detection in LIDAR data.

3.1. Domain adaptation

A landcover classification problem is now considered. As the quantity of remote sensing data is growing, the requirement of disposing of labeled information for each image acquired is impossible to fulfill. As such, one must make the best use of the labeled information that is available for other similar scenes. The idea might seem tempting, but re-using labeled information as such generally leads to catastrophic results: from one acquisition to the other, the spectra are distorted by either the acquisition parameters, the atmospheric conditions or by the differences in scale/appearance of the objects in different geographical regions [15]. The compensation for such distortions, or *shifts* is one of the lively areas of machine learning, *domain adaptation* [16].

OT is providing a natural solution to this problem, by allowing to devise a non-linear transformation in the spectral dimensions of the image that helps in matching both distributions [17]. This transformation is directly obtained through the barycentric interpolation presented in Eq. (4). Moreover, one can also use the class information in one of the domain to promote a better transportation for instance by using an appropriate class-based regularization of the coupling matrix [12].

3.2. Remote sensing time series analysis

Because of time warping, the question of comparing and averaging remote sensing time series is still open. For example, in vegetation monitoring two time series of a similar culture can be delayed in time (due to regional climate, seeding date, ...) even though they represent the same phenomenon. Therefore estimating a reliable difference between time series

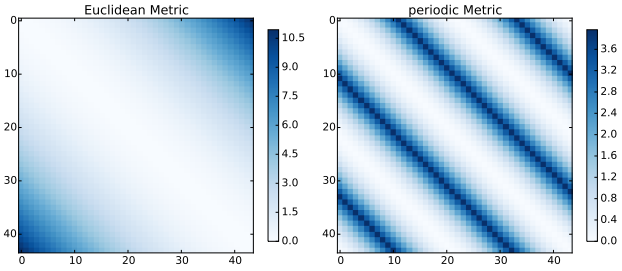

Fig. 1. **Cost functions used for averaging time series** While the first represents a classical Euclidean metric, the second one is a periodic metric

is a critical problem. To this end the well-known Dynamic Time Warping (DTW) approach has been develop and widely used to extract best possible alignments between time series and to compare them on the basis of such alignments. However, DTW is not a metric in the sense that it does not respect the triangular inequality . Hence computing averages under DTW is not trivial, though more or less heuristic approaches exist [18]. In this application, we exploit optimal transport to compute averages of EVI (Enhanced Vegetation Index) times series over two years (2005-2006) issued from MODIS images on the Amazon forest (Figs. 2a and 2c, corresponding to time series of forests with unimodal and bimodal cycles of growth and decay, respectively).
The series are normalized so that they sum to one. Thus they can be considered as probability measure in the time domain. Then, the Wasserstetin barycenter is computed using Eq. 6 with the iterative Bregman projection technique [14]. Because of the specific periodic nature of the observation (collection of measurements lasting 2 years), a different cost function can be used, which embeds a periodicity notion: transporting masses occurring at the same moment of the year does not cost anything as illustrated in Fig. 1.

Extracted barycenters for L_2 (mean), L_1 (median) and Wasserstein distance with euclidean and periodic metric are depicted in Figs. 2b and 2d. From these series, one observes that L_2-based averages are noisy because of existing time delays between series of the same class, yielding unsatisfactory means. Using Wassersteinbarycenters, the corresponding averages are more in accordance with phenological cycles, but still suffer of a lack of consistency at both ends of the time series and during the second growth cycle in 2d. The periodic Wasserstein mean enables to remove these flaws in the averaging and provide very meaningful averages, showing the potential of optimal transport to interpret remote sensing time series.

3.3. Change detection in LiDAR data

In this application, we aim at monitoring changes affecting the coastal cliff face to understand the ongoing erosion pro-

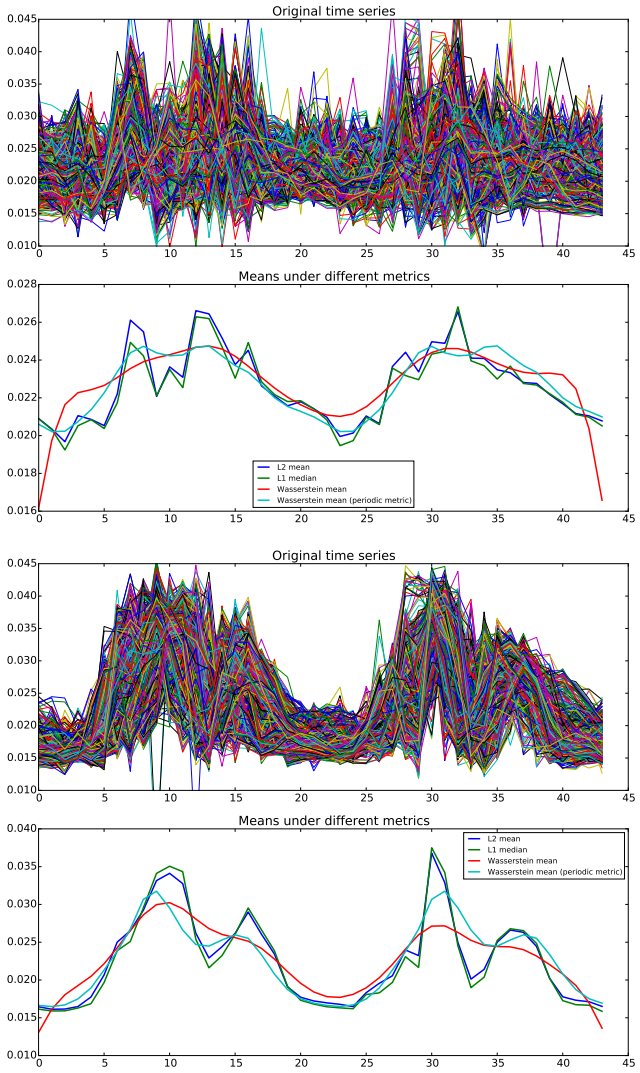

Fig. 2. **Average of EVI time series on 2005-2006 years**. (a)-(c): raw time series ; (b-d): averages based on L_2, Wasstertein and periodic Wasstertein

cess . To this end, terrestrial laser scanning (LiDAR) measurements are regularly performed in order to assess into detail the morphological structures of cliffs. However at the moment, changes are extracted using differences of Digital Elevation Models computed from the 3D point clouds LiDAR point cloud [19]. This is obviously not optimal, since it requires an interpolation step that could disturb areas especially those affected by local rockfalls, where the higher resolution of the LiDAR scan would be the most beneficial.

Here we rather propose to use an optimal transport directly on the LiDAR point cloud to highlight changes. In order to do so, we simply compute a displacement interpolation of the first cloud (acquired on a chalk cliff of upper Normandy (France) on Oct., 7th 2010, Fig. 3a) on the second (acquired

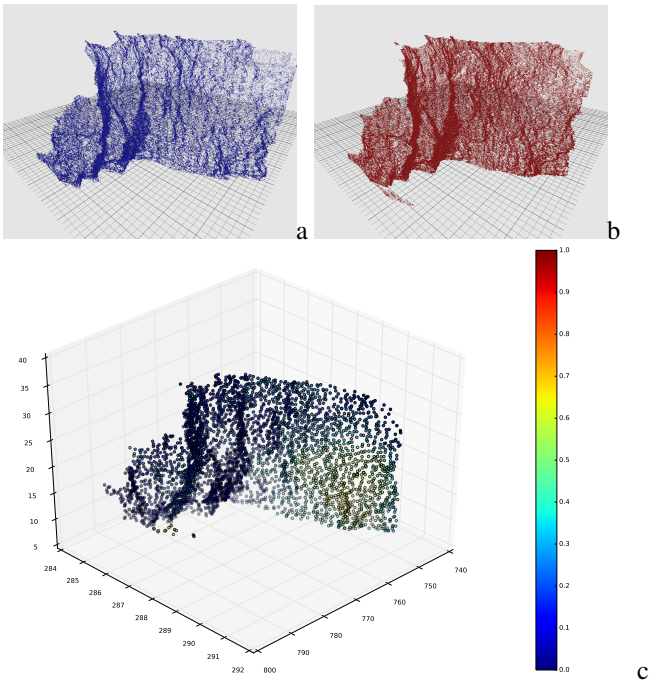

Fig. 3. **Change detection in LiDAR data**. (a)-(b) : the two scans; (c) : magnitude of the changes detected.

on July, 6th, 2011 over the same region, Fig. 3b). Note that this operation does not require to have the same number of points nor any kind of landmarks. The quantity of changes is simply given, for each of the original point, by the magnitude of change, $|\mathbf{X}_1 - \hat{\mathbf{X}}_1|^2$. Results are shown in Fig. 3, where the two 3D point clouds are represented in the top panel, and the magnitude of changes are depicted in the bottom panel. The lower right part of the area has strongly been affected by rockfalls, which is consistent with on site observations and with the study in [19]. However here, the important precision in quantifying altered areas using optimal transport makes this approach a very good alternative for future studies in LiDAR change detection.

4. CONCLUSION

In this work, we presented Optimal Transport as a way of performing data fusion for remote sensing image processing problems. We shown the potential of Optimal Transport in three challenging applications: domain adaptation, time series averaging and LiDAR change detection. In all application Optimal Transport emerged as a valid alternative to current approaches to process non-registered, complex and locally evolving data.

5. REFERENCES

[1] J. Zhang, "Multi-source remote sensing data fusion: status and trends," *Int. J. Data Fusion*, vol. 1, no. 1, pp. 5–24, 2010.

[2] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1560–1584, 2015.

[3] L. Kantorovich, "On the translocation of masses," *C.R. (Doklady) Acad. Sci. URSS (N.S.)*, vol. 37, pp. 199–201, 1942.

[4] C. Villani, *Optimal transport: old and new*, vol. 338, Springer Science & Business Media, 2008.

[5] Y. Rubner, C. Tomasi, and L.J. Guibas, "A metric for distributions with applications to image databases," in *Proc. ICCV*, Jan 1998, pp. 59–66.

[6] W. Wang, D. Slepčev, J. A. Basu S.and Ozolek, and G. K. Rohde, "A linear optimal transportation framework for quantifying and visualizing variations in sets of images," *Int. J. Comp. Vision*, vol. 101, no. 2, pp. 254–269, 2013.

[7] Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas, "Convolutional wasserstein distances," *ACM Trans. Graph.*, vol. 34, no. 4, pp. 66:1–66:11, 2015.

[8] M. Cuturi and A. Doucet, "Fast computation of Wasserstein barycenters," in *Proc. ICML*, jun 2014.

[9] N. Bonneel, M. van de Panne, S. Paris, and W. Heidrich, "Displacement interpolation using lagrangian mass transport," *ACM Trans. Graph.*, vol. 30, no. 6, pp. 158:1–158:12, Dec. 2011.

[10] S. Ferradans, N. Papadakis, G. Peyré, and J-F. Aujol, "Regularized discrete optimal transport," *SIAM J. Imaging Sciences*, vol. 7, no. 3, 2014.

[11] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transportation," in *Proc. NIPS*, pp. 2292–2300. 2013.

[12] N. Courty, R. Flamary, and D. Tuia, "Domain adaptation with regularized optimal transport," in *Proc. ECML*, 2014.

[13] M. Cuturi and G. Peyré, "A Smoothed Dual Approach for Variational Wasserstein Problems," *SIAM J. Imaging Sciences*, Dec. 2015.

[14] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré, "Iterative Bregman Projections for Regularized Transportation Problems," *SIAM J. Scientific Computing*, vol. 2, no. 37, pp. A1111–A1138, 2015.

[15] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson, "Advances in hyperspectral image classification," *IEEE Signal Proc. Mag.*, vol. 31, pp. 45–54, 2014.

[16] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: a survey of recent advances," *IEEE Signal Proc. Mag.*, vol. 32, no. 3, pp. 53–69, 2015.

[17] D. Tuia, R. Flamary, A. Rakotomamonjy, and N. Courty, "Multitemporal classification without new labels: a solution with optimal transport," in *Proc. Multitemp*, 2015.

[18] F. Petitjean, A. Ketterlin, and P. Gançarski, "A global averaging method for dynamic time warping, with applications to clustering," *Pattern Rec.*, vol. 44, no. 3, pp. 678–693, 2011.

[19] P. Letortu, S. Costa, J.-M. Cador, C. Coinaud, and O. Cantat, "Statistical and empirical analyses of the triggers of coastal chalk cliff failure," *Earth Surf. Proc. Landforms*, 2015.