NOAA Technical Report NOS NGS 30

# On Data Snooping and Multiple Outlier Testing

Rockville, Md.
1984

**U.S. DEPARTMENT OF COMMERCE**
**National Oceanic and Atmospheric Administration**
National Ocean Service

NOAA Technical Publications

National Ocean Service/National Geodetic Survey
Subseries

Geodetic Survey (NGS) of the National Ocean Service (NOS), Office d Geodetic Services, NOAA, establishes and maintains the basic national horizontal and vertical networks of geodetic control and provides Government-wide leadership in the improvement of geodetic surveying methods and instrumentation, coordinates operations to assure network development, and provides specifications and criteria for survey operations by Federal, State, and other agencies.

NGS engages in research and development for the improvement of knowledge of the figure of the Earth and its gravity field, and has the responsibility to procure geodetic data from all sources, process these data, and make them generally available to users through a central data base.

NOAA geodetic publications and relevant geodetic publications of the former U.S. Coast and Geodetic Survey are sold in paper form by the National Geodetic Information Center. To obtain a price list or to place an order, contact:

National Geodetic Information Center (N/CG17x2)
Charting and Geodetic Services
National Ocean Service
National Oceanic and Atmospheric Administration
Rockville, MD 20852

When placing an order, make check or money order payable to: National Geodetic Survey. Do not send cash or stamps.

Publications can also be purchased over the counter at the National Geodetic Information Center, 11400 Rockville Pike, Room 14, Rockville, Md. (Do not send correspondence to this address.)

An excellent reference source for all Government publications is the National Depository Library Program, a network of about 1,300 designated libraries. Requests for borrowing Depository Library material may be made through your local library. A free listing of libraries in this system is available from the Library Division, U.S. Government Printing Office, 5236 Eisenhower Ave., Alexandria, VA 22304 (703-557-9013).

NOAA Technical Report NOS NGS 30

# On Data Snooping and Multiple Outlier Testing

Johan J. Kok

National Geodetic Survey
Rockville, Md.
1984

**U. S. DEPARTMENT OF COMMERCE**
**Malcolm Baldrige, Secretary**

**National Oceanic and Atmospheric Administration**
John V. Byrne, Administrator

National Ocean Service
Paul M. Wolff, Assistant Administrator

Charting and Geodetic Services
R. Adm. John D. Bossler, Director



For sale by the National Geodetic Information Center, NOAA, Rockville, Md. 20852

CONTENTS

# ON DATA SNOOPING AND MULTIPLE OUTLIER TESTING

Johan J. Kok*

National Geodetic Survey
Charting and Geodetic Services
National Ocean Service, NOAA
Rockville, Md. 20852

ABSTRACT. Data snooping, using Studentized or un-Studentized one-dimensional conventional hypothesis tests, is described as a special case of general multivariate linear hypothesis tests. The derivation of test statistics is based on the ideas of Allen J. Pope. Special attention is given to the concepts of internal and external reliability of networks, as defined by W. Baarda, for un-Studentized data snooping. Finally a heuristic procedure for the detection of multiple (or simultaneous) outliers in one adjustment, called "iterated data snooping," is described. The results of an experiment with simulated errors in the observations are given.

## 1. INTRODUCTION

Three types of tests that are in use for the detection of isolated blunders (outliers) in the observations for geodetic networks can be divided into Studentized tests, using either Student's t-distribution or the tau-distribution, and in un-Studentized tests, using the standard normal distribution of Gauss-Laplace. Allen J. Pope showed, during a series of seminars at the National Geodetic Survey (NGS) (Pope 1982), that the three types of tests mentioned can all be derived algebraically from general multivariate linear hypothesis tests as found in, e.g., Hamilton (1964), Graybill (1976), and Koch (1980).

Section 2 of this report generally follows Pope's derivation. Section 3 gives a short review of the concepts of internal and external reliability of networks. These were defined by Baarda (1968, 1977) as measures of the quality, as well as the optimization of design of geodetic networks with respect to the possible presence of outliers in the observations.

---

*Prepared during a grant period (June 1981 through May 1982) while serving as a Visiting Senior Scientist in Geodesy, National Research Council, National Academy of Sciences, Washington, D.C.

Permanent address: Delft University of Technology, Department of Geodesy, Thijsseweg 11, Delft, The Netherlands.

As the derivation of the tests includes multidimensional hypothesis tests for simultaneous outliers of unspecified relative size, this has been the start of the idea of iterated data snooping, described in section 4. Iterated data snooping is an attempt to overcome the difficulties encountered when more than one outlier is present at the same time. It is known that, because the residuals in a least-squares adjustment are not robust with respect to outliers in the observations, one outlier can mask others. Also, an outlier may cause the rejection of good observations, especially in un-Studentized outlier tests where the rejection limit (critical value for the test) is not affected by the estimated variance of unit weight $\hat{\sigma}_o^2$.

In iterated data snooping we try to make the residuals (and consequently the statistics for data snooping tests) more robust by subsequent removal of suspected observations using the sequential adjustment technique (backwards). As will be shown, this can be done without recomputing or updating the triangular (Cholesky) factors of the normal matrix, so information about possible simultaneous outliers can be generated during a single adjustment computation.

## 2. MULTIVARIATE LINEAR HYPOTHESIS TESTS

### 2.1. The Use of Variance Ratios

Consider an adjustment of an overdetermined geodetic network by the method of variation of parameters. The (linearized) observations equations are

$$A\Delta X = \Delta L + V, \tag{1}$$

where $\Delta L = L - L\{X^o\}$; ("observed" minus "computed" observations")

$r$ = degrees of freedom, or redundancy,

$\Sigma_1 = Q_1 . \sigma_o^2$; $L \sim N[\mu, Q_1 . \sigma_o^2]$.

A least-squares adjustment based on the observation equations has the solution

$$\hat{X} = X^o + \Delta\hat{X} = X^o + (A^t PA)^{-1} A^t P\Delta L; \quad \text{with } P = Q_1^{-1}. \tag{2}$$

Its covariance matrix is

$$\Sigma_{\hat{x}} = (A^tPA)^{-1} \cdot \sigma_o^2 \qquad \text{or} \qquad \hat{\Sigma}_{\hat{x}} = (A^tPA)^{-1} \cdot \hat{\sigma}_o^2 \,, \tag{3}$$

with $\hat{\sigma}_o^2 = \frac{V^tPV}{r}$ as the unbiased estimate of $\sigma_o^2$.

Adjusted observations are

$$\hat{L} = L^o + \Delta\hat{L} = L^o + A\Delta\hat{X} \tag{4}$$

and least-squares residuals to the observations

$$V = \hat{L} - L = \Delta\hat{L} - \Delta L. \tag{5}$$

Sometimes $\Sigma_l$ is a diagonal matrix, and consequently also $P = Q_l^{-1}$, but this is not necessary. Normally distributed observations are not required for the least-squares adjustment, but this is assumed here for the purpose of hypothesis testing.

A constrained least-squares adjustment of the same problem, which is introduced for the purpose of testing general linear hypotheses concerning the parameters (unknowns) or functions thereof, has observation equations

$$A\Delta X = \Delta L + V_c \tag{6a}$$

$$C^t\Delta X = L_c \tag{6b}$$

where matrix $C^t$ is of full rank c.

The degrees of freedom are now f = r + c, with

r = degrees of freedom, unconstrained adjustment,

c = number of constraints.

From the theory of the linear model (Hamilton 1964, Graybill 1976) we have the following quantities and their distributions after applying the least-squares adjustment. The unconditional sum of least-squares of residuals of the unconstrained adjustment (1) is

$$V^tPV \sim \chi_r^2 \cdot \sigma_c^2, \tag{7}$$

and the additional sum of squares due to the constraints in the constrained adjustment based on eqs. (6a) and (6b) is

$$(V_c^t P V_c - V^t P V) \sim \chi_c^2 \cdot \sigma_o^2 \tag{8}$$

It can be shown that the sums of squares $V^tPV$ and $(V_c^tPV_c - V^tPV)$ are independently distributed as $\chi^2$.

Also we have the identity

$$(V_c^t P V_c - V^t P V) = (C^t \Delta\hat{X} - L_c)^t \cdot (C^t Q_{\hat{X}} C)^{-1} \cdot (C^t \Delta\hat{X} - L_c). \tag{9}$$

Then, using $\hat{\sigma}_o^2 = (V^tPV)/r$, we have the variance ratio

$$T = \frac{(C^t \Delta\hat{X} - L_c)^t \cdot (C^t Q_{\hat{X}} C)^{-1} \cdot (C^t \Delta\hat{X} - L_c)}{c \cdot \hat{\sigma}_o^2} = \frac{V_c^t P V_c - V^t P V}{V^t P V} \cdot \frac{r}{c} \,. \tag{10}$$

With eqs. (7) and (8), we see that T is distributed as $F_{c,r}$, denoted as

$$T = \frac{V_c^t P V_c - V^t P V}{V^t P V} \cdot \frac{r}{c} \sim \frac{\chi_c^2 / c}{\chi_r^2 / r} \equiv F_{c,r} \,. \tag{11}$$

In case the variance factor $\sigma_o^2$ is known, the estimate $\hat{\sigma}_o^2 = (V^tPV)/r$ in eqs. (10) and (11) can be replaced by $\sigma_o^2$, and we get

$$T' = \frac{(C^t \Delta\hat{X} - L_c)^t \cdot (C^t Q_{\hat{X}} C)^{-1} \cdot (C^t \Delta\hat{X} - L_c)}{c \cdot \sigma_o^2} = \frac{V_c^t P V_c - V^t P V}{c \cdot \sigma_o^2} \sim \frac{\chi_c^2}{c} \equiv F_{c,\infty} \tag{12}$$

We will refer to tests based on the statistic of eq. (11) as the "Studentized" case and to that of eq. (12) as the "un-Studentized" case.

Statistics of type T and T' in eqs. (11) and (12) can be used for testing the hypothesis that a parameter (unknown) or a combination or function of parameters in the adjustment are in agreement with known values. This is done by choice and specification of the constraint equations (6b). In adjustments of geodetic networks we rarely have this knowledge (e.g., theoretical values) of the parameters.

However, the apparatus of general linear hypothesis tests can also be used for hypotheses on parameters that are not originally included in the adjustment model, but instead are introduced after the adjustment as additional parameters. In the following this possibility is worked out. It is included for reference, because a similar strategy is followed for outlier testing, described in section 2.2. Let us now denote the observation equations for an unconstrained adjustment of a geodetic network by

$$A\Delta X_1 = \Delta L + V \ . \tag{13}$$

For reference purposes a least-squares adjustment based on this system of equations will be called the "nominal" adjustment, with r degrees of freedom, it has a vector of residuals denoted by $V_{(nominal)}$.

For the case that model errors are supposed to be present in the observations, the model is extended with additional parameters $X_2$ for such unknowns

$$A\Delta X_1 + B\Delta X_2 = \Delta L + V \tag{14}$$

with degrees of freedom $f = r - c$, when c is the number of additional parameters $X_2$. The constrained adjustment for this enhanced adjustment model has observation equations

$$A\Delta X_1 + B\Delta X_2 = \Delta L + V_c \tag{15a}$$

$$C_1^t \Delta X_1 + C_2^t \Delta X_2 = L_c . \tag{15b}$$

A valid choice of a set of explicit absolute constraints on the parameters $X_2$ is, e.g.,

$$C^t = (C_1^t \ C_2^t) = (0 \ I_c) \text{ and } L_c = 0,$$

where

$$I_c = (\varepsilon_i \ \varepsilon_j \ \varepsilon_k \ldots\ldots) \text{ with } \varepsilon_i = \text{i-th unit vector, etc.}$$

The consequence of this specific choice is that the parameters $X_2$, just added to the adjustment, are at the same time constrained to zero. The results of the constrained (enhanced) adjustment, using eqs. (15a, 15b), will be the same as those of the nominal adjustment, using (13) on its own.

Consequently, the degrees of freedom of the constrained enhanced adjustment equal that of the nominal adjustment $f' = (r-c) + c = r$, and the residuals $V_c \equiv V_n$.

If $C_2^t X_2 - L_c \overset{def}{=} \mu_c$, the hypotheses will generally have the form

- null hypothesis $H_o$: $\mu_c = 0$;

- alternative hypothesis $H_a$: $\mu_c \neq 0$.

Such hypotheses do not specify anything about the functional model (design matrix A) or the stochastic model (covariance matrix $\Sigma_1 = Q_1 . \sigma_o^2$), so it is assumed that these are correct. The variance factor, however, may be known or unknown, depending upon the type of test that will be used: Studentized or un-Studentized. Studentization involves only $\sigma_o^2$, so for both types of tests the relative size of the variances described by $Q_1$ must be known. The dimension of the hypothesis is determined by the number of additional parameters, and it equals the rank of the matrices B and $C^t$. Usually the hypotheses describe additional small physical effects, which were not included in the model of the nominal adjustment, but possibly are present. If only one parameter $X_2$ is used to describe the hypothesis, then $C^t$ is a lxu vector and B is a nxl vector, with rank 1. Consequently, the hypothesis in this case is one-dimensional. If the hypothesis is described by use of c additional parameters $X_2$, then $C^t$ is a cxu matrix, and B a nxc matrix, both with rank c. The hypothesis is then c-dimensional.

For testing hypotheses, to be formulated in the enhanced adjustment model, we use the statistic of eq. (10):

$$T_1 = \frac{(C^t \Delta\hat{X} - L_c)^t . (C^t Q_{\hat{X}} C)^{-1} . (C^t \Delta\hat{X} - L_c)}{c.\hat{\sigma}_o^2} \sim F_{c,f}, \tag{16}$$

where

$$C^t = (0 \; I_c) \text{ and } L_c = 0$$

$$\hat{\sigma}_o^2 = \frac{V^t PV}{f}; \quad f = r - c = \text{degrees of freedom, unconstrained adjustment.}$$

Then

$$T_1 = \frac{(C^t\Delta\hat{X} - L_c)^t\ (C^tQ_{\hat{X}}C^{-1}(C^t\Delta\hat{X} - L_c)}{V^tPV} \cdot \frac{f}{c} = \frac{\Delta\hat{X}_2{}^t Q_{\hat{X}_2}^{-1}\Delta\hat{X}_2}{V^tPV}\ \frac{f}{c} \sim F_{c,f} \tag{17}$$

because $C^t\Delta\hat{X} = \begin{pmatrix} 0 & I_c \end{pmatrix} \begin{pmatrix} \Delta\hat{X}_1 \\ \Delta\hat{X}_2 \end{pmatrix} = \Delta\hat{X}_2$ and $L_c = 0$,

also

$$(C^tQ_{\hat{X}}C) = \begin{pmatrix} 0 & I_c \end{pmatrix} \begin{pmatrix} Q_{\hat{X}_1} & Q_{\hat{X}_{12}} \\ Q_{\hat{X}_{12}} & Q_{\hat{X}_2} \end{pmatrix} \begin{pmatrix} 0 \\ I_c \end{pmatrix} = Q_{\hat{X}_2}.$$

Now consider that

$$N = \begin{pmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{pmatrix} = \begin{pmatrix} A^tPA & A^tPB \\ B^tPA & B^tPB \end{pmatrix}$$

then

$$\begin{aligned} Q_{\hat{X}_2} &= (N^{-1})_{22} = \bar{N}_{22}^{-1} = (N_{22} - N_{21}\ N_{11}^{-1}\ N_{12})^{-1} \\ &= (B^t(P-PA(A^tPA)^{-1}A^tP)B)^{-1} \\ &= (B^tPQ_{V_c}\ PB)^{-1} \end{aligned}$$

or $$Q_{\hat{X}_2} = (B^tQ_{\bar{V}_c}\ B)^{-1}, \tag{18}$$

where $\bar{V}_c = PV_c$ and $Q_{\bar{V}_c} = PQ_{V_c}\ P = P-PA\ (A^tPA)^{-1}A^tP.$

Also $$U = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} = \begin{pmatrix} A^tPL \\ B^tPL \end{pmatrix}$$

then

$$\bar{U}_2 = U_2 - N_{21} N_{11}^{-1} U_1$$

$$= B^t PL - B^t PA(A^t PA)^{-1} A^t PL$$

$$= B^t P(L - A(A^t PA)^{-1} A^t PL)$$

$$= -B^t PV_c$$

or $$\bar{U}_2 = -B^t \bar{V}_c . \qquad (19)$$

Finally we have

$$\Delta\hat{X}_2 = (N^{-1})_{22} . \bar{U}_2 = (B^t Q_{\bar{V}_c} B)^{-1} (- B^t \bar{V}_c) . \qquad (20)$$

Now eq. (17) becomes, with use of eqs. (18) and (20):

$$T_1 = \frac{\Delta\hat{X}_2^t . Q_{\hat{X}_2} . \Delta\hat{X}_2}{V^t PV} . \frac{f}{c} = \frac{\bar{V}_c^t B(B^t Q_{\bar{V}_c} B)^{-1} (B^t Q_{\bar{V}_c} B) (B^t Q_{\bar{V}_c} B)^{-1} B^t \bar{V}_c}{V^t PV} . \frac{f}{c}$$

or

$$T_1 = \frac{\bar{V}_c^t B(B^t Q_{\bar{V}_c} B)^{-1} B^t \bar{V}_c}{V^t PV} . \frac{f}{c} \sim F_{c,f} . \qquad (21)$$

By use of the identity (9) eq. (17) can be written as

$$T_1 = \frac{V_c^t PV_c - V^t PV}{V^t PV} . \frac{f}{c} \sim F_{c,f}. \qquad (22)$$

so it follows then from eqs. (21) and (22) that

$$V_c^t PV_c - V^t PV = \bar{V}_c^t B(B^t Q_{\bar{V}_c} B)^{-1} B^t \bar{V}_c \qquad (23a)$$

or

$$V^t PV = V_c^t PV_c - \bar{V}_c^t B(B^t Q_{\bar{V}_c} B)^{-1} B^t \bar{V}_c \qquad (23b)$$

Then the test statistic $T_1$ can also be written as

$$T_1 = \frac{\bar{V}_c^t B(B^t Q_{\bar{V}_c} B)^{-1} B^t \bar{V}_c}{V_c^t P V_c - \bar{V}_c^t B(B^t Q_{\bar{V}_c} B)^{-1} B^t \bar{V}_c} \cdot \frac{f}{c} \sim F_{c,f} \, . \qquad (24)$$

This last equation is very useful, because all quantities $\bar{V}_c$ and $Q_{\bar{V}_c}$ can be obtained from the nominal adjustment with

$$\bar{V}_c = PV_c = PV_n \qquad \text{and} \qquad Q_{\bar{V}_c} = PQ_{V_c} P = PQ_{V_n} P \, ,$$

and matrix B can be specified without actually performing the enhanced adjustment itself.

In Pope (1982) the statistic $T_2$ is introduced, which is defined as

$$T_2 = \frac{\bar{V}_c^t B(B^t Q_{\bar{V}_c} B)^{-1} B^t \bar{V}_c}{V_c^t P V_c} \cdot \frac{r}{c} \sim \tau_{c,r} \, . \qquad (25)$$

The distribution $\tau_{c,r}$ is "new" and not known, but it can be expressed as a function of the F-distribution, because $T_2$ can be written as a function of $T_1$ as follows

$$T_2 = \frac{\bar{V}_c^t B(B^t Q_{\bar{V}_c} B)^{-1} B^t \bar{V}_c}{V_c^t P V_c} \cdot \frac{r}{c} = \frac{V^t P V}{V_c^t P V_c} \cdot \frac{r}{f} \cdot \frac{V_c^t P V_c - V^t P V}{V^t P V} \cdot \frac{f}{c} \, ,$$

or with use of eq. (22):

$$T_2 = \frac{V^t P V}{V_c^t P V_c} \cdot \frac{r}{f} \cdot T_1 = \frac{1}{1 + \frac{V_c^t P V_c - V^t P V}{V^t P V}} \cdot \frac{r}{f} \cdot T_1 \, .$$

Then $$T_2 = \frac{1}{1 + \frac{c}{f} \cdot T_1} \cdot \frac{r}{f} \cdot T_1 = \frac{r . T_1}{f + c . T_1} \sim \frac{r . F_{c,f}}{f + c . F_{c,f}} \, .$$

Thus

$$T_2 = \frac{\bar{V}_c^t B (B^t Q_{\bar{V}_c} B)^{-1} B^t \bar{V}_c}{V_c^t P V_c} \cdot \frac{r}{c} \sim \frac{r.F_{c,f}}{f + c.\ F_{c,f}} = \tau_{c,r} \ . \qquad (26)$$

If the variance factor $\sigma_0^2$ is known, we may write eq. (24) as

$$T_3 = \frac{\bar{V}_c^t B (B^t Q_{\bar{V}_c} B)^{-1} B^t \bar{V}_c}{c.\sigma_0^2} \sim \frac{\chi_c^2}{c} \ (\equiv F_{c,\infty}) \ , \qquad (27)$$

which is the un-Studentized statistic for general linear hypotheses in the enhanced adjustment model, where $T_1$ in eq. (24) and $T_2$ in eq. (25) are Studentized ones. In all three cases there are several possibilities for the choice and specification of $B.\Delta X_2$. In all cases, however, the matrix $(B^t Q_{\bar{V}_c} B)$ must be regular and invertible. The rank of the total matrix $Q_{\bar{V}_c} = P - PA(A^tPA)^{-1} A^tP$ is $r = n-p$, where n is the number of observations and p is the number of independent unknowns in the nominal adjustment (or the rank of $A^tPA$). Consequently the rank of $(B^t Q_{\bar{V}_c} B)$ can maximally be $r = n-p$. If $(B^t Q_{\bar{V}_c} B)$ is singular, then the (local) redundancy is insuffìcent to test the hypothesis specified.

For one-dimensional hypotheses, where $C^t$ is a 1 x u vector and B is a n x 1 vector, the statistics $T_1$, $T_2$ and $T_3$ are (dropping subscript c):

$$T_1 = \frac{\bar{V}^t b (b^t Q_{\bar{V}} b)^{-1} b^t \bar{V}}{V^t PV - \bar{V}^t b (b^t Q_{\bar{V}} b)^{-1} b^t \bar{V}} \cdot (r-1) \sim F_{1,r-1} \qquad (28)$$

$$T_2 = \frac{\bar{V}^t b (b^t Q_{\bar{V}} b)^{-1} b^t \bar{V}}{V^t PV} \cdot r \sim \tau_{1,r} \qquad (29)$$

$$T_3 = \frac{\bar{V}^t b (b^t Q_{\bar{V}} b)^{-1} b^t \bar{V}}{\sigma_0^2} \sim \frac{\chi_1^2}{1} \ (\equiv F_{1,\infty}) \qquad (30)$$

Considering that

$F_{1,r-1}$ is the square of Student's $t_{r-1}$;

$\tau^2_{1,r}$ is the square of univariate tau $\tau_r$; and

$F_{1,\infty} = \chi^2_1$ is the square of standard-normal $n[0,1]$,

one can also write

$$t_1 = \frac{-(b^t Q_{\bar{v}} b)^{-\frac{1}{2}} b^t \bar{V}}{(V^t PV - \bar{V}^t b(b^t Q_{\bar{v}} b)^{-1} b^t \bar{V})^{\frac{1}{2}}} \cdot \sqrt{r-1} \sim t_{r-1} \tag{31}$$

$$t_2 = \frac{-b^t \bar{V}}{\hat{\sigma}_o \cdot (b^t Q_{\bar{v}} b)^{\frac{1}{2}}} \sim \tau_r \tag{32}$$

$$t_3 = \frac{-b^t \bar{V}}{\sigma_o \cdot (b^t Q_{\bar{v}} b)^{\frac{1}{2}}} \sim n\,[0,1] \tag{33}$$

For eq. (32) see Pope (1976), and for eq. (33) see Baarda (1968).

One-dimensional hypotheses of the type

$H_o$: $\Delta X_2 = 0$

$H_a$: $\Delta X_2 \neq 0$,

tested by use of statistic $t_3$ in eq. (33), play an important role in Baarda's B-method of testing. By use of chosen levels of significance $\alpha$ and power $1.-\beta$, quantification of reliability of networks is based on the same type of hypotheses and tests. This is reviewed in section 3. A geometrical derivation and interpretation of the one-dimensional un-Studentized test can be found in Baarda (1968) and Van Mierlo (1981). An illustration of the applicability of the general one-dimensional linear hypothesis tests to deformation analysis is given in Kok (1981, 1982), where the vector b in the hypothesis is used to describe a systematic pattern of changes in coordinates due to deformation.

## 2.2. Tests for the Detection of Outliers (Data Snooping)

For the purpose of testing for outliers in the observations in a least-squares adjustment the enhanced model of section 2.1 and linear hypotheses tests can be used as follows. For this purpose the enhanced model of which we are thinking is no longer one which contains additional small physical effects, but instead is one which is enhanced by the addition of c blunder parameters for a selected set of c observations.

Vector $\Delta X_2$ is the vector of (unknown) blunder parameters, each parameter representing the blunder in one of the observations. We then have

$$H_o: \Delta X_2 = 0 \text{ (null hypothesis)}$$

$$H_a: \Delta X_2 \neq 0 \text{ (alternative hypothesis)}.$$

If the alternative hypothesis is an assumption of c outliers of unspecified relative sizes being present in the vector of observations, then each column of B assigns the effect of a single blunder parameter to the appropriate observation. In this case the c columns of B will be unit vectors. If the hypothesis involves knowledge about the relative size of the errors, (e.g., a systematic disturbance of observations) then the columns of B will not be unit vectors.

Suppose a hypothesis involves c different outliers of unknown sizes. Then matrix B is

$$B = \{b_i \; b_j \; \ldots.\} = \{\varepsilon_i \; \varepsilon_j \; \ldots.\} = I_c.$$

Each column of B (unit vector) assigns the effect of the particular blunder parameter to just one observation, $\varepsilon_i$ to observation $L_i$, $\varepsilon_j$ to observation $L_j$, and so forth. The hypothesis is c-dimensional, because the rank of B is c, and there are c blunder parameters $\Delta X_2$.

Test statistics of eqs. (24), (25), and (27) become respectively

$$T_1 = \frac{\bar{V}^t I_c (I_c{}^t Q_{\bar{V}} I_c)^{-1} I_c{}^t \bar{V}}{V^t PV - \bar{V}^t I_c (I_c{}^t Q_{\bar{V}} I_c)^{-1} I_c{}^t \bar{V}} \cdot \frac{r-c}{c} \sim F_{c,r-c} \quad (34)$$

$$T_2 = \frac{\bar{V}^t I_c (I_c{}^t Q_{\bar{V}} I_c)^{-1} I_c{}^t \bar{V}}{V^t PV} \cdot \frac{r}{c} \sim \tau^2_{c,r} \quad (35)$$

$$T_3 = \frac{\bar{V}^t I_c (I_c{}^t Q_{\bar{V}} I_c)^{-1} I_c{}^t \bar{V}}{c.\sigma_o^2} \sim \frac{\chi^2_c}{c} \quad (36)$$

where $\bar{V} = PV$ and $Q_{\bar{V}} = PQ_V P = P - PA(A^t PA)^{-1} A^t P$; and $I_c{}^t \bar{V}$ and $(I_c{}^t Q_{\bar{V}} I_c)$ are selected elements of $\bar{V}$ and $Q_{\bar{V}}$, related to those observations that are involved in the hypothesis. All these quantities can be computed from a nominal adjustment. Forming hypotheses involving such simultaneous possible outliers is generally not easy because of the lack of available information.

A special application, however, is given in Kok (1982). This is the testing of the coordinates of known points in the case of densification networks and deformation networks (in one-, two-, and three-dimensional coordinate systems), one at a time. Usually this is done by one-dimensional tests, where each coordinate successively is considered to be possibly an outlier. Using statistics from eqs. (34), (35), or (36), we can involve in each hypothesis the coordinates of one point, which will be one (x), two (x, y), or three (x, y, z) respectively. For one-dimensional coordinate systems (e.g., leveling, gravity) these tests coincide with standard data snooping, but for two- or three-dimensional coordinate systems the hypothesis for each point is two- or three-dimensional. This seems to be a more realistic strategy than testing each coordinate separately.

Baarda (1968) introduced "conventional" alternative hypotheses, where only one outlier at a time is assumed to be present. Each conventional hypothesis is then one-dimensional and the vector $b = \varepsilon_i$, a unit vector corresponding to that observation involved in the hypothesis. If we form these conventional hypotheses for all observations successively, this results in a set of n conventional $H_{a_i}$, each of them being one-dimensional. Testing these $H_{a_i}$ consecutively is called "a data snooping strategy."

We have $H_{o_i}: \quad \Delta X_2 = 0; \qquad (i = 1, \ldots, n)$

$H_{a_i}: \quad \Delta X_2 \neq 0.$

Test statistics for one-dimensional hypotheses of eqs. (31), (32), and (33) then become

$$t_1 = \frac{-(b_i{}^t Q_{\bar{V}} b_i)^{-\frac{1}{2}} b_i{}^t \bar{V}}{(V^t PV - \bar{V}^t b_i (b_i{}^t Q_{\bar{V}} b_i)^{-1} b_i{}^t \bar{V})^{\frac{1}{2}}} \cdot \sqrt{r-1} = \frac{-Q_{\bar{V}_i}^{-\frac{1}{2}} \bar{V}_i}{(V^t PV - \bar{V}_i^t Q_{\bar{V}_i}^{-1} \bar{V}_i)^{\frac{1}{2}}} \cdot \sqrt{r-1}$$

or

$$t_1 = \frac{-\bar{\bar{V}}_i}{(V^t PV - \bar{\bar{V}}_i^2)^{\frac{1}{2}}} \sqrt{r-1} \sim t_{r-1} \tag{37}$$

where $\bar{\bar{V}}_i = \dfrac{\bar{V}_i}{Q_{\bar{V}_i}^{\frac{1}{2}}}$

$$t_2 = \frac{-b_i{}^t \bar{V}}{\hat{\sigma}_o \cdot (b_i{}^t Q_{\bar{V}} b_i)^{\frac{1}{2}}} = \frac{-\bar{V}_i}{\hat{\sigma}_o \cdot Q_{\bar{V}_i}^{\frac{1}{2}}} = \frac{-\bar{V}_i}{\hat{\sigma}_{\bar{V}_i}} \sim \tau_r \tag{38}$$

$$t_3 = \frac{-b_i{}^t \bar{V}}{\sigma_o \cdot (b_i{}^t Q_{\bar{V}} b_i)^{\frac{1}{2}}} = \frac{-\bar{V}_i}{\sigma_o \cdot Q_{\bar{V}_i}^{\frac{1}{2}}} = \frac{-\bar{V}_i}{\sigma_{\bar{V}_i}} \sim n\,[0,\, 1] \tag{39}$$

These are the test statistics for either Studentized data snooping, using Student's t-distribution (Heck, 1980) or the tau-distribution (Pope, 1976), or for un-Studentized data-snooping, using the standard normal distribution (Baarda, 1968). The data snooping tests are then for each observation, respectively:

$$\text{if} \quad t_1 = \frac{|\bar{\bar{V}}_i|}{(V^t PV - \bar{\bar{V}}_i^2)^{\frac{1}{2}}} \cdot \sqrt{r-1} > t_{r-1;1-\alpha} \quad \text{then reject } H_{o_i} \tag{40a}$$

$$\text{if} \quad t_2 = \frac{|\bar{V}_i|}{\hat{\sigma}_{\bar{V}_i}} > \tau_{r;\,1-\alpha} \quad \text{then reject } H_{o_i} \tag{40b}$$

$$\text{if} \quad t_3 = \frac{|\bar{V}_i|}{\sigma_{\bar{v}_i}} > x_{1-\alpha} \text{ then reject } H_{o_i} \tag{40c}$$

If $H_{o_i}$ is rejected in favor of $H_{a_i}$, this means that at least the i-th observation $L_i$ has to be checked for a possible gross error.

In the most uncommon case of uncorrelated observations, where P is a diagonal matrix, statistics $t_1$, $t_2$, and $t_3$ are often replaced by

$$t_1' = \frac{-V_i}{(V^t PV - \tilde{V}_i^2)^{\frac{1}{2}}} \sqrt{r-1} \sim t_{r-1}; \tag{41}$$

where $\tilde{V}_i = Q_{v_i}^{\frac{1}{2}} . V_i$,

$$t_2' = \frac{-V_i}{\hat{\sigma}_{v_i}} \sim \tau_r; \tag{42}$$

$$t_3' = \frac{-V_i}{\sigma_{v_i}} \sim n[0, 1]; \tag{43}$$

which are slightly easier to compute.

### 2.3. Least-Squares Estimates for Outliers

To test for the presence of outliers (using one- or multidimensional hypotheses), matrix B in the enhanced adjustment model of section 2.1:

$$A\Delta X_1 + B\Delta X_2 = L + V,$$

is built up by c unit vectors, where c is the number of possible (simultaneous) outliers involved in the hypothesis, and consequently the dimension of the hypothesis and of the test. Using the least-squares solution of eq. (20):

$$\Delta\hat{X}_2 = (B^t Q_{\bar{V}} B)^{-1} (-B^t \bar{V}),$$

we can compute least-squares estimates of outliers from the results of the nominal adjustment. A vector of blunder estimates $\hat{\nabla L}$ can be found in general by

$$\hat{\nabla L} = B\Delta\hat{X}_2 = B(B^t Q_{\bar{V}} B)^{-1}(- B^t\bar{V}). \tag{44}$$

Considering that B is a n x c matrix of unit vectors, this is

$$\hat{\nabla L} = I_c\Delta\hat{X}_2 = I_c(I_c{}^t Q_{\bar{V}} I_c)^{-1}(- I_c{}^t\bar{V}), \tag{44a}$$

where $(I_c{}^t Q_{\bar{V}} I_c)$ is a selected c x c submatrix of

$$Q_{\bar{V}} = PQ_V P = P - PA(A^t PA)^{-1} A^t P.$$

In the common case of one-dimensional conventional hypotheses, B is a n x 1 unit vector $b_i = \varepsilon_i$, and we have

$$\hat{\nabla L} = \varepsilon_i\, Q_{\bar{V}_i}^{-1}(- \bar{V}_i) = \varepsilon_i \frac{-\bar{V}_i}{Q_{\bar{V}_i}}. \tag{45}$$

In this case the vector $\hat{\nabla L}$ has only one nonzero element $\hat{\nabla L}_i$, which can be written as the least-squares estimate of a single blunder in L

$$\hat{\nabla L}_i = \frac{-V_i}{Q_{\bar{V}_i}}. \tag{46}$$

Since eqs. (44), (45), and (46) do not include $\sigma_o^2$ or $\hat{\sigma}_o^2$, the blunder estimates are independent of the type of data snooping test applied.

For the least-squares estimate of a blunder in observation $L_i$, Förstner (1979) gives

$$\hat{\nabla L}_i = \frac{-V_i}{r_i}$$

where $r_i = (Q_V P)_i$, which is called "Redundanz-Anzahl"

Considering only uncorrelated observations (P is diagonal), then formula (46) yields the same

$$\hat{\nabla}L_i = \frac{-\bar{V}_i}{Q_{\bar{v}_i}} = \frac{-(PV)_i}{(PQ_vP)_i} = \frac{-V_i}{(Q_vP)_i} = \frac{-V_i}{r_i} \ . \tag{47}$$

## 3. RELIABILITY OF NETWORKS

### 3.1. The Concept

When applying tests for the detection of outliers to a network adjustment, the choice of the level of significance of the test ($\alpha$) determines the probability for the occurrence of type-I errors, i.e., the rejection of the null hypothesis $H_o$ when it is true. By the choice of $\alpha$ we control the probability that remeasurements or investigations for error sources will be performed when there is no error present.

Ideally we also wish to have control over the type-II errors, i.e., the acceptance of the null hypothesis $H_o$ when the alternative hypothesis $H_a$ is true. Probability of making type-II errors ($\beta$) is computable for a specified size of a model error, and it establishes the power of each test ($1-\beta$). The power describes the probability that model errors of specific size will be detected when present.

The concept of reliability of networks as defined by Baarda (1960, 1968, 1977) can be described as a quantification and analysis of the probabilities of type-II errors, when applying tests of one-dimensional conventional hypotheses (data snooping) to the observations in a network.

By fixing the power of all one-dimensional data snooping tests on a chosen level (usually $1-\beta_o = 0.80$), the size of so-called "marginally detectable errors" can be computed for each of the conventional $H_{a_i}$, i.e., for each of the observations contributing to the adjustment. Essential for the B-method of testing is that we use the same level of power ($1-\beta_o$) for one-dimensional tests as for the "global" test of the variance of unit weight $\hat{\sigma}_o^2$, by use of the statistic

$$T = \frac{V^tPV}{r.\sigma_o^2} = \frac{\hat{\sigma}_o^2}{\sigma_o^2} \sim \frac{\chi_r^2}{r} \ . \tag{48}$$

Also it is essential that the size of $\lambda$ (non-centrality parameter) be fixed at a common level. As a result of these choices the levels of significance of the one-dimensional data snooping tests $\alpha_o$ and that of the r-dimensional test of the variance of unit weight $\alpha$ become interdependent.

The sizes of marginally detectable errors, computed for all observations, are measures of the capability of the network to detect outliers of that size in an observation with a probability (power) of $1-\beta_o$. They constitute the "internal reliability" of the network. The influence of each of the marginally detectable errors on the unknowns in the adjustment (usually only on the coordinates) is called the "external reliability." For a given network configuration (geometry) and stochastic model (covariance matrix of the observations), the reliability is computable, but it can then only be controlled by changing the level of significance $\alpha_o$, which means also changing the control over type-I errors.

Changing the network configuration and/or the stochastic model (choice of instruments, measurement procedures), however, will influence the type-II error control, because reliability changes. Consequently, reliability of networks can be an important tool in the process of optimal network design. Baarda developed the concept of reliability of networks using the un-Studentized conventional hypothesis tests of eq. (40c). So far it is not available for Studentized types of data snooping in an operational form. The review of computation of reliability in sections 3.2 and 3.3 is thus restricted to the un-Studentized case.

## 3.2. Internal Reliability

For one-dimensional un-Studentized tests in data snooping, we have eq. (40c), which is applied successively for each conventional alternative hypothesis

$$H_{a_i} : \quad \tilde{\nabla} L = B \Delta X_2 = \varepsilon_i \Delta X_2 \neq 0,$$

with statistic

$$w_i = t_3 = \frac{-\bar{V}_i}{\sigma_{\bar{V}_i}} \sim n[0,1].$$

The following tests are made:

If $|w_i| > x_{1-\alpha_o}$ then reject $H_o$.

In figure 1 the test is depicted for a certain value of w.

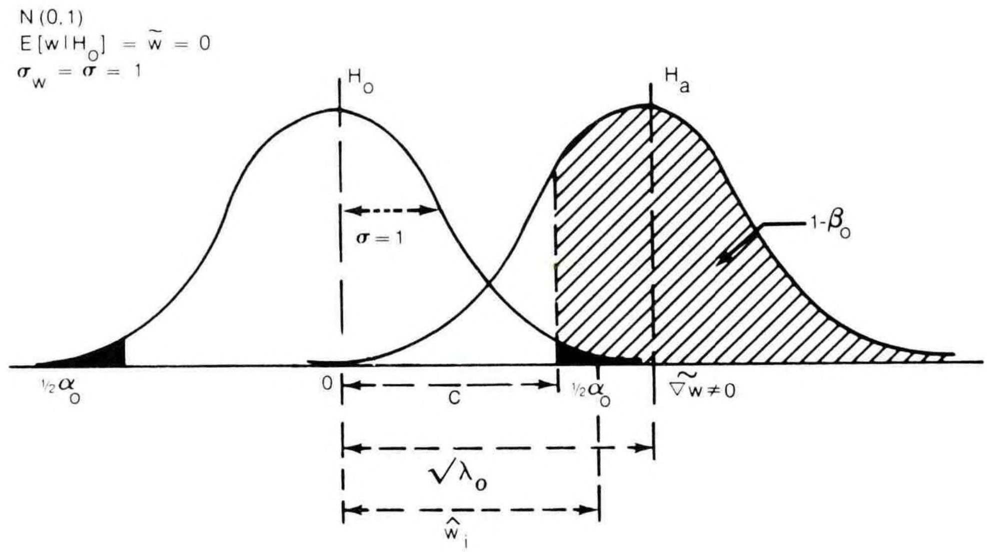


Figure 1.--Un-Studentized w-test.

If the $H_{a_i}$ is true (observation $L_i$ is an outlier $\tilde{L}_i + \nabla\tilde{L}_i$), then we have the non-central normal distribution, indicated in figure 1 under $H_a$. The normal distribution is shifted over a distance $\sqrt{\lambda_o}$, which is called the non-centrality parameter. Under the null hypothesis the mathematical expectation of $w_i$ is zero

$$E\{w_i | H_o\} = 0.$$

Under the alternative hypothesis this is

$$E\{w_i | H_a\} = \nabla\tilde{w} = \sqrt{\lambda}$$

If $\lambda$ is known, then the power of the test is determined, for a certain level of significance $\alpha_o$, by

$$P\{w_i > c | H_a\} = 1-\beta = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-(c+\sqrt{\lambda})} e^{-\frac{1}{2}x^2} dx + \frac{1}{\sqrt{2\pi}} \int_{-(\sqrt{\lambda}-c)}^{+\infty} e^{-\frac{1}{2}x^2} dx \tag{49}$$

The first integral in eq. (49) is that part of the "left tail" of the normal distribution under $H_a$, that falls in the left rejection region. For large values of $1-\beta$ it is very small, and in practice it is usually neglected. The second integral in eq. (49) can be written as

$$\frac{1}{\sqrt{2\pi}}\int_{-(\sqrt{\lambda}-c)}^{+\infty} e^{-\frac{1}{2}x^2}\,dx = \frac{1}{\sqrt{2\pi}}\int_{-(\sqrt{\lambda}-c)}^{0} e^{-\frac{1}{2}x^2}\,dx + \frac{1}{\sqrt{2\pi}}\int_{0}^{+\infty} e^{-\frac{1}{2}x^2}\,dx$$

or

$$\frac{1}{\sqrt{2\pi}}\int_{-(\sqrt{\lambda}-c)}^{+\infty} e^{-\frac{1}{2}x^2}\,dx = \frac{1}{\sqrt{2\pi}}\int_{-(\sqrt{\lambda}-c)}^{0} e^{-\frac{1}{2}x^2}\,dx + 0.5. \qquad (50)$$

We have then for the power

$$1-\beta = 0.5 + \frac{1}{\sqrt{2\pi}}\int_{(\sqrt{\lambda}-c)}^{0} e^{-\frac{1}{2}x^2}\,dx. \qquad (51)$$

Specifying a certain level of errors $|\nabla L_i| = |\nabla L|$ for all observations of the same type, e.g., as a function of their standard deviation, would make it possible to compute $(1-\beta)_i$ for each $H_{a_i}$ and it would give an impression of internal reliability in terms of probabilities. For such computations (involving as many integrals of eq. (51) as there are observations in the adjustment), it is necessary to standardize the error level $|\nabla L|$ for each $H_{a_i}$ by

$$\sqrt{\lambda}_i = |\nabla L| \cdot \sqrt{\frac{Q_{\bar{v}_i}}{\sigma_o^2}}\,, \qquad (52)$$

where $\bar{v}_i = (PV)_i$ and $Q_{\bar{v}_i} = (PQ_vP)_{ii} = (P-PA(A^tPA)^{-1}A^tP)_{ii}$. If values $(1-\beta)_i$ are of the same order of magnitude for all observations, then the internal reliability is homogeneous. Otherwise it is not.

In practice the process is turned around by choosing a common level of power $(1-\beta_o)$ for all tests, usually $1-\beta_o = 0.80$. Then eq. (51) becomes

$$\frac{1}{\sqrt{2\pi}} \int_{-(\sqrt{\lambda_0}-c)}^{0} e^{-\frac{1}{2}x^2} dx = 0.30 . \tag{53}$$

In this formula $\lambda_0$ is the only unknown quantity, because the critical value c is a function of $\alpha_0$ and n [0, 1] only, so $\sqrt{\lambda_0}$ can be computed by integration. Because of the common level $(1-\beta_0)$ and the use of standardized statistics, $\sqrt{\lambda_0}$ has to be computed only once for the whole adjustment.

Instead of computing the power of each individual one-dimensional data snooping test, one can now compute the magnitudes of blunders that can be detected with the same probability $(1-\beta_0)$. Considering eq. (46) for the least-squares estimate of a blunder in the i-th observation

$$\hat{\nabla}L_i = \frac{-\bar{v}_i}{Q_{\bar{v}_i}}$$

and using

$$\hat{w}_i = \frac{-\bar{v}_i}{\sigma_0 \cdot Q_{\bar{v}_i}^{\frac{1}{2}}} \tag{54}$$

or

$$- \bar{v}_i = \hat{w}_i \cdot Q_{\bar{v}_i}^{\frac{1}{2}} \cdot \sigma_0^2, \tag{54a}$$

we can also write

$$\hat{\nabla}L_i = \frac{\hat{w}_i \cdot Q_{\bar{v}_i}^{\frac{1}{2}} \cdot \sigma_0}{Q_{\bar{v}_i}} = \frac{\hat{w}_i \cdot \sigma_0}{Q_{\bar{v}_i}^{\frac{1}{2}}}. \tag{55}$$

Equation (55) can be considered as a transformation of the standardized statistic w to the sample space.

Replacing the "distance" $\hat{w}_i$ by $\sqrt{\lambda_o}$ (see fig. 1) yields

$$\tilde{\nabla L} = \varepsilon_i \cdot \sqrt{\frac{\lambda_o}{Q_{\bar{v}_i}}} \cdot \sigma_o \tag{56a}$$

or

$$\frac{\tilde{\nabla L}}{\sigma_o} = \varepsilon_i \cdot \sqrt{\frac{\lambda_o}{Q_{\bar{v}_i}}} \, , \tag{56b}$$

which is given in Baarda (1968) as

$$\left(\frac{\tilde{\nabla x}^i}{\sigma}\right) = (c_p^i) \cdot \sqrt{\frac{\lambda_o}{n_p}} \tag{56c}$$

where $n_p = (c_p^i)^t . \ (P-PAN^{-1}A^tP)(c_p^i)$, and $\tilde{\nabla x}^i = (\tilde{\nabla L})_i$. In vector $\frac{\nabla L}{\sigma_o}$ of eq. (56) only the i-th component is nonzero because of $\varepsilon_i$ being the i-th unit vector. So we obtain

$$\left|\frac{\nabla \tilde{L}_i}{\sigma_o}\right| = \sqrt{\frac{\lambda_o}{Q_{\bar{v}_i}}} \tag{57}$$

which is defined as the marginally detectable error of observation $L_i$. These marginally detectable errors can be computed for all the observations contributing to the adjustment (redundant observations), to provide measures of internal reliability. In other words:

> "A blunder in observation $L_i$ of size $|\tilde{\nabla L}_i|$ will cause $|w_i| > x_{1-\alpha_o}$ (rejection of $H_{o_i}$) with the specific probability $(1-\beta_o)$, under the assumption that the other redundant observations are free of blunders."

Since the marginally detectable errors $|\nabla\tilde{L}_i|$ in eq. (57) do not depend on the vector of observations, nor on the residuals, they can be computed as soon as the configuration of the network and the stochastic model is known.

The relation between the power of the tests and marginally detectable errors is illustrated in figure 2, for a certain convention $H_{a_i}$. From this figure it can

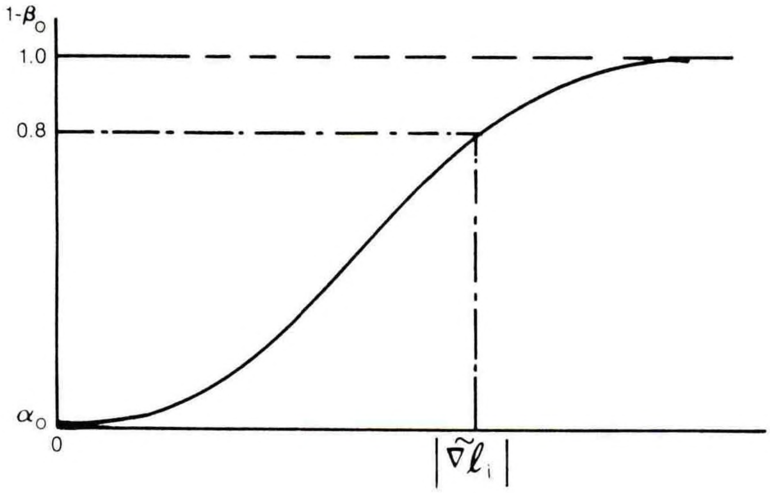


Figure 2.--Power of an individual w-test.

be seen that errors smaller than $|\tilde{\nabla L}_i|$ have a decreasing probability of discovery, and that for errors larger than $|\tilde{\nabla L}_i|$ this probability increases.

The definition of internal reliability of networks, using conventional $H_a$ and marginally detectable errors, was first introduced by Baarda (1968) as a feature of his B-method of testing. Essential for the B-method is the interdependency of the levels of significance for these one-dimensional data snooping tests and the r-dimensional test of the variance of unit weight $\hat{\sigma}_o^2$. By use of the statistic

$$T = \frac{v^t P v}{r \cdot \sigma_o^2} = \frac{\hat{\sigma}_o^2}{\sigma_o^2} \sim F_{r,\infty} \quad (= \frac{\chi_r^2}{r}),$$

the test of $\hat{\sigma}_o^2$ is the following:

$$\text{If} \qquad T = \frac{\hat{\sigma}_o^2}{\sigma_o^2} > F_{r,\infty;1-\alpha}, \text{ then reject } H_o. \tag{58}$$

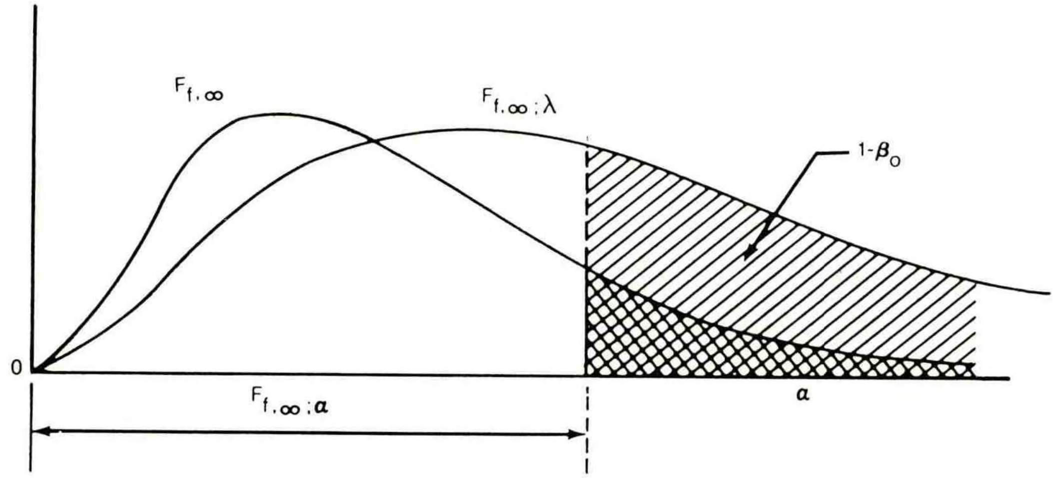


Figure 3.--Central and non-central F-distribution.

Applying the same power $(1-\beta_0)$ as was used for the data snooping tests, and fixing the non-centrality parameter $\lambda$ at the same level as $\lambda_0$ of the one-dimensional case, establishes a relation between $\alpha_0$ (one-dimensional) and $\alpha$ (r-dimensional) levels of significance. This is symbolically denoted as

$$\lambda_0 = \lambda(\alpha,\ \beta_0,\ r,\ \infty) = \lambda(\alpha_0,\ \beta_0,\ 1,\ \infty). \tag{59}$$

When the power $(1-\beta_0)$ is chosen (usually $1-\beta_0 = 0.80$), then $\alpha$ and $\alpha_0$ are the only free variables in eq. (59). Having chosen one of them, the other can be computed through $\lambda = \lambda_0$. By using this procedure a single outlier has the same probability of discovery (rejection of $H_0$) in both types of tests.

In case the tests with statistics in eq. (36),

$$T_3 = \frac{\bar{V}^t I_c (I_c^t Q_{\bar{V}} I_c)^{-1} I_c^t \bar{V}}{c.\sigma_0^2} \sim \frac{\chi_c^2}{c},$$

are used for testing c-dimensional ("conventional") hypotheses concerning the coordinates of known stations treated by weighted constraints in an adjustment of densification networks or of deformation networks, reliability regions can be computed (see Kok 1982). These are established by the invariant quadratic form

$$\bar{Q} = (I_c^t Q_{\bar{V}} I_c)^{-1} .\ \lambda_0 .\ \sigma_0^2 , \tag{60}$$

where

$$\lambda_o = \lambda(\alpha_o, \beta_o, 1, \infty) = \lambda(\alpha, \beta_o, r, \infty) = \lambda(\alpha_c, \beta_o, c, \infty) \text{ and}$$

$(I_c^t Q_{\bar{v}} I_c)$ is a selected (c x c) submatrix of $Q_{\bar{v}}$.

If the network (coordinate system) is one-dimensional, the test and consequently the reliability region coincide with ( c = 1) the square of the one-dimensional data snooping test and its marginally detectable error:

$$\text{If } w^2 = \frac{\bar{v}_i^t \cdot Q_{\bar{v}_i}^{-1} \cdot \bar{v}_i}{\sigma_o^2} > F_{1,\infty;1-\alpha_o} \quad \text{then reject } H_{o_i}, \tag{61}$$

and

$$\bar{Q} = \frac{\lambda_o}{Q_{\bar{v}_i}} \cdot \sigma_o^2 = \tilde{\nabla L}_i^2 . \tag{62}$$

For two-dimensional networks (c = 2) the test of each known station is as follows:

$$\text{If } T_3 = \frac{\bar{v}^t I_c (I_c^t Q_{\bar{v}} I_c)^{-1} I_c \bar{v}}{2 \cdot \sigma_o^2} > F_{2,\infty;1-\alpha_2} \quad \text{then reject } H_{o_i}, \tag{63}$$

with reliability region

$$\bar{Q} = (I_c^t Q_{\bar{v}} I_c)^{-1} \cdot \lambda_o \cdot \sigma_o^2,$$

which is a 2 x 2 submatrix (inverted) of $Q_{\bar{v}}$, and it establishes an ellipse describing the sensitivity of the test with respect to a disturbance of the station in all possible directions. In the case of three-dimensional networks (c = 3) we use the test:

$$\text{If } T_3 = \frac{\bar{v}^t I_c (I_c^t Q_{\bar{v}} I_c)^{-1} I_c \bar{v}}{3 \cdot \sigma_o^2} > F_{3,\infty;1-\alpha_3} \quad \text{then reject } H_{o_i} \tag{64}$$

with reliability region

$$\bar{Q} = (I_c^t Q_{\bar{v}} I_c)^{-1} \cdot \lambda_o \cdot \sigma_o^2,$$

now being a 3 x 3 submatrix (inverted) of $Q_{\bar{v}}$ establishing an ellipsoid with the same meaning as the ellipses for two-dimensional cases.

### 3.3. External Reliability

Internal reliability is described by the errors that are marginally detectable with a specific probability. It is used during the design of networks and it predicts the effectiveness of the tests that will be used during the adjustment. Usually, however, the final results of network adjustments are not adjusted observations, but coordinates in a one-, two-, or three-dimensional coordinate system. A good and homogeneous internal reliability does not automatically guarantee reliable coordinates. Generally the influence of each observation - and thus of each error in an observation - on the coordinates is different. The influence that each of the marginally detectable errors $|\tilde{\nabla}L_i|$ has on all coordinates of the network is known as external reliability.

The vector of marginally detectable errors $\tilde{\nabla}L_{(i)}$ for one conventional $H_{a_i}$, computed by use of eq. (56a), is

$$\tilde{\nabla}L_{(i)} = \varepsilon_i \sqrt{\frac{\lambda_o}{Q_{\bar{v}_i}}} \cdot \sigma_o. \tag{65}$$

Let $L' = L + \tilde{\nabla}L$ be a vector of observations under $H_{a_i}$, then the vector of differences between solutions

$$\hat{X}' = (A^tPA)^{-1}A^tP(L + \tilde{\nabla}L_{(i)}) \text{ and } \hat{X} = (A^tPA)^{-1}A^tPL$$

is

$$\nabla\hat{X}_{(i)} = \hat{X}' - \hat{X} = (A^tPA)^{-1}A^tP(L + \tilde{\nabla}L_{(i)}) - (A^tPA)^{-1}A^tPL,$$

or

$$\nabla\hat{X}_{(i)} = (A^tPA)^{-1}A^tP\,\tilde{\nabla}L_{(i)} = (A^tPA)^{-1}A^tP\,\varepsilon_i\,(\sqrt{\frac{\lambda_o}{Q_{\bar{v}_i}}}) \cdot \sigma^o \tag{66}$$

This is the general formula for computation of one vector $\nabla\hat{X}$ of influences of marginally detectable errors $\tilde{\nabla L}$, under hypothesis $H_{a_i}$, on all unknowns. Since there are n (= number of observations) hypotheses $H_{a_i}$ possible, then n vectors $\nabla\hat{X}$ must be evaluated. If part of the unknowns are not coordinates, but "nuisance parameters" such as orientation unknowns and scale factors. the total system of normal equations can be partitioned as

$$\begin{pmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{pmatrix} \cdot \begin{pmatrix} \hat{X}_1 \\ \hat{X}_2 \end{pmatrix} = \begin{pmatrix} A^t_{11} \\ A^t_{12} \end{pmatrix} \cdot PL \ , \tag{67}$$

where $\hat{X}_1$ are the coordinates and $\hat{X}_2$ the nuisance parameters.

We now have for each $H_{a_i}$:

$$\begin{pmatrix} \nabla\hat{X}_1 \\ \nabla\hat{X}_2 \end{pmatrix} = \begin{pmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{pmatrix}^{-1} \cdot \begin{pmatrix} A^t_{11} \\ A^t_{12} \end{pmatrix} \cdot P.\tilde{\nabla L}_{(i)} \tag{68}$$

The influence of $\tilde{\nabla L}_{(i)}$ on coordinates only is

$$\nabla\hat{X}_1 = \begin{pmatrix} M_{11} & M_{12} \end{pmatrix} \cdot \begin{pmatrix} A^t_{11} \\ A^t_{12} \end{pmatrix} \cdot P.\tilde{\nabla L}_{(i)} \tag{69}$$

where

$$\begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} = \begin{pmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{pmatrix}^{-1}$$

The computation of n vectors $\nabla\hat{X}_1$, to depict the external reliability of a network, is a large computational effect, and for networks of substantial size it is difficult to judge this information properly. Another disadvantage is the fact that vectors $\nabla\hat{X}_1$ are dependent on the coordinate definition, which is determined by the actual choice of minimum constraints (datum) that has been made.

Baarda (1977) introduced a distortion parameter $\bar{\lambda}_i^{1/2}$, which is the length of the normalized vector $\nabla X_{(i)}$, computed by

$$\bar{\lambda}_t^{1/2} = (\nabla\hat{X}^t_{(i)} \cdot Q^{-1}_{\hat{x}} \cdot \nabla\hat{X}_{(i)})^{1/2} \cdot \frac{1}{\sigma_o} \tag{70}$$

where $Q^{-1}_{\hat{x}} = (N^{-1})^{-1} = A^tPA$. For the influence on coordinates only, we use

$$\bar{\lambda}_i^{1/2} = (\nabla\hat{X}^t_1 \cdot Q^{-1}_{\hat{x}_1} \nabla\hat{X}_1)^{1/2} \cdot \frac{1}{\sigma_o} \tag{71}$$

where $Q^{-1}_{\hat{x}_1} = (N^{-1})^{-1}_{11} = \bar{N}_{11} = N_{11} - N_{12} N^{-1}_{22} N_{21}$. The amount of information depicting the external reliability is now reduced to n parameters $\sqrt{\bar{\lambda}_i}$, one for each hypothesis $H_{a_i}$. If these parameters are of the same order of magnitude, then the network is homogeneous with respect to external reliability.

Because eq. (71) still necessitates the computation of all vectors $\nabla\hat{X}_1$ first, this formula is not recommended for actual computation. A more computer-oriented formula is

$$\bar{\lambda}_i = (\tilde{\nabla}L^t_{(i)} \cdot Q^{-1}_{\ell} \cdot \tilde{\nabla}L_{(i)}) \cdot \frac{1}{\sigma_o^2} - \lambda_o - (\tilde{\nabla}Y^t_{(i)} \cdot N^{-1}_{22} \cdot \tilde{\nabla}Y_{(i)}) \cdot \frac{1}{\sigma_o^2} \tag{72}$$

where $\tilde{\nabla}Y_{(i)} = A^tP\,\tilde{\nabla}L_{(i)}$;   $\lambda_o = \lambda(\alpha_o, \beta_o, 1, \infty)$ and $Q^{-1}_{\ell} = P$.

In cases where there are no noise parameters, but instead all unknowns are coordinates (e.g., leveling networks), we simply use

$$\bar{\lambda}_i = (\tilde{\nabla}L^t_{(i)} \cdot Q^{-1}_{\ell} \cdot \tilde{\nabla}L_{(i)}) \cdot \frac{1}{\sigma_o^2} - \lambda_o, \tag{73}$$

which does not even require the (partial) inverse of the normal equation matrix.

For certain purposes it may be useful to compute influences of marginally detectable errors on functions of the unknowns. This may also be considered a form of external reliability. For example, one can compute the influence on the adjusted observations by

$$\nabla\hat{L}_{(i)} = A(A^tPA)^{-1}A^tP \,.\, \tilde{\nabla}L_{(i)}, \tag{74}$$

or even on the residuals to the observations by

$$\nabla V_{(i)} = (A(A^tPA)^{-1}A^tP - I).\ \tilde{\nabla}L_{(i)}. \tag{75}$$

Replacing vector $\tilde{\nabla}L_{(i)}$ in eq. (75) by the vector of least-squares error estimates $\hat{\nabla}L_{(i)}$, for the same hypothesis $H_{a_i}$, computed by use of eq. (45) yields the following:

$$\hat{\nabla}V_{(i)} = (A(A^tPA)^{-1}A^tP - I).\ \hat{\nabla}L_{(i)}. \tag{76}$$

Because the vector $\hat{\nabla}L_{(i)}$, computed by

$$\hat{\nabla}L_{(i)} = \varepsilon_i . \frac{-\bar{V}_i}{Q_{\bar{V}_i}} \quad ,$$

has only one nonzero component $|\hat{\nabla}L_i|$ and this is the least-squares estimate of an isolated blunder in observation $L_i$, eq. (76) gives us the influence of a single error estimate on all the residuals in the adjustment. Quantities $\hat{\nabla}V$ will be used in section 4.2 for iterated data snooping.

## 4. MULTIPLE OUTLIER TESTING

### 4.1. Some Difficulties of Data Snooping

It is known that in least-squares adjustments one gross error in the observations may mask other gross errors, in the sense that the influence of that error can be such that the residuals to other erroneous observations V, and also $\bar{V} = PV$, do not reveal these other errors. The residuals are not robust with respect to errors in the observations. This is also illustrated by eq. (76), which gives the influence of one (estimate of an) error on the vector of residuals

$$\hat{\nabla}V = (A(A^tPA)^{-1}A^tP - I)\ \hat{\nabla}L_{(i)},$$

as a result of the least-squares adjustment itself.

Others (e.g., Meissl 1980, Fuchs 1981) have successfully located outliers by use of adjustment by minimizing the sum of absolute residuals ($\ell_1$-norm). If we wish to maintain the property of minimum variance of results of the adjustment, however, then a subsequent adjustment by the method of least-squares ($\ell_2$-norm) must still be performed. The so-called "Danish Method" has the effect of "robustizing" the residuals by use of a reweighting strategy (Krarup et al. 1980). This also means the normals have to be formed and solved repeatedly.

A second difficulty with data snooping procedures is the fact that, in practice, the largest statistic does not always coincide with an outlier. This may be caused by incorrect relative weights of observations in spite of all the care taken.

A third difficulty is the fact that a large blunder can easily cause the statistics of other observations to exceed the rejection limit also (critical value of the one-dimensional data snooping tests), especially if un-Studentized tests are used.

At Delft University it has become standard procedure to combine data snooping of observations with testing of condition equations, in order to decide which observation is indeed an outlier. For this purpose a separate adjustment is performed with the method of condition equations. In the case of outliers spatially not very close to each other, it is possible to localize more than one of them in a single adjustment run. It is known from experience that only part of the conditions, in the vicinity of the observations flagged for rejection in data snooping is actually used in the decision process.

The following may be of use for testing individual condition equations or groups of them without using a program for adjustment by condition equations, but instead using the nominal adjustment results of the parametric method. Consider an enhanced constrained adjustment (sec. 2.1) with observation equations

$$A\ X_1 + (QU^t)X_2 = L + V$$

$$C_1 X_1 + C_2\ X_2 = 0.$$

Choose

$$C = (C_1 \; C_2) = (0 \; I_r) \text{ and}$$

U = the coefficient matrix of r independent (linearized) conditions.

Then following the line of thought of section 2.1, we have after least-squares adjustment of the nominal adjustment,

$$\hat{X}_2 = (UQQ_{\bar{V}}QU^t)^{-1}(-UQ\bar{V})$$

or

$$\hat{X}_2 = (UQ(P-AN^{-1}A^tP)QU^t)^{-1}(-UQPV)$$

$$\hat{X}_2 = (U(Q-AN^{-1}A^t)U^t)^{-1}(-UV). \tag{78}$$

Using the identity

$$Q_{\bar{V}} = P-PAN^{-1}A^tP = U^tM^{-1}U,$$

where

$$M = UQU^t,$$

and also using for the misclosures

$$-UV = W,$$

eq. (78) can also be written as

$$\hat{X}_2 = (UQ(U^tM^{-1}U)QU^t)^{-1}W$$

$$\hat{X}_2 = ((UQU^t)M^{-1}(UQU^t))^{-1}W$$

$$\hat{X}_2 = (MM^{-1}M)^{-1}W$$

or

$$\hat{X}_2 = M^{-1} W \quad \text{(correlates)}. \tag{79}$$

Then

$$(QU^t)\ \hat{X}_2 = QU^t M^{-1} W = V. \tag{79a}$$

The statistic $T_3$ for an un-Studentized r-dimensional test becomes

$$T_3 = \frac{\bar{V}^t QU^t (UQ(P-PAN^{-1}A^tP)QU^t)^{-1} UQ\bar{V}}{r \cdot \sigma_o^2} \sim F_{r,\infty},$$

also

$$T_3 = \frac{\bar{V}^t U^t (UQU^t M^{-1} UQU^t)^{-1} UV}{r. \sigma_o^2} = \frac{W^t M^{-1} W}{r. \sigma_o^2} = \frac{\hat{\sigma}_o^2}{\sigma_o^2} \sim F_{r,\infty}. \tag{80}$$

If we use an individual column of $U^t$ in eqs. (77) and following, denoted as $U_i^t$, then

$$\hat{X}_2 = -((U_i QU^t) \quad M^{-1} (UQU_i^t))^{-1} U_i V$$

or

$$\hat{X}_2 = -M_{ii}^{-1} \ U_i V = M_{ii}^{-1} W_i. \tag{81}$$

Equation (79a) yields

$$(QU_i^t)\ \hat{X}_2 = QU_i^t M_{ii}^{-1} W_i = V_{(i)} \tag{82}$$

and $V_{(i)}$ are then the residuals due to the i-th condition only.
Statistic $T_3$ is now

$$T_3 = \frac{V^t U_i^t (U_i QU^t M^{-1} UQU_i^t)^{-1} U_i V}{\sigma_o^2} = \frac{W_i^t M_{ii}^{-1} W_i}{\sigma_o^2} \sim F_{1,\infty} = \chi_1^2 \tag{83}$$

and it can be used to test this individual condition, while actually $T_3$ can be computed from the nominal adjustment (parametric model) by

$$T_3 = \frac{V^t U_i^t (U_i (Q-AN^{-1}A^t) U_i^t)^{-1} U_i V}{\sigma_o^2} \sim F_{1,\infty} = \chi_1^2 \tag{83a}$$

## 4.2. A Heuristic Procedure for Multiple Outlier Detection (Iterated Data Snooping)

Considering eq. (76) for the influence of a single error (least-squares estimate) $\hat{\nabla L}_i$ on the vector of residuals in a nominal adjustment

$$\hat{\nabla V} = (AN^{-1}A^tP - I)\ \varepsilon_i .\ \hat{\nabla L}_i ,$$

We can update the vector of residuals by

$$V' = V - (AN^{-1}A^tP - I)\ \varepsilon_i .\ \hat{\nabla L}_i \tag{84}$$

and

$$\bar{V}' = \bar{V} - (PAN^{-1}A^tP - P)\ \varepsilon_i .\ \hat{\nabla L}_i . \tag{84a}$$

**This means that the estimate of the isolated blunder in the i-th observation is removed, including its influence on all the residuals. Pope pointed out that applying eq. (84) is identical to removing observation $L_i$ by a sequential adjustment technique, and consequently the co-factor matrix of weighted residuals $Q_{\bar{V}}$ should also be updated by a corresponding sequential formula (see Creusen 1965, Whiting and Pope 1976):**

$$Q_{\bar{V}'} = Q_{\bar{V}} - Q_{\bar{V}}\ \varepsilon_i(\varepsilon_i^t Q_{\bar{V}} \varepsilon_i)^{-1}\ \varepsilon_i^t Q_{\bar{V}}$$

or

$$Q_{\bar{V}'} = Q_{\bar{V}} - Q_{\bar{V}}\ \varepsilon_i (Q_{\bar{V}}^{-1})_{ii}\ \varepsilon_i^t Q_{\bar{V}} . \tag{85}$$

Suppose that the estimate $\hat{\nabla L}_i = \hat{\nabla L}_k$ was computed for the observation with $t_3 = |w_k| = |w_i|_{max}$. Then applying eqs. (84a) and (85) allows us to test for the presence of other outliers after observation $L_k$ has been removed from the adjustment. The standardized residual $\bar{V}_k'$ will at that time be zero, as well as the co-factor $Q_{\bar{V}_k}'$, because

$$\bar{V}_k' = \bar{V}_k - \varepsilon_k^t (PAN^{-1}A^tP - P)\varepsilon_k .\ \hat{\nabla L}_k$$

or $$\bar{v}'_k = \bar{v}_k - Q_{\bar{v}_k} \cdot \frac{-\bar{v}_k}{Q_{\bar{v}_k}} = \bar{v}_k - \bar{v}_k = 0,$$

and $$Q_{\bar{v}'_k} = \varepsilon_k^t (Q_{\bar{v}} - Q_{\bar{v}}\, \varepsilon_k (Q_{\bar{v}}^{-1})_{kk} \varepsilon_k^t\, Q_{\bar{v}}) \varepsilon_k$$

or $$Q_{\bar{v}'_k} = Q_{\bar{v}_k} - Q_{\bar{v}_k} (Q_{\bar{v}}^{-1})_{kk}\, Q_{\bar{v}_k} = Q_{\bar{v}_k} - Q_{\bar{v}_k} = 0.$$

Repeating the standard data snooping, using the updated vector of weighted residuals $\bar{v}'$ and the updated cofactor matrix $Q_{\bar{v}'}$ of eqs. (84) and (85), will result in updated statistics w', of which $|w'_k|$ can no longer be the largest, because of $\bar{v}'_k=0$ and $Q_{\bar{v}'_k}=0$. Instead, another observation will now have the largest statistic, say

$$|w'_m| = \frac{|\bar{v}'_m|}{\sigma_o \cdot \sqrt{Q_{\bar{v}'_m}}} = |w_i|_{max} \; ; \; (i = 1, 2, \ldots, n \text{ and } Q_{\bar{v}'_i} \neq 0). \tag{86}$$

Consider observation $L_k$ as the first candidate for being an outlier; then observation $L_m$ is the next candidate, provided that in the updated situation the one-dimensional data snooping test for $L_m$ still gives rejection of $H_o$. During the first data snooping the statistic $w_m$ for observation $L_m$ was not necessarily the second largest of the statistics $w_i$; it even may have been smaller than the critical value for the test, due to the masking effect of the adjustment. We will call the process of repeatedly updating, and testing by standard data snooping subsequently on the updated results, a method of "iterated data snooping." The recursive formulas for such a process are

$$\bar{v}'^{(n+1)} = \bar{v}'^{(n)} - (PAN^{-1}A^tP - I)\, \varepsilon_k \cdot \hat{\nabla}L_k^{(n)} \tag{87}$$

$$Q_{\bar{v}'}^{(n+1)} = Q_{\bar{v}'}^{(n)} - Q_{\bar{v}'}^{(n)} \varepsilon_k \, (Q_{\bar{v}'}^{(n)-1})_{kk}\, \varepsilon_k^t \, Q_{\bar{v}'}^{(n)} \tag{88}$$

where
$$\hat{\nabla}L_k^{(n)} = \varepsilon_k \cdot \frac{-\bar{v}'^{(n)}}{Q_{\bar{v}'}^{(n)}} \quad \text{for } L_k \text{ with } |w_k'^{(n)}| = |w_i'^{(n)}|_{max} \tag{89}$$

and $\quad |w_k'^{(n)}| \geq \sqrt{F_{1,\infty;1-\alpha_o}}$ (90)

For the start of the process (n = o) we have $\bar{V}'^{(n)} = \bar{V}$ and $Q_{\bar{V}'}^{(n)} = Q_{\bar{V}}$.
At each step of this process the maximum test statistic and the estimate of the presumed outlier will not be influenced by masking effects of errors that were removed at preceding steps.

Because each deletion of a (suspected) observation uses one degree of freedom and the rank of matrix $Q_{\bar{V}}$ is r, we can repeat the deleting and updating r times maximally. After r deletions of observations, the remaining set of observations is just sufficient to determine the network without redundancy, and all residuals $\bar{V}_i$ and cofactors $Q_{\bar{V}_i}$ will be zero.

If a fixed level of significance $\alpha_o$ is used, the rejection procedure and updating will be terminated as soon as eq. (90) is no longer satisfied.
At each step of the process a candidate is added to the list of suspected observations. However, at such a step there may be more than one candidate with maximum $w_i'$, and at the same time these candidates can have equal cofactors $Q_{\bar{V}_i'}$.
This is a tie-breaking problem for the procedure, and in order to proceed with the search for other candidates with smaller errors, one of those candidates has to be removed, but this is not necessarily the erroneous one. In fact the (conventional) alternative 1-dimensional hypotheses for these observations cannot be separated, and if one of them is removed by the updating, the others will also get values $\bar{V}_i' = 0$ and $Q_{\bar{V}_i}' = 0$. They should also be added to the list of suspected observations. Also they cannot become candidates in later steps any more, because they become nonredundant ("no check") observations after the updating.

The process, as described above, has the following drawbacks.

(1) Although a list of estimates of the sizes of possible errors is built up by applying eq. (89) at each successive step, previous estimates have already been computed without knowledge about suspects that are found later.

(2) Matrix $Q_{\bar{V}}$ has to be computed completely and also will be updated for all its elements, because of eq. (88); consequently it must be kept in storage as a n x n full matrix.

(3) Good observations may have had $|w'_i|_{max}$ at a certain step, and then been added to the suspect list and removed from the adjustment.

For computational reasons, especially with respect to the second drawback mentioned, iterated data snooping can also be formulated differently. This will result in the same suspect list, and also the first drawback mentioned will no longer exist.

Instead of using the recursive formulas (87) and (88), the updated $\bar{v}'^{(n+1)}$ and $Q_{\bar{v}'}^{(n+1)}$ can also be found by updating the "original" $\bar{v}$ and $Q_{\bar{v}}$ of the nominal adjustment. Each "new" suspect then increases the "conventional" hypothesis by one dimension, supposing c simultaneous outliers of unknown sizes.
We have

$$\hat{\nabla}v = (AN^{-1}A^tP - I)\hat{\nabla}L, \qquad (91)$$

where the vector of error estimates $\hat{\nabla}L$ has been computed from

$$\hat{\nabla}L = I_c\,\hat{x}_2 = I_c(I_c^tQ_{\bar{v}}I_c)^{-1}(-I_c^t\bar{v}), \qquad (92)$$

and $I_c$ is a n x c matrix composed of columns of unit vectors, each referring to an observation having $|w'_k| = |w'_i|_{max}$ at one of the subsequent steps, which are the entities in the list of suspects.

The updated vector of weighted residuals (for c suspects) is now computed by

$$\bar{v}' = \bar{v} - (PAN^{-1}A^tP - P)\hat{\nabla}L, \qquad (93)$$

and the updated co-factor matrix by

$$Q_{\bar{v}}' = Q_{\bar{v}} - Q_{\bar{v}}\,I_c(I_c^tQ_{\bar{v}}I_c)^{-1}I_c^tQ_{\bar{v}}$$

or

$$Q_{\bar{v}}' = Q_{\bar{v}} - Q_{\bar{y}}I_cQ_{\bar{\bar{v}}}^{-1}I_c^tQ_{\bar{v}} \qquad (94)$$

where $Q_{\bar{\bar{v}}}^{-1} = (I_c^tQ_{\bar{v}}I_c)^{-1}$ is a selected c x c submatrix of $Q_{\bar{v}}$, inverted.
Because of the one-dimensional data snooping in each step, applied to the

current updates $\bar{V}'$ and $Q_{\bar{V}}'$, only $(Q_{\bar{V}}')_{ii}$ = diag. $\{Q_{\bar{V}}'\}$ is actually needed. Equation (94) can then be simplified to

$$Q_{\bar{V}_i}' = (Q_{\bar{V}}')_{ii} = (Q_{\bar{V}})_{ii} - (Q_{\bar{V}} I_c Q_{\bar{V}}^{=-1} I_c^t Q_{\bar{V}})_{ii} \tag{95}$$

Equations (93) and (95) can be evaluated efficiently by use of the sparse Cholesky factor of the normal matrix of the nominal adjustment (see sec. 4.4), and storage of matrix $Q_{\bar{V}}$, respectively $Q_{\bar{V}}'$, can be avoided. Now at each step c error estimates are solved simultaneously as additional unknowns in eq. (92), and c suspects are deleted from the nominal results. Consequently the residuals $I_c^t \bar{V}'$ and cofactors $I_c^t Q_{\bar{V}}' I_c$ are zero for those c suspects after the updating. It should be noted that deletion of observations during iterated data snooping is only performed for building a list of suspects and computing estimates of possible outliers. The solution of the nominal adjustment $\hat{X}_1$ itself is not updated explicitly.

In the author's opinion it is "dangerous" to update the solution, thus creating automatic deletion of the suspects from the total adjustment, especially in low-redundant situations as in geodetic networks. It should be left to the user of the adjustment program to decide - after proper investigation of the observational data - which of the entries in the suspect list that is provided by the procedure is actually a blunder.

Figure 4 illustrates the procedure of iterated data snooping, computationally performed by repeatedly updating the nominal adjustment results.
The multi-dimensional test of $\hat{\sigma}_0^2/\sigma_0^2$ has also been incorporated (B-method). The critical value for this test depends on the degrees of freedom in the adjustment. As a consequence of the decrease in the degrees of freedom during successive steps in the iterated data snooping, this critical value is $F_{r-c,\infty;1-\alpha'}$.
The value $\alpha'$ is then computed for each step from the relation

$$\lambda_0' = \lambda(\alpha', \beta_0, r-c, \infty) = \lambda(\alpha_0, \beta_0, 1, \infty),$$

where a constant level of $\alpha_0$ is used.

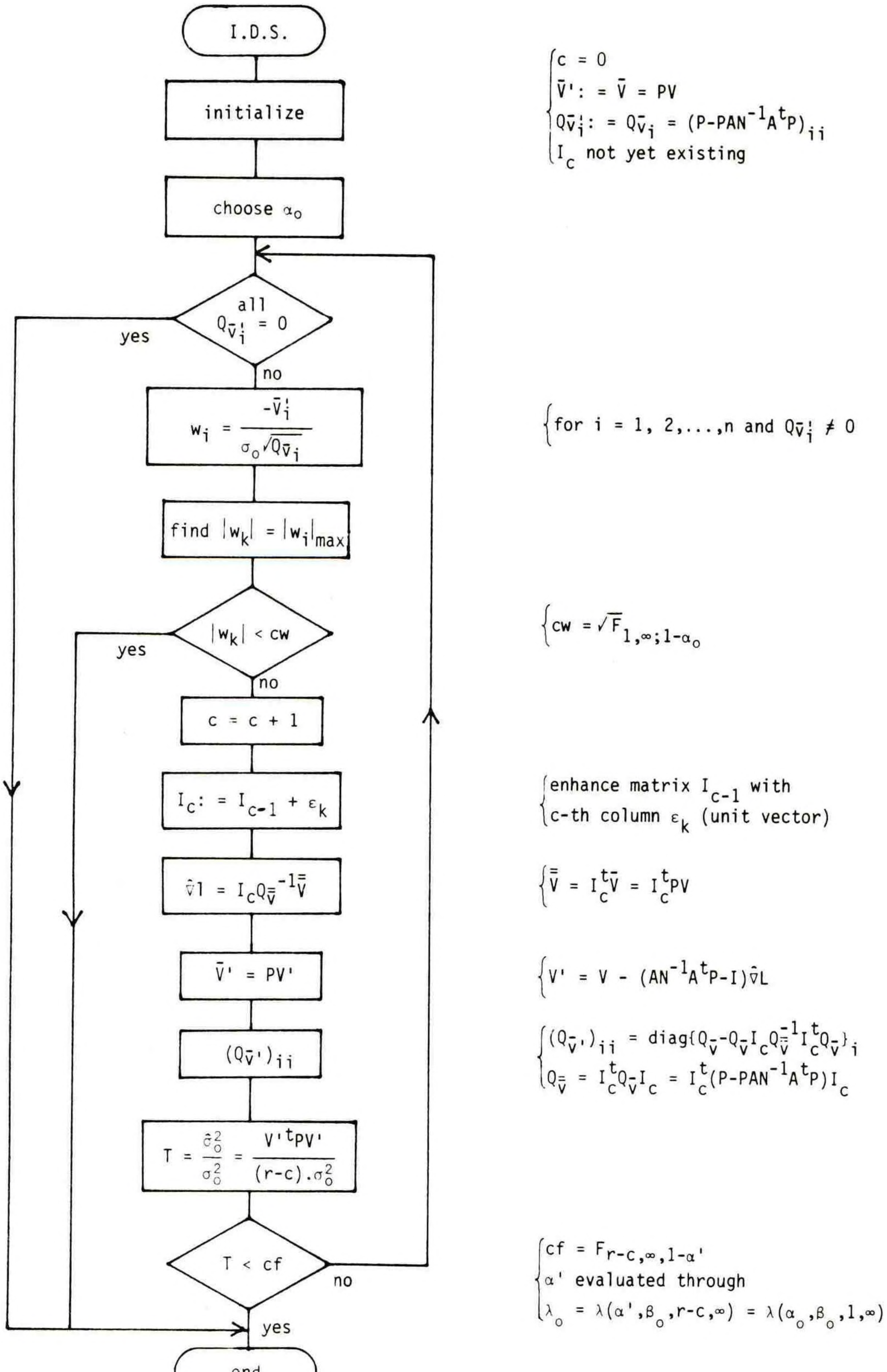


Figure 4.--The process of iterated data snooping (un-Studentized)

A source listing of subroutine HEROB3 (Heuristic Robustizing, version 3) is given in the appendix. It is listed for illustrative purposes only, because for the updating computations with sparse matrices the subroutine library SCAN-II is used, which is implemented on the IBM system of the National Geodetic Survey.

### 4.3. Some Experimental Results

The process of iterated data snooping, by use of subroutine HEROB3 in the subset for two-dimensional networks of the SCAN-II adjustment system, was tried experimentally on two different networks.

(1) Network 1 (fig. 5) is a triangulation network of 16 stations with generated observations, which were made available by Heck (1982). There are no outliers simulated in the original data set.

(2) Network 2 (fig. 6) is a traverse network of 29 stations with real data, made available by the Geodetic Computing Centre of the Delft University of Technology.

Several combinations of blunders were simulated in the observations of the networks, for testing the iterated data snooping procedure. The largest combinations for both networks (eight blunders of different magnitudes in Network 1, and six blunders of different magnitudes in Network 2) are summarized in tables 1 through 4. Additionally a simulation with two extra blunders in Network 2 was performed (table 5).

Table 1.--Simulated errors in Network 1

| | Obs.No. | St.-tg. | Observation [1] | Error [2] | $w_i$ first run [3] |
|---|---|---|---|---|---|
| 1 | 26 | 35-41 | 51.3642 | + 500 | 318.53 |
| 2 | 48 | 41-11 | 269.9067 | - 99 | 1.90 |
| 3 | 2 | 3-7 | 38.6994 | - 30 | - 27.01 |
| 4 | 27 | 35-13 | 134.4900 | - 20 | - 88.83 |
| 5 | 50 | 43-21 | 0.0025 | + 25 | 0.98 |
| 6 | 61 | 47-37 | 118.1463 | + 14 | 4.25 |
| 7 | 59 | 47-35 | 39.2156 | - 11 | 31.70 |
| 8 | 42 | 39-45 | 246.5389 | - 8 | 21.85 |

[1]Units are grades.
[2]Units are $10^{-4}$ grades.
[3]Critical value for $\alpha_o$=0.001 is $\sqrt{F_{1,\infty;1-\alpha_o}} = 3.29$

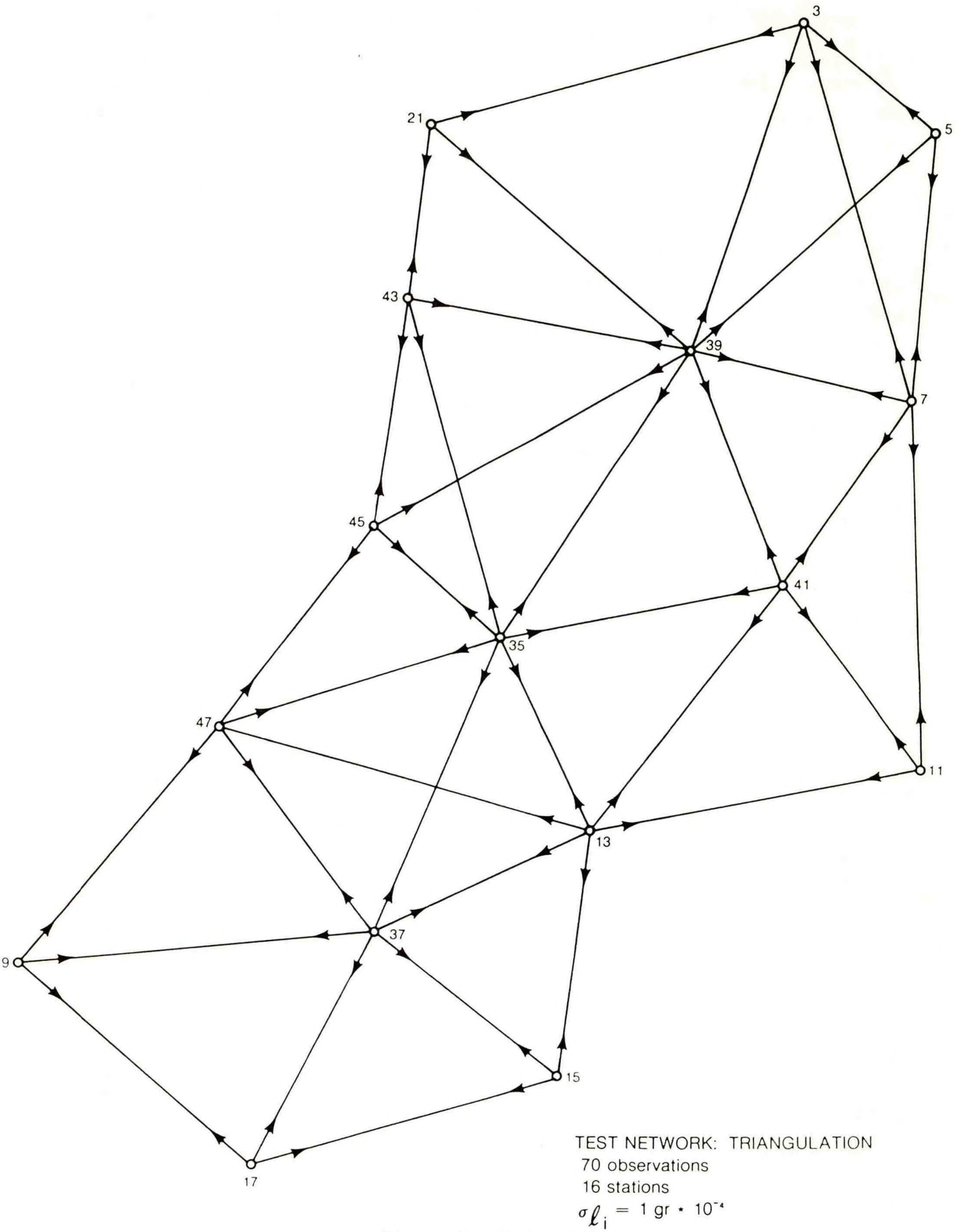


Figure 5.--Network 1.

Table 2.--Results of iterated data snooping (Network 1)

| | | | Iterated data snooping | | | | | | | | F-tests | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\hat{\nabla}L_1$ | $\hat{\nabla}L_2$ | $\hat{\nabla}L_3$ | $\hat{\nabla}L_4$ | $\hat{\nabla}L_5$ | $\hat{\nabla}L_6$ | $\hat{\nabla}L_7$ | $\hat{\nabla}L_8$ | $\alpha'$ | $\hat{\sigma}_o^2/\sigma_o^2$ | F-value |
| STEP | $\lvert w_i \rvert_{max}$ | i | 35-41 | 41-11 | 3-7 | 35-13 | 43-21 | 47-37 | 47-35 | 39-45 | | | |
| 1 | 318.5 | 26 | 487.5 | - | - | - | - | - | - | - | 0.14 | 4042.1 | 1.30 |
| 2 | 51.8 | 48 | 500.2 | -99.6 | - | - | - | - | - | - | 0.14 | 137.6 | 1.31 |
| 3 | 20.6 | 2 | 499.6 | -100.9 | -28.4 | - | - | - | - | - | 0.13 | 36.0 | 1.33 |
| 4 | 13.8 | 27 | 494.4 | -102.0 | -28.7 | -22.7 | - | - | - | - | 0.13 | 21.7 | 1.34 |
| 5 | 11.1 | 50 | 494.8 | -102.0 | -30.4 | -22.5 | 22.8 | - | - | - | 0.12 | 14.4 | 1.36 |
| 6 | 11.0 | 61 | 495.3 | -101.7 | -30.4 | -22.0 | 23.0 | 16.4 | - | - | 0.11 | 9.5 | 1.38 |
| 7 | 7.2 | 59 | 497.4 | -101.6 | -30.5 | -19.5 | 22.5 | 14.8 | -10.2 | - | 0.11 | 4.14 | 1.41 |
| 8 | 3.8 | 42 | 498.3 | -101.7 | -30.6 | -18.4 | 23.1 | 15.1 | -11.2 | -6.0 | 0.10 | 1.79 | 1.43 |
| 9 | 2.9 | 14 | STOP[1] | | | | | | | | 0.09 | 1.13[1] | 1.46 |

[1] $\lvert w_i \rvert_{max} < 3.29$ and $\hat{\sigma}_o^2/\sigma_o^2 < F_{r-8,\infty;1-\alpha'}$ (=1.46)

The results for Network 1, which is a "monolytic" network (only one type of observation) with a high local redundancy, is very satisfying.
All simulated errors are recovered, and in the final list of estimated errors of the suspects (table 2) the largest difference with the actually simulated error is less than three times the standard deviation of the observation.

Note that the masking effect can very well be recognized in observation 48 where the test statistic $w_i$ of the first run - in spite of a blunder of $99.10^{-4}$ grades - does not exceed the critical value of 3.29. Due to the relatively large errors in observations 26 and 48, out of a total of 70 observations 54 had a test statistic $|w_i| > 3.29$ in the nominal adjustment.

Network 2 involves two different types of observations, distance measurements and direction measurements. The results of two trials of iterated data snooping will be summarized. For the first one, six errors of various sizes were applied to the data, four errors in distances of respectively 10 meters and 1 meter, and two errors in directions of 0.010 gr and 0.007 gr. (See table 3).

Table 3.--Simulated errors in Network 2

| | Obs. No. | Type | St.-tg. | Observation[1] | Error |
|---|---|---|---|---|---|
| 1 | 4 | distance | 183-185 | 921.495 | - 10.00 |
| 2 | 2 | distance | 105-213 | 1224.685 | - 10.00 |
| 3 | 8 | distance | 189-239 | 936.965 | - 1.00 |
| 4 | 13 | distance | 195-199 | 693.027 | + 1.00 |
| 5 | 88 | direction | 237-251 | 168.3218 | + 0.0100 |
| 6 | 80 | direction | 233-241 | 289.7501 | + 0.0070 |

[1] Units of distances are meters; units of directions are grades.

In six steps of iterated data snooping, all simulated errors are detected, and the error estimates, which are solved in the last step, are close to the values of the simulation. The results of these six steps are summarized in table 4.

However, in spite of these results we still must treat the outcome of the procedure as a list of suspects only, that hopefully is in agreement with actual errors. The list is built on the assumption that at each step of the procedure the

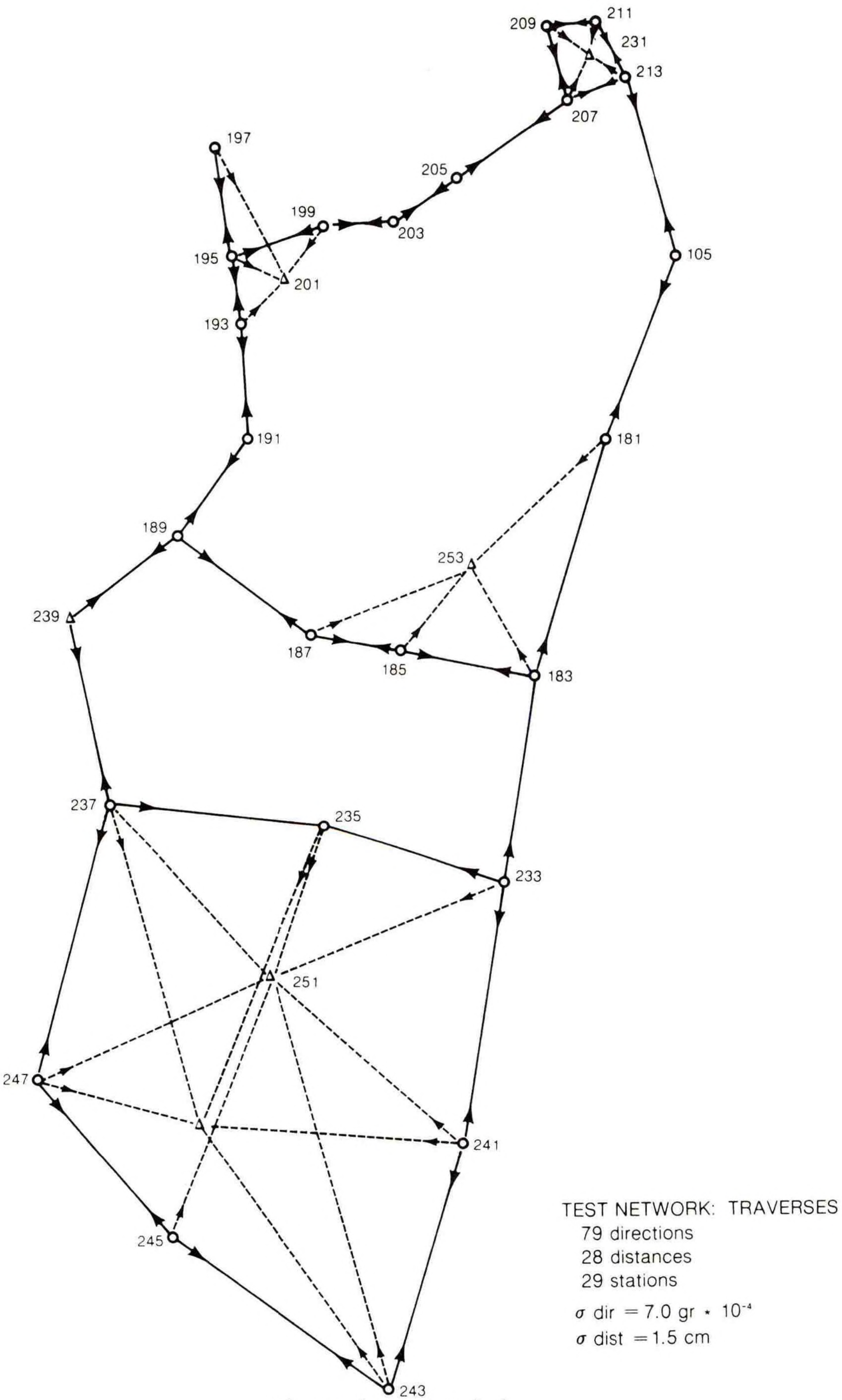


Figure 6.--Network 2.

maximum updated test statistic $|w_i|_{max}$ coincides with a blunder. We cannot prove that this will always be true. An extension of the method, which is indicated in section 5, may improve the reliability of the suspects list, but is not programmed and tested yet.

Table 4.--Results of iterated data snooping (Network 2)

| | | | $\hat{\nabla}L_1$ | $\hat{\nabla}L_2$ | $\hat{\nabla}L_3$ | $\hat{\nabla}L_4$ | $\hat{\nabla}L_5$ | $\hat{\nabla}L_6$ |
|---|---|---|---|---|---|---|---|---|
| Step | $\|w_i\|_{max}$ | i | 183-185 | 105-213 | 189-239 | 195-199 | 237-251 | 233-241 |
| 1 | 335.53 | 4 | -10.13 m | - | - | - | - | - |
| 2 | 103.68 | 2 | -10.23 m | - 9.99 m | - | - | - | - |
| 3 | 15.43 | 8 | -10.06 m | -10.14 m | -1.04 m | - | - | - |
| 4 | 16.57 | 13 | - 9.95 m | - 9.92 m | -1.14 m | 1.03 m | - | - |
| 5 | 6.91 | 88 | - 9.95 m | - 9.90 m | -1.04 m | 1.05 m | 0.0111 gr | - |
| 6 | 5.34 | 80 | - 9.95 m | - 9.89 m | -1.07 m | 1.05 m | 0.0101 gr | 0.0082 gr |
| 7 | 2.81 | 74 | STOP | (<3.29) | | | | |

The following may illustrate the fact that the list of suspects, resulting after the iterated data snooping as presented here, sometimes will not be in complete agreement and may find another pattern of error estimates which fulfills the enhanced adjustment model. We added two more errors to the list of simulated errors of Network 2 (table 3):

- in distance 189-191 an error of + 1.00 m, and
- in direction 207-213 an error of + 0.0200 gr.

The process is now terminated after seven steps (instead of eight, which could be expected) and the resulting list of suspects and error estimates is given in table 5.

We see that two simulated errors are not detected now:

- in direction 233-241 of 0.0070 gr, and
- in distance 189-191 of 1.00 m.

Also one entity in the list of suspects (table 5) is not really an outlier (distance 191-193) and one of the estimates is not in good agreement with the simulated error (distance 105-213). Still the list of suspects includes six out of the eight outliers that were simulated, and this is performed in one single run of the adjustment program.

Table 5.--Results of the third simulation (Network 2)

| | Obs. No. | St-tg. | Error estimate | Simulated error? |
|---|---|---|---|---|
| 1 | 4 | 183-185 | - 9.99 m | yes |
| 2 | 2 | 105-213 | - 7.54 m | yes |
| 3 | 13 | 195-199 | + 0.95 m | yes |
| 4 | 8 | 189-239 | - 1.07 m | yes |
| 5 | 64 | 207-213 | + 0.0202 gr | yes |
| 6 | 10 | 191-193 | + 3.20 m | no |
| 7 | 88 | 237-251 | + 0.0078 gr | yes |

### 4.4. Computational Remarks

The second variant of iterated data snooping, described in sec. 4.2, is computationally attractive, because it is repeatedly using the Cholesky factor of the nominal adjustment. Most modern programs for least-squares adjustment of networks utilize the sparsity of the normal matrix, and of the triangular matrices obtained after factorization of the normal matrix by the method of Gauss or of Cholesky. Supposing the Cholesky factorization performed:

$A^tPA = N = U^tU$, where U is the sparse Cholesky factor;

then in each step of iterated data snooping we have to evaluate the equations

$$\hat{\nabla}L = I_C\,\hat{X}_2 = I_C(I_C^t Q_{\bar{V}} I_C)^{-1}\cdot(-\,I_C^t\bar{V}) = I_C\,Q_{\bar{\bar{V}}}^{-1}\cdot -\,\bar{\bar{V}}, \tag{92}$$

where $Q_{\bar{V}} = P - PAN^{-1}A^tP,$

$$\bar{V}' = \bar{V} - \hat{\nabla}\bar{V} = \bar{V} - (PAN^{-1}A^tP - P)\,\hat{\nabla}L, \tag{93}$$

and $$Q_{\bar{V}'_i} = (Q_{\bar{V}'})_{ii} = (Q_{\bar{V}})_{ii} - (Q_{\bar{V}} I_C Q_{\bar{\bar{V}}}^{-1} I_C^t Q_{\bar{V}})_{ii}. \tag{95}$$

In all three equations selected elements of $Q_{\bar{V}}$ of the nominal adjustment, computable by use of the Cholesky factor U, are used.

For standard data snooping procedures, where only the diagonal elements of $Q_{\bar{v}}$ are needed for the test statistics $t_1$, $t_2$, or $t_3$ of eqs. (40a), (40b), and (40c), there are different ways for an efficient computation of these quantities by using sparsity and avoiding complete inversion of N.

In SCAN-II.3 (the software system used for the test computations of sec. 4.3) the diagonal of $Q_{\bar{v}}$ is computed as follows: Let the factorization of N in $N = U^t U$ be performed in sparse matrix technique, with storage of U in a profile structure or, as in SCAN-II, in a general sparse matrix storage scheme. Then we have to compute the diagonal elements

$$Q_{\bar{v}_i} = (P - PAN^{-1}A^tP)_{ii} = (P)_{ii} - (PAN^{-1}A^tP)_{ii}$$

Suppose that PA = B is already computed, then

$$Q_{\bar{v}_i} = (P)_{ii} - (BN^{-1}B^t)_{ii} = (P)_{ii} - (B_i N^{-1} B_i^t),$$

where $B_i$ is the i-th row of B.

The second part, $B_i N^{-1} B_i^t$, involves the inverse of matrix N.
Then let

$$N^{-1} B_i^t = z,$$

or z is the solution of the system

$$N.z = B_i^t \text{ or } (U^t U)\ z = B_i^t.$$

Now let U.z = y; then this system can be written also as

$$U^t y = B_i^t,$$

which is a lower triangular (sparse) system of equations, to be solved by forward substitution

$$y = U^{-t} B_i^t.$$

The vector product $y^t.y$ now yields

$$y^t.y = B_i U^{-1}.U^{-t} B_i^t = B_i N^{-1} B_i^t,$$

so $Q_{\bar{v}_i}$ can be found by

$$Q_{\bar{v}_i} = (Q_{\bar{v}})_{ii} = (P)_{ii} - y^t.y = (P)_{ii} - (B_i N^{-1} B_i^t).$$

The normal matrix has not been inverted explicitly. This technique is known as successive one-column bordering of the reduced normal matrix.

However, for evaluating diagonal elements of the inverse $N^{-1}$, as well as functions thereof, such as $Q_{\bar{v}_i}$, Hanson (1978) showed that for sparse normal matrices his technique with computation of the elements of the inverse $N^{-1}$ within the band or profile only, is superior. The number of multiplications (and thus of computing time) is significantly smaller.

The algorithm for computing the part of the inverse within the band (in-place recursive partitioning) can be adapted for general sparse matrix storage schemes. Then the elements of the inverse that are needed (and computed) only, do replace the nonzero elements of the Cholesky factor itself, so no additional storage is needed. For details on the computation of "sparse inverses" the reader is referred to Hanson (1978).

Let $\tilde{N}^{-1}$ be the sparse inverse of the normal matrix N, then $Q_{\bar{v}_i}$ can be evaluated through

$$Q_{\bar{v}_i} = (Q_{\bar{v}})_{ii} = (P)_{ii} - (B_i \tilde{N}^{-1} B_i^t).$$

It can be shown (see, e.g., Kok et al. 1980) that the elements of the sparse inverse $\tilde{N}^{-1}$ are the only ones needed in the vector-matrix-vector products $B_i \tilde{N}^{-1} B_i^t$. Release 4 of the SCAN-II system is now also using this sparse inverse algorithm for standard data snooping and for the computation of standard ellipses. Computing times improved considerably, because the in-place recursive partitioning for the sparse inverse computation fully utilizes the symmetry and the sparsity (by avoiding zero products) and the method of successive one-column bordering does not.

For iterated data snooping, proposed in this report, the situation is slightly more complicated, because only in the first step we have the situation of $Q_{\bar{v}_i}$ being the diagonal elements directly of $P - PAN^{-1}A^tP$. During the updating in the following steps, we have to deal with multiple outliers (estimates), and eqs. (92), (93), and (95) involve then selected submatrices of $Q_{\bar{v}}$, denoted as $Q_{\bar{\bar{v}}} = I_c^t Q_{\bar{v}} I_c$, so we also will need off-diagonal elements of the matrix $Q_{\bar{v}}$. So far we have used the bordering method, in sparse matrix mode, to perform the process efficiently. In short, the three main computations during each step of iterated data snooping are as follows:

Equation (92) involves the matrix $Q_{\bar{\bar{v}}}$, which is a selected c x c submatrix of $Q_{\bar{v}}$, where the elements are to be selected according to suspects that are indicated through the statistical tests in successive steps.
Let $L_i$ and $L_j$ be two subsequent suspects; then $Q_{\bar{\bar{v}}}$ is a 2 x 2 matrix.

$$I_c = (\varepsilon_i \ \varepsilon_j); \quad Q_{\bar{\bar{v}}} = I_c^t Q_{\bar{v}} I_c = \begin{pmatrix} Q_{\bar{v}_{ii}} & Q_{\bar{v}_{ij}} \\ Q_{\bar{v}_{ji}} & Q_{\bar{v}_{jj}} \end{pmatrix} \text{ and } I_c^t P I_c = \begin{pmatrix} P_{ii} & P_{ij} \\ P_{ji} & P_{jj} \end{pmatrix}$$

Both matrices are symmetric, and often P is a diagonal matrix, so then

$$P_{ij} = P_{ji} = 0.$$

Let $$Q_{\bar{\bar{v}}} = I_c^t(P - BN^{-1}B^t)I_c = I_c^t P I_c - I_c^t BN^{-1}B^t I_c.$$

For $Q_{\bar{v}_{ii}}$ we use again the system $N.z = B_i^t$, or $(U^tU).z = B_i^t$.
Again let $U.z = y$, then

$$U^t y = B_i^t,$$

with solution y computed by forward substitution

$$y = U^{-t}B_i^t.$$

The vector product

$$y^t.y = B_i U^{-1}.U^{-t}B_i^t = B_i N^{-1}B_i^t$$

is used for $Q_{\bar{v}_{ii}}$, through

$$Q_{\bar{v}_{ii}} = P_{ii} - y^t.y = P_{ii} - B_i N^{-1} B_i^t.$$

For $Q_{\bar{v}_{jj}}$ the computation is similar. The system to be solved is now

$$N.z = B_j^t, \text{ or } (U^t U) z = B_j^t$$

where $B_j^t = A^t P.\varepsilon_j$, or the j-th column of $A^t P$. Let Uz = x, then $U^t x = B_j^t$, with solution $x = U^{-t} B_j^t$. The vector product $x^t.x$ yields the diagonal element $Q_{\bar{v}j}$ by

$$Q_{\bar{v}_{jj}} = P_{jj} - x^t.x = P_{jj} - B_j U^{-1}. U^{-t} B_j^t = P_{jj} - B_j N^{-1} B_j^t.$$

The off-diagonal element $Q_{\bar{v}_{ij}} = Q_{\bar{v}_{ji}}$ is found by use of the vector product $y^t.x$, which then is used for

$$Q_{\bar{v}_{ij}} = P_{ij} - y^t.x = P_{ij} - B_i U^{-1}.U^{-t} B_j^t = P_{ij} - B_i N^{-1} B_j^t.$$

For later steps in the procedure, when c > 2, a similar strategy is followed.

The error estimate vector $\hat{X}_2$ follows as the solution of the system

$$(I_c^t Q_{\bar{v}} I_c). \hat{X}_2 = - I_c \bar{V},$$

or

$$Q_{\bar{\bar{v}}}. \hat{X}_2 = - \bar{\bar{V}},$$

by using full matrix technique, because $Q_{\bar{\bar{v}}}$ is always a (small) full matrix. Then the vector of error estimates for all observations $\hat{\nabla}L$ follows from

$$\hat{\nabla}L = I_c. \hat{X}_2,$$

with $\hat{\nabla}L_i$ and $\hat{\nabla}L_j$ as the only nonzero components.

The updating of the vector of weighted residuals $\bar{V}$, by eq. (93) again uses the bordering method. We write

$$\bar{V}' = \bar{V} - (PAN^{-1}A^{t}P - P)\hat{\nabla L} = \bar{V} - BN^{-1}B^{t}.\hat{\nabla L} + P.\hat{\nabla L}. \tag{93a}$$

Consider the matrix-vector product $BN^{-1}B^{t}.\hat{\nabla L}$ which is the most important part. Let $B^{t}.\hat{\nabla L} = \hat{\nabla Y}$, a vector; then we have

$$B(N^{-1}.\hat{\nabla Y}) = B(U^{-1}U^{-t}).\hat{\nabla Y}.$$

The part $(U^{-1}U^{-t}).\hat{\nabla Y} = z$ is the solution of the system

$$(U^{t}U).z = \hat{\nabla Y}.$$

Then again using $U.z = x$, we can solve

$$U^{t}.x = \hat{\nabla Y}$$

by forward substitution, resulting in

$$x = U^{-t}.\hat{\nabla Y}.$$

Solving z, by applying back substitution on the system

$$U.z = y\,,$$

results in

$$z = U^{-1}y = U^{-1}U^{-t}.\hat{\nabla Y} = N^{-1}.\hat{\nabla Y}.$$

Premultiplying z by matrix B, then finally yields the vector

$$Bz = BN^{-1}.\hat{\nabla Y} = BN^{-1}B^{t}.\hat{\nabla L},$$

which can then be applied in eq. (93a).

The computational effort for updating of $\bar{V}$ is mainly established by performing one forward and one backward substitution process, using the sparse Cholesky factor U of the nominal adjustment and some additional sparse matrix products.

The updating of the diagonal of $Q_{\bar{v}}$, as in eq. (95), is done by essentially the same method, but now applying it repeatedly for more right hand sides. The final result of eq. (95):

$$(Q_{\bar{v}'})_{ii} = (Q_{\bar{v}})_{ii} - (Q_{\bar{v}} I_c Q_{\bar{\bar{v}}}^{-1} I_c^t Q_{\bar{v}})_{ii}$$

is the diagonal of matrix $Q_{\bar{v}}'$, but intermediately the matrix $Q_{\bar{v}} I_c$ is needed. This n x c matrix is built up by selected columns of $Q_{\bar{v}}$ (or rows, because of symmetry). The procedure, described for an element $Q_{\bar{v}_{jj}}$, is followed until the forward substitution, resulting in

$$x = U^{-t} B_j^t = U^{-t} A^t P \varepsilon_j .$$

Now backward substitution is performed on U.z = x, which yields

$$z = U^{-1} . x = U^{-1} . U^{-t} A^t P \varepsilon_j = N^{-1} A^t P \varepsilon_j .$$

Premultiplying z by matrix B = PA, finally results in

$$B.z = PA.z = PAN^{-1} A^t P \varepsilon_j$$

$$Q_{\bar{v}} \cdot \varepsilon_j = P \cdot \varepsilon_j - PAN^{-1} A^t P \varepsilon_j$$

which is the j-th column of matrix $Q_{\bar{v}}$. These steps are repeated for each column of matrix $I_c$, thus building up $Q_{\bar{v}} \cdot I_c$ by successive columns. For eq. (95) matrix $Q_{\bar{\bar{v}}}$, which was already computed, must be inverted as a (small) full matrix. The diagonal of $Q_{\bar{v}}'$ is then

$$(Q_{\bar{v}}')_{ii} = (Q_{\bar{v}})_{ii} - ((Q_{\bar{v}} I_c) . Q_{\bar{\bar{v}}}^{-1} . (I_c^t . Q_v))_{ii} .$$

Of the matrix product $(Q_{\bar{v}} I_c) . Q_{\bar{\bar{v}}}^{-1} . (I_c^t Q_v)$ only the diagonal is actually needed, so this can be done efficiently as well.

Because $I_c$ is built up during the interated data snooping, by columns of unit vectors for new suspects successively, it is not necessary to compute $Q_{\bar{v}} I_c$ completely in each step. Always the last column is needed and it can be added to the previous columns, which were computed in preceding steps, provided these were kept in storage.

## 5. FINAL REMARKS

It is obvious that, because of the close relationship between Studentized and un-Studentized tests for outliers as derived from general linear hypothesis tests as shown in sec. 2, the method of iterated data snooping can also be adapted for Studentized tests.

Some aspects of iterated data snooping need further attention and possibly improvements can be made. The third simulation in sec. 4.2 showed clearly that in the procedure presented, the risk of rejecting "good" observations is still present. A possibility to improve the procedure could be the following extension, which is an idea that is used by Benciolini, Mussio and Sansò (1982). First apply iterated data snooping as described in sec. 4.2, and use a relatively large value of the level of significance for the one-dimensional tests (e.g. $\alpha_o = 0.05$ or $\alpha_o = 0.10$ ), to ensure that the remaining set of observations is virtually free of outliers.

Then, as a second step, apply the procedure again, starting with the last updated results, and re-enter each of the suspects in the same order as it was found in the first step. In this second step, each of the suspects is tested without the presence of other outliers. In case a suspect is not an outlier but has been rejected in the first step as a result of the influence of other outliers, it will now be accepted by the data snooping test. The second step should be applied with a smaller value of $\alpha_o$ (e.g., $\alpha_o = 0.001$).

An aspect that also needs more attention is the definition of the reliability of a network, if iterated data snooping is used. Can the same quantification, based on conventional hypotheses, as given in section 3, be used when applying iterated data snooping?

Also during the procedure, as the degrees of freedom become smaller because they are used for the estimation of possible outliers, reliability of the remaining (current) subset of observations is changing. Computing updated values of marginally detectable errors $\hat{\nabla} L_i$ during the process, which is relatively simple because $Q_{\bar{v}_i}'$ as diagonal elements of $Q_{\bar{v}}'$ are already computed for the test-statistics, can give an indication of the effectiveness that can be expected in following steps of updating and testing. If observations other than the entries in the current list of suspects, which were testable originally ($\hat{\nabla}L_i \neq \infty$ and $Q_{\bar{v}_i} \neq 0$), become untestable ($\nabla L_i' = \infty$ and $Q_{\bar{v}_i}' = 0$) because of inseparability, then an error in these observations cannot be found. But also in case $Q_{\bar{v}_i}' \neq 0$, the value of $\tilde{\nabla}L_i'$ may be so large that errors will not be detected any more.

The examples of iterated data snooping in section 4.2 are an indication that the method may be useful in practical applications. Certainly we need more experience in using it in practice, to see if lists of suspects resulting from the procedure are really helpful for the detection of multiple outliers in different types of geodetic networks. Also the lists of suspects could possibly be a start for other more refined methods for multiple outlier testing.

## 6. ACKNOWLEDGMENT

The author thanks the participants of the "Pope seminars", during which several aspects of testing and reliability were given attention, for their help and support. He is especially grateful to Allen J. Pope, who gave several suggestions for use in multiple outlier testing and who gave the basis for the derivation of the different types of tests from general linear hypothesis tests, as used in section 2 of this report.

## 7. REFERENCES

Baarda, W., 1960: Precision, Accuracy and Reliability of Observations. Geodetic Computing Centre, Delft University of Technology, Delft.

Baarda, W., 1968: A Testing Procedure for Use in Geodetic Networks. Netherlands Geodetic Commission, Delft.

Baarda, W., 1977: Measures for the accuracy of geodetic networks. Proceedings IAG Symposium, Sopron, 1977, pp. 419-436.

Benciolini, B., Mussio, L., Sansò, F., 1982: An approach to gross error detection more conservative than Baarda snooping. Proceedings Symposium of Commission III of the ISP, Helsinki, 1982.

Creusen, M.W.F.J., 1965: A Sequential Procedure to Solve Linear Systems. Report Automatic Operation of Raytheon Co., Alexandria, Va.

Förstner, W., 1979: Das Programm TRINA zur Ausgleichung und Gütebeurteilung geodätischer Lagenetze. Zeitschrift für Vermessungswesen, 104. Jg., Stuttgart.

Fuchs, H., 1981: Adjustment by minimizing the sum of absolute residuals. Proceedings International Symposium on Geodetic Networks and Computations, Munich, 1981, pp. 29-48.

Graybill, F.A., 1976: Theory and Application of the Linear Model. Duxbury Press, North Scituate, Massachusetts.

Hamilton, W.C., 1964: Statistics in Physical Science. The Ronald Press Co., New York, N.Y.

Hanson, R.H., 1978: A posteriori error propagation. Proceedings of the Second International Symposium on Problems Related to the Redefinition of North American Geodetic Networks, Washington, D.C., 1978, pp. 427-445.

Heck, B., 1980: Statistische Ausreisserkriterien zur Kontrolle Geodätischer Beobachtungen. Paper (B-10) VIII. Int. Kurs für Ingenieursvermessung, Zürich.

Koch, K.-R., 1980: Parameterschätzung und Hypothesentests in Linearen Modellen. Dümmler, Bonn.

Kok, J.J., Ehrnsperger, W., Rietveld, H., 1980: The 1979 adjustment of the UELN and its analysis of precision and reliability. Proceedings Second International Symposium on Problems Related to the Redefinition of North American Vertical Geodetic Networks, Ottawa, 1980, pp. 455-483.

Kok, J.J., 1981: The B-method of testing applied to deformation measurements. In: A Comparison of Different Approaches into the Analysis of Deformation Measurements, edited by A. Chrzanowski, Proceedings XVI-th International Congress FIG, Montreux, 1981, pp. 602.3/1-24.

Kok, J.J., 1982: Statistical Analysis of deformation problems, using Baarda's testing procedures. In: Forty Years of Thought, Delft University of Technology, Delft, pp. 469-488.

Krarup, T., Juhl, J., Kubik, K., 1980: Götterdämmerung over least-squares adjustment. Proceedings 14th International Congress of the ISP, Commission III, Hamburg, 1980, pp. 369-378.

Meissl, P., 1980: Adjustment of leveling networks by minimizing the sum of absolute residuals. Proceedings Second International Symposium on Problems Related to the Redefinition of North American Vertical Geodetic Networks, Ottawa, 1980, pp. 393-415.

Pope, A.J., 1976: The statistics of residuals and the detection of outliers. NOAA Technical Report, NOS65 NGS1, National Geodetic Information Center, NOS/NOAA, Rockville, Md.

Pope, A.J., 1982: Personal notes, not published. (National Geodetic Survey, NOS/NOAA, Rockville, Md.).

Van Mierlo, J., 1981: A review of model checks and reliability. Proceedings International Symposium on Geodetic Networks and Computations, Munich, 1981, pp. 308-321.

Whiting, M.C., Pope, A.J., 1976: Adjustment of geodetic field data using a sequential method. NOAA Technical Memorandum, NOS NGS3, National Geodetic Information Center, NOS/NOAA, Rockville, Md.

APPENDIX.--SOURCE LISTING OF A SUBROUTINE FOR ITERATED DATA SNOOPING

```
C  ***********************************************************************
C                                                                        *
C       ***  SOFTWARE SYSTEM   S C A N - II   SUBROUTINE  ***            *
C                                                                        *
C       1982:  J. J. KOK / NATIONAL GEODETIC SURVEY   NOAA/NOS           *
C                                                                        *
C  ***********************************************************************
C
C
C   *** ITERATED DATA-SNOOPING, USING L.S. BLUNDER ESTIMATES  ***
C
C   THE SUBROUTINE USES THE FOLLOWING EXTERNAL ROUTINES
C   FROM THE SCAN-II SUBROUTINE LIBRARY:
C
C      SAF005 - MATRIX PRODUCT:  SPARSE SYM. MATRIX * VECTOR
C      SAF043 - VECTOR PRODUCT:  VECTOR * VECTOR (TRANSP)
C      SAF105 - MATRIX PRODUCT:  SPARSE MATRIX * VECTOR
C      SAF114 - MATRIX INVERSION (FULL) AND SOLUTION,
C               BY THE METHOD OF GAUSS/JORDAN
C      SAF222 - COMPUTATION OF CONFIDENCE LEVELS AND CRITICAL VALUES
C      SAF341 - FORWARD SUBSTITUTION OF SYSTEM:   U(TRANSP).Y = B,
C               WHERE U(TRANSP) IS THE SPARSE CHOLESKI FACTOR TRANSPOSED,
C                       OBTAINED BY SUBROUTINE SAF340,
C                     B IS A SPARSE VECTOR,
C                     Y IS THE SOLUTION VECTOR OF U(T).Y = B
C      SAF142 - BACKWARD SUBSTITUTION OF SYSTEM:  U.X = Y,
C               WHERE U IS THE SPARSE CHOLESKI FACTOR,
C                     Y IS THE SOLUTION VECTOR OF THE LOWER-TRIANGULAR
C                       SYSTEM:  U(T).Y = B, COMPUTED BY SAF341,
C                     X IS THE SOLUTION VECTOR OF SYSTEM:   U.X = Y, AND
C                       ALSO THE SOLUTION VECTOR OF THE TOTAL SYSTEM:
C                       (U(TRANSP)   U).X = B
C
C
C  * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C
      SUBROUTINE HEROB3(NA,A,NG,GBA,NB,B,V,W,NP,VR,NPI,PI,NABL,VEC,VEC2,
     +                  VJ,MAT,QV,ISEQ,SIGM,CV,NS,KL,DS06)
C
      DOUBLE PRECISION A(1),GBA(1),B(1),V(1),W(1),NP(1),VR(1),PI(1),
     +            VJ(1),NABL(1),VEC(1),VEC2(1),SIGM,CV,LW,WW,WM,NAB,X,
     +            MAT(1),QV(1),VPV
C
      INTEGER NA(1),NG(1),NB(1),NPI(1),ISEQ(1),ABC,DS06,NS,KL,DF
      LOGICAL INT
      REAL ALFA,ALFO,BETA,LAMDA,CF,CW
C
C INITIALISING:
      N=NA(2)
      NE=NS+1
      NL=NA(3)
      DF=N-NL+NS+1
      BETA=0.8E00
      INT=.FALSE.
      KL=0
      CALL SAF005(NPI,PI,V,VJ)
      DO 10 I=1,N
      IF (NP(I).LT.1.D-10) NP(I)=0.D0
```

```
   10 NABL(I)=0.D0
C
C STANDARDIZED RESIDUALS PV AND VPV:
      DO 100 I=1,N
C
C MULTI-DIMENSIONAL F-TEST:
      DF=DF-1
      ALFA=0.E00
      ALFO=1.E-03
      CALL SAF222(ALFA,ALFO,BETA,CF,CW,LAMDA,DF,DS06)
      WRITE(DS06,1111) ALFA,ALFO,BETA,CF
 1111 FORMAT(/'    ALFA =',F8.3,'     ALFO =',F8.3,'     BETA =',F8.3,
     +'     CF =',F8.3)
      IF (VPV.LE.CF) GOTO 200
C
C DATA SNOOPING (FIND W-MAX):
      WM=0.D0
      DO 20 J=1,N
      WW=DABS(W(J))
      IF (WW.LE.WM) GOTO 20
      K=J
      WM=WW
   20 CONTINUE
      IF (WM.LE.CV) GOTO 200
      KL=KL+1
      ISEQ(KL)=K
C
C COMPUTATION OF L.S. ERROR ESTIMATES AND UPDATING OF VARIANCES Q(V):
      CALL EST2(NG,GBA,NB,B,NPI,PI,VJ,ISEQ,NABL,KL,NS,VEC,
     +             VEC2,VR,MAT,NP,SIGM,ABC,DS06)
      IF (ABC.LT.500) STOP
      DO 25 J=1,N
      NP(J)=QV(J)-NP(J)
      IF (NP(J).LT.1.D-10) NP(J)=0.D0
   25 NP(J)=DSQRT(NP(J))
C
C COMPUTATION OF UPDATED RESIDUALS V':
      CALL SAF105(NB,B,NABL,VEC)
      DO 40 J=1,NS
   40 VEC(J)=0.D0
      CALL SAF341(NG,GBA,VEC,NE,NL,INT,ABC,DS06)
      CALL SAF142(NG,GBA,VEC,NE,NL)
      CALL SAF105(NA,A,VEC,VR)
      DO 50 J=1,N
      X=VR(J)
   50 VR(J)=V(J)-X+NABL(J)
C
C COMPUTATION OF VPV AND W-STATISTICS (UPDATED):
      CALL ZSV(NPI,PI,VR,W,ISEQ,KL)
      CALL SAF043(VR,W,VPV,1,N)
      VPV=VPV/(DF*SIGM*SIGM)
C
 1060 FORMAT(/'   DEGR. OF FREEDOM =',I4,'    F-VALUE =',F15.3)
      DO 60 J=1,N
      IF (NP(J).NE.0.D0) GOTO 55
      W(J)=0.D0
      GOTO 60
   55 W(J)=-W(J)/(SIGM*NP(J))
   60 CONTINUE
  100 CONTINUE
```

```
C
  200 WRITE(6,1030) KL
 1030 FORMAT(/'  # OF ERROR-ESTIMATES =',I4)
  500 RETURN
C
      END
C
C
C
C  * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C
C    ****   INTERNAL SUBROUTINE FOR HEROB3: ESTIMATES AND VARIANCES ***
C
C  * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C
      SUBROUTINE EST2(NG,GBA,NB,B,NPI,PI,VJ,ISEQ,NABL,KL,NS,RSJ,RSK,
     +                H,MAT,QV,SIGM,ABC,DS06)
C
      DOUBLE PRECISION GBA(1),B(1),PI(1),VJ(1),NABL(1),RSJ(1),RSK(1),
     +                H(1),TRAP(1300),VEC(50),ELM,MAT(1),QV(1),SIGM
      INTEGER NG(1),NB(1),NPI(1),ISEQ(1),ABC,DS06,RANK
      LOGICAL INT,BB(50)
C
C INITIALISATION:
      ABC=500
      INT=.FALSE.
      N=NB(3)
      NL=NB(2)
      NE=NS+1
      NN=(KL+1)*(KL+2)/2
      DO 02 K=1,NN
   02 TRAP(K)=0.D0
C
      DO 100 I=1,KL
      IE=I*(I-1)/2
      JJ=ISEQ(I)
      VEC(I)=VJ(JJ)
      DO 05 K=1,N
   05 H(K)=0.D0
      H(JJ)=1.D0
      CALL SAF105(NB,B,H,RSJ)
      DO 06 K=1,NS
   06 RSJ(K)=0.D0
      CALL SAF341(NG,GBA,RSJ,NE,NL,INT,ABC,DS06)
      IF (ABC.LT.500) GOTO 500
      JE=NPI(JJ+3)+1
      JL=NPI(JJ+4)
      DO 50 J=1,I
      DO 20 K=JE,JL
      IF (NPI(K).NE.ISEQ(J)) GOTO 20
      TRAP(IE+J)=PI(K-N-4)
      GOTO 21
   20 CONTINUE
   21 KK=ISEQ(J)
      IF (KK.NE.JJ) GOTO 40
      CALL SAF043(RSJ,RSJ,ELM,NE,NL)
      IF (I.LT.KL) GOTO 45
      CALL SAF142(NG,GBA,RSJ,NE,NL)
      CALL ZTV(NB,B,RSJ,H,NE)
C
```

```
      KE=NPI(JJ+3)+1
      KLA=NPI(JJ+4)
      DO 22 K=KE,KLA
      II=NPI(K)
   22 H(II)=H(II)-PI(K-N-4)
      KF=JJ+1
      DO 24 K=KF,N
      KE=NPI(K+3)+1
      KLA=NPI(K+4)
      DO 23 KJ=KE,KLA
      II=NPI(KJ)
      IF (II.EQ.JJ) H(K)=H(K)-NPI(KJ-N-4)
      IF (II.GT.JJ) GOTO 24
   23 CONTINUE
   24 CONTINUE
C
      JE=(KL-1)*N
      DO 30 JJ=1,N
      II=JE+JJ
   30 MAT(II)=-H(JJ)
      GOTO 45
   40 DO 25 K=1,N
   25 H(K)=0.DO
      H(KK)=1.DO
      CALL SAF105(NB,B,H,RSK)
      DO 26 K=1,NS
   26 RSK(K)=0.DO
      CALL SAF341(NG,GBA,RSK,NE,NL,INT,ABC,DS06)
      IF (ABC.LT.500) GOTO 500
      CALL SAF043(RSJ,RSK,ELM,NE,NL)
   45 II=IE+J
      TRAP(II)=TRAP(II)-ELM
   50 CONTINUE
  100 CONTINUE
C
      IE=KL*(KL+1)/2
      DO 110 I=1,KL
      II=IE+I
  110 TRAP(II)=VEC(I)
      RANK=KL
      NR=KL+1
      CALL SAF114(TRAP,NR,NR,1,RANK,RSJ,RSK,BB)
      IF (RANK.EQ.KL) GOTO 115
      WRITE(DS06,1000) KL,RANK
 1000 FORMAT(//'     ***  WARNING:   MATRIX QV IS SINGULAR'/
     +          '                    ORDER =',I4,'    RANK =',I4//
     +          '           EXEC OF SUBROUTINE HEROB3 TERMINATED'//)
      ABC=1
      GOTO 500
  115 DO 120 I=1,KL
      II=IE+I
      K=ISEQ(I)
  120 NABL(K)=TRAP(II)
      CALL MTM(MAT,TRAP,QV,RSJ,RSK,KL,N)
C
  500 RETURN
C
      END
C
C
```

```
C
C  * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C
C    *** SUBROUTINE ZTV (MAT-PRODUCT):  SPARSE MATRIX (TR) * VECTOR
C
C  * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C
      SUBROUTINE ZTV(NZ,Z,V,RES,NE)
C
      DOUBLE PRECISION Z(1),V(1),RES(1)
      INTEGER NZ(1),NE
C
      NR=NZ(3)
      NC=NZ(2)
      DO 10 J=1,NR
   10 RES(J)=0.D0
C
      DO 100 I=NE,NC
      IE=NZ(I+3)+1
      IL=NZ(I+4)
      DO 20 K=IE,IL
      J=NZ(K)
   20 RES(J)=RES(J)+Z(K-NC-4)*V(I)
  100 CONTINUE
C
      RETURN
      END
C
C  * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C
C    *** SUBROUTINE MTM (MAT-PRODUCT):
C           DIAG( MAT(TR) * SYM.MAT * MAT)
C
C  * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C
      SUBROUTINE MTM(MAT,TRAP,DIAG,VEC,W,N,M)
C
      DOUBLE PRECISION MAT(1),TRAP(1),DIAG(1),VEC(1),W(1),SUM
C
      DO 100 I=1,M
      DO 10 J=1,N
      II=(J-1)*M+I
   10 VEC(J)=MAT(II)
C
      DO 50 J=1,N
      SUM=0.D0
      DO 30 K=1,N
      JJ=J*(J-1)/2+K
      IF (K.GT.J) JJ=K*(K-1)/2+J
   30 SUM=SUM+VEC(K)*TRAP(JJ)
   50 W(J)=SUM
C
      SUM=0.D0
      DO 60 J=1,N
   60 SUM=SUM+W(J)*VEC(J)
      DIAG(I)=SUM
  100 CONTINUE
C
      RETURN
      END
```

```
C
C  * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C
C   *** SUBROUTINE ZSV (MATRIX-PRODUCT):
C           SYM. SPARSE MAT * VECTOR, OMITTING SPECIFIC ROWS OF MAT
C
C  * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C
      SUBROUTINE ZSV(NZ, Z, VEC, RES, ISEQ, KL)
C
      DOUBLE PRECISION Z(1), VEC(1), RES(1), SUM
      INTEGER NZ(1), ISEQ(1)
C
      NR=NZ(2)
      NC=NZ(3)
C
      DO 100 I=1, NR
      SUM=0. DO
      DO 10 J=1, KL
   10 IF(ISEQ(J). EQ. I) GOTO 90
      IE=NZ(I+3)+1
      IL=NZ(I+4)
      DO 50 K=IE, IL
      J=NZ(K)
   50 SUM=SUM+Z(K-NR-4)*VEC(J)
      IF (I. EQ. NR) GOTO 90
      IE=I+1
      DO 80 II=IE, NR
      JE=NZ(II+3)+1
      JL=NZ(II+4)
      DO 70 K=JE, JL
      J=NZ(K)
      IF (J. GT. I) GOTO 80
      IF (J. LT. I) GOTO 70
      SUM=SUM+Z(K-NR-4)*VEC(II)
   70 CONTINUE
   80 CONTINUE
   90 RES(I)=SUM
  100 CONTINUE
C
      RETURN
C
      END
C
C
```

**U.S. DEPARTMENT OF COMMERCE**
**National Oceanic and Atmospheric Administration**

National Ocean Service
Charting and Geodetic Services
National Geodetic Survey (N/CG17x2)
Rockville, Maryland 20852

**Official Business**

POSTAGE AND FEES PAID
U.S. DEPARTMENT OF COMMERCE
COM-210
THIRD CLASS MAIL

U.S.MAIL