submitted to *Geophys. J. Int.*

Gradient-boosted equivalent sources

Santiago R. Soler[1,2] and Leonardo Uieda[3]

[1]*Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Ciudad Autónoma de Buenos Aires, Argentina (santiago.*

[2]*Instituto Geofísico Sismológico Volponi, Universidad Nacional de San Juan, San Juan, Argentina*

[3]*Department of Earth, Ocean and Ecological Sciences, School of Environmental Sciences, University of Liverpool, UK*

SUMMARY

The equivalent source technique is a powerful and widely used method for processing gravity and magnetic data. Nevertheless, its major drawback is the large computational cost in terms of processing time and computer memory. We present two techniques for reducing the computational cost of equivalent source processing: block-averaging source locations and the gradient-boosted equivalent source algorithm. Through block-averaging, we reduce the number of source coefficients that must be estimated while retaining the minimum desired resolution in the final processed data. With the gradient boosting method, we estimate the sources coefficients in small batches along overlapping windows, allowing us to reduce the computer memory requirements arbitrarily to conform to the constraints of the available hardware. We show that the combination of block-averaging and gradient-boosted equivalent sources is capable of producing accurate interpolations through tests against synthetic data. Moreover, we demonstrate the feasibility of our method by gridding a gravity dataset covering Australia with over 1.7 million observations using a modest personal computer.

Key words: Gravity anomalies and Earth structure; Magnetic anomalies: modelling and interpretation; Geopotential theory; Inverse theory; Statistical methods; Australia.

1 INTRODUCTION

Measurements of anomalies in potential fields, like gravity disturbances and total-field magnetic anomalies, are widely used in geophysical exploration for their low cost of acquisition. These data can be surveyed using ground, airborne, shipborne, or satellite systems. During ground surveys, the data are often gathered following irregular paths or networks along the surface of the terrain, leading to highly variable elevations in mountainous regions. Airborne and satellite surveys gather data along flight lines, producing closely spaced measurements along almost straight lines but with larger spacing between adjacent lines. Measurement height can also change because of the vertical movement of the aircraft. Processing of the data often involves interpolation onto a regular grid at constant height, both to improve visualization for interpretation purposes as well as to prepare the data for further processing and modelling (e.g., reduction-to-the-pole, derivative calculations, upward continuation, Euler deconvolution).

Several methods exist in the literature for interpolation in two dimensions, for example continuous curvature splines in tension (Smith & Wessel 1990), bi-harmonic (thin-plate) splines (Sandwell 1987), and kriging (Hansen 1993). These general-purpose methods have limitations when it comes to interpolating potential field data, namely (i) they are not able to take into account the variable height of the observation points and (ii) the interpolating functions are not necessarily harmonic, which is the underlying assumption behind many processing techniques (e.g., upward continuation and vertical derivatives).

A widely used method for interpolating gravity and magnetic data is the equivalent sources technique (also known as equivalent layer, radial basis functions, or Green's functions interpolation). First introduced by Dampney (1969), the method consists in fitting a model of finite elementary sources to the data and using this model to predict new data values. Besides interpolation, equivalent sources have been used for reduction-to-the-pole of magnetic data (Silva 1986; Nakatsuka & Okuma 2006; Guspí & Novara 2009), upward continuation (Emilia 1973; Li & Oldenburg 2010), joint processing of gravity gradient data (Barnes & Lumley 2011), modelling the lithospheric magnetic field (Kother et al. 2015), recovering the magnetic induction vector from total-field magnetic anomalies (Li et al. 2020), and more.

It is also worth mentioning the least-squares collocation method (LSC), which is widely used in geodesy (Tscherning 2015, and references therein). LCS is often applied to combine and interpolate different linear functionals of the disturbing gravity potential (gravity anomalies, gravity disturbances, deflections of the vertical, geoid height, et cetera). Like equivalent sources, collocation also requires the solution of a large linear system of the order of the number of observed data. As such, it's practical application suffers from the same computational challenges.

Many variants of the equivalent sources technique have been proposed, often attempting to obtain faster or more accurate solutions. The key factors that vary between them are: (i) the type of source, (ii) the location of the sources, and (iii) the solution strategy.

The most commonly used type of source is a point mass for gravity or dipole for magnetics (e.g., von Frese et al. 1981; Silva 1986; Mendonça & Silva 1994; Siqueira et al. 2017). However, right-rectangular prisms (e.g., Barnes & Lumley 2011; Jirigalatu & Ebbing 2019; Li et al. 2020) and tesseroids (Bouman et al. 2016) have also been used successfully. In fact, even point sources with a simple inverse distance function, instead of actual gravity or magnetic fields, can be used as equivalent sources (Cordell 1992).

The location of sources often follows one of two strategies. The most common approach is to distribute sources on a regular grid at a constant depth (e.g., Leão & Silva 1989; Barnes & Lumley 2011; Oliveira et al. 2013). Alternatively, sources can be placed beneath each data point (e.g., Cordell 1992; Siqueira et al. 2017). Some recent work by Li et al. (2020) places the sources in two overlapping layers at different depths.

The coefficients of the equivalent source model are often estimated through damped least-squares. This imposes a heavy computational load when the number of data points is large (e.g., airborne and satellite surveys). To reduce the computational load, Mendonça & Silva (1994) built the solution iteratively by incorporating one data point at a time using the "equivalent data concept". Leão & Silva (1989) processed the input data using a moving window, only fitting the data inside the window and predicting observations at its center. Li & Oldenburg (2010) and Barnes & Lumley (2011) apply different operations to generate a sparse representation of the sensitive matrix (respectively, wavelet compression and quadtree discretization), which significantly improves the speed of the least-squares solution. Oliveira et al. (2013) parametrized the equivalent layer as a piecewise bivariate polynomial function, reducing the number of parameters in the solution. Siqueira et al. (2017) developed an iterative solution in which the sensitivity matrix is transformed into a diagonal matrix with constant terms through the "excess mass criterion". Jirigalatu & Ebbing (2019) applied the Gauss-FFT method to speed up the forward modelling operations and solved the least-squares problem using steepest descent to avoid calculating the Hessian matrix and solving linear systems.

Many of the existing methods solve under-determined problems, requiring a much larger number of equivalent sources than the number of data points. Some achieve greater efficiency by restricting their applications to specific data types (Siqueira et al. 2017), interpolating only on regular grids (Leão & Silva 1989), or requiring already gridded data (Takahashi et al. 2020), to name a few. Furthermore, many of the optimizations proposed are also complex to implement in a computer program, limiting their wider adoption.

In the present study, we propose two strategies for reducing the computational load of the equivalent sources technique:

(i) Reduce the number of equivalent sources for oversampled surveys through a *block-averaging* strategy while maintaining the quality of the solution.

(ii) Fit the equivalent source model iteratively along overlapping windows using a *gradient boosting* algorithm (Friedman 2001).

The first strategy consists in dividing the survey area into horizontal blocks and assigning a single source to each block, located at the median horizontal location of the data points. For airborne, shipborne, and satellite surveys, which are oversampled along tracks, this can greatly reduce the size of the inverse problem while retaining the same quality of interpolation.

The gradient boosting algorithm allows us to fit the equivalent source model iteratively by operating on individual overlapping windows. As a result, our method solves several much smaller least-squares problems instead of a large one. This has some similarities with the strategy used by Leão & Silva (1989) but without the requirement for sources and predictions to be on regular grids.

Through tests on synthetic data, we show that: (i) the *block-averaged* sources are able to achieve the same accuracy as other traditional equivalent source layouts while using a fraction of the number of sources, and (ii) the *gradient boosting* algorithm greatly reduces the computational memory required to fit very large datasets without sacrificing prediction accuracy. Finally, a combination of both strategies is used to process a collection of approximately 1.7 million ground gravity data measurements from Australia.

2 METHODOLOGY

2.1 The equivalent sources technique

We will follow the "generalized equivalent sources" of Cordell (1992) and assume that any harmonic function $d(\mathbf{p})$ can be approximated by a sum of M discrete point source effects

$$d(\mathbf{p}) = \sum_{j=1}^{M} \frac{c_j}{\|\mathbf{p} - \mathbf{q}_j\|} \,, \tag{1}$$

in which $\mathbf{p}$ and $\mathbf{q}_j$ are, respectively, the position vectors in a 3D Cartesian space of data and sources, c_j is a scalar coefficient related to the point source located at $\mathbf{q}_j$, and $\|\cdot\|$ represents the L_2 norm. The horizontal and vertical distribution of sources is discussed in section 2.4.

In case we have values of the harmonic function at N discrete points $\{\mathbf{p}_1\ \mathbf{p}_2\ \ldots\ \mathbf{p}_N\}$, we can write a set of N equations of the form

$$d_i = \sum_{j=1}^{M} \frac{c_j}{\|\mathbf{p}_i - \mathbf{q}_j\|} \quad \forall i = 1, 2, \ldots, N \;, \tag{2}$$

where d_i is the calculated value at point $\mathbf{p}_i$. These equations can also be expressed in matrix form as

$$\mathbf{d} = \mathbf{A}\mathbf{c} \;, \tag{3}$$

where $\mathbf{d}$ is a column vector containing the N predicted values at the observation points, $\mathbf{c}$ is a column vector containing the M coefficients c_j, and $\mathbf{A}$ is the $N \times M$ Jacobian matrix, whose elements are

$$a_{ij} = \frac{1}{\|\mathbf{p}_i - \mathbf{q}_j\|} \tag{4}$$

For a given set of N observed data $\mathbf{d}^o$, we can find a least-squares solution to Eq. 3 and obtain the values of $\mathbf{c}$ that best fit the observations. These coefficients can, in turn, be used to predict the value of the harmonic function at any other point outside of the sources by evaluating Eq. 1. Gridding and upward continuation can thus be achieved by predicting values on points that fall on a regular grid or at different heights, respectively.

2.2 Damped least-squares solution

We can obtain the values of the source coefficients $\mathbf{c}$ that best fit the observed field values $\mathbf{d}^o$ by minimizing the goal function

$$\phi(\mathbf{c}) = [\mathbf{d}^o - \mathbf{A}\mathbf{c}]^{\mathrm{T}} \, \mathbf{W} \, [\mathbf{d}^o - \mathbf{A}\mathbf{c}] + \lambda_d \, \mathbf{c}^{\mathrm{T}}\mathbf{c} \;, \tag{5}$$

where $\mathbf{W}$ is a $N \times N$ diagonal matrix of data weights and λ_d is a positive *damping* parameter with the same units as the Jacobian matrix elements. The second term on the right-hand side of Eq. 5 is the zeroth-order Tikhonov regularization (Tikhonov 1977), also known as a damping regularization, that is used to stabilize the solution.

The damping parameter controls the amount of regularization that will be applied. An overly large value would generate a smooth solution that fails to reproduce the high frequency components of the data, while an overly small value would result in over-fitting, thus failing to produce realistic interpolation results (Martinez & Li 2016). The range of acceptable values for the damping parameter λ_d will depend on the values of the Jacobian matrix $\mathbf{A}$ and the coefficients. Consequently, this range will vary (often dramatically) between datasets, making it difficult to choose an appropriate value in practice.

To solve this issue, we first scale the Jacobian matrix so that its elements are dimensionless and each column has unit variance. We define a diagonal matrix $\mathbf{S}$

$$\mathbf{S} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_M \end{bmatrix}_{M \times M} , \tag{6}$$

in which σ_j is the standard deviation of the j-th column of $\mathbf{A}$. We then write the forward problem in Eq. 3 as

$$\mathbf{d} = \mathbf{A}\mathbf{S}^{-1}\mathbf{S}\mathbf{c} = \left[\mathbf{A}\mathbf{S}^{-1}\right]\left[\mathbf{S}\mathbf{c}\right] = \mathbf{B}\mathbf{m} \tag{7}$$

where $\mathbf{B} = \mathbf{A}\mathbf{S}^{-1}$ is the scaled and dimensionless Jacobian matrix and $\mathbf{m} = \mathbf{S}\mathbf{c}$ is a vector containing scaled coefficients with the same units as the data.

The goal function defined in Eq. 5 can be rewritten as

$$\phi(\mathbf{m}) = \left[\mathbf{d}^o - \mathbf{B}\mathbf{m}\right]^{\mathrm{T}} \mathbf{W} \left[\mathbf{d}^o - \mathbf{B}\mathbf{m}\right] + \lambda\, \mathbf{m}^{\mathrm{T}}\mathbf{m} \,, \tag{8}$$

where λ is a *dimensionless* damping parameter and regularization is applied on the scaled coefficients $\mathbf{m}$ instead of $\mathbf{c}$. Using a dimensionless damping parameter allows us to narrow the range of values of λ that would generate the most accurate predictions, irrespective of the dataset and its units. From experience, we recommend searching for suitable λ values between 10^{-6} and 10^{4} varying by order-of-magnitude. The choice of the damping and other hyper-parameters, like the source depth, could be done through well-established statistical methods, such as cross-validation.

The vector of scaled coefficients $\hat{\mathbf{m}}$ that minimizes the goal function can be found by solving the *normal equation system* (Menke 1989)

$$\left[\mathbf{B}^{\mathrm{T}}\mathbf{W}\mathbf{B} + \lambda\mathbf{I}\right] \hat{\mathbf{m}} = \mathbf{B}^{\mathrm{T}}\mathbf{W}\mathbf{d}^o. \tag{9}$$

Once the scaled coefficients are obtained, the estimated unscaled coefficients $\hat{\mathbf{c}}$ can be calculated by removing the scaling factor

$$\hat{\mathbf{c}} = \mathbf{S}^{-1}\hat{\mathbf{m}} \,. \tag{10}$$

The forward modeling operations used to perform predictions (e.g., for interpolation and upward continuation) are left unchanged by using vector $\hat{\mathbf{c}}$ instead of $\hat{\mathbf{m}}$.

2.3 Gradient boosting

Gradient boosting was first introduced by Friedman (2001, 2002) as a method for fitting additive parametric models of the form

$$d = \sum_{k=1}^{K} \alpha_k f(\mathbf{c}_k), \tag{11}$$

where α_k is a scalar coefficient called the *step-size* and f is a function of the parameter vector $\mathbf{c}_k$. For linear problems, these additive models can be written as the matrix equation

$$\mathbf{d} = \sum_{k=1}^{K} \mathbf{A}_k \mathbf{c}_k \; . \tag{12}$$

Because of the linearity of the $f(\mathbf{c}_k)$ functions, the α_k step-size parameters can be incorporated into the parameter vector $\mathbf{c}_k$.

We can transform our equivalent source problem in Eq. 3 into an additive model by following these steps:

(i) Define a set of M equivalent sources distributed throughout the survey area (see section 2.4 for details).

(ii) Define a set of K overlapping windows of equal size that cover the survey area.

(iii) Create K separate sets of equivalent sources, one for each window. Each set will be formed by the portion of the original M sources that fall inside the respective window. Since the windows overlap, the total number of sources from all sets will be greater than M.

(iv) Define vector $\mathbf{c}_k$ as the M_k coefficients of the equivalent sources of the k-th window.

(v) Define matrix $\mathbf{A}_k$ as the $N \times M_k$ Jacobian matrix between the sources in the k-th window and all N data points of the survey.

(vi) Model the predicted data as a superposition of the effects of the K separate sets of equivalent sources (i.e., Eq. 12).

The gradient boosting algorithm works by fitting each component of the additive model, one at a time, to the residuals of the previous component. Friedman (2001) demonstrates that this corresponds to a steepest-descent optimization in the so-called "function space". The adaptation of the gradient

boosting method to find the damped least-squares solutions for the K parameter vectors $\mathbf{c}_k$ in Eq. 12 is presented in Algorithm 1.

Algorithm 1: Gradient boosting solution for damped least-squares regression.

1 Define the residual vector $\mathbf{r}_0 = \mathbf{d}^o$
2 **for** $k = 1$ **to** K **do**
3 Calculate the $N \times M_k$ Jacobian matrix $\mathbf{A}_k$
4 $\mathbf{B}_k = \mathbf{A}_k \mathbf{S}_k^{-1}$
5 $\hat{\mathbf{m}}_k = \left[\mathbf{B}_k^\mathrm{T} \mathbf{W}_k \mathbf{B}_k + \lambda \mathbf{I}\right]^{-1} \mathbf{B}_k^\mathrm{T} \mathbf{W}_k \mathbf{r}_{k-1}$
6 $\hat{\mathbf{c}}_k = \mathbf{S}_k^{-1} \hat{\mathbf{m}}_k$
7 $\mathbf{d}_k = \mathbf{A}_k \hat{\mathbf{c}}_k$
8 $\mathbf{r}_k = \mathbf{r}_{k-1} - \mathbf{d}_k$
9 **end for**

After all $\mathbf{c}_k$ coefficients vectors are estimated, we can predict the effect of the additive equivalent source model on any point through the summation

$$d(\mathbf{p}) = \sum_{k=1}^{K} \sum_{j=1}^{M_k} \frac{c_{kj}}{\left\|\mathbf{p} - \mathbf{q}_{kj}\right\|} , \tag{13}$$

in which c_{kj} is the j-th element of the $\mathbf{c}_k$ vector and the $\mathbf{q}_{kj}$ is the position vector of the j-th source of the k-th window.

To improve the convergence of the algorithm, Friedman (2002) suggests introducing randomness into the fitting process. We achieve this by randomizing the order in which the K windows are used in the gradient boosting algorithm. Section 3.3 explores the effect of randomization in the convergence rate of the algorithm and the accuracy of the interpolation.

The $\mathbf{A}_k$ matrices have only $N \times M_k$ elements (where M_k is the number sources on the $k-$th window), which can be considerably smaller than the $N \times M$ elements of $\mathbf{A}$. Therefore, the gradient boosting method allows us to fit equivalent source models that would produce Jacobian matrices that are larger than the available computer memory. Furthermore, we can increase or decrease the size of the overlapping windows as needed depending on the number of sources in the model and the available computer memory.

We can improve the efficiency of the algorithm further by:

(i) Using only the N_k data points that fall within the k-th window for fitting the sources (steps 4 and 5 of algorithm 1). By doing so, we can replace the $N \times M_k$ Jacobian matrix $\mathbf{A}_k$ with the smaller

$N_k \times M_k$ matrix $\tilde{\mathbf{A}}_k$. We still use all N data points when calculating the predicted data and residuals (steps 7 and 8 of algorithm 1).

(ii) The forward modeling operation performed in step 7 can be done by a summation (Eq. 2) instead of a matrix-vector product, which allows us to avoid computing and storing the larger $N \times M_k$ matrix $\mathbf{A}_k$ at any point.

Algorithm 2 is the final *gradient-boosted equivalent sources algorithm* which incorporates these changes. Figure 1 shows a sketch of the algorithm steps applied a set of observation points that simulate a ground survey and locating one source below each data point.

Algorithm 2: Gradient-boosted equivalent sources algorithm.

Define the residual vector $\mathbf{r}_0 = \mathbf{d}^o$
for $k = 1$ **to** K **do**
  Select weights $\tilde{\mathbf{W}}_k$ and residuals $\tilde{\mathbf{r}}_{k-1}$ for data points inside the k-th window
  Calculate Jacobian matrix $\tilde{\mathbf{A}}_k$ with data points and sources inside the k-th window
  $\mathbf{B}_k = \tilde{\mathbf{A}}_k \mathbf{S}_k^{-1}$
  $\hat{\mathbf{m}}_k = \left[\mathbf{B}_k^{\mathrm{T}} \tilde{\mathbf{W}}_k \mathbf{B}_k + \lambda \mathbf{I}\right]^{-1} \mathbf{B}_k^{\mathrm{T}} \tilde{\mathbf{W}}_k \tilde{\mathbf{r}}_{k-1}$
  $\hat{\mathbf{c}}_k = \mathbf{S}_k^{-1} \hat{\mathbf{m}}_k$
  Calculate $\mathbf{d}_k$, where $d_{ki} = \sum_{j=1}^{M_k} \frac{c_{kj}}{\left\|\mathbf{p}_i - \mathbf{q}_{kj}\right\|} \quad \forall\, i = 1 \text{ to } N$
  $\mathbf{r}_k = \mathbf{r}_{k-1} - \mathbf{d}_k$
end for

It is worth noting that two sets of equivalent sources obtained through two adjacent overlapping windows have some portion of the sources on the same locations, specifically the ones that fall on the intersection between the two windows. We can interpret this as the gradient-boosting algorithm fitting the source coefficients multiple times: one time for every window that covers each source. This fact can be exploited in order to save computer memory. Instead of storing all of the $\mathbf{c}_k$ vectors (Eq. 12), we can initialize a single $\mathbf{c}$ vector with zeros, where each element represents the coefficient of each one of the original M sources. After each iteration of the gradient-boosting algorithm, we add the estimated coefficients $\hat{\mathbf{c}}_k$ to the corresponding elements of vector $\mathbf{c}$. Because the forward modelling function is linear, we can safely compute the resulting field through Eq. 1 instead of Eq. 13. This way, the memory needed to store the entire set of estimated coefficients is limited to a single vector of M elements.

Our gradient boosting algorithm for overlapping windows is similar to the "bootstrap inversion" used in von Frese et al. (1988), which also iteratively fits portions of an equivalent source model to the

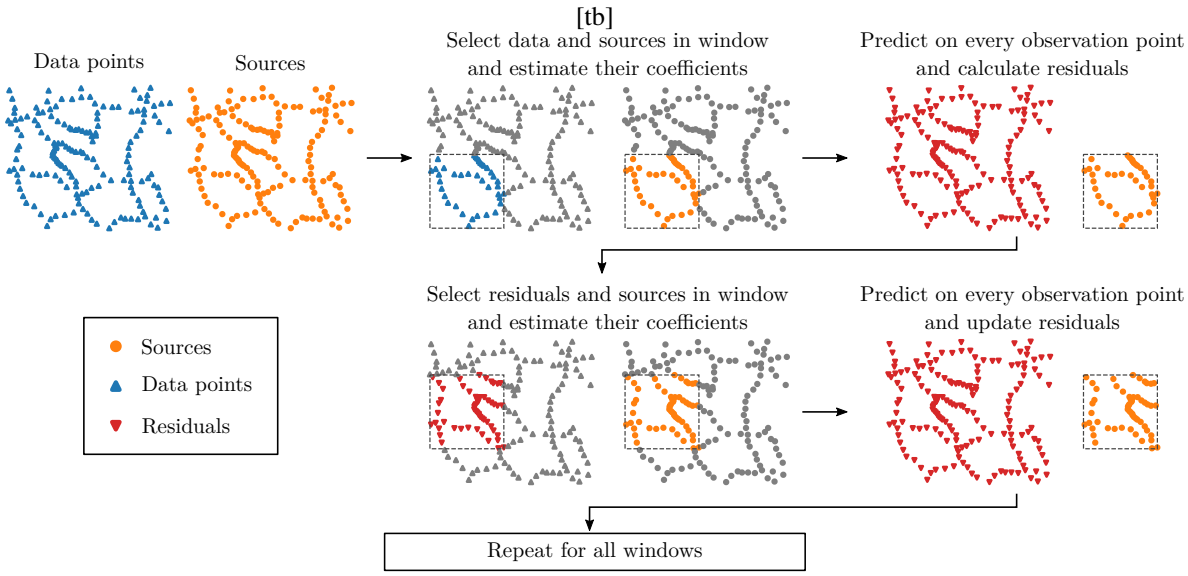


Figure 1. Sketch of the gradient-boosted equivalent source algorithm. Data points are represented by blue upwards-facing triangles, equivalent sources by orange dots, data residuals by red downwards-facing triangles, and the current window by black dashed lines. The algorithm starts by selecting the data and sources inside the first window and estimating the source coefficients using the selected data points. Then, the effect of the estimated sources is predicted on all data points and used to calculate the residuals. Another window is used to select residuals and sources and estimate the coefficients using the selected residuals instead of the original data. Again, the effect of the estimated sources is predicted on all data points and the residuals are updated. These steps are repeated for every window in a randomized order.

data residuals. The key differences are that in our method: (i) the sources in the overlapping portions of the windows are fitted more than once, allowing the algorithm to self-correct for poor solutions to any given window; (ii) we use only data points within the window when fitting, what enables the use of larger datasets.

2.4 Location of sources

The ideal number of sources and their locations, both horizontally and vertically, has been debated since the inception of the equivalent sources technique with Dampney (1969). The choices made regarding these parameters can play an important role on the accuracy of the predictions and the computational resources needed to estimate the source coefficients. An ideal distribution of sources should simultaneously be able to reproduce the measured data on the survey points, make accurate predictions on non-surveyed locations, and minimize the required computational resources.

A large number of evenly distributed sources along the survey region are capable of reproducing the observed data. Nevertheless, the computational load can be prohibitive and such underdetermined problems are prone to overfitting the data, leading to poor predictive power when interpolating and extrapolating. On the other hand, using few sources will reduce the computational requirements but the model may be incapable of reproducing the full spectral content of the measured data.

[tb]

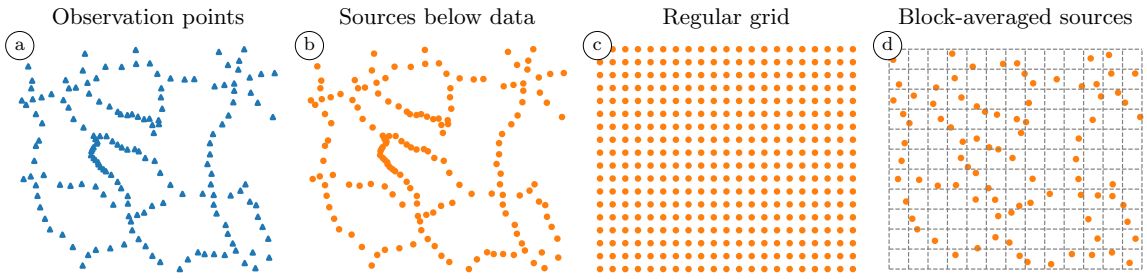


Figure 2. Sketch of different horizontal layouts for equivalent source models. Blue points represent the locations of observations and orange points represent the locations of equivalent sources according to different layout strategies. (a) Set of 166 observation points that simulate a ground survey. (b) Location of the 166 sources obtained through the *sources below data* layout. (c) Location of the 378 sources obtained through the *regular grid* layout. (d) Location of the 87 sources obtained through the *block-averaged sources* layout. Grey dashed lines represent the spatial blocks within which the median observation location is calculated.

Particular survey characteristics also play a role in the choice of equivalent source distribution. In a ground survey, observations are usually located along irregular paths and scattered points. The coverage of the survey region is often uneven, leaving large areas without any observation. On the other hand, observations from airborne surveys are located along almost straight and closely spaced flight lines. Measurements are usually taken at a high temporal frequency, leading to observation points along the flight lines that are several times closer to each other than the flight line spacing. This creates a bias in the sampling, which can cause aliasing artifacts in gridded products.

2.4.1 Horizontal source layouts

The most widely used layouts for distributing equivalent sources horizontally are:

(i) *Sources below data points*: one equivalent source is placed at the horizontal location of each data point (Fig. 2b). Therefore, the number of sources is equal to the number of observations ($M = N$).

(ii) *Regular grid*: a homogeneous distribution of point sources below the survey region (Fig. 2c). A padding region is often added to help reduce edge effects. In practice, it often leads to underdetermined problems since a large number of sources is required ($M > N$).

For ground surveys, the *regular grid* layout needs a sufficiently small grid spacing to be able to fit the observed data. This creates an unnecessarily large number of sources in areas where no observations exist. In contrast, the *sources below data* layout is more likely to accurately fit the observed data with many fewer sources, reducing the computational load. But when applied to airborne surveys, the *sources below data* layout may place an undesirably large number of sources along the flight paths. This could lead to aliasing effects on the predicted values, such as the stripes parallel to flight lines that are often observed when gridding airborne magnetic data. The *regular grid* layout can avoid this effect

[tb]

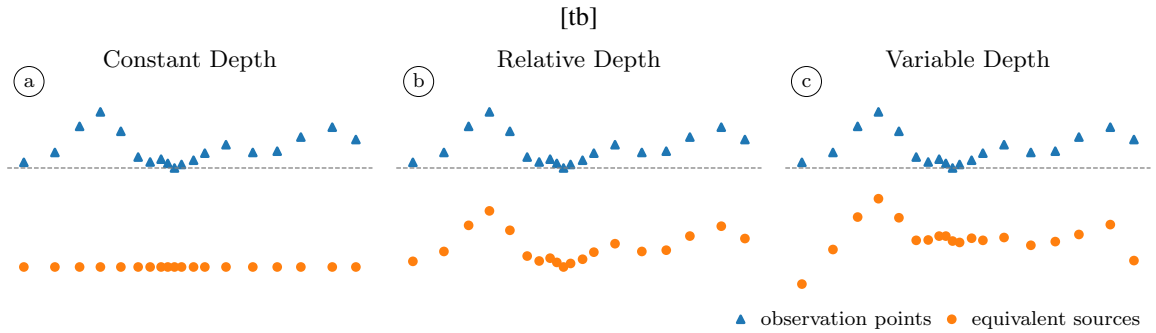


Figure 3. Examples of different strategies for assigning depths to equivalent sources. Here we assign one source for each observation point, located at the same horizontal coordinates as the data points. Source depths are (a) a *constant depth* at a chosen vertical coordinate, (b) a *relative depth* determined by uniformly shifting downward the vertical coordinate of data points, and (c) a *variable depth* determined by shifting the vertical coordinates of the observation points by an amount proportional to the average distance to neighbouring sources. The distance between data points and their respective sources (a) depends on observation height, (b) is constant, and (c) is proportional to the horizontal distribution of sources. Notice how the closely spaced sources in the middle of the profile (c) are shallower than their counterparts in (b).

by evenly distributing sources and using a continuous source layer (e.g., right-rectangular prisms or tesseroids).

We propose a new way of distributing equivalent sources horizontally that could simultaneously reduce the computational load and mitigate some of the drawbacks of existing layouts. In the *block-averaged sources* layout, point sources are placed in the average position of data points that fall within specified spatial blocks (Fig. 2d). This is done by:

(i) Dividing the survey region into rectangular blocks of equal size.

(ii) Computing the median horizontal position of the observation points that fall inside each block. Blocks without any observation point are omitted.

(iii) Assign one point source to each of the median horizontal positions calculated in step (ii).

The number of sources created by this new layout will be less than the number of observations if the block size is chosen appropriately (i.e., making sure that blocks are large enough to contain more than a single data point). The overdetermined problem that arises from this layout has a lower computational load and is less prone to overfitting the data since the model complexity is lower. Moreover, the block averaging process can balance the spacing between sources along a flight line and between adjacent lines, helping to reduce aliasing effects in the generated grids. In Section 3.1, we demonstrate through tests on synthetic data that the block-averaged sources layout is able to interpolate with comparable accuracy to other layouts while using a fraction of the equivalent sources.

2.4.2 Depth of sources

It is widely known from potential theory that the depth of a point source influences the wavelength of the observed field at the surface. This makes the source depth a key parameter affecting the outcome of interpolation and other operations done with equivalent sources. Several different strategies for assigning the depths of equivalent sources have been proposed in the literature. Here, we will highlight the following (Fig. 3):

(i) *Constant depth*: The simplest option is to locate all sources at the same depth (Fig. 3a). If the measurements were taken at significantly different altitudes, some measurements will be more distant to the sources than others, which may create problems for reproducing short wavelengths in high altitude points.

(ii) *Relative depth*: The depths of sources are determined by shifting the vertical coordinate of data points downward by a fixed amount (Fig. 3b). The sources will not all be at the same vertical coordinate, but they will all be at the same vertical distance from the observation points.

(iii) *Variable depth*: The depths of sources are proportional to the horizontal distance to the nearest neighbouring data points or sources (Fig. 3c). Different variations of this strategy have been proposed before, for example Cordell (1992), Guspí et al. (2004), and Guspí & Novara (2009). The rationale for this strategy is that if a survey has data points clustered in some areas, we may want the sources below those areas to be shallower in order to preserve the shorter wavelengths that can be measured.

Our approach to the *variable depth* strategy will be:

$$z = z_{obs} + \Delta z + \alpha h, \tag{14}$$

in which z is the vertical coordinate (positive downwards) of an equivalent source, Δz is a relative depth shift that is the same for all sources, α is an dimensionless depth factor, h is the median horizontal distance to the k nearest neighbouring sources, and z_{obs} is a vertical observation coordinate that will depend on the horizontal layout strategy. For *sources below data*, it is the vertical coordinate of the data point corresponding to the given source. For *regular grid*, it can be interpolated from the vertical coordinates of all data points. Finally, for *block-averaged sources* it will be the median vertical coordinate of the data within the corresponding block.

In Section 3.1, we test the effectiveness each of these strategies on synthetic data.

[tb]

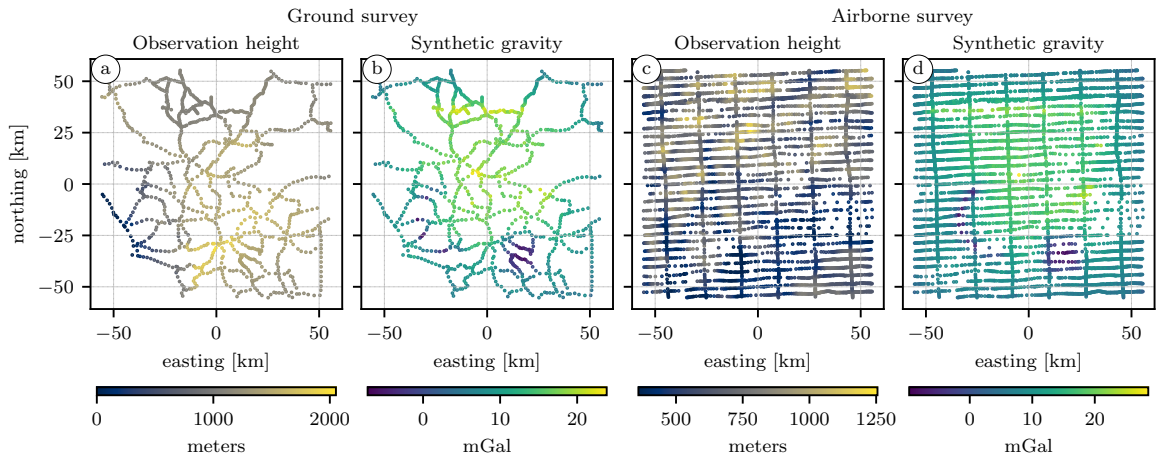


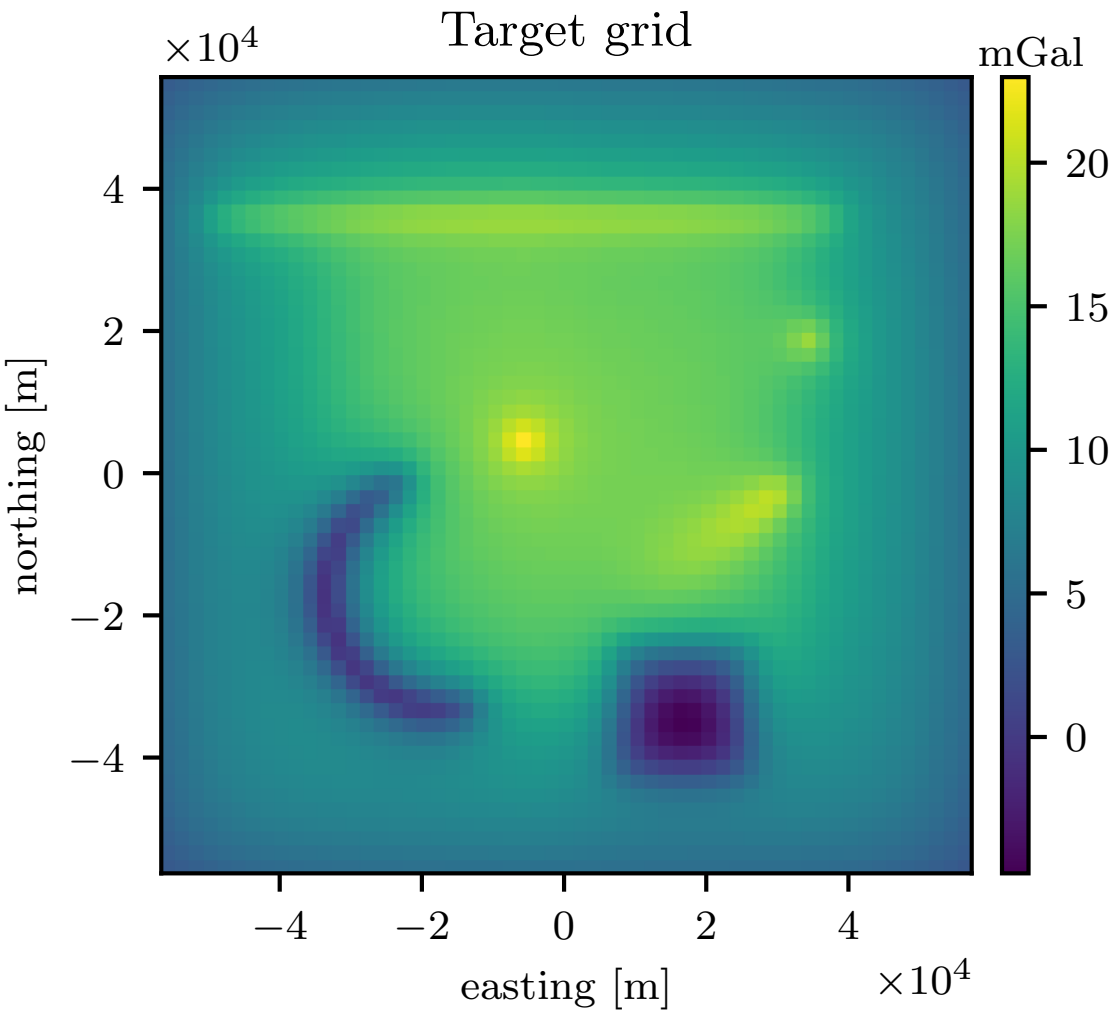
Figure 4. Observation heights and gravity values for the synthetic ground (a-b) and airborne (c-d) surveys. Heights are given in meters above the zero height plane. The synthetic gravity data are contaminated with pseudo-random Gaussian noise with zero mean and 1 mGal standard deviation.

3 TESTS ON SYNTHETIC DATA

We have used synthetic gravity datasets to test the interpolation accuracy of the difference horizontal and vertical source distribution strategies as well as the gradient-boosted equivalent sources method. To generate the data, we created a model of 64 right-rectangular prisms, distributed in a 111319 m×111319 m area with depths varying between 10000 m and zero. The density contrast of prisms ranges from $-900\,\mathrm{kg\,m^{-3}}$ to $500\,\mathrm{kg\,m^{-3}}$. The model includes prisms of different shapes, sizes, and depths to create gravity disturbances with a variety of wavelengths.

We created two synthetic datasets from the model, one simulating a ground survey and another an airborne acquisition (Fig. 4). To create the synthetic ground survey, we selected measurement positions from a portion of a public domain gravity dataset for Southern Africa, available through the NOAA National Centers for Environmental Information (NCEI). For the synthetic airborne survey, we used a portion of the Great Britain Aeromagnetic Survey acquired by Hunting Geology and Geophysics Ltd and Canadian Aeroservices Ltd between 1955 and 1965 and made publicly available by the British Geological Survey (BGS). In both cases, we rescaled the horizontal coordinates of each survey portion to span an area of 111319 m×110576 m, matching the model dimensions. The ground survey contains 963 observations distributed at heights between 0 m and 2052.2 m (Fig. 4a). The airborne survey has 5673 observations at heights between 359 m and 1255 m (Fig. 4c).

The vertical component of the gravitational acceleration generated by the model was computed using the method of Nagy et al. (2000, 2002) with recent modifications by Fukushima (2020), as implemented in the open-source software Harmonica (Uieda et al. 2020b). We generated a *target grid* of 57×56 points with a spacing of 2 km and located 2000 m above the zero height plane (Fig. 5)



[tb]

Figure 5. Pseudo-color map of the target grid of synthetic gravity data. The grid is composed of 57×56 points with a spacing of 2 km. The grid height is 2000 m above the zero height plane.

to serve as a reference when calculating the interpolation error. We then generated synthetic ground (Fig. 4b) and airborne (Fig. 4d) data to which we added pseudo-random Gaussian noise with zero mean and 1 mGal standard deviation.

3.1 Source distribution strategies

We investigated the effect on interpolation accuracy of different strategies for distributing the equivalent sources horizontally and vertically. To do this, we used the damped least-squares solution described in Section 2.2 (without gradient boosting) to interpolate the synthetic datasets (Fig. 4) and compared the results against the target grid (Fig. 5). This process was repeated for each combination of horizontal layout (*sources below data* and *block-averaged sources*) and depth type (*constant*,

[p]

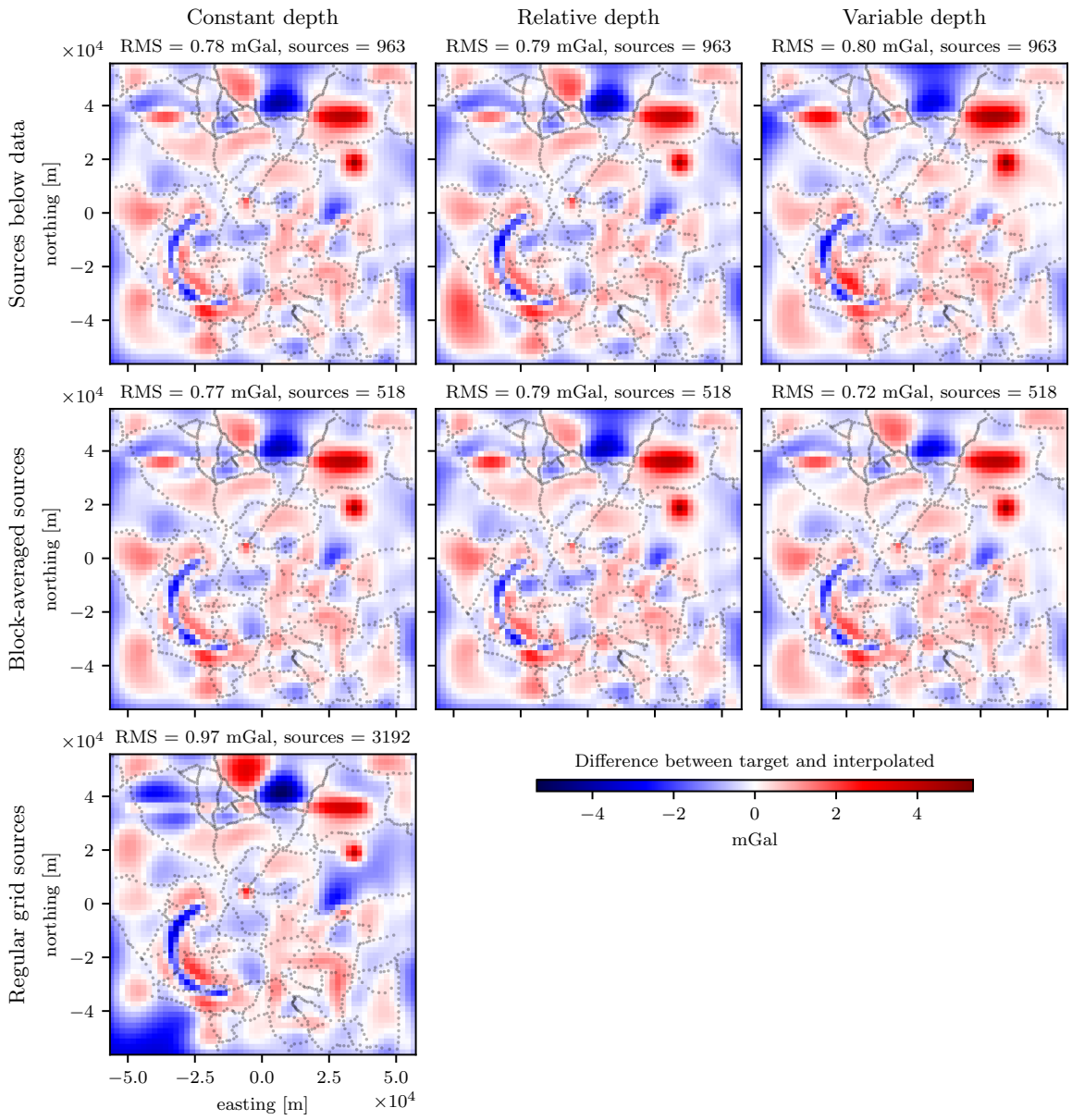


Figure 6. Pseudo-color maps of the differences between the target grid and the interpolated synthetic ground survey data produced by each source distribution strategy. The black dots represent the horizontal location of the synthetic data points. The RMS error and total number of equivalent sources is reported for each strategy at the top of the respective maps.

relative, and *variable*) and for regular grid sources with a constant depth, totalling 7 different combinations.

Each source distribution strategy requires certain hyper-parameters to be chosen in order to build the set of point sources. For example, using a constant depth needs the definition of the depth and using block-averaged sources requires the definition of the block size. The predictive capabilities of the equivalent sources depend on the choice of these hyper-parameters. To ensure that our comparisons

[p]

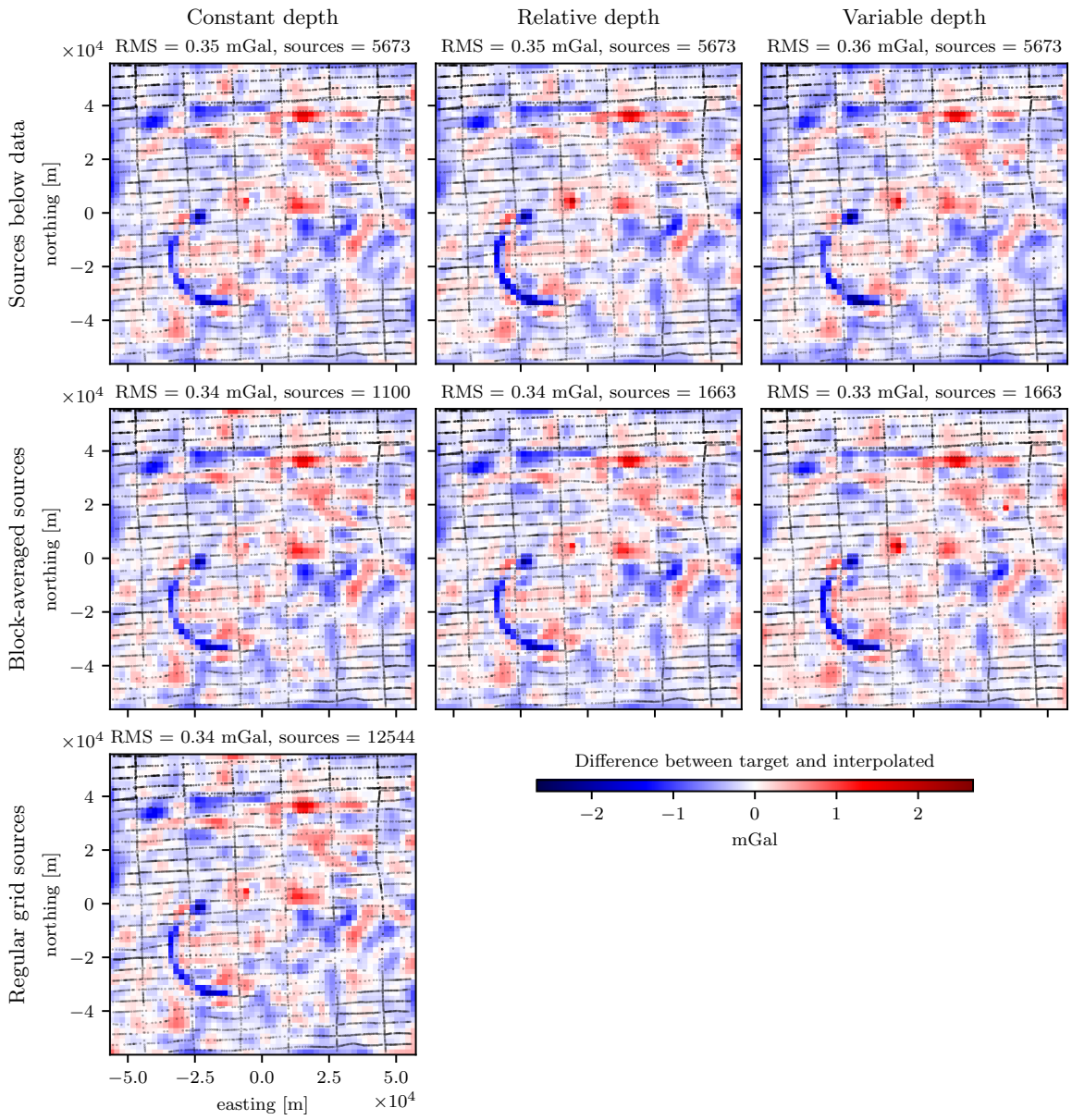


Figure 7. Pseudo-color maps of the differences between the target grid and the interpolated synthetic airborne survey data produced by each source distribution strategy. The black dots represent the horizontal location of the synthetic data points. The RMS error and total number of equivalent sources is reported for each strategy at the top of the respective maps.

are fair, we perform an exhaustive search over combinations of hyper-parameter values (including the damping parameter from Eq. 8) to obtain the best prediction that can be achieved by each source distribution strategy. Here, the best prediction is defined as the one that minimizes the root mean-square error (RMS) between interpolated values and the target grid (Fig. 5). The parameter values used in these searches and the one producing the smallest RMS error are outlined in Tables A1 and A2.

Fig. 6 and 7 show the differences between the target grid and the best prediction achieved by each

source distribution strategy for the ground and airborne synthetic surveys, respectively. For the synthetic ground survey, the horizontal layouts produced similar RMS values of approximately 0.8 mGal regardless of the depth type, with the exception of the regular grid layout which produced a larger RMS of 0.97 mGal. The differences between the target grid and the interpolated values are larger in regions of poor data coverage. Edge effects are present for all strategies but are noticeably smaller for the combination of block-averaged sources with a variable depth based on the nearest neighbour distance. For the synthetic airborne survey, all strategies (including the regular grid) produced similar RMS errors of approximately 0.3 mGal. The maps of the differences between the target grid and interpolation results are visually indistinguishable from each other.

3.2 Window size and overlap in gradient boosting

We assessed the trade-offs in interpolation accuracy and computation time of the gradient-boosted equivalent sources algorithm as a function of the two key controlling factors: the window size and the amount of overlap between adjacent windows. The comparisons were performed against a regular least-squares solution (Eq. 9) using the synthetic airborne data (Fig. 4c-d). To avoid biasing the results, we used the same locations of equivalent sources for both the regular and gradient-boosted interpolations, namely block-averaged sources with a block size of 2000 m and a relative depth of 9000 m.

3.2.1 Window size

The size of the windows controls the size of the Jacobian matrices $\tilde{\mathbf{A}}_k$ by limiting the number of data points and equivalent sources used in each step of the gradient-boosting algorithm (Alg. 2). Thus, using smaller windows will reduce the total amount of computer memory required to estimate the source coefficients. Nevertheless, smaller windows may produce less accurate interpolations by failing to achieve the global minimum of the goal function in Eq. 8. The window size might also impact the computation time in non-intuitive ways since smaller windows generate smaller least-squares problems but also require more gradient-boosting iterations.

We calculated the interpolation RMS error (between the interpolated grid and the target grid in Fig. 5) and computation time for a fixed window overlap of 50% and several window sizes. To avoid any biases introduced by the shuffling of windows, the calculations were repeated using different seeds for the pseudo-random number generator used in the shuffling. Fig. 8a shows the RMS error and Fig. 8c shows the computation time required for estimating the source coefficients, both as functions of the window size.

These results show that the interpolation error for gradient-boosting is generally larger than the

[tb]

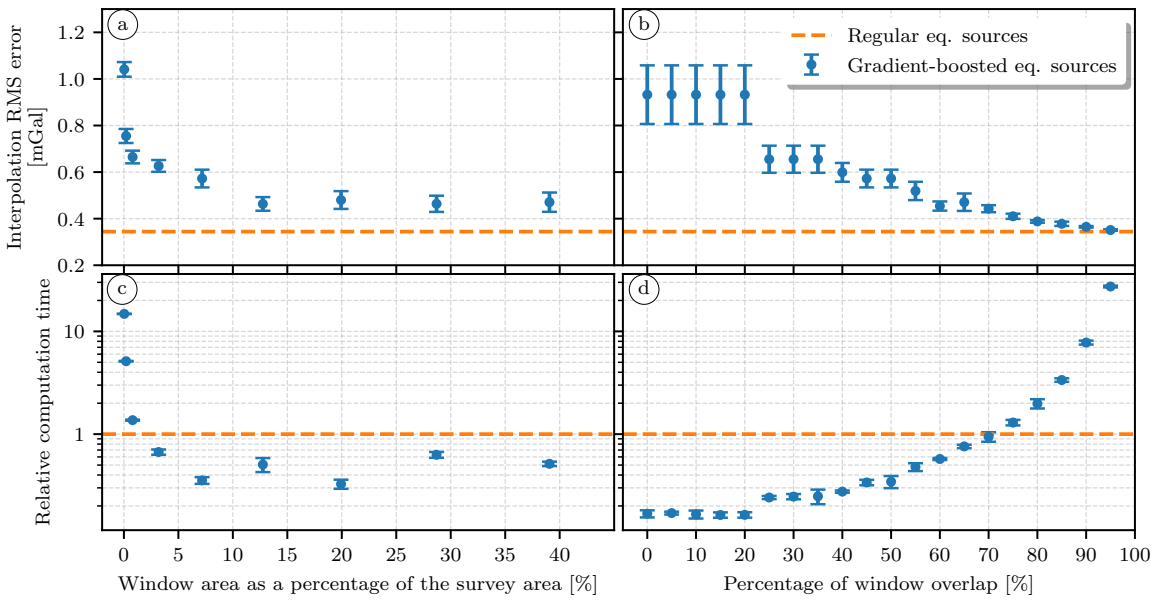


Figure 8. Interpolation RMS error (a-b) and relative computation time (c-d) for regular least-squares equivalent sources (orange dashed lines) and gradient-boosted equivalent sources (blue dots and error bars). Window overlap is given as a percentage of the window size (an overlap of 50% means that two adjacent windows share an area half of the size of the entire window). For gradient-boosting, the RMS errors and computation times are the means (error bars are 1 standard deviation) of results using different seeds for the pseudo-random number generator. Computation time is the ratio between the time required to estimate the source coefficients for the gradient-boosted and the regular equivalent sources.

error for regular equivalent sources. The error decreases asymptotically to within $\sim 40\%$ of the regular equivalent sources for windows with an area greater than $\sim 10\%$ of the survey area. The computation time similarly decreases with window size, with the gradient-boosting being generally faster than the regular equivalent sources for windows with an area greater than $\sim 5\%$ of the survey area. As the window size increases, both RMS error and computation time appear to stabilize to nearly constant levels.

3.2.2 Window overlap

The amount of overlap between adjacent windows plays an important role in the performance of the gradient-boosted equivalent sources. It controls the number of iterations and how many times a particular source is used in the least-squares fitting process. The experiments in the previous section showed that 50% overlap was sufficient to achieve acceptable interpolation accuracy. However, we studied separately the impacts of the amount of window overlap on both accuracy and computation time.

We performed a similar experiment to the one in section 3.2.1 but this time kept the window size fixed to 30000 m and varied the amount of overlap from 0% to 95% with a step size of 5%. All other

experimental procedures remained unchanged. Fig. 8b shows the RMS error and Fig. 8d shows the computation time required for estimating the source coefficients, both as functions of the window overlap.

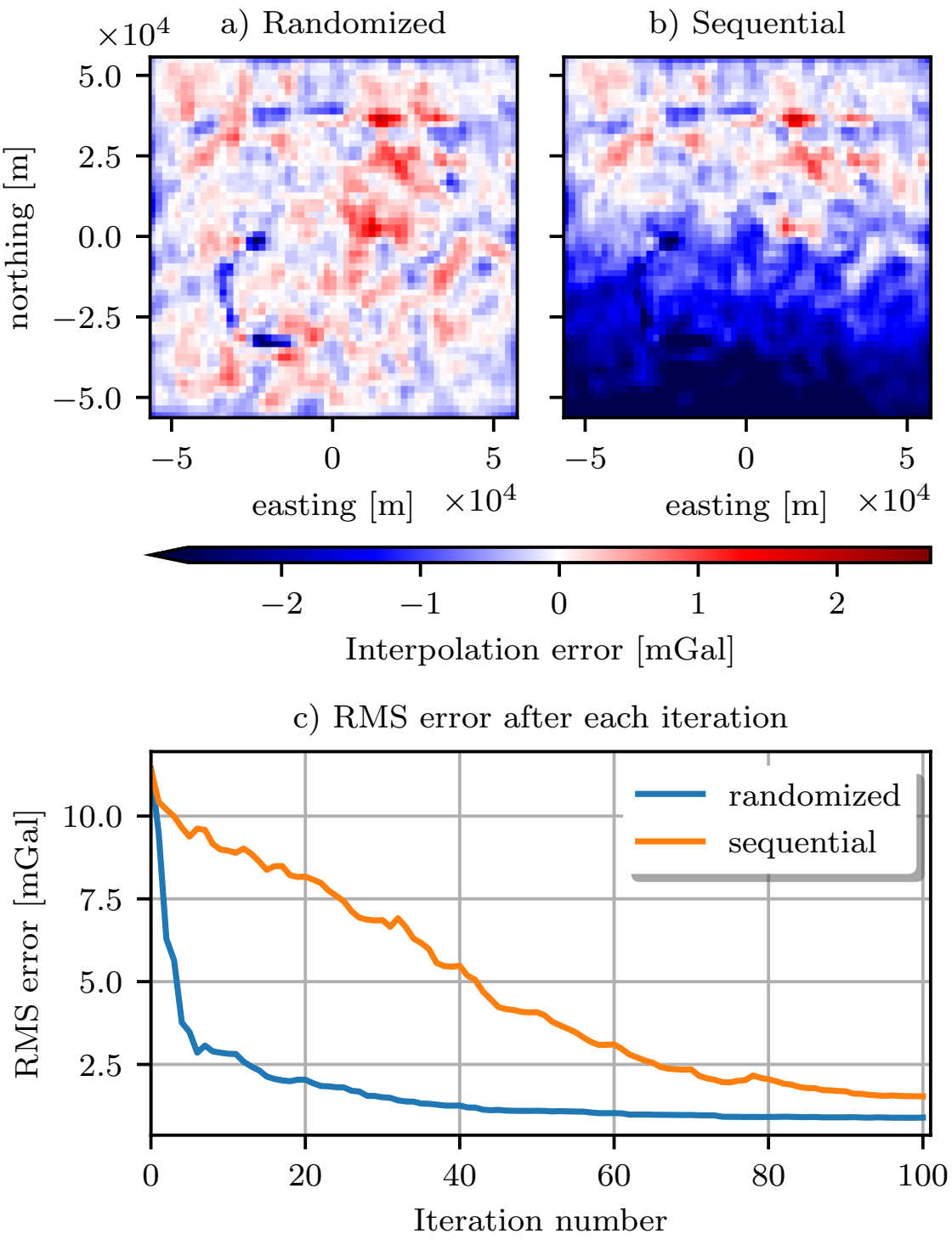
Our results show that the interpolation RMS error decreases with the amount of overlap, reaching the same accuracy as the regular equivalent sources at approximately 90% overlap. On the other hand, the computation time increases with the amount of overlap, becoming larger than that of the regular equivalent sources for overlaps greater than 70%. This is expected since increasing the overlap adds iterations to the gradient boosting algorithm without decreasing the individual least-squares problem sizes to compensate.

3.3 Interpolation with gradient boosting

Finally, we applied the gradient-boosted equivalent sources to interpolate the synthetic airborne survey (Fig. 4). As previously, we used the block-averaged sources layout with a block size of 2 km. Based on the results from section 3.2, we adopted a window overlap of 50% and a window size of 20 km.

We estimated the relative depth of the sources and the damping parameter by comparing the predictions against the values of the target grid. The search explored *depth* values between 1000 m and 19000 m and *damping* values between 1e-06 and 10 by steps of one order of magnitude. The most accurate predictions achieved a RMS error of 0.38 mGal with a depth of 3000 m and a damping of 0.1. It is worth noting that the RMS error achieved by the gradient-boosted equivalent sources is comparable to the ones obtained by the regular equivalent sources in Section 3.1. To highlight the importance of randomizing the order of windows in the gradient-boosting iterations, we preformed the interpolation once more using the same values of *damping* and *depth* but this time iterating over windows in sequential order (South to North, West to East).

Figs. 9a-b show the differences between the target grid and the interpolation results for windows in randomized and sequential order, respectively. The differences for randomized windows resemble those for regular least-squares equivalent sources seen in Figs. 6 and 7. On the other hand, the differences for sequential windows show a clear trend of large negative differences in the South decreasing towards the North. This trend is correlated with the order in which windows are executed, with differences decreasing in absolute value towards the end of the algorithm. Fig. 9c shows the RMS error of the fitting process after each iteration for both window orders, clearly indicating that a randomized window order leads to faster convergence of the algorithm.



[tb]

Figure 9. Interpolation error for gradient-boosted equivalent sources using randomized (a) and sequential (b) window order. (a and b) Pseudo-color maps of the differences between the target grid and the interpolated synthetic airborne survey data. The color scale has been cropped to the same range as Fig. 7. (c) Root-mean squared error after each iteration of the gradient-boosting algorithm.

[p]

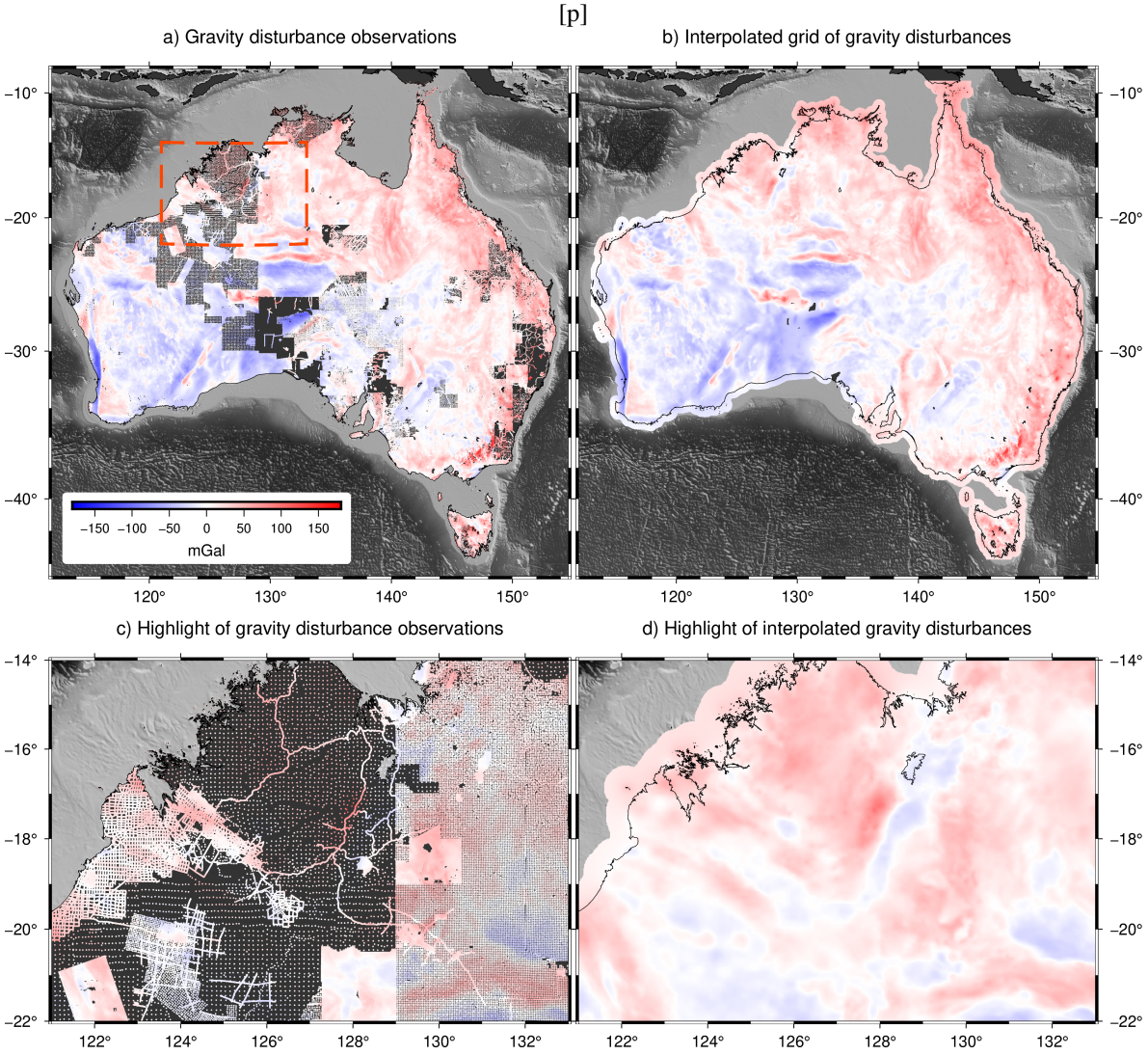


Figure 10. Pseudo-color maps of observed (a and c) and interpolated (b and d) gravity disturbance of Australia. The observed values in a and c are plotted as colored circles. The red rectangle marks the boundaries of the highlight maps in c and d. Observations are part of a compilation by Wynne (2018) of over 1.7 million ground gravity measurements. Interpolated values were obtained through gradient-boosted equivalent sources and calculated on a regular grid at 2127.58 m over the WGS84 ellipsoid.

4 GRIDDING GRAVITY DATA FROM AUSTRALIA

This section will demonstrate how gradient-boosted equivalent sources can be used to interpolate large datasets onto regular grids at uniform height. For this purpose, we selected an open-access compilation of ground gravity surveys over Australia made by Wynne (2018) and filtered and referenced to the WGS84 ellipsoid by Uieda (2021). It contains over 1.7 million data points and covers most of the Australian territory at variable point spacings. Our goal is to create a 1 arc-minute resolution grid of gravity disturbances at a constant geometric height of 2127.58 m (the largest height of observations).

We computed the gravity disturbance by removing the normal gravity of the WGS84 ellipsoid from the observed gravity data (Fig. 10). Here, normal gravity was computed at each observation

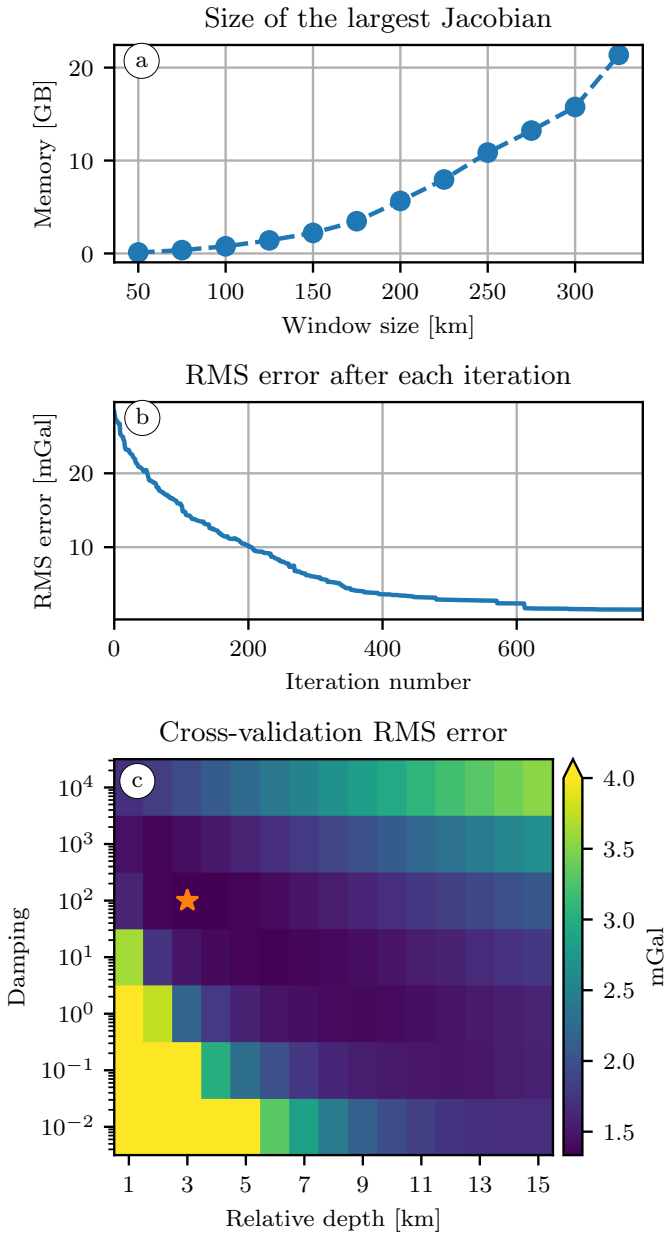
point through the closed-form formula of Li & Götze (2001) using the Boule software (Uieda & Soler 2020). Finally, we converted the observations to planar Cartesian coordinates by applying a Mercator projection.

We start the interpolation process by defining a set of block-averaged sources using a block size of 1.8 km, resulting in a total of 796744 point sources. The block size was chosen to match the desired resolution of the final grid (1 arc-minute is approximately 1.8 km at the equator). Based on the results obtained in Section 3.1, we have chosen to use the *relative depth* strategy. The window overlap was once again fixed at 50%. To determine the size of the windows, we calculated the amount of computer memory needed to store the largest Jacobian matrix for different values of window size (Fig. 11a). We have chosen a size of 225 km in order to limit the amount of memory needed to under 16 Gigabytes.

We determined the depth of the sources and the damping parameter by applying K-Fold cross-validation through the scikit-learn library (Pedregosa et al. 2011). The method randomly divides the original data into k sets (folds), fits the model using only data from $k-1$ folds, and validates the model by comparing its predictions against the one remaining fold. This process is carried out once for each one of the k folds, leading to an estimated mean cross-validation RMS error for the model. To speed up the computation, we only performed the cross-validation on a subset of the data corresponding to an area of $300\,\text{km}\times300\,\text{km}$ containing 14934 points. We ran the cross-validation repeatedly for combinations of *depth*, ranging from $1000\,\text{m}$ to $15000\,\text{m}$, and *damping*, from 0.01 to 10000 in steps of one order of magnitude. Figure 11c shows the resulting cross-validation RMS errors and highlights the minimum value of $1.33\,\text{mGal}$, which corresponds to a relative depth of $3000\,\text{m}$ and a damping equal to 100.

Finally, we proceeded to estimate the source coefficients using the entire dataset and the parameters previously determined. The estimated source coefficients were then used to predict the values of the gravity disturbance on a regular grid of 2442×2085 points at $2127.58\,\text{m}$ above the ellipsoid. On a modest workstation with 16 cores and 16 Gigabytes of RAM, estimating the 796744 coefficients with gradient-boosting took ~ 1.3 hours and the prediction step took ~ 18 minutes.

Fig. 10 shows the original data distribution and the interpolated grid. Grid points that are further than 50 km from the nearest data point are masked to avoid unrealistic extrapolations. Fig. 11b shows the RMS error against the observed data after each iteration of the algorithm. Fig. 12 shows the difference between the observed and predicted gravity disturbances. The inset figure shows a histogram of these residuals, which are approximately normally distributed around zero.



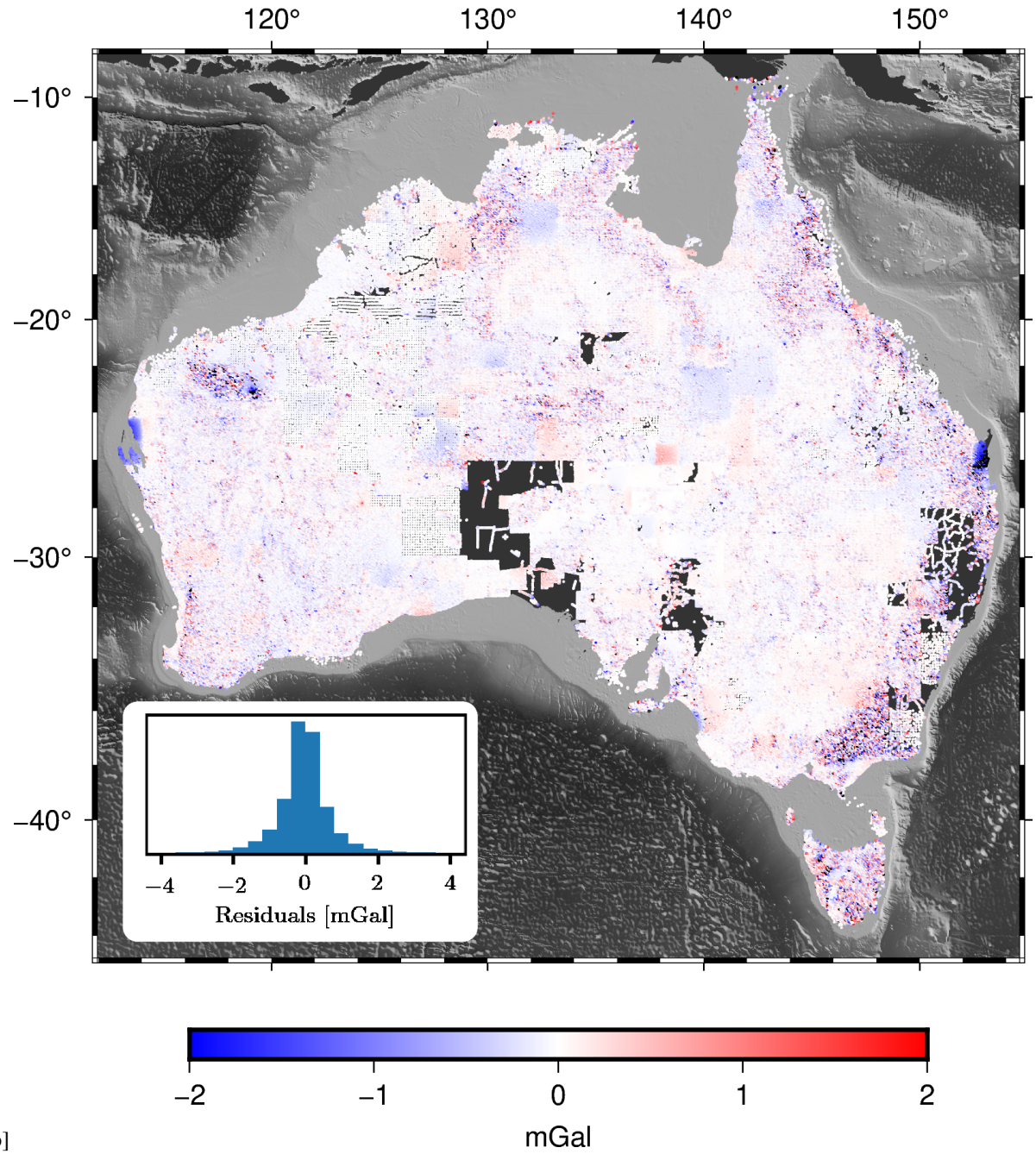
[tbh!]

Figure 11. (a) Amount of computer memory needed to store the largest Jacobian matrix for different window sizes. Our implementation uses double precision floating point numbers (64 bits) for the Jacobian. (b) Root-mean square error against the observed data after each iteration of the gradient-boosting algorithm. (c) K-Fold cross-validation root-mean square errors obtained for each pair of damping and depth parameters. The orange star highlights the minimum.

5 DISCUSSION

5.1 Location of sources

The results of our tests on synthetic data (Figs. 6 and 7) show that there are no significant differences in interpolation accuracy between source distribution strategies, both in terms of the interpolation RMS errors and from visual inspection of the difference maps. Therefore, we conclude that all source



[tb]

Figure 12. Residuals. Differences between the gravity disturbance data from Australia and the predicted values by the estimated equivalent sources on the same observation points. The color map has been truncated to improve the visualization around the largest portion of residual values. The inset plot shows a histogram of the residuals.

distribution strategies are able to produce comparable interpolations. Nevertheless, the *block-averaged sources* strategy makes use of fewer sources when compared with other strategies, which reduces the computational load of estimating the sources coefficients and forward modelling. To ensure that the interpolation is able to reproduce the high frequencies in the data, the block size used in the averaging should be chosen to match the desired grid resolution.

The choice of source depth strategy does not appear to significantly impact the interpolation RMS error. In the particular case of a sparse ground survey with block-averaged sources, the use of a variable depth visibly reduced edge effects and artifacts in areas of poor data coverage. At a first glance, the choice of a depth strategy would not seem to impact the computation time. However, when searching for the set of hyper-parameters that produce the most accurate interpolation (e.g., through cross-validation), one must solve the inverse problem once for every possible combination of parameters. A depth strategy like the *variable depth* requires a higher number of hyper-parameters (depth shift Δz, depth factor α, and the number of nearest neighbours k from Eq. 14) than other strategies which only require a single parameter. Having more parameters means increasing the dimensions of the parameter space and thus increasing the number of possible combinations. Thus, we recommend using a *constant depth* or a *relative depth* when processing large datasets in order to minimize computation time.

5.2 Gradient boosting

From Fig. 8a and 8c, we can see that the gradient-boosted equivalent sources produce slightly less accurate interpolation results but are able to achieve smaller computation times than regular equivalent sources. The reduction of the accuracy might be due to the gradient boosting algorithm failing to converge to the global minimum of the goal function. As the windows increase in size, interpolation error decreases because more data points are included into the least-squares fitting of the source coefficients. At the same time, the fitting process becomes faster because of a reduction in the number of iterations. Our results indicate that it is desirable to maximize the window size, which can be done up to the point that the Jacobian matrices still fit within the available computer memory.

The results shown in Figs. 8b and 8d indicate that using window overlap values between 40% and 70% strike a balance between accuracy and computation time. This corroborates our initial choice of 50% overlap, which is good enough for producing accurate predictions in reasonable times.

Finally, the results in Fig. 9 highlight the importance of randomizing the order in which the overlapping windows are iterated. Running the gradient boosting algorithm sequentially produces less accurate predictions and decreases the convergence rate of the method.

5.3 Australia gravity data

The application of the gradient-boosted equivalent sources to the Australian gravity dataset demonstrates that the method is able to interpolate and upward-continue large datasets in a reasonable amount of time using only modest computational resources. The resulting grid (Fig. 10) preserves the high resolution of the original data while avoiding aliasing artifacts due to the block averaging of the source locations. Some parts of the grid are smoother and have lower amplitudes than the original data (e.g.,

some southwestern parts), which is expected from the upward continuation that was performed to have the grid at a constant height. From the cross-validation analysis on a subset of the data, we estimate that the interpolation error is approximately 1.33 mGal.

The largest residuals in Fig. 12 are located in regions with high-amplitude short-wavelength features in the observed data. This is expected since the method involves some degree of smoothing because of the use of damping and the source depths. There are also low-amplitude long-wavelength residual signals that seem to coincide with some of the windows of the gradient-boosting method. A possible cause of these features is inability of the equivalent-sources within a window to adequately fit the long-wavelength components of the data. We note, however, that all of these long-wavelength residuals are smaller than 1 mGal in amplitude and do not represent a significant source of errors.

The elongated valley around the minimum of the cross-validation RMS errors (Fig. 11c) shows that there is ambiguity in the choice of damping and source depths. One could choose a large damping with a small depth or a small damping with a large depth to achieve roughly the same interpolation result. This is expected since both parameters control the smoothness of the interpolation.

6 CONCLUSIONS

The equivalent source technique has been proven to be well suited for interpolating gravity disturbances and magnetic anomalies. The two main reasons that make it to stand out from other 2D interpolation methods is the fact that the equivalent sources take into account the height of the observations and that the interpolated values will always be harmonic functions. The main challenge of using equivalent sources in practice is the high computational load of estimating the coefficients of the equivalent sources, specially the computer memory needed to store the Jacobian matrix.

We present two strategies that could be simultaneously applied to interpolate datasets with millions of points on modest hardware: block-averaging source locations, which reduces the number of equivalent sources needed for the interpolation, and the gradient-boosted equivalent source algorithm, which breaks down the inverse problem into smaller sets of equivalent sources defined by overlapping windows. Both methods were tested against synthetic datasets in order to compare their accuracy and how they perform in terms of computational efficiency.

Our results show that the block-averaged sources reduce the computational load needed to estimate source coefficients in comparison to two traditional strategies (placing sources below data points or on regular grids). We also show that this reduction of the number of sources does not affect the accuracy of the predictions. The use of block-averaged sources may also prevent aliasing of the interpolated values, specially when the observations are unevenly sampled (e.g., airborne and shipborne surveys). Special attention must be payed when choosing the size of the blocks for averaging. As a thumb rule,

we recommend choosing a size approximately equal to the resolution of the regular grid where the values will be interpolated.

Tests that compared strategies for the vertical location of the sources showed that any one of the three strategies tested (*constant depth*, *relative depth* and *variable depth*) produces comparable accuracy of interpolation. Nevertheless, we are more prone to recommending either the *constant depth* or the *relative depth* for most applications because they involve less hyper-parameters that would need to be configured before the actual interpolation.

Gradient-boosted equivalent sources were shown to heavily reduce the computer memory needed to estimate source coefficients, making it possible to interpolate large datasets with millions of points that would otherwise produce Jacobian matrices larger than the available memory. The interpolations obtained though this new method achieve close to the same accuracy than the regular equivalent sources, while reducing the computation time by approximately a factor of three. We also show that an overlap of 50% between adjacent windows achieves a good compromise between accuracy and computation time. The size of the overlapping windows should be chosen as the maximum value possible that creates Jacobian matrices that still fit into computer memory. Moreover, randomizing the order in which the windows are iterated increases the convergence rate of the algorithm and is essential to producing accurate predictions.

The gradient-boosting method can be used in conjunction with any horizontal source layout, depth strategy, or source type (e.g., point sources, prisms, tesseroids) because it does not rely on assumptions about the sources. Future research should investigate the application of gradient boosting to other equivalent source methods.

7 DATA AND CODE AVAILABILITY

The Python source code used to produce all results and figures presented here is available at https://doi.org/10.6084/m9.figshare.13604360 and https://github.com/compgeolab/eql-gradient-boosted under the BSD 3-clause open-source license.

The gradient-boosted equivalent sources implementation is based on the equivalent source code in the Harmonica library (Uieda et al. 2020b). Other software used in this study includes: Pooch (Uieda et al. 2020a) for downloading and caching datasets, Verde (Uieda 2018) for block reductions and coordinate manipulations, Boule (Uieda & Soler 2020) for normal gravity calculations, xarray (Hoyer & Hamman 2017) and Numpy (Harris et al. 2020) for handling multidimensional arrays and numerical computations, Numba (Lam et al. 2015) for just-in-time compilation and parallelization, scikit-learn (Pedregosa et al. 2011) for cross-validation, Matplotlib (Hunter 2007) and PyGMT (Uieda et al. 2020c) for generating the figures and maps, and the Jupyter notebook programming environment

(Kluyver et al. 2016). Harmonica, Boule, Pooch, and Verde are part of the Fatiando a Terra project (Uieda et al. 2013).

All datasets used are open-access and publicly available. The synthetic surveys were generated using a public domain gravity dataset for Southern Africa distributed by the NOAA NCEI (https://www.ngdc.noaa.gov/mgg/gravity/gravity.html) and the Great Britain Aeromagnetic Survey distributed by the British Geological Survey (BGS) under an Open Government License (https://www.bgs.ac.uk/products/geophysics/aeromagneticRegional.html). The shaded relief in Fig. 10 is the SRTM15+ dataset by Tozer et al. (2019). The Australian ground gravity data is based on a compilation distributed by Geoscience Australia under a Creative Commons Attribution 4.0 International Licence (Wynne 2018) which was filtered and referenced to the WGS84 ellipsoid by Uieda (2021) and is distributed under the same license (https://doi.org/10.6084/m9.figshare.13643837).

8 ACKNOWLEDGEMENTS

We are indebted to the developers and maintainers of the open-source software without which this work would not have been possible. We would also like to thank Editor Frederik Simons, Assistant Editor Fern Storey, and two anonymous reviewers for their constructive comments. S.R. Soler is supported by a scholarship from CONICET, Argentina. This work contains British Geological Survey materials © UKRI. S.R. Soler and L. Uieda jointly developed the initial idea, analysed the results, and wrote the paper. S.R. Soler produced all results and developed the software implementation with the assistance of L. Uieda.

REFERENCES

Barnes, G. & Lumley, J., 2011. Processing gravity gradient data, *Geophysics*, **76**(2), I33–I47. doi:10.1190/1.3548548.

Bouman, J., Ebbing, J., Fuchs, M., Sebera, J., Lieb, V., Szwillus, W., Haagmans, R., & Novak, P., 2016. Satellite gravity gradient grids for geophysics, *Scientific Reports*, **6**(1). doi:10.1038/srep21050.

Cordell, L., 1992. A scattered equivalent-source method for interpolation and gridding of potential-field data in three dimensions, *Geophysics*, **57**(4), 629–636. doi:10.1190/1.1443275.

Dampney, C. N. G., 1969. The equivalent source technique, *Geophysics*, **34**(1), 39–53. doi:10.1190/1.1439996.

Emilia, D. A., 1973. Equivalent Sources Used As An Analytic Base For Processing Total Magnetic Field Profiles, *Geophysics*, **38**(2), 339–348. doi:10.1190/1.1440344.

Friedman, J. H., 2001. Greedy function approximation: a gradient boosting machine, *The Annals of Statistics*, **29**(5), 1189–1232. doi:10.1214/aos/1013203451.

Friedman, J. H., 2002. Stochastic gradient boosting, *Computational Statistics & Data Analysis*, **38**(4), 367–378. doi:10.1016/s0167-9473(01)00065-2.

Fukushima, T., 2020. Speed and accuracy improvements in standard algorithm for prismatic gravitational field, *Geophysical Journal International*, **222**(3), 1898–1908. doi:10.1093/gji/ggaa240.

Guspí, F. & Novara, I., 2009. Reduction to the pole and transformations of scattered magnetic data using Newtonian equivalent sources, *Geophysics*, **74**(5), L67–L73. doi:10.1190/1.3170690.

Guspí, F., Introcaso, A., & Introcaso, B., 2004. Gravity-enhanced representation of measured geoid undulations using equivalent sources, *Geophysical Journal International*, **159**(1), 1–8. doi:10.1111/j.1365-246X.2004.02364.x.

Hansen, R. O., 1993. Interpretive gridding by anisotropic kriging, *Geophysics*, **58**(10), 1491–1497. doi:10.1190/1.1443363.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del R'ıo, J. F., Wiebe, M., Peterson, P., G'erard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., & Oliphant, T. E., 2020. Array programming with NumPy, *Nature*, **585**(7825), 357–362. doi:10.1038/s41586-020-2649-2.

Hoyer, S. & Hamman, J., 2017. xarray: N-D labeled arrays and datasets in Python, *Journal of Open Research Software*, **5**(1). doi:10.5334/jors.148.

Hunter, J. D., 2007. Matplotlib: A 2D graphics environment, *Computing in Science & Engineering*, **9**(3), 90–95. doi:10.1109/MCSE.2007.55.

Jirigalatu, J. & Ebbing, 2019. A fast equivalent source method for airborne gravity gradient data, *Geophysics*, **84**(5), G75–G82. doi:10.1190/geo2018-0366.1.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., & Willing, C., 2016. Jupyter Notebooks – a publishing format for reproducible computational workflows, in *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pp. 87 – 90, IOS Press.

Kother, L., Hammer, M. D., Finlay, C. C., & Olsen, N., 2015. An equivalent source method for modelling the global lithospheric magnetic field, *Geophysical Journal International*, **203**(1), 553–566. doi:10.1093/gji/ggv317.

Lam, S. K., Pitrou, A., & Seibert, S., 2015. Numba, in *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC - LLVM '15*, ACM Press. doi:10.1145/2833157.2833162.

Leão, J. W. D. & Silva, J. B. C., 1989. Discrete linear transformations of potential field data, *Geophysics*, **54**(4), 497–507. doi:10.1190/1.1442676.

Li, D., Liang, Q., Du, J., Sun, S., Zhang, Y., & Chen, C., 2020. Transforming Total-Field Magnetic Anomalies Into Three Components Using Dual-Layer Equivalent Sources, *Geophysical Research Letters*, **47**(3). doi:10.1029/2019gl084607.

Li, X. & Götze, H.-J., 2001. Ellipsoid, geoid, gravity, geodesy, and geophysics, *Geophysics*, **66**(6), 1660–1668.

639 doi:10.1190/1.1487109.

640 Li, Y. & Oldenburg, D. W., 2010. Rapid construction of equivalent sources using wavelets, *Geophysics*, **75**(3),
641 L51–L59. doi:10.1190/1.3378764.

642 Martinez, C. & Li, Y., 2016. Denoising of gravity gradient data using an equivalent source technique, *GEO-*
643 *PHYSICS*, **81**(4), G67–G79. doi:10.1190/geo2015-0379.1.

644 Mendonça, C. A. & Silva, J. B. C., 1994. The equivalent data concept applied to the interpolation of potential
645 field data, *Geophysics*, **59**(5), 722–732. doi:10.1190/1.1443630.

646 Menke, W., 1989. *Geophysical Data Analysis: Discrete Inverse Theory*, Academic Press.

647 Nagy, D., Papp, G., & Benedek, J., 2000. The gravitational potential and its derivatives for the prism, *Journal*
648 *of Geodesy*, **74**, 552–560. doi:10.1007/s001900000116.

649 Nagy, D., Papp, G., & Benedek, J., 2002. Corrections to "The gravitational potential and its derivatives for the
650 prism", *Journal of Geodesy*, **76**(8), 475–475. doi:10.1007/s00190-002-0264-7.

651 Nakatsuka, T. & Okuma, S., 2006. Reduction of magnetic anomaly observations from helicopter surveys at
652 varying elevations, *Exploration Geophysics*, **37**(1), 121–128. doi:10.1071/EG06121.

653 Oliveira, Jr., V. C., Barbosa, V. C. F., & Uieda, L., 2013. Polynomial equivalent layer, *Geophysics*, **78**(1),
654 G1–G13. doi:10.1190/geo2012-0196.1.

655 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P.,
656 Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay,
657 E., 2011. Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research*, **12**, 2825–2830.

658 Sandwell, D. T., 1987. Biharmonic spline interpolation of GEOS-3 and SEASAT altimeter data, *Geophysical*
659 *Research Letters*, **14**(2), 139–142. doi:10.1029/GL014i002p00139.

660 Silva, J. B. C., 1986. Reduction to the pole as an inverse problem and its application to low-latitude anomalies,
661 *Geophysics*, **51**(2), 369–382. doi:10.1190/1.1442096.

662 Siqueira, F. C. L., Jr., V. C. O., & Barbosa, V. C. F., 2017. Fast iterative equivalent-layer technique for
663 gravity data processing: A method grounded on excess mass constraint, *Geophysics*, **82**(4), G57–G69.
664 doi:10.1190/geo2016-0332.1.

665 Smith, W. H. F. & Wessel, P., 1990. Gridding with continuous curvature splines in tension, *Geophysics*, **55**(3),
666 293–305. doi:10.1190/1.1442837.

667 Takahashi, D., Jr., V. C. O., & Barbosa, V. C. F., 2020. Convolutional equivalent layer for gravity data pro-
668 cessing, *Geophysics*, **85**(6), G129–G141. doi:10.1190/geo2019-0826.1.

669 Tikhonov, A. N., 1977. *Solutions of Ill Posed Problems*, Vh Winston.

670 Tozer, B., Sandwell, D. T., Smith, W. H. F., Olson, C., Beale, J. R., & Wessel, P., 2019. Global
671 Bathymetry and Topography at 15 Arc Sec: SRTM15+, *Earth and Space Science*, **6**(10), 1847–1864.
672 doi:10.1029/2019ea000658.

673 Tscherning, C. C., 2015. Least-squares collocation, in *Encyclopedia of Geodesy*, pp. 1–5, Springer Interna-
674 tional Publishing.

675 Uieda, L., 2018. Verde: Processing and gridding spatial data using Green's functions, *Journal of Open Source*

Software, **3**(29), 957. doi:10.21105/joss.00957.

Uieda, L., 2021. Ground gravity data compilation for Australia filtered by survey quality and packaged in CF-compliant netCDF. figshare. doi:10.6084/M9.FIGSHARE.13643837.

Uieda, L. & Soler, S. R., 2020. Boule v0.2.0: Reference ellipsoids for geodesy, geophysics, and coordinate calculations. Zenodo. doi:10.5281/zenodo.3939204.

Uieda, L., Jr, V. C. O., & Barbosa, V. C. F., 2013. Modeling the Earth with Fatiando a Terra, in *Proceedings of the 12th Python in Science Conference*, pp. 96 – 103. doi:10.25080/Majora-8b375195-010.

Uieda, L., Soler, S., Rampin, R., van Kemenade, H., Turk, M., Shapero, D., Banihirwe, A., & Leeman, J., 2020a. Pooch: A friend to fetch your data files, *Journal of Open Source Software*, **5**(45), 1943. doi:10.21105/joss.01943.

Uieda, L., Soler, S. R., Pesce, A., Oliveira Jr, V. C., & Shea, N., 2020b. Harmonica: Forward modeling, inversion, and processing gravity and magnetic data. Zenodo. doi:10.5281/zenodo.3628742.

Uieda, L., Tian, D., Leong, W. J., Toney, L., Newton, T., & Wessel, P., 2020c. PyGMT: A Python interface for the Generic Mapping Tools. Zenodo. doi:10.5281/zenodo.4253459.

von Frese, R. R., Hinze, W. J., & Braile, L. W., 1981. Spherical earth gravity and magnetic anomaly analysis by equivalent point source inversion, *Earth and Planetary Science Letters*, **53**(1), 69–83. doi:10.1016/0012-821x(81)90027-3.

von Frese, R. R. B., Ravat, D. N., Hinze, W. J., & McGue, C. A., 1988. Improved inversion of geopotential field anomalies for lithospheric investigations, *Geophysics*, **53**(3), 375–385. doi:10.1190/1.1442471.

Wynne, P., 2018. NetCDF Ground Gravity Point Surveys Collection. Commonwealth of Australia (Geoscience Australia). doi:10.26186/5C1987FA17078.

APPENDIX A: SOURCE POSITION PARAMETERS USED FOR THE TESTS ON SYNTHETIC DATA

Tables A1 and A2 show the parameter values tested and their optimal values for creating the equivalent source distributions tested in Section 3.1. The optimal values were used to produce the results in Figs 6 and 7.

Table A1. Parameters used to produce each source distribution for interpolating the synthetic ground survey data. Also contains the set of parameters that generates the smallest RMS error for each source distribution and their corresponding RMS.

Source layout	Depth type	Parameters	Values	Best	RMS
Source Below Data	Constant	Depth (m) Damping	1000 to 17000, step size 2000 10^{-4}, 10^{-3},. . . , 10^{2}	7000 10^{-1}	0.78
	Relative	Depth (m) Damping	1000 to 17000, step size 2000 10^{-4}, 10^{-3},. . . , 10^{2}	9000 10^{-1}	0.79
	Variable	Depth (m) Depth factor k neighbours Damping	0 to 1400, step size 200 0.1, 0.5, 1, 2, 3, 4, 5 and 6 1, 5, 10 and 15 10^{-4}, 10^{-3},. . . , 10^{2}	1000 1 15 1	0.80
Block Averaged Sources	Constant	Depth (m) Block size (m) Damping	1000 to 17000, step size 2000 1000, 2000, 3000 and 4000 10^{-4}, 10^{-3},. . . , 10^{2}	7000 3000 10^{-1}	0.77
	Relative	Depth (m) Block size (m) Damping	1000 to 17000, step size 2000 1000, 2000, 3000 and 4000 10^{-4}, 10^{-3},. . . , 10^{2}	7000 3000 10^{-1}	0.79
	Variable	Depth (m) Depth factor k neighbours Block size (m) Damping	0 to 1400, step size 200 0.1, 0.5, 1, 2, 3, 4, 5 and 6 1, 5, 10 and 15 1000, 2000, 3000 and 4000 10^{-4}, 10^{-3},. . . , 10^{2}	600 1 15 3000 10^{-1}	0.72
Grid Sources	Constant	Depth (m) Grid spacing (m) Damping	1000 to 9000, step size 2000 1000, 2000, 3000 and 4000 10^{1}, 10^{2}, 10^{3} and 10^{4}	3000 2000 10^{2}	0.97

Table A2. Parameters used to produce each source distribution for interpolating the synthetic airborne survey data. Also contains the set of parameters that generates the smallest RMS error for each source distribution and their corresponding RMS.

Source layout	Depth type	Parameters	Values	Best	RMS
Source Below Data	Constant	Depth (m) Damping	1000 to 17000, step size 2000 10^{-4}, 10^{-3},. . . , 10^{2}	7000 10^{-2}	0.35
	Relative	Depth (m) Damping	1000 to 17000, step size 2000 10^{-4}, 10^{-3},. . . , 10^{2}	9000 10^{-2}	0.35
	Variable	Depth (m) Depth factor k neighbours Damping	50 to 1450, step size 200 1 to 6, step size 1 1, 5, 10 and 15 10^{-4}, 10^{-3},. . . , 10^{2}	1450 1 15 1	0.36
Block Averaged Sources	Constant	Depth (m) Block size (m) Damping	1000 to 17000, step size 2000 1000, 2000, 3000 and 4000 10^{-4}, 10^{-3},. . . , 10^{2}	9000 3000 10^{-4}	0.34
	Relative	Depth (m) Block size (m) Damping	1000 to 17000, step size 2000 1000, 2000, 3000 and 4000 10^{-4}, 10^{-3},. . . , 10^{2}	9000 2000 10^{-3}	0.34
	Variable	Depth (m) Depth factor k neighbours Block size (m) Damping	50 to 1450, step size 200 1 to 6, step size 1 1, 5, 10 and 15 1000, 2000, 3000 and 4000 10^{-4}, 10^{-3},. . . , 10^{2}	50 2 15 2000 10^{-2}	0.33
Grid Sources	Constant	Depth (m) Grid spacing (m) Damping	1000 to 9000, step size 2000 1000, 2000 and 3000 10^{-3}, 10^{-2},. . . , 10^{2}	7000 1000 10^{-1}	0.34