A METHOD FOR ASSESSMENT OF OPTIMAL CHOICE OF PARAMETRIC MODEL IN LEAST SQUARES COLLOCATION.

Odumosu, J. O.[(1)],Onuigbo, I. C. [(1)], Nwadialor, I. J. [(1)] and Kemiki, O. A.[(2)]

[(1)] Department of Surveying and Geoinformatics, Federal University of Technology, Minna

[(2)] Department of Estate Management and Valuation, Federal University of Technology, Minna

ABSTRACT

The process of choosing a suitable parametric model prior least squares collocation suffers from a high degree of arbitrariness. Although, the congruency test (wherein$\sigma^2{}_0 = \sigma^2$)gives an overall impression on the validity of the model, but further testing is always required, even when the congruency test points to the contrary, since it is possible that effects from different modeling errors cancel each other out, in the computation of$\widehat{\sigma^2}$. Besides, the procedure for estimating the variance components from the parametric model is computationally tasking. A semi automated assessment procedure is therefore herein presented by considering some rudimentary statistics of residuals followed by a student's t- test on the mean of residuals. The model is simpler as it eliminates the need for variance component estimation and faster to implement compared to when strict reliance is placed on the congruency test in parametric model selection.

Keywords: *Least squares collocation*, *parametric models*, *stochastic models*, *adjustment computation*.

1.0 INTRODUCTION

The least squares collocation (LSC) is a useful mathematical tool used in gravity prediction, filtering of anomalies and estimation of parameters that define the mathematical model (Ruffhead, 1987). In gravimetric network design, the LSC is a preferred mathematical tool for network adjustment. The process of LSC begins with the formation of observation equations that describe the physical behavior of the outcomes in relation to certain system parameters hence the need for formulation of suitable parametric equation that will accommodate all the essential parameters that affect the system. The process of choosing a suitable parametric model prior least squares collocation suffers from a high degree of arbitrariness. Earlier studies have shown that the parametric model used affects the estimated value of the variance component and consequently might lead to unreliable results since the variance component is often taken as a measure of accuracy (Nafisi, 2003).

A common method of determining suitable parametric models has been the use of the congruency test wherein ${\sigma^2}_0 = \sigma^2$ (Dermanis and Rossikopoulos, 1991) which requires that first a variance component estimate (VCE) be doneafter which an “f”-test is performed between the a-priori value and a-posteriori values of the variance of weighted observations. Unfortunately, VCE is a computationally tedious procedure and is itself dependent on an arbitrary choice of parametric model (Guo and xu, 2015). Besides, the variance component estimation procedure, pre-supposes that *no biases* or *systematic effects* are present in the data. Any unmodelled bias effect (which depends on the parametric equation) may propagate into the estimated variances and give unreliable results (Persson, 1981; Koch, 1999).

The 3 σ –rule is a simple and widely used heuristic for outlier detection especially in statistical based analysis (Lehmann, 2013). Although, the procedure is generally used for outlier detection and geodetic data filtering, it similarly gives an analyst an idea of the suitability of the underlying parametric model upon which the design matrix is built by taking advantage of the statistical properties of a least squares as a minimum variance estimate that is distribution-free.

In this work, asemi-automated assessment procedure is presented by considering some rudimentary statistics of residuals followed by a student’s t- test on the mean of residuals to determine the most suitable among eight (8) tested parametric models using the south Western part gravity points of Nigeria as a case study.

2.0 Statistical checks for assessment of model fit.

Before performing a least squares estimate, certain statistical checks could be used to assess the fit of the parametric model. Assessing model fit is an essential task in a least squares estimation since the quality of the results obtained depends on the correctness or otherwise of imposed model. A good parametric model has the following assumptions (Vanicek et al, 2001):

1) The mathematical expectation of the resulting residuals has zero error i.e mean 0 error
2) Residuals have constant variance
3) The error is normally distributed
4) Observational errors are independent.

Considering the above stated assumptions, an analysis of the residuals obtained from a particular parametric model could be used as a precursory guide in determination of model suitability since the residuals of a least squares estimate are of great significance (Lehmann, 2012). Analysis of residuals involves certain statistics which include:

(i) **Normalized residuals**:

Given the Gauss Markov model, the residuals are given as (1)

$$V = A\widehat{\widehat{X^a}} - L^b \quad (1)$$

Where: A = design matrix and has a rank 3 if the equations are consistent

X^a = Vector of adjusted parameters

L^b = Vector of observations

V = Vector of residuals

Since the variance (σ^2) is given as (2)

$$\hat{\sigma}^2 = \frac{V^T PV}{n-m} \quad (2)$$

Where P = observational weight

n = number of equations

m = number of parameters

The normalized residuals in a least squares are then computed as given in (3):

$$V_{norm} = \sum \frac{V_i}{\hat{\sigma}\sqrt{q_{vv,ii}}} \quad (3)$$

Where

$q_{vv,ii}$ = corresponds to the ith diagonal of the cofactor matrix of residuals (Q_{vv})

$Q_{vv} = P^{-1} - A(A^T PA) - A^T$

P = Weight matrix

(ii) **Sum of squares of residuals** (SSR):

The sum of squares of residuals is given as (4)

$$\mathrm{SSR} = \sum v^2 = \sum((A\widehat{\widehat{X^{a}}} - L^b)^2) \qquad (4)$$

(iii) Standard error (SE):

$$SE = \frac{SSR}{df} \qquad (5)$$

(iv) Error distribution

Although, the least squares method is a distribution free method, computation of confidence intervals and hypothesis testing of the resulting data requires that the error are normally distributed i.e$v \sim N(0, \sigma^2 P^{-1})$which is a pre-requisite for equation (6) to hold

$$v = -Q_{vv} P\, L^a \qquad (6)$$

Where L^a = vector of adjusted observations.

Identifying the data distribution for non-repeated observations could be problematic. The standard geodetic approach is to assume a balance in the number of positive and negative values of residuals as a standard tool for determining the normality or otherwise of residual distribution.

If the sum of positive residuals balances with the sum of non-positive values, then the residuals are considered normally distributed and vice versa.

(v) 3σ-rule for outlier detection

As a general rule of thumb, observations larger than 3σ are often considered as outliers in data adjustment (Lehmann, 2012). However, it is logically reasonable to state that the better the model fit, the more points the model accommodates, hence the lower the number of outliers. The rule is mathematically described by (6)

$$\text{If } v_i \geq 3\sigma \qquad (7)$$

Then i is an outlier observation.

(vi) The student-t test statistic

From the vector of residuals, the t-statistics is then used to further ascertain correctness of the model. Since the least squares estimate is aimed at minimizing the sum of squares of observational residuals, then (8) holds as stated below:

$\sum v^2$ = minimum (8a)

$v^T PV = minimum$ (8b)

$\therefore H_0: E\{v\} \leq 0$ (8c)

$H_A: E\{v\} \geq 0$ (8d)

Taking equation 8c and d, the Null and alternative hypothesis respectively become:

H_0: The mean of residuals is less than or equal to zero therefore the model is optimal

H_A: The mean of residuals is greater than zero therefore the model is not optimal

To test the hypothesis, the one sample student t-statistic given by (9) is then used thereby statistically validating the correctness of the model.

$t = \frac{\bar{x} - \mu}{SE}$ (9)

Where:

μ= hypothesismean

$\bar{x}$= Sample mean

SE = standard error.

3.0 Parametric Model

Determining a suitable parametric model that could efficiently represent the gravity field of a study area is an essential step in LSC since the design matrix is dependent on it. Since the parametric model is expected to include all the known parameters that affect the system. In this work, eight (8) models were tested in order to determine the optimal amongst them using the statistical criteria earlier discussed. The tested models are given in equations 10 (a – h).

$$\Delta g = x_0 + x_1\varphi + x_2\lambda \qquad (10a)$$

$$\Delta g = x_0 + x_1\gamma + x_2\gamma^2 \qquad (10b)$$

$$\Delta g = x_0 + x_1 h + x_2 h^2 \qquad (10c)$$

$$\Delta g = x_0 + x_1 h + x_2\gamma + x_3 h\gamma \qquad (10d)$$

$$\Delta g_{ij} = \delta\gamma_{ij} - \delta g_{ij} + \Delta g_i \qquad (10e)$$

$$\Delta g = x_0 + x_1\varphi + x_2\lambda + x_3 h + x_4\gamma \qquad (10f)$$

$$\Delta g = x_0 + x_1\varphi + x_2\lambda + x_3\varphi\lambda + x_4 h + x_5\gamma + x_6 h\gamma \qquad (10g)$$

$$\Delta g = x_0 + x_1\varphi + x_2\lambda + x_3\varphi\lambda \qquad (10h)$$

Where:

Δg = gravity anomaly

$x_0 . . . x_n$ = parameter coefficients

φ, λ, h = 3D station geodetic coordinates in degrees, degrees and meters respectively.

γ = station normal gravity computed based on the WGS84 normal gravity formulae.

i and j = occupied station and control/reference station respectively.

4.0 Data used:

A total of 4511 gravity data points located across the entire country was collected from the BGI (Bureau Gravimetric International) database. The data were collected from various sources and archived by the BGI (see table 1 for various data sources). The data was then filtered for noise and non-homogeneity using the standard cross over adjustment technique (Odumosu et al, 2016) leaving 2634 nationally homogenous points. The Baarda's outlier detection technique and 3σ-rule was then used to remove outlier observations. A total of 1074 nationally consistent points were left after filtering of which the 193 points located within South Western part of Nigeria.

These 193 points were used in this study in an attempt to model the gravity field of the study area using the LSC.

Table 1: Gravity data from various sources (Collected from BGI)

S/N	**Data Source**	**Number of points**	**Accuracy (mgals)**	**Date of Observation**
1	IGSN-71	1	0.040	1-11-1961
2	Euro/African Secondary Calibration Line survey	4	0.032	1-11-1965
3	British Antarctic Survey	63	Not available	1-11-1975
4	Academy of Science, France	11	Not available	1-11-1938
5	Princeton University, USA	220	Not available	1-1-1969
6	University of Leeds	789	Not available	1-11-1984
7	Geological Survey of Nigeria	987	Not available	1-11-1961
8	Shell Exploration Company	69	Not available	1-1-1965
9	University of Ibadan	192	Not available	1-1-1978
10	Ahmadu Bello University, Zaria	151	Not available	1-1-1978
11	University of Ibadan	303	Not available	1-1-1978
12	University of Calabar and Leeds	1074	Not available	1-1-1984
13	Cratchley, C. R (1960)	460	Not available	1-1-1960
14	Anonymous observer	69	Not available	Unknown
15	Garcia, G	117	Not available	1-11-1967

5.0 Methodology:

Figure 1 pictorially describes the methodology herein presented for model optimality determination. The process starts by computing the OLS based on the selected parametric model and evaluating the residuals obtained. From the obtained residuals, the various test statistics as earlier discussed are carried out in sequence all to determine model optimality.

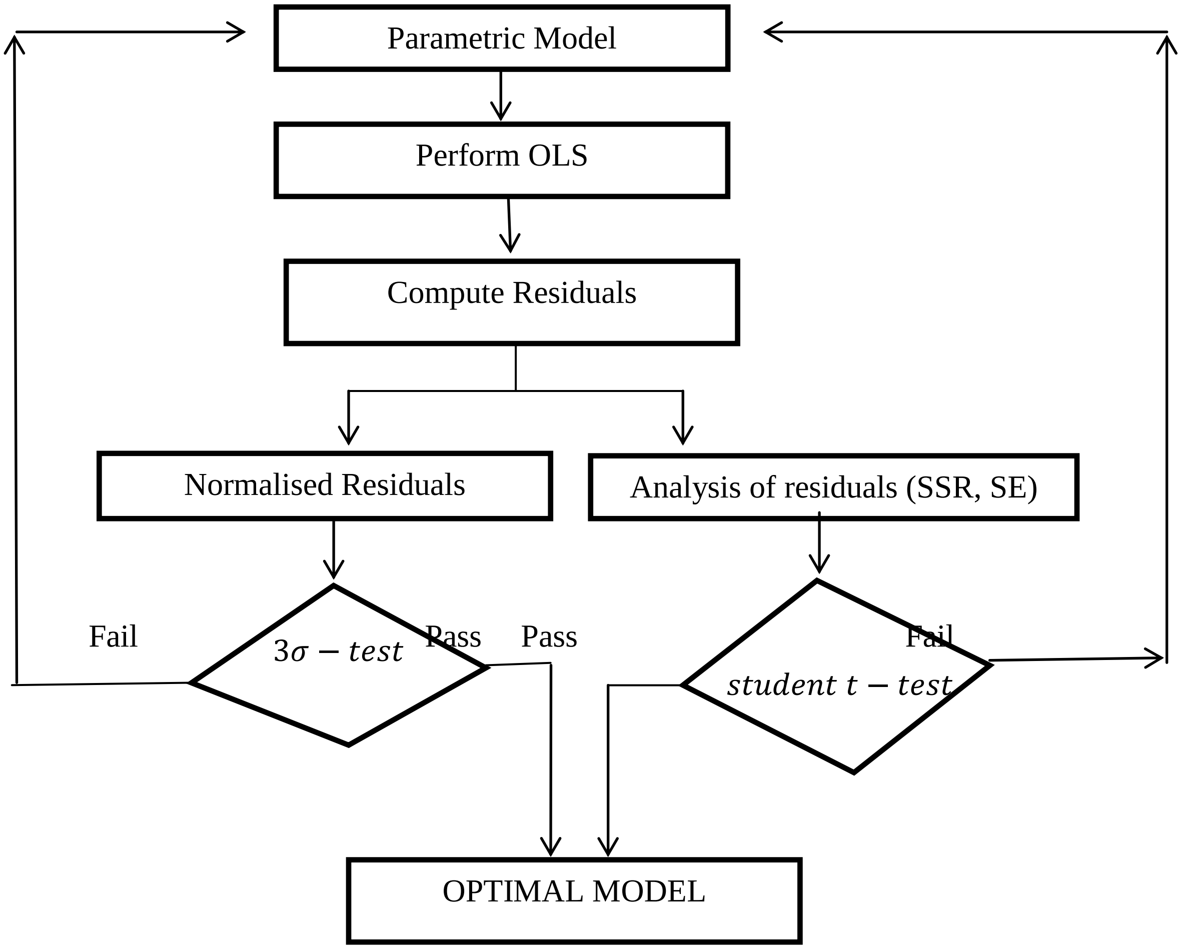

Figure 1: Flow chart of model execution methodology.

Models that indicate fewer outliers based on the 3σ-test are considered suitable since they have effectively fitted well into most of the data used. But the 3σ-test cannot be used as the only basis for suitability determination. Further tests are therefore before a decision can be taken. The further tests are done by analyzing the residuals using the sum of squares of residuals (SSR) and standard error (SE). Amongst the two analysis the sum of squares of residuals is more important since the principal objective of a least squares is to minimize the sum of squares of weighted residuals.

After the residuals have been analysed the student-t test is performed to test the hypothesis statistically. The t distribution provides a good way to perform one sample tests on the mean when the population variance is not known provided the population is normal or the sample is sufficiently large so that the Central Limit Theorem applies.

Therefore, if the P-value falls within the significance level (i.e$P \leq \alpha$ (0.05)) we do not reject the null hypothesis. Similarly, when the t critical value is greater than the t-observed i.e ($t_{crit} \geq t_{obs}$), the null hypothesis is sustained that the mean of residuals is less than zero therefore the chosen parametric model is optimal.

6.0 Results and discussion of results

Presented in table 2 is a summary of results obtained from each model. The least squares adjustment was performed using both the weighted and unweighted models and the results are as presented in table 2.

Also results obtained from the t-test for each model is presented in table3. It is seen that although model 5 gave the least standard error and sum of squares of residuals which makes it appear to be the most suitable model, a look at the normality of the residuals indicate that the model residuals are not normally distributed but rather lump-sided. This non-normality is further seen in the hypothesis test of the model which rather than give the highest probability gave an average probability of 50% conformity of the data with the null hypothesis.

From table 2, we also observe that the more the number of outliers detected (based on the 3σ-test) in a model, the lesser the suitability of the chosen model because the outlier reduce the number of data points fitted into the model. However, this alone is not sufficient to take a decision about a model's suitability since model 5 though with most suitable SSR and minimal number of outliers the residual distribution and its percentage probability (as seen in table 3) suggest that it is not the most optimal model.

Table 2: Analysis of residuals

	Min Resi	**Max Resi**	**Range_res**	**Distr**	<= 3σ	> 3σ	**Outliers**	**Stddev**	**Σv**	**Σv^2**	**Std Error**	**RMS**
					Norm_resi							
Model equation 1												
OLS	-37.25	29.05	66.30	Normal	191	0	0	12.54	2.80E-10	29562.92	12.54	12.44
Weighted LS	-40.48	29.21	69.69	Normal	117	74	74	0.94	-100.90	31089.94	12.86	12.76
Model equation 2												
OLS	-42.54	26.24	68.78	Normal	191	0	0	13.09	-1.68	32047.49	13.06	12.95
Weighted LS	-50.77	23.77	74.54	Normal	131	60	60	1.02	-571.66	36141.74	13.87	13.76
Model equation 3												
OLS	-46.72	25.02	71.74	Normal	191	0	0	12.42	0.00	28982.49	12.42	12.32
Weighted LS	-53.43	21.32	74.75	Normal	131	60	60	0.99	-653.99	32729.05	13.19	13.09
Model equation 4												
OLS	-35.76	6.40	42.16	Normal	191	0	0	12.41	-0.0076	28811.5576	12.41	12.28
Weighted LS	-53.30	19.61	72.91	Normal	130	61	61	1.01	-694.68	34924.6365	13.67	13.52
Model equation 5												
OLS	8.88 E -15	4.26 E-14	3.11E-14	Lump sided	191	0	0	2.89 E -14	0.00	0.00	0.00	0.00
Weighted LS	-1.95 E-14	3.55 E-15	2.31E-14	Lump sided	190	1	1	7.03 E -16	0.00	0.00	0.00	0.00
Model equation 6 (5 parameters)												
OLS	-18.13	32.97	51.10	Normal	190	1	1	8.33	2.600	12584.372	8.23	8.12
Weighted LS	-27.69	38.73	66.42	Normal	123	68	68	0.70	-64.650	14657.986	8.88	8.76
Model equation 7 (7 parameters model)												
OLS	-16.85	22.41	39.26	Normal	190	1	1	6.25	-52.58	7181.48	6.25	6.13
Weighted LS	-16.42	21.65	38.07	Normal	122	69	69	0.56	-55.24	7235.25	6.27	6.15
Model equation 8												
OLS	-22.37	27.63	50.00	Normal	191	0	0	9.33	0.00	16267.81	9.33	9.23
Weighted LS	-24.18	27.56	51.74	Normal	118	73	73	0.73	-31.91	16425.19	9.37	9.27

Table 3: Result of hypothesis test

Model No	8	7	6	5	4	3	2	1
Count	191	191	191	191	191	191	191	191
Mean	-0.61	-0.15	0.34	0.01	0.39	0.07	0.51	0.73
Stddev	12.70	12.41	12.94	12.82	12.40	12.52	12.53	12.52
Std error	0.91	0.89	0.93	0.92	0.89	0.90	0.90	0.90
Hypothesis mean	0	0	0	0	0	0	0	0
alpha	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
tails	1	1	1	1	1	1	1	1
df	190	190	190	190	190	190	190	190
t-stat	-0.67	-0.17	0.37	0.01	0.43	0.07	0.57	0.81
p value	0.75	0.57	0.36	0.50	0.33	0.47	0.29	0.21
t-critical	1.65	1.65	1.65	1.65	1.65	1.65	1.65	1.65
Significance	ACCEPT	ACCEPT	ACCEPT	ACCEPT	ACCEPT	ACCEPT	ACCEPT	ACCEPT

A glance at table 3 suggests that the model 7 despite the significant number of outliers it identifies. Comparing the results of model 3 on both tables, we see that the SSR of the model is the next most suitable after model 5 (which is the principal objective function of a least squares). Furthermore, it shows a 50% probability of statistical suitability which is next to the highest P value of 75%.

Model 8 is seen to show the statistically most significant result but from previous analysis does not seem to conform to the factors that make for model suitability. We therefore can see that the statistical significance test alone does not suffice to determine model suitability hence stressing the importance of further testing.

Model 7 which though rank second best fit in terms of analysis of residuals and also statistical test is considered most suitable. This is because the model maintains consistently suitable results for both the analysis of residuals and the test statistics. It is therefore chosen as the most suitable model describing the gravity data distribution across the study area and thus adopted as the chosen parametric model for the least squares collocation (LSC).

7.0 Conclusion and Recommendation

This work presents a simple and easily implementable method for confirming model suitability to be used in LSC. The method incorporates the use of analysis of model residuals with statistical t-tests in checking mathematical models. It is concluded that integration of the residual analysis with the statistical test provides model users a more scientific and rationale means of judging model suitability than mere dependence on only the statistical test.

References

Dermanis A and Rossikopoulos, D.(1991). Statistical Inference in Integrated Geodesy. IUGG XXth General Assembly, International Association of Geodesy, Vienna, August 11-24, 1991.

Guo, D. M and Xu, H, Z. (2015). Application of variance component estimation to calibrate geoid error models.*Springer Plus* (2015) 4:434.

Koch, K. R. (1999): Parameter Estimation and Hypothesis Testing in Linear Models,Second edition, Springer-Verlag, Berlin Heidelberg, Germany.

Lehmann, R 2012, 'Improved critical values for extreme normalized and studentized residuals in gauss-markov models', *Journal of Geodesy*, vol. 86, issue 12, pp.1137-1146.

Lehmann, R. (2013). “On the formulation of the alternative hypothesis for geodetic outlier detection.” *Journal of Geodesy*, 87(4), 373–386.

Nafisi, V. (2003): Detection of systematic errors by variance components. *Survey Review,* vol. 37, no. 288, pp. 155-161.

Odumosu, J. O, Musa, A. A., Onuigbo I. C and Ojigi, L, M. (2017). Crossover Analysis of BGI gravity datasets across Nigeria. *Nigerian Journal of Geodesy*. Volume 1, No. 1, 105 – 112. Special Issue.

Persson, C. G. (1981): On the estimation of variance components in linear models – withspecial reference to geodetic applications. Royal Institute of Technology, Division ofGeodesy, Stockholm.

Ruffhead, A. (1987). An introduction to least -squares collocation, Survey Review, 29 (224), 85-94.

Vaníček, P,Craymer, M. R andKrakiwsky, E. J (2001). 'Robustness analysis of geodetic horizontal networks', *Journal of Geodesy*, vol. 75, issue 4, pp. 199-209.