Principal component analysis for filtering and leveling of geophysical data

Journal Article**Author(s):**

Davydenko, Alexander Y.; Grayver, Alexander

Publication date:

2014-10

Permanent link:

<https://doi.org/10.3929/ethz-b-000089387>

Rights / license:

In Copyright - Non-Commercial Use Permitted

Originally published in:

Journal of Applied Geophysics 109, <https://doi.org/10.1016/j.jappgeo.2014.08.006>

Principal component analysis for filtering and levelling of geophysical data.

Alexander Y. Davydenko^a, Alexander V. Grayver^{b,*}

^a*Formerly Irkutsk State Technical University, Lermontova str., 664074, Irkutsk, Russia;
presently Private Research Consultant*

^b*Institute of Geophysics, ETH Zurich, Sonneggstrasse 5, 8092 Zurich, Switzerland*

Abstract

In this study, we investigate the use of multivariate statistical methods for geophysical data filtering. For this purpose, a measured scalar field is vectorized using a moving window technique and mean vector and covariance matrix are calculated by employing memory-efficient numerical algorithms. These multivariate statistics are then used to conduct principal component analysis (PCA). Namely, covariance matrix is decomposed into a set of eigenvalues and eigenvectors. By selecting a subset of eigenvectors, a PCA-based filter is realized. We demonstrate how properties of the filter are determined by the chosen subset of the eigenvectors, which in turn depend on spectro-spatial properties of the field. In particular, we presented approaches to construct low-pass and spatial directional PCA-based filters. As an application, we aim at suppressing levelling errors commonly occurring in airborne data sets. The devised PCA-filter was analyzed using real aeromagnetic dataset and synthetic levelling errors. The scenarios of statistically dependent and

*Tel.: +41(44)633-3154, E-mail: agrayver@erdw.ethz.ch
URL: davydenkoay@gmail.com (Alexander Y. Davydenko),
agrayver@erdw.ethz.ch (Alexander V. Grayver)

independent levelling errors were studied. Finally, we successfully applied it to real aero-electromagnetic data levelling.

Keywords:

Multivariate statistics, principal component analysis, moving window, levelling, trend removal

2000 MSC: 86A20, 86A32

1. Introduction

Multivariate statistical methods provide a large variety of techniques aimed at both analyzing and filtering data consisting of multiple variables (Härdle and Simar, 2007). These methods find numerous applications in geophysics (Reynolds, 2011). For instance, full-tensor gradiometry data can be viewed as a set of statistical variables and analyzed using multivariate statistical methods. These methods, however, cannot be directly applied to scalar fields such as those obtained from aero-magnetic or aero-electromagnetic surveys. Nevertheless, by using a moving window approach (also known as sliding window or rolling window), data can be represented as a set of statistical variables and thus become suitable for multivariate statistical methods. One of such methods is principal component analysis (PCA), which has found numerous applications in geophysics (Egbert and Booker, 1986; Chung and Nigam, 1999; Guo et al., 2009; Minsley et al., 2012; Smirnov and Egbert, 2012). Here, we apply the combination of the moving window approach and PCA to analyze structure of the field and suppress levelling errors in airborne geophysical data.

Levelling errors typically occur during collection of geophysical data along

19 profile lines over large areas and result in a random, possibly spatially cor-
20 related, displacement of measurements from neighboring profile lines. This
21 creates a characteristic washboard or strip pattern perpendicular to the sur-
22 vey profile direction. The effect has long been recognized when working with
23 aero-magnetic (Yarger et al., 1978) or aero-electromagnetic (Valleau, 2000)
24 data sets. Levelling errors can be caused by diurnal Earth’s magnetic field
25 drift (Minty, 1991), temperature variations in measuring devices (Siemon,
26 2009), flight direction changes or calibration errors (Huang, 2008) and even-
27 tually hinder the interpretation and can cause severe artifacts when inverting
28 data.

29 Several methods exist to remove or suppress this type of errors. Some
30 approaches are based on FFT filtering (Minty, 1991; Nelson, 1994; Ferraccioli
31 et al., 1998; Siemon, 2009), whereas other techniques make use of different
32 fitting methods such as least squares, polynomials or B-splines (Yarger et al.,
33 1978; Mauring et al., 2002; Beiki et al., 2010). Many of the methods depend
34 on so called tie-lines – a few lines measured perpendicular to the main sur-
35 vey profiles. Differences at the cross points between the original profiles and
36 the tie-lines are then used to quantify levelling errors. If no tie-lines are
37 available, some authors suggest choosing a reference line that is presumably
38 free of levelling errors (Huang, 2008) or applying robust median filters us-
39 ing a moving window (Mauring and Kihle, 2005). Alternatively, one could
40 build a directional filter using discrete wavelet transform (Fedi and Quarta,
41 1998; Paoletti et al., 2007). Few researchers have addressed the problem of
42 suppressing levelling errors by means of multivariate statistical methods.

43 We present a method to suppress levelling errors that is based on the

44 multivariate statistical analysis. The method does not require tie-lines and
 45 is readily valid for survey areas with arbitrary contours, gaps inside, and
 46 quasi-regular observation layouts. We start with the definition of the multi-
 47 variate statistics in a moving window, derive PCA-based filter, and discuss
 48 the problem of the dependent noise in data and consequences from the statis-
 49 tical point of view. We then shortly describe differential polynomial fitting
 50 (DPF) approach (Beiki et al., 2010) used to assess efficiency of the PCA-
 51 filter. Finally, we test the devised methodology on the real magnetic field
 52 with synthetic levelling errors, compare its efficiency with DFP and apply it
 53 to the real aero-electromagnetic data from Siberia.

54 **2. Theory and method**

55 *2.1. Data in moving window*

56 Before applying multivariate statistical methods, measured fields need to
 57 be vectorized and represented as a set of random variables. To this end,
 58 we introduce the moving window approach. Denote $\mathbf{d} \in \mathbb{R}^{N_d}$ a real vector
 59 consisting of N_d data measurements. Every measurement d^k is assigned a pair
 60 of coordinates $(x, y)_k$. Data are measured and arranged in N_p quasi-parallel
 61 profile lines. For clarity, we will assume that the lines are directed eastward.
 62 If this condition is not fulfilled, one can rotate profiles to point eastward.
 63 Figure 1 shows a schematic example of a geophysical survey layout.

64 We define a rectangular window of size $n \times m$ points centred at a mea-
 65 surement point d^k , $k = 1 \dots N_d$ with n, m being odd numbers. Figure 1
 66 illustrates an example of the rectangular window of size 5×3 centred at the
 67 k -th measurement point. A subset of points situated in the moving window

at any position is called a window sampling. By cascading all the window rows, we can write down the measurement points in a window sampling as a vector of size $N = n \times m$. In other words, for a rectangular window of size $n \times m$ points centred at a measurement point k , we construct a window sampling vector $\mathbf{w}^k \in \mathbb{R}^N$ as

$$\begin{pmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,n} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m,1} & d_{m,2} & \cdots & d_{m,n} \end{pmatrix} \rightarrow \mathbf{w}^k = \begin{pmatrix} d_{1,1} \\ d_{1,2} \\ \cdots \\ d_{1,n} \\ d_{2,1} \\ \cdots \\ d_{m,1} \\ d_{m,2} \\ \cdots \\ d_{m,n} \end{pmatrix}. \quad (1)$$

Every window sampling vector $\mathbf{w}^k$ is a point in the N -dimensional space. Each dimension of such space is formed by the respective element of the moving window. Figure 2(a) shows a synthetic signal $\mathbf{d}$ consisting of 13 data values. Choosing a moving window of size 3×1 yields eleven window samplings, excluding positions of the window at first and last measurements where one data point inside window is missing. Figure 2(b) shows the window sampling vectors plotted as points in the three-dimensional space. The points colored in blue and green are the values which form window sampling vectors $\mathbf{w} \in \mathbb{R}^3$ at the 3rd and 11th positions. This simplified example serves to show

that the moving window approach is in fact a mapping of the original scalar signal into some higher dimensional vector space. Such space captures the behavior of the signal in some vicinity of each measurement and facilitates application of various multivariate statistical methods.

Details on extending this approach to quasi-regular survey layouts are given in Appendix A.

2.2. Multivariate statistics in moving window

Adopting moving window approach introduced in the previous section enables us to consider a scalar geophysical field $\mathbf{d}$ as a vector field, i.e. at any position there exists vector $\mathbf{w} \in \mathbb{R}^N$ as defined by equation (1). Vector $\mathbf{w}$ is a multivariate random variable that is characterized by its probability density function $f(\mathbf{w})$ with respective expected value $\boldsymbol{\mu}_w$ and covariance matrix $\mathbf{C}_w$. These quantities can be written as following (Tabachnick et al., 2001)

$$\boldsymbol{\mu}_w = E(\mathbf{w}) = \int_{-\infty}^{\infty} \mathbf{w} f(\mathbf{w}) d\mathbf{w} \quad (2a)$$

$$\mathbf{C}_w = E(\mathbf{w}\mathbf{w}^T) - \boldsymbol{\mu}_w \boldsymbol{\mu}_w^T. \quad (2b)$$

In practice, we want to get discrete estimations of the multivariate statistics presented in equation (2). Therefore, we first define a moving window with N points and construct sampling vectors for all N_d points in the data set $\mathbf{d}$. These vectors can be combined into a matrix $\mathbf{F}$ of size $N_d \times N$ where k th row is given by $(\mathbf{w}^k)^T$. Following the approach adopted in this work, every column of this matrix is a random variable, whereas the whole matrix is a statistical sampling with the uniquely defined mean vector and covariance

102 matrix. Calculating mean values for all columns and arranging them in a
 103 vector yields sample mean vector that approximates $\boldsymbol{\mu}_w$ from equation (2a).
 104 More precisely, this can be written as

$$\hat{\boldsymbol{\mu}}_w = \frac{1}{N_d} \mathbf{1}^T \mathbf{F}, \quad (3)$$

105 where $\mathbf{1}^T$ is a row vector with all elements equal one such that $\mathbf{1}^T \mathbf{F}$ is a
 106 vector which contains sums of the columns of $\mathbf{F}$. Accordingly, the covariance
 107 matrix is then given by (cf. equation 2b)

$$\hat{\mathbf{C}}_w = \frac{1}{N_d} \mathbf{F}^T \mathbf{F} - \hat{\boldsymbol{\mu}}_w \hat{\boldsymbol{\mu}}_w^T. \quad (4)$$

108 As was mentioned in the previous section, a set of window sampling vec-
 109 tors is a point cloud in the N -dimensional space characterized by its discrete
 110 joint probability distribution. The sample mean vector in equation (3) gives a
 111 point in the space that minimizes sum of Euclidean distances from all points
 112 to it

$$\hat{\boldsymbol{\mu}}_w = \underset{\mathbf{v} \in \mathbb{R}^N}{\operatorname{argmin}} \|\mathbf{v} - \mathbf{w}^k\|_2, \quad (5)$$

113 whereas sample covariance matrix provides the estimation of linear depen-
 114 dence between all pairs of random variables (or, equivalently, columns of $\mathbf{F}$).
 115 Mean vector and covariance matrix are the most widely used statistics and
 116 in particular case of the normal distribution provide the complete charac-
 117 terization of the respective joint probability distribution (Härdle and Simar,
 118 2007).

119 In contrast to univariate statistics such as mean value or variance, the
 120 multivariate statistics (3) and (4) calculated in a moving window provide a
 121 more complete description of the field since they are valid in some vicinity
 122 of a measurement point determined by the window size.

123 2.3. Efficient implementation

124 For data sets consisting of $10^6 - 10^7$ measurements and large moving
 125 windows, matrix $\mathbf{F}$ becomes huge and prohibitive to form and store explicitly.
 126 To alleviate this limitation, we exploit an algorithm which avoids explicit
 127 formation of the matrix $\mathbf{F}$. Careful inspection of equations (3) and (4) reveals
 128 that they can be rewritten in the following way

$$\hat{\boldsymbol{\mu}}_w = \frac{1}{N_d} \sum_{k=1}^{N_d} \mathbf{w}^k \quad (6)$$

129 and

$$\hat{\mathbf{C}}_w = \frac{1}{N_d - 1} \sum_{k=1}^{N_d} \mathbf{w}^k (\mathbf{w}^k)^T - \hat{\boldsymbol{\mu}}_w \hat{\boldsymbol{\mu}}_w^T. \quad (7)$$

130 Equations (6) and (7) require that all N values in a moving window are
 131 present at any point k . In practice this is often not the case. Clearly, if
 132 a moving window is located at the edge of the survey area or some points
 133 inside the area are missing (see Figure 3), equations above become invalid.
 134 To mitigate this problem, the normalizing factors need to be adjusted as
 135 follows

$$\hat{\boldsymbol{\mu}}_w = \mathbf{b} \circ \sum_{k=1}^{N_d} \mathbf{w}^k \quad (8)$$

136 and

$$\hat{\mathbf{C}}_w = (\mathbf{B} - \mathbf{1}) \circ \sum_{k=1}^{N_d} \mathbf{w}^k (\mathbf{w}^k)^T - \hat{\boldsymbol{\mu}}_w \hat{\boldsymbol{\mu}}_w^T, \quad (9)$$

137 where $\circ$ denotes point-wise multiplication (known as Hadamard product),
 138 vector $\mathbf{b}$ and matrix $\mathbf{B}$ contain reciprocals of the number of occurrences. For
 139 instance, $b_i = 1/(N_d - N_i^0)$, where N_i^0 counts how many times i -th element in
 140 a moving window was missing. Analogously, $B_{ij} = 1/(N_d - N_i^0 - N_j^0)$. It has
 141 long been recognized that the sample covariance matrix calculated for data

142 with gaps can become indefinite (Schwertman and Allen, 1979). This, how-
 143 ever, contradicts the basic definition of the covariance matrix. Therefore, we
 144 check if the matrix $\mathbf{C}_w$ calculated using equation (9) becomes indefinite and
 145 – if it does – calculate the nearest positive semi-definite matrix by adopting
 146 the algorithm presented in Higham (1988).

147 2.4. Stochastic model of the field

148 Often geophysical data processing methods rely on the additive model,
 149 thus assuming measured signal to be a sum of the signal we are looking for
 150 and some noise. Following this assumption, vector $\mathbf{w}$ can be written as

$$\mathbf{w} = \mathbf{w}_a + \mathbf{w}_b, \quad (10)$$

151 with $\mathbf{w}_a$ and $\mathbf{w}_b$ being the signal and noise (e.g., levelling errors), respectively.

152 Here, we examine this assumption from the stochastic point of view
 153 (Vakhromeyev and Davydenko, 1987) and within the framework of meth-
 154 ods used in this work. An appropriate stochastic model allows to take into
 155 account the dependence between random variables (e.g., between $\mathbf{w}_a$ and $\mathbf{w}_b$
 156 in equation 10). In general, these dependences are encapsulated in the joint
 157 probability density function. The latter, however, is not available for real
 158 data and hence we work with the first two statistical moments, namely the
 159 mean vector and covariance matrix. In this case, using equation (10) and
 160 linearity of the expected value operator $E(\mathbf{w})$, equations (2) can be rewritten
 161 as

$$\boldsymbol{\mu}_w = E(\mathbf{w}_a + \mathbf{w}_b) = E(\mathbf{w}_a) + E(\mathbf{w}_b) = \boldsymbol{\mu}_a + \boldsymbol{\mu}_b \quad (11a)$$

$$\mathbf{C}_w = \mathbf{C}_a + \mathbf{C}_b + \mathbf{C}_{ab} + \mathbf{C}_{ab}^T, \quad (11b)$$

where the respective covariance matrices are given by

$$\mathbf{C}_a = E(\mathbf{w}_a \mathbf{w}_a^T) - \boldsymbol{\mu}_a \boldsymbol{\mu}_a^T \quad (12a)$$

$$\mathbf{C}_{ab} = E(\mathbf{w}_a \mathbf{w}_b^T) - \boldsymbol{\mu}_a \boldsymbol{\mu}_b^T. \quad (12b)$$

Matrix $\mathbf{C}_{ab}$ define a linear relationship between the signal and noise. For instance, this relationship can originate from the dependence between the measurement error and accuracy of the sensor positioning. For some cases, quantities (11) and (12) can be calculated using stochastic modeling approach presented in Vakhromeyev and Davydenko (1987). In most cases, however, the problem of determining all relationships for a set of random variables becomes infeasible since factors causing them (e.g. thermal fluctuations, variations in sensor altitude, etc.) are difficult to express analytically. Therefore, it is reasonable to view the covariance matrix as $\mathbf{C}_w = \mathbf{C}_a + \mathbf{C}_d$, where $\mathbf{C}_d = \mathbf{C}_b + \mathbf{C}_{ab} + \mathbf{C}_{ab}^T$ is now considered to be generally dependent noise. Recalling that the total energy of the field is given by the trace of the covariance matrix, we get that the trace of $\mathbf{C}_{ab}$ determines amount of dependency and its influence on a filter's properties.

In section 4 we will demonstrate that the levelling error has relatively low energy when compared to the measured field and the developed filters remain efficient even if significant degree of dependency is present.

2.5. PCA Filtering

Mean vector and covariance matrix can be used to construct flexible and efficient statistical filters. For this purpose, we employed principal component analysis (PCA). PCA is a linear transformation that finds a basis or a new coordinate system where the largest variance of a set of random variables lies

183 along the first coordinate (first principal axis spanned by the first eigenvec-
184 tor), second largest variance is associated with the second coordinate (second
185 principal axis) and so on (Jolliffe, 2005; Guo et al., 2009). For the multivari-
186 ate normal distribution, principal components coincide with the axes of its
187 concentration ellipsoid. The useful property of this basis is its orthogonality
188 and, hence, any multivariate data projected into this basis becomes linearly
189 uncorrelated. In case data obeys normal distribution, the orthogonal basis
190 guarantees that it will be statistically independent, which readily implies
191 zero correlation (note that the opposite is not true for non-Gaussian data)
192 (Hyvärinen and Oja, 2000).

193 In order to find an orthogonal basis for all multivariate random samples
194 $\mathbf{w}^k, k = 1 \dots N_d$ in matrix $\mathbf{F}$, we decompose symmetric positive semi-definite
195 covariance matrix calculated using equation (4) or (7) by applying eigenvalue
196 decomposition

$$\hat{\mathbf{C}}_w = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T, \quad (13)$$

197 where $\mathbf{\Lambda} \in \mathbb{R}^{N \times N}$ is the matrix with eigenvalues on its main diagonal ap-
198 pearing in decreasing order

$$\boldsymbol{\lambda} = \text{diag}(\mathbf{\Lambda}) = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \dots \\ \lambda_N \end{pmatrix}, \quad (14)$$

199 and columns of the orthogonal matrix $\mathbf{V} \in \mathbb{R}^{N \times N}$ are eigenvectors. Eigen-
200 vectors $\mathbf{V}$ span the orthogonal space of principal components, whereas eigen-
201 values represent data variance along the corresponding dimensions. The set
202 $\boldsymbol{\lambda}$ of the eigenvalues of any matrix comprises its spectrum.

203 Matrix $\mathbf{F}$ can now be transformed into matrix of principal components
 204 $\mathbf{X}$ by using the following equation

$$\mathbf{X} = \mathbf{G}\mathbf{V}, \quad (15)$$

205 where $\mathbf{G} = \mathbf{F} - \mathbf{1}\hat{\boldsymbol{\mu}}_w^T$ is the centered matrix $\mathbf{F}$. Employing the orthogonality
 206 of $\mathbf{V}$, the original matrix $\mathbf{F}$ is then given by

$$\mathbf{F} = \mathbf{X}\mathbf{V}^T + \mathbf{1}\hat{\boldsymbol{\mu}}_w^T. \quad (16)$$

207 Inasmuch as k th row of the matrix $\mathbf{F}$ represents a window sampling vector
 208 $(\mathbf{w}^k)^T$ formed at the k th measurement position, the same row of the matrix
 209 $\mathbf{X}$ forms principal component vector $(\mathbf{x}^k)^T$.

210 The filtering properties of the PCA stem from the fact that we can select
 211 any subset r of eigenvectors $\mathbf{V}$ forming thus a reduced basis denoted by $\mathbf{V}_r$.
 212 A vector $\mathbf{w}^k$ is projected into the reduced space and back by substituting $\mathbf{V}_r$
 213 in equations (15) and (16)

$$\mathbf{X}_r = \mathbf{G}\mathbf{V}_r, \quad (17a)$$

$$\mathbf{F}_r = \mathbf{G}\mathbf{V}_r\mathbf{V}_r^T + \mathbf{1}\hat{\boldsymbol{\mu}}_w^T. \quad (17b)$$

214 The filtered values at the measurement positions lie in the central column of
 215 the matrix $\mathbf{F}_r$ with index $c = (N + 1)/2$.

216 Since we are merely interested in the central column of $\mathbf{F}_r$ and want to
 217 avoid forming $\mathbf{X}_r$ and $\mathbf{F}_r$, filtered values at the point k were obtained by
 218 exploiting the following equation

$$(\mathbf{w}_r^k)_c = (\mathbf{w}^k - \hat{\boldsymbol{\mu}}_w)^T \mathbf{a}_r + (\hat{\boldsymbol{\mu}}_w)_c, \quad (18)$$

219 where $(\hat{\boldsymbol{\mu}}_w)_c$ is the central element of the mean vector and $\mathbf{a}_r$ is the central
 220 column of the matrix $\mathbf{V}_r \mathbf{V}_r^T$ from equation (17).

221 One can alternatively think of this procedure as a convolution of the
 222 window function with some filter whose kernel is given by $\mathbf{a}_r$. The filter
 223 kernel $\mathbf{a}_r$ depends on a pre-selected subset r of eigenvectors in matrix $\mathbf{V}_r$.
 224 The subset r can be chosen depending on required properties of the filter. The
 225 easiest way to determine r is to link eigenvectors to the respective eigenvalues
 226 by recalling that the latter correspond to the total variances along respective
 227 dimensions. In practice, this property very often can be used to extract
 228 regional (i.e., slowly-varying and smooth) component of the field by forming
 229 $\mathbf{V}_r$ using few eigenvectors associated with the largest eigenvalues. Another
 230 approach is based on the analysis of the spatial anisotropy (see Appendix B
 231 for details) and will be used for extracting levelling errors in the next sections.

232 Now we have everything at hand to formulate the generic PCA-based
 233 filter using the following algorithm

- 234 Step 1. Choose a rectangular window of size $N = m \times n$ elements.
- 235 Step 2. Calculate multivariate statistics (sample covariance matrix $\hat{\mathbf{C}}_w$ and
 236 mean vector $\hat{\boldsymbol{\mu}}_w$) as described in section 2.3.
- 237 Step 3. Decompose covariance matrix using equation (13).
- 238 Step 4. Select a subset r of eigenvectors using one of the discussed criteria
 239 and construct matrix $\mathbf{V}_r$.
- 240 Step 5. Construct filter kernel $\mathbf{a}_r$ and calculate filtered values for all $\mathbf{w}^k$:
 241 $k = 1 \dots N_d$ using equation (18).

242 2.6. Estimation of the signal-to-noise ratio

243 We can now estimate the influence of the noise in the covariance matrix
 244 $\mathbf{C}_w$, which helps analyze and understand properties of the PCA-filter. Sub-
 245 stituting eigenvalue decompositions of the corresponding covariance matrices
 246 in equation (11), multiplying by $\mathbf{V}$ from both sides and using orthogonality
 247 of $\mathbf{V}$ (i.e. $\mathbf{V}\mathbf{V}^T = \mathbf{V}^T\mathbf{V} = \mathbf{I}$) yields

$$\mathbf{\Lambda} = \mathbf{V}^T \mathbf{C}_a \mathbf{V} + \mathbf{V}^T \mathbf{C}_d \mathbf{V}. \quad (19)$$

248 Denoting respective main diagonals $\boldsymbol{\lambda} = \text{diag}(\mathbf{\Lambda})$, $\boldsymbol{\tau}_a = \text{diag}(\mathbf{V}^T \mathbf{C}_a \mathbf{V})$, $\boldsymbol{\tau}_d =$
 249 $\text{diag}(\mathbf{V}^T \mathbf{C}_d \mathbf{V})$ and using identity $\boldsymbol{\lambda} = \boldsymbol{\tau}_a + \boldsymbol{\tau}_d$, we can estimate the contri-
 250 bution of noise in the spectrum of the covariance matrix $\mathbf{C}_w$ as

$$\delta_i = \frac{(\tau_d)_i}{\lambda_i} = \frac{\lambda_i - (\tau_a)_i}{\lambda_i}. \quad (20)$$

251 The analysis given in the previous paragraph can be complemented with
 252 another measure. In particular, we can quantify the difference between eigen-
 253 vectors of the covariance matrices $\mathbf{C}_w$ and $\mathbf{C}_a$ by using the following equation

$$\eta_i = \|\mathbf{v}_i - (\mathbf{v}_a)_i\|_2, \quad (21)$$

254 where $\mathbf{v}_i$ and $(\mathbf{v}_a)_i$ are i th columns of the $\mathbf{V}$ and $\mathbf{V}_a$ matrices, respectively.

255 We use quantities η and δ to assess PCA-filter in section 4.

256 3. Differential polynomial fitting

257 To facilitate assessment of the presented methodology, we will compare it
 258 with differential polynomial fitting (DPF) as described in Beiki et al. (2010).
 259 Here we give a short description of DPF using second-order polynomials,

260 whereas details can be found in the original work. Similarly to our work, the
 261 DPF method employs a 2D moving window approach. Authors define the
 262 second-order polynomials at a point (x, y) inside a window as

$$p(x)_{1D} = a_2x^2 + a_1x + a_0, \quad (22a)$$

$$p(x, y)_{2D} = a_{2,0}x^2 + a_{0,2}y^2 + a_{1,1}xy + a_{1,0}x + a_{0,1}y + a_{0,0}, \quad (22b)$$

263 where $p(x)_{1D}$ is one-dimensional polynomial for the sentral profile in the
 264 moving window. Once the fitting is done, the levelling error is calculated as
 265 $e = p(0, 0)_{2D} - p(0)_{1D} = a_{0,0} - a_0$. The difference e is zero if the neighboring
 266 profiles are on the same level. Therefore, all variations across profiles that
 267 are failed to be described by the second-order polynomials are assumed to
 268 be due to levelling error.

269 4. Numerical experiment

270 4.1. Experiment setup and principal component analysis of the field

271 In the following sections, we will present an analysis and application of
 272 the derived algorithms. For this purpose, we have used aeromagnetic data
 273 collected southwest of the McFaulds Lake Area, Canada (Figure 4a). The
 274 data was retrieved from the Natural Resources Canada data server. The ge-
 275 ology of the area is generally well-studied due to its prospects for Ni-Cu-PGE
 276 magmatic sulphide and chromite deposits contained in the ultramafic com-
 277 plex of the Precambrian McFaulds Lake (Dyer and Burke, 2012). The main
 278 structures influencing local magnetic field are related to gabbro, anortosite
 279 and ultramafic rocks (see Figure 3 in Dyer and Burke (2012)). The tonalite

280 and granodiorite in southern, and metasediments in northern parts along the
281 central SW-NE diagonal are present. The extensive positive anomaly to the
282 south is related to basaltic and andesitic flows, tuffs and breccias.

283 The data considered in this section consists of 150 lines with 397 measure-
284 ments along each line spaced at the distance of 50 m. The interval between
285 lines is 100 m and thus the whole area has size of 15×19.85 km.

286 It is instructive to study the main aspects of PCA using the field T^M .
287 To this end, we used a moving window of size 1500×1500 m (equivalently,
288 $N = 31 \times 15 = 465$ points) to calculate sample covariance matrix $\hat{\mathbf{C}}^M$ as well
289 as its eigenvalue decomposition (13) (Figure 4(b-d)).

290 As can be seen from Figure 4(c), first eigenvector contains approximately
291 70% of the total field variance, whereas all four eigenvectors explain more
292 than 90% of the total field variance. The behavior of the 2nd and 4th eigen-
293 vectors indicate that they span parts of the field with predominant NE di-
294 rection. Accordingly, 3rd principal component bear information on the NW
295 direction. Restoring original field for every component individually using
296 equation (18) confirms our argumentation (see Figures 5(a-d)). We can also
297 restore the field from all four components (Figure 5e) and calculate the resid-
298 ual field (Figure 5f), which is characterized by small scale heterogeneities.

299 *4.2. Generation of levelling errors*

300 To test the PCA-filter, we generated statistically independent and de-
301 pendent levelling errors (see discussion in section 2.4). The multivariate
302 correlated Gaussian noise T^L that is independent of T^M was generated using
303 the algorithm described in Appendix C. The value 15 nT was used as the
304 standard deviation for the generated noise since it is close to the diurnal

305 magnetic field variations in this area. The independent levelling errors are
 306 shown in Figure 6(a).

307 The levelling error T_D^L , statistically dependent on T^M , was obtained using
 308 the following expression

$$T_D^L = kT^LT^G, \quad (23)$$

309 where $T^G = T^M(pc_2) + T^M(pc_3)$ (see Figures 5b,c) and scaling factor $k =$
 310 $\sqrt{\text{var}(T^L)/\text{var}(T^G)}$ ensures that the variance of the two fields is equal. The
 311 result is demonstrated in Figure 6(b).

312 The noise-contaminated data denoted by T and T^D for independent and
 313 dependent noise realizations, respectively, are shown in Figures 6(c,d). Our
 314 goal is to filter the contaminated fields and suppress the leveling error. Before
 315 doing that, we study the influence of the noise on the multivariate statistics,
 316 namely its contribution to the spectrum of the covariance matrix (represented
 317 by eigenvalues and eigenvectors).

318 We calculated sample covariance matrix $\hat{\mathbf{C}}_w$ and vector mean $\hat{\boldsymbol{\mu}}_w$ for the
 319 fields T^M , T^L , T_D^L and T using a rectangular window of size 1500×1500 m or,
 320 equivalently, $N = 31 \times 15 = 465$ points. The calculated statistics are shown
 321 in Figure 7. We can see some subtle differences between mean vectors $\hat{\boldsymbol{\mu}}^M$
 322 and $\hat{\boldsymbol{\mu}}$ calculated for the initial and distorted fields, respectively, but their
 323 covariance matrices are visually indistinguishable (see Figures 7a and 7c).
 324 The differences between them, however, do exist and they become apparent
 325 if we use PCA.

326 Figure 8 shows three eigenvectors for the respective covariance matrices.
 327 Clearly, the first and eighth eigenvectors for the initial T^M (Figure 8a) and
 328 distorted T (Figure 8c) fields are virtually identical. The levelling error

329 starts to gain significant contribution in eigenvectors associated with smaller
 330 eigenvalues. For instance, we can see that the 21st eigenvector of the matrix
 331 $\hat{\mathbf{C}}$ (Figure 8c) resembles levelling error pattern (Figure 8b) indicating that
 332 this principal component is dominated by the leveling error and should be
 333 filtered out eventually.

334 The argumentation given in the previous paragraph is further comple-
 335 mented with a more rigorous analysis from section 2.6. Figure 9(a) shows
 336 that the levelling errors have little contribution in the eigenvectors that cor-
 337 respond to the roughly first 12 largest eigenvalues. The rest of the spectra
 338 exhibits substantial differences due to the presence of the leveling errors in
 339 the data.

340 Figure 9(b) shows the plot of the δ values (see equation 20). Clearly, δ_i^L
 341 tends to grow for smaller eigenvalues. The contribution of the noise overall
 342 increases for smaller eigenvalues and eventually dominates the spectra.

343 4.3. Application of the PCA-based filter

344 The numerical results from the previous section are in a good agreement
 345 with the theory, prompting application of the PCA-based filter to extract
 346 levelling errors. The efficiency of the PCA-filter is compared with DPF ap-
 347 proach shortly presented in section 3. To assess performance of the two
 348 methods quantitatively, we calculate root-mean square error between two
 349 fields

$$\text{rms} = \sqrt{\frac{\sum_{i=1}^n (T_i^a - T_i^b)^2}{n}}.$$

350 In order to extract the leveling error, we developed an approach called
 351 directional PCA (DPC). More specifically, we exploit the fact that the lev-

352 levelling error exhibits much larger variability in the direction perpendicular
 353 to the profiles. As we have shown, this behaviour is reflected in the spatial
 354 structure of the covariance matrix eigenvectors (see Figure 8). This prompts
 355 us to introduce a directional PCA filter that is capable of extracting parts
 356 of the field with predominant variability in one direction. Such filter is the
 357 same as described at the end of section 2.5, except that at Step 4 we exploit
 358 analysis of variance to select eigenvectors with predominant variability in y
 359 direction automatically. The detailed description of the directional filter is
 360 given in Appendix B.

361 We begin by applying DPF and DPC to the true levelling errors T^L, T_D^L
 362 (Figures 6a,b) in order to validate filters. For an efficient filter, the output
 363 should be very close to the input field in this case. First two rows in Table 1
 364 summarize the results. As we can see, DPC performs better for independent
 365 error, whereas both methods have produced close results for the dependent
 366 levelling error. Next, we apply both methods to the noise-contaminated fields
 367 shown in Figure 6(c,d). The results are shown in the third and fourth rows of
 368 the Table 1. Because of the relatively low correlation values and large misfit
 369 we conclude that both methods exhibit poor performance here. To improve
 370 the accuracy, we first eliminate the regional low-frequency part from the
 371 fields T and T^D . Since the workflow for the independent and dependent
 372 cases was identical and results look similar, Figure 10 shows only the case
 373 of independent noise. To eliminate low-frequency part of the field, we have
 374 used a dozen of eigenvectors corresponding to the largest eigenvalues which
 375 account for more than 90% of the total field variance in the moving window.
 376 These eigenvectors were used to construct the PCA-based filter as described

377 in section 2.5. After applying this filter, the field shown in Figure 10(a)
 378 was obtained. According to the quantities given in Figures 9, this field is
 379 virtually free of levelling error. The residual field T^Δ (Figure 10b) contains
 380 both levelling errors and some part of the initial magnetic field (Figure 4a).
 381 The DPC and DPF methods were applied to it and results are shown in
 382 Figures 4(c, d) and Table 1. Clearly, after subtracting the low-frequency part,
 383 both methods produced much better results with DPC slightly outperforming
 384 DPF for both independent and dependent levelling errors.

385 Finally, subtracting the estimated error from the field T gives us the final
 386 filtered data shown in Figure 10(e). It is evident that the influence of the
 387 levelling error has been substantially reduced, while preserving even small-
 388 scale features of the original field.

389 **5. Field experiment**

390 In this section, we apply the developed approach to real aero-elektromagnetic
 391 data. The survey area is located within the Baikal-Vitim Fold System in
 392 the southeastern part of Siberia (Buryatia Republic, Russia) with Paleozoic
 393 rocks and Precambrian sequences being widespread in the region (Ruzhent-
 394 sev et al., 2012). Aeroelectromagnetic data helps map paleovalleys in basaltic
 395 and granite bedrocks, which are filled predominantly with Neogenic sedi-
 396 ments known to be perspective for ore deposits. The data has been collected
 397 by using four-frequency airborne system EM-4H (Volkovitskiy et al., 2008).
 398 Data acquisition was done at the frequencies of 130, 520, 2080 and 8320 Hz
 399 with approximate dipole moments of 20000, 10000, 4000 and 2000 Am², re-
 400 spectively. The data set consist of 339 lines and 434 measurements along

each line resulting in 147 000 data measurements with spacing of 100×100 m. The average flight height is 123 m with standard deviation, minimum and maximum being 5.7, 96.6 and 194.6 m, respectively. After initial pre-processing, all fields are transformed into apparent resistivity maps taking into account the flight height and source-receiver positions relative to each other. Since the processing applied to the data was identical for all frequencies, we report results only for the highest frequency of 8320 Hz in which the levelling errors are most pronounced.

We filtered natural logarithm of the apparent resistivity because (i) it is desirable to avoid working with original quadrature and in-phase components since they depend on subsurface resistivity and height of the flight in a complicated non-linear way (Siemon, 2009) and (ii) statistical distribution of the apparent log-resistivity is close to normal – a favourable property for the multivariate statistical methods employed in this work. Finally, because we need to perform many arithmetic operations with fields, the logarithm is better suited since the positivity of the argument is preserved.

The logarithm of apparent resistivity for frequency of 8120 Hz is shown in Figure 11(a). One can observe levelling errors on the order of 10-100 Ωm . We start with extracting low-frequency part of the field by applying PCA filter described in section 2.5. The result of this filter is shown in Figure 11(b). Again, only the components corresponding to the few largest eigenvalues were used to construct the filter. These components account for around 80% of the total data variance. Next, the directional filter was applied to the residual field show in Figure 11(c) to extract the component associated with dominant variability in the y direction. The result of the directional

426 filtering is shown in Figure 12(a) and the final levelled apparent resistivity is
 427 presented in Figure 12(b). Comparing filtered and initial fields (Figure 11a),
 428 we observe the substantial improvement which confirms that the developed
 429 methodology works efficiently for real data. The computational time for this
 430 example was several minutes on the laptop with four CPU cores.

431 6. Conclusions

432 In this work we have derived the generic PCA-based filter based on the
 433 multivariate statistics calculated using the moving window approach. As a
 434 relevant application, we have developed a methodology to suppress levelling
 435 errors. To this end, we first remove the low-frequency part of the field by
 436 applying the principal component analysis and projecting our data into a
 437 reduced subspace. This is done by selecting a few components corresponding
 438 to the largest eigenvalues of the covariance matrix. To extract levelling error
 439 from residual field we proposed to utilize a directional PCA-filter based on
 440 analysis of variance of principal components.

441 We have applied this technique to real world aero-magnetic data from
 442 Canada which was deteriorated with synthetic levelling error. We have
 443 demonstrated that the method levels data efficiently for both statistically in-
 444 dependent and dependent noise. It was shown that removing low-frequency
 445 part of the field prior to levelling by means of PCA-filter improves accuracy
 446 substantially. We have compared PCA-based directional filter with differen-
 447 tial polynomial fitting. Both methods performed similarly with PCA-based
 448 giving slightly better results.

449 Although data levelling has been used as an example here, the presented

450 algorithms are general and can be used to solve a variety of relevant prob-
451 lems related to data processing and analysis. The important advantage of
452 PCA-based filter is that it is based on the field structure reflected in the
453 eigenvectors of the covariance matrix. We showed that the same work-flow
454 can be used for detrending or filtering high-frequency noise. Since vector
455 mean and covariance matrix give an extensive statistical characterization of
456 the data within the moving window, a thorough analysis of principal compo-
457 nents help reveal subtle effects in the data. Directional filtering based on the
458 analysis of variance inside a moving window is straightforward to generalize
459 for multiple directions yielding a tool to detect the directions of the dominant
460 variability in the data and suppress or amplify corresponding components of
461 the field.

462 Few adjustments may prove desirable in the future to make this method-
463 ology efficient for data sets collected over very large areas with substantially
464 different geological settings. For instance, the filter properties can be im-
465 proved by performing cluster analysis prior to filtering in order to split a
466 data set into several subsets with significant differences in statistical proper-
467 ties. Processing these subsets independently would then allow to keep local
468 statistical characteristics of the field and make presented filters more effi-
469 cient. The directional filter can be less efficient if anomalies with amplitude
470 and direction coinciding with those of the levelling error are present in the
471 field. In this case, assumptions used in the directional filter may be violated
472 and other criteria would be required. The application of the robust modifi-
473 cations of the PCA may be necessary for data sets containing outliers. We
474 plan to report on these elaborated aspects in our future works.

475 7. Acknowledgements

476 We acknowledge Aerogeophisica Inc. for permission to publish AEM
 477 data. We are grateful to Alexey Trusov and Pavel Babayants for a number of
 478 constructive discussions on AEM data processing. The presented algorithms
 479 are a part of the geophysical software package GelioSML.

480 Appendix A. Application to the quasi-regular survey layouts

481 In section 2.1 we assumed that measurements are carried on a regular
 482 grid. In practice this is never the case and thus some sort of interpolation is
 483 necessary to fulfill this requirement. For many practical scenarios, it is feasi-
 484 ble to weaken the regularity requirement and avoid additional interpolation
 485 which inevitably introduces some errors.

486 Denote Δy the distance between two neighboring profiles and Δx between
 487 two neighboring measurement points along a profile. Further, take $\overline{\Delta y}$ and
 488 $\overline{\Delta x}$ to be average distances for Δy and Δx , respectively. As long as following
 489 condition holds for all Δy and Δx

$$\begin{aligned} 0.5\overline{\Delta y} < \Delta y < 1.5\overline{\Delta y} \quad \forall \Delta y \\ 0.5\overline{\Delta x} < \Delta x < 1.5\overline{\Delta x} \quad \forall \Delta x, \end{aligned} \tag{A.1}$$

490 it is possible to work with such data set as if it was regular since one can
 491 assign $N = m \times n$ measurement points to a window unambiguously. Figure 13
 492 shows a schematic example of a quasi-regular acquisition network. Whether
 493 to follow this assumption or interpolate a field on a regular grid depends
 494 eventually on a data set under consideration.

495 Appendix B. Directional PCA-filter

496 A directional filter is aimed at extracting parts of the field with domi-
497 nant variability in a certain direction. This can be accomplished by selecting
498 covariance matrix eigenvectors associated with field components that have
499 dominant spatial variation in a specific direction. For instance, levelling er-
500 rors have a pronounced direction-dependent variability such that they usually
501 vary faster in the direction perpendicular to the original survey profiles than
502 along profiles. Assuming that profiles are directed eastward, or equivalently
503 along x-axis, the levelling error exhibits faster variations along y-axis. This
504 characteristic behaviour is reflected in the covariance matrix eigenvectors and
505 suggests introducing a directional filter.

506 Following equation (1), any eigenvector can be represented either as a
507 vector $\mathbf{v}_i$, $\forall i \in [1, N]$ or as a two-dimensional array of size $m \times n$. In
508 this section, we work with the latter representation, namely eigenvectors are
509 folded into two-dimensional arrays. If the values in the array exhibit the
510 maximum variance in y direction over all other directions, it is reasonable to
511 assume that the respective component is dominated by the levelling errors
512 and should be filtered out.

513 In order to perform such analysis automatically, a quantitative criterion
514 to estimate the variance of the data in a 2D array along a specified direction is
515 required. Single-factor analysis of variance (ANOVA) (Gelman, 2005) offers
516 a proper tool for this. To employ ANOVA, we split an array formed by the
517 eigenvector into m groups $\mathbf{g}_{1...m}$ (Figure 14). The groups are then supplied to
518 ANOVA, which tests whether the mean values of the group values are drawn
519 from the general population with the same mean value. To this end, ANOVA

520 calculates F-statistic F given by the ratio of the between-group variance to
 521 the within-group variance (Weir and Cockerham, 1984)

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}, \quad (\text{B.1})$$

where MS_{between} and MS_{within} are mean squares of the sum of squares defined as

$$MS_{\text{between}} = \frac{SS_{\text{between}}}{m - 1}$$

$$MS_{\text{within}} = \frac{SS_{\text{within}}}{N - m},$$

and sum of squares are calculated as

$$SS_{\text{between}} = \sum_{i=1}^m \frac{(\sum_{j=1}^n g_{ij})^2}{n} - \frac{(\sum_{i=1}^N d_i^k)^2}{N}$$

$$SS_{\text{within}} = \sum_{i=1}^N (d_i^k)^2 - \sum_{i=1}^m \frac{(\sum_{j=1}^n g_{ij})^2}{n}.$$

522 F-statistic is then compared against the critical value $F_c = \mathcal{F}^{-1}(m -$
 523 $1, N - m, p)$ drawn from the theoretical inverse probability density function
 524 of the F-distribution with $m - 1$ and $N - m$ being the number of degrees
 525 of freedom and $p = 0.05$ the significance level. If $F > F_c$, it is accepted
 526 that the mean values of the groups exhibit statistically significant difference
 527 indicating substantial variability. If $F \leq F_c$, the between-group variance
 528 exhibits no statistically significant difference and the variability is uniform.
 529 Finally, repeating this analysis for all eigenvectors and angles between 0 and
 530 179 degrees, we select a subset l of eigenvectors dominated by the levelling
 531 error. An eigenvector $v_{1...N}$ belongs to subset l if the following conditions are

532 fulfilled

$$\begin{cases} \operatorname{argmax}_{\alpha=0\dots 179^\circ} F(\alpha) = F(0) \\ F(0) > F_c(0) \end{cases} \quad (\text{B.2})$$

533 where $F(\alpha)$ is F-statistic calculated for groups g constructed for angle α .
 534 First condition checks that the maximum variability is in y direction, whereas
 535 second inequality ensures that the variability in y direction is statistically
 536 significant. Once the subset l is formed, standard PCA-based filter presented
 537 in section 2.5 is constructed and applied to the field.

538 Note that the introduced directional filter is not limited to extracting
 539 eastward component of the field and can easily be generalized to any other
 540 direction. For instance, if we were interested in extracting component of the
 541 field with spatial domination along 45° , we would use $F(45)$ in the right-hand
 542 side of equation (B.2).

543 **Appendix C. Drawing random samples from the multidimensional** 544 **Gaussian distribution**

545 We aim at drawing a random normal zero-mean vector $\mathbf{n} \in \mathbb{R}^m$ from
 546 distribution $\mathcal{N}(\mathbf{0}, \mathbf{C})$, where $\mathbf{C} \in \mathbb{R}^{m \times m}$ is a predefined covariance matrix.

547 To this end, we specify auto-correlation function

$$R(x) = \exp(-ax), \quad (\text{C.1})$$

548 where a is the constant factor used to control correlation radius. In this
 549 study, we have used $a = 0.0002$. Figure 15 shows the graph of the correlation
 550 function defined above. Further, we construct covariance matrix $\mathbf{C} = \sigma^2 \mathbf{R}$,

551 where σ is standard deviation of noise and an element r_{ij} of the correlation
 552 matrix $\mathbf{R}$ is given by

$$r_{ij} = R(\Delta x | j - i |), \quad (\text{C.2})$$

553 where Δx is the measurement interval along a profile line.

554 The eigenvalue decomposition is then applied to the covariance matrix
 555 $\mathbf{C} = \mathbf{V}\mathbf{D}\mathbf{V}^T$. Recalling that the eigenvectors form orthogonal basis, we
 556 can generate a vector $\mathbf{y} \in \mathbb{R}^m$ of independent (hence uncorrelated) random
 557 normal numbers $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$ and project it into original coordinate system
 558 as

$$\mathbf{n} = \mathbf{V}\mathbf{y}, \quad (\text{C.3})$$

559 where $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$. The interested reader is referred to Cressie (1993) for
 560 details.

561 Figure 16 shows few random samples drawn from the multivariate normal
 562 distribution using the introduced method.

563 Beiki, M., Bastani, M., Pedersen, L.B., 2010. Leveling HEM and aeromag-
 564 netic data using differential polynomial fitting. *Geophysics* 75, L13–L23.

565 Chung, C., Nigam, S., 1999. Weighting of geophysical data in princi-
 566 pal component analysis. *Journal of Geophysical Research: Atmospheres*
 567 104, 16925–16928. URL: <http://dx.doi.org/10.1029/1999JD900234>,
 568 doi:10.1029/1999JD900234.

569 Cressie, N., 1993. *Statistics for spatial data* .

570 Dyer, R.D., Burke, H.E., 2012. Preliminary Results from the McFaulds Lake
 571 (“Ring of Fire”) Area Lake Sediment Geochemistry Pilot Study, North-

ern Ontario. Open File Report 6269. Ontario Geological Survey. Ontario,
Canada.

Egbert, G.D., Booker, J.R., 1986. Robust estimation of geomagnetic transfer
functions. *Geophysical Journal International* 87, 173–194.

Fedi, M., Quarta, T., 1998. Wavelet analysis for the regional-residual and
local separation of potential field anomalies. *Geophysical Prospecting* 46,
507–525.

Ferraccioli, F., Gambetta, M., Bozzo, E., 1998. Microlevelling procedures
applied to regional aeromagnetic data: an example from the Transantarctic
Mountains (Antarctica). *Geophysical Prospecting* 46, 177–196.

Gelman, A., 2005. Analysis of variance - why it is more important than ever.
The Annals of Statistics 33, 1–53.

Guo, H., Marfurt, K.J., Liu, J., 2009. Principal component spectral analysis.
Geophysics 74, P35–P43.

Härdle, W., Simar, L., 2007. *Applied multivariate statistical analysis*.
Springer.

Higham, N.J., 1988. Computing a nearest symmetric positive semidefinite
matrix. *Linear algebra and its applications* 103, 103–118.

Huang, H., 2008. Airborne geophysical data leveling based on line-to-line
correlations. *Geophysics* 73, F83–F89.

Hyvärinen, A., Oja, E., 2000. Independent component analysis: algorithms
and applications. *Neural networks* 13, 411–430.

- 594 Jolliffe, I., 2005. Principal component analysis. Wiley Online Library.
- 595 Mairing, E., Beard, L.P., Kihle, O., Smethurst, M.A., 2002. A comparison of
596 aeromagnetic levelling techniques with an introduction to median levelling.
597 Geophysical Prospecting 50, 43–54.
- 598 Mairing, E., Kihle, O., 2005. Leveling aerogeophysical data using a moving
599 differential median filter. Geophysics 71, L5–L11.
- 600 Minsley, B.J., Smith, B.D., Hammack, R., Sams, J.I., Veloski, G., 2012.
601 Calibration and filtering strategies for frequency domain electromagnetic
602 data. Journal of Applied Geophysics 80, 56–66.
- 603 Minty, B., 1991. Simple micro-levelling for aeromagnetic data. Exploration
604 Geophysics 22, 591–592.
- 605 Nelson, J.B., 1994. Leveling total-field aeromagnetic data with measured
606 horizontal gradients. Geophysics 59, 1166–1170.
- 607 Paoletti, V., Fedi, M., Florio, G., Rapolla, A., 2007. Localized cultural
608 denoising of high-resolution aeromagnetic data. Geophysical prospecting
609 55, 421–432.
- 610 Reynolds, J.M., 2011. An introduction to applied and environmental geo-
611 physics. John Wiley & Sons.
- 612 Ruzhentsev, S., Minina, O., Nekrasov, G., Aristov, V., Golionko, B., Doron-
613 ina, N., Lykhin, D., 2012. The baikal-vitim fold system: Structure and
614 geodynamic evolution. Geotectonics 46, 87–110.

- 615 Schwertman, N.C., Allen, D.M., 1979. Smoothing an indefinite variance-
616 covariance matrix. *Journal of Statistical Computation and Simulation* 9,
617 183–194.
- 618 Siemon, B., 2009. Levelling of helicopter-borne frequency-domain electro-
619 magnetic data. *Journal of Applied Geophysics* 67, 206–218.
- 620 Smirnov, M.Y., Egbert, G., 2012. Robust principal component analysis of
621 electromagnetic arrays with missing data. *Geophysical Journal Interna-*
622 *tional* 190, 1423–1438.
- 623 Tabachnick, B.G., Fidell, L.S., Osterlind, S.J., 2001. Using multivariate
624 statistics .
- 625 Vakhromeyev, G., Davydenko, A.Y., 1987. Calculation of potential fields
626 for stochastic models of heterogeneous geological objects. *Geophysical*
627 *prospecting* 35, 288–311.
- 628 Valteau, N.C., 2000. HEM data processing: a practical overview. *Exploration*
629 *Geophysics* 31, 584–594.
- 630 Volkovitskiy, A., Karshakov, E., Trusov, A., 2008. Four frequency AEM sys-
631 tem EM-4H, in: 5th International conference on airborne electromagnetics,
632 Haikko Manor, Finland.
- 633 Weir, B.S., Cockerham, C.C., 1984. Estimating F-statistics for the analysis
634 of population structure. *evolution* , 1358–1370.
- 635 Yarger, H.L., Robertson, R.R., Wentland, R.L., 1978. Diurnal drift removal
636 from aeromagnetic data using least squares. *Geophysics* 43, 1148–1156.

Table 1: Correlation coefficient R and RMS error between original and filtered fields obtained by applying the differential polynomial fitting (DPF) and directional principal component analysis (DPC) using the moving window of size 1500×1500 m.

Field	Method	Independent levelling error		Dependent levelling error	
		R	rms (nT)	R	rms (nT)
T^L, T_D^L	DPF	0.89	6.88	0.99	3.10
	DPC	0.99	2.46	0.97	3.64
T, T_D	DPF	0.52	20.0	0.46	23.4
	DPC	0.51	22.8	0.47	23.2
T^Δ, T_D^Δ	DPF	0.75	10.6	0.71	11.3
	DPC	0.80	9.4	0.72	10.9

Figure 1: Sketch of a survey layout. Black points indicate measurement positions grouped into N_p profile lines. The dashed line depicts a rectangular window of size 5×3 centered at the measurement point k .

Figure 2: (a) Signal shown with the line and two positions of the moving window centered at points 3 and 11 with respective signal values marked by the blue triangles and green squares. (b) Three-dimensional space formed by window sampling vectors extracted from the signal in (a). The blue and green points in this space correspond to the $\mathbf{w}^3$ and $\mathbf{w}^{11}$ window sampling vectors, respectively.

Figure 3: Survey layout with a gap inside. Two moving window positions with missing values inside are shown with dashed-line rectangles.

Figure 4: (a) Magnetic field data T^M collected in the McFauld's Lake area, Canada. Coordinates are given in UTM convention. (b) Covariance matrix $\hat{\mathbf{C}}^M$ calculated using the moving window of size 1500×1500 m. (c) Normalized cumulative contributions $s_i = \left(\sum_{j=1}^i \lambda_j \right) / \left(\sum_{j=1}^N \lambda_j \right)$ to the total field variance with the contribution of the first four components indicated by the dashed line. (d) First four eigenvectors of the covariance matrix in (b) corresponding to the first four principal axes.

Figure 5: (a-d) Magnetic fields obtained from the first four principal components individually. (e) Sum of the four magnetic fields (a-d). (f) Residual magnetic field after subtracting (e) from the field T^M .

Figure 6: Synthetic levelling errors (a) independent of the field T^M and (b) dependent on it; (c-d) Noise-contaminated T^M field.

Figure 7: Covariance matrices $\hat{\mathbf{C}}^M$, $\hat{\mathbf{C}}^L$, $\hat{\mathbf{C}}$ and mean vectors $\hat{\boldsymbol{\mu}}^M$, $\hat{\boldsymbol{\mu}}^L$, $\hat{\boldsymbol{\mu}}$ calculated for the respective fields from Figures 4(a) and 6(a,c) using the moving window of size 31×15 points.

Figure 8: First, 8th and 21st eigenvectors of the covariance matrices (a) $\hat{C}^M$, (b) $\hat{C}^L$ and (c) $\hat{C}$ from Figure 7.

Figure 9: (a) Relative discrepancy between eigenvectors of the covariance matrices $\hat{\mathbf{C}}$ (black line), $\hat{\mathbf{C}}^D$ (gray line) and $\hat{\mathbf{C}}^M$. (e) Relative contribution of the independent noise T^L (black line) and dependent noise T_D^L (gray line) into the spectrum of the matrix $\hat{\mathbf{C}}$.

Figure 10: (a) Low-frequency part of the field T that is free of levelling error, denoted by T^{low} . (b) The residual field $T - T^{low}$. (c) Estimated leveling error (cf. Figure 4c) obtained by applying DPF to the field T^{Δ} in (b). (d) Same as (c), but obtained by applying DPC filter. (e) Residual field $T - DPC(T^{\Delta})$. (f) Comparison of the true and recovered leveling errors along the two cross-profile lines indicated by A and B in (d).

Figure 11: (a) Natural logarithm of the apparent resistivity calculated for the frequency of 8320 Hz for aero-electromagnetic survey carried out in Vitim ore province, Russia. A zoom-in of the field values in the black rectangle is shown in the right bottom corner. (b) Low frequency part of the apparent resistivity extracted by applying the PCA filter. (c) Difference between (a) and (b).

Figure 12: (a) Levelling errors extracted by applying directional filter to the field from Figure 11(c). (b) Original field with levelling error subtracted.

Figure 13: Sketch of a quasi-regular survey layout. Black points indicate measurement positions grouped into N_p profile lines. The dashed rectangle show a rectangular window of size 5×3 centered at the measurement point k . Compare with Figure 1.

Figure 14: (a) A window sampling of size $m \times n$ subdivided into m groups along horizontal lines of the 2D array. (b) Same as (a), but division is carried out along dashed lines rotated by 45° . Accordingly, these groups are used to estimate the variance in corresponding perpendicular direction.

Figure 15: (left) Correlation function calculated along profile line with 400 points spaced at 50 m. (right) Covariance matrix that is constructed using plotted correlation function and $\sigma = 15$.

Figure 16: Few random realizations drawn from the multivariate normal distribution using the covariance matrix shown in Figure 15.