DEOS Progress Letter

Edited by Roland Klees

no 98.1

DEOS

DEOS Progress Letter

98.1

DEOS Progress Letter

98.1

Edited by Roland Klees

Delft University Press / 1998

Published and distributed by:

Delft University Press
Postbus 98
2600 MG Delft
The Netherlands
Telephone: +31 15 2783254
fax: +31 15 2781661
E-mail: DUP@DUP.TUDelft.NL

ISBN 90-407-1832-6

Printed in the Netherlands

CONTENTS

On the Information Content and Regularisation of Lunar Gravity Field Solutions 1
Rune Floberghagen and Johannes Bouman

Integration of a priori gravity field models in boundary element formulations to geodetic boundary value problems 21
Roland Klees and Rüdiger Lehmann

Fast numerical solution of the vector Molodensky problem 31
Roland Klees, Christian Lage and Christoph Schwab

Stabilization of global gravity field solutions by combining satellite gradiometry and airborne gravimetry 43
Johannes Bouman and Radboud Koop

The shift operators and translations of spherical harmonics 57
Martin van Gelderen

A gravity mission for Earth sciences 69
Roland Klees and Radboud Koop

A procedure for combining gravimetric geoid models and independent geoid data, with an example in the North Sea region 89
Roger Haagmans, Arnoud de Bruijne and Erik de Min

A strategy for geoid detremination in the Indonesian archipelago 101
Kosasih Prijatna

DEOS Progress Letter 98.1: 1–19

On the Information Content and Regularisation of Lunar Gravity Field Solutions

Rune Floberghagen and Johannes Bouman

Abstract

The quality of the lunar gravity model GLGM-2 is analysed on the basis of the solution bias and contribution of the observations to the actual solution. Alternative parameter choices for the regularisation of lunar gravity solutions are presented and applied to both GLGM-2 and prospective future gravity field solutions from satellite-to-satellite tracking. Finally, the use of the mean square error matrix as quality measure is advocated, as opposed to the commonly used variance-covariance matrix.

1 Introduction

Much progress has been booked in lunar gravity field modelling over the past few years due to the availability of new satellite tracking data. The Clementine mission, Nozette *et al.* (1994), and in particular Lunar Prospector, Binder (1998), have added significantly to the data sets obtained during the Apollo era of lunar exploration. Although the Clementine orbit was far from ideal for gravity modelling purposes, the mission gave the first high-quality near-side and high-inclination satellite tracking data since the Apollo missions. Due to the eccentricity of the orbit, the spacecraft could also be tracked slightly beyond the poles of the Moon. The final gravity field product was the 70×70 model GLGM-2, Lemoine *et al.* (1997). Lunar Prospector in turn, through its low, polar orbit, has given scientists excellent near-side data, and will continue to do so through its extended, very low, mission. The first official gravity product is a 75×75 expansion called LP75G, Konopliv *et al.* (1998), which, as does GLGM-2, also includes both the Clementine data and data from earlier earlier missions (Lunar Orbiter I – V, Apollo 15/16 sub-satellites). The final (?) post-Prospector gravity model is expected to be a 90×90 model, due in early 1999 (Konopliv, priv. comm.). Nevertheless, none of these missions, or any mission for which only deep space tracking by means of Earth-based stations is available is able to overcome the very fundamental problem in lunar gravity modelling: the lack of a fully global, high-quality data set giving satisfactory sensitivity to satellite orbit perturbations over a large range of orbital frequencies.

Another problem faced in lunar gravimetry (as in any kind of satellite-based gravity modelling) is the fact that observations are made at satellite altitude, which requires a downward continuation, to determine the selenopotential at the surface (e.g. in terms of selenoidal undulations or gravity anomalies). Such a downward continuation is known to be an error amplifying

operation, giving rise to instability in the related inverse problem. Furthermore, gravity field modelling suffers from the facts that each measurement type has its own sensitivity depending on orbital frequency and that measurement and dynamical modelling errors (gravitational and non-gravitational) are always present.

In view of these limiting factors in lunar gravity modelling and the numerical problems faced in the related inverse problems, the scope of this paper is to analyse existing models in terms of quality and information content. It is aimed to assess the true value of the existing satellite tracking data for gravity modelling purposes, as well as their limitations. Given the aforementioned limiting factors, it is clear that regularisation (constraints) is a fundamental aspect of lunar gravity field modelling. In lack of an adequate data set, regularisation itself becomes an involved issue. A second aim of this paper is therefore to investigate the role of regularisation, and in particular the choice of the *optimal* regularisation parameter. Historically speaking, regularisation of planetary gravity fields has typically been a scaled variant of Kaula's rule-of-thumb, Kaula (1966). However, the choice of the scaling factor for this rule has been (and still is) a matter of discussion, as long as the true power law is not known until the gravity field itself is known.

Applying Kaula's rule inevitably means global smoothing, that is, smoothing is also applied where it is not wanted (over-smoothing may take place). For this reason, variants of the global Kaula constraint have been developed, cf. Konopliv and Sjogren (1996), in which fictitious measurements are added in areas where the error exceeds the signal. These additional observations are typically based on some a priori power model (e.g. Kaula). Such constraints are known to allow higher peak effects and may be viewed as an indirect Kaula-type constraint. A completely different approach is advocated in this paper. Instead of relying on a priori knowledge of the selenopotential, it is proposed to disregard the physics of the Moon, and hence not depend on largely unknown a priori coefficient power estimates, and rather look at the problem from a numerical point of view. Two heuristic methods for regularisation parameter determination are proposed and investigated for gravity reduction purposes, being the L-curve and the Quasi-optimality criterion, both of which seek to balance the data error (deduced from the available tracking data) and the solution bias (which is introduced by the actual regularisation process). Furthermore, one a posteriori parameter choice rule is discussed for comparison, the method of Quasi-solutions.

Finally, this paper advocates alternative error measures to be used in error assessment studies of gravity modelling. In particular in the case of the Moon, where the current data distribution is severely heterogeneous, and a significant bias is introduced by the very fact that regularisation is required to enable any extended solution at all, this bias should be accounted for. The mean square error matrix, which is the sum of the propagated error and the squared bias, is therefore proposed as an alternative to pure error variance-covariance propagation.

The first model investigated is GLGM-2, as this is the model currently used by both the lunar science community and space mission planners in orbit design. However, in view of the ongoing efforts for missions involving inter-satellite tracking, which will enable spacecraft tracking over the lunar far-side and hence a global set of observations, simulated satellite-to-satellite tracking solutions will also be used on occasion. An important point is that such missions may give nearly self-contained gravity solutions requiring only little amount of regularisation, with due benefits for the lunar geosciences and low lunar satellite dynamics.

2 The linear model and its solution

In general the relationship between the observations $y^e = y + e$, with y the 'exact' observation and e the data noise, and the unknowns x is non-linear, i.e. $E\{y^e\} = E\{y\} = A(x)$, where A is the functional relationship between the observations and the unknowns. In this paper the unknowns are parameterised in terms of global spherical harmonics basis functions. $E\{\}$ is the expectation operator and it is assumed that $E\{e\} = 0$ and $E\{ee^T\} = P^{-1}$, i.e. model errors are not considered in the present analysis. Inversion for the selenopotential is done by linearising the relationship between observations and unknowns and solving for corrections to the a priori model in an iterative fashion. Hence, in practical cases one has

$$y^e = y + e = Ax + e \tag{1}$$

where $E\{y^e\} = E\{y\} = Ax$, where A is now the design matrix. The vector of unknowns of course contains the corrections to the gravity field coefficients $\{\bar{C}_{lm}, \bar{S}_{lm}\}$, other unknowns are not considered here. The least-squares solution $\hat{x}_{ls}$ is the best linear unbiased estimate of the solution. Taking into account a possible weighting of different data sets in the solution, represented by the weight matrix P (inverse error covariance of the measurements), the familiar least-squares solution reads

$$\hat{x}_{ls} = \left(A^T P A\right)^{-1} A^T P y^e. \tag{2}$$

If the problem at hand is perfectly observable, $\hat{x}_{ls}$ may be considered a stable solution to our inverse problem. However, the complications of the measurements being inhomogeneously distributed in space and in time, as well as the damping of orbital perturbations with altitude causes $\hat{x}_{ls}$ to be unstable. That is, small data errors may cause large errors in the solution. In theory, this may be handled by either 1) limiting the number of unknowns to the ones with large singular values (i.e. those for which direct estimation is largely possible), or 2) applying regularisation to the normal equation system. While the first method may work for the estimation of small spherical harmonics expansions, it is known that the lunar potential is dominated by mass concentrations which are impossible to represent in terms of low-degree harmonics. Secondly, for gravity solutions with a resolution close to the satellite altitude, the signal of coefficients or groups of coefficients (lumped coefficients) will remain below the data noise level and it might still not be possible to compute a stable $(A^T P A)^{-1}$. This is further illustrated by computing the singular value decomposition of the design matrix $A = U\Sigma V^T$, where U and V are orthonormal matrices spanning the space of the observations and the measurements, respectively, and Σ is a diagonal matrix with elements $\{\sigma_i | \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n > 0\}$ and $\lim_{i\to\infty} \sigma_i = 0$. Let $P = I$ (not essential in this context), then

$$\begin{aligned}
\hat{x}_{ls} &= (V\Sigma U^T U \Sigma V^T)^{-1} V \Sigma U^T y^e \\
&= (V\Sigma^2 V^T)^{-1} V \Sigma U^T y^e = V\Sigma^{-1} U^T y^e \\
&= V\Sigma^{-1} U^T (y + e) \\
&= x + V\Sigma^{-1} U^T e.
\end{aligned} \tag{3}$$

Thus, the least-squares solution consists of the exact solution x from exact observations y and a term which depends on the noise e. Because the elements of Σ tend to zero, due to the essentially one-hemisphere data coverage and the satellite altitude, the noise is amplified with a large number and instability occurs.

A priori information on the coefficient power as given by Kaula may be used to increase stability. With the elements of the diagonal Kaula matrix K given by $10^{10} \times l^4$, the regularised least-squares solution becomes

$$\hat{x}_\alpha = (A^T P A + \alpha K)^{-1} A^T P y^e, \tag{4}$$

where α is the regularisation parameter scaling Kaula's rule. Eq. 4 is equivalent to the least-squares collocation solution, Marsh *et al.* (1988). However, due to the constraint, it is also biased towards zero, which in principle is impossible for a collocation solution. Since in this study no assumption is made on the mean value of the estimated parameters, the collocation framework for error assessment is considered unsuitable, as the bias should be taken into account. Hence, it is more appropriate to consider Eq. 4 as a biased estimator. The expectation of Eq. 4 is

$$\begin{aligned} E\{\hat{x}_\alpha\} &= (A^T P A + \alpha K)^{-1} A^T P E\{y^e\} \\ &= (A^T P A + \alpha K)^{-1} A^T P A x \\ &\neq x, \end{aligned}$$

and the bias introduced by the constraints is

$$E\{\hat{x}_\alpha - x\} = \delta x = -(A^T P A + \alpha K)^{-1} \alpha K x. \tag{5}$$

The bias may become larger than the coefficient value. One difficulty with the bias computation is that the true solution x is involved. Obviously, these coefficients are unknown and estimating the bias with biased coefficients will give too optimistic results since the power of these coefficients is too low, Xu (1992);Xu and Rummel (1994). This is another reason why a pure variance-covariance propagation may lead to optimistic estimates of the true solution quality. For GLGM-2, which forms the test case of the present analysis, the coefficient biases are larger than 50% of the coefficient value for nearly all harmonics beyond degree and order 30, compare Section 5.2.1.

3 Determination of the regularisation parameter for biased estimators

Three methods for the determination of the optimal regularisation parameter α under the presence of a bias are considered, being the L-curve, the Quasi-optimality criterion and the method of Quasi-solutions, cf. Hansen and O'Leary (1993); Morozov (1984); Ivanov (1962). The first two methods are so-called heuristic methods while the latter is an a posteriori method. The ideas of both heuristic methods are roughly the same. Consider the minimisation of

$$J(\alpha) = \|Ax - y\|_P^2 + \alpha \|x\|_K^2$$

which leads to Eq. 4. The norm on the left is the least-squares problem. Minimisation of this term gives the smallest errors, but the norm of the solution is unconstrained. The second term controls the solution smoothness; however, this smoothness condition also causes the solution to become biased. The parameter α controls the compromise between smoothness of the solution, i.e. $\|x\|_K$ remains small, and data fit, $\|Ax - y\|_P$ small. Both the L-curve and the Quasi-optimality criterion aim to find an α such that other α's close to it yield a comparable solution. In other words, the methods seek to balance the data error with the regularisation error.

3.1 L-curve

The L-curve is a plot, for all valid α, of the norm $\|\hat{x}_\alpha\|_K$ of the stable solution versus the corresponding residual norm $\|A\hat{x}_\alpha - y\|_P$. It turns out that the L-curve, when plotted in *log-log* scale, has an L-shaped appearance. The vertical part of the curve corresponds to smaller α. The emphasis of minimising $J(\alpha)$ is on $\|A\hat{x}_\alpha - y\|_P$, allowing $\|\hat{x}_\alpha\|_K$ to become large. The horizontal part of the L-curve corresponds to solutions where the residual norm $\|A\hat{x}_\alpha - y\|_P$ is most sensitive to the scaling parameter because $\hat{x}_\alpha$ is dominated by the bias, compare Hansen and O'Leary (1993); Hansen (1997). The corner of the L-curve is optimal in the sense that the change of both norms is equal for changing α. This point may be located by maximum curvature. However, it may be shown that the minimum of the function

$$\psi(\alpha) = \|\hat{x}_\alpha\|_K \|A\hat{x}_\alpha - y\|_P \tag{6}$$

gives the same scaling parameter, Regińska (1996). Note that Eq. 6 cannot be computed directly since the design matrix A is not available. This is generally the case in gravity field estimation from satellite tracking data where normal matrices for single arcs are combined to form a final normal equation system. However, the data errors behave like σ_i^{-1}, compare Eq. 3, which can be determined by eigenvalue techniques. Hence, real observations y^e may be approximated by Ax^e where $x^e = x + e$. The obvious choice for the exact solution x is the GLGM-2 solution itself, the errors e are assumed to be given by $3N(0,1)/\sigma_i \times$ coefficient sigma, with $N(0,1)$ the standard normal distribution. An upper bound for the second norm in Eq. 6 is now given by

$$\begin{aligned} \|A\hat{x}_\alpha^e - y^e\|_P = \|A\hat{x}_\alpha^e - Ax^e\|_P &\leq \|A\| \, \|\hat{x}_\alpha^e - x^e\|_K \\ &= \sqrt{\lambda_1}\|\hat{x}_\alpha^e - x^e\|_K \end{aligned}$$

where λ_1 is the largest eigenvalue of A^TPA which can easily be obtained with the power method, Kreyszig (1988).

3.2 Quasi-optimality

The idea of the Quasi-optimality criterion is that if α is too small, $\hat{x}_\alpha$ is dominated by the data error which is now sensitive to small changes in α. On the other hand, if α is too large, $\hat{x}_\alpha$ is dominated by the bias which is now sensitive to small changes in α. The optimal α is obtained when the size of both norms is about equal. Hence, the L-curve and Quasi-optimality are alike. Morozov (1984) derives the Quasi-optimality equations as follows. Let $\hat{x}_{\alpha_1}$ be the solution of the minimisation problem

$$\min_\alpha \|Ax - y\|_P^2 + \alpha\|x - x_0\|_K^2$$

where x_0 is an initial guess, e.g. $x_0 = 0$. Furthermore, let $\hat{x}_{\alpha_2}$ be the solution of

$$\min_\alpha \|Ax - y\|_P^2 + \alpha\|x - \hat{x}_{\alpha_1}\|_K^2.$$

The two consecutive solutions $\hat{x}_{\alpha_1}$ and $\hat{x}_{\alpha_2}$ are related as

$$\begin{aligned} \hat{x}_{\alpha_1} &= (A^TPA + \alpha K)^{-1}(A^TPy + \alpha K x_0) \\ \hat{x}_{\alpha_2} &= (A^TPA + \alpha K)^{-1}(A^TPy + \alpha K \hat{x}_{\alpha_1}) \\ &= \hat{x}_{\alpha_1} - \alpha(A^TPA + \alpha K)^{-1}(x_0 - \hat{x}_{\alpha_1}) \\ &= \hat{x}_{\alpha_1} - \alpha\frac{d\hat{x}_\alpha}{d\alpha} \end{aligned}$$

The latter equality can be checked by straightforward calculation. The optimal value of α is obtained by minimising the change from one solution to the next, that is, $\alpha\|d\hat{x}_\alpha/d\alpha\|$ is minimised.

3.3 Quasi-solutions

The third parameter choice rule considered is the method of Quasi-solutions. The regularisation parameter α is chosen such that the solution $\hat{x}_\alpha$ satisfies

$$\|\hat{x}_\alpha\|^2 = c^2,$$

where c is some a priori norm bound on the signal $\hat{x}$. Because $\hat{x}_\alpha$ are the potential coefficients, c can be determined with Kaula's rule for the Moon's gravity field. Numerically, the regularisation parameter can be obtained by Newton iteration, Press *et al.* (1992). In practice, one solves

$$z(\alpha) = \|\hat{x}_\alpha\|^2 - Rc^2 < \varepsilon$$

with $R \in \mathbb{R}^+$ to scale the power and ε some small number.

4 Quality assessment

The quality assessment of the computed solution is an important task in gravity field modelling. It is, however, also a difficult task since none of the quality assessment methods is capable of describing the 'quality' in all circumstances. It matters whether one wants to use the gravity field model, for example, for orbit determination or for selenoid/geoid determination. The latter requires global quality measures while the former does not if one is interested in the orbit of one specific satellite. Furthermore, quality assessment is hampered by the fact that model errors exist and that the solution is biased.

This paper is concerned with global quality measures. The measures make use of the gravity field solution, and the estimated errors of the coefficients are propagated to selenoid errors. The contribution of the observations to the solution, signal-to-noise ratio, bias-to-noise ratio, etc. are measures for the coefficients themselves and will be briefly discussed hereafter.

4.1 Mean square error matrix

The mean square error matrix is introduced as a measure of the error in lunar gravity models. Since the solution is biased, it has two parts: the propagated error Q_x and the squared bias. For the same reason, it is also considered a more realistic error measure than the more commonly used variance-covariance matrix. Nevertheless, effective use of the mean square error matrix requires reliable estimation of the solution bias, a problem mentioned in Section 2. Error propagation yields, compare Eq. 4,

$$Q_x = (A^TPA + \alpha K)^{-1}A^TPA(A^TPA + \alpha K)^{-1}. \tag{7}$$

The mean square error matrix MSEM is therefore

$$MSEM = Q_x + \delta x \delta x^T \tag{8}$$

with the bias term given by Eq. 5. Identical to the case of error variance-covariance matrices for unbiased estimators, the MSEM may be subject to error propagation, for example to selenoid height errors or to gravity anomaly errors, cf. Haagmans and van Gelderen (1991).

4.2 Ratio measures

The size of the estimated coefficients with respect to their uncertainty indicates how well a certain coefficient is resolved. This is called the signal-to-noise ratio (SNR) and is defined as

$$SNR_{lm} := \frac{|K_{lm}|}{\sigma_{lm}}$$

with K_{lm} the estimated coefficients and σ_{lm} their uncertainty, which is the square root of the corresponding diagonal element of the MSEM. Ideally the SNR is larger than one for each coefficient, in which case there is more signal than noise.

A second ratio measure is the bias-to-noise ratio (BNR). One could use

$$BNR_{lm} := \frac{[\delta x \delta x^T]_{lm}}{[Q_x]_{lm}}$$

or $BNR := \text{trace}(\delta x \delta x^T)/\text{trace}(Q_x)$. If the BNR is small, this means the bias can be neglected.

Finally, it might be interesting to look at the bias-to-signal ratio (BSR):

$$BSR_{lm} := \left| \frac{\delta x_{lm}}{K_{lm}} \right|.$$

The larger the BSR, the smaller the signal with respect to the bias.

4.3 Contribution measures

It is rather straightforward to derive that the regularised solution, Eq. 4, is also the solution of

$$E\{\begin{pmatrix} y^\varepsilon \\ z \end{pmatrix}\} = \begin{pmatrix} A \\ I \end{pmatrix} x, \quad D\{\begin{pmatrix} y^\varepsilon \\ z \end{pmatrix}\} = \begin{pmatrix} P^{-1} & 0 \\ 0 & [\alpha K]^{-1} \end{pmatrix}$$

with $z = 0$, that is, zero observations for all unknowns. If $E\{0\} = x$ holds true, the solution is unbiased and the error covariance matrix is $Q_x = (A^T P A + \alpha K)^{-1}$. The contribution of the observations to the solution of the unknowns is now defined as

$$contr_y := Q_x Q_{x,y}^{-1} = (A^T P A + \alpha K)^{-1} A^T P A \tag{9}$$

where $Q_{x,y}^{-1}$ is the least-squares normal matrix. The diagonal elements of $contr_y \in \{0, 1\}$. Zero means that the observations do not contribute to the determination of the unknown, i.e. all information comes from the constraint, whereas a unit value implies that the observations completely determine that specific unknown. Eq. 9 equals the gain matrix in Kalman filtering. It can be shown that the diagonal elements of Eq. 9 equal

$$contr_{y_i} = 1 - [Q_x]_{ii} \cdot [\alpha K]_{ii},$$

compare Bouman (1997). Numerically, this is an easy and stable computation.

A contribution measure for biased solutions has been developed by Bouman (1997):

$$\begin{aligned} contr_y &:= MSEM \times MSEM^{-1}\Big|_{\alpha=0} \\ &= (A^T P A + \alpha K)^{-1}(A^T P A + \alpha^2 K x x^T K)(A^T P A + \alpha K)^{-1} A^T P A. \end{aligned} \tag{10}$$

Eqs. 9 and 10 are equal if $K^{-1} \approx \text{diag}(xx^T)$ for $\alpha = 1$ and neglecting the off-diagonal terms of xx^T. Note that the diagonal elements of Eq. 10 may become larger than one due to the bias term. It is, therefore, somewhat more complicated to interpret $contr_y$. Furthermore, three matrix multiplications are involved, yielding greater sensitivity to numerical round-off errors.

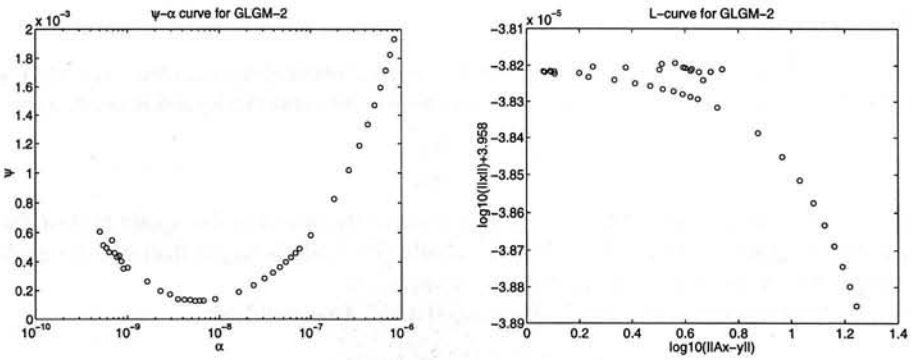

Fig. 1. L-curve results for GLGM-2 normal matrix. The minimal ψ is easily found (left), whereas the L-shape is less clearly present (right). The optimal α is found to be 0.545×10^{-8}, corresponding to a Kaula rule of $13545 \times 10^{-5}/l^2$.

5 Results and discussion

5.1 Parameter choice rules

5.1.1 GLGM-2

Quasi-solutions. The method of Quasi-solutions did not give any reasonable regularisation parameter. For the GLGM-2 case the method proved to be extremely sensitive to the choice of the scaling parameter R. The Newton iteration only converges for a small range of R values. Taking R, for example, 10% larger or smaller resulted in divergence. Even when the iteration converged, the final regularisation parameter was of the order 10^{15} or larger, and therefore useless. Our conclusion, therefore, is that the method of Quasi-solutions is not suitable for GLGM-2 regularisation. Although the method may seem an attractive parameter choice rule at first sight, it turns out to be of no value. The problem of selecting a proper regularisation parameter shifts to the problem of choosing a proper scaling parameter R for the signal bound.

L-curve and Quasi-optimality criterion. The regularisation parameter α estimated by the L-curve method, Fig. 1, amounts to 0.545×10^{-8}, which corresponds to a Kaula rule of $\{\bar{C}_{lm}, \bar{S}_{lm}\} \sim 0 \pm 13545 \times 10^{-5}/l^2$. Clearly, this is a much more relaxed constraint (by a factor ~ 900) than what has been applied in GLGM-2. The Quasi-optimality criterion, on the other hand, predicts an even slightly smaller α, but the method fails to produce a clear minimum for the distance (in the functional space) between two neighbouring x_α's.

The α as found by the L-curve method has been applied in an inversion of the GLGM-2 normal equation to produce a fictitious new lunar gravity model depicted in terms of selenoid heights and corresponding selenoid height errors from the propagated covariance matrix in Fig. 2, which are both shown in Hammer projections centered at 270° longitude, such that the near-side is depicted on the right hand side of the plot, and vice versa, the entire far-side is shown on the left hand side of the plot. Fully white areas in the selenoid height plots are off the scale.

Two main conclusions may be drawn from these plots: first, the L-curve and Quasi-optimality methods are not methods for the prediction of the gravity coefficient power, i.e. they are mathematical optimisation methods for the total error based on the observability of

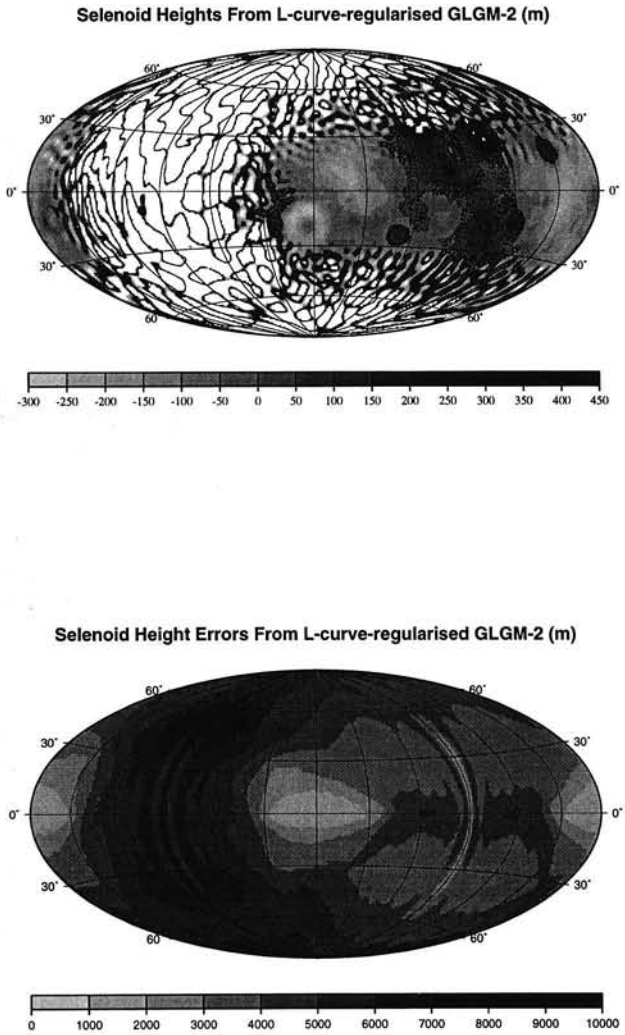

Fig. 2. Selenoid heights on the basis of the GLGM-2 normal equation, solved by optimal regularisation according to the L-curve method (top), and the corresponding selenoid height errors from the propagated error covariance matrix (bottom). The global rms of the errors amounts to 4699 m.

each coefficient (which depends primarily on the data coverage, observation type, orbit characteristics, etc.). The methods seek to balance the regularisation error with the data error, while at the same time remain as close as possible to the *true* inverse. Such a small α therefore yields truly large selenoid errors in areas with no measurements. The fact that lunar gravity solutions from Earth-based Doppler measurements depend for a large part on the constraint, implies that these methods optimise the recovery of the gravitational potential only in areas largely covered with measurements.

Therefore, while the error plot depicts realistic errors in a modern lunar gravity model, no guarantee is given for the global usefulness of the solution obtained in terms of selenophysics or satellite orbit modelling. That is, none of these direct methods for the estimation of the optimal regularisation parameter is able to ensure solid estimation of global basis function parameters based on near-side data only. Evidently, some realistic power constraint, either in the form of 1) a global coefficient Kaula rule, or 2) some selenographical constraint adding fictitious measurements in far-side and high-latitude areas, needs to be applied. The former is the traditional way, but it also carries the property of smoothing where smoothing is not wanted. For present-day solutions, this argues in the favour of the selenographical type of constraint, and this should be investigated and compared to the Kaula rule in the near future. Both these latter methods will suffer from a truly large bias, but find their merit in the fact that they are able to produce a solution suitable for geophysical interpretation and orbit determination.

It deserves mentioning that in the case of GLGM-2 the size of the harmonic expansion is extreme compared to the actual information available in the tracking data, compare Section 5.3. For a much smaller expansion, e.g. 15×15, the overall parameter observability is significantly better, and the derived selenoid error will accordingly be much smaller. Moreover, the L-curve method and the Quasi-optimality criterion have been tested for cases with reduced data errors, or similarly, improved data coverage, Bouman (1998), and have proven their ability.

5.1.2 SST

Given the fact that knowledge of the gravity field of the Moon is a necessary tool and sometimes a boundary constraint for a number of other lunar sciences, overcoming the present problems in lunar gravimetry remains one of the highest priorities in lunar science, Lemoine *et al.* (1997). Currently, one satellite mission to the Moon is intended to carry a instrument for inter-satellite tracking between a mother spacecraft and a sub-satellite, SELENE Project Team (1996). In view of these on-going mission preparations, the L-curve method has furthermore been applied in the analysis of the required regularisation for gravity field solutions derived from satellite-to-satellite tracking data. The gravity field solutions investigated here are all 70×70 spherical harmonics expansions.

To this end, several cases of 2-way low-low SST have been studied, all flying at 100 km altitude, but with varying inclination, and, hence, polar gaps. The investigated inclinations are 90°, 85° and 80°. Tandem configurations, i.e. configurations with purely a separation in mean anomaly of 3°, or a spacing of $\sim$ 100 km are chosen, although it may be proven that so-called *en echelon*, or butterfly, configurations will outperform the more simple tandem in terms of estimation error. The assumed tracking precision is 0.1 mm/s range-rate. Relatively sparse ground tracking, range and range-rate from two ground stations, at 3 m and 0.3 mm/s weights is further necessary for orbit determination. The gravity field solutions, however, are largely dominated by the SST link. All solutions, i.e. one for each inclination, are combined solutions from 4 one-week arcs, with one state-vector and one solar radiation pressure coefficient estimated per

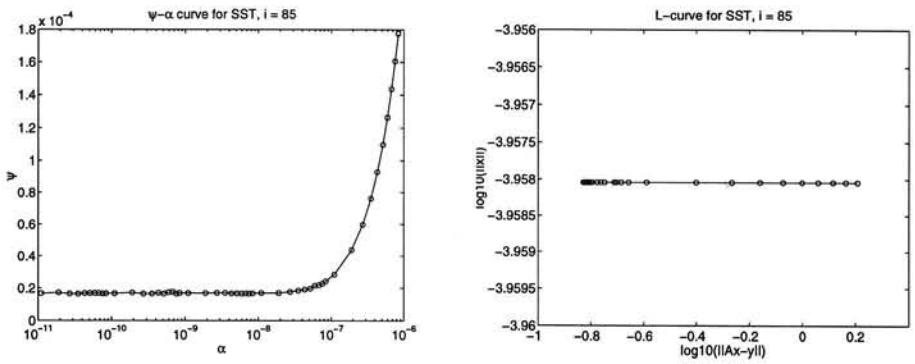

Fig. 3. L-curve results for an SST-based gravity field solution at 85° inclination. No obvious minimum is found for ψ, and the L-curve remains flat.

week, in addition to the potential coefficients.

Fig. 3 shows that with respect to the pure least-squares solution the reduction in data error by regularisation does not compensate for the increase in bias. The $\psi(\alpha)$ curves exhibits no clear minimum, and the L-curve remains in the flat region. Hence, from an L-curve point of view, regularisation is not required in this case. Nevertheless, it is expected that further expansion of the harmonic series, e.g. up to degree and order 120, will show an increase in data error, and hence increase the demand for regularisation. This work is still on-going.

5.2 Bias computations

5.2.1 GLGM-2

In biased estimation one is confronted with the problem of determining the true value of the coefficient biases. The aim is evidently to establish some sort of measure for the true effect of regularisation on the estimated spherical harmonics expansion. When, as in the present case of GLGM-2, the coefficient solutions are significantly biased, applying these solutions in bias computations will lead to severe underestimation of the bias. This is illustrated here, where the GLGM-2 bias is computed for two cases: one in which the coefficient bias is directly applied, and one in which the bias is computed based on the sign of the GLGM-2 coefficients, but the amplitude is based on Kaula's rule, $15 \times 10^{-5}/l^2$, a case further referred to as the case of *synthetic bias*.

For GLGM-2, it is seen from Fig. 4 that the case of a synthetic bias, which is considered a realistic bias estimate, by far exceeds the coefficient solution for a vast range of harmonics. This may be seen as yet another proof that the satellite tracking data incorporated in GLGM-2 do not contain enough information to determine such a large spherical harmonics expansion as GLGM-2. In the case of simple coefficient biases, the bias-to-signal ratio will tend to unity for the higher orders, simply because the coefficients are fully determined by the Kaula constraint.

5.2.2 GLGM-2 regularised by the L-curve metod

The bias-to-coefficient ratios have also been determined for the fictitious solution of the GLGM-2 normal equation system, with the regularisation parameter α as determined by the L-curve

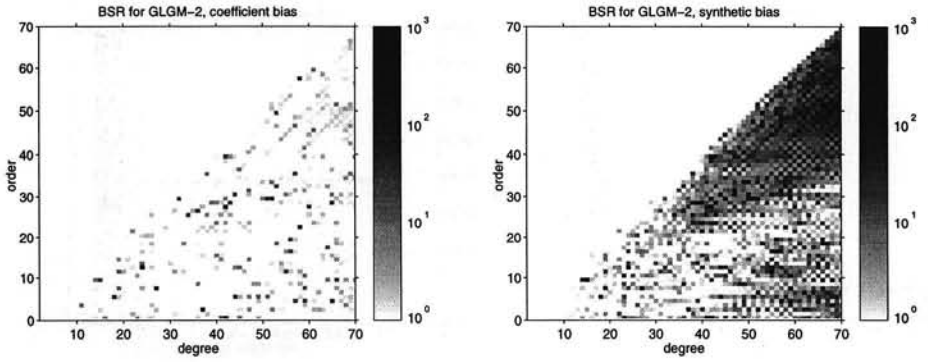

Fig. 4. Bias-to-signal ratios for the $\bar{C}_{lm}$ parameters in the GLGM-2 normal matrix. The coefficient biases are on the left, and the synthetic biases on the right. The BSR of the $\bar{S}_{lm}$ coefficients are similar.

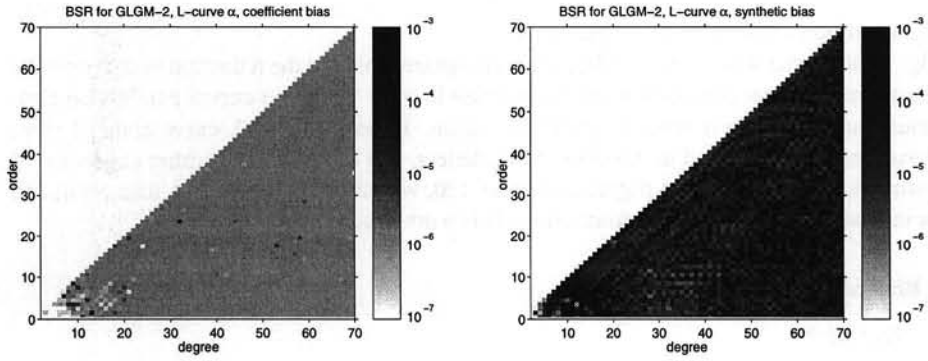

Fig. 5. Bias-to-signal ratios for the $\bar{C}_{lm}$ parameters in the GLGM-2 normal matrix, with the regularisation parameter α determined by the L-curve method. The coefficient biases are on the left, and the synthetic biases on the right. The BSR of the $\bar{S}_{lm}$ coefficients are similar.

method, i.e. $\alpha = 0.545 \times 10^{-8}$.

Figure 5 shows that, because the bias is very small, both for the case of pure coefficient biases as well as for the case of synthetic biases, the bias-to-signal ratios remain small, and effectively, the effect of the bias on the gravity field solution remains small. Nevertheless, also in this case, there is a substantial difference, up to several orders of magnitude, between the coefficient bias case and the synthetic bias case.

5.3 Information content

5.3.1 GLGM-2

Signal-to-noise ratio, no bias. Assuming that the solution is unbiased, the square root of the diagonal elements of $(A^T P A + \alpha K)^{-1}$ are the standard deviations of K_{lm}. Their ratio is plotted in Fig. 6, the power indicates the number of significant digits a specific coefficient has. It is clear that most of the coefficients are well below the measurement noise.

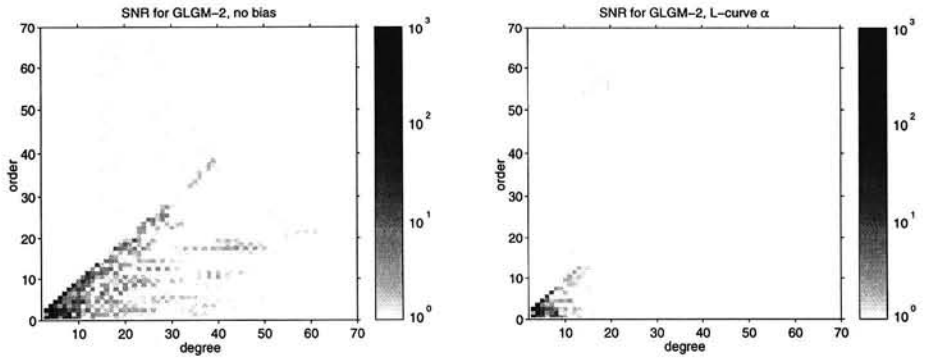

Fig. 6. Signal-to-noise ratio for GLGM-2, no bias (left) and L-curve α (right). Shown is the SNR of the $\bar{C}_{lm}$ coefficients.

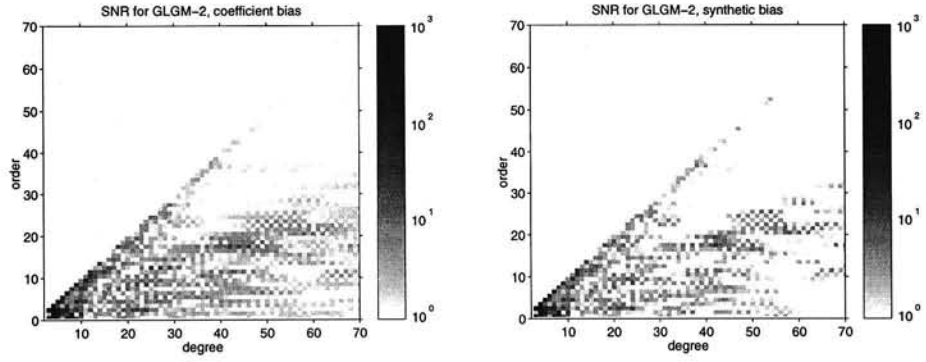

Fig. 7. Signal-to-noise ratio for GLGM-2, coefficient bias (left) and synthetic bias (right). Shown is the SNR of the $\bar{C}_{lm}$ coefficients.

This situation is even worse for the L-curve α, Fig. 6. Although the coefficients have more power compared to the original GLGM-2 solution, the error increases as well, and this increase apparently overwhelms the coefficient power increase. The bias has been neglected since it is small, Fig. 5.

Signal-to-noise ratio, bias. The expected errors of K_{lm} are the diagonal elements of Eq. 8. The ratio $|K_{lm}|/\sigma_{lm}$ has been plotted in Fig. 7 for the coefficient and synthetic bias respectively. One sees that the synthetic bias approximately yields the same results as the unbiased case, while the coefficient bias seems to be too optimistic.

Therefore, it is concluded that using the unbiased assumption does not influence the SNR compared to the unbiased case, as long as the bias estimate is 'correct'. The true bias is unknown, however, which one should realize interpreting any statement concerning the quality.

Contribution, unbiased case. The contribution of the observations to the original GLGM-2 solution is depicted in Fig. 8. The results for the $\bar{S}_{lm}$ coefficients are not shown since they are

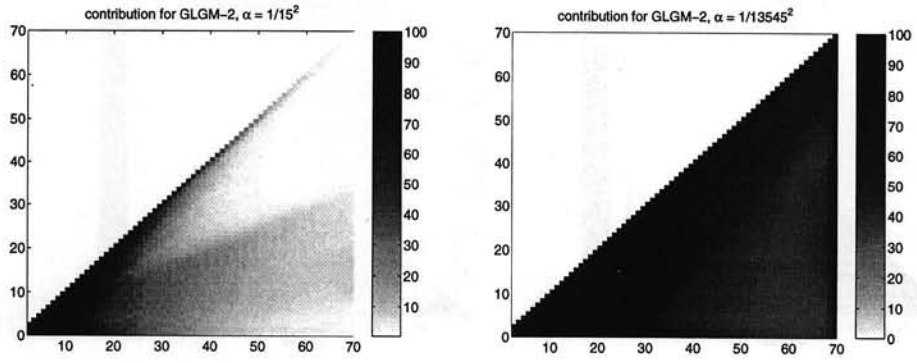

Fig. 8. Contribution of the observations to the original GLGM-2 solution, $\alpha = 1/15^2$ (left) and the L-curve solution, $\alpha = 1/13545^2$ (right). Shown are the $\bar{C}_{lm}$ coefficients, the $\bar{S}_{lm}$ coefficients are similar.

similar to the $\bar{C}_{lm}$ coefficients. It is clear that only the coefficients up to degree and order 15 have an observation contribution larger than 50%. With the exception of the sectorial terms, the contribution quickly drops to low levels. This means that these coefficients can hardly be recovered from the measurements and are likely to be biased.

As the regularisation parameter determined with the L-curve is much smaller than the original α, the contribution of the observations is expected to be much larger. This holds true indeed, Fig. 8. The minimum contribution is 35%, and approximately 68% of the coefficients has a contribution larger than 50%.

Contribution, biased case. Using the contribution measure, Eq. 10, for the biased case is for several reasons not trivial: 1) The bias has to be estimated, which is not straightforward, i.e. the choice of the bias will affect the computation; 2) The contribution can become both larger than 100% and also negative, which is difficult to interpret. On the theoretical side, the bias term of Eq. 10 destroys the symmetry of the matrix. Furthermore, for the same reason, it is not trivial to prove that the diagonal elements are necessarily positive, nor that the overall $contr_y$ matrix is positive (semi-)definite. Numerical round-off errors, due to the involved matrix multiplications may also play a part.

Therefore, it is concluded that although the contribution measure (10) for biased estimates is equivalent to Eq. 9 for unbiased estimators, further research is required to understand its behaviour and interpretation. Results are therefore not shown.

5.3.2 SST

Signal-to-noise ratio. The signal-to-noise ratio for a satellite-to-satellite tracking mission with an inclination of 85 degrees is shown in Fig. 9. Compared to the GLGM-2 solutions, the improvement is dramatic. Almost all coefficients can be determined. The current solution is a least-squares solution, there is therefore no bias. These results show that from an SST mission, coefficients above degree and order 70 may be determined as well, as the SNR still is greater than one. However, solving for higher degrees yields unstable solutions and again regularisation will be necessary.

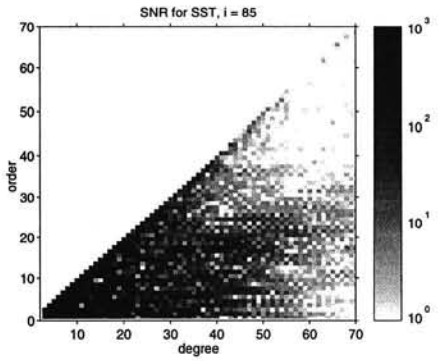

Fig. 9. Signal-to-noise ratio for SST, inclination is 85°. Shown is the SNR of the $\bar{C}_{lm}$ coefficients.

Contribution. Obviously, the contribution is 100% for all coefficients of a l.s. solution.

5.4 Error propagation

Error variance-covariances as well as mean square error matrices are propagated to formal errors in selenoidal undulations for both GLGM-2 and the SST-based solution..

5.4.1 GLGM-2

The propagated error covariance for GLGM-2 in terms of selenoid heights is shown in Fig. 10, yielding an overall rms value of 15.59 m. Taking the bias contribution into account as well, one obtains Fig. 11. For the coefficient bias case, the overall rms error amounts to 8.35 m, notably smaller than what results for the pure covariance propagation, while the synthetic case yields an rms value of 11.97 m. On the other hand, the synthetic bias, likely to represent a more realistic measure of the true bias, yields a larger error extremes in the selenoid than does the error covariance, with differences up to a factor of 3.5. The overall rms value remains lower, however, due to the fact that a large portion of the far-side exhibits smaller errors. Such behaviour is explained by the bias term. For some coefficients the bias is positive, while for others it is negative. It is therefore not evident that the use of the MSEM instead of the error covariance necessarily leads to larger formal selenoid errors. Xu (1992) reports similar results for geopotential estimation from satellite gravity gradiometry, and also shows that the degree-wise contribution to the error may increase or decrease depending on the choice of the bias.

Furthermore, the similarity between covariance-based and MSEM-based selenoid errors is less evident from the synthetic bias. In this case, the error appears more scattered, which may be explained from the use of sign according to the model, but amplitudes coming from the Kaula rule.

5.4.2 SST

The L-curve results for the SST-based solutions, Section 5.1.2, all showed that no regularisation is necessary for inclinations in the range of $90° \pm 10°$, when limiting the spherical harmonics expansion to 70×70. As an example, the L-curve was depicted for the 85° inclination case, Fig.

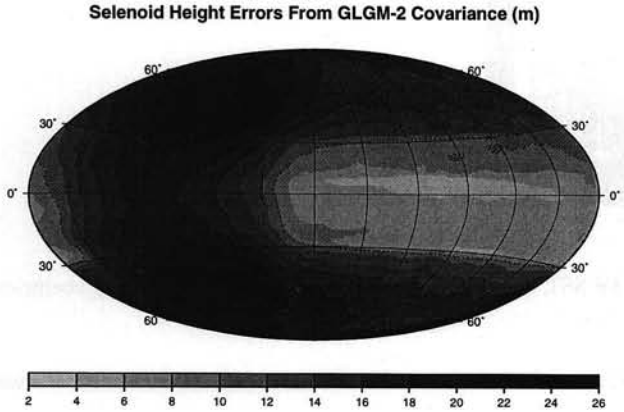

Fig. 10. Formal errors in selenoidal heights based on the GLGM-2 error covariance matrix.

3. For the same case, the error covariance is here propagated to selenoid errors, Fig. 12. As no model errors are considered in the present research, this represents the limit case of the achievable accuracy from 4 weeks of satellite-to-satellite tracking, i.e. roughly two global sweeps of the Moon surface (ground track spacing is only about 1° due to the slow lunar rotation). The overall rms error amounts to 0.48 m, which proves the enormous benefit of the SST-based solutions above solutions derived from more conventional Doppler observations collected by deep space antennae. Compared to the current situation, SST may improve our knowledge of the lunar potential by several orders of magnitude and also relieve lunar science from its strict dependence of a priori power estimates. Furthermore, the quality dichotomy between the far-side and the near-side is reduced to simply the small additional contribution of the Earth-based Doppler observations.

6 Conclusions

The fundamental problem in lunar gravity field modelling - the lack of a global satellite tracking data set - seriously affects the quality of the solutions derived till date. Downward continuation appears to be less of a problem for spherical harmonics expansion up to degree and order 70×70, since the SST-based solutions do not require much regularisation and because the orbits of Moon orbiting spacecraft may be very low, down to 50 – 100 km.

Solving for a full 70×70 solution from the tracking data of the GLGM-2 model is only possible through the use of a true coefficient power constraint. The alternative methods proposed in this paper, the L-curve method and the Quasi-optimality criterion predict a very small regularisation parameter, and hence yield gravity field solutions with km-level selenoid errors in

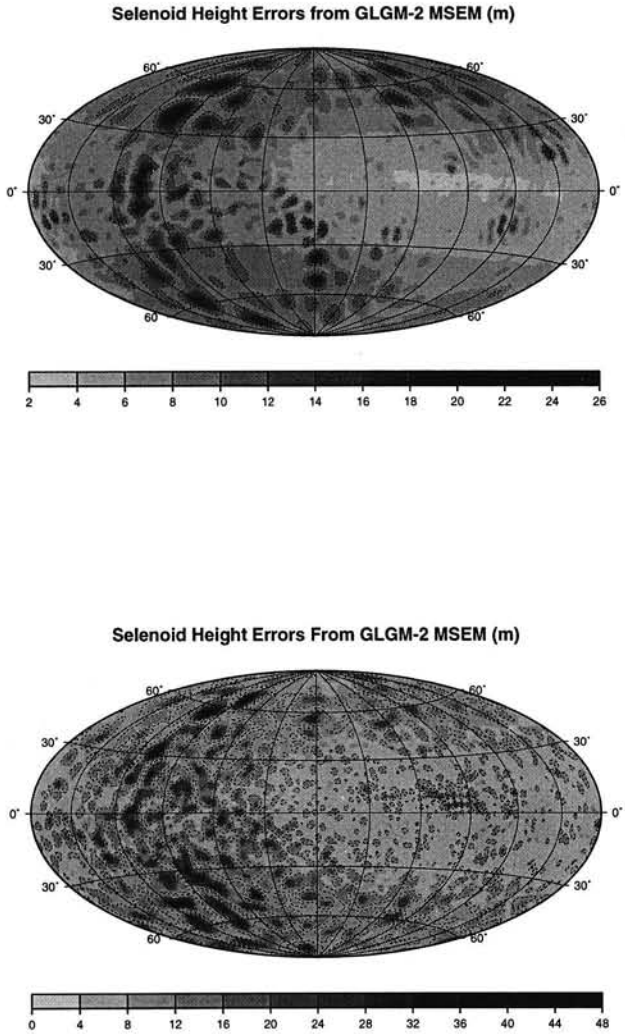

Fig. 11. Formal errors in selenoidal heights based on the GLGM-2 MSEM, using the coefficient biases (top), and using a synthetic bias, where the sign is taken according to the GLGM-2 solution, but the amplitude from Kaula's rule (bottom).

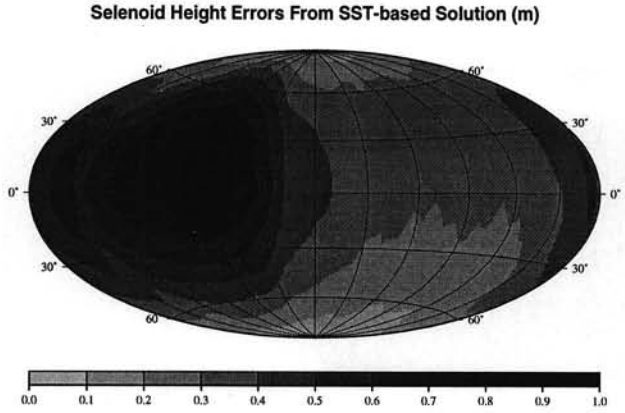

Fig. 12. Formal errors in selenoidal heights based on the error covariance matrix of the SST-based gravity field solution at $i = 85°$.

areas not covered with spacecraft observations. The heuristic regularisation parameter estimation algorithms are clearly not intended to predict the realistic power in gravity field solutions extended far beyond what is available in the tracking data, but merely balance the observation error with the bias the analyst introduces by regularisation. The L-curve solution moreover shows that the derived selenoid is selenophysically meaningful in areas covered by measurements only, which leads to the conclusion that the estimation of fully global basis functions, up to such high degree and order, is not possible.

The mean square error matrix is advocated as a more realistic quality measure than the simple error covariance matrix, as it also contains the bias contribution. A problem is, however, the estimation of the true solution bias, as a bias estimate on the basis of the biased coefficient solution always is too optimistic. The use of synthetic biases, based on a Kaula rule, nevertheless appears to give more realistic formal errors, and exceed those coming from the pure error covariance by a factor of two in the case of GLGM-2.

7 Further work

The heuristic regularisation parameter estimation methods will be applied to gravity models derived from Lunar Prospector tracking, for example LP75G. Moreover, it will be interesting to apply the L-curve and Quasi-optimality methods to extended SST-based gravity field solutions, for example a solution up to degree and order 120. Finally, it might be worthwhile to investigate iterative methods for regularisation parameter estimation, as a complement to the heuristic and a posteriori methods investigated thus far.

Acknowledgement The authors would like to thank Frank Lemoine of the NASA/Goddard Space Flight Center for providing the GLGM-2 normal equation. Furthermore, our appreciation goes to Radboud Koop and Pieter Visser for the many useful discussions and remarks. This work is supported by the Delft University of Technology's Centre for High Performance and Applied Computing (HPαC).

References

Binder, A. B. (4 September 1998). Lunar Prospector: Overview. *Science Magazine*, **281**(5382), 1476–1480.

Bouman, J. (1997). Quality assessment of geopotential models by means of redundancy decomposition? *DEOS Progress Letters*, **97.1**, 49–54.

Bouman, J. (1998). Quality of regularization methods. DEOS Report no 98.2, Delft Institute for Earth-Oriented Space Research.

Haagmans, R. and van Gelderen, M. (1991). Error variances-covariances of GEM-T1: their characteristics and implications in geoid computation. *Journal of Geophysical Research*, **96**(B12), 20011–20022.

Hansen, P. (1997). *Regularization Tools, A Matlab package for analysis and solution of discrete ill-posed problems, Version 2.1 for Matlab 5.0.* Department of Mathematical Modelling, Technical University of Denmark. http://www.imm.dtu.dk/~pch.

Hansen, P. and O'Leary, D. (1993). The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM J. Sci. Comput.*, **14**(6), 1487–1503.

Ivanov, V. (1962). Integral equations of the first kind and an approximate solution for the inverse problem of potential. *Soviet Math. Doklady*, **3**, 210–212.

Kaula, W. (1966). *Theory of satellite geodesy*. Blaisdell Pub. Co.

Konopliv, A. S. and Sjogren, W. L. (1996). *Venus Gravity Handbook*. Jet Propulsion Laboratory, Pasadena, California, jpl publication 96-2 edition.

Konopliv, A. S., Binder, A. B., Hood, L. L., Kucinskas, A. B., Sjogren, W. L., and Williams, J. G. (4 September 1998). Improved Gravity Field of the Moon from Lunar Prospector. *Science Magazine*, **281**, 1476–1480.

Kreyszig, E. (1988). *Advanced engineering mathematics*. John Wiley and Sons, sixth edition.

Lemoine, F., Smith, D., Zuber, M., Neumann, G., and Rowlands, D. (1997). A 70th degree lunar gravity model (GLGM-2) from Clementine and other tracking data. *Journal of Geophysical Research*, **102**(E7), 16339–16359.

Marsh, J., Lerch, F., Putney, B., Christodoulidis, D., Smith, D., Felsentreger, T., Sanchez, B., Klosko, S., Pavlis, E., Martin, T., Williamson, J. R. R., Colombo, O., Rowlands, D., Eddy, W., Chandler, N., Rachlin, K., Patel, G., Bhati, S., and Chinn, D. (1988). A new gravitational model for the earth from satellite tracking data: GEM-T1. *Journal of Geophysical Research*, **93**(B6), 6169–6215.

Morozov, V. (1984). *Methods for solving incorrectly posed problems*. Springer-Verlag.

Nozette, S., Rustan, P., Pleasance, L. P., Horan, D. M., Regeon, P., Shoemaker, E. M., Spudis, P. D., Acton, C. H., Baker, D. N., Blamont, J. E., Buratti, B. J., Corson, M. P., Davies, M. E., Duxbury, T. C., Eliason, E. M., Jakosky, B. M., Kordas, J. F., Lewis, I. T., Lichtenberg, C. L., Lucey, P. G., Malaret, E., Massie, M. A., Resnick, J. H., Rollins, C. J., Park, H. S., McEwen, A. S., Priest, R. E., Pieters, C. M., Reisse, R. A., Robinson, M. S., Simpson, R. A., Smith, D. E., Sorenson, T. C., Breugge, R. W. V., and Zuber, M. T. (1994). The Clementine Mission to the Moon: Scientific Overview. *Science Magazine*, **266**, 1835–1862.

Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (1992). *Numerical recipes in C: the art of scientific computing*. Cambridge University Press, second edition.

Regińska, T. (1996). A regularization parameter in discrete ill-posed problems. *SIAM J. Sci. Comput.*, **17**(3), 740–749.

SELENE Project Team (1996). SELENE PROJECT (SELenological and ENgineering Explorer), ISAS/NASDA Joint Moon Orbiting Satellite Project. Technical report, ISAS/NASDA, Tokyo, Japan.

Xu, P. (1992). The value of minimum norm estimation of geopotential fields. *Geophysical Journal International*, **111**, 170–178.

Xu, P. and Rummel, R. (1994). A simulation study of smoothness methods in recovery of regional gravity fields. *Geophysical Journal International*, **117**, 472–486.

DEOS Progress Letter 98.1: 21–29

Integration of a priori gravity field models in boundary element formulations to geodetic boundary value problems *

Roland Klees and Rüdiger Lehmann[1]

[1] Institute of Mine Surveying and Geodesy, Technical University of Freiberg, Germany

Abstract

Current high resolution geopotential models of the Earth are based on a combination of satellite and terrestrial data. Satellite data are well-suited to recover the long-wavelength features of the geopotential up to some degree N, whereas terrestrial gravity and height data fix the medium and short wavelengths. Usually, the recovering of the medium and short-wavelengths from terrestrial data is formulated as a boundary value problem (BVP) for the difference between the Earth's geopotential and the long-wavelength geopotential model as derived from satellite data commonly referred to as the disturbance potential. Since a number of geopotential coefficients of the satellite model cannot be improved by terrestrial data, we should fix them when solving the BVP. Then we are faced with a constrained (overdetermined) BVP for the Laplace equation. This has implications for the representation formula and/or the choice of the trial & test space in Galerkin boundary element methods.

We consider multipole representation, modified kernel functions, and modified trial spaces. The latter are the best choice for theoretical and numerical reasons. We propose a general method to construct a system of base functions that fix an a priori given set of geopotential coefficients. In addition, we address the problem of compression rates and stability, which implies the use of multiscale base functions. Various implementations are tested and compared for the altimetry-gravimetry II BVP.

1 Introduction

The recovery of the geopotential from terrestrial data is usually formulated as a geodetic boundary value problem (GBVP). After linearization around a suitable approximate solution, the problem is formulated in terms of the disturbing potential as a linear exterior boundary value

* Presented at the IV Hotine-Marussi Symposium, 14-17 September, Trento, Italy, 1998

problem (BVP) for the Laplace equation:

$$\Delta U(x) = 0 \qquad x \in \text{ext } \Gamma \tag{1}$$

$$(BU)(x) = g(x) \qquad x \in \Gamma \tag{2}$$

$$U(x) = O(|x|^{-1}) \qquad |x| \to \infty \tag{3}$$

U denotes the disturbing potential, B a linear differential operator of order ≤ 1, and g is a given function on the boundary Γ. See (Heck 1997) for details. The long wavelength part of the geopotential, however, cannot be precisely determined from terrestrial data mainly due to the inhomogeneous data distribution. This part of the spectrum is much better inferred from satellite data, and the result is usually expressed in terms of a series expansion in outer solid spherical harmonics of the geopotential up to a maximum degree and order. In terms of the boundary value problem formulation with terrestrial data this means that at least a subset of the expansion coefficients describing the long wavelength part of the geopotential cannot be improved, and any contribution of the solution of the BVP to these terms is likely to be purely noise or reflects discretization and approximation errors. Therefore, these coefficients should better be fixed to their values derived from satellite data.

Fixing to *zero* of certain expansion coefficients is also necessary in order to guarantee existence and uniqueness of the solution of some GBVPs. For instance, the vector Molodensky problem requires that no terms of order 1 are present in the disturbing potential, and the altimetry-gravimetry I, II BVPs require that the zero-order term is not present (Sacerdote & Sansò 1987).

For this reason, we have to impose additional contraints to the solution of the BVP, which ensures that the disturbing potential is orthogonal to a set of in total M surface spherical harmonics on a Brouillon sphere with radius R:

$$\Delta U(x) = 0 \qquad x \in \text{ext } \Gamma \tag{4}$$

$$(BU)(x) = g(x) \qquad x \in \Gamma \tag{5}$$

$$\int_{x \in \mathcal{S}_R} U(x) Y_{l_j m_j}(x/R) d\mathcal{S}_R(x) = 0 \qquad j = 1 \ldots M \tag{6}$$

$$U(x) = O(|x|^{-1}) \qquad |x| \to \infty \tag{7}$$

If the set is complete up to degree L, equations (6)-(7) simplify to the stronger decay condition $U = O(|x|^{-L-2})$.

The usual practice in geodesy is to solve the constrained BVP (4)-(7) locally by ignoring far-field data and implicitly assuming that these data do not violate any solvability condition. This may cause a bias in the solution with a long-wavelength pattern.

The aim of this contribution is to include in one way or another constraints of this type in the Boundary Element Method (BEM) approach to GBVPs. This would allow (i) to include satellite geopotential models, and (ii) to ensure well-posedness of the BVP and the numerical scheme.

2 Generalized BEM setting

In order to reformulate the BVP as integral equation over the boundary we first of all have to choose a proper representation of the solution of the BVP in terms of layer potentials with density u:

$$U(x) = \int_{y\in\Gamma} K(x,y)\, u(y)\, d\Gamma(y), \quad x \in \text{ext}\,\Gamma \tag{8}$$

After inserting the representation into the boundary condition and observing the corresponding jump relations we usually end up with a second kind integral equation for the layer density u:

$$Au := \lambda(x)\, u(x) + \int_{\Gamma} u(y)\, k(x,y)\, d\Gamma(y) = f(x), \quad x \in \Gamma \tag{9}$$

The Galerkin method is the proper discretization method to (9). It provides an approximation to the weak form of the integral equation $Au = f$: Given a dense sequence $\{V_N\}_{N=0}^{\infty}$ of finite dimensional subspaces of $L^2(\Gamma)$, we solve

$$u_N \in V_N: \qquad \langle Au_N, v\rangle = \langle f, v\rangle, \quad \forall\, v \in V_N \tag{10}$$

More details on BEM can be found in (Hackbusch 1995), special topics relevant to geodesy are treated in (Klees 1997, Lehmann 1997).

¿From the standard theory of BEM it is known that continuity, Garding inequality, and injectivity of the operator A ensure the unique solvability of this system, provided that N is sufficiently large. However, in our case the standard theory is not applicable to the weak form, since the operator assigned to the constrained BVP (4)-(7) will not be injective on $L^2(\Gamma)$. There are various possibilities how to solve this problem:

(i) We choose a representation formula that fulfils by definition the constraints (6). For instance, if the set of constraints is complete up to degree and order L, a multipole representation of order L fulfils the constraints, which in this case are equivalent to the stronger decay condition $U(x) = O(|x|^{-L-2})$ as $|x| \to \infty$:

$$U(x) = \frac{1}{4\pi} \int_{y\in\Gamma} u(y) \frac{\partial^{L+1}}{\partial n(y)^{L+1}} \left(\frac{1}{|y-x|}\right) d\Gamma(y) \tag{11}$$

(ii) We *modify* the kernel of a *classical* representation formula such that the constraints (6) are fulfilled. When applied to the Stokes' kernel this procedure is known in geodesy as the modified Stokes' kernel approach. For instance, when assuming that the set of constraints is complete up to degree L we may modify the single layer kernel by subtracting the first $L+1$ terms of a series expansion in outer solid spherical harmonics of the inverse distance.

$$U(x) = \frac{1}{4\pi} \int_{y\in\Gamma} u(y) \left(\frac{1}{|y-x|} - \sum_{l=0}^{L} \frac{|y|^l}{|x|^{l+1}} P_l\left(\frac{\langle x,y\rangle}{|x||y|}\right)\right) d\Gamma(y) \tag{12}$$

(iii) We incorporate the constraints (6) into the weak formulation by penalization with a Lagrange multiplier. This approach is discussed by (Klees et al., these proceedings) in another context and will not be considered here.

(iv) We directly impose the constraints to the trial space V_N resulting in, what we call, the *modified trial space* $\tilde{V}_N$. This approach is not trivial, since we first have to express the constraints on U into corresponding constraints on u, which, of course, results in a kind of *orthogonality condition* for u w.r.t. some globally supported functions on Γ, which span a linear space, say, $\mathcal{N}$. Then, we have to construct a basis of the modified trial space $\tilde{V}_N := V_N \cap N^{\perp}$, which, due to the dimension of V_N, is not trivial from a numerical point of view. Moreover the basis should be stable. This is the approach we want to focus on in the next section.

3 The modified trial space approach

The modified trial space approach is based on a *proper* weak formulation of the integral equation and a Galerkin discretization. In order to do that we first have to find an equivalent formulation of the constraints on U in terms of the layer density u. For the *single layer density* we can easily show that the constraints on U are equivalent to the orthogonality of the single layer density u to the restrictions to the boundary of the set of homogeneous harmonic polynomials, $\{H_{l_j m_j} : j = 1 \dots M\}$:

$$\langle U, Y_{l_j m_j} \rangle_{L^2(S_R)} = 0 \Leftrightarrow \langle u, H_{l_j m_j}\big|_{\Gamma} \rangle_{L^2(\Gamma)} = 0, \quad j = 1 \dots M \tag{13}$$

Therefore, a proper weak formulation of the integral equation $Au = f$ must be given in terms of a subspace of codimension M of $L^2(\Gamma)$, namely the space $L^2(\Gamma) \cap \mathcal{N}^{\perp}$, where $\mathcal{N}$ is the linear space spanned by the set $\{ H_{l_j m_j}\big|_{\Gamma} : j = 1 \dots M \}$. The corresponding approximate solution is

$$u_N \in V_N \cap \mathcal{N}^{\perp} \qquad \langle Au_N, v \rangle = \langle f, v \rangle \quad \forall\, v \in V_N \cap \mathcal{N}^{\perp}$$

In order to construct a basis of $\tilde{V}_N$ we remember the definition of the modified trial space: it consists of functions from V_N that are orthogonal to $\mathcal{N}$. Therefore, if the set $\{b_i : i = 1 \dots N\}$ span V_N, and the set $\{ H_{l_j m_j}\big|_{\Gamma} : j = 1 \dots M \}$ span $\mathcal{N}$, and if $\mathbf{H}$ is the matrix defined by

$$\mathbf{H} = (H_{ji}) \text{ with } H_{ji} = \langle b_i, H_{l_j m_j}\big|_{\Gamma} \rangle, \; i = 1 \dots N, \; j = 1 \dots M, \tag{14}$$

then a basis of $\tilde{V}_N$ is given by a basis of the *nullspace* of $\mathbf{H}$. That means we have to find a $N \times N - M$ matrix $\mathbf{B}$ such that $\mathbf{HB} = \mathbf{0}$. The solution $\mathbf{B}$ is by far not unique, and we may use the degrees of freedom in order to choose a $\mathbf{B}$ with some desirable properties such as easy computability, sparsity, i.e., small support of the base functions, optimal compression rates, and stability. In the following we outline the construction of various special solutions to $\mathbf{HB} = \mathbf{0}$, which have different properties:

(i) We apply a Gauss-Jacobi elimination to $\mathbf{H}$, which transforms $\mathbf{H}$ into its Hermitian normal form

$$\mathbf{H} \to \begin{pmatrix} \mathbf{I} & -\mathbf{K} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \Rightarrow \mathbf{B}_{HN} = \begin{pmatrix} \mathbf{K} \\ \mathbf{I} \end{pmatrix},$$

where $\mathbf{K}$ is a $M \times N - M$ matrix and $\mathbf{I}$ is the $N - M \times N - M$ identity matrix. Since $\mathbf{B}_{HN}$ is obviously sparse the base functions spanning $\tilde{V}_N$ have local support. However, the supporting panels do not border each other, which lowers the compression rate during the matrix assembly considerably.

(ii) By predefining the sparsity pattern of $\mathbf{B}$, we can force the new base functions to have contiguous support, which is important from a numerical point of view in order to allow compression of the system matrix by truncation. We have designed an algorithm, which provides such a basis automatically, but we want to skip the details. We call this basis a single scale basis, and denote the corresponding matrix with $\mathbf{B}_{SS}$.

(iii) Instead of a single scale basis we may use a multiscale basis, which is likely to behave more stable than a single scale basis, which would provide better condition numbers for the system matrix as the discretization becomes finer. For instance, second generation wavelets may be used as base functions (see Schneider 1995, Kleemann et al. 1996). The associated matrix is called $\mathbf{B}_{MS}$.

4 Numerical study

We tested and compared the methods described before for the altimetry-gravimetry II BVP in spherical and constant radius approximation:

$$\Delta U(x) = 0 \qquad x \in \text{ext } S_R \tag{15}$$

$$-\frac{\partial U}{\partial r}(x) = \delta g(x) \qquad x \in S_R^S \tag{16}$$

$$\left(-\frac{\partial U}{\partial r} - \frac{2}{R}U\right)(x) = \Delta g(x) \qquad x \in S_R^L \tag{17}$$

$$U(x) = O(|x|^{-2}) \quad |x| \to \infty, \tag{18}$$

where S_R is the sphere with radius R, S_R^S is the sea part and S_R^L the land part. However, we want to stress that our numerical approach does not take any advantage of this simplification. Note the stronger decay condition (18), which forces mass conseration, and, at the same time, ensures uniqueness, but in general not existence. We assume that a reference field up to degree *zero* is known, i.e., we force the disturbing potential to have no zero order term, which means physically mass conservation. We use a single layer representation formula for the disturbing potential. The sphere is triangulated into an equiangular grid with $\bar{N}$ parallels. On each grid cell we use a piecewise constant approximation to the single layer density, which implies that the (unmodified) trial space V_N consists of piecewise constant functions defined on the triangulation; its dimension is $N = 2\bar{N}^2$. The modified trial space $\tilde{V}_N$ then consists of piecewise constant functions that are orthogonal to the restriction to the sphere of the zero order homogeneous harmonic polynomial, i.e., to the surface spherical harmonic of degree zero. It has dimension $N - 1$. The weak form of the integral equation has been discretized by the Galerkin method. The following methods have been investigated:

Table 1. Condition number of the stiffness matrix

stiffness matrix	$\bar{N}=8$ $N=128$	$\bar{N}=16$ $N=512$	$\bar{N}=32$ $N=2048$
ordinary single layer kernel $\mathbf{A}:\mathbf{R}^N\to\mathbf{R}^N$	34	2016	143
modified single layer kernel $\bar{\mathbf{A}}:\mathbf{R}^N\to\mathbf{R}^N$	3.0	3.1	3.1
modified single scale trial $\mathbf{AB}_{SS}:\mathbf{R}^{N-1}\to\mathbf{R}^N$	125	481	1870
modified multiscale trial $\mathbf{AB}_{MS}:\mathbf{R}^{N-1}\to\mathbf{R}^N$	3.0	3.2	3.1
modified multiscale trial & test $\mathbf{B}_{MS}^T\mathbf{AB}_{MS}:\mathbf{R}^{N-1}\to\mathbf{R}^{N-1}$	3.2	3.4	3.3

(i) Unmodified trial space (ignoring the constraint!):

$$\mathbf{A}:\mathbf{R}^N\to\mathbf{R}^N \qquad \mathbf{Au_N}=\mathbf{f}$$

(ii) Unmodified trial space with posterior removal of the mass term:

$$\mathbf{A}:\mathbf{R}^N\to\mathbf{R}^N \qquad \mathbf{Au_N}=\mathbf{f},$$

and afterwards mass conservation is forced by replacing u_N with $\hat{u}_N = u_N - \langle u_N, H_{00}|_\Gamma\rangle$.

(iii) Modified trial space with single scale ($B = B_{SS}$) and multiscale (Haar-) base functions ($B = B_{MS}$):

$$\begin{aligned} B &: \mathbf{R}^{N-1}\to\mathbf{R}^N & AB\tilde{u}_N &= f \\ AB &: \mathbf{R}^{N-1}\to\mathbf{R}^N & u_N &= B\tilde{u}_N \end{aligned}$$

(iv) Modified trial & test space with multiscale (Haar-) base functions ($B = B_{MS}$):

$$\begin{aligned} B^TAB : \mathbf{R}^{N-1}\to\mathbf{R}^{N-1} \qquad (B^TAB)\tilde{u}_N &= B^Tf \\ u_N &= B\tilde{u}_N \end{aligned}$$

5 Results and discussions

In a first test we investigated the condition number of the system matrix expressed as the ratio of the maximum and minimum non-zero singular value (Table 1). Obviously only the modified kernel approach and the modified multiscale basis systems seem to guarantee stability of the linear system. However, for the modified kernel this is only because the boundary surface is spherical. For non-spherical surfaces we expect a bad conditioning due to an almost zero singular value. The modified single scale basis seems not to be stable since the condition number

Table 2. Sparsity of the truncated stiffness matrix in terms of percentage of zero elements (5 valid decimal digits of the solution are guaranteed after truncation

stiffness matrix	$\bar{N} = 8$ $N = 128$	$\bar{N} = 16$ $N = 512$	$\bar{N} = 32$ $N = 2048$
ordinary single layer kernel $\mathbf{A} : \mathbf{R}^N \to \mathbf{R}^N$	0%	0%	0%
modified single layer kernel $\bar{\mathbf{A}} : \mathbf{R}^N \to \mathbf{R}^N$	0%	0%	0%
modified single scale trial $\mathbf{AB}_{SS} : \mathbf{R}^{N-1} \to \mathbf{R}^N$	0%	0%	1%
modified multiscale trial $\mathbf{AB}_{MS} : \mathbf{R}^{N-1} \to \mathbf{R}^N$	1%	11%	36%
modified multiscale trial & test $\mathbf{B}_{MS}^T\mathbf{AB}_{MS} : \mathbf{R}^{N-1} \to \mathbf{R}^{N-1}$	5%	21%	56%

becomes worse with finer discretization. This is what has to be expected for a single scale basis spanning the modified trial space. The relatively large condition number for $\bar{N} = 16$ in case of the *unmodified* single layer approach may reflect the non-uniqueness of the altimetry-gravimetry II BVP if the zero-order term is still present.

Next we investigated the sparsity of the system matrix after truncation, i.e., after replacing small entries by zeros such that about five valid decimal digits for the solution is guaranteed (Table 2). The sparsity is expressed in the percentage of zero elements of the truncated stiffness matrix. Obviously only the modified multiscale trial & test space (Haar wavelets) yields a significant sparsity of about 56% for $\bar{N} = 32$.

We also computed solutions of the altimetry-gravimetry II BVP and compared them with true values. In order to do so we did a simple numerical simulation: the sample potential is generated by two point masses inside the sphere; the data have been computed without any noise. The linear system of equations have been calculated and solved for all methods. Thereafter, potential values at the centers of the grid cells have been computed from the approximated single layer density by integration over the boundary surface, and these potential values have been compared with their true values.

Table 3 shows the maximum absolute, the mean absolute and the rms error for all tested methods. They are evaluated in terms of relative potential differences (true-computed) at the centers of the grid cells. By far the best results are obtained by the modified trial space approach. They are about one order of magnitude better than for the unmodified single layer approach. Surprisingly is that a simple posterior removal of the zero-order term improves the result by only a factor of two. Single scale and multiscale trial and trial & test space yield comparable results. This is due to the rather coarse resolution ($\bar{N} \leq 32$), which implies that the instability of the single scale basis does not show up yet in the results. For finer discretizations we expect a significant loss of decimal digits for the single scale basis. Concerning the numerical test we want to emphasize that there are no short wavelength features in the boundary data

Table 3. Error statistics: relative potential residuals on the boundary surface ($\bar{N} = 32, N = 2048$)

linear system of equations	max$\lvert\cdot\rvert$	mean$\lvert\cdot\rvert$	rms
ordinary single layer kernel			
$\mathbf{A}\mathbf{u}_N = \mathbf{f}$	0.0468	0.0186	0.0211
posterior removal			
$\mathbf{A}\mathbf{u}_N = \mathbf{f}, \hat{u}_N = u_N - \langle u_N, H_{00}\vert_\Gamma \rangle$	0.0289	0.0081	0.0101
modified single layer kernel			
$\bar{\mathbf{A}}\bar{\mathbf{u}}_N = \mathbf{f}$	0.0184	0.0023	0.0041
modified single scale trial			
$(\mathbf{A}\mathbf{B}_{SS})\tilde{\mathbf{u}}_N = \mathbf{f}, \mathbf{u}_N = \mathbf{B}_{SS}\tilde{\mathbf{u}}_N$	0.0177	0.0008	0.0017
modified multiscale trial			
$(\mathbf{A}\mathbf{B}_{MS})\tilde{\mathbf{u}}_N = \mathbf{f}, \mathbf{u}_N = \mathbf{B}_{MS}\tilde{\mathbf{u}}_N$	0.0177	0.0008	0.0017
modified multiscale trial & test			
$(\mathbf{B}_{MS}^T\mathbf{A}\mathbf{B}_{MS})\tilde{\mathbf{u}}_N = \mathbf{B}_{MS}^T\mathbf{f}, \mathbf{u}_N = \mathbf{B}_{MS}\tilde{\mathbf{u}}_N$	0.0179	0.0010	0.0018

since the gravitational field is generated by two mass points in the deep interior of the sphere. Unmodelled short wavelength features may propagate differently into the solution for the various methods we investigated. This can be one reason why the results for the various methods are that pronounced.

6 Conclusions

The most important conclusion is that the benefit of using a global reference field in terms of accuracy is significant, even when a very low degree field is used. Not only discretization errors and far-field effects are suppressed, but also the well-posedness of the BVP can be guaranteed in some cases. Therefore, the method developed so far, e.g., the Haar multiscale basis, although by no means perfect, is already sufficient to guarantee well-posedness of some real GBVPs.

As the various methods are concerned, we conclude that modified multiscale basis systems are superior to all other methods in terms of stability and accuracy. However, the construction of suitable multiscale base functions spanning the modified trial space is far from being trivial. Currently, we have to limit to mass conservation for conceptual reasons, which is the most trivial constraint. The use of higher degree reference geopotential models requires the construction of multiscale bases that are orthogonal to the restriction to the boundary of a set of homogeneous harmonic polynomials. This has still to be done.

References

Hackbusch, W. (1995): *Integral equations: theory and numerical treatment.* Birkhäuser Verlag Basel Boston Berlin.

Heck, B. (1997): Formulation and linearization of boundary value problems: From observables to a mathematical model. In: F. Sansò, R. Rummel (Eds.): *Geodetic Boundary Value Problems in View of the One Centimeter Geoid.* Lecture Notes in Earth Sciences, **65**. Springer Verlag Berlin Heidelberg.

Kleemann, B.H., A. Rathsfeld, R. Schneider (1996): Multiscale methods for boundary integral equations and their application to boundary value problems in scattering theory and geodesy. In: W. Hackbusch, G. Wittum (eds.): *Boundary Elements: Implementation and Analysis of Advanced Algorithms*, Notes on Numerical Fluid Mechanics, **54**, Vieweg Verlag Braunschweig.

Klees, R. (1997): Topics on boundary element methods. In: F. Sansò, R. Rummel (Eds.): *Geodetic Boundary Value Problems in View of the One Centimeter Geoid.* Lecture Notes in Earth Sciences, **65**. Springer Verlag Berlin Heidelberg.

Klees, R., C. Lage, C. Schwab: Fast numerical solution of the vector Molodensky problem. Paper presented at the IV Hotine-Marussi Symposium, 14-17 September 1998, Trento, Italy.

Lehmann, R. (1997): Studies on the Use of Boundary Element Methods in Physical Geodesy. Publ. German Geodetic Commission, Series A, **113**. Munich.

Sacerdote, F., F. Sansò, (1987): Further remarks on the altimetry gravimetry problems. *Bull. Geod.*, **61**, 183–201

Schneider, R. (1995): Multiskalen- und Wavelet-Matrixkompression: Analysisbasierte Methoden zur effizienten Lösung großer vollbesetzter Gleichungssysteme. Habilitation thesis. Technical University Darmstadt.

DEOS Progress Letter 98.1: 31–42

Fast numerical solution of the vector Molodensky problem*

Roland Klees, Christian Lage[1] and Christoph Schwab[1]

[1] Seminar for Applied Mathematics, ETH Zürich, Switzerland

Abstract

When *standard* boundary element methods (BEM) are used in order to solve the linearized vector Molodensky problem we are confronted with two problems: (i) the absence of $O(|x|^{-2})$ terms in the decay condition is not taken into account, since the single layer ansatz, which is commonly used as representation of the disturbing potential, is of the order $O(|x|^{-1})$ as $x \to \infty$. This implies that the standard theory of Galerkin BEM is not applicable since the injectivity of the integral operator fails; (ii) the $N \times N$ stiffness matrix is dense, with N typically of the order 10^5. Without fast algorithms, which provide suitable approximations to the stiffness matrix by a sparse one with $O(N \cdot \log^s N)$, $s \geq 0$, non-zero elements, high-resolution global gravity field recovery is not feasible.

We propose solutions to both problems. (i) A proper variational formulation taking the decay condition into account is based on some closed subspace of co-dimension 3 of $L^2(\Gamma)$. Instead of imposing the constraints directly on the boundary element trial space, we incorporate them into a variational formulation by penalization with a Lagrange multiplier. The conforming discretization yields an augmented linear system of equations of dimension $N + 3 \times N + 3$. The penalty term guarantees the well-posedness of the problem, and gives precise information about the incompatibility of the data. (ii) Since the upper left submatrix of dimension $N \times N$ of the augmented system is the stiffness matrix of the standard BEM, the approach allows to use all techniques to generate sparse approximations to the stiffness matrix such as wavelets, fast multipole methods, panel clustering etc. without any modification. We use a combination of panel clustering and fast multipole method in order to solve the augmented linear system of equations in $O(N)$ operations. The method is based on an approximation of the kernel function of the integral operator by a degenerate kernel in the far field, which is provided by a multipole expansion of the kernel function.

We demonstrate the potential of the method by solving a Robin problem on the sphere with a nullspace of dimension 3. For $N = 65538$ unknowns the matrix assembly takes about 600 s and the solution of the sparse linear system using GMRES without any preconditioning takes about 8 s. 30 iterations are sufficient to make the error smaller than the discretization error.

* Presented at the IV Hotine-Marussi Symposium, 14-17 September, Trento, Italy, 1998

1 Introduction

The determination of the exterior gravity field of the Earth from terrestrial observations is usually formulated in terms of a boundary value problem (BVP) for the Laplace-Poisson equation. Depending on the type of observations several boundary value problems can be defined. However, after linearization around a suitable approximate solution all problems are more or less special cases of the exterior oblique derivative BVP for the Laplace operator; the boundary surface is either the Earth's surface, a suitable approximation to it like a telluroid or an ellipsoid of revolution. Numerical solutions of the linearized BVP are usually based on various additional approximation steps like, e.g., spherical approximation and constant radius approximation.

Here we consider Galerkin methods for integral equation formulations of the linearized BVP which avoid any of the aforementioned approximations. The price to pay for this is that the kernel functions are non-isotropic and the boundary surface is non-spherical. Therefore, the assembly of the linear system of equations becomes more elaborate; moreover, since the system matrix is dense, sparse solvers cannot be used any more to solve for the huge number of unknowns.

There is another aspect which has to be taken into account in the formulation of geodetic BVPs. Usually, the low frequency components of the geopotential are accurately obtained by satellite measurements. That means that a number of coefficients in the spherical harmonics series expansion of the geopotential is determined with a precision that cannot be improved by terrestrial data. This is accounted for in the formulation of the geodetic BVP in the form of additional constraints to the perturbation problem. The same holds if the geodetic BVP lacks well-posedness. For instance, the vector Molodensky BVP requires the first order terms in the expansion of the geopotential in spherical harmonics to vanish in order to ensure uniqueness of the solution; for the same reason the Altimetry-Gravimetry I & II BVPs require that no zero order term is present. Finally, if the measured data is not in the range of the operator the problem may even not have any solution at all.

Therefore, a numerical approach has to be designed that can handle these peculiarities of geodetic BVPs. As far as Galerkin methods to integral equations are concerned this implies the following questions: (i) how to properly handle the conditions that ensure well-posedness of the problem, (ii) how to properly include satellite-derived geopotential models, and (iii) how to design a fast algorithm which is suitable for high resolution global geopotential recovery with a performance that is almost independent of (i) and (ii)?

Our solution to (i) and (ii) is based on a new saddle point formulation which avoids to modify the trial and test spaces. The solution to (iii) is a fast algorithm that combines ideas of panel clustering and fast multipole methods, and which is easy to combine with the saddle point formulation.

The outline of the paper is the following: We start with the formulation of our model problem, which in terms of the problems (i)-(iii) is closely related to the geodetic situation. Then, we will briefly discuss its integral equation formulation and the proper weak formulation and conforming approximation in a modified trial space, see also Klees & Lehmann (cf. these proceedings). Finally, we discuss the fast algorithm and demonstrate its performance based on a simple numerical study.

2 The mathematical model

Our model problem reads as follows: Given a function f on the surface of the unit sphere $\Gamma \subset \mathbb{R}^3$; let n denote the unit normal vector field on Γ pointing into the exterior to Γ. We wish to solve the boundary value problem

$$\begin{aligned} \triangle U(x) &= 0 && x \in \text{ext } \Gamma \\ U(x) + \langle \nabla U(x), n(x) \rangle &= f(x) && x \in \Gamma \\ U(x) &= \frac{c}{|x|} + O(|x|^{-3}), \quad |x| \to \infty, && c \in \mathbb{R}\backslash\{0\} \end{aligned} \tag{1}$$

The homogeneous problem with $f = 0$ admits 3 eigensolutions which span the nullspace $\mathcal{N}$. Since (1) is a regular elliptic boundary value problem, Fredholm's alternative holds. Thus, uniqueness implies existence, and the former requires that the data f satisfies 3 compatibility conditions, i.e., the data f must be orthogonal to the nullspace of the homogeneous adjoint BVP which, due to Fredholm's alternative, has dimension 3 as well. Moreover, the problem has a unique solution $U \perp \mathcal{N}$ if f satisfies this compatibility condition.

The main difference between the model (1) and the linearized vector Molodensky problem is the spherical geometry and the boundary operator which involves the normal derivative instead of the oblique derivative. However, our approach does rely neither on the normal derivative nor on the spherical geometry of the boundary surface. In fact, the saddle point formulation and the fast algorithm are applicable without any modification for oblique derivative problems and non-spherical geometries, as well. The decision to use the model (1) has been done for simplicity reasons.

In order to reformulate the BVP (1) as an integral equation, we choose the single layer ansatz with kernel $k(z) = (4\pi|z|)^{-1}$:

$$U(x) = \int_{y \in \Gamma} k(x-y)\, u(y)\, d\Gamma(y), \quad x \in \text{ext } \Gamma \tag{2}$$

where u is the unknown density function. Inserting (2) into the boundary condition (1) yields a weakly-singular boundary integral equation for the unknown density u:

$$Au := \frac{1}{2}u(x) + \int_\Gamma \frac{\partial k(x-y)}{\partial n(x)} u(y) d\Gamma(y) + \int_\Gamma k(x-y) u(y) d\Gamma(y) = f(x), \quad x \in \Gamma \tag{3}$$

The principal symbol of the integral operator A is positive definite, which implies that A is strongly elliptic. Moreover, it can be shown (Mikhlin and Prößdorf, 1986) that A is bijective from $L^2(\Gamma) \to L^2(\Gamma)$. Notice, however, that the absence of the $O(|x|^{-2})$-terms in the decay condition is not taken into account by (2) since the single layer potential is of order $O(|x|^{-1})$ as $|x| \to \infty$.

3 Weak formulation and approximation

We use the Galerkin method in order to discretize the boundary integral equation (3). Note that we could use collocation as well, but this would not be the proper discretization method

for the linearized geodetic BVPs, where we usually have to deal with Cauchy-singular and hypersingular operators A. We consider the weak form of the integral equation (3):

$$u \in L^2(\Gamma): \qquad \langle Au, v\rangle = \langle f, v\rangle \quad \forall v \in L^2(\Gamma), \tag{4}$$

where $\langle\cdot,\cdot\rangle$ denotes the $L^2(\Gamma)$-inner product. The Galerkin method in abstract form reads: Given a dense sequence $\{V_N\}_{N=0}^{\infty}$ of finite dimensional subspaces of $L^2(\Gamma)$, find

$$u_N \in V_N: \qquad \langle Au_N, v\rangle = \langle f, v\rangle \quad \forall v \in V_N. \tag{5}$$

Hence, for a given basis $\{b_1, \ldots, b_N\}$ of V_N, we have to solve the linear system of equations $\mathbf{Au} = \mathbf{f}$ where the stiffness matrix $\mathbf{A}$ and the right-hand side $\mathbf{f}$ are defined by

$$(\mathbf{A})_{ij} := \langle b_i, Ab_j\rangle, \text{ and } (\mathbf{f})_i := \langle b_i, f\rangle, \quad i,j = 1\ldots N. \tag{6}$$

It is known that continuity, Garding inequality, and injectivity of the operator A ensure the unique solvability of this system, provided that N is sufficiently large (Hildebrandt and Wienholtz, 1964). However, in our case the standard theory is not applicable to the weak form, since the latter does not take into account the constraint $U \perp \mathcal{N}$ which means that the injectivity fails. Therefore, in order to make the standard theory applicable, the *proper* weak formulation of $Au = f$ must not be based on $L^2(\Gamma)$ but on some closed subspace of co-dimension 3 of $L^2(\Gamma)$:

$$u \in L^2(\Gamma) \cap \mathcal{N}^{\perp}: \qquad \langle Au, v\rangle = \langle f, v\rangle \quad \forall v \in L^2(\Gamma) \cap \mathcal{N}^{\perp} \tag{7}$$

The corresponding conforming approximate solution is

$$u_N \in V_N \cap \mathcal{N}^{\perp}: \qquad \langle Au_N, v\rangle = \langle f, v\rangle \quad \forall v \in V_N \cap \mathcal{N}^{\perp} \tag{8}$$

Therefore, we need the subspace $\mathcal{N}$. In our case it is easy to show that the condition of vanishing $O(|x|^{-2})$-terms in the expansion of U is equivalent to the orthogonality of the density u to the restriction to the boundary Γ of the homogeneous harmonic polynomials of degree 1. This implies that $\mathcal{N}$ is the linear space spanned by the restriction to the boundary of the 3 homogeneous harmonic polynomials of degree 1:

$$\mathcal{N} = \text{span}\{H_{1,m}|_{\Gamma} : m = -1, 0, 1\} \tag{9}$$

4 The saddle point formulation

The conforming Galerkin discretization (8) is difficult to realize in practice. The reason is that the homogeneous harmonic polynomials of degree 1 which span $\mathcal{N}$ are globally supported, and for the computations a basis of $V_N \cap \mathcal{N}^{\perp}$ must be generated. Since the dimension of V_N is typically very large (in the experiments below about 10^5 gravity field parameters have to be solved for), it is a non-trivial matter how to do that stably and efficiently. Moreover, the support of the base functions spanning $V_N \cap \mathcal{N}^{\perp}$ will be larger than the support of the base functions spanning V_N which increases the computational effort. (Klees & Lehmann 1998)

have discussed this problem in another context, and have proposed the method of modified multiscale trial & test spaces. However, this solution strategy is currently limited to constraints involving homogeneous harmonic polynomials of degree 0.

Here, we propose a different approach: We reformulate (7) as a *saddle point problem* analogous to what is done in incompressible fluid flow. The constraint $u \perp \mathcal{N}$ will not be imposed directly on the boundary element space V_N, but will rather be incorporated into the variational formulation by penalization with a Lagrange multiplier p. This leads to an augmented system which reads:

$$(u,p) \in L^2(\Gamma) \times \mathcal{N}: \qquad \begin{array}{lcll} \langle Au, v\rangle + \langle Ap, v\rangle & = & \langle f, v\rangle & \forall v \in L^2(\Gamma) \\ \langle u, q\rangle & = & 0 & \forall q \in \mathcal{N} \end{array} \tag{10}$$

and the conforming Galerkin approximation to (10) is:

$$(u_N,p_N) \in V_N \times \mathcal{N}: \qquad \begin{array}{lcll} \langle Au_N, v\rangle + \langle Ap_N, v\rangle & = & \langle f, v\rangle & \forall v \in V_N \\ \langle u_N, q\rangle & = & 0 & \forall q \in \mathcal{N} \end{array} \tag{11}$$

(u,p) is called the saddle point of the variational system. The conforming approximation defines a linear system of equations of dimension $N+3$. The upper left matrix is the usual $N \times N$ stiffness matrix of the unconstrained problem, the upper right and the transposed of the lower left matrix have dimension $N \times 3$; their elements are inner products of the bases of $A\mathcal{N}$ and of $\mathcal{N}$, respectively, with the basis of V_N.

A major advantage of the saddle point formulation is that all techniques to generate sparse approximations to the matrix $\langle Au_N, v\rangle$ such as wavelets, fast multipole methods, panel clustering etc. can be used here without any modification. Moreover, if the data happen to be in $A\mathcal{N}^\perp$, then, of course, $p = 0$. In practice, however, f is not exactly in $A\mathcal{N}^\perp$ due to various data and approximation errors. Then, the saddle point formulation (10) is still well-posed and the size of p gives precise information about the degree of incompatibility of the data f. Note that the proper weak formulation (7) would not have a solution if $f \notin A\mathcal{N}^\perp$. Finally, the assembly of the matrices $\langle Ap_N, v\rangle$ and $\langle u_N, q\rangle$ is of order $O(N)$, and therefore, does not make the numerics much more elaborate.

5 The fast algorithm

In BEM the stiffness matrix is a dense $N \times N$-matrix, since the kernel function $k(x-y)$ links every point $x \in \Gamma$ to every point $y \in \Gamma$. Hence, storage and time consumptions of the method are of order $O(N^2)$ provided that iterative solvers could be applied efficiently which limits the application of BEM in practice. In the eighties Hackbusch and Nowak (Hackbusch and Nowak, 1989) developed the panel clustering method in order to overcome this grave drawback. Independently, Rokhlin proposed the fast multipole method (Rokhlin, 1985). Both methods are based on an approximation of the kernel factorizing the x, y-dependency. By this, the x-integration is separated from the y-integration reducing the amount of work substantially.

In our approach, we use a blend of panel clustering and fast multipole method. Suppose that

the kernel k may be replaced by a degenerate kernel k_m

$$k(x,y) \approx k_m(x,y;x_0,y_0) = \sum_{(\mu,\nu)\in\mathcal{I}_m} \kappa_{\mu\nu}(x_0,y_0) X_\mu(x;x_0) Y_\nu(y;y_0) \tag{12}$$

with parameters $m \in \mathbb{N}$, $x_0, y_0 \in \mathbb{R}^3$ such that the error bound

$$|k(x,y) - k_m(x,y;x_0,y_0)| \leq C_\eta \, \eta^m \, |k(x,y)| \tag{13}$$

is valid for $0 < \eta < 1$ and all $x, y \in \mathbb{R}^3$ satisfying

$$|y - y_0| + |x - x_0| \leq \eta \, |y_0 - x_0|. \tag{14}$$

Here, $\mathcal{I}_m$ denotes a finite index set.

There are several possibilities to choose an approximation by degenerate kernels (Lage, 1998). In our experiments described in Section 6 approximation (12) was obtained by applying a truncated multipole expansion, i.e.,

$$\mathcal{J}_m := \{\mu \in \mathbb{N}_0 \times \mathbb{Z} : |\mu_2| \leq \mu_1, \, \mu_1 < m\}, \quad \mathcal{I}_m := \{(\mu,\nu) \in (\mathcal{J}_m)^2 : \mu_1 + \nu_1 < m\} \tag{15}$$

$$\kappa_{\mu\nu}(x_0,y_0) := \kappa_{\mu+\nu}(x_0,y_0) := \frac{1}{4\pi \, C_{\mu_1+\nu_1}^{\mu_2+\nu_2} |y_0 - x_0|^{\mu_1+\nu_1+1}} Y_{\mu_1+\nu_1}^{\mu_2+\nu_2}\left(\frac{y_0 - x_0}{|y_0 - x_0|}\right) \tag{16}$$

$$X_\mu(x;x_0) := C_{\mu_1}^{\mu_2} |x - x_0|^{\mu_1} Y_{\mu_1}^{-\mu_2}\left(\frac{x - x_0}{|x - x_0|}\right), \quad Y_\nu(y;y_0) := X_\nu(-y;-y_0) \tag{17}$$

with

$$C_l^p := \frac{i^{|p|}}{\sqrt{(l-p)!(l+p)!}}, \quad Y_l^p(x) := P_l^{|p|}(\cos\theta) \, \mathrm{e}^{ip\phi} \tag{18}$$

for $x = (\cos\phi \sin\theta, \sin\phi \sin\theta, \cos\theta)^T \in \mathbb{S}_2$. The functions X_μ and Y_ν are solid spherical harmonics of positive degree whereas the expansion coefficients $\kappa_{\mu\nu}$ are homogeneous harmonic polynomials of negative degree. Note that the multipole expansion is nothing else but an efficient representation of the Taylor expansion of $|y - x|^{-1}$. While for arbitrary kernel functions k, the index set $\mathcal{J}_m$ of a truncated Taylor expansion contains $O(m^3)$ indices, only $O(m^2)$ coefficients must be stored to evaluate the Taylor expansion of $|y - x|^{-1}$ using the multipole ansatz according to (15)-(17). The expansion for the adjoint kernel of the double layer potential is obtained from (15)-(17) by applying the $\frac{\partial}{\partial n}$-Operator to $X_\mu(\cdot, x_0)$.

In order to derive an efficient approximation of the stiffness matrix $\mathbf{A}$ from the approximation of the kernel, we have to define appropriate regions on the boundary surface Γ, such that the approximation error could be controlled by (13),(14). Let $\mathcal{P}(\Gamma)$ denote the set of all subsets of Γ and $\mathcal{C} \subset \mathcal{P}(\Gamma) \times \mathcal{P}(\Gamma)$ a finite set defining a partition of $\Gamma \times \Gamma$. The elements of the first and second component of $\mathcal{C}$, i.e.,

$$\mathcal{X} := \mathcal{X}_\mathcal{C} := \{\sigma \subset \Gamma : \exists \tau \subset \Gamma, (\sigma,\tau) \in \mathcal{C}\} \tag{19}$$

$$\mathcal{Y} := \mathcal{Y}_\mathcal{C} := \{\tau \subset \Gamma : \exists \sigma \subset \Gamma, (\sigma,\tau) \in \mathcal{C}\}, \tag{20}$$

are called clusters. A pair of clusters $(\sigma, \tau) \in \mathcal{C}$ is η-admissible, iff

$$\check{r}_\sigma + \check{r}_\tau \leq \eta |\check{c}_\sigma - \check{c}_\tau|, \tag{21}$$

where $\check{r}_M$ and $\check{c}_M$ denote for $M \subset \mathbb{R}^3$ the Čebyšev radius and center, respectively. Using this property we split the partition $\mathcal{C}$ into a *far field*

$$\mathcal{F} := \mathcal{F}_\mathcal{C}(\eta) := \{(\sigma, \tau) \in \mathcal{C} : (\sigma, \tau) \text{ is } \eta\text{-admissible}\} \tag{22}$$

and a *near field*

$$\mathcal{N} := \mathcal{N}_\mathcal{C}(\eta) := \mathcal{C} \backslash \mathcal{F}_\mathcal{C}(\eta) \tag{23}$$

which implies a corresponding splitting of the stiffness matrix $\mathbf{A}$ into a near field contribution $\mathbf{N}$ and a far field contribution $\mathbf{F}$:

$$(\mathbf{N})_{i,j} := \sum_{(\sigma,\tau)\in\mathcal{N}} \int_\sigma b_i(x) \int_\tau k(x,y)\, b_j(y)\, dydx \tag{24}$$

$$(\mathbf{F})_{i,j} := \sum_{(\sigma,\tau)\in\mathcal{F}} \int_\sigma b_i(x) \int_\tau k(x,y)\, b_j(y)\, dydx \tag{25}$$

Since the domains of integration of the far field part are well-separated, i.e., satisfy (14) with $x_0 := \check{c}_\sigma$ and $y_0 := \check{c}_\tau$, the kernel k can be replaced by its approximation k_m which in turn yields an approximation of $\mathbf{F}$:

$$\mathbf{F} \approx \sum_{(\sigma,\tau)\in\mathcal{F}} \mathbf{X}_\sigma \mathbf{F}_{\sigma\tau} \mathbf{Y}_\tau, \tag{26}$$

where the matrices $\mathbf{X}_\sigma$, $\mathbf{Y}_\tau$, and $\mathbf{F}_{\sigma\tau}$ are defined by

$$(\mathbf{X}_\sigma)_{i,\mu} := \int_\sigma b_i(x) X_\mu(x; c_\sigma) dx, \quad (\mathbf{Y}_\tau)_{\nu,j} := \int_\tau b_j(y) Y_\nu(y; c_\tau) dy \tag{27}$$

$$(\mathbf{F}_{\sigma\tau})_{\mu,\nu} := \begin{cases} \kappa_{\mu\nu} & \text{if } (\mu,\nu) \in \mathcal{I}_m \\ 0 & \text{else} \end{cases} \tag{28}$$

In other words, the stiffness matrix is approximated by a near field matrix $\mathbf{N}$ and a finite sum of rank-$|\mathcal{J}_m|$ modifications corresponding to the approximation of the kernel by degenerate kernels. The matrices $\mathbf{X}_\sigma$ only depend on x, the matrices $\mathbf{Y}_\tau$ only on y, and the matrices $\mathbf{F}_{\sigma\tau}$ contain the expansion coefficients $\kappa_{\mu\nu}$.

Essential for the efficiency of the algorithm is (i) the construction of a partition $\mathcal{C}$ such that the near field matrix $\mathbf{N}$ is a sparse matrix, i.e., contains only $O(N)$ entries, and (ii) the fast evaluation of the approximate far field contribution (26), in particular the fast evaluation of the matrix vector product

$$\mathbf{v} = \sum_{(\sigma,\tau)\in\mathcal{F}} \mathbf{X}_\sigma\, \mathbf{F}_{\sigma\tau} \mathbf{Y}_\tau\, \mathbf{u}. \tag{29}$$

The key is a hierarchical organization of clusters. Let $\mathcal{P}$ denote the given panelization of Γ. We subdivide $\mathcal{P}$ into two about equally large sets recursively until the subsets contain $O(1)$ panels. This defines a binary tree with root $\mathcal{P}$. Each node of the tree represents a subset of $\mathcal{P}$ which in turn implies a subset of Γ, i.e. the binary tree defines a hierarchical decomposition of Γ into clusters.

By traversing the tree a suitable partition $\mathcal{C} = \mathcal{F} \cup \mathcal{N}$ is constructed:

```
partition(σ, τ, F, N){
  if (σ is a leaf) or (τ is a leaf) then
    N ← {(σ, τ)} ∪ N
  else if ((σ, τ) η-admissible) then
    F ← {(σ, τ)} ∪ F
  else if (r̆_σ < r̆_τ) then
    for all children τ' of τ   partition(σ, τ', F, N)
  else
    for all children σ' of σ   partition(σ', τ, F, N)
}
```

The matrix vector product (29) is evaluated in three steps:

1. evaluate $\mathbf{u}_\tau := \mathbf{Y}_\tau \mathbf{u}$ for all $\tau \in \mathcal{Y}$,
2. evaluate $\mathbf{v}_\sigma := \begin{cases} \mathbf{F}_{\sigma\tau} u_\tau & \text{for } (\sigma,\tau) \in \mathcal{F}, \\ 0 & \text{otherwise} \end{cases}$ for all $\sigma \in \mathcal{X}$,
3. evaluate $\mathbf{v} = \sum_\sigma \mathbf{X}_\sigma \mathbf{v}_\sigma$.

The first and the last step could be accelerated by using so-called shift operations. We find

$$\mathbf{Y}_\tau = \sum_{\tau' \text{ child of } \tau} \mathbf{D}_{\tau\tau'} \mathbf{Y}_{\tau'}, \tag{30}$$

with matrices $\mathbf{D}_{\tau\tau'}$, i.e.,

$$\mathbf{u}_\tau = \begin{cases} \mathbf{Y}_\tau \mathbf{u} & \text{for } \tau \text{ a leaf}, \\ \sum_{\tau' \text{ child of } \tau} \mathbf{D}_{\tau\tau'} \mathbf{u}_{\tau'} & \text{otherwise}. \end{cases} \tag{31}$$

Hence, to evaluate $\mathbf{u}_\tau$ for all $\tau \in \mathcal{Y}$ we only have to assemble matrices $\mathbf{Y}_\tau$ if τ is a leaf. These matrices are sparse with $O(|\mathcal{J}_m|) = O(m^2)$ entries. The products $\mathbf{D}_{\tau\tau'}\mathbf{u}_{\tau'}$ are handled by efficient algorithms without assembling $\mathbf{D}_{\tau\tau'}$ explicitly (Greengard and Rokhlin, 1997). The same holds for step 3. With matrices $C_{\sigma\sigma^*}$ defined by

$$\mathbf{X}_{\sigma^*} = \sum_{\sigma \text{ child of } \sigma^*} \mathbf{X}_\sigma \mathbf{C}_{\sigma\sigma^*}, \tag{32}$$

and vectors $\bar{\mathbf{v}}_\sigma := \mathbf{v}_\sigma + \mathbf{C}_{\sigma\sigma^*} \bar{\mathbf{v}}_{\sigma^*}$, σ child of σ^*, it follows that

$$\mathbf{v} = \sum_\sigma \mathbf{X}_\sigma \mathbf{v}_\sigma = \sum_{\sigma \text{ a leaf}} \mathbf{X}_\sigma \bar{\mathbf{v}}_\sigma. \tag{33}$$

Again, only matrices $\mathbf{X}_\sigma$ for leaves $\sigma \in \mathcal{X}$ must be assembled.

An analysis of the complexity (cf. (Hackbusch and Nowak, 1989), (Rokhlin, 1985)) shows that the number of operations necessary to perform the matrix vector product (29) is of order $O(m^4 N)$, with N the number of unknowns.[1] The storage consumptions are of order $O(m^2 N)$. To ensure that the error of the far field approximation is asymptotically equal to the order of the discretization error, we have to choose $m = O(\log N)$.

6 Numerical experiment

We did some numerical test computations in order to demonstrate the performance of the method. The "true" potential is given by

$$U(x) = |x|^{-1} + x_1 x_2 \, |x|^{-5} \tag{34}$$

We approximated the unit sphere by planar triangles. Piecewise linear polynomials have been used as trial and test functions. The linear system of equations (LSE) was solved using a GMRES solver without any preconditioning. About 30 iterations were necessary to keep the error lower than the discretization error, independent of the number of unknowns. For our cluster algorithm the matrix-vector operations for the calculation of the far field contribution have been done in every iteration step. The necessary information about the $\mathbf{X}_\sigma$, $\mathbf{Y}_\tau$ and $\mathbf{F}_{\sigma\tau}$ matrices have been stored in core on the workstation. The quality of the solution has been checked at a grid of points with distance 0.5 to the surface of the unit sphere.

The results were obtained on a *SUN* Ultra-Enterprise 4000/5000 on a single processor (*UltraSPARC*, 248MHz), 2 GB RAM using the *SUN* C++ 4.2 Compiler and the class library *Concepts-1.3* for boundary elements.

Figure 1 shows the CPU-time for the matrix assembly for the standard BEM (dashed line) and our fast algorithm (solid lines). The latter depends on the order m of the multipole expansion. The computations have been done for $m = 3 \ldots 7$. The results are shown as function of the number of unknowns, i.e., of the resolution. The finest resolution (65538 unknowns, 131072 panels) is equivalent to 0.5 degrees. The dependency on m is minor, because N dominates. Compared with the standard method a speed-up of up to 3 orders of magnitude can be expected for the finest resolution.

Figure 2 shows the relative mean absolute error in the potential in exterior points located at a distance of 0.5 from the surface of the unit sphere. The solid lines represent the cluster-BEM solution for $m = 3 \ldots 7$, the dashed line represents the standard-BEM solution. Only for $m = 6, 7$ we observe an almost monotone decreasing error with increasing number of unknowns. This indicates that small values of m corresponding to low expansion orders produce approximation errors that dominate the total error budget if the discretization becomes finer. At a certain discretization level $m = 5$ gives a better accuracy than $m = 7$. This can be explained by the influence of the discretization error which dominates at this discretization level the total error budget. Therefore, variations in order of the discretization error can be expected.

[1] With a new approach to evaluate the products $\mathbf{F}_{\sigma\tau}\mathbf{u}_\tau$ using exponential expansions this could be reduced to $O(m^3 N)$ (Greengard and Rokhlin, 1997).

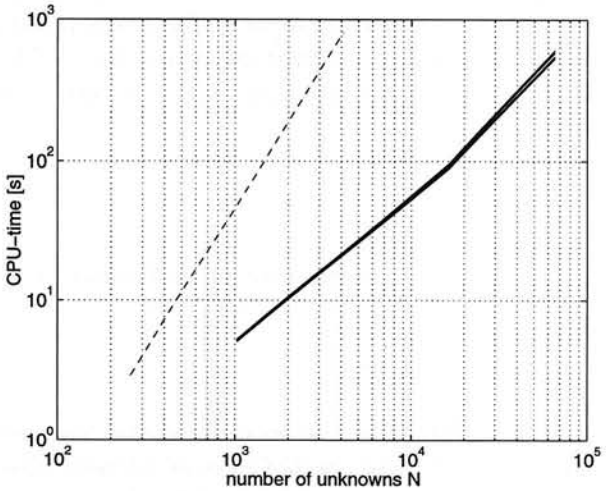

Fig. 1. CPU-time for matrix assembly ($m = 3, 4, 5, 6, 7$): standard BEM (dashed line) versus fast algorithm (solid lines).

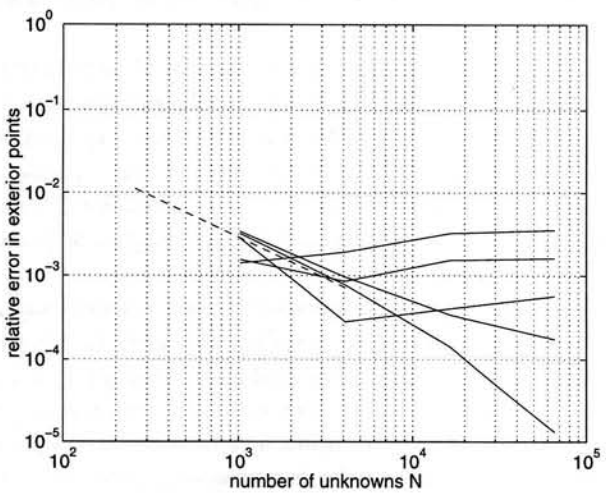

Fig. 2. Relative mean absolute error in a set of points with distance 0.5 from the surface of the unit sphere: standard BEM (dashed line) versus fast algorithm for $m = 3, 4, 5, 6, 7$ (solid lines)

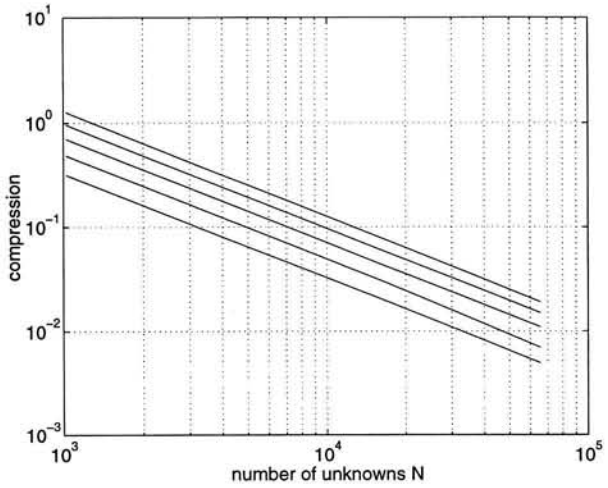

Fig. 3. Compression of the stiffness matrix for $m = 3, 4, 5, 6, 7$.

Figure 3 shows the compression rate as a function of the number of unknowns. A compression factor of 0.01 means that the total of entries to store the necessary information of the $\mathbf{X}_\sigma$, $\mathbf{F}_{\sigma\tau}$, $\mathbf{Y}_\tau$ matrices is equal to 1% of the entries of the dense stiffness matrix $\mathbf{A}$.

In Figure 4 we show the number of necessary matrix entries for the cluster-BEM and the standard-BEM as a function of the potential error in exterior points. It clearly shows that the higher the accuracy requirements are the more storage could be saved with the cluster-BEM.

7 Summary

The saddle point formulation and the fast algorithm are well-suited for solving geodetic BVPs. The former guarantees not only the well-posedness of the problem but also allows to properly include a priorily given geopotential model. The fast algorithm has the potential to speed up the assembly and solution of the linear system by $2-3$ orders of magnitude, and to reduce the storage requirements by about the same amount. Therefore, it will make high resolution global gravity field recovery feasible. The flexibility and efficiency of our method does not degrade significantly if oblique derivatives and more complex boundary surfaces are taken into account, and if higher order gravity fields have to be recovered from terrestrial data. Currently, we are working on the convergence analysis including the existence and uniqueness of the saddle point (u, p). Besides, we want to apply our algorithm to the IAG test data set which is currently being developed within subcommission 2 of the IAG Section IV special commission 1.

References

Greengard, L. and Rokhlin, V. (1997). A new version of the fast multipole method for the laplace equation in three dimensions. In Iserles A., editor, *Volume 6*, Acta Numerica, pages 229–269. Cambridge University Press.

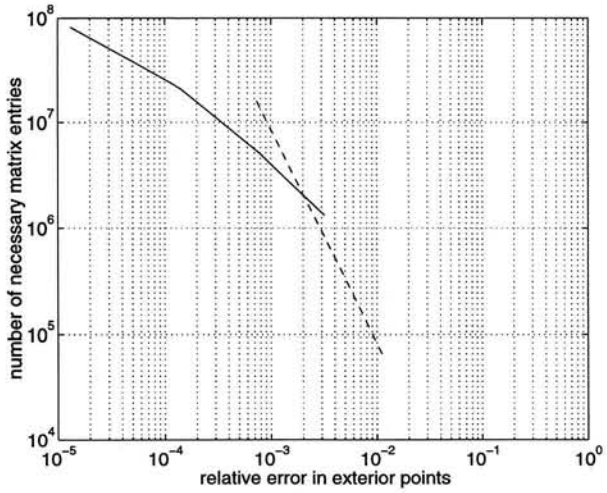

Fig. 4. Number of necessary matrix entries as function of the potential error in exterior points: standard BEM (dashed line) versus fast algorithm ($m = 7$) (solid line).

Hackbusch, W. and Nowak, Z.P. (1989). On the Fast Matrix Multiplication in the Boundary Element Method by Panel Clustering. *Numerische Mathematik*, 54(4):463–491.

Lage, C. (1998) Fast evaluation of singular kernel functions by cluster methods. Technical report, Seminar für Angewandte Mathematik, ETH Zürich, CH-8092 Zürich. In preparation.

Mikhlin, S. G. and Prößdorf. (1986). *Singular Integral Operators*. Springer, Berlin.

Rokhlin, V. (1985). Rapid solutions of integral equations of classical potential theory. *J. Comput. Phys.*, 60:187–207.

Hildebrandt, S. and Wienholtz, E. (1964). Constructive proofs of representation theorems in seperable hilbert spaces. *Comm. and Pure Appl. Math.*, 17:369–373.

DEOS Progress Letter 98.1: 43–55

Stabilization of global gravity field solutions by combining satellite gradiometry and airborne gravimetry*

Johannes Bouman and Radboud Koop

Abstract

The expected high resolution and precision of a global gravity field model derived from satellite gradiometric observations is unprecedented compared to nowadays satellite-only models. However, a dedicated gravity field mission will most certainly fly in a non-polar (sun-synchronous) orbit, such that small polar regions will not be covered with observations. The resulting inhomogeneous global data coverage, together with the downward continuation problem and coloured noise, leads to unstable global solutions and regularization is mandatory. Regularization gives rise to a bias in the solution, mainly in the polar areas.

Undoubtedly, the combination with gravity related measurements in polar areas, like airborne gravimetry, will improve the quality of the solution. Open questions are, for example, how accurate gravity anomalies must be, what spatial sampling is required, and how large the area with observations should be.

In order to answer these questions, a gravity field solution from gradiometry-only will be compared with a solution from gradiometry combined with several airborne gravimetric scenarios. Special attention is given to the quality improvement and bias reduction relative to the gradiometry-only solution. The coefficients of a spherical harmonic series are the unknowns and their errors are propagated to, for example, geoid heights.

1 Introduction

An accurate and high resolution knowledge of the earth's global gravity field is needed in geosciences. In geodesy, for example, the gravity field is needed for levelling with GPS, in oceanography it is important for studying ocean circulation and last but not least in geophysics a better knowledge of the earth's gravity field yields better boundary conditions in the study of the earth's interior.

The determination of the earth's gravity field is very convenient using satellite methods since a satellite orbiting the earth samples practically the whole globe within a relative short time span. A very promising satellite technique for global gravity field determination is satellite gravity gradiometry. With this technique one can in principle determine all frequencies up to

* Presented at the IV Hotine-Marussi Symposium, Trento, Italy, 1998

high degree and order, typically $L = 180 - 250$. Due to certain constraints on the satellite (power supply and disturbances due to heat fluctuations), the orbit of a gradiometric mission will most likely be a sun-synchronous dawn-dusk orbit, leading to numerical instability of the global gradiometric inversion due to the polar gaps. Combining the gradiometric data with gravimetric data in the polar regions, for example obtained with airborne gravimetry, should give more stable solutions.

In general the determination of the earth's gravity field at the earth's surface from satellite observations is unstable, and therefore ill-posed, because of the downward continuation problem. A stable solution can be obtained by regularizing the solution. This is well known and often Kaula's rule is used, which can be interpreted as a constraint on the signal. Inherent to the regularization is the regularization error or bias, Louis (1989); Xu (1992). A proper quality description takes into account this bias, and it is reasonable to expect that the bias decreases for a combined gradiometric-gravimetric solution compared to a gradiometric-only solution.

The purpose of this paper is to compare the quality of the different gravity field models. In particular we are interested in the effect of varying the precision, resolution and coverage of the additional gravimetric data.

The description of the gradiometric missions, the gravity anomaly data, and the observation model in Section 2 is followed by a summary of the method of regularization and the related errors in Section 3. Section 4 lists the results and Section 5 presents the conclusions.

2 Model and mission description

2.1 Observation model

The unknowns to be solved for are the normalized harmonic coefficients $\bar{C}_{lm}, \bar{S}_{lm}$ of a (truncated) spherical harmonic expansion of the gravitational potential:

$$V = \frac{GM}{R} \sum_{l=0}^{L} \left(\frac{R}{r}\right)^{l+1} \sum_{m=-l}^{l} \bar{Y}_{lm}(\theta, \lambda), \tag{1}$$

with

$$\bar{Y}_{lm}(\theta, \lambda) = \begin{cases} \bar{C}_{lm} \cos m\lambda \bar{P}_{lm}(\cos\theta), & m \geq 0 \\ \bar{S}_{l|m|} \sin |m|\lambda \bar{P}_{l|m|}(\cos\theta), & m < 0 \end{cases}, \tag{2}$$

where GM is the gravitational constant times mass of the earth, R the radius of a reference sphere enclosing all masses, l, m degree and order, $\bar{P}_{lm}(\cos\theta)$ the fully normalized Legendre functions and r, θ, λ the geocentric polar coordinates. For the maximum degree and order to be resolved we take $L = 180$, corresponding to a spatial resolution of $\approx 1°$, which is a typical resolution to be achieved from a gradiometry mission.

Gradiometry. The observations we consider are gravity anomalies and gravity gradients, i.e. the second order derivatives of the gravitational potential. The latter could for example be the change in distance between two falling proof masses around the earth. A local satellite coordinate system is x, y, z with x along-track, y cross-track and z radial. Observing the distance changes in these three directions yields the observables V_{xx}, V_{yy} and V_{zz}. By a proper coordinate transformation these values can be related to (1), see e.g. Koop (1993).

In particular we do not use the gradiometric observations themselves, the actual gravity gradients, but their along track Fourier spectrum. Let's assume that the orbit is circular, that

there are no data gaps and that after a number of revolutions the ground-track of the satellite repeats exactly. Considering the observations V_{xx} etc. as a time series along the orbit one may compute the Fourier coefficients of these observations, the lumped coefficients. These lumped coefficients are linear combinations of the potential coefficients $\bar{C}_{lm}, \bar{S}_{lm}$. Due to the assumptions (circular orbit etc.), the normal matrix becomes block-diagonal, e.g. Koop (1993); Schrama (1990). The above approach is the time-wise in the frequency domain method, with the advantage that for example coloured noise can easily be accounted for Rummel *et al.* (1993), compare Section 2.2.

Gravimetry. The most likely technique to observe gravity in the polar areas is airborne gravimetry. In our error propagations we do not use the actual airborne gravimetry observations directly, but we assume that after data processing a grid of point values of gravity anomalies at the earth's surface is available, compare Schwarz and Li (1997).

The unknowns and observations are connected by the linear model

$$E\{g\} = Af, \; D\{g\} = P^{-1} \tag{3}$$

with g the observations, f the unknowns, A the design matrix and P^{-1} the error covariance matrix of the observations. The linear model (3) is used for the satellite gradiometric as well as the airborne gravimetric observations. The unknowns are the corrections to the initial or reference potential coefficients, compare Section 4.1.

2.2 Input specifications

Two satellite gradiometric missions are considered. One with only V_{zz} observed and one with the three diagonal components V_{xx}, V_{yy}, V_{zz} observed. For the gradiometric missions we have chosen a satellite height of 250 km, a mission duration of six months and coloured noise with a PSD which is flat at the level of $10^{-3}E/\sqrt{Hz}$ for frequencies between $0.005Hz$ and $0.1Hz$, and which behaves as $1/f$ for the low frequencies below $0.005Hz$. Frequencies below 2 cpr (cycles per revolution) are not taken into account. Effectively, the lower order blocks of the normal matrix become unstable because of the coloured noise. The high pass filtering of the spectrum affects the lower orders for all degrees, Koop (1993).

The gravity anomalies are assumed to have uncorrelated errors with a standard deviation of 5, 10, or 20 mgal. Although a standard deviation of 2 mgal is an accuracy obtainable with nowadays airborne gravimetry, Schwarz (1998); Tscherning (1998), we used a minimum of 5 mgal. The errors of the gravity anomalies are correlated along-track, and a more pessimistic error assumption might compensate for the neglection of the correlation. Concerning airborne gravimetry one is referred to Schwarz and Li (1997). The anomalies are located in the polar areas and cover circular areas (polar caps) with radii $0.125°, 2.5°$ or $5°$ from the poles. The gravity anomalies are given as point values in a grid with a spacing of $0.125°$. The anomalies will be denoted with dg.

A mission like GOCE, apart from SGG observations, will make use of SST observations too, but such observations are not considered here. The polar gaps have less influence on these measurements, and only lower degrees up to e.g. 70 will be estimable from SST measurements.

3 Regularization

3.1 Least-squares solution

Usually the unknowns f can be solved by a least-squares approach minimizing the observation error

$$\min_f \|g - Af\|_P^2 \tag{4}$$

which leads to the estimate $\hat{f}$ of f

$$\hat{f} = (A^T P A)^{-1} A^T P g. \tag{5}$$

This approach, however, is no longer suitable here because $A^T P A$ is badly conditioned. The instability reflects the fact that we are dealing with an inverse problem which is ill-posed. There are four reasons for this:

- *Satellite height.* The observation noise is amplified due to the downward continuation, since we solve for the gravitational potential at the earth's surface.
- *Orbit inclination.* Every inclination not equal to 90° results in two polar gaps without observations. Hence, a global solution has to be derived from 'local' measurements.
- *Type of observation.* Every kind of observation related to the gravity potential (like gravity, satellite position or gravity gradients) will have, in the frequency domain, a different sensitivity for different frequencies. For instance, for V_{zz} the sensitivity decreases with increasing l, whereas it is constant for all orders m per degree. Or V_{yy} which has an increasing sensitivity for increasing order m. Sometimes a particular observation is not sensitive to a certain gravity field parameter at all, like V_{yz} and V_{xy} in a polar orbit which are not sensitive to the zonal harmonics or V_{yy} from which the zonal harmonics can only poorly be determined, in particular at the equator.
- *Noise characteristic.* The errors of the gradiometer can be characterized as coloured noise, compare Section 2.2. Consequently, the low order blocks of the normal matrix are unstable.

Note that the downward continuation does not have any effect, because we only solve for coeffcients up to degree 180. For V_{zz}, for example, the term $(R/r)^{l+1}$ is compensated by the term $(l+1)(l+2)$, with the given height and maximum degree. The degree truncation itself acts as regularization or stabilization. The major cause of the instability is the polar gap.

3.2 Tikhonov regularization

Least-squares does not provide a stable solution, several other methods do. One of these methods is Tikhonov regularization Tikhonov and Arsenin (1977). Instead of minimizing (4) we use

$$\min_f \|g - Af\|_P^2 + \alpha \|f\|_K^2. \tag{6}$$

One sees that in this case the solution $\hat{f}$ has to satisfy the constraint that the total power of the signal is finite. The positive real number α is the compromise between the constraint and the minimization of the observation error. The solution of (6) yields

$$\hat{f}_r = (A^T P A + \alpha K)^{-1} A^T P g. \tag{7}$$

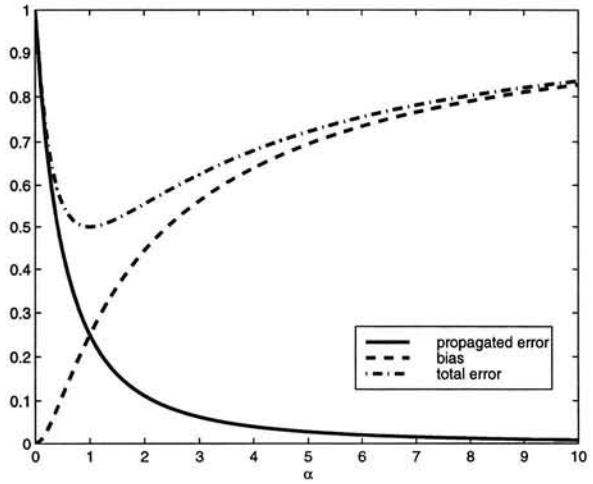

Fig. 1. Errors as function of α.

In this case K is a diagonal matrix with elements $10^{10}l^4$ which is the inverse of the well-known Kaula rule for degree-order variances.

The combination of gradiometry, gravimetry and regularization simply is

$$\hat{f}_r = ([A^TPA]_{sgg} + [A^TPA]_{ga} + \alpha K)^{-1}([A^TPg]_{sgg} + [A^TPg]_{ga}) \tag{8}$$

where sgg and ga stand for satellite gravity gradiometry and gravity anomalies respectively. Thus, the combined solution is regularized as well.

3.3 Propagated error and bias

The total error or Mean Square Error Matrix $MSEM$ consists of the propagated error

$$Q_f = (A^TPA + \alpha K)^{-1}A^TPA(A^TPA + \alpha K)^{-1} \tag{9}$$

and the regularization error (bias), Xu (1992),

$$E\{\hat{f}_r - f\} = \Delta f = -(A^TPA + \alpha K)^{-1}\alpha K f \tag{10}$$

i.e.:

$$MSEM = Q_f + \Delta f \Delta f^T. \tag{11}$$

An optimal α can be found by minimization of the trace of the $MSEM$. Other choices for an optimal α probably exist, the present one is the expected distance (2-norm) from f to $\hat{f}_r$, Hoerl and Kennard (1970). The trace of the propagated error has the form $1/(1+\alpha)^2$, which is a decreasing function for increasing α. The bias squared has the form $\alpha^2/(1+\alpha)^2$ which is an increasing function for increasing α. The sum of the two gives a function with one minimum: the optimal α, compare Fig. 1.

Remarks.

- The bias is usually neglected or said to be 100% maximum Marsh *et al.* (1988). The latter is not true, since the bias in a single coefficient does not only depend on the size of the coefficient but, due to correlation, also on the size of other coefficients, cf. Eq. (10). The bias may therefore become smaller or larger than 100% for specific coefficients.

- As can easily be seen, the same solution (7) is obtained by adding zero observations for all coefficients with weight matrix αK:[1]

$$E\{\begin{pmatrix} g \\ 0 \end{pmatrix}\} = \begin{pmatrix} A \\ I \end{pmatrix} f, \quad D\{\begin{pmatrix} g \\ 0 \end{pmatrix}\} = \begin{pmatrix} P^{-1} & 0 \\ 0 & (\alpha K)^{-1} \end{pmatrix} \tag{12}$$

  This would change the propagated error to $Q_{f_0} = (A^T P A + \alpha K)^{-1}$ and reduces the bias to zero since one would have $E\{0\} = f$ or $E\{f\} = 0$. However, as Kaula's rule already shows, $E\{0\} \neq f$, and accordingly it is not allowed to write Eq. (12). That is why the above approach has not been used.

- To avoid confusion it has to be mentioned that for error propagation no observations are needed. Only $A^T P A$, αK and f are required. Consequently, the results should be considered as an approximation of the actual feasible accuracy using real or simulated measurements.

4 Results

4.1 Assumptions

The results to be presented here are obtained applying a number of assumptions. First of all we have to assume that f is known, OSU91A (truncated at degree and order 180) is taken as our 'ground truth' for that purpose. The results are presented with respect to the reference field GRS80. Note that the choice of the reference field influences the bias computation. Using a higher degree and order reference field, for example JGM3, yields lower bias values. In practice, of course, we don't know the bias and a conservative estimate, like the application of GRS80 is giving, seems to be appropriate in our opinion.

The results presented will be optimistic because we did not take into account model errors, aliasing or the fact that the errors of the gravity anomalies are correlated. Furthermore, we are forced to use a block-diagonal Mean Square Error Matrix due to computer constraints, although the bias term really yields a full MSEM. In the Appendix it is explained what consequences this has for the error propagation. The main effect is that there is almost no east-west variation for geoid errors and that the errors tend to be symmetric with respect to the equator.

4.2 Combination with anomalies in areas of variable size

4.2.1 Results for V_{zz}

First we look at the combination of the second radial derivative of the gravitational potential combined with gravimetric data located in two polar caps of equal and increasing size, Fig.

[1] Limiting the regularization to those coefficients that are not well determined leads to truncated singular value decomposition methods, compare for example Bouman (1998).

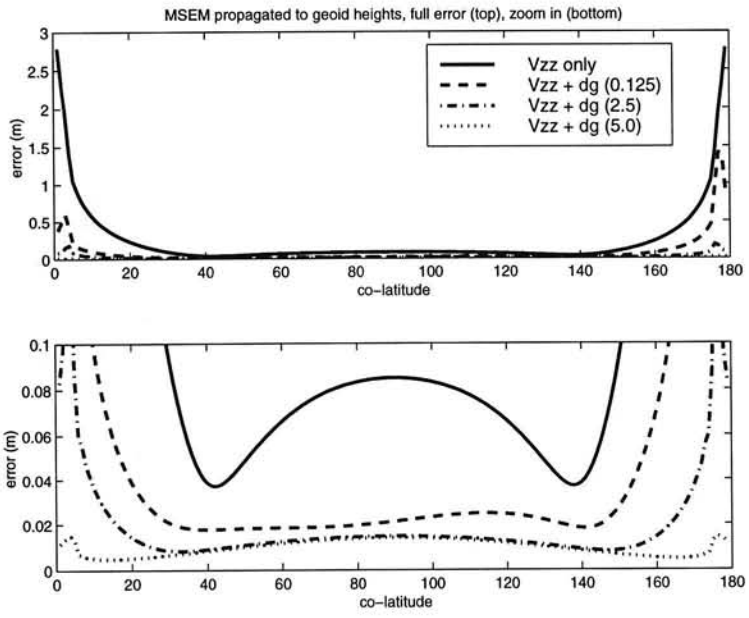

Fig. 2. Combination of V_{zz} with dg in area of variable size.

2. The three sizes of the caps are $0.125°, 2.5°$ and $5.0°$. The effect of the gravity data is most noticeable in the polar areas but the *MSEM* propagated to geoid heights decreases at other latitudes as well. This is mainly caused by the bias reduction in those areas. Since the east-west variation of the geoid height errors is negligible, one meridian is shown only. The geoid error decrease towards the poles is probably due to the increase of the number of measurements per square km. An equi-angular grid is assumed, and therefore the measurement density increases towards the poles. Although this may be an unrealistic assumption, it is considered satisfactory for our first results.

The bias is not only reduced for geoid heights. Comparing the bias in the $\bar{C}_{lm}$ coefficients for the V_{zz} case and the combination with dg ($5.0°$ cap size) one sees a dramatic reduction, Fig. 3. (Results for the $\bar{S}_{lm}$ coefficients are similar and thus not shown.) Without the gravity anomalies there is a large bias in the low orders (as expected), where regularization is needed. The combined solution has an almost homogeneous bias error.

A further extension of the area were gravity is measured is unnecessary. On the one hand the global basis functions have been constrained by the gravity data in the polar regions, on the other hand the accuracy of the gravity data is not enough to expect much improvement at lower latitudes. Compare Fig. 4 where the geoid height errors for the combination of V_{zz} with dg on a global basis and dg ($5.0°$ cap size) are displayed. The geoid improvement is negligible.

4.2.2 *Results for* V_{diag}

The combination of V_{xx}, V_{yy}, V_{zz}, or V_{diag} for short, with gravity anomalies in the polar areas only has local effect, Fig. 5. However, looking at the bias with respect to the size of the coefficients a substantial improvement comes from the combination, compare Fig. 6. Again,

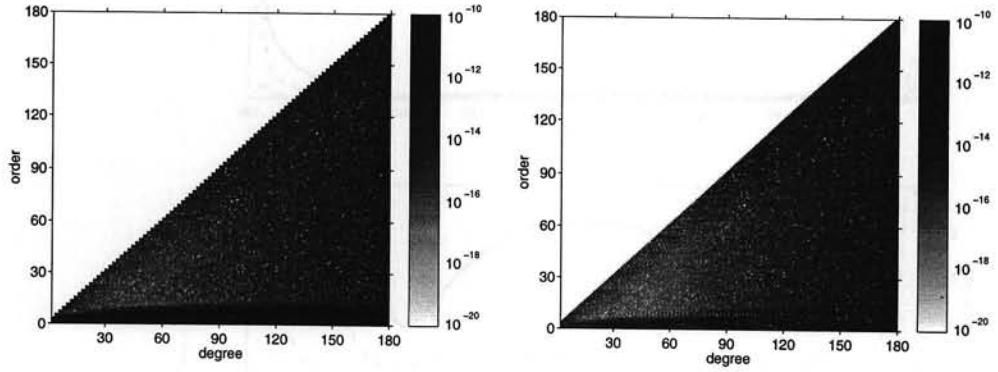

Fig. 3. Bias in the $\bar{C}_{lm}$ coefficients for V_{zz} (left) and for $V_{zz} + dg(5.0°$ cap size) (right).

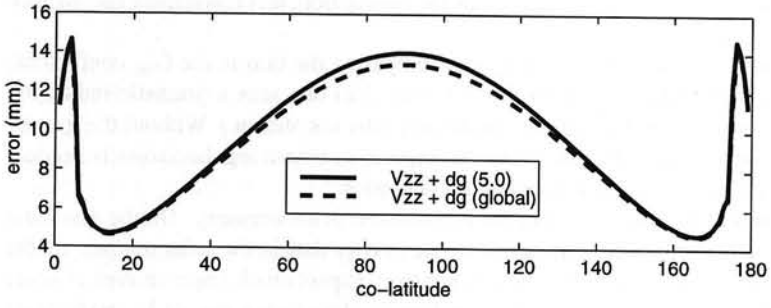

Fig. 4. Geoid height errors for the combination of V_{zz} with $dg(5.0°$ cap size) and V_{zz} with dg globally.

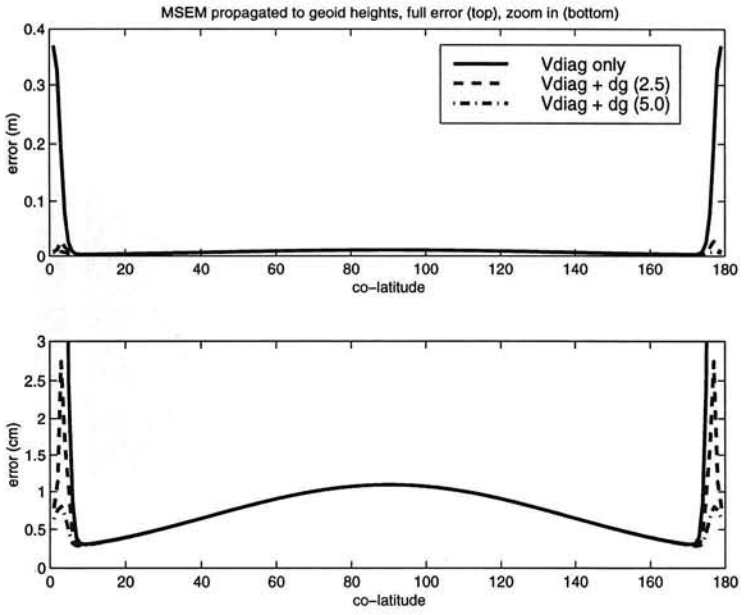

Fig. 5. Combination of V_{diag} with dg in area of variable size.

the combined solution yields a homogeneous error, the bias is several orders smaller than the signal. The bias in the gradiometry-only solution, in contrast, may become up to two orders larger than the signal for the low order coefficients (small m).

4.3 Combination with anomalies of variable measurement accuracy

The combination of gravity anomalies in both polar caps with a size of five degrees and for different dg measurement accuracy is illustrated in Fig. 7. A decrease in precision of a factor two means a decrease of weight of the normal matrix $[A^TPA]_{ga}$ with respect to $[A^TPA]_{sgg}$ of a factor four. When a homogeneous geoid precision over the whole earth is required, an anomaly precision of 5-10 mgal is sufficient. Note that in our approach a decrease of precision is equivalent to a decrease of resolution. For example, a grid spacing of a quarter of a degree instead of $0.125°$, yields four times less measurements which, in our approach, corresponds to a weight decrease of a factor four.

4.4 Summary

In summary, the bias in the coefficients is greatly reduced by adding gravity data in the polar regions to gradiometric observables. Moreover, if the accuracy of dg is chosen "correctly", a homogeneous geoid height precision is obtained. Looking at Table 1, one notices that the size of the regularization parameter decreases going from a less favourite configuration to a better configuration. The third column lists the quotient of the diagonal elememts of the bias part of the *MSEM* and the diagonal elements of the propagated error. The bias is negligible with respect to the propagated error when gravity anomalies at both poles in a dense grid with an area size of

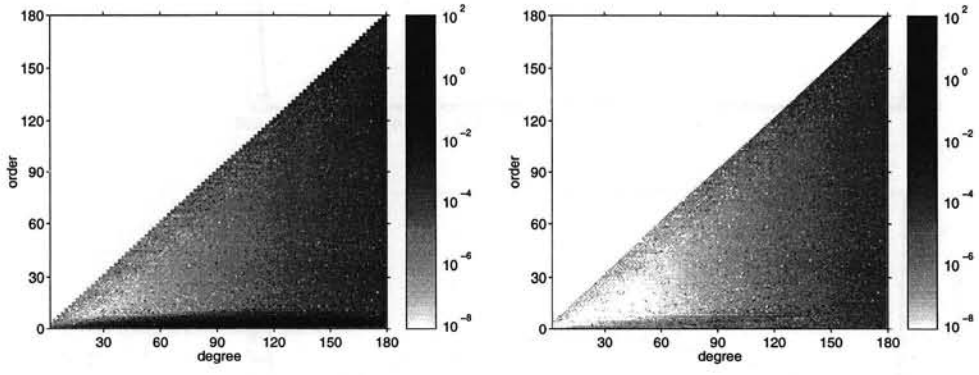

Fig. 6. Bias with respect to the size of the $\bar{C}_{lm}$ coefficients for V_{diag} (left) and for $V_{diag} + dg(5.0°$ cap size) (right).

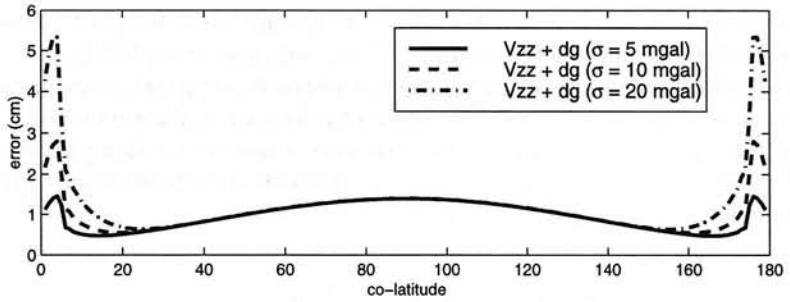

Fig. 7. Combination of V_{zz} with dg of variable measurement accuracy.

Table 1. Summary of some combinations.

Combination	α	Bias w.r.t. prop. error	Error w.r.t. $V_{zz} + 5.0$
V_{zz}	1.49	4.0	245
$V_{zz} + \Delta g(5.0°)$	0.70	10^{-5}	1
$V_{zz} + \Delta g(5.0°), \sigma = 20$ mgal	0.86	10^{-4}	1.14
$V_{zz} + \Delta g(5.0°), \Delta\theta = 0.5°$	0.86	10^{-4}	1.14
V_{diag}	7.57	1.4	2.17
$V_{diag} + \Delta g(5.0°)$	0.72	10^{-5}	0.61

five degrees are available. In column four, the trace of the *MSEM* has been compared with that of the reference case: V_{zz} combined with anomalies in a 5° cap. The improvement with respect to the gradiometry-only case is impressive, while the other cases yield comparable errors.

5 Conclusions

An anomaly precision of 5-10 mgal is sufficient, and therefore airborne gravimetry data seems useful. The areas where dg and V_{ij} are known do not have to overlap, specifically for a polar gap of 6.6°, the measurement of gravity anomalies in polar caps of 5° or even less is sufficient. When only V_{zz} gradiometric observables are available, there is precision improvement for geoid heights at lower latitudes, outside the caps, due to the combination with dg. Moreover, the bias in the lower orders for all degrees is substantially reduced. When all three diagonal elements of the gravity potential tensor have been measured, V_{diag}, there is no improvement for geoid heights at lower latitudes. However, the bias in the lower order coefficients for all degrees is substantially reduced, the polar gap is hardly noticeable anymore.

In summary: the addition of gravity in the polar regions to gradiometric observables makes the polar gap problem disappear for the SGG only case!

Acknowledgement The computations were partially performed in C++. The matrix library `newmat` developed by R. Davies facilitated this. This work is supported by the Delft University of Technology's Centre for High Performance and Applied Computing (HPαC). The information on airborne gravimetry by Klaus-Peter Schwarz and Christian Tscherning is highly appreciated. Finally, we acknowledge the remarks by Roland Klees, who critically reviewed this paper.

A Error propagation with a block-diagonal matrix

It is shown under what conditions the propagation of a block-diagonal Mean Square Error Matrix to for example geoid heights results in symmetry with respect to the equator and/or no east-west variation. Since gradiometric measurements have a homogeneous precision this is also what one would expect. The only variation is due to the polar gaps and the decrease of the number of observations for each latitude towards the equator.

The error propagation of the *MSEM*, which is the error matrix of gravity potential coeffi-

cients, to linear functionals of the potential in general has the form, cf. Haagmans and van Gelderen (1991)

$$\begin{aligned} cov(p,q) &= \sum_{m=0}^{L}\sum_{k=0}^{L}[A_{mk}\cos m\lambda_p\cos k\lambda_q + B_{mk}\sin m\lambda_p\cos k\lambda_q \\ &\quad + C_{mk}\cos m\lambda_p\sin k\lambda_q + D_{mk}\sin m\lambda_p\sin k\lambda_q] \end{aligned} \tag{13}$$

with

$$\begin{aligned} A_{mk} &= \sum_{l=m}^{L}\sum_{n=k}^{L}\lambda_l\lambda_n cov(\bar{C}_{lm},\bar{C}_{nk})\bar{P}_{lm}(\cos\theta_p)\bar{P}_{nk}(\cos\theta_q) \\ B_{mk} &= \sum_{l=m}^{L}\sum_{n=k}^{L}\lambda_l\lambda_n cov(\bar{S}_{lm},\bar{C}_{nk})\bar{P}_{lm}(\cos\theta_p)\bar{P}_{nk}(\cos\theta_q) \\ C_{mk} &= \sum_{l=m}^{L}\sum_{n=k}^{L}\lambda_l\lambda_n cov(\bar{C}_{lm},\bar{S}_{nk})\bar{P}_{lm}(\cos\theta_p)\bar{P}_{nk}(\cos\theta_q) \\ D_{mk} &= \sum_{l=m}^{L}\sum_{n=k}^{L}\lambda_l\lambda_n cov(\bar{S}_{lm},\bar{S}_{nk})\bar{P}_{lm}(\cos\theta_p)\bar{P}_{nk}(\cos\theta_q) \end{aligned} \tag{14}$$

where

l, n	degree,
m, k	order,
p, q	points on the earth's surface,
λ_l, λ_n	eigenvalues, e.g. R for geoid heights,
A_{mk}, etc.	Fourier coefficients of a two dimensional series.

Let's consider point variances only, that is, $p = q$, and assume that the *MSEM* has a block-diagonal structure, that is, $B_{mk} = C_{mk} = 0$ and $m = k$. The propagated error, which now is denoted as $cov(\theta, \lambda)$ since it is a function of one point only, then is

$$cov(\theta,\lambda) = \sum_{m=0}^{L}\left[A_m\cos^2 m\lambda + D_m\sin^2 m\lambda\right] \tag{15}$$

with

$$\begin{aligned} A_m &= \sum_{l=m}^{L}\sum_{n=m}^{L}\lambda_l\lambda_n cov(\bar{C}_{lm},\bar{C}_{nm})\bar{P}_{lm}(\cos\theta)\bar{P}_{nm}(\cos\theta) \\ D_m &= \sum_{l=m}^{L}\sum_{n=m}^{L}\lambda_l\lambda_n cov(\bar{S}_{lm},\bar{S}_{nm})\bar{P}_{lm}(\cos\theta)\bar{P}_{nm}(\cos\theta). \end{aligned} \tag{16}$$

The normal matrix, $(A^TPA + \alpha K)^{-1}$, becomes block-diagonal when observing gravity gradients in a circular orbit with exact repeat and no data gaps. Moreover, $cov(\bar{C}_{lm},\bar{C}_{nm}) = cov(\bar{S}_{lm},\bar{S}_{nm})$ for $m = 1,\ldots,L$. The normal matrix for the gravity anomalies obtains the same structure when the anomaly distribution and precision is symmetric with respect to the equator. Then Eq. (15) becomes

$$cov(\theta) = \sum_{m=0}^{L} A_m \tag{17}$$

with A_m as before, Eq. (16). There is, therefore, no east-west variation, the propagated error is independent of longitude. In our case there is some minor dependence on longitude because the bias term, $\Delta f\Delta f^T$, yields unequal C and S covariances. The variation, however, is negligible.

A further consequence of the aforementioned data distribution is the separation of even and odd degrees, that is, the error covariance is zero when $|l - n|$ is odd. Recalling the property

$$P_{lm}(-t) = (-1)^{l+m} P_{lm}(t) \tag{18}$$

the following four cases occur:

1. m is even, l, n are even; $P_{lm}(-t) = P_{lm}(t)$ and $P_{nm}(-t) = P_{nm}(t)$,
2. m is even, l, n are odd; $P_{lm}(-t) = -P_{lm}(t)$ and $P_{nm}(-t) = -P_{nm}(t)$,
3. m is odd, l, n are even; $P_{lm}(-t) = -P_{lm}(t)$ and $P_{nm}(-t) = -P_{nm}(t)$,
4. m is odd, l, n are odd; $P_{lm}(-t) = P_{lm}(t)$ and $P_{nm}(-t) = P_{nm}(t)$.

Because l and n have the same parity, the Legendre functions for a specific m are always simultaneously symmetric or anti-symmetric with respect to the equator. The combination of two of these functions, as in A_m and D_m, is therefore always north-south symmetric: $cov(\theta, \lambda) = cov(\pi - \theta, \lambda)$. Again the bias term does destroy the exact north-south symmetry. For the combination solutions we also computed the bias for the degrees not having the same parity. However, when the bias is small compared to Q_f, Eq. (9), north-south symmetry will occur as is evident from the figures.

References

Bouman, J. (1998). Quality of regularization methods. DEOS Report no 98.2, Delft Institute for Earth-Oriented Space Research.

Haagmans, R. and van Gelderen, M. (1991). Error variances-covariances of GEM-T1: their characteristics and implications in geoid computation. *Journal of Geophysical Research*, **96**(B12), 20011–20022.

Hoerl, A. and Kennard, R. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**(1), 55–67.

Koop, R. (1993). Global gravity field modelling using satellite gravity gradiometry. Publications on geodesy. New series no. 38, Netherlands Geodetic Commission.

Louis, A. (1989). *Inverse und schlecht gestellte Probleme*. Teubner.

Marsh, J., Lerch, F., Putney, B., Christodoulidis, D., Smith, D., Felsentreger, T., Sanchez, B., Klosko, S., Pavlis, E., Martin, T., Williamson, J. R. R., Colombo, O., Rowlands, D., Eddy, W., Chandler, N., Rachlin, K., Patel, G., Bhati, S., and Chinn, D. (1988). A new gravitational model for the earth from satellite tracking data: GEM-T1. *Journal of Geophysical Research*, **93**(B6), 6169–6215.

Rummel, R., van Gelderen, M., Koop, R., Schrama, E., Sansò, F., Brovelli, M., Miggliaccio, F., and Sacerdote, F. (1993). Spherical harmonic analysis of satellite gradiometry. Publications on geodesy. New series no. 39, Netherlands Geodetic Commission.

Schrama, E. (1990). Gravity field error analysis: application of GPS receivers and gradiometers on low orbiting platforms. TM 100769, NASA.

Schwarz, K. (1998). Private communication.

Schwarz, K. and Li, Z. (1997). An introduction to airborne gravimetry and its boundary value problems. In F. Sansò and R. Rummel, editors, *Geodetic Boundary Value Problems in View of the One Centimeter Geoid*, volume 65 of *Lecture notes in earth sciences*, pages 312–358. Springer-Verlag.

Tikhonov, A. and Arsenin, V. (1977). *Solutions of ill-posed problems*. Winston and Sons.

Tscherning, C. (1998). Private communication.

Xu, P. (1992). The value of minimum norm estimation of geopotential fields. *Geophysical Journal International*, **111**, 170–178.

DEOS Progress Letter 98.1: 57–67

The shift operators and translations of spherical harmonics

Martin van Gelderen

Abstract

Solid and surface spherical harmonics functions have very simple transformation properties with respect to the gradient and angular momentum operators. These properties can be utilized for the derivation of translation relations of the spherical harmonic functions.

1 Introduction

Already many papers have been published about the transformational properties of the spherical harmonics functions. To cite a few: Hobson (1955), Rose (1957), Aardoom (1969), Giacaglia (1980) and Epton and Dembart (1994). The formulas presented in this paper are not new, but they are derived in a particular straightforward manner which we believe to be much simpler than often found in other literature.

First we give some definitions, then we show the properties of the operators applied and finally we show how they can be used to derive translation relations for spherical harmonic functions and their coefficients.

2 General properties

Since many definitions can be found for the spherical harmonic functions, we first start with the definitions used in this paper. For simplicity, the complex spherical harmonics are used; defined as in e.g. Edmonds (1957). We start with the *associated Legendre function*:

$$P_{\ell,m}(t) = \frac{1}{2^\ell \ell!}(1-t^2)^{m/2}\frac{d^{\ell+m}}{dt^{\ell+m}}(t^2-1)^\ell.$$

It has the following symmetry with respect to order m

$$P_{\ell,-m}(t) = (-1)^m \frac{(\ell-m)!}{(\ell+m)!}P_{\ell,m}(t).$$

The *spherical harmonics* are defined as

$$Y_{\ell,m}(\theta,\lambda) = P_{\ell,m}(\cos\theta)\,\mathrm{e}^{im\lambda}.$$

Often we will write

$$Y_{\ell,m}(\mathbf{x}) = Y_{\ell,m}(\mathbf{x}/\|\mathbf{x}\|) \equiv Y_{\ell,m}(\theta,\lambda);$$

with $\mathbf{x} \in \mathbb{R}^3$ and where $\|\mathbf{x}\|$ is the Eucledian norm of $\mathbf{x}$. For the spherical harmonics a *normalisation*

$$\overline{Y}_{\ell,m} = \beta_{\ell,m} Y_{\ell,m} \quad \text{with} \quad \beta_{\ell,m} = (-1)^m \sqrt{\frac{(2\ell+1)(\ell-m)!}{4\pi(\ell+m)!}} \tag{1}$$

can be used such that the spherical harmonics are ortho-normal:

$$\int\int_{\sigma} \overline{Y}_{\ell,m}(\mathbf{x})\, \overline{Y}^*_{\ell',m'}(\mathbf{x})\, d\mathbf{x} = \delta_{\ell'\ell}\, \delta_{m'm}.$$

The asterix * denotes the complex conjugate; the integration is taken over the (unit) sphere. For the *regular solid spherical harmonics* $\|\mathbf{x}\|^\ell Y_{\ell,m}(\mathbf{x})$ and the *irregular solid spherical harmonics* $\frac{1}{\|\mathbf{x}\|^{\ell+1}} Y_{\ell,m}(\mathbf{x})$ the following abbreviations are introduced:

$$\begin{aligned} S_{\ell,m}(\mathbf{x}) &= (-1)^m(\ell-m)!\, \frac{1}{\|\mathbf{x}\|^{\ell+1}} Y_{\ell,m}(\mathbf{x}) \\ R_{\ell,m}(\mathbf{x}) &= (-1)^m \frac{1}{(\ell+m)!}\, \|\mathbf{x}\|^\ell Y_{\ell,m}(\mathbf{x}). \end{aligned} \tag{2}$$

With respect to the sphere, they are only orthogonal; but the 'normalisation' used here will render very simple formulas. From the definitions it is easily derived that the following symmetry relations hold:

$$\begin{aligned} Y^*_{\ell,m} &= (-1)^m \frac{(\ell+m)!}{(\ell-m)!}\, Y_{\ell,-m} & Y_{\ell,m}(\mathbf{x}) &= (-1)^\ell Y_{\ell,m}(-\mathbf{x}) \\ \overline{Y}^*_{\ell,m} &= (-1)^m\, \overline{Y}_{\ell,-m} & \overline{Y}_{\ell,m}(\mathbf{x}) &= (-1)^\ell \overline{Y}_{\ell,m}(-\mathbf{x}) \\ S^*_{\ell,m} &= (-1)^m S_{\ell,-m} & S_{\ell,m}(\mathbf{x}) &= (-1)^\ell S_{\ell,m}(-\mathbf{x}) \\ R^*_{\ell,m} &= (-1)^m R_{\ell,-m} & R_{\ell,m}(\mathbf{x}) &= (-1)^\ell R_{\ell,m}(-\mathbf{x}). \end{aligned} \tag{3}$$

Apart from the usual geocentric cartesian frame $\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z$ a new frame $\mathbf{e}_-, \mathbf{e}_0, \mathbf{e}_+$ is used (Van Gelderen, 1999) with:

$$\begin{aligned} \mathbf{e}_- &= \frac{1}{\sqrt{2}}(\mathbf{e}_x - i\mathbf{e}_y) \\ \mathbf{e}_0 &= \mathbf{e}_z \\ \mathbf{e}_+ &= -\frac{1}{\sqrt{2}}(\mathbf{e}_x + i\mathbf{e}_y). \end{aligned}$$

All *covariant* vector components v_x, v_y, v_z of the vector $\mathbf{v}$ transform in the same way:

$$\begin{aligned} v_- &= \frac{1}{\sqrt{2}}(v_x - iv_y) \\ v_0 &= v_z \\ v_+ &= -\frac{1}{\sqrt{2}}(v_x + iv_y). \end{aligned} \tag{4}$$

For the radial distance $r \equiv \|\mathbf{x}\|$ we then obtain

$$r^2 = x_x^2 + x_y^2 + x_z^2 = x_0^2 - 2x_+x_-.$$

From the well-known recursion relations of the Legendre functions, see e.g. Ilk (1983), the following relations for the solid spherical harmonics are derived:

$$\begin{aligned}
(\ell+m)x_0R_{\ell,m} &= r^2R_{\ell-1,m} + (\ell-m+1)\sqrt{2}x_+R_{\ell,m-1} \\
(\ell+m)\sqrt{2}x_-R_{\ell,m} &= (\ell-m+1)x_0R_{\ell,m-1} - r^2R_{\ell-1,m-1} \\
2mx_0R_{\ell,m} &= (\ell-m+1)\sqrt{2}x_+R_{\ell,m-1} - (\ell+m+1)\sqrt{2}x_-R_{\ell,m+1} \\
(\ell-m+1)x_0S_{\ell,m} &= r^2S_{\ell+1,m} - (\ell+m)\sqrt{2}x_+S_{\ell,m-1} \\
(\ell+m+1)x_0S_{\ell,m} &= r^2S_{\ell+1,m} - (\ell-m)\sqrt{2}x_-S_{\ell,m+1} \\
2mx_0S_{\ell,m} &= (\ell+m)\sqrt{2}x_+S_{\ell,m-1} - (\ell-m)\sqrt{2}x_-S_{\ell,m+1}.
\end{aligned} \tag{5}$$

3 The ladder operators

In this section operators are introduced which change the spherical harmonics by one degree or order. Two differential operators are used: the *gradient operator* ∇ and the *angular momentum operator* $\mathbf{L}$. The gradient operator with respect to the cartesian basis reads:

$$\begin{aligned}
\nabla f &= \mathbf{e}_x\frac{\partial f}{\partial x} + \mathbf{e}_y\frac{\partial f}{\partial y} + \mathbf{e}_z\frac{\partial f}{\partial z} \\
&\equiv (\mathbf{e}_x\nabla_x + \mathbf{e}_y\nabla_y + \mathbf{e}_z\nabla_z)f;
\end{aligned}$$

f is a funtion in $\mathbb{R}^3$. The gradient operator can be split up into a radial and a surface part:

$$\nabla = \mathbf{e}_r\nabla_r + \frac{1}{r}\nabla_{\text{surf}}$$

The operator $\mathbf{L}$ is a tangential vector operator, i.e. the vector $\mathbf{L}f$ is tangential to the sphere, defined as

$$\begin{aligned}
\mathbf{L} &= -i\mathbf{e}_r \times \nabla = -ir\mathbf{e}_r \times \nabla_{\text{surf}} \qquad \Leftrightarrow \\
\nabla &= \mathbf{e}_r\nabla_r + \frac{1}{r}\nabla_{\text{surf}} = \mathbf{e}_r\nabla_r - \frac{i}{r}(\mathbf{e}_r \times \mathbf{L}),
\end{aligned} \tag{6}$$

with $\mathbf{e}_r$ the radial basis vector. The vector $\mathbf{L}f$ is always perpendicular to $\mathbf{e}_r$ and ∇f, which can directly be seen from its definition; see e.g. Jackson (1967) for more definitions and properties of these operators. The components of both operators with respect to the $\{\mathbf{e}_-, \mathbf{e}_0, \mathbf{e}_+\}$ are defined as (4) since always covariant differentiation is used:

$$\begin{aligned}
\nabla_\pm &= \mp\frac{1}{\sqrt{2}}(\nabla_x \pm i\nabla_y), \qquad \nabla_0 = \nabla_z \\
\mathrm{L}_\pm &= \mp\frac{1}{\sqrt{2}}(\mathrm{L}_x \pm i\mathrm{L}_y), \qquad \mathrm{L}_0 = \mathrm{L}_z.
\end{aligned}$$

First the operators $\mathrm{L}_{-,0,+}$ are applied to the spherical harmonics. Their action on the $Y_{\ell,m}$ is straightforward; this is related to the fact that they are the joint eigenfunctions of the operators

$\mathbf{L}^2$ and L_0; see Edmonds (1957). Since $\mathbf{L}$ is a pure tangential operator, also the relations for the solid harmonics are found directly:

$$\begin{aligned} \mathrm{L}_- Y_{\ell,m} &= -\frac{(\ell+m)(\ell-m+1)}{\sqrt{2}}\, Y_{\ell,m-1} \\ \mathrm{L}_0 Y_{\ell,m} &= m\, Y_{\ell,m} \\ \mathrm{L}_+ Y_{\ell,m} &= \frac{1}{\sqrt{2}} Y_{\ell,m+1} \end{aligned} \tag{7}$$

$$\begin{aligned} \mathrm{L}_- \overline{Y}_{\ell,m} &= \sqrt{\frac{(\ell+m)(\ell-m+1)}{2}}\, \overline{Y}_{\ell,m-1} \\ \mathrm{L}_0 \overline{Y}_{\ell,m} &= m\, \overline{Y}_{\ell,m} \\ \mathrm{L}_+ \overline{Y}_{\ell,m} &= -\sqrt{\frac{(\ell-m)(\ell+m+1)}{2}}\, \overline{Y}_{\ell,m+1} \end{aligned} \tag{8}$$

$$\begin{aligned} \mathrm{L}_- S_{\ell,m} &= \frac{\ell+m}{\sqrt{2}}\, S_{\ell,m-1} \\ \mathrm{L}_0 S_{\ell,m} &= m\, S_{\ell,m} \\ \mathrm{L}_+ S_{\ell,m} &= -\frac{\ell-m}{\sqrt{2}} S_{\ell,m+1} \end{aligned} \tag{9}$$

$$\begin{aligned} \mathrm{L}_- R_{\ell,m} &= \frac{\ell-m+1}{\sqrt{2}}\, R_{\ell,m-1} \\ \mathrm{L}_0 R_{\ell,m} &= m\, R_{\ell,m} \\ \mathrm{L}_+ R_{\ell,m} &= -\frac{\ell+m+1}{\sqrt{2}} R_{\ell,m+1} \end{aligned} \tag{10}$$

Now the $\boldsymbol{\nabla}$ operator is applied to the surface spherical harmonics. It is decomposed into a radial and surface component (6). In components, see Van Gelderen (1999),

$$\begin{pmatrix} \nabla_- \\ \nabla_0 \\ \nabla_+ \end{pmatrix} = \frac{1}{r} \begin{pmatrix} x_- \\ x_0 \\ x_+ \end{pmatrix} \nabla_r + \frac{1}{r^2} \begin{pmatrix} x_-\mathrm{L}_0 - x_0\mathrm{L}_- \\ x_-\mathrm{L}_+ - x_+\mathrm{L}_- \\ x_0\mathrm{L}_+ - x_+\mathrm{L}_0 \end{pmatrix}.$$

The action of the first part on spherical harmonics is straighforward; for the second part the equations (9-10) are used. By applying the recurrence relations (5) the outcome can be reduced to very simple expressions:

$$\begin{aligned} \nabla_- S_{\ell,m} &= -\frac{1}{\sqrt{2}} S_{\ell+1,m-1} \\ \nabla_0 S_{\ell,m} &= -S_{\ell+1,m} \\ \nabla_+ S_{\ell,m} &= -\frac{1}{\sqrt{2}} S_{\ell+1,m+1} \end{aligned} \tag{11}$$

and

$$\begin{aligned}\nabla_- R_{\ell,m} &= -\frac{1}{\sqrt{2}} R_{\ell-1,m-1} \\ \nabla_0 R_{\ell,m} &= R_{\ell-1,m} \\ \nabla_+ R_{\ell,m} &= -\frac{1}{\sqrt{2}} R_{\ell-1,m+1}\end{aligned} \tag{12}$$

With (1-2) we then find

$$\begin{aligned}\nabla_- \frac{1}{r^{\ell+1}} Y_{\ell,m} &= \frac{(\ell-m+1)(\ell-m+2)}{\sqrt{2}} \frac{1}{r^{\ell+2}} Y_{\ell+1,m-1} \\ \nabla_0 \frac{1}{r^{\ell+1}} Y_{\ell,m} &= -(\ell-m+1) \frac{1}{r^{\ell+2}} Y_{\ell+1,m} \\ \nabla_+ \frac{1}{r^{\ell+1}} Y_{\ell,m} &= \frac{1}{\sqrt{2}} \frac{1}{r^{\ell+2}} Y_{\ell+1,m+1}\end{aligned} \tag{13}$$

$$\begin{aligned}\nabla_- \frac{1}{r^{\ell+1}} \overline{Y}_{\ell,m} &= -\sqrt{\frac{(2\ell+1)(\ell-m+1)(\ell-m+2)}{2(2\ell+3)}} \frac{1}{r^{\ell+2}} \overline{Y}_{\ell+1,m-1} \\ \nabla_0 \frac{1}{r^{\ell+1}} \overline{Y}_{\ell,m} &= -\sqrt{\frac{(2\ell+1)(\ell+m+1)(\ell-m+1)}{(2\ell+3)}} \frac{1}{r^{\ell+2}} \overline{Y}_{\ell+1,m} \\ \nabla_+ \frac{1}{r^{\ell+1}} \overline{Y}_{\ell,m} &= -\sqrt{\frac{(2\ell+1)(\ell+m+1)(\ell+m+2)}{2(2\ell+3)}} \frac{1}{r^{\ell+2}} \overline{Y}_{\ell+1,m+1}\end{aligned} \tag{14}$$

$$\begin{aligned}\nabla_- r^\ell Y_{\ell,m} &= \frac{(\ell+m)(\ell+m-1)}{\sqrt{2}} r^{\ell-1} Y_{\ell-1,m-1} \\ \nabla_0 r^\ell Y_{\ell,m} &= (\ell+m) r^{\ell-1} Y_{\ell-1,m} \\ \nabla_+ r^\ell Y_{\ell,m} &= \frac{1}{\sqrt{2}} r^{\ell-1} Y_{\ell-1,m+1}\end{aligned} \tag{15}$$

$$\begin{aligned}\nabla_- r^\ell \overline{Y}_{\ell,m} &= -\sqrt{\frac{(2\ell+1)(\ell+m)(\ell+m-1)}{2(\ell-1)}} r^{\ell-1} \overline{Y}_{\ell-1,m-1} \\ \nabla_0 r^\ell \overline{Y}_{\ell,m} &= \sqrt{\frac{(2\ell+1)(\ell+m)(\ell-m)}{(2\ell-1)}} r^{\ell-1} \overline{Y}_{\ell-1,m} \\ \nabla_+ r^\ell \overline{Y}_{\ell,m} &= -\sqrt{\frac{(2\ell+1)(\ell-m)(\ell-m-1)}{2(\ell-1)}} r^{\ell-1} \overline{Y}_{\ell-1,m+1}\end{aligned} \tag{16}$$

The components of the operators ∇ and $\mathbf{L}$ are sometimes called *ladder operators* since they relate a spherical harmonic function to another of one degree and/or order higher or lower. The $\mathrm{L}_\pm$ are real ladder operators for the surface spherical harmonics in the sense that they only make one step in the m (order) for a fixed degree. This is related to the fact that all the surface

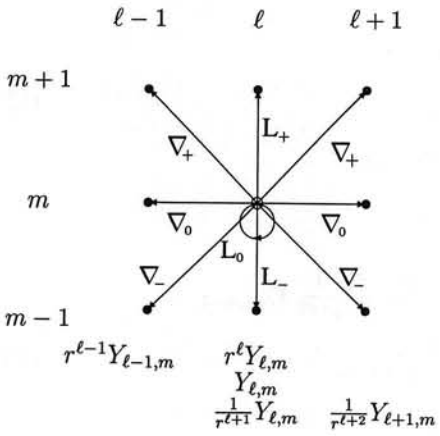

Fig. 1. The ladder operators of spherical harmonics functions

spherical harmonics of fixed degree ℓ form a basis for a $2\ell+1$ dimensional representation of the Lie algebra so3; cf. Hamermesh (1964). The $\nabla_\pm$ also go one step up or down in the m-direction, but always increase (irregular solid spherical harmonics) or decrease (regular solid spherical harmonics) the degree.

The spherical harmonics linked up to eachother by the ladder operators are depicted in Figure 1.

As $S_{0,0} = Y_{0,0} = \frac{1}{r}$, all irregular solid spherical harmonics can be obtained from the iterative use of the ∇-operators using (11,13,14):

$$S_{\ell,m} = (-1)^\ell \, 2^{|m|/2} \, {\nabla_\pm}^{|m|} {\nabla_0}^{\ell-|m|} \frac{1}{r} \tag{17}$$

$$\frac{1}{r^{\ell+1}} Y_{\ell,m} = (-1)^{\ell-m} \, 2^{|m|/2} \, \frac{1}{(\ell-m)!} \, {\nabla_\pm}^{|m|} {\nabla_0}^{\ell-|m|} \frac{1}{r}, \tag{18}$$

$$\frac{1}{r^{\ell+1}} \overline{Y}_{\ell,m} = (-1)^\ell 2^{|m|/2} \sqrt{\frac{(2\ell+1)}{4\pi(\ell+m)!(\ell-m)!}} \, {\nabla_\pm}^{|m|} {\nabla_0}^{\ell-|m|} \frac{1}{r}, \tag{19}$$

where $\nabla_\pm$ denotes ∇_+ for $m \geq 0$ and ∇_- for $m < 0$. Since $S_{\ell,m}$ is a harmonic function we have:

$$\Delta S_{\ell,m} = ({\nabla_0}^2 - 2\nabla_+\nabla_-) S_{\ell,m} = 0 \Leftrightarrow 2\nabla_+\nabla_- S_{\ell,m} = {\nabla_0}^2 S_{\ell,m}.$$

This property can be used to reduce combination of powers of ∇_- and ∇_+:

$$(\sqrt{2}\nabla_+)^m (\sqrt{2}\nabla_-)^{m'} = \begin{cases} (\sqrt{2}\nabla_+)^{m-m'} \, {\nabla_0}^{2m'} & m \geq m' \\ (\sqrt{2}\nabla_-)^{m'-m} \, {\nabla_0}^{2m} & m < m' \end{cases} \tag{20}$$

4 Translation relations

The inverse distance is expanded into spherical harmonics as (Hobson, 1955)

$$\frac{1}{\|\mathbf{x}-\mathbf{y}\|} = \sum_{\ell,m} R^*_{\ell,m}(\mathbf{y}) S_{\ell,m}(\mathbf{x}) = \sum_{\ell,m} R_{\ell,m}(\mathbf{y}) S^*_{\ell,m}(\mathbf{x}) \qquad \|\mathbf{y}\| < \|\mathbf{x}\|. \tag{21}$$

Translated spherical harmonics can be obtained easily from this expansion with (17):

$$\begin{aligned} S_{\ell,m}(\mathbf{x}-\mathbf{y}) &= (-1)^\ell 2^{|m|/2} {\nabla_\pm}^{|m|} {\nabla_0}^{\ell-|m|} \frac{1}{\|\mathbf{x}-\mathbf{y}\|} \\ &= \sum_{\ell'=0}^{\infty} \sum_{m'=-\ell'}^{\ell'} R^*_{\ell',m'}(\mathbf{y}) (-1)^\ell 2^{|m|/2} \, {\nabla_\pm}^{|m|} {\nabla_0}^{\ell-|m|} S_{\ell',m'}(\mathbf{x}). \end{aligned}$$

With, using (20),

$$\begin{aligned} (-1)^\ell 2^{|m|/2} {\nabla_\pm}^{|m|} {\nabla_0}^{\ell-|m|} S_{\ell',m'}(\mathbf{x}) &= (-1)^{\ell+\ell'} 2^{(|m+m'|)/2} \, {\nabla_\pm}^{|m+m'|} {\nabla_0}^{\ell+\ell'-|m+m'|} \frac{1}{\|\mathbf{x}\|} \\ &= S_{\ell+\ell',m+m'}(\mathbf{x}) \end{aligned}$$

this can be written as

$$S_{\ell,m}(\mathbf{x}-\mathbf{y}) = \sum_{\ell'=0}^{\infty} \sum_{m'=-\ell'}^{\ell'} R^*_{\ell',m'}(\mathbf{y}) S_{\ell+\ell',m+m'}(\mathbf{x}). \tag{22}$$

This is the translation relation for the irregular solid spherical harmonics. For the translation of the regular spherical harmonics it is less straightforward to find the relation directly; see Rose (1958) or Epton and Dembart (1994). Much easier is to start from the expansion of the inverse distance and apply (22):

$$\begin{aligned} \frac{1}{\|\mathbf{x}-\mathbf{y}\|} &= \sum_{\ell,m} R^*_{\ell,m}(\mathbf{y}) S_{\ell,m}(\mathbf{x}) = \sum_{\ell,m} R^*_{\ell,m}(\mathbf{y}-\boldsymbol{\Delta}) S_{\ell,m}(\mathbf{x}-\boldsymbol{\Delta}) \\ &= \sum_{\ell,m} R^*_{\ell,m}(\mathbf{y}-\boldsymbol{\Delta}) \sum_{\ell',m'} R^*_{\ell'-\ell,m'-m}(\boldsymbol{\Delta}) S_{\ell',m'}(\mathbf{x}) \\ &= \sum_{\ell',m'} S_{\ell',m'}(\mathbf{x}) \sum_{\ell,m} R^*_{\ell',m'}(\mathbf{y}-\boldsymbol{\Delta}) R^*_{\ell-\ell',m-m'}(\boldsymbol{\Delta}). \end{aligned}$$

Since the spherical harmonics are a set of independent basis vectors, the expansion coefficients of a function with respect to them are unique. Confrontation of the last with the first line of the equation above gives:

$$R^*_{\ell,m}(\mathbf{y}) = \sum_{\ell',m'} R^*_{\ell',m'}(\mathbf{y}-\boldsymbol{\Delta}) R^*_{\ell-\ell',m-m'}(\boldsymbol{\Delta}) \Leftrightarrow$$

$$R_{\ell,m}(\mathbf{x}+\mathbf{y}) = \sum_{\ell',m'} R_{\ell',m'}(\mathbf{y}) R_{\ell-\ell',m-m'}(\mathbf{x}). \tag{23}$$

Using the symmetry relations (3), the following set of relations can be derived from (22-23):

$$
\begin{aligned}
S_{\ell,m}(\mathbf{x}-\mathbf{y}) &= \sum_{\ell'=0}^{\infty}\sum_{m'=-\ell'}^{\ell} R^*_{\ell',m'}(\mathbf{y}) S_{\ell+\ell',m+m'}(\mathbf{x}) \qquad (24)\\
&= \sum_{\ell'=0}^{\infty}\sum_{m'=-\ell'}^{\ell} (-1)^{m'} R_{\ell',m'}(\mathbf{y}) S_{\ell+\ell',m-m'}(\mathbf{x})\\
&= \sum_{\ell'=\ell}^{\infty}\sum_{m'=-\ell'}^{\ell} R^*_{\ell'-\ell,m'-m}(\mathbf{y}) S_{\ell',m'}(\mathbf{x})\\
&= \sum_{\ell'=\ell}^{\infty}\sum_{m'=-\ell'}^{\ell} (-1)^{m'-m} R_{\ell'-\ell,m-m'}(\mathbf{y}) S_{\ell',m'}(\mathbf{x})\\
R_{\ell,m}(\mathbf{x}-\mathbf{y}) &= \sum_{\ell'=0}^{\ell}\sum_{m'=-\ell'}^{\ell} (-1)^{\ell'} R_{\ell',m'}(\mathbf{y}) R_{\ell-\ell',m-m'}(\mathbf{x}) \qquad (25)\\
&= \sum_{\ell'=0}^{\ell}\sum_{m'=-\ell'}^{\ell} (-1)^{\ell'+m'} R^*_{\ell',m'}(\mathbf{y}) R_{\ell-\ell',m+m'}(\mathbf{x}).
\end{aligned}
$$

or

$$
\begin{aligned}
S_{\ell,m}(\mathbf{x}+\mathbf{y}) &= \sum_{\ell'=0}^{\infty}\sum_{m'=-\ell'}^{\ell} (-1)^{\ell'} R^*_{\ell',m'}(\mathbf{y}) S_{\ell+\ell',m+m'}(\mathbf{x}) \qquad (26)\\
&= \sum_{\ell'=0}^{\infty}\sum_{m'=-\ell'}^{\ell} (-1)^{\ell'+m'} R_{\ell',m'}(\mathbf{y}) S_{\ell+\ell',m-m'}(\mathbf{x})\\
&= \sum_{\ell'=\ell}^{\infty}\sum_{m'=-\ell'}^{\ell} (-1)^{\ell'} R^*_{\ell'-\ell,m'-m}(\mathbf{y}) S_{\ell',m'}(\mathbf{x})\\
&= \sum_{\ell'=\ell}^{\infty}\sum_{m'=-\ell'}^{\ell} (-1)^{\ell'+m'-m} R_{\ell'-\ell,m-m'}(\mathbf{y}) S_{\ell',m'}(\mathbf{x})\\
R_{\ell,m}(\mathbf{x}+\mathbf{y}) &= \sum_{\ell'=0}^{\ell}\sum_{m'=-\ell'}^{\ell} R_{\ell',m'}(\mathbf{y}) R_{\ell-\ell',m-m'}(\mathbf{x}) \qquad (27)\\
&= \sum_{\ell'=0}^{\ell}\sum_{m'=-\ell'}^{\ell} (-1)^{m'} R^*_{\ell',m'}(\mathbf{y}) R_{\ell-\ell',m+m'}(\mathbf{x}).
\end{aligned}
$$

From the relations above, the equivalent relations for the $Y_{\ell,m}$ and the $\overline{Y}_{\ell,m}$ are found directly with (1) and (2). The inverse distance reads

$$
\begin{aligned}
\frac{1}{\|\mathbf{x}-\mathbf{y}\|} &= \sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell} \frac{(\ell-m)!}{(\ell+m)!} \|\mathbf{y}\|^{\ell} Y^*_{\ell,m}(\mathbf{y}) \frac{1}{\|\mathbf{x}\|^{\ell+1}} Y_{\ell,m}(\mathbf{x})\\
&= \sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell} \frac{4\pi}{2\ell+1} \|\mathbf{y}\|^{\ell} \overline{Y}^*_{\ell,m}(\mathbf{y}) \frac{1}{\|\mathbf{x}\|^{\ell+1}} \overline{Y}_{\ell,m}(\mathbf{x}).
\end{aligned}
$$

Translation of the irregular solid harmonics:

$$\frac{1}{\|\mathbf{x}-\mathbf{y}\|^{\ell+1}} Y_{\ell,m}(\mathbf{x}-\mathbf{y}) = \sum_{\ell',m'} \frac{(\ell+\ell'-m-m')!}{(\ell-m)!(\ell'+m')!} \|\mathbf{y}\|^{\ell'} Y^*_{\ell',m'}(\mathbf{y}) \frac{1}{\|\mathbf{x}\|^{\ell+\ell'+1}} Y_{\ell+\ell',m+m'}(\mathbf{x}).$$

and for the normalized $\overline{Y}_{\ell,m}$

$$\frac{1}{\|\mathbf{x}-\mathbf{y}\|^{\ell+1}} \overline{Y}_{\ell,m}(\mathbf{x}-\mathbf{y}) = \sum_{\ell',m'} \sqrt{\frac{4\pi(2\ell+1)}{(2\ell+2\ell'+1)(2\ell'+1)}} \cdot \sqrt{\frac{(\ell-m)!(\ell'+m')!(\ell+\ell'+m+m')!}{(\ell+m)!(\ell'-m')!(\ell+\ell'-m-m')!}} \|\mathbf{y}\|^{\ell'} \overline{Y}^*_{\ell',m'}(\mathbf{y}) \frac{1}{\|\mathbf{x}\|^{\ell+\ell'+1}} \overline{Y}_{\ell+\ell',m+m'}(\mathbf{x}).$$

And for the regular solid harmonics we have

$$\|\mathbf{x}-\mathbf{y}\|^{\ell} Y_{\ell,m}(\mathbf{x}-\mathbf{y}) = \sum_{\ell',m'} (-1)^{\ell'} \frac{(\ell+m)!}{(\ell'+m')!} \frac{1}{(\ell-\ell'+m-m')!} \cdot \|\mathbf{y}\|^{\ell'} Y_{\ell',m'}(\mathbf{y}) \|\mathbf{x}\|^{\ell-\ell'} Y_{\ell-\ell',m-m'}(\mathbf{x})$$

and for the normalized $\overline{Y}_{\ell,m}$:

$$\|\mathbf{x}-\mathbf{y}\|^{\ell} \overline{Y}_{\ell,m}(\mathbf{x}-\mathbf{y}) = \sum_{\ell',m'} (-1)^{\ell'} \sqrt{\frac{4\pi(2\ell+1)}{(2\ell-2\ell'+1)(2\ell'+1)}} \cdot \frac{1}{\sqrt{(\ell-\ell'-m+m')!(\ell-\ell'+m-m')!}} \sqrt{\frac{(\ell-m)!(\ell+m)!}{(\ell'-m')!(\ell'+m')!}} \cdot \|\mathbf{y}\|^{\ell'} \overline{Y}_{\ell',m'}(\mathbf{y}) \|\mathbf{x}\|^{\ell-\ell'} \overline{Y}_{\ell-\ell',m-m'}(\mathbf{x}).$$

Translation relations for coefficients

Often a spherical harmonic expansion is used to represent a function in 3D space. If the origin of the expansion is shifted, all its coefficients change. This is directly derived from the properties of the spherical harmonics. We take the following example. For a mass density ρ contained in a volume V, the total potential is

$$G \int_V \frac{\rho(\mathbf{y})}{\|\mathbf{x}-\mathbf{y}\|} \, d\mathbf{y}.$$

We define the potential function ϕ and apply (21):

$$\begin{aligned}
\phi(\mathbf{x}) &= G \int_V \frac{\rho(\mathbf{y}-\mathbf{x}_0)}{\|\mathbf{x}-\mathbf{y}\|} \, d\mathbf{y} \\
&= G \int_V \rho(\mathbf{y}-\mathbf{x}_0) \sum_{\ell,m} R^*_{\ell,m}(\mathbf{y}) S_{\ell,m}(\mathbf{x}) \, d\mathbf{y} \\
&= \sum_{\ell,m} G \int_V \rho(\mathbf{y}-\mathbf{x}_0) R^*_{\ell,m}(\mathbf{y}) \, d\mathbf{y} \; S_{\ell,m}(\mathbf{x}) \\
&= \sum_{\ell,m} \underbrace{G \int_V \rho(\mathbf{y}-\mathbf{x}_0) R^*_{\ell,m}(\mathbf{y}-\mathbf{x}_0) \, d\mathbf{y}}_{\equiv M_{\ell,m}(\mathbf{x}_0)} S_{\ell,m}(\mathbf{x}-\mathbf{x}_0),
\end{aligned} \tag{28}$$

where $\mathbf{x}_0$ is a point of reference. The coefficients $M_{\ell,m}(\mathbf{x}_0)$ are the *multipole coefficients* of the potential due to the mass distribution $\rho(\mathbf{x})$ with respect to the origin $\mathbf{x}_0$. It is actually a Laurent series. If there is no mass outside a sphere of radius a around $\mathbf{x}_0$ then convergence is guarenteed for $\|\mathbf{x}\| > a$. Likewise we obtain the Taylor expansion

$$\phi(\mathbf{x}) = \sum_{\ell,m} \underbrace{G \int_V \rho(\mathbf{y}-\mathbf{x}_0) S^*_{\ell,m}(\mathbf{y}-\mathbf{x}_0)\, d\mathbf{y}}_{\equiv L_{\ell,m}(\mathbf{x}_0)} R_{\ell,m}(\mathbf{x}-\mathbf{x}_0);$$

where the $L_{\ell,m}(\mathbf{x}_0)$ are the *local expansion coefficients* of the potential of the mass distribution $\rho(\mathbf{x})$ with respect to the origin $\mathbf{x}_0$. If there is only mass *outside* the sphere of radius b around $\mathbf{x}_0$, then we have convergence for $\|\mathbf{x}\| < b$.

The translation relations for the coefficients are obtained by inserting (from (26))

$$S_{\ell,m}(\mathbf{x}-\mathbf{x}_0) = \sum_{\ell',m'} (-1)^{\ell'} R^*_{\ell'-\ell,m'-m}(\boldsymbol{\Delta}) S_{\ell',m'}(\mathbf{x}-\mathbf{x}_0-\boldsymbol{\Delta})$$

into (28)

$$\phi(\mathbf{x}) = \sum_{\ell',m'} S_{\ell',m'}(\mathbf{x}-(\mathbf{x}_0+\boldsymbol{\Delta})) \underbrace{\sum_{\ell,m} (-1)^{\ell'} M_{\ell,m}(\mathbf{x}_0) R^*_{\ell'-\ell,m'-m}(\boldsymbol{\Delta})}_{= M_{\ell,m}(\mathbf{x}_0+\boldsymbol{\Delta})}.$$

Likewise the translation relations for local coefficients and the relation for the multipole to local expansioncoefficients are obtained:

$$\begin{aligned}
M_{\ell,m}(\mathbf{x}_0+\boldsymbol{\Delta}) &= \sum_{\ell',m'} (-1)^{\ell'} M_{\ell',m'}(\mathbf{x}_0) R^*_{\ell-\ell',m-m'}(\boldsymbol{\Delta}) && \|\mathbf{x}\| > \|\boldsymbol{\Delta}\| + a \\
L_{\ell,m}(\mathbf{x}_0+\boldsymbol{\Delta}) &= \sum_{\ell',m'} L_{\ell',m'}(\mathbf{x}_0) R_{\ell-\ell',m-m'}(\boldsymbol{\Delta}) && \|\mathbf{x}\| < b - \|\boldsymbol{\Delta}\| \\
L_{\ell,m}(\mathbf{x}_0+\boldsymbol{\Delta}) &= \sum_{\ell',m'} (-1)^{\ell'+m'} M_{\ell',m'}(\mathbf{x}_0) S^*_{\ell+\ell',m-m'}(\boldsymbol{\Delta}) && \|\mathbf{x}\| > \|\boldsymbol{\Delta}\| - a.
\end{aligned}$$

The translation of the center of expansion is only allowed if the convergence criteria for the new expansion are fulfilled. This leads to the criteria indicated above.

With the translation (26) also a double expansion of the inverse distance can be constructed:

$$\begin{aligned}
\frac{1}{\|\mathbf{x}-\mathbf{y}\|} &= \sum_{\ell,m} R^*_{\ell,m}(\mathbf{y}) S_{\ell,m}(\mathbf{y}) \\
&= \sum_{\ell,m} R^*_{\ell,m}(\mathbf{y}-\mathbf{y}_0) S_{\ell,m}(\mathbf{y}-\mathbf{y}_0) \\
&= \sum_{\ell,m} R^*_{\ell,m}(\mathbf{y}-\mathbf{y}_0) \sum_{\ell',m'} (-1)^{\ell'} R^*_{\ell',m'}(\mathbf{x}-\mathbf{x}_0) S_{\ell+\ell',m+m'}(\mathbf{x}_0-\mathbf{y}_0) \\
&= \sum_{\substack{\ell,m \\ \ell',m'}} \xi_{\ell,\ell',m,m'}(\mathbf{x}_0,\mathbf{y}_0) R^*_{\ell,m}(\mathbf{y}-\mathbf{y}_0) R^*_{\ell',m'}(\mathbf{x}-\mathbf{x}_0)
\end{aligned} \tag{29}$$

with the coefficients

$$\xi_{\ell,\ell',m,m'}(\mathbf{x}_0,\mathbf{y}_0) = (-1)^{\ell'} S_{\ell+\ell',m+m'}(\mathbf{x}_0 - \mathbf{y}_0).$$

Where $\mathbf{x}$ is close to the expansion centre $\mathbf{x}_0$ and $\mathbf{y}$ to $\mathbf{y}_0$. More exactly we can state that for the convergence of the expansion it is required $\sup \|\mathbf{x} - \mathbf{x}_0\| + \sup \|\mathbf{y} - \mathbf{y}_0| < \|\mathbf{x}_0 - \mathbf{y}_0\|$.

Applying (29) to the potential ϕ, point $\mathbf{x}_0$ can be used as the local expansion centre for the potential and $\mathbf{y}_0$ as the local centre for the multipole coefficients of the mass distribution. Obviously this new expansion directly relates to the multipole and local expansion:

$$\phi(\mathbf{x}) = \sum_{\ell',m'} \sum_{\ell,m} \underbrace{\underbrace{G \int_V R^*_{\ell,m}(\mathbf{y} - \mathbf{y}_0) d\mathbf{y}}_{M_{\ell,m}(\mathbf{y}_0)} \xi_{\ell,\ell',m,m'}(\mathbf{x}_0,\mathbf{y}_0)}_{L^*_{\ell',m'}(\mathbf{x}_0)} R^*_{\ell',m'}(\mathbf{x} - \mathbf{x}_0).$$

References

Aardoom, L. (1969). Some transformation properties for the coefficients in a spherical harmonics expansion of the earth's gravitational potential. *Tellus*, **4**, 572–584.

Edmonds, A. R. (1957). *Angular Momentum in Quantum Mechanics*. Princeton University Press, Princeton, New Jersey.

Epton, M. A. and Dembart, B. (1994). Mutlipole translation theory for the three-dimensional laplace and helmholz equations. *SIAM J. Sci. Comp.*, **16**(4), 865–897.

Gelderen, M. van (1999). Tensor spherical harmonics and group theory for physical geodesy. to be published.

Giacaglia, G. E. O. (1980). Transformations of spherical harmonics and applications to geodesy and satellite theory. *Studia geoph. et geod.*, **24**, 1–11.

Hamermesh, M. (1964). *Group Theory, and its Application to Physical Problems*. Addison-Wesley, Reading, Massachusets.

Hobson, E. W. (1955). *The theory of spherical and ellipsoidal harmonics*. Cambridge University Press.

Ilk, K.-H. (1983). Ein Beitrag zur Dynamik ausgedehnter Körper – Gravitationswechselwirkung –. Reihe C, Heft 288,Deutsche Geodätische Kommission.

Jackson, J. (1967). *Classical Electrodynamics*. John Wiley & Sons, New York.

Rose, M. E. (1957). *Elementary theory of angular momentum*. John Wiley & Sons.

Rose, M. E. (1958). The electrostatic interaction of two arbitrairy charge distributions. *J. Math. and Phys.*, **37**, 215–222.

DEOS Progress Letter 98.1: 69–87

A gravity mission for Earth sciences*

Roland Klees and Radboud Koop

Abstract

In spite of the developments in accurate gravimetry and the use of artificial earth orbiting satellites for global gravity field determination, the quality of our current global gravity field models remains inhomogeneous and too poor for some applications in geodesy, geodynamics and oceanography. These applications would greatly benefit from the determination of a global gravity field model of homogeneous and high accuracy as it seems feasible now from the dedicated ESA gravity field mission GOCE, which is currently under development. The main objective of GOCE is the recovery of the Earth's gravity field with a homogeneous accuracy of less than 5 cm in terms of geoid heights and less than 2 mgal in terms of gravity anomalies at a resolution of about 100 km. Error covariance propagation studies have already shown the feasibility of GOCE to reach its goals. Now more sophisticated end-to-end closed-loop simulation studies are going on, aiming at a detailed modelling of the satellite and all its instruments in order to carefully estimate the effect of all kinds of instrument errors on the final gravity field recovery. Once the satellite will be operational millions of observations have to be processed and reduced to geophysical end products. Data reduction and analysis methods methods and software have to be developed, for which different strategies for geophysical parameter estimation should be considered. A key issue here is the quality assessment of the measurement data as well as of the scientific data products like geoid heights, geoid slopes, gravity anomalies or harmonic coefficients.

1 Introduction

Many fundamental questions concerning the exact nature of the dynamics of the solid Earth are unsettled. In this context, the analysis of seismic wave propagation, magnetic field, and Earth gravity field provides the most valuable source of information about the nature and composition of our planet, and about evolutionary processes which continue to shape it. Snieder (1998) stresses the role of seismology in this context; Wortel (1998) tells us more about the dynamics of processes. Our focus will be on the gravity field of the Earth, which is the subject of the research program D of the Vening Meinesz Research School of Geodynamics (VMSG). More

* Presented at "Views on the dynamic Earth", Symposium 20 November 1998, Utrecht, The Netherlands

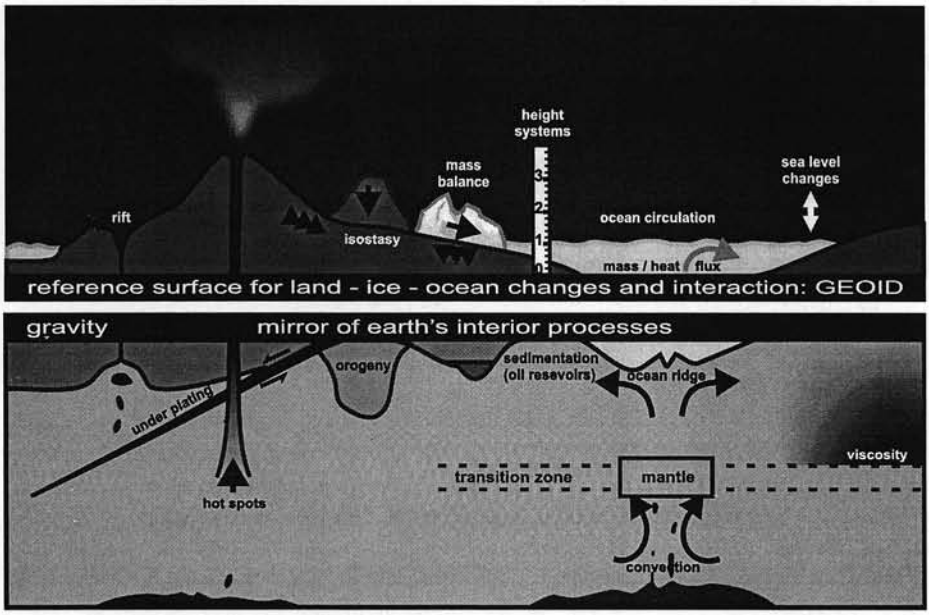

Fig. 1. The dual role of the gravity field in Earth sciences

precise, we want to discuss current research at the Delft Institute for Earth-Oriented Space Research (DEOS) related to the improvement of our knowledge about the gravity field driven by requirements from geodynamics. The geophysical implications of an improved model of the static and time-varying gravity field is addressed by Wahr (1998).

Let us first make a few remarks concerning the role of the gravity field in Earth sciences. First of all, the gravity field is a mirror of various processes in the Earth's interior. In order to understand this we simply have to remember that the mass distribution inside the Earth and the Earth's rotation generate the gravity field. If there would be no geodynamics at all, the Earth gravity field would be that of a slowly rotating fluid in hydrostatic equilibrium. Therefore, any difference between the gravity field of the real Earth and that of the equilibrium figure reflects the anomalous density structure inside the Earth. These density anomalies are due and related to a number of processes and features over a wide range of scales from global to regional. Examples are the structure of the lithosphere, orogenic processes, the existence and characteristics of sedimentary basins, and the temperature and viscosity variations in the upper mantle (cf. figure 1).

The second role of the gravity field in Earth sciences, which is in some sense dual to the first one, is to serve as a reference surface of all topographic processes. This becomes obvious when one recalls that the geoid, the equipotential or level surface of the Earth's gravity field at mean sea level, represents the hypothetical ocean surface at rest. So it is the surface, which heights are referred to, and it determines in which direction water flows. Thus, it is the natural

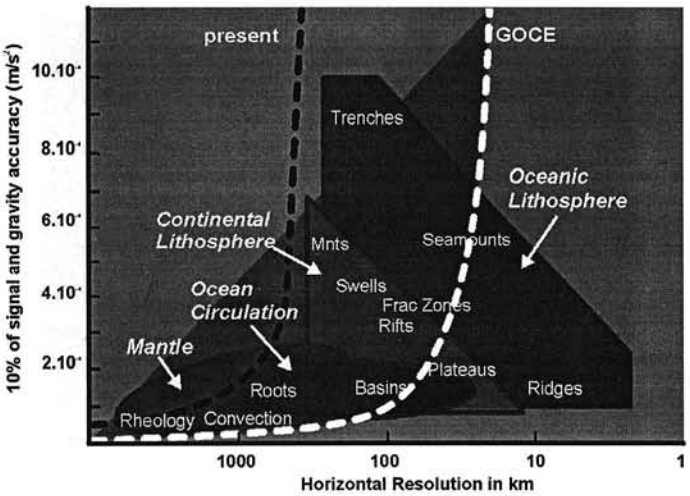

Fig. 2. Required accuracy as function of horizontal resolution necessary to resolve various geophysical features

reference surface for the topography of land and ice surfaces (several km) and their temporal variations (several m) as well as for the topography of the oceans (up to 2 m).

2 The goals

The gravity signal and the spatial pattern of geodynamic processes and geophysical features determine the requirements we have to impose on gravity field models in terms of accuracy and resolution. Figure 2 illustrates the required accuracy as a function of horizontal resolution necessary in order to resolve the quoted geodynamical and tectonic features. The read dashed line indicates that the gravity signal of most of the characteristic features of interest cannot be resolved yet. This weakness in gravity field knowledge is related to the limitations of current observation techniques, mainly terrestrial gravimetry, satellite altimetry, and conventional satellite tracking. For instance, after more than 50 years of *terrestrial gravimetry*, surface gravity data are very precise, but still highly incomplete, inhomogeneous with many gaps (high mountain areas, shallow water areas, polar regions, lakes), and often contaminated by systematic error. *Satellite altimetry* measures so to say the ocean geoid but is much too approximate, since actually the real and not the idealized ocean surface is measured; it deviates from the geoid at the meter level. Gravity field modelling by *satellite orbit analysis* of many, mostly non-geodetic satellites using various ground-based tracking techniques at many observatories, can only resolve the long wavelength features, i.e., wavelengths of a few thousand kilometers and longer. Therefore neither the accuracy nor the resolution of current geopotential models can be expected to improve significantly by additional data from conventional gravity field sensors.

The aim must be to move the read line in figure 2 further to the right. This will be achieved

	Accuracy		Spatial Resolution
	Geoid	Gravity	(half wavelength)
Ocean Circulation - Small scale - Bassin scale	2 cm <1 cm		60-250 km 1000 km
Geodynamics - Continental lithosphere (thermal structure, post-glacial rebound) - Mantle composition, rheology - Ocean lithosphere and interaction with asthenosphere (subduction processes)		1-2 10^{-5}m/s^2 1-2 10^{-5}m/s^2 5-10 10^{-5}m/s^2	50-400 km 100-5000 km 100-200 km
Geodesie - Ice and land vertical movements - Rock basement under polar ice sheets - World-wide height system	2 cm <5 cm	1-5 10^{-5}m/s^2	100-200 km 50-100 km 50-100 km

Fig. 3. GOCE scientific requirements (from ESA 1996)

by the dedicated gravity field mission GOCE. More precise, the aim of GOCE is to determine the global gravity field with a resolution of 100 km or better and a global homogeneous accuracy of $(1-10)\cdot 10^{-5}\, m/s^2$ and $1-5$ cm in terms of gravity anomalies and geoid heights, respectively (cf. figure 3). This is an improvement of the *resolution* with a factor 3 to 4 and of the *accuracy* of several orders of magnitude over the state-of-the-art global gravity field models.

3 The idea

A significant improvement can only be expected from new satellite techniques. However, any satellite technique for gravity field mapping has one pitfall, which is illustrated in figure 4. The bottom panel shows the gravity signal at the Earth's surface. It mainly reflects the topography, thus containing many short-wavelength features. The top panel shows what signal is left at an altitude of 250 km. Obviously, small scale features are highly damped, and only the dominant large scale features are still visible. Therefore, altitude acts as a low pass filter, and the gravity signal of small scale mass inhomogeneities can hardly be seen at satellite altitude.

Basically there are two possibilities to counteract the damping effect: first of all, we can fly the satellite as low as possible. The lowest altitude, however, is limited to $200-250$ km for technical, financial, and safety reasons, and this is not sufficient to meet the goals given in figure 3. The second possibility is indicated in figure 5. The top panel shows the information content of the second radial derivative of the potential at 250 km altitude. Obviously, this quantity has much more power in the short wavelengths compared to the first derivative (bottom panel of figure 5) , which is of course due to the differentiation. Derivatives of the potential are provided by differencing techniques. For instance, second derivatives of the potential can be derived from acceleration differences between two proof masses over very short distances (some decimeters). This is called "differential accelerometry". In general, the term "gravity gradiometry" is used to indicate the measurement of second order potential derivatives. These quantities are much more sensitive to the fine structures of the gravitational field, and this sensitivity increases with

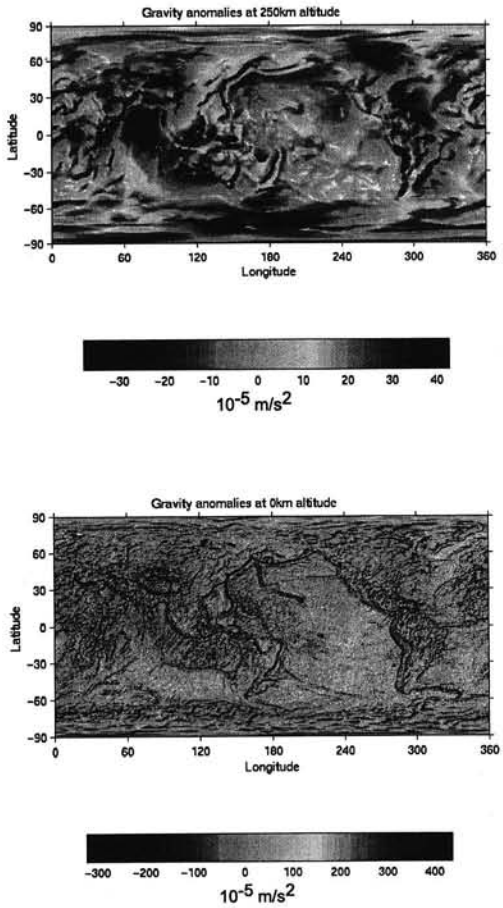

Fig. 4. The attenuation effect

decreasing distance between the proof masses.

4 The concept

The idea can now be transformed into a concept as shown in figure 6. In a low orbiting satellite at, say, 250 km altitude, test masses are tied by springs to their equilibrium position, the center of mass, on three perpendicular axes, but each mass is otherwise free to move along its spring axis. At the positions of the test masses, the compensation of gravitational force and centrifugal force is not complete, resulting in small tidal accelerations that move the test masses away from their equilibrium positions. These displacements are a measure of the gravitational acceleration at the location of the proof mass; the acceleration difference, divided by the distance between the proof masses is a first order approximation of the gravity gradient in the direction of the

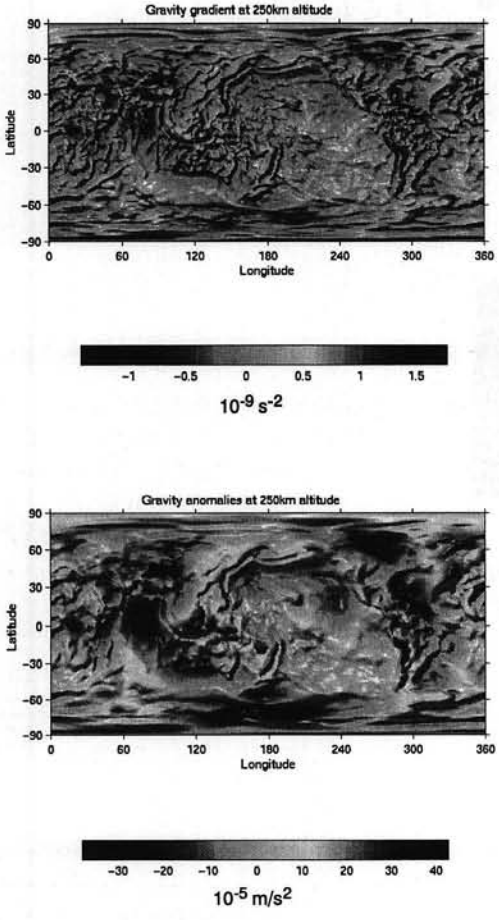

Fig. 5. The attenuation effect counteracted

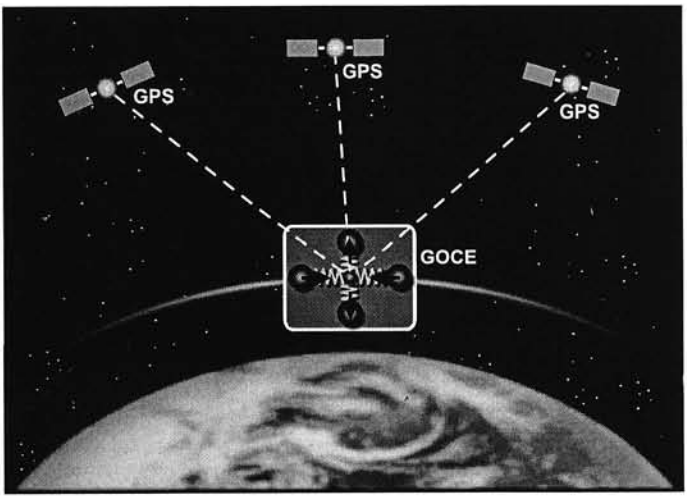

Fig. 6. GOCE concept

axes. In practice, residual *non-inertial accelerations* of the frame are unavoidable (e.g., due to atmospheric drag); they are eliminated when forming the *difference* of the displacements of pairs of masses on one axis. The sum of the displacements then measures these non-inertial accelerations, and that information may be used to restitute the orbit via thrusters, i.e., to compensate for non-inertial accelerations. *Rotation* of the frame, i.e., angular velocities and angular accelerations, also affect the difference of the displacements. We may, however, correct for frame rotation if acceleration differences are taken in all possible spatial combinations ("full tensor gradiometer"). The displacements are very small, typically about 10^{-7} m.

They have to be determined to about 7 significant digits to meet the mission goals, say, once every second (in that time the satellite moves about 8 km), so we are talking about length scales of the order of 10^{-14} m. That is truly remarkable when one recalls that the radius of an atomic nucleus is only one order of magnitude smaller.

The present concept of the GOCE gradiometer has 3 pairs of aligned accelerometers (figure 7). One pair is pointing along track, one pair perpendicular to the orbit plane, and one pair pointing towards the Earth. This configuration is able to recover the three diagonal terms of the gravity gradient tensor. It will also provide the non-diagonal terms, but with degraded performance, since the proof masses will also move a little bit in the directions perpendicular to the sensitive axis, but the springs in these directions are much stiffer. Finally, the linear acceleration due to surface forces such as atmospheric drag and solar radiation pressure, and the rotation of the gradiometer frame, i.e., the angular velocity and the angular accelerations are recovered. The accelerometers are of course not simple springs but the springs are realised either by electrostatic suspension or through magnetic levitation with superconducting coils. But even then, the high accuracy level can only be maintained over a certain time period of less than, say, 200 s, corresponding to a spatial resolution of some 800 km, due to instrument drifts. That means

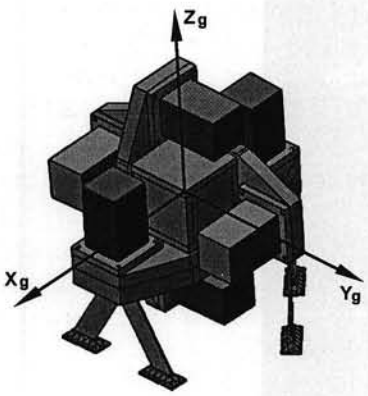

Fig. 7. GOCE 3 axes gradiometer

wavelengths above 800 km can only be determined with reduced accuracy. To build such a gradiometer is an extremely demanding task, but an even bigger challenge is how such a sensitive instrument can be isolated from the mechanical, electromagnetic, and thermal environment in the spacecraft. Therefore, instrument calibration is one of the most delicate issues. For more details we refer to (Morgan & Paik 1988).

In order to recover the gravity field we have to know where the observations have been taken, i.e., we have to know the orbit of the GOCE satellite at any time epoch. This is done by tracking simultaneously the satellites of the NAVSTAR-GPS (cf. figure 6). The GPS tracking data, however, have also another purpose. Any GPS satellite and the GOCE satellite can be treated a pair of moving proof masses in the total gravitational field. From tracking the position of the low proof mass (GOCE satellite) w.r.t. the position of the high proof mass (GPS satellite) and the known GPS satellite orbits, we may recover the Earth's gravitational field according to the principle of gradiometry (cf. figure 6). Compared to gravity gradiometry, the differencing effect is less pronounced since the distance between the GPS proof mass and the GOCE proof mass is about 20000 km. Consequently, only the long wavelength features of the geopotential can be recovered. This is complementary to the characteristics of the gravity gradiometer measurements, which are strong at the medium and short wavelengths, but less strong at the long wavelengths and weak at the very long wavelengths. Therefore, the GOCE mission will make use of both concepts in order to resolve the entire spectrum up to a maximum resolution.

5 The design

Scientific mission objectives and mission concept have to be converted into a mission design (e.g., satellite orbit) and system design (e.g., payload, attitude and drag control, satellite system, ground segment, launcher). Our contribution to find an "optimal" design, i.e., a design that

meets the scientific goals and is feasible from a technical and financial point of view, is mainly devoted to a detailed gravity field error analysis for various possible options. The goal of such an error analysis is to quantify for any given mission and system design and observational error characterstic, the expected accuracy of recovered potential coefficients and gravity field functionals. Every time the design has been changed, a new error analysis has to be done. Till now, the error analysis is mostly based on a covariance propagation using a more or less adequate linear observation model connecting observations and gravity field parameters. This allows on a case-by-case basis, without simulating any observations, to study the effects on gravity field functionals like geoid heights and gravity anomalies of, e.g., satellite altitude and orbit, stochastic model, observation type. Let us give two examples of mission and system design aspects for which error propagation studies are being done.

The first example illustrates the role of satellite altitude, one of the important mission design parameters. From a scientific point of view low altitudes are preferred in order to counteract the attenuation effect. On the other hand, at low altitudes aerodynamic forces and torques are also higher. This requires higher thrust levels to compensate for atmospheric drag, i.e., higher electric power making the mission much more expensive. For GOCE a mean orbit of 250 km has been chosen, mainly from spacecraft constraints. It is the altitude that can be maintained by ion propulsion with a power demand of the order of 500 W; altitudes below $200 - 250$ km are not allowed because of the requirement for the spacecraft not to re-enter before 7 days in case of failure. The task is to investigate what the relation is between satellite altitude and scientific mission requirements. This relation depends on many parameters, among them the assumed measurement noise level and the type of observation. Figure 8 shows the result of an error propagation study. It indicates the expected geoid commission error as a function of the satellite altitude for (i) various observation types (i.e., full tensor, diagonal, cross-track component) and (ii) various measurement noise levels (white noise over the entire measurement frequency band). For instance, when measuring the three diagonal elements of the tensor at an altitude of 250 km a measurement noise level of about $5 \cdot 10^{-4}$ E is required in order to keep the averaged geoid commission error over 1×1 degree blocks below 2 cm. When increasing the altitude by only 40 km, the gradiometer performance has to be improved by a factor of 5.

These results are based on the assumption that the measurement noise is white over the entire measurement spectrum. This is a best case scenario. More realistic is that this specification can only be met over a subband above 27 cycles per revolution (cpr), which corresponds to a frequency of $5 \cdot 10^{-3}$ Hz. Below 27 cpr the exact noise characteristic is still unknown, but currently it is expected that the noise increases somehow proportional to $1/f$, where f denotes the frequency. Finally, below, say, 4 cpr ($8 \cdot 10^{-4}$ Hz) the measurements are likely to contain no useful information at all due to gradiometer drift. It is therefore useful to look into the effect of any band-limitation of the instrument. Figure 9 shows an extreme example of the effect on the estimated potential coefficients using gravity gradient observations only. If no information is available below 27 cpr, no improvement of the long wavelengths and only little improvements of the short wavelengths would be achieved compared to the current situation. If information is available above 4 cpr the situation becomes more favorable. Of course, this is still

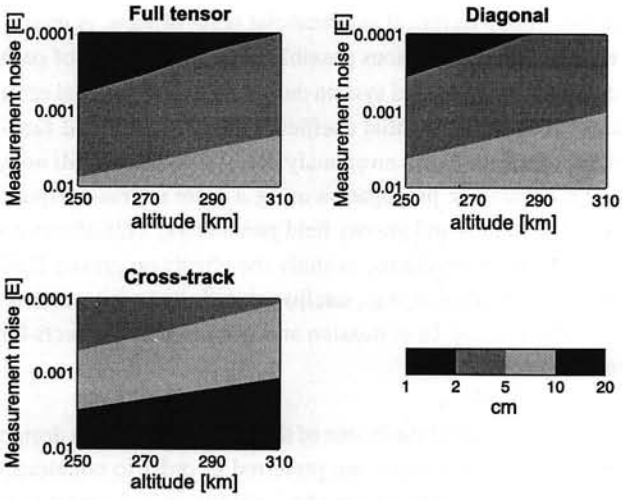

Fig. 8. Geoid commission error over 1×1 degree blocks as function of satellite altitude and gradiometer measurement noise

worse compared with the most favorate situation of a white noise error spectrum over the entire measurement spectrum. This also indicates the importance of the measurement bandwidth and the stochastic model of the observations for proper error propagation studies.

6 Data center

Currently we are involved in phase A of GOCE, which aims at (i) the finalization of mission and system design, especially mission duration, gravity gradiometry and satellite-to-satellite tracking performance objectives, selection of gradiometer type, drag free control approach, demonstration of performance, (ii) the definition of the satellite development programme, e.g., launcher and ground control, and (iii) making cost estimates.

In ESA's mission development program phase A will be followed by phases B-E. It is likely that at a certain point in the program a data center will be set up for scientific data analysis (figure 10). Such a data center will consist of five task units: the sensor unit, the data processor, the end product unit, the quality assessment unit, and the simulator. The sensor unit will provide the instrument readouts (raw data), calibration data and various corrections. Moreover, it computes calibrated and corrected gravity gradients and GPS observations, and provides information about linear accelerations, angular velocity and angular accelerations incl. a stochastic model for the various types of observations. The data processor unit, which forms the heart of the data center, consists of all tools for data analysis and synthesis. The quality assessment unit aims at comparing the results with ground truth, e.g., orbits, regional geoid models, gravity and geoid profiles, and test field data. In addition, it contains tools for statistical testing and post mission calibration. The end product unit will provide the user with various end products such as gridded geoid heights and gravity anomalies, geoid slopes, and propagated error estimates.

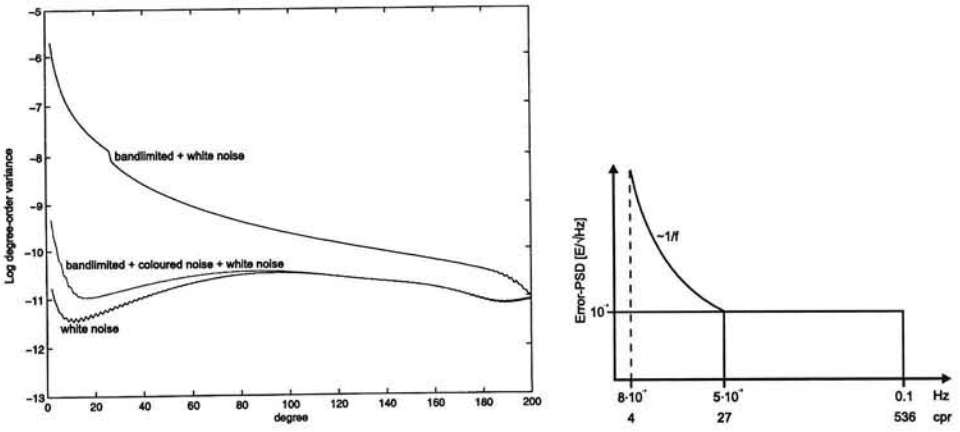

Fig. 9. Effect of band-limitation of the GOCE gradiometer on degree-order variances of the gravitational potential

Finally, the simulator allows to perform full-scale simulations prior and during the mission.

Full-scale simulations prior to the mission are important in order to investigate whether or not the mission goals will be achieved depending on the mission and system design. Full-fledged simulations of the GOCE mission are currently being done under contract of ESA and in collaboration with the industrial prime contractor by the SID consortium, a cooperation between DEOS, the Dutch Space Research Organisation (SRON) and the Institute for Astronomical and Physical Geodesy at the Technical University of Munich (IAPG). The main goal is a realistic description of the quality of the observations and a proper propagation of the observation errors to any type of gravity field functionals, such as potential coefficients, gravity anomalies, geoid heights, and geoid slopes. In order to end up with a realistic error budget, the various error sources have to be identified and described, e.g., sensor errors (e.g., gradiometer, star camera, GPS antenna and receiver), control unit errors (e.g., drag and attitude control), and environmental effects (e.g., orbit, gravity field and non-conservative forces). Moreover, the interaction between sensors, control loops, actuators, and other subsystems have to be taken into account ("closed-loop" simulation).

The closed-loop simulation (figure 11) starts with a given set of gravity gradients and information about satellite position and orientation, disturbing forces, and a model for various instrumental errors, e.g., various misalignment errors, scaling errors, and non-perfect drag and attitude control. The motion of the coupled system of proof masses is modelled by a system of differential equations ("forward step"). From the solution of the equations of motion and after adding read-out errors linear accelerations, Euler accelerations, and gravity gradients *as measured* are computed ("backward step"). The linear accelerations are due to non-gravitational forces mainly atmospheric drag and solar radiation pressure. They are fed into the drag free control system (DFC). Then, a pair of ion thrusters corrects for these disturbances such that the orbit is reconstituted. From the measured Euler accelerations and the observations of a star camera any attitude motion of the satellite can be computed. This information is fed into

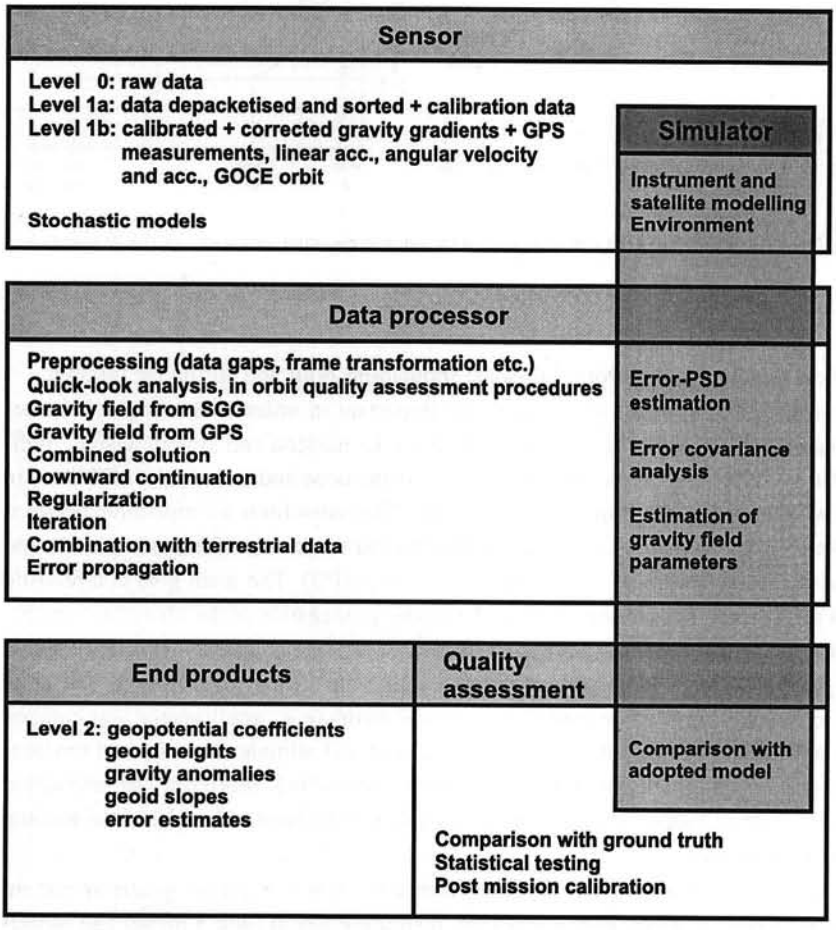

Fig. 10. Scheme of the GOCE data center

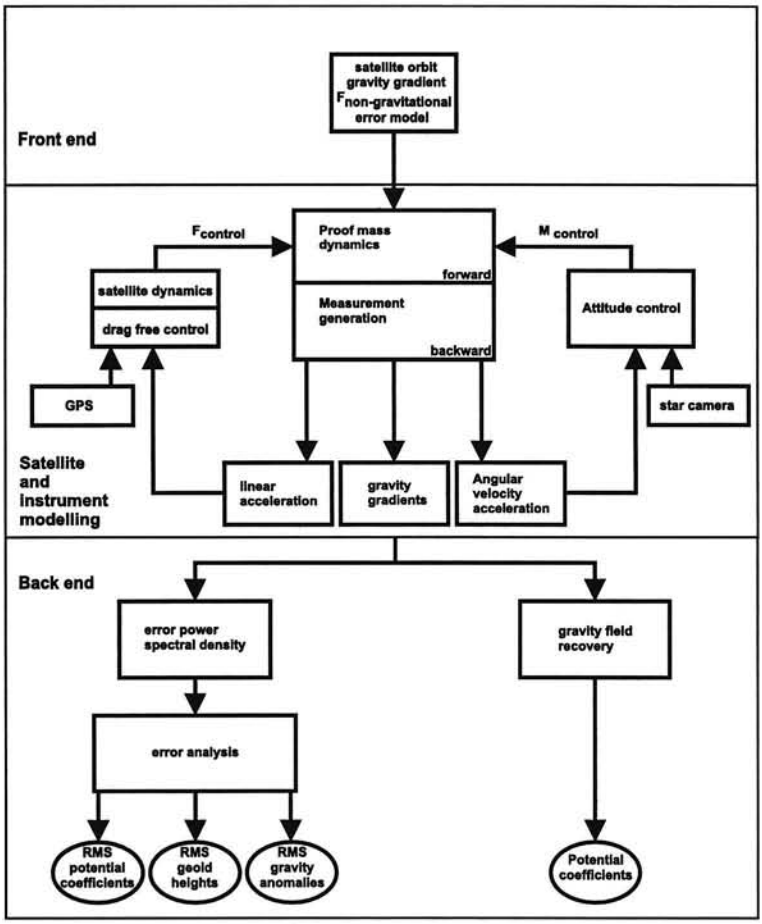

Fig. 11. GOCE end-to-end closed-loop simulator

the attitude control system (ACS), causing a set of cold-gas proportional thrusters to control this attitude motion. However, DFC and ACS cause control forces and moments, respectively, which in turn affect for instance the satellite orbit, the satellite dynamics, and the gradiometer signal. They are fed into the gradiometer forward model, which closes the loop ("closed-loop simulation"). We refer to (Sneeuw et al. 1998) for more details.

The result of such a closed-loop simulation is a time series of output gravity gradients. We compare them with the input gravity gradients, and from the difference, a realistic stochastic model of the gradiometer readouts in terms of error power spectral densities (error PSD's) is computed. Figure 12 shows, e.g., the error power spectral density of the influence of a misaligment between the sensitivity axes of the two component accelerometer that measures V_{yy} in the presence of drag. For comparison, the mission requirements of 5 mE/$\sqrt{\text{Hz}}$ and $1/f$ behaviour below 27 cpr ($5 \cdot 10^{-3}$ Hz) and the error power spectral density of the numerical

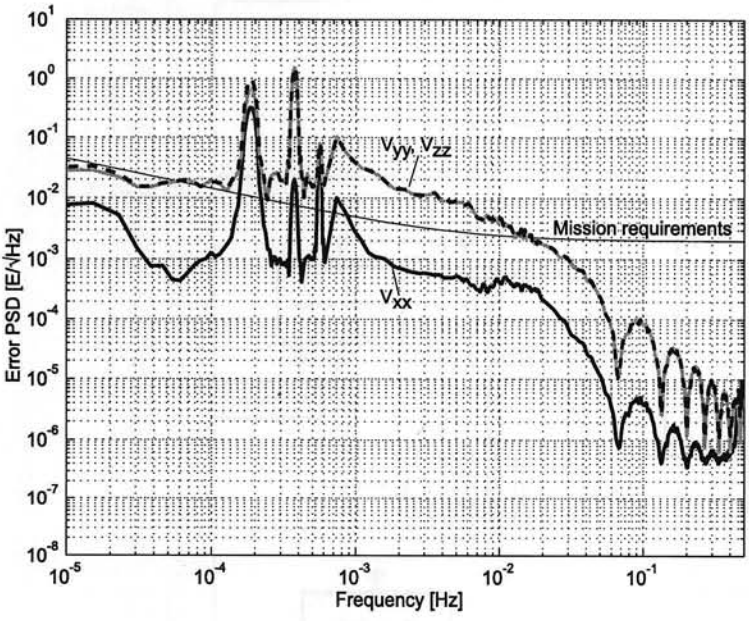

Fig. 12. Effect of misaligment on the gravity gradients

integration errors (solid line) are shown. The latter can be seen as the simulator noise but can be made arbitrarily small by choosing a smaller step size. In a second step we propagate the error PSD into geoid errors or gravity anomaly errors, in order to investigate whether or not the mission goals are met. For instance, figure 13 shows how the alignment error propagates into geoid heights in the presence of drag.

Similar calculations are currently being done for many other errors of the sensors, the control units, the actuators, and other subsystems. This allows to identify weaknesses and limitations of the mission and system design already prior to the mission, and to adapt the design to the mission goals.

As the data processor concerns we are confronted with a number of *theoretical* and *numerical* problems, most of them have not yet been fully solved. The *numerical* problems are caused by the huge number of observations and unknowns to be solved for (figure 14), which are extremely demanding in terms of computer power and storage requirements. They require efficient algorithms and supercomputing facilities. For instance, when using a sophisticated functional model, which allows for instance for real, perturbed orbits, satellite maneuvers, and data gaps, all entries of the normal matrix are non-zero. Currently there is no operational approach for the assembly of the observation equations and the full normal equations. The solution of normal equations itself does not pose any problem from a mathematical point of view: iterative solvers have to be used, e.g., conjugate gradient methods or multigrid techniques. A critical point could be the number of iterations. However, since the normal matrix shows a dominant block-diagonal

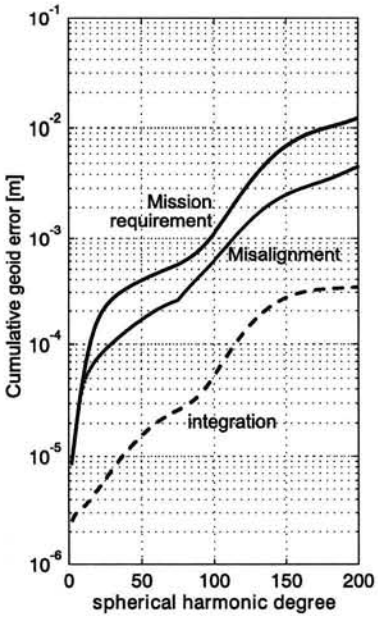

Fig. 13. Propagation of the misalignment error (12) into geoid height errors

structure with some resonance bands, we may exploit this to design efficient preconditioners in order to reduce the number of iterations. This has still to be investigated.

Our current approach is to use a simplified functional model, which only holds if orbit, mission length, maximal resolution, and sampling fulfil certain requirements. To be more specific, we assume that we have an uninterrupted time series of observations available along a circular repeat orbit with a prime number of revolutions in a repeat cycle, where this prime number has to be larger than twice the maximal degree of the potential field. Then, the normal equation matrix has a block-diagonal structure even when coloured noise and/or band-limited stochastic behaviour of the observations is assumed. This allows to solve the normal equations very easily order by order. We assemble the observation vector along the "actual" orbit and take a realistic stochastic behaviour of the measurements (e.g., coloured noise or even band-limitation) properly into account. Then, the strategy is to reduce the influence of model errors on the gravity field parameters by iteration. Questions of convergence and speed of onvergence will become important. This is currently being investigated numerically but should also be proved analytically. We did an experiment in order to investigate whether and how fast the iteration scheme converges. A non-polar, non-circular GOCE-like orbit was simulated along which gravity gradients were computed using a potential model up to degree and order 80. From this time series of gravity gradients the potential coefficients were estimated up to degree and order 80, and the relative differences between input coefficients and estimated coefficients were computed. This proces was then iterated. The results are shown in figure 15. The lowest curve in the figure is the one-step solution for a non-polar, circular orbit. It seems that after a few iterations the

degree n	# unknows	# observations	Storage requirements (GB)	
			Design matrix	normal matrix
70	5037	19600	0.8	0.2
180	32757	129600	34.0	8.4
240	58077	230400	104.5	26.4

Fig. 14. Number of observations and unknowns and storage requirements for estimation of potential coefficients from gravity gradiometry observations. The number of observations has been estimated using the Nyquist sampling theorem; during a real mission the number of observations may be 10 times as large

influence of the non-perfect functional model (in this case the assumption of a circular orbit) on the estimated gravity field parameters is negligible. However, before a definite answer can be given, more numerical experiments have to be done.

From a *theoretical* point of view the major questions are related to the problem of, e.g., non-homogeneous data distribution, downward continuation, combination with terrestrial data sets, and choice of base functions.

A homogeneous data distribution is required in order to provide a homogeneous accuracy and resolution over the entire Earth. However, a simple spacecraft design (power production by means of solar panels, minimal effect of thermal and mechanical noise due to occultation), requires an orbit such that the orbital plane will remain fixed w.r.t. the sun during the mission (a so-called sun-synchronous orbit). This implies that the satellite flies in a slightly non-polar orbit, leaving a data gap of some degrees around the poles, the so-called polar gaps. A number of problems are related to the polar gaps. For instance, in terms of spherical harmonic analysis they may cause leakage of the spherical harmonic spectrum; next the normal equations may become unstable. If they need to be regularized, bias will be introduced. Finally, aliasing will become important since we only estimate spherical harmonics up to a maximum degree.

We investigated various regularization techniques and have put special attention to the bias, the propagated error, and aliasing, see (van Gelderen 1997), (Bouman 1998). Let us take as a simple example the problem of the bias. The left panel of figure 16 shows the bias in the potential coefficients due to polar gaps of 13.2 degree diameter. Obviously the bias is almost concentrated on the low order potential coefficients. The bias is closely related to the loss of power of the spherical harmonics due to the limitation of the domain. Therefore one can try to construct a system of base functions that minimizes the loss of power in the polar areas, cf. (Albertella et al. 1998). Another strategy to reduce the bias is to add terrestrial gravity data in the polar areas. We investigated how dense the sampling must be and what precision the gravity measurements should have in order to remove the bias in the low order coefficients, cf. (Bouman & Koop 1998). We have assumed a quality and coverage that can be achieved by available observation techniques such as airborne gravimetry. The panel on the right of fig-

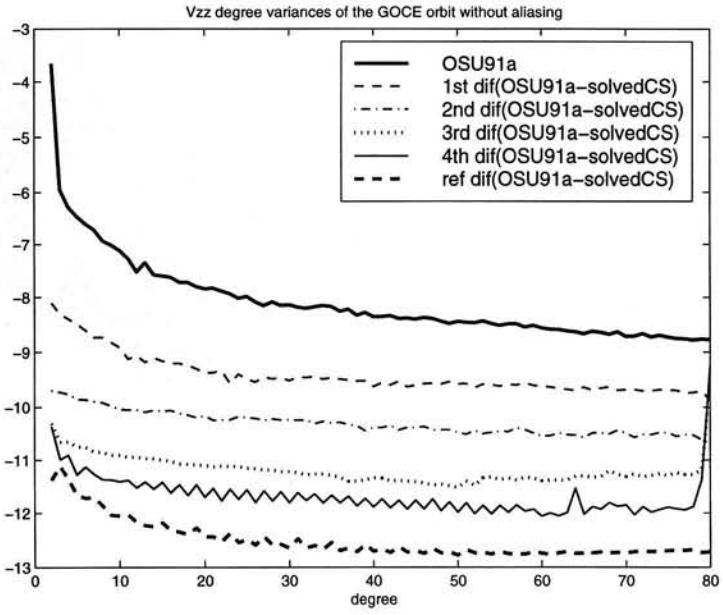

Fig. 15. Semi-analytical approach to potential coefficient estimation from gravity gradiometry observations

ure 16 indicates that for moderate spatial density and quality of the terrestrial gravity data the bias does almost vanish.

Assuming that a high degree geopotential model has been estimated from gravity gradiometry and SST data, we have to compute functionals of the geopotential, usually at the Earth's surface. Then, another problem arises, due to the downward continuation of the geopotential from satellite altitude to the Earth's surface. If the geopotential would be exactly known at satellite altitude it would be known everywhere outside the Earth's masses due to its harmonicity. However, since the estimated geopotential coefficients at satellite altitude are erroneous, errors in the coefficients of degree n will be amplified by a factor of $1.0424^{(n+1)}$ if geoid heights have to be evaluated at the Earth's surface. This amplification is worst for the shortest wavelengths (i.e., for large n up to maximum degree $n \approx 240$) and less pronounced for the long wavelengths. For instance, errors in the high degree potential coefficients will be amplified by more than 3 orders of magnitude. Therefore, downward continuation may lead to weak solutions at the Earth's surface, requiring some regularization in order to control the errors.

Finally, the integration with terrestrial data sets is an important issue. Terrestrial data sets provide the very short wavelengths above a resolution of degree $200 - 240$ as expected from satellite gradiometry and satellite-to-satellite tracking. In principle this can be done by a classical least-squares approach taking the stochastic properties of the measurements and of the estimated gravity field parameters properly into account. Alternatively to this discrete approach

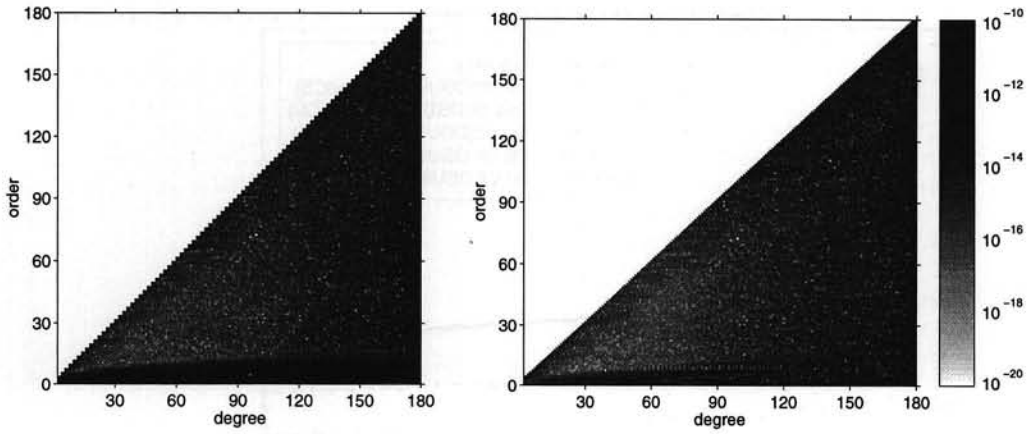

Fig. 16. Bias in potential coefficients due to polar gaps with (right panel) and without (left panel) terrestrial gravity anomalies

we may formulate the problem of improving the gravity field from terrestrial observations as a boundary value problem with a priori constraints. When using integral equation techniques, we are faced with the problem of how to construct efficiently a suitable system of base functions that fulfil the constraints, and, at the same time, are stable. First investigations have shown that multiscale bases with compactly supported base functions fulfil these requirements. However, when applied to the geodetic situation, we are confronted with some conceptual problems of wavelet basis functions due to the lack of smoothness of the Earth's topography over the support of the coarse scale wavelet base functions, see (Klees & Lehmann 1998).

Acknowledgement We would like to thank Axel Smits for making the figures, and Johannes Bouman, Martin van Gelderen, José v/d IJssel, and Pieter Visser for their support.

References

Bouman, J. (1998): *Quality of regularization methods. DEOS Report*, no. 98.2, Delft University Press, The Netherlands, pp 104.

Bouman, J., R. Koop (1998): Quality improvement of global gravity field models by combining satellite gradiometry and airborne gravimetry. Paper presented at the IV Hotine-Marussi Symposium, 14-17 September 1998, Trento, Italy.

Gelderen, M. van (1997): Error propagation for satellite gradiometry *DEOS Progress Letters*, no. 97.1, Delft Institute for Earth-Oriented Space Research, Delft, The Netherlands, 33–41.

Klees, R., R. Lehmann (1998): Integration of a priori gravity field models in boundary element formulations to geodetic boundary value problems. Paper presented at the IV Hotine-Marussi Symposium, 14-17 September 1998, Trento, Italy.

Morgan, S.H., H.J. Paik (1988): *Superconducting gravity gradiometry mission*. NASA Technical Memorandum 4091, Vol. II: Study Team Technical Report.

Albertella, A., N. Sneeuw, F. Sansó (1998): The Slepian problem on the sphere Paper presented at the IV Hotine-Marussi Symposium, 14-17 September 1998, Trento, Italy.

Sneeuw, N., Chr. Gerlach, J. Mueller, H. Oberndorfer, R. Rummel, R. Koop, P. Visser, P. Hoyng, A. Selig, M. Smit

(1998): Simulation of the GOCE gravity field mission Paper presented at the IV Hotine-Marussi Symposium, 14-17 September 1998, Trento, Italy.

Snieder, R. (1998): The deep Earth: bridging the gap from kinematics to dynamics. Paper presented at the symposium on "Views on the Dynamic Earth", Vening Meinesz Research School of Geodynamics (VMSG), Utrecht, The Netherlands, 20 November 1998.

Wahr, J. (1998): Time variable gravity from the GRACE satellite mission: what will it tell us about the Earth?. Paper presented at the symposium on "Views on the Dynamic Earth", Vening Meinesz Research School of Geodynamics (VMSG), Utrecht, The Netherlands, 20 November 1998.

Wortel, R. (1998): Surface-depth relations of the crust-lithosphere system. Paper presented at the symposium on "Views on the Dynamic Earth", Vening Meinesz Research School of Geodynamics (VMSG), Utrecht, The Netherlands, 20 November 1998.

DEOS Progress Letter 98.1: 89–99

A procedure for combining gravimetric geoid models and independent geoid data, with an example in the North Sea region

Roger Haagmans, Arnoud de Bruijne and Erik de Min[1]

[1] Survey Department of Rijkswaterstaat, Delft, The Netherlands

Abstract

The main goal of the study is to obtain a consistent height and depth reference system in the form of a geoid for the Dutch mainland and marine area. For this purpose a procedure has been developed and tested to combine available gravity data and external data from satellite altimetry, GPS and levelling in an optimal manner. After all, local gravimetric geoid models from a combination of a global geopotential model and local gravity anomalies, usually contain errors of dm-level on wavelengths longer than 50 km. One of the main causes for this is the limited precision of the global models. External geoid information on discrete points, like GPS and levelling sites on land and altimeter tracks combined with permanent sea surface topography at sea is often available with cm-precision. These points can be used to correct for the medium and longer wavelength errors in the gravimetric geoid. The problem is to find an adequate functional representation of the correction surface. The authors have developed a method to investigate the form of this correction, and find empirical representations depending on area size. Once a class of functions have been selected the most suitable can be found from a statistical testing procedure.

The initial purely gravimetric geoid is adjusted in longer wavelengths by means of the external geoid data. The preliminary North Sea geoid GEONZ97 has a precision of better than 4 cm at sea and along the Dutch coast. As soon as instantaneous GPS height components in off-shore applications reach a comparable accuracy, tides and meteorological response can be eliminated in an efficient and effective way simply by subtracting the geoid.

1 Introduction

The main objective of this study (De Bruijne *et al.* (1997)) was to develop and implement a procedure for determining a consistent height and depth reference system for the Netherlands in the form of a unified land and marine geoid. This geoid is mainly needed for coastal zone management by Rijkswaterstaat. For instance for maintenance of important sea lanes, or maintenance of coastal areas depending on sediment transport related to (surface) currents. An accurate geoid and permanent sea surface topography model (PST) are necessary tools to achieve this.

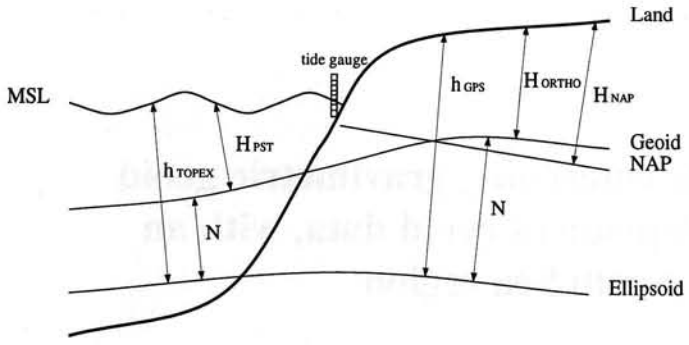

Fig. 1. The land and sea geoid concept.

It also allows to convert depths related to mean sea level (MSL) into depths related to Lowest Astronomical Tide for use in nautical charts.

The setting for the geoid is:

$$h - H = N, \tag{1}$$

where h is the measured geometrical height, H is the known orthometric height, and N is the unknown geoid height that can be approximated with a gravimetric one. Specifically, at sea H_{PST} and h_{TOPEX}, and on land H_{ORTHO} from levelling and h_{GPS} are needed. In figure 1, the relations between all components are visible together with MSL and Normaal Amsterdams Peil (NAP). The hypothesis used in the preliminary setup is that the PST is negligibly small compared to the required accuracy of 10 cm. This level can hopefully be achieved for instantaneous GPS heights off-shore in the near future. The neglection implies the MSL from altimetry to coincide with the marine geoid, and consequently the equipotential for height reference to coincide with current MSL at tide gauges. The land height datum NAP refers to the MSL at the time of definition, and in a strict sense the NAP heights do not refer to an equipotential anymore, due to land subsidence and sea level change. Therefore, the NAP heights are transformed to orthometric heights referring to an equipotential at current MSL in the tide gauges. At sea the spatial resolution and quality of the altimeter measurements is not homogeneous. Therefore, only good quality TOPEX altimetry is used as external geoid data, rather than a direct altimetric result. This allows to optimally exploit details in the gravimetric geoid.

2 Procedure

The procedure we followed for determining the North Sea geoid can be summarized as follows, cf. De Bruijne *et al.* (1997).

- Compute detailed gravimetric geoid N_g from gravity data
- Determine possible empirical models to correct the expected error in the gravimetric geoid
- Include external altimetry, GPS and levelling data
- Correct the gravimetric geoid N_g with the external data, yielding MSL

The following sections explain this procedure. The second item is dealt with extensively in section 4.

3 Gravimetric computation

For the determination of the gravimetric geoid N_g, we refer to De Min (1996). Here, we suffice with the main formula

$$N_g = \frac{GM}{\gamma r}\sum_{n=2}^{360}\left(\frac{a}{r}\right)^n \lambda_n w_n \sum_{m=0}^{n}\Delta C_{nm}Y_{nm}(\varphi,\lambda) + \frac{R}{4\pi\gamma}\int_{\sigma_0}\{\sum_{n=2}^{\infty}\frac{2n+1}{n-1}(1-\lambda_n w_n)P_n(\cos\Psi)\}\Delta g\, d\sigma, \quad (2)$$

where G is the gravitational constant, M the mass of the earth, γ the normal gravity, r the radius, a the semi-major axis, λ_n the eigenvalue for the specific gravity field quantity, w_n the spectral weights, ΔC_{nm} the EGM96 global geopotential coefficients, Y_{nm} the spherical harmonics of degree n and order m at latitude φ and longitude λ, R the mean radius of the earth, σ_0 the integration area up to a given capsize, P_n the Legendre polynomials, Ψ the spherical distance and Δg the local gravity data. The first part of equation (2) is a weighed spherical harmonic expansion and the second part involves numerical Stokes' integration up to a capsize, based on a weighed difference between local gravity data and EGM96 (Δg), in order to resolve all details. The weights w_n that take care of an optimal combination of the longer and shorter wavelength contents, belong to the Meissl/Wong&Gore (MWG) modification, cf. Heck and Grüninger (1987), De Min (1996), De Bruijne *et al.* (1997) and section 4. The gravity data involved are:

- the EGM96 global geopotential model referring to GRS80 (Lemoine *et al.* (1996) and Rapp (1996));
- 6'x10' block mean free-air gravity values in Europe (Weber (1984)), and 3'x5' block mean free-air gravity values in parts of the computation area, predicted from various data sets (De Min (1996)).

4 Correcting the error in the gravimetric geoid

4.1 Background

In gravimetric geoid computation procedures, local gravity data are usually combined with a global geopotential model (GGM). In principle an optimal combination based upon realistic error characteristics of the globally available gravity data and a GGM should lead to the best possible gravimetric geoid result, however this is limited due to systematic errors at regional scales from terrain reductions, height datums, etc., cf. Pavlis (1988). In regional computations, however, dense gravity data are only available in a restricted area. This spatial limitation and the fact that one likes to exploit fully the advantages of the local gravity data and the global model lead to a practical optimal choice in a combined solution, compare also discussions in Haagmans and Van Gelderen (1991). The advantage of the local gravity is that it provides all details at small and medium scales in the geoid solution. The advantage of the global models is that the long wavelength solution is best resolved from satellite data instead of gravity data. In case of a specific weighing between local gravity data and a global model (see figure 3), the

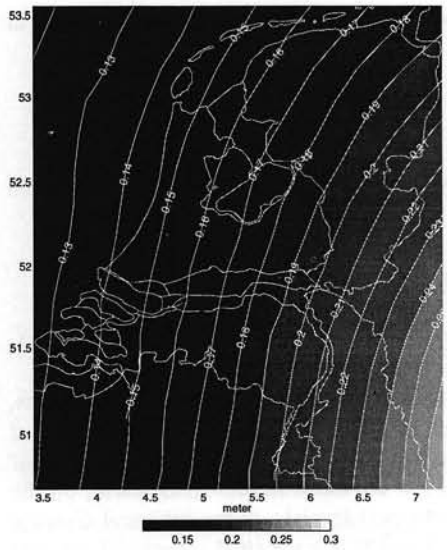

Fig. 2. Difference between total gravimetric geoid in the Netherlands based on OSU91 and EGM96.

differences in the total gravimetric geoid computations based upon OSU91 and EGM96 for the Netherlands may differ as shown in figure 2. From figure 2 we can observe that the difference only exhibits a longer wavelength pattern and no smaller scale details which is of practical importance for GPS and levelling applications. Accepting the fact that regional gravimetric geoid solutions may always be contaminated with errors at medium scales (cf. also Sideris and Li (1992)) we try to find a procedure to combine the gravimetric geoid with external geoid data from GPS (h) and levelling (H) on land and altimetry (h) with a permanent sea surface topography model (H) at sea, all with proper quality measures. The problem is to find an adequate functional description for correction surface F_c in equation (3).

$$h - H = N = N_g + F_c. \tag{3}$$

Generally, these are chosen to be trend functions of bi-linear type or similar, cf. Sideris and She (1995) and Forsberg *et al.* (1997). Usually, a profound reasoning for choosing such a specific function is missing. Therefore, we tried to find a procedure that can be applied for all gravimetric geoid results and that is applicable and adaptive for areas of different size.

4.2 Determination of the correction surface

The procedure we followed can be divided into several steps:

- Generate likely GGM geoid error surfaces
- Find adequate empirical representations for the surface out of a set of functions
- Fit the empirical function to the residuals between the gravimetric and external geoid
- Apply a statistical test procedure for finding the best representation

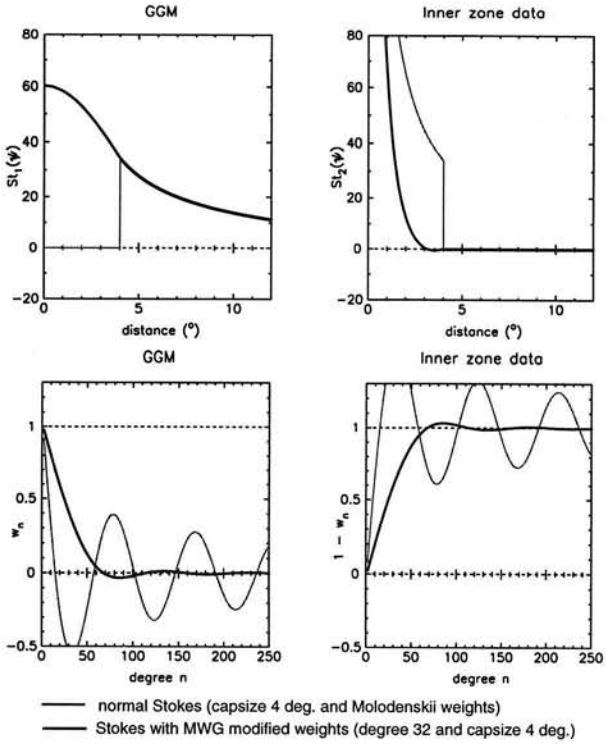

Fig. 3. Normal Stokes weights (gray) and MWG weights (black).

The possible shape of the correction surface will be analyzed based upon the chosen GGM with its formal error description and the assigned weighing in the procedure. First, several possible sets of error coefficients per degree n and order m (E_{nm}) are generated from the formal standard deviations of the coefficients of EGM96, assuming the errors per coefficient to be normally distributed. For each set, error geoid surfaces can be obtained from the error coefficients weighed with w_n, as shown in equation (4), cf. De Min (1996):

$$N_e = \frac{GM}{\gamma r} \sum_{n=2}^{360} \left(\frac{a}{r}\right)^n \lambda_n w_n \sum_{m=0}^{n} E_{nm} Y_{nm}(\varphi, \lambda). \tag{4}$$

The weights w_n can be in an idealized case Shannon weights, being 1 up to 360 and 0 for higher degrees, or Molodenskii weights in case of spatial truncation of Stokes' function, or weights according to the MWG (Meissl/Wong&Gore) kernel modification; the latter two cases are shown in figure 3. An example of degree variances based on generated error coefficients for EGM96 is shown in the right part of figure 4, together with the MWG weights for coefficients. In the left part of figure 4 the signal and error degree variances of OSU91 and EGM96, and the difference between OSU91 and EGM96 are shown for comparison. For the North Sea area 10 error surfaces were randomly generated according to equation (4) with MWG weights for degree parameter 32 and spherical capsize of 4°. The surfaces show a range of 16–23 cm and an rms of 3.6–6.4 cm, cf. figure 5.

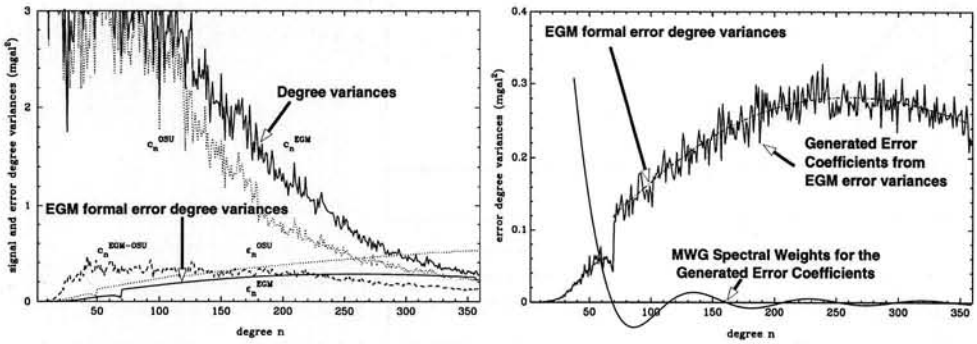

Fig. 4. Generation of random error coefficients based on EGM96 including the MWG weights (right) based on the formal coefficient error variances (left).

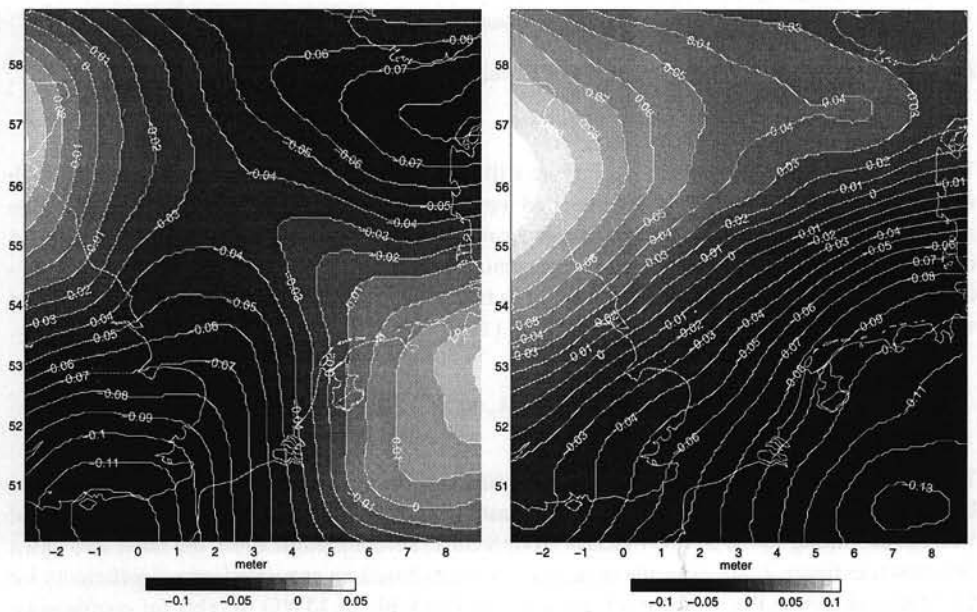

Fig. 5. Examples of randomly generated geoid error surfaces based on EGM96 and MWG modification.

The next step is to select a class of functions for empirical modelling of these surfaces. Generally, this can be e.g. polynomials, wavelets, harmonic base functions depending on the surface characteristics and the area extent. For the North Sea a bi-linear trend function and trigonometric functions are selected, based on a Fourier analysis, which are symbolically represented in equations (5) and (6):

$$a_{00} + b_{00}\lambda_l + c_{00}\varphi_k + d_{00}\varphi_k\lambda_l, \text{ and} \tag{5}$$

$$\sum_{i=1}^{I}\sum_{j=1}^{J} a_{ij}\cos(i\lambda_l)\cos(j\varphi_k) + b_{ij}\sin(i\lambda_l)\cos(j\varphi_k) + c_{ij}\cos(i\lambda_l)\sin(j\varphi_k) + d_{ij}\sin(i\lambda_l)\sin(j\varphi_k)\,; \tag{6}$$

λ_l and φ_k indicate longitude and latitude increments relative to a chosen origin in the area. From the Fourier analysis of the 10 surfaces it appeared that the maximum limit for I and J is 2. In 60% a 12 parameter model and in 40% a 28 parameter model was necessary for reducing the unmodelled negligible residual below a 1 cm rms. Examination of figure 4 reveals that the error estimates for EGM96 may be too optimistic by a factor of 2–3 in the range between degree 2–70 from comparison with OSU91. Thus, the previous results need to be scaled to a 2–3 cm unmodelled residual, which is in the range of the precision of the external geoid data, so that no extension of the correction model is necessary. N.B. it is in principle possible to extend the model with more bias parameters in case land data from different height datums are involved.

The final step is to fit the parameters to the residuals $N - N_g$ of equation (3) in a least squares adjustment, with an overall model test, and iterative data snooping. The model can be extended and tested against others in order to select the optimal one within the class of functions, following the principles developed for deformation analysis (De Heus *et al.* (1995)). Careful analysis of the geoid error surface and suited correction functions limits the number of possible and acceptable correction surface parameters. This procedure has been successfully applied for the computation of the preliminary North Sea geoid GEONZ97, De Bruijne *et al.* (1997).

5 Connecting the gravimetric geoid with external data

In De Min (1996), De Bruijne *et al.* (1997) and section 4 it is described that the gravimetric geoid N_g has to be corrected for its longer wavelengths, based on external data:

$$N = N_g + F_c. \tag{7}$$

The correction function F_c depends on the error in the gravimetric geoid and the control data (altimetry, GPS and levelling). Tests resulted in the choice of a sum of a bilinear and a trigonometric surface, expressed in 12 or at most 28 parameters, cf. subsection 4.2.

Figure 6 (left side) contains the residuals with an rms of 8.5 cm between the external geoid and the gravimetric geoid values, after removal of the mean of 69 cm with respect to two year averaged TOPEX and GPS/levelling heights. For details, see De Bruijne *et al.* (1997). A least squares adjustment procedure including iterative data-snooping resulted in the acceptance of the 28 parameter model in view of the required accuracy. Figure 6 (right side) shows the residuals after the correction; the rms of all accepted points is 3.2 cm, and the mean of all points is -0.1 cm and the rms 4.2 cm.

The 28 parameter correction surface is shown on the left side of figure 7. Combining the gravimetric geoid and the correction surface yields the preliminary North Sea geoid GEONZ97 shown on the right side of figure 7.

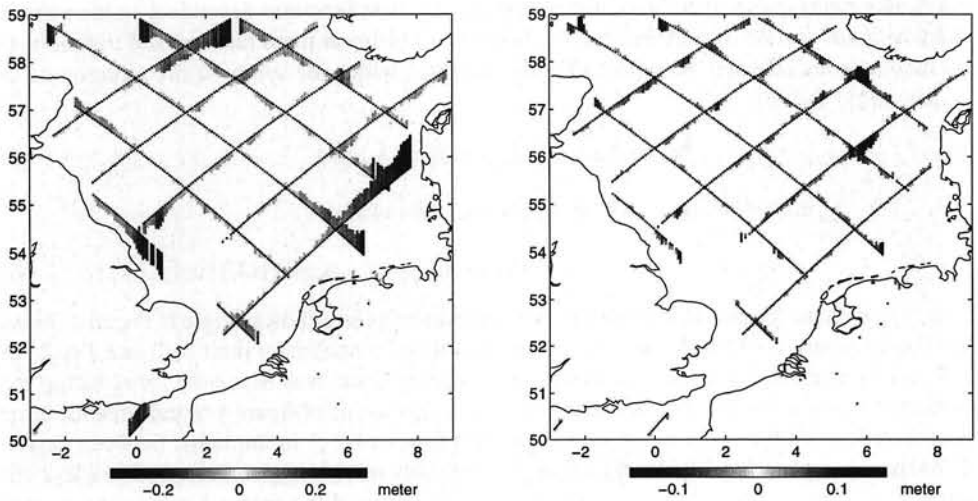

Fig. 6. External data minus gravimetric (left) and corrected (right) geoid at external data locations.

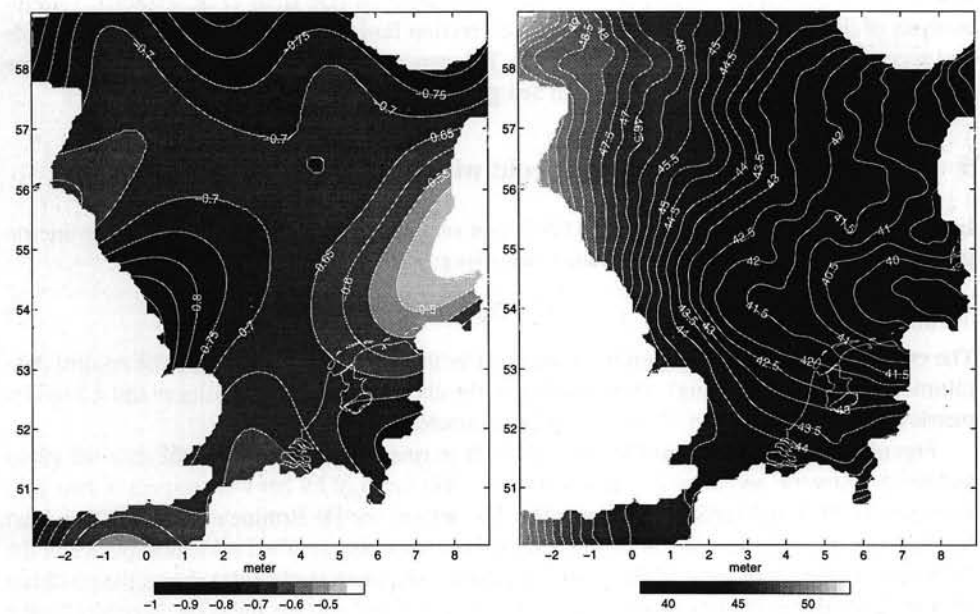

Fig. 7. Correction surface for the gravimetric geoid (left) and preliminary geoid GEONZ97 (right).

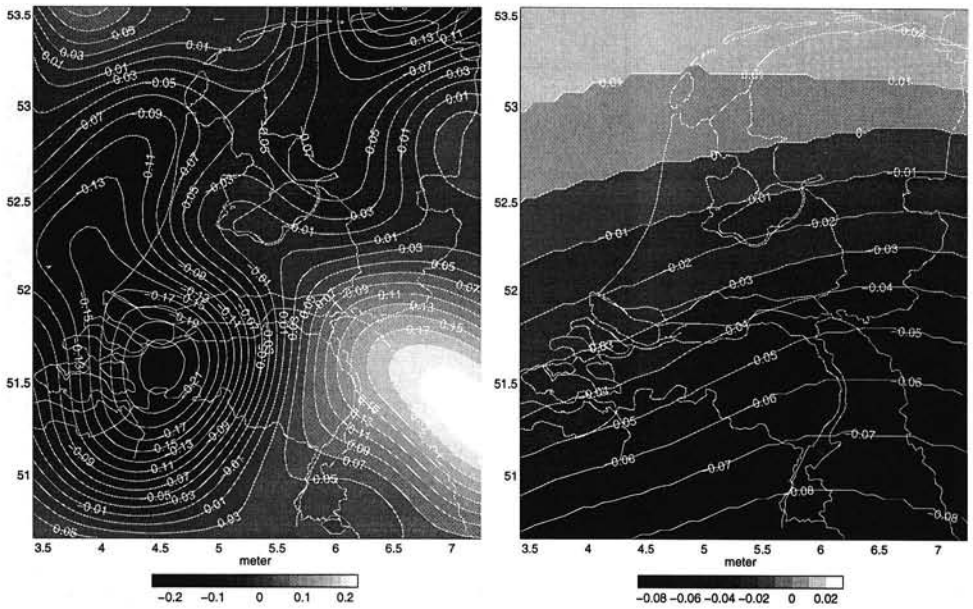

Fig. 8. Randomly generated geoid error surfaces based on EGM96 with Molodenskii weights (left) and with MWG modification (right).

6 Conclusions

A procedure is proposed for correcting the longer wavelength errors in the gravimetric geoid by means of an adequately chosen empirical function based upon geoid error surfaces generated from the formal errors of a GGM. It depends mainly on the weighing between local data and a global model. A standard approach with Molodenskii weights results for the Netherlands in a rather irregular geoid error surface (see figure 8 (left)), that is rather complex to model. The MWG modification shows a smooth trend surface (see figure 8 (right)). Modelling this by means of external geoid data results in the elimination of the trend surface, but also of the difference between two gravimetric geoid solutions as shown in figure 2: the final geoids will be practically identical. Thus a proper weighing or kernel modification is important. The procedure can easily be extended to larger areas, avoiding unnatural blending of neighbouring solutions.

This procedure has been successfully applied to determine - as a first attempt - a consistent geoid for land and sea in the Dutch region, within a ± 10 cm level at sea and a few cm's on land. However, some aspects can be improved in future computations. The statistical testing procedures lead to the choice of a 28 parameter model, but the overall model test was not fully accepted. Further extension of the model possibly leads to a better fit, but is not expected to be realistic. One of the reasons can be that the assumption of H_{PST} to be small and of random nature may be invalid, since a model based upon the major tidal component and wind predicts a PST with a trend pattern as shown in figure 9, cf. Prandle (1984). Comparing the left plots of figure 7 and figure 9 leads to the suggestion that the correction surface absorbs part of the systematic effect of the PST. So, inclusion of a state of the art PST model may very well lead

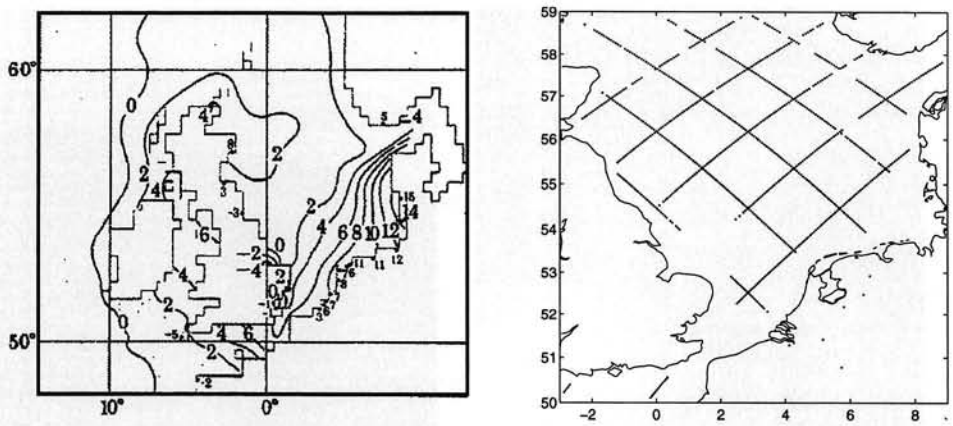

Fig. 9. Model PST in cm (left) and accepted points in the final adjustment (right).

to acceptance of the 12 parameter model and improvements near the coast. From the right plot of figure 6 and figure 9 and the original data distribution, it appears that data rejection took place at locations where residuals are large due to poor marine gravity coverage, and poor tidal modelling of altimetry close to the coast. Improving these aspects and inclusion of GPS and levelling of more countries can lead to an accurate unified geoid as a height and depth reference for the whole region.

Acknowledgement The authors wish to thank all colleagues at Bureau Gravimétrique International, the British Geological Survey, Ohio State University, NIMA, NASA, CNES, University of Hannover, Shell, NAM, and the Survey Department of Rijkswaterstaat for providing data sets and models.

References

de Bruijne, A. J. T., Haagmans, R. H. N., and de Min, E. J. (1997). A preliminary north sea geoid model GEONZ97. Technical report, Survey Department of Rijkswaterstaat, Delft, The Netherlands. MDGAP-9735.

Forsberg, R., Kaminskis, J., and Solheim, D. (1997). Geoid of nordic and baltic region from gravimetry and satellite altimetry. In *Gravity, Geoid and Marine Geodesy*, number 117 in IAG, pages 540–547, Berlin. Springer.

Haagmans, R. H. N. and van Gelderen, M. (1991). Error variances-covariances of GEM-T1: their characteristics and implications in geoid computation. *Journal of Geophysical Research*, **96**(B12), 20011–20022.

Heck, B. and Grüninger, W. (1987). Modification of stokes' integral formula by combining two classical approaches. In *XIX IUGG General Assembly*, volume 2, pages 319–337. IAG.

de Heus, H., Joosten, P., Martens, M., and Verhoef, H. (1995). Strategy for the analysis of the groninger gasfield levellings: an overview. In *Land subsidence*, pages 301–311. Balkema, Rotterdam.

Lemoine, F. G., Smith, D. E., Kunz, L., Smith, R., Pavlis, E. C., Pavlis, N. K., Klosko, S. M., Chinn, D. S., Torrence, M. H., Williamson, R. G., Cox, C. M., Rachlin, K. E., Wang, Y. M., Kenyon, S. C., Salman, R., Trimmer, R., Rapp, R. H., and Nerem, R. S. (1996). The development of the NASA GSFC and NIMA joint geopotential model. Proceedings paper for the International Symposium on Gravity, Geoid, and Marine Geodesy (GRAGEOMAR 1996), The University of Tokyo, Tokyo, Japan.

de Min, E. J. (1996). *De geoïde voor Nederland.* Ph.D. thesis, Technische Universiteit Delft, Delft. (in Dutch).

Pavlis, N. (1988). Modelling and estimation of a low degree geopotential model from terrestrial gravity data. Technical Report 386, Department of Geodetic Science and Surveying, OSU, Columbus, Ohio.

Prandle, D. (1984). A modelling study of the mixing of ^{137}cs in the seas of the european continental shelf. *Philosophical transactions of the Royal Society of London; physical sciences and engineering*, **310**(A), 407–436.

Rapp, R. H. (1996). Global models for the 1 cm geoid; present status and near term prospects. Prepared for international summer school of theoretical geodesy: boundary value problems and the modelling of the earth's gravity field in view of the one centimeter geoid. Como, Italy.

Sideris, M. and She, B. (1995). A new, high-resolution geoid for canada and part of the u.s. by the 1D-FFT method. *Bulletin Géodésique*, **69**(2), 92–108.

Sideris, M. G. and Li, Y. (1992). Improved geoid determination for levelling by GPS. In *Sixth International Geodetic Symposium on Satellite Positioning*, volume 2, pages 873–882, Columbus, Ohio.

Weber, G. (1984). *Hochauflösende mittlere Freiluftanomalien und gravimetrische Lotabweichungen für Europa*. Ph.D. thesis, Universität Hannover, Hannover. Published as volume 135 in Wissenschaftliche Arbeiten der Fachrichtung Vermessungswesen der Universität Hannover (in German).

DEOS Progress Letter 98.1:101-122

A Strategy for Geoid Determination in the Indonesian Archipelago

Kosasih Prijatna[1]

Faculty of Civil Engineering and Geosciences, Delft University of Technology
Thijsseweg 11, NL-2629 JA, Delft, The Netherlands

Abstract

The long term goal is to determine an accurate geoid for the Indonesian region. The most important step towards this goal is a thorough analysis of availability of data in this region, and the search for a proper method for geoid determination adapted to the specific situation. Therefore, an analysis is done of available terrestrial and satellite data for both land and marine areas. Based upon these data sets and expected extensions, geoid determination scenarios are proposed for future computation of a unified geoid for the Indonesian Archipelago.

1 Introduction

A geoid of the Indonesian region was derived by Kahar with a precision of about 4 meter. The main purpose of geoid determination at that time was the need to know of the geometrical relationship between earth surface and reference ellipsoid for geodetic computations, Kahar (1981). Recently, specific geodetic, oceanographic, and geophysical applications demand a more precise, *dm*- or even *cm*-level, high resolution geoid in the region as a reference surface. In geodesy, the prospect of establishing a highly accurate geoid for Indonesia can be found in practice when the costs of levelling can be reduced tremendously by using GPS in combination with the geoid. In this case, the subject of geoid determination as studied by Kahar needs to be reconsidered.

According to Khafid (1997), a precise geoid computation in the Indonesian region is influenced by the following facts :

- Physical terrain or bathymetry characteristics, e.g. mountainous terrain, complex tectonics and archipelago-type geography.
- Establishment of a high resolution mean free air gravity anomaly data base covering the entire area of Indonesia and of its surroundings.
- Need for a Digital Terrain Model (DTM) in order to correct for the terrain effects.
- Unified national vertical datum, such a reference does not exist.

---

[1] On leave from Department of Geodetic Engineering, Institut Teknologi Bandung, Jl. Ganesa 10, Bandung-40132, Indonesia

- Insight into the oceanographic and tidal setting in the Indonesian waters for tide gauges and satellite altimetry.

Improvements have been achieved in the methodology and measurement techniques. For the ocean part the accuracy of satellite altimetry improved dramatically so that it can be used perfectly in combination with marine gravity data. For the land part new terrestrial gravity data has been collected, in combination with the high precision Global Positioning System (GPS). Besides, improvements in high resolution topographic mapping are obtained with the airborne or satellite technique Synthetic Aperture Radar (SAR). Also theoretical developments in geoid computation evolved and practical experiences were gained in the geoid computation for the Netherlands, De Min (1996b), and for the North Sea, De Bruijne et al. (1997). Those two recent studies investigate the proper procedure for precise geoid computation in both land and sea areas. Within this report, we try to apply procedures following the approach which was used by those two researches for geoid determination in the Indonesian archipelago adapted to its specific situation.

2 Stokes' Approach in Local Geoid Determination

The geoid determination of the Netherlands and the North Sea by De Min (1996b) and De Bruijne et al. (1997) were carried out based upon the well known Stokes' approach. This section reviews the method of geoid determination by those two researches, and also some additional aspects which are important to be considered as they are related to the Indonesian region situation.

2.1 Stokes' Approach

The Stokes' solution is to determine the disturbing potential T that satisfies Laplace's equation:

$$\frac{\partial^2 T}{\partial x^2}+\frac{\partial^2 T}{\partial y^2}+\frac{\partial^2 T}{\partial z^2}=0, \tag{1}$$

and also fulfills a linearized Stokes' boundary value problem in spherical and constant radius approximation :

$$\Delta g=-\frac{\partial T}{\partial R}-\frac{2}{R}T, \tag{2}$$

with R denoting the mean radius of the earth, and Δg the difference between the reduced actual gravity on the geoid and normal gravity on the ellipsoid. The Δg is called gravity anomaly.

According to Stokes, solution of the boundary value problem is:

$$T=\frac{R}{4\pi}\iint_{\sigma}\Delta g\, S(\Psi)\, \mathrm{d}\sigma, \tag{3}$$

where $S(\Psi)$ is Stokes' function, Ψ is the angular distance between gravity anomaly data point and point of computation, and $\mathrm{d}\sigma$ is the surface element of a unit sphere. A closed formula of the Stokes' function is Heiskanen & Moritz (1967):

$$S(\Psi)=\sum_{n=2}^{\infty}\frac{2n+1}{n-1}P_n(\cos\Psi)$$
$$=1+\frac{1}{\sin\frac{1}{2}\Psi}-6\sin\frac{1}{2}\Psi-5\cos\Psi-$$
$$-3\cos\Psi\ln(\sin\frac{1}{2}\Psi+\sin^2\frac{1}{2}\Psi), \qquad (4)$$

where $P_n(\cos\Psi)$ is known as Legendre's polynomial. By applying Brun's formula, the geoid height N above a reference ellipsoid can be determined as follows :

$$N=\frac{R}{4\pi\gamma}\iint_{\sigma}\Delta g\,S(\Psi)\,\mathrm{d}\sigma, \qquad (5)$$

where γ is mean value of gravity.

Both equations (3) and (5) assume that Ilk (1996):

- the reference potential on the ellipsoid is equal to the gravity potential on the geoid,
- the mass of reference ellipsoid is equal to the true mass of the earth,
- the ellipsoid's centre coincides with the earth's centre of mass.

In a more general form, instead of equations (3) and (5), we may write :

$$N=N_0+\frac{R}{4\pi\gamma}\iint_{\sigma}\Delta g\,S(\Psi)\,\mathrm{d}\sigma, \qquad (6)$$

where

$$N_0=\frac{\delta GM}{R\gamma}-\frac{\Delta W_0}{\gamma} \qquad (7)$$

Here, δGM and ΔW_0 are introduced. They are the unknown difference between the value of gravitational constant GM of the actual earth and its value of the adopted reference ellipsoid, and the difference between the potential of the geoid and the potential of the reference ellipsoid, respectively. The N_0, which is a constant, may be neglected for the computation of a local relative geoid, De Min (1996b).

From equation (5), N can be evaluated if the gravity anomaly function Δg :

- refers to geoid surface, and no masses outside the geoid,
- represents a continuous Δg-field,
- covers the whole earth surface.

But in practice, we have a different situation :

- gravity anomalies are measured discretely,
- the measured gravity anomalies refer to the actual earth surface,
- gravity anomalies are only available within limited coverage.

In order for the Stokes' integral to be evaluated numerically to compute N from the observed gravity anomalies,

- the gravity anomaly data has to be reduced to geoid surface,
- the discrete gravity anomaly data has to be assigned to represent surface element gravity anomaly values,
- a combination solution has to be applied by means of combining geopotential coefficients, i.e. global geopotential model and gravity anomaly data.

2.2 Data Preparation: Gravity Reduction and Representation

2.2.1 Gravity Reduction

In principle, we are interested in how to bring the gravity value from the actual earth surface to the geoid. On the other hand, we also have to consider the Stokes' formula requirement that there are no masses outside the geoid (topography and atmosphere). The masses can be removed or condensed to the geoid, depending on the adopted approach. To perform this step, the density function of the masses must be known. Therefore, hypothesis of the density structure of the masses outside the geoid is necessary.

There are many ways to do such a gravity reduction in geoid computation, for example: Bouguer reduction, Helmert's second condensation, and Rudzki's methods. Each method is different, depending on how the topographic masses above the geoid are treated. Theoretically, all gravity reduction methods will result in an identical geoid height provided that they are applied properly.

Except for Rudzki's method, the removal or shifting of masses in gravity reduction will change gravity potential, and hence the geoid, Heiskanen & Moritz (1967). The equipotential surface derived from Stokes' formula is called *cogeoid*, not the geoid (see Fig. 1).

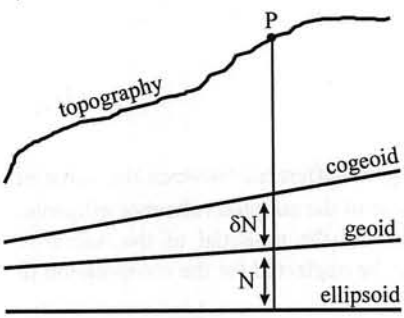

Fig. 1. Indirect effect on geoid δN and cogeoid

It is important to underline that every gravity reduction will result in a different cogeoid. In order to produce a geoid, the observed gravity anomaly Δg_0 must be reduced to the geoid by adding the reduction term δA; followed by a transformation to the cogeoid after applying a small correction $\delta\Delta g$ called the indirect effect on gravity. Furthermore, the computed cogeoid height is also corrected by a term δN called indirect effect on the geoid. Now, the expression of equation (5) can be modified as follows, Wichiencharoen (1982) and Sideris (1994):

$$N = \frac{R}{4\pi\gamma} \iint_{\sigma} (\Delta g_0 + \delta A + \delta \Delta g)\, S(\Psi)\, d\sigma + \delta N \qquad (8)$$

In precise geoid computation, both indirect effects on gravity and geoid should also be taken into account, Wichiencharoen (1982).

2.2.2 Gravity Anomaly Representation

As mentioned before, equation (5) can be applied if the given gravity anomalies on the geoid represent a continuous field function. In practice, however, the gravity is measured pointwise. Then, in order to approximate the actual gravity anomaly field, an appropriate representation of the field in the form of surface elements is required. In this case, each surface element is represented by one gravity anomaly value. A very common approach is to define mean (equiangular) block gravity anomaly as surface element value (see Figure 2).

Because of this discretization procedure, the integral operator in the Stokes' formula expressed in equation (5) becomes summation operator, De Min (1996b):

$$N(P) = \frac{R}{4\pi\gamma} \sum_{n=1}^{I} \overline{\Delta g}_i \times \times \int_{\lambda=\lambda_Q-\frac{\Delta\lambda}{2}}^{\lambda_Q+\frac{\Delta\lambda}{2}} \int_{\varphi=\varphi_Q-\frac{\Delta\varphi}{2}}^{\varphi_Q+\frac{\Delta\varphi}{2}} S(\Psi_{PQ}) \cos\varphi_Q \, d\lambda_Q \, d\varphi_Q \qquad (9)$$

N(P) is the computed geoid height at point P, and $Q(\varphi_Q, \lambda_Q)$ is the center point of block i. Whereas the $\overline{\Delta g}_i$ represents mean gravity anomaly of the block, and it can be determined by using simple arithmetic mean of the available data Δg_j within the block as follows:

$$\overline{\Delta g}_i = \frac{1}{I} \sum_{j=1}^{I} \Delta g_j \qquad (10)$$

A better formulation is the use of "interpolate-average" technique as shown by Rummel (1991) and De Min (1996b) as follows:

$$\overline{\Delta g}_i = \frac{1}{2\pi(1-\cos\Psi_i)} \cdot \int_{\Psi=0}^{\Psi_i} \int_{\alpha=0}^{2\pi} \Delta g(\Psi,\alpha)\, \sin\Psi \, d\alpha \, d\Psi \qquad (11)$$

The determination of $\overline{\Delta g}_i$ is simple if the gravity anomaly data cover the earth's surface with sufficient density. However, the problem will arise if the available data are very few in number or they are sparsely distributed or even if there is a "no data" area. There are several ways that can be applied to approximate the mean block gravity anomaly value if there is no data available inside the block. The approach can be interpolation/prediction,

Moritz (1980) and Tscherning (1994), or gravity anomaly derived from geopotential coefficient data.

From geopotential coefficient data, in spherical approximation, the block value within the area of A can be approximated by :

$$\overline{\Delta g}_i \approx \frac{1}{A}\iint_A \Delta g^{ggm}\, dA, \tag{12}$$

where

$$\Delta g^{ggm} = \gamma \sum_{n=2}^{n_{max}} (n-1). \cdot \sum_{m=0}^{n} (\overline{C}_{nm}\cos m\lambda + \overline{S}_{nm}\sin m\lambda)\overline{P}_{nm}(\sin\varphi), \tag{13}$$

and $\overline{C}_{nm}, \overline{S}_{nm}$ are the fully normalised geopotential coefficients of the anomalous potential, $\overline{P}_{nm}$ are the fully normalised Legendre functions, and n_{max} is the maximum degree of the geopotential coefficient model.

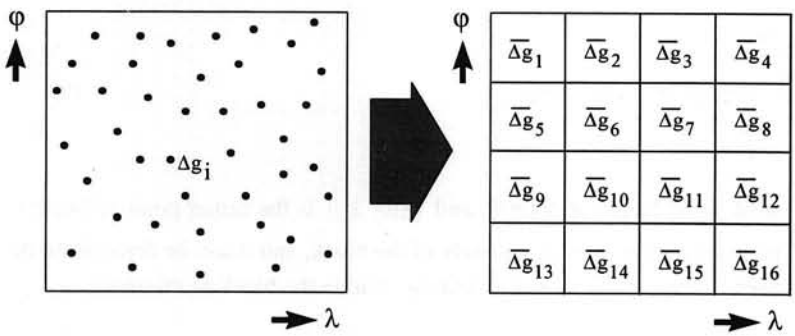

Fig. 2. Gravity anomaly representation.

Another alternative approach is the use of least squares prediction method. This method to predict point gravity anomalies is defined through the following equation, Heiskanen & Moritz (1967):

$$\overline{\Delta g} = \mathbf{C_{ij}}\left(\mathbf{C_{jj}}\right)^{-1}\underline{\Delta g} \tag{14}$$

Here, $\mathbf{C_{ij}}$ is he row vector containing signal cross-covariance between the gravity anomaly being predicted $\overline{\Delta g}$ and the vector of observed gravity anomaly $\underline{\Delta g}$. Also $\mathbf{C_{jj}}$ is the auto-covariance matrix of the observed gravity anomaly. It can be seen that for optimal prediction, the statistical behaviour of the gravity anomalies, represented by their covariance matrices, must be known.

In the current situation, the sea and ocean areas in general are not well covered by marine gravity data. Fortunately, the contrary holds for altimeter data. Based upon the altimeter and

gravity data, one can compute free-air gravity anomalies at sea. There are several techniques used to carry out this computation, for example the use of least-squares collocation, Basic & Rapp (1992) and Li & Sideris (1997), the inverse Vening Meinesz formula, Rummel & Haagmans (1990), Sandwell (1992), and Hwang (1997).

A schematic flow diagram of data preparation procedures can be seen in Diagram 1.

2.3 Combination Solution

Since the gravity data usually does not cover the whole earth but is only available locally, then equation (5) cannot be used directly to calculate geoid heights in practice. To overcome this problem, for the computation of geoid height N, usually two kinds of data sets are combined, i.e. local gravity anomaly data and geopotential coefficients. Based on those data sets, there are two ways to calculate geoid height, Rapp & Rummel (1975):

Method 1:

$$N = N_1 + N_2 = \frac{R}{4\pi\gamma}\iint_{\sigma_c}\left(\Delta g - \Delta g_{ref}\right) S(\Psi)\, d\sigma_c + N_{ref}\,, \tag{15}$$

where

$$\Delta g_{ref} = \gamma \sum_{n=2}^{n_{max}} (n-1). \cdot \sum_{m=0}^{n} (\bar{C}_{nm}\cos m\lambda + \bar{S}_{nm}\sin m\lambda)\bar{P}_{nm}(\sin\varphi)\,, \tag{16}$$

$$N_{ref} = R\sum_{n=2}^{n_{max}}\sum_{m=0}^{n}(\bar{C}_{nm}\cos m\lambda + \bar{S}_{nm}\sin m\lambda)\bar{P}_{nm}(\sin\varphi) \tag{17}$$

The σ_c indicates the cap size, so the integral is extended only up to $\Psi = \Psi_c$ in which the gravity anomaly data are evaluated. The N_1 and N_2 represent the short and long wavelength contributions of gravity field respectively. This method is also called the "remove-restore" technique.

Method 2 :

$$N = N_1 + N_2 = \frac{R}{4\pi\gamma}\iint_{\sigma_c}\Delta g\, S(\Psi)\, d\sigma_c + + \frac{R}{2\gamma}\sum_{n=2}^{\infty} Q_n(\Psi_o)\,\Delta g_n(\varphi,\lambda)\,, \tag{18}$$

where Δg_n is the degree anomaly computed from

$$\Delta g_n = \gamma(n-1). \cdot \sum_{m=0}^{n} \left(\overline{C}_{nm}\cos m\lambda + \overline{S}_{nm}\sin m\lambda\right)\overline{P}_{nm}(\sin\varphi), \tag{19}$$

and the Molodenski's truncation coefficient $Q_n(\Psi_o)$:

$$Q_n(\Psi_o) = \int_{\Psi_o}^{\pi} S(\Psi)\, P_n(\cos\Psi)\, \sin\Psi\, d\Psi\,. \tag{20}$$

For both methods, the geoid height N can be computed and leads to identical results provided that they are applied properly. But in practice, the first method is usually chosen since it requires less computational effort.

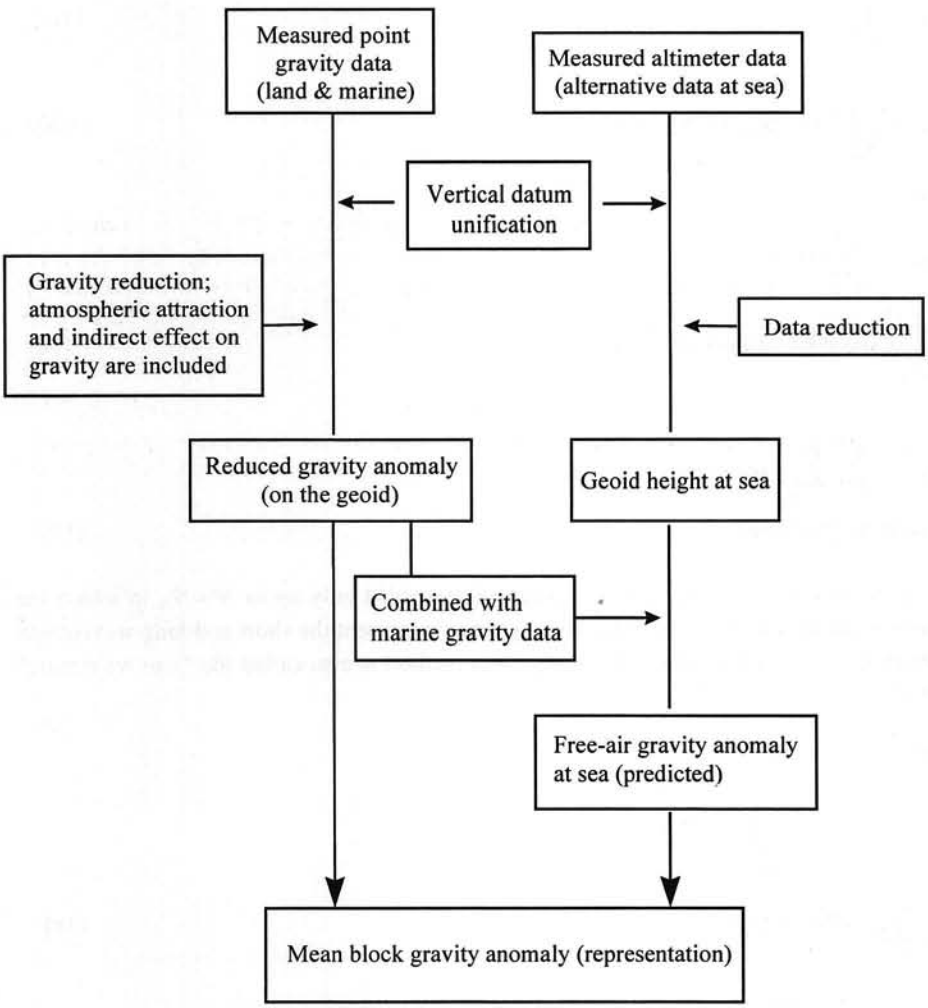

Diagram 1. Data preparation procedure for geoid determination

Furthermore, it is important to find a proper choice of combination such that the global geopotential model information and the local gravity data are well combined in terms of reducing the truncation error. In view of the second method, equation (18) can be rewritten as, De Min (1996a):

$$N = N_1 + N_2 = \frac{R}{4\pi\gamma} \iint_{\sigma} S_1(\Psi)\, \Delta g_1\, d\sigma + \\ + \frac{R}{4\pi\gamma} \iint_{\sigma} S_2(\Psi)\, \Delta g_2\, d\sigma\,, \tag{21}$$

where

$$S_1(\Psi) = \sum_{n=2}^{\infty} \frac{2n+1}{n-1}(1 - w_n)\, P_n(\cos\Psi)\,, \tag{22a}$$

$$S_2(\Psi) = \sum_{n=2}^{\infty} \frac{2n+1}{n-1} w_n\, P_n(\cos\Psi)\,, \tag{22b}$$

and always $S_1(\Psi) + S_2(\Psi) = S(\Psi)$. There are many possibilities to choose the weight w_n which determine the proper combination of the two data sets.The different choices are known as kernel modifications. In the computation of precise geoid for the Netherlands, the chosen Stokes' kernel modification is the *combined Meissl/Wong&Gore* model. This model has the following properties:

- it can be tuned to select which degrees are mainly used from global geopotential model,
- the kernel is exactly zero at the inner zone boundary, and the spectral weights stay close to zero for higher degree n.

Following the above description, the N_1 in equation (15) is modified as :

$$N_1 = \frac{R}{4\pi\gamma} \iint_{\sigma_c} \left(\Delta g - \Delta g_{ref}\right) S_1^{MWG}(\Psi)\, d\sigma_c \tag{23}$$

where $$S_1^{MWG}(\Psi) = \begin{cases} M(\Psi) & 0 \le \Psi \le \Psi_c \\ \\ 0 & \Psi_c < \Psi \le \pi \end{cases} \tag{24}$$

and

$$M(\Psi) = S(\Psi) - \sum_{n=2}^{L} \frac{2n+1}{n-1} P_n(\cos\Psi) - \\ - \left(S(\Psi_c) - \sum_{n=2}^{L} \frac{2n+1}{n-1} P_n(\cos\Psi_c) \right)$$

In precise geoid computation, one should also pay attention to the aspect of numerical integration of equation (23) in view of equation (9). As we know, the gravity anomaly data used in practical computation are often in the form of mean block values. In evaluating numerical integration, this will introduce discretization error especially for inner most ($\Psi < \Psi_0^*$) residual gravity data. The situation of the inner zone σ_0 and the inner most zone σ_0^* can be seen in Figure 3.

To reduce that kind of error, the contribution of the inner most data to the geoid height can be calculated by least-squares collocation method. Implicitly, the least-square collocation does two steps at once: least-squares prediction and Stokes' integral. It allows to create automatically a smooth gravity function through the given point values. Hence, we use point gravity data instead of mean block values as input to the computation. A refined collocation formula for this purpose has been shown by *de Min (1995)*. On the other hand, the collocation method has a numerical problem of large inversion matrix for larger number of the data. Furthermore, the contribution of the data in the rest of the inner zone ($\Psi_0^* < \Psi < \Psi_0$) can be evaluated by numerical integration. Several numerical methods of evaluating Stokes' integral based on Fast Fourier Transform (FFT) approach have been proposed by many authors. A more refined technique for evaluation of convolution integral on the sphere was introduced by applying one-dimensional FFT [*Haagmans et al, 1993*]. So all practical advantages of both methods in evaluating equation (23), collocation and numerical integration, can be fully exploited.

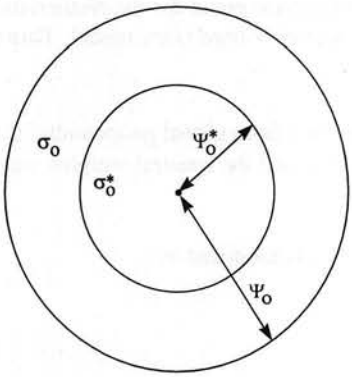

Fig. 3. The inner zone σ_0 and the inner most zone σ_0^*

The solution of geodetic boundary problem described previously is based on spherical and constant radius approximation. To have a better solution, an ellipsoidal correction δN_e should also be added to the result of equation (15). The magnitude of this correction is in the range of cm-level for the selected combined Meissl/Wong&Gore kernel modification [*de Bruijne et al, 1997*]. For Stokes' formula the correction can be computed from [*Pavlis,1988*]:

$$\delta N_e = \frac{R}{4\pi\gamma} \int_{\sigma_o} S_1^{MWG}(\Psi)\, \varepsilon_{\overline{\Delta g}}\, d\sigma \qquad (25)$$

where $\varepsilon_{\overline{\Delta g}}$ is the higher order term of equation (2) series expansion containing ellipsoidal flattening term.

Then a complete expression to derive gravimetric geoidal height N from combination solution can be written as :

$$N = N_1 + N_2 + \delta N_e + \delta N \tag{26}$$

2.4 Comparison with Independently Derived Geoid Heights

The geoid height derived by equation (26) is still contaminated by errors in the lower frequencies of the selected global geopotential model. The magnitude of the error is in the order of a decimeter. Note that a 70 cm difference in this study comes from a 70 cm difference in the reference ellipsoids of TOPEX/Poseidon and GRS'80. If the computed gravimetric geoid is compared to external and independently derived precise geoid information, we may see the error pattern of the long wavelengths.

Besides gravimetric approach, precise geoid heights can also be determined geometrically by means of GPS/levelling combination on land area, and altimetric techniques at sea. Recently, the TOPEX/Poseidon altimeter data are considered very precise. If the geometric geoid is well defined and referred to a regional reference equipotential surface, then correction to the gravimetric geoid can be modelled and computed. The procedure for correcting the gravimetric geoid is shown by *de Bruijne et al (1997)*.

The schematic flow diagram of evaluating gravimetric geoid using this combination solution approach can be seen in Diagram 2. The main input of this procedure are the global geopotential model and the mean block gravity anomaly as a result of the data preparation stage (Diagram 1).

3 Data Availability

The definition of the area of investigation is the Indonesian region and its surroundings. Approximately, it has the following extension in geographical coordinates:

$$-15^{o} \leq \text{latitude } \varphi \leq 10^{o}$$

$$90^{o} \leq \text{longitude } \lambda \leq 145^{o}$$

The data availability related to geoid determination within the area defined above is of major importants for the current choice of the strategy for geoid determination at present and for possible future projects for data collection. The current situation is as follows:

- **Gravity data**

The gravity data covering Indonesian archipelago and its surroundings are available at several databases such as *Bureau Gravimetrique International* (BGI), *National Oceanic and Atmospheric Administration* (NOAA), *Badan Koordinasi Survey dan Pemetaan Nasional* (Bakosurtanal) and several other institutions. For the land part and very few sea part, Bakosurtanal or *National Gravity Committee* of Indonesia may provide point or line gravity data, while BGI or NOAA provides line gravity data for sea areas (see Figure 4). Several

lines of marine gravity data in Banda Sea area are also available from the Snellius II expedition [*Strang van Hees, 1987* and *Woodside et al, 1989*]. As explained in the previous chapter, the gravity data (free-air gravity anomaly) at sea can also be derived from satellite altimeter data. Recently, there are altimeter data available from several satellite altimetry missions such as Geosat, ERS-1, ERS-2 and TOPEX/Poseidon. Several institutions also derive and provide free-air gravity anomaly at sea based upon these data such as *National Geodetic Data Center*/NOAA (USA), *National Chiao-Tung University* (Taiwan), *Kort & Matrikelstyrelsen* (Denmark), and *DEOS, Delft University of Technology* (The Netherlands).

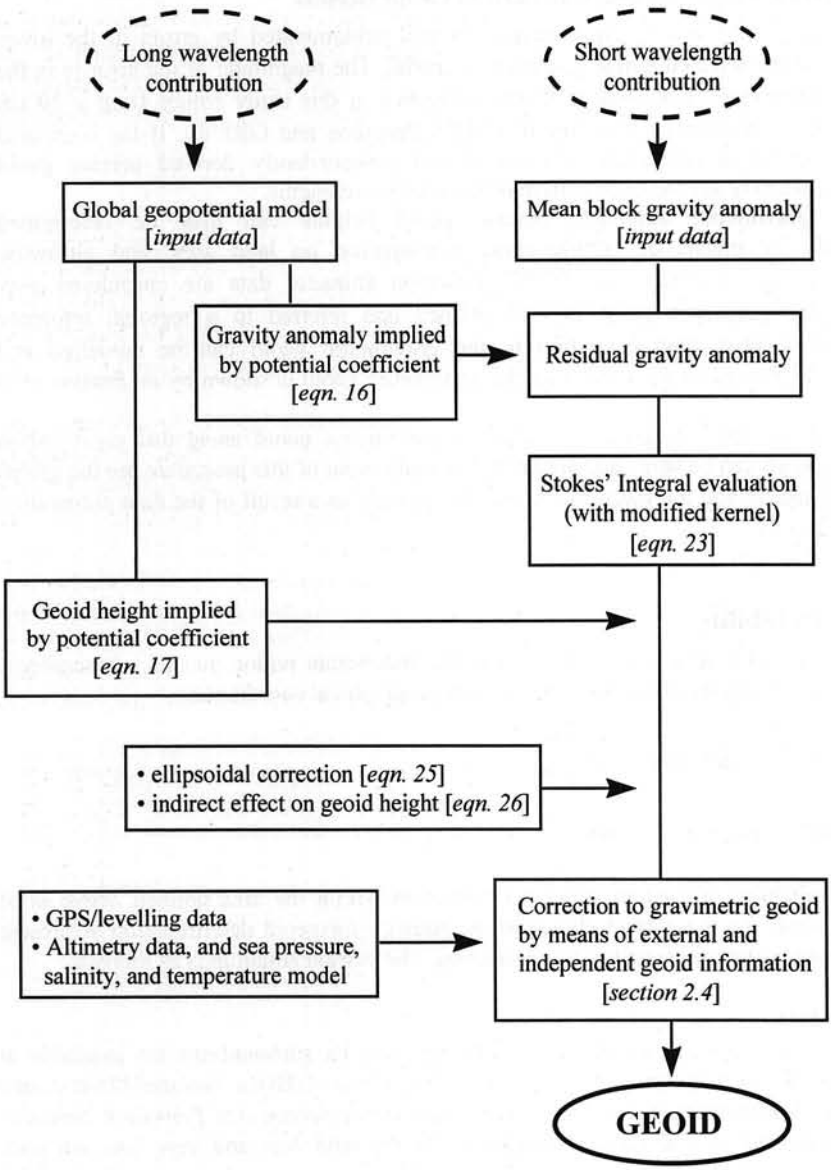

Diagram 2. Gravimetric geoid computation by combination solution approach

The University of Leeds, in collaboration with South East Asian countries and other international institutions, conducted the so called *South East Asia Gravity Project* (SEAGP) during 1991-1995. Based upon the SEAGP data, *Geophysical Exploration Technology* (GETECH) processed and compiled all available land and marine gravity data, and satellite altimetry derived gravity in this area into a unified data set. From the SEAGP original data coverage in Figure 5, it can be seen that there is very poor marine gravity data in the Indonesian archipelago. Fortunately, the sea part is well covered by satellite altimeter data. It can also be seen that there are gap areas in land parts especially in Kalimantan, Sulawesi and Irian islands. Filling the gap areas by means of airborne gravimetric survey should be considered in the future. The GETECH's derived gravity product is in the form of 5'x5' grid of free-air, bouguer, and isostatic gravity anomalies. This does not mean that the data set has 5'x5' data resolution. The data set is referred to the IGSN71 gravity datum, processed using the WGS'84 gravity formula and terrain corrected to 167 km.

- *Digital Terrain Model*

In gravity reduction to correct for terrain effects, and in gravity interpolation/prediction to smooth the gravity field, Digital Terrain Model (topography and bathymetry data) of the region is required. Up to now, Bakosurtanal provides the high resolution DTM only for Jawa, i.e. still under construction, and parts of the Sumatera islands. Again, as mentioned in the beginning, a unified national vertical datum does not exist in the region. In this case, each island has its own vertical datum.

Besides gravity data, GETECH also provides a 5'x5' grid of topography and bathymetry for the whole world called Global DTM5. The derived topography and bathymetry for Indonesian region can be seen in Figure 6. In constructing this data set, various elevation data sets were used: ETOPO5, height data from gravity stations, topographic maps, national DTMs, bathymetry, satellite derived heights, and shoreline, GETECH (1995). Even though it covers land and sea globally, it does not mean that the height of grid points in the data set refer to a unique vertical datum. It is suspected that the height component of the DTM is adopted directly from the original data source. Another weakness of this data set is the resolution of the data. With 5'x5' or about 10 km x 10 km grid, the high frequency information of topography cannot be recovered. It is expected in the future that the realisation of the high resolution SAR derived DTM will be carried out.

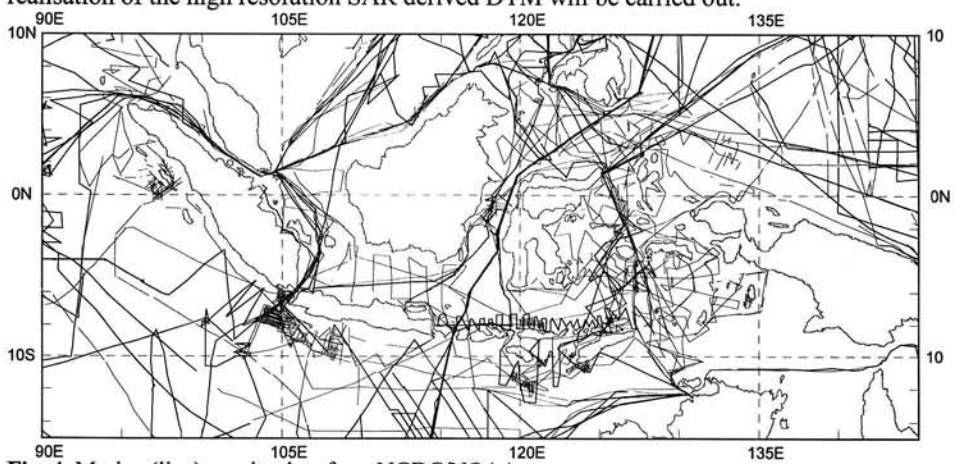

Fig. 4. Marine (line) gravity data from NGDC/NOAA.

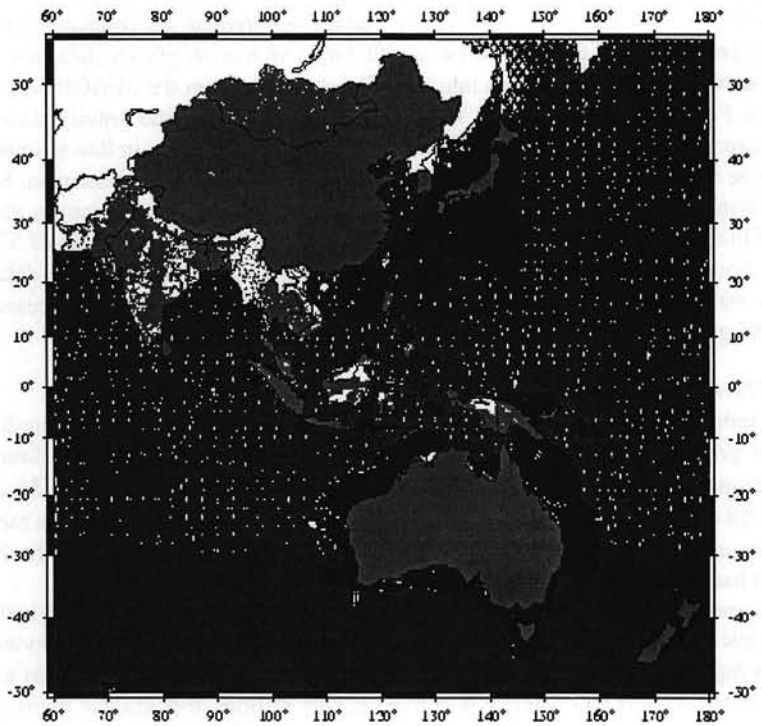

Fig. 5. Data coverage for determination of GETECH's gravity anomaly data set. Land gravity data (light gray), altimeter data (medium gray), and marine gravity data (dark gray).

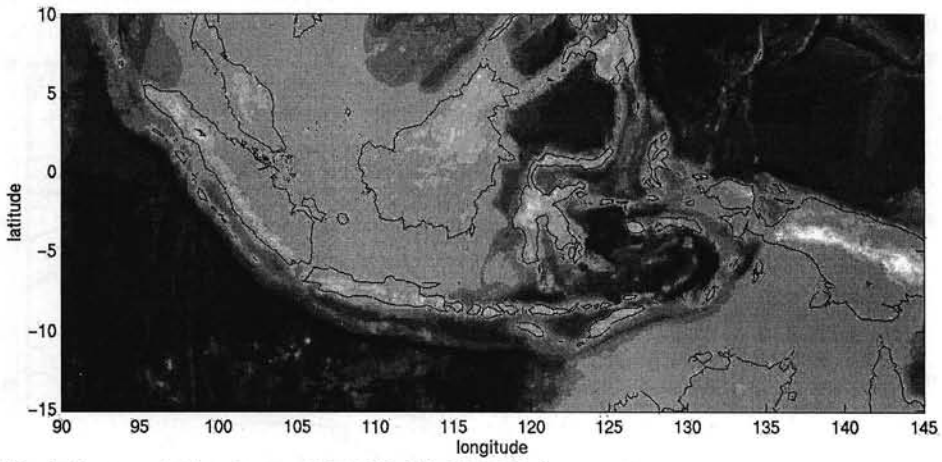

Fig. 6. Topography based upon GETECH global DTM5 (in meters).

- *Geopotential Coefficient Model*

The use of potential coefficient models for the calculation of geoid heights has been carried out for years. Today a high resolution global model up to degree 360 is used routinely. Improvements on the model have been carried out, especially in reducing the long wavelength error in the model, Rapp (1996). The most recent model is *Earth Gravity Model 1996* (EGM96). This model has smaller formal error degree variances for all degrees than those of the previous model, OSU91, De Bruijne et al. (1997). In deriving the OSU91 model, there were very few data from Indonesian region contributed in the computation. But in the EGM96, through the SEAGP, the data from the region was included so that one can expect more accurate potential coefficient derived geoid height in the Indonesian region.

- *Geometrically Derived Geoid Height*

Some independently precise determined geoid heights can be used as a comparison to results computed by gravimetric method. This independent information can be derived geometrically from combination GPS/levelling technique at land area, and very precise altimetry measurement at sea such as the TOPEX/Poseidon data with a permanent sea surface topography model. There are several GPS/levelling points distributed over several major islands. Since each island has its own vertical datum, then the geoid height at those points do not refer to a unique equipotential surface. In order for those points to be used as control to the regionally determined geoid, a unified vertical datum in the region is required.

- *Auxiliary Data*

Several types of oceanographic data are required in reducing satellite altimeter data, and sea-surface topography determination. Sea tide data are required in reducing altimeter data to derive geoid height. There are some tide gauge stations maintained by Bakosurtanal and NOAA in the region. For ocean areas, the tide can also be modelled through a most recent global model called Provost model, Provost et al. (1994). The validity of this new tide model in the Indonesian waters should be tested. Following Khafid (1997), the sea surface topography can be determined oceanographically (ocean levelling technique) based upon sea current, temperature, salinity, pressure, and density data.

4 Strategies for Geoid Determination in the Indonesian Region

As mentioned previously, GETECH's SEAGP gravity data set covers the South East Asian region in a 5'x5' grid of gravity anomalies, i.e. free-air, bouguer, and isostatic gravity anomalies. These gridded gravity anomalies are derived by using combination of terrestrial point gravity (land and sea) and satellite altimeter data at sea. The data set was not derived for precise geoid determination purpose, but for other geophysical applications such as regional geological interpretations, basin analysis, and continental margin studies, GETECH (1997). It is not clear how the data set was derived. Of course there are some criteria that have to be fulfilled to develop a gravity anomaly data set for precise high resolution geoid determination. Therefore, before it is used to compute geoid heights, several questions related to the data set are remain open :

- What is the precision and reliability of original gravity data set used to derive gridded data ?
- What is the density and distribution of original data set to derive gridded data ?
- What is the vertical datum used to unify the gravity anomalies in a unique height reference system ?

- What is the formula used in deriving (point) free-air, bouguer, and isostatic anomalies ? What is the numerical approach to evaluate the analytical formula in the gravity reductions ?
- How are the satellite altimeter data reduced to geoid height at sea ? What is the tide data or model used ? Is the sea surface topography taken into account ?
- Are tide gauge measurement used and in what way ?
- How are different data (combined and) weighted ?
- What is the Digital Terrain Model used in gravity reduction ? What is the quality and resolution of the DTM ?
- What is the formula used in transforming geoid height at sea to free-air gravity anomaly?
- How to unify the vertical datum of both land and sea gravity anomalies ?
- What is the approach used in gridding the original point gravity data into a 5'x5' grid data set ?

Those questions should be answered in order to have more detail information about the data set so that the difficulty in estimating accuracy of the computed geoid heights can be reduced.

Another alternative of gravity data set is the use of point (original) gravity and altimeter data. The main point here is data treatment in preparation for geoid height computation. Here, we assume that all of the gravity data have already been referred to a selected gravity datum. There are some problems that should be considered and anticipated for further investigations and improvements:

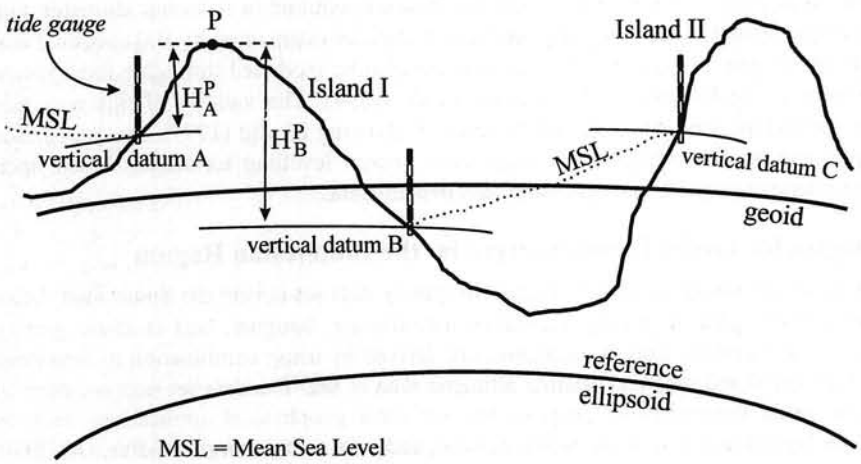

Fig. 7. Inhomogeneous vertical datums and the geoid

- *Unified vertical datum does not exist in the region.*

Up to now, every island in Indonesian region has its own vertical height datum. This situation is described in Figure 7. The effect of vertical datum inconsist-ency on the geoid computation is shown by Khafid (1997). A one meter difference of height datum in Kalimantan island can cause up to 12 cm error in geoid computation. The error magnitude is

significant in precise geoid determination. In this research he also attempted to unify the vertical datum by means of oceanographic and geodetic approaches. The most interesting is the oceano-graphic approach, i.e. ocean levelling technique, since it is independent of local geoid information.
However, the results show that the (dynamic) sea surface topography values for datum trans-formation is only available for deep water areas. Alternatively, another possible approach is the use of satellite altimetry data in combination with a geoid derived from potential coefficients purely derived from the analysis of satellite tracking data. However, the spatial resolution is rather limited. The recent geopotential model has a maximum degree of at most 70. The future satellite geopotential model, based on satellite gradiometry technique, may provide such a model up to degree and order of 200-300 which is equivalent to a spatial resolution of between 50 and 100 km (half-wavelength), Koop (1993). Although it still needs to be refined, the two methods can be the best candidate to solve vertical datum problem in geoid computation in Indonesian region.

- *Mass density model and DTM data in land gravity reduction.*

As can be seen from Figure 6, the topographical setting of the land part of the Indonesian archipelago is mountainous. The maximum height can reach more than 4000 m above sea level at Irian island. It is well known that besides its height variation, the area is also located at a tectonically active region, Bowin et al. (1980) and Katili (1989). This is characterised by the presence of tens of active volcanoes. This situation makes us aware about the topographic mass density variations and the needs of gravity reductions. The formulas used in gravity reductions usually adopt the mean mass (above geoid) density value of $2.67\ \mathrm{g\ cm^{-3}}$. For the Indonesian region situation, this density value could be not appropriate anymore, especially in precise geoid computation. Thus, the need of mass density modelling in gravity reductions becomes important for future investigation. Another important aspect in gravity reduction is the topography information. For mountainous region, an accurate and high resolution DTM is necessary to compute the topography attraction effects on gravity. Anticipating the high resolution topography information in the near future, refinement of gravity reduction and indirect effects models should be considered, especially in the inclusion of the higher order or non-linear terms which is usually ignored in the computations. Besides for gravity reductions, a good knowledge of mass density and topography information can also be useful in gravity interpolation.

- *Representation technique to derive mean block gravity anomalies in the region.*

Before deriving mean block gravity anomalies, we have to know the density distribution of gravity data in the computation area. As mentioned previously, there are still data gap areas in land parts, especially in Kalimantan, Sulawesi, and Irian islands. A proper handling of this problem also needs more investigation. For sea areas, a suitable approach for transforming altimeter data to gravity anomalies should be searched. Fortunately, there are several institutions that provide altimetry derived free-air gravity anomaly at sea. The use of this data sets should also be tested. The selected block size should describe the resolution of gravity data in the region.

With an appropriate data preparation approach, we may expect an optimal data representation for the region of computation. In the following, based on data availability and

their characteristics, three alternative procedures for geoid determination in Indonesian region will be discussed.

4.1 Procedure A : Utilisation of GETECH's Data Set

In this first procedure, we deal with the possibility of utilising the bouguer or isostatic gravity anomaly contained in GETECH's data set. If it is used directly to compute geoid heights for Indonesian region, then at least two assumptions should be made :

- All of the reduced gravity anomalies are referred to a homogeneous vertical datum.
- The 5'x5' grid values represent the 5'x5' mean (equiangular) block gravity anomaly values.

Another required data set is the global geopotential coefficients. In this case the EGM96 model is selected. However, the potential coefficients of EGM96 refer to its own ellipsoid, Rapp (1996). As mentioned, GETECH's gravity data refer to the WGS84 datum, so transformation of all EGM96's coefficients with respect to degree n should be made, de Bruijne et al. (1997):

$$\overline{\Delta C}_{WGS84} = \frac{GM_{EGM96}}{GM_{WGS84}}\left(\frac{a_{EGM96}}{a_{WGS84}}\right)^{n}\overline{\Delta C}_{EGM96} \tag{27}$$

where $\overline{\Delta C}$ represents the fully normalised geo-potential coefficients $\bar{C}_{nm}$ and $\bar{S}_{nm}$ of degree n and order m with respect to an ellipsoidal reference, GM is the gravitational constant, and a is the semi-major axis of the reference ellipsoid.

Based on those two data sets, i.e. gravity anomaly and geopotential coefficient data, geoid heights are then evaluated by using the combination solution approach as described in section 2.3. In order to yield an optimal combination solution, the following should be considered for future investigations :

- The search of appropriate maximum degree L of modified kernel or other kind of modifications in evaluating equation (24).
- The search of optimum cap size σ_c .

By means of comparing to external and independent geoid heights, the long wavelength (not globally, but locally) error in the geopotential coefficients can be computed. The computed correction might be still contaminated by the effect of vertical datum inconsistencies.

Since the history of GETECH's gravity data processing is not clearly understood, and also we imposed two assumptions mentioned above, it is very difficult to estimate the accuracy of the resulted geoid heights.

4.2 Procedure B : Utilisation of Original Data Set

Instead of using gravity anomalies from GETECH's data set, a better alternative is the use of original data sets, i.e. point gravity data and satellite altimetry data. In this way we can take some advantages on the use of more accurate data sets and better models such as newly determined topography from SAR, newer tidal model etc. As described in the beginning of this section, the procedure of the whole geoid computation can be written as follows :

- *Data preparation* :

We have first to define a homogeneous vertical datum for the whole region (land and sea) of computation. In preparing the data before reduction to geoid, all point gravity anomalies must also be referred to the same system, i.e. gravity datum. The practical procedure for calculating gravity anomalies from gravity data can be found in Heck (1990). As mentioned previously, there are many ways to carry out gravity reductions. The important thing is how to do the reduction in a proper way. As we know, there are fewer terrestrial gravity data available at sea. Fortunately, the sea areas are well and densely covered by satellite altimetry data, i.e. ERS and Geosat altimeter data. After being reduced to geoid, then the altimeter data in combination with marine gravity anomalies are processed to derive marine free-air gravity anomalies. Here, the TOPEX/Poseidon data are not included in the computation. These data will be used as an external and independently geoid information in the later stage. Another alternative for the sea area is the use of available satellite altimetry derived free-air gravity anomaly data set provided by some institutions. As the next step, the point gravity anomalies are transformed to the mean block values. Unfortunately, there are some data gap areas at several islands, especially in Kalimantan, Sulawesi, and Irian islands. As described in section (2.2.2), there are several possible techniques to compute the gravity anomaly values in the data gap areas. The following should be considered in gravity data preparation for future improvements and investigations :

1. Application of Khafid's approach in vertical datum unification.
2. Gravity reduction procedures by means of a mathematical model, mass density model, and topography information.
3. Methods of determination of marine free-air gravity anomaly from satellite altimetry data in combination with terrestrial marine gravity data.
4. Suitable technique for gravity anomaly interpolation/prediction in data gap areas.
5. Selection of optimal block size.

- *Combination solution* :

As before, the data needed in this stage are mean block gravity anomalies and geopotential coefficients. Here, we also propose the use of geopotential coefficients from EGM96 model. Geoid heights are then evaluated by using the combination solution approach as described in section 2.3. Not like in *procedure A*, here the use of least-squares collocation in the inner most zone for reducing discretization error proposed by De Min (1995) can be applied since the point gravity anomaly data are available. Of course this approach is only applicable for such an area with sufficient data density. The following should be considered in evaluating combination solution to calculate geoid height for future improvements and investigations :

1. The search of appropriate maximum degree L of modified kernel or other kind of modifications in evaluating equation (23).
2. The search of optimum cap size σ_c.
3. The use of available marine free-air gravity anomaly data set in geoid computation.

- *Comparison with external and independent geoid information* :

The realisation of vertical datum unification allows the external and independent precise geoid data (GPS/levelling and TOPEX/Poseidon data) to refer to one equipotential surface. As explained previously, the reason of comparing the gravimetric geoid to the external and independent one is to correct the computed gravimetric geoid heights with the long wavelength error contained in the geopotential model. Since the error magnitude is in decimeter level typically, then this correction will be optimal if the accuracy of the independent geoid height is in the order of less than one decimeter level. On the other hand, the computed gravimetric geoid heights have to be precise too. Otherwise, the correction not only contains long wavelength error but also other type of errors. Related to the current Indonesian region situation, the goal of this procedure is difficult to be achieved due to many factors.

To compute the geoid for the whole Indonesian region, it needs not only gravity data from the Indonesian region itself but also the data from several neighbouring countries such as Malaysia, Australia, Singapore, Papua Nugini, Phillipines, and Brunei Darussalam.

4.3 Procedure C : Island-by-Island Geoid Determination

As we have seen from gravity data availability, there are still some data gap areas at several islands. In *procedure A* and *procedure B*, these gap areas are filled by means of interpolation or prediction. Of course, this approach introduces errors. The procedure of vertical datum unification also introduces errors. Therefore, both errors will propagate to the computed geoid heights. The error magnitude in geoid heights caused by those two sources could be reduced by limiting area of computation. Here, the strategy is *island-by-island* geoid determination. Based on data availability, compared to other islands, Jawa and Sumatera are the best. For both islands, there are two main advantages of using this approach :

- The vertical datum unification procedure is only applied to the data located at sea and at the neighbouring islands.
- The gravity data density and quality can be more uniform.

The procedure of the geoid height computation can follow the *procedure B*. In this way, we may expect best results for Jawa or Sumatera islands based upon the current data availability. Unlike applying the previous approach (*procedures A* and *B*), this approach may exploit the whole procedure of geoid heights computation described in the section 2. In case of other islands, the *procedure A* or *procedure B* can be applied.

5 Conclusions and Recommendations

The main constraints in precise geoid determination in Indonesian archipelago are :

- Vertical datum unification problem.
- Data availability.

Considering those two aspects, three possible procedures for geoid height computations are proposed. The most simple but the least accurate approach is *procedure A*. This can be easily understood since the GETECH's data set is not prepared for geoid determination. The second approach, *procedure B*, is expected to give better results than the first approach, but the procedure is very laborious. Moreover, there are still some weaknesses in this procedure,

i.e. the quality of reference height datum over the whole region, and the effects of data gap areas. These can be said to be the main limiting factors of precise geoid determination for Indonesian archipelago. The better approach but the less area coverage of the computational area is *procedure C*. This procedure tries to reduce the computed geoid height errors by confining the area of computation by means of avoiding data gap areas, and also reducing the errors in inter-island vertical datum unification. Applying the last approach to the Jawa and Sumatera islands, the best derived geoid of part of Indonesian region can be expected.

The recent advances in GPS technique and gravimetry instrumentation make the gravity measurement by means of airborne gravimetry more promising. At the same time, the SAR technique also provides a more refine high resolution topography data which are needed in the gravity reduction. This technique should be taken into account for filling the gravity gap areas in the future consideration. This technique may also provide solution of a unique reference for a proper land-sea gravity transition problem. Besides, these gravity data are also useful for other geophysical purposes.

Another problem is in modelling. In general, various assumptions in geoid modelling are no longer hold in Indonesian geoid determination. The refinement or improvement on the mathematical models in geoid computation is also very important. There are many assumptions imposed in the derivation of the Stokes' boundary value problem solution. Besides, the resolution and the accuracy of data become higher and better. Another imperfectness is the simplification of mass density modelling in gravity reduction. Therefore, refinement or improvement on those two models is necessary for the future work.

As an alternative to the Stokes' approach, the use of Molodenski's boundary value problem solution in geoid determination for Indonesian archipelago should also be considered in the future investigation. The main advantage of this approach is the independency of the mass density knowledge. In this way we get the so called *quasigeoid* instead of the geoid. At sea or ocean, the quasigeoid is equal to the geoid, but this is not the case in land areas. If we relate to the local adopted height system, it means that we have to change the system from orthometric height to normal height. Of course, it is also possible to transform the quasigeoid to the geoid, but again the knowledge of mass density is required.

Precise geoid determination for Indonesian archipelago is a very laborious task and a long term process. Following this report, the next tasks can be several investigations to test the proposed procedures (*A* and *B*) in small test areas. Besides testing and validating the data used, the investigation should also more concentrate on the geoid modelling as described in the previous sections. From this stage, it is expected to yield a more optimal and suitable geoid determination procedure to be applied in Indonesian archipelago.

Acknowledgments This study is carried out at Delft institute for Earth-Oriented Space research (DEOS), section of Fysische, Meetkundige en Ruimtegeodesie, Delft University of Technology, under the frame work of cooperation between the Delft University of Technology (DUT) and Institut Teknologi Bandung (ITB). The author wishes to thank Prof. Karl F.Wakker and Prof. Roland Klees for hosting me at the DUT, and Prof. Joenil Kahar (ITB) for offering the opportunity to perform this challenging study. I would like also to thank Roger Haagmans and Arnoud de Bruijne for their valuable discussions during this study.

References

Basic, T. & R.H. Rapp, 1992, *Oceanwide Prediction of Gravity Anomalies and Sea Surface Heights Using Geos-3, Seasat, and Geosat Altimeter Data and ETOPO5U Bathymetric Data*, Report No.416, Dept. of Geodetic Science and Surveying, The Ohio State University, Columbus.

Bowin, C., G.M. Purdy, C. Johnston, G. Shor, L. Lawver, H.M.S. Hartono, P. Jezek, 1980, Arc-Continent Collision in Banda Sea Region, *The American Association of Petroleum Geologists Bulletin*, Vol.64, No.6.

De Bruijne, A.J.T., R.H.N.Haagmans, E.J.de Min, 1997, *A preliminary North Sea Geoid model GEONZ97*, MD-rapport MDGAP-9735, Meetkundige Dienst, Rijkwaterstaat, Delft, The Netherlands.

GETECH, 1995, *Global Digital Terrain Model; Global DTM5*, Geophysical Exploration Technology, Department of Earth Sciences, University of Leeds, Leeds, UK.

Haagmans, R., E.de Min, M. van Gelderen, 1993, Fast evaluation of convolution integrals on the sphere using 1D FFT, and a comparison with existing methods for Stokes' integral, *Manuscripta Geodaetica*, 18: 227-241.

Heck, B., 1990, An evaluation of some systematic error sources affecting terrestrial gravity anomalies, *Bulletin Geodesique*, 64: 88-108.

Heiskanen, W. & H. Moritz, 1967, *Physical Geodesy*, W.H. Freeman and Company, San Francisco.

Hwang, C., 1997, *Inverse Vening Meinesz formula and deflection-geoid formula: applications to the predictions of gravity and geoid over the South China Sea*, Dept. of Civil Engineering, National Chiao-Tung University, Taiwan.

Ilk, K.H., 1996, *Introduction to the gravity field theory*, Lecture notes, Second Tropical School of Geodesy, Bandung, Indonesia.

Kahar, J., 1981, *Analysis of geoid in Indonesian Region*, PhD Dissertation, Institut Teknologi Bandung, Indonesia.

Katili, J.A., 1989, Review of past and present geotectonic concepts of eastern Indonesia, *Netherlands Journal of Sea Research*, Vol.24-No.(2/3).

Khafid, 1997, *On the unification of Indonesian Local Height Systems*, PhD Dissertation (draft), Institut fuer Astronomische und Physikalische Geodaesie, Technischen Universitaet Muenchen, Germany.

Koop, R., 1993, *Global gravity field modelling using satellite gravity gradiometry*, PhD Dissertation, Delft University of Technology, The Netherlands.

de Min, E.J., 1995, A comparison of Stokes' numerical integration and collocation, and new combination technique, *Bulletin Geodesique*, 69:223-232.

de Min, E.J., 1996a, *The Netherlands geoid computation procedure*, Presented at European Geophysical Society, XXI General Assembly, The Hague, The Netherlands.

de Min, E.J., 1996b, *De geoide voor Nederland*, PhD Dissertation, Delft University of Technology, The Netherlands.

Moritz, H., 1980, *Advanced Physical geodesy*, Kalrsruhe: Wichmann Verlag.

Pavlis, N.K., 1988, *Modeling and estimation of a low degree geopotential model from terrestrial gravity data*, Department of Geodetic Science and Surveying, Report 386, Ohio State University, Columbus.

Provost, C.L., M.L. Genco, F. Lyard, 1994, Spectroscopy of the world ocean tides from a finite element hydrodynamic model, *Journal of Geophysical Research*, Vol.99, No.C12.

Rapp, R.H., 1996, *Global models for the 1 cm geoid; present status and near term prospects*, Prepared for International Summer School of Theoretical Geodesy: Boundary value problems and the modelling of the earth's gravity field in view of the one centimeter geoid, Como, Italy.

Rapp, R. & R.Rummel, 1975, *Methods for the computation of detailed geoids and their accuracy*, Department of Geodetic Science, Report 233, The Ohio State University, Columbus.

Rummel, R. and R.H.N. Haagmans, 1990, Gravity Gradients from Satellite Altimetry, *Marine Geodesy*, Vol. 14, No.1

Rummel, R., 1991, *Fysische Geodesie II*, Collegediktaat, Faculteit der Geodesie, Technische Universiteit Delft, The Netherlands.

Sandwell, D.T., 1992, Antarctic marine gravity field from high-density satellite altimetry, *Geophysical Journal International*, 109.

Sideris, M.G., 1994, Geoid determination by FFT techniques, *International School for the Determination and Use of the Geoid*, International Geoid Service, DIIAR, Milano, Italy.

Strang van Hees, G.L., 1987, *Gravity in the Banda Sea, The Snellius II Expedition in Indonesia 1985*, Delft University of Technology, The Netherlands.

Tscherning, C.C., 1994, Geoid determination by least-squares collocation using GRAVSOFT, *Internatinal School for the Determination and Use of the Geoid*, International Geoid Service, DIIAR, Milano, Italy.

Wichiencharoen, C., 1982, *The indirect effects on the computation of geoid undulations*, Department of Geodetic Science and Surveying, Report 336, The Ohio State University, Columbus.

Woodside, J.M., D.Jongsma, M.Thommeret, G.L.Strang van Hees, Puntodewo, 1989, Gravity and magnetic field measurements in the Eastern Banda sea, *Netherlands Journal of Sea Research*, Vol.24-No.(2/3).

3029676

ISBN 90-407-1832-6 Delft University Press