Diva: a Data Analysis Software with Generalized-Cross Validation and Quality Control

Ch. Troupin [a], M. Rixen [b], D. Sirjacobs [a] and J.-M. Beckers [a]

[a] GHER - MARE, Sart-Tilman B5, University of Liège, BELGIUM
[b] NURC - NATO Undersea Research Centre, Viale San Bartolomeo 400, 19126 La Spezia, ITALY

Abstract

The Diva software aims to determine the values of a given field on a regular grid, knowing the value of this field at discrete data points.
The method is based on the minimisation of a variational principle; the problem is solved with the help of a finite element mesh. An error field is also provided by the software, allowing the user to know where the analysis is valuable.
Some tools were added recently to the core of Diva: a fitting tool of the correlation length, an estimation of the Signal-to-Noise ratio (λ) using a Generalized Cross Validation (GCV) and a data quality control (QC).

1 Introduction

In oceanography a typical concern consists in determining a field $\varphi(\mathbf{r})$ on a regular grid of positions $\mathbf{r}$, knowing N_d data in locations $\mathbf{r_j}, j = 1, \ldots, N_d$.
This is called the *gridding problem* and has numerous applications: data analysis, graphical display, forcing or initialization of models, ...
Because of data error and often close data points, the field is always reconstructed with the help of *approximation*, never with a strict interpolation.
Instead of working with the data themselves, we work with *anomalies* (φ') with respect to a *background field* (φ_b):

$$\varphi(\mathbf{r}) = \varphi_b(\mathbf{r}) + \varphi'(\mathbf{r}). \qquad (1)$$

The background field φ_b is defined *a priori* and the anomalies are calculated with respect to this reference field.
Diva is based on the **V**ariational **I**nverse **M**ethod (VIM), which was initially designed for climatology [Brasseur et al. (1996)]: in that case, you have generally high resolution vertical profiles in all seasons, but irregular horizontal coverage. Thus a spatial analysis on horizontal planes is needed. In case of large number of data, the numerical cost of Optimal Interpolation (OI) is too high, hence the development of a new method.
Diva relies on a efficient finite-element solver, which allows to perform the analysis *only* at points where the sea is present.

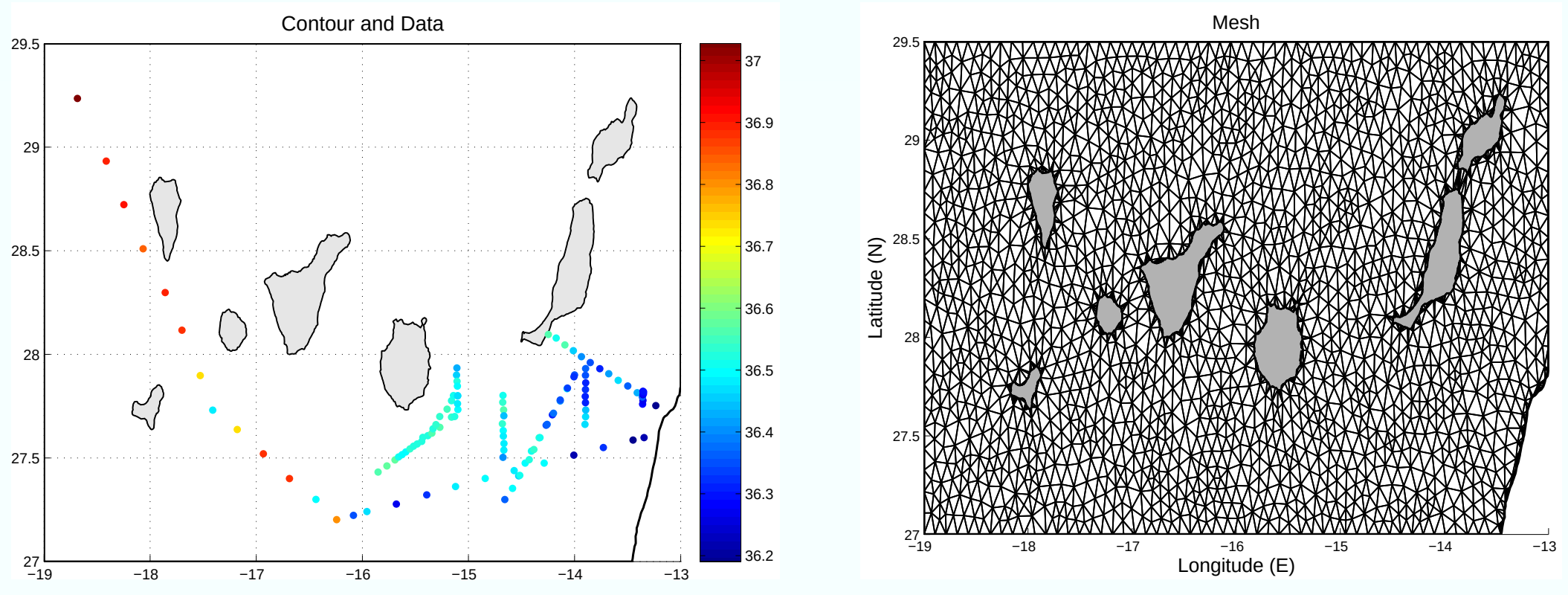

FIGURE 1: *Example: data (salinity), contour and mesh.*

2 Diva theory

2.1 Formulation

In Diva, the field φ has to minimize the variational principle:

$$J\left[\varphi\right] = \sum_{j=1}^{Nd} \mu_j \left[d_j - \varphi(x_j, y_j)\right]^2 + \|\varphi - \varphi_b\|^2 \qquad (2)$$

with

$$\|\varphi\| = \int_D (\alpha_2 \boldsymbol{\nabla\nabla}\varphi : \boldsymbol{\nabla\nabla}\varphi + \alpha_1 \boldsymbol{\nabla}\varphi \cdot \boldsymbol{\nabla}\varphi + \alpha_0 \varphi^2)\, dD \qquad (3)$$

over the region of interest. The parameters $\alpha_i\,(i = 0, 1, 2)$ can be evaluated as functions of the correlation length L and the signal-to-noise ratio λ:

- $\alpha_0 = \frac{1}{L^4}$ penalizes the field itself (anomalies),
- $\alpha_1 = \frac{2\xi}{L^2}$ penalizes the gradients (no trends),
- α_2 penalizes the variability (regularization),

2.2 Weights on data

Each data d_i can be assigned with a weight μ_i showing the confidence you have in it. μ is computed as a function of the signal-to-noise ratio and the correlation length according to

$$\mu = \frac{4\pi\,\lambda}{L^2}. \qquad (4)$$

In Diva, this weighting is done by adding a fourth column to the data input file (`x | y | data value`).

2.3 Background field

Interpolation (and extrapolation) works on anomalies with respect to a background field (1). Diva allows you to work with different background fields, depending on the analysis you want to perform:

1. no treatment is applied, assuming your data are already anomalies;
2. the average value is subtracted from the data values;
3. the linear regression (plane) is subtracted from the data values;
4. an additional semi-normed field ($\alpha_0 = 0$ and large L) obtained by two chained Diva executions is subtracted.

2.4 Error field

In Diva, error fields are calculated by analogy with OI: since analysis in OI is equivalent to analysis with VIM and since error field of OI equals analysis of covariance fields, error field of VIM equals analysis (by VIM) of covariance fields [Rixen et al. (2000)].
In practice, the data input of the analysis tool is a vector containing the covariance of data points with the point in which the error estimate has to be calculated.

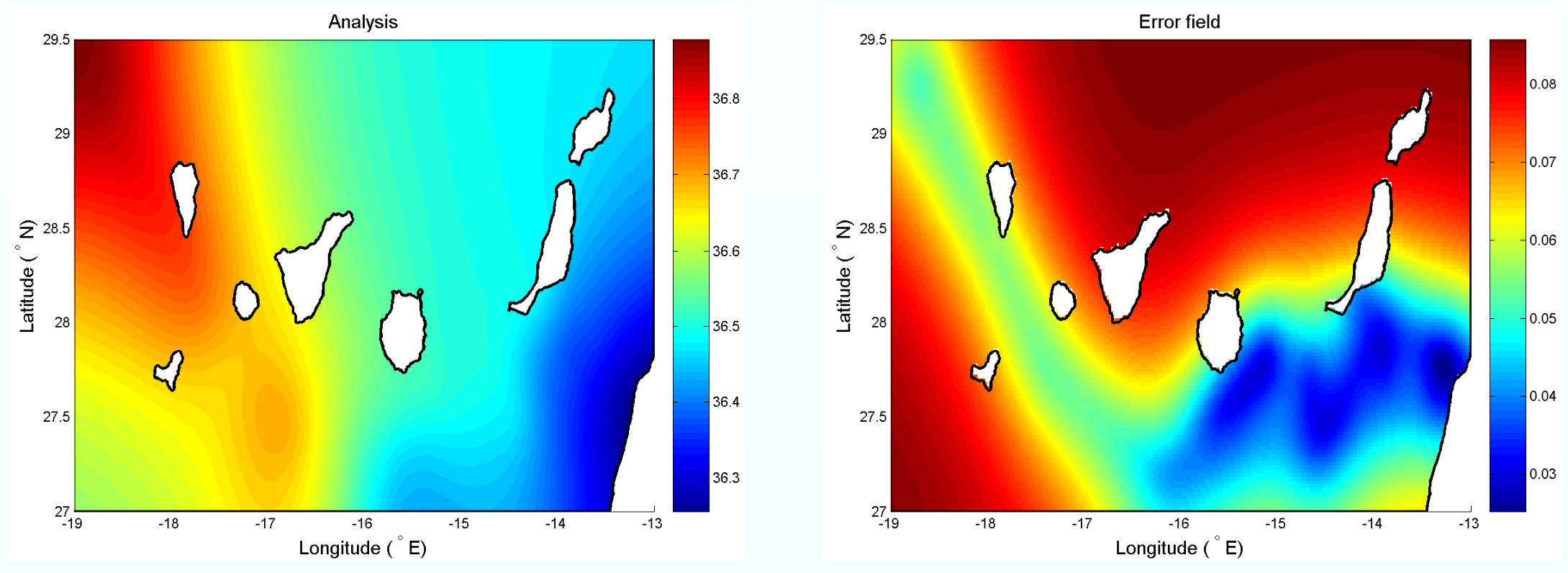

FIGURE 2: *Analysed and error fields with the data of Fig.1*

3 Additional tools

3.1 Parameters fit

The module `divafit` uses the data for a direct fitting of the covariance function. The fit is more effective when having a sufficiently large data set. The best estimates are given as output and can be used as parameter values for running Diva. Estimates of the correlation length are rather robust while those of the signal-to-noise ratio λ are neither precise nor robust, especially for large values. Here cross validation should be used.

3.2 Ordinary Cross Validation

The purpose is to optimize the parameter λ by looking for its value for which the analysis has a minimal error. For this reason, we need to find a proxy norm that we will be able to minimize. We will work with the difference of the analysed field at the data points, $\tilde{d}_i$, with respect to the original data field, d_i:

$$\theta_i^2 = (d_i - \tilde{d}_i)^2. \qquad (5)$$

Trying to minimize this norm would lead us to an infinite signal-to-noise ratio. To avoid this, we need to calculate the difference of the data value with respect to the analysed field in which the data under investigation was not taken into account. This method is called the *Ordinary Cross Validation* (OCV). The estimate is made robust by repeating the analysis over a large number of data points.

3.3 Generalized Cross Validation (GCV)

OCV is generally too expensive to be performed when working with a large number of data. According to [Craven and Wahba (1979)], modifying the error estimate as follows:

$$\hat{\theta}_i^2 = \frac{(d_i - \tilde{d}_i)^2}{(1 - A_{ii})^2}, \qquad (6)$$

allows to keep the data during the analysis and will reduce the computing cost. In this formulation, $\mathbf{A}$ is the matrix operator that provides the analysis at data points: $\tilde{\mathbf{d}} = \mathbf{Ad}$. The error estimator is made robust by taking the average over all data points, defining the *generalized cross validator* as

$$\Theta^2 = \frac{1}{N}\sum_{i=1}^{N} \hat{\theta}_i^2.$$

Assuming temporarily $\epsilon_i^2 = \epsilon^2$, hence having all misfits with the same weight, the generalized cross validator is obtained:

$$\Theta^2 = \frac{\|\,\mathbf{d} - \tilde{\mathbf{d}}\,\|^2}{N\left(1 - \frac{1}{N}\mathrm{trace}\,(\mathbf{A})\right)^2} = \frac{\|\,(\mathbf{I} - \mathbf{A})\mathbf{d}\,\|^2}{(1/N)\,(\mathrm{trace}\,(\mathbf{I} - \mathbf{A}))^2} \qquad (7)$$

For uncorrelated observational errors, we can show that the variance of expected misfit at point i is

$$\left\langle \left(d_i - \tilde{d}_i\right)^2 \right\rangle = \epsilon_i^2(1 - A_{ii}). \qquad (8)$$

Combining (7) and (8), and assuming a spatial average corresponding to a statistical expectation, we obtain an estimation of the expected variance of the noise:

$$\epsilon^2 \simeq \Theta^2\left(1 - \frac{1}{N}\mathrm{trace}\,(\mathbf{A})\right). \qquad (9)$$

When the observational errors are uncorrelated but vary in space, we should replace the residual measure $r = (\mathbf{d} - \tilde{\mathbf{d}})^{\mathsf{T}}(\mathbf{d} - \tilde{\mathbf{d}})$ by

$$r = (\mathbf{d} - \tilde{\mathbf{d}})^{\mathsf{T}}\,\hat{\mathbf{R}}^{-1}\,(\mathbf{d} - \tilde{\mathbf{d}}) \qquad (10)$$

to take into account the relative noise level; in (10), $\hat{\mathbf{R}} = \frac{1}{\epsilon^2}\mathbf{R}$, with $\mathbf{R}$, the error covariance matrix of the data.

3.4 Quality control

The purpose of *quality control* (QC) is to have at our disposal criteria to decide whether a given data is valuable or should be discarded. The criteria will involve the difference of the analysis with respect to the data and compare it with a value $\Delta_i^{(n)}$ that depends on the analysis:

$$|d_i - \tilde{d}_i| > 3\Delta_i^{(n)}. \qquad (11)$$

The actual value of the misfit can be compared with the expected standard deviation $\left\langle \left(d_i - \tilde{d}_i\right)^2 \right\rangle$, leading to:

$$\Delta_i^{(1)} = \epsilon_i\sqrt{1 - A_{ii}}, \qquad (12)$$

which is the most expensive version if A_{ii} is not explicitly known.
If A_{ii} is replaced by its average $\frac{1}{N}\mathrm{trace}\,(\mathbf{A})$, we obtain a second criterion based on:

$$\Delta_i^{(2)} = \epsilon_i\sqrt{\left(1 - \frac{1}{N}\mathrm{trace}\,(\mathbf{A})\right)}. \qquad (13)$$

This version requires only a few analysis of a random vector if trace $(\mathbf{A})$ cannot be evaluated explicitly.
Finally, in case the noise is not calculated from Θ^2, another estimate is then, according to (9)

$$\Delta_i^{(3)} = \frac{\epsilon_i}{\epsilon}\left(1 - \frac{1}{N}\mathrm{trace}\,(\mathbf{A})\right)\Theta, \qquad (14)$$

which can be calculated directly from the rms value of the misfit (or residual) and Θ. This version can easily be used simultaneously with test (13) using the results of the GCV.
Quality control corresponding to (12), (13) or (14) are respectively implemented in Diva with `divaqc`, `divaqcbis` and `divaqcter`.

4 Conclusion

We presented the theorical background of the Diva software as well as the new tools recently developed: the parameters fit, based on the generealized cross validation, and the quality control.
The updated version is available through
http://modb.oce.ulg.ac.be/modb/diva.html

Acknowledgements

We acknowledge the Fonds de la Recherche Scientique - FNRS and the Fonds pour la formation à la Recherche dans l'Industrie et dans l'Agriculture (FRIA) and the European Commission for their financial support. We specially thank J. Arístegui and P. Sangrà for providing the data used in the examples.

References

[Brasseur et al. (1996)] Brasseur, P., J.-M. Beckers, J.-M. Brankart, and R. Schoenauen. Seasonal temperature and salinity fields in the Mediterranean Sea: Climatological analyses of a historical data set. *Deep Sea Research*, **43**:159-192, 1996.

[Craven and Wahba (1979)] Craven, P. and G. Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, **31**: 377-403, 1979.

[Rixen et al. (2000)] Rixen, M., J.-M. Beckers, J.-M. Brankart, and P. Brasseur. A numerically efficient data analysis method with error map generation. *Ocean Modelling*, **2**:45-60, 2000.

Contact: Charles TROUPIN (FRIA doctoral student), ctroupin@ulg.ac.be