NOAA
Professional Paper 12

A Priori Prediction of Roundoff Error Accumulation in the Solution of a Super-Large Geodetic Normal Equation System

Peter Meissl

Rockville, Md.
June 1980

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

NOAA
Professional Paper 12

A Priori Prediction of Roundoff Error Accumulation in the Solution of a Super-Large Geodetic Normal Equation System

Peter Meissl
National Geodetic Survey
National Ocean Survey
Rockville, Md. 20852

June 1980

U.S. DEPARTMENT OF COMMERCE
Philip M. Klutznick, Secretary
National Oceanic and Atmospheric Administration
Richard A. Frank, Administrator

Foreword

This report is written for a geodetically minded reader. For this reason background material on the theory of roundoff errors is included that could otherwise be found in journals, or even in monographs. Numerical analysts will find many portions of the text long-winded; other sections may interest them. Structural analysts may also find this work applicable to their discipline. The presentation is a compromise between complete documentation and readability. Therefore, I have omitted listing the computer programs and documenting some of the steps used in predicting roundoff.

I have tried to make the report complete as regards the main line of thought. However, some complementary discussions of network theory are only referenced or quoted from other sources. A thorough account of this theory would require a separate monograph that I hope to write later.

Mention of a commercial company or product does not constitute an endorsement by NOAA/National Ocean Survey. Use for publicity or advertising purposes of information from this publication concerning proprietary products or the tests of such products is not authorized.

Library of Congress Catalog number 80-600088.

Acknowledgments

This study was performed primarily at NOAA/National Ocean Survey's National Geodetic Survey (NGS) during an 8-month stay in 1977, when the author served as a Senior Scientist in Geodesy, while on leave from the Technical University at Graz, Austria. The National Research Council of the National Academy of Sciences directed this program for the purpose of providing research opportunities in geodesy at federally-supported research centers.

I express my thanks to my scientific adviser Capt. John D. Bossler who assisted and encouraged me. I am also indebted to other NGS staff members, especially Joseph F. Dracup, John G. Gergen, Allen J. Pope, Dr. Charles R. Schwarz, and T. Vincenty, for valuable suggestions, information, and discussions. Gergen spent many hours introducing me to the details of the U.S. geodetic network and providing me with necessary data. William H. Dillinger, Robert H. Hanson, and John F. Isner contributed computer programing support. In addition, Dillinger carried out roundoff experiments on small subnetworks. My appreciation is also expressed to Prof. R.P. Tewarson, of the State University of New York at Stony Brook, who reviewed my manuscript for NOAA.

My visit was a fascinating experience. The new adjustment of the North American Horizontal Datum has brought together a group of highly skilled and specialized geodesists who, under Bossler's leadership as project manager of the new adjustment of the North American Datum (NAD), are spearheading research in geodesy.

In preparing the final manuscript, much assistance was received from my coworker Dr. N. Bartelme. Finally, I wish to extend my gratitude to Eleanor Andree of NGS who edited the text and helped to overcome various obstacles in the publication of my manuscript.

To all the individuals and institutions named above, I express my thanks and appreciation.

Blank page retained for pagination

Contents

Abstract 1

1. Introduction and Summary of Results 1

1.1 Purpose 1
1.2 Method 1
1.2.1 Two computer families 2
1.2.2 Elementary and local roundoff errors 3
1.2.3 Global roundoff errors 3
1.3 Data 4
1.3.1 Estimation of the inverse 5
1.3.2 Estimation of the number of nonzero locations and elementary operation steps 5
1.3.3 Estimation of the coordinate shifts 6
1.3.4 Estimation of local roundoff errors during triangular decomposition 6
1.4 Evaluation 7
1.5 Results 8
1.6 Verification 9

2. Roundoff Errors and Ways to Analyze Them 9

2.1 Elementary Roundoff Errors 9
2.2 Normalized Floating Point Arithmetic with Fixed Length Mantissas . 10
2.3 Rounding on the CDC 6600 10
2.4 Rounding on the IBM 360 11
2.5 Local Roundoff Errors 11
2.6 Linear Roundoff Error Propagation 11
2.7 Backward Analysis 13
2.8 Stochastic Assumptions on Roundoff Errors 14
2.8.1 Addition and subtraction 14
2.8.1.1 Model for addition and subtraction on the CDC 6600 14
2.8.1.2 Model for addition and subtraction on the IBM 360 15
2.8.2 Multiplication and division 15
2.8.3 Local roundoff error of the square root 16

3. Cholesky's Algorithm Applied to the Normal Equations of Geodetic Networks 16

3.1 Cholesky's Algorithm for a General Symmetric Positive Definite System 16
3.2 Partial Reduction by Cholesky's Algorithm 17
3.3 Geodetic Normal Equations 18
3.4 Geodetic Interpretation of the Partial Cholesky-Reduced System 19
3.5 Problem of Station Ordering 21
3.5.1 Ordering for small bandwidth 21
3.5.2 Ordering for small profile 22
3.5.3 Identifying nonzero coefficients for a certain reduction state . . 22
3.5.4 Nested dissection 23
3.5.5 Helmert blocking 26

4. Roundoff Errors for a General Positive Definite System 27

4.1 Roundoff Errors during the Triangular Decomposition Phase 27
4.1.1 Left-hand side local roundoff errors arising during triangular decomposition 27
4.1.2 Right-hand side local roundoff errors during triangular decomposition 29
4.1.3 Global roundoff errors caused by triangular decomposition... 29
4.1.4 Preliminary estimates of the global roundoff errors in the U.S. network caused by triangular decomposition 30
4.2 Roundoff Errors During Back Substitution 33
4.2.1 Local errors during back substitution..... 33
4.2.2 Global roundoff errors resulting from back substitution..... 34
4.2.3 Preliminary estimates of the global roundoff errors in the U.S. network resulting from back substitution 34
4.3 Taking into Account Helmert Blocking 34
4.4 Effect of Scaling the Normals upon Roundoff Error Propagation ... 35

5. Properties of the U.S. Network Relevant to the Roundoff Study 36

5.1 General Overview 37
5.1.1 Size of the network..... 37
5.1.2 Type of observations 37
5.1.3 Inhomogeneity of the network 37
5.1.4 Structure of the network..... 37
5.2 Estimating the Inverse of the Normal Equation Matrix 41
5.2.1 Description of the model 41
5.2.2 Further simplifications..... 45
5.2.2.1 Assumed angular observations 46
5.2.2.2 Assumed base lines other than those in the transcontinental traverses 46
5.2.2.3 Observations of the transcontinental traverses..... 46
5.2.2.4 Doppler position observations..... 47
5.2.3 Remarks on the computer programs for the simulation study.. 47
5.2.4 Results of the simulation study..... 47
5.3 Estimating the Local Features of the Covariance Matrix 54
5.4 Supporting Evidence for Estimates of Global and Local Covariances from Test Calculations..... 54
5.5 Supporting Evidence for Estimates of Global and Local Covariances from the Mathematical Theory of Regular Networks 55
5.5.1 Regular model of the U.S. network 55
5.6 Model Adopted for the Covariance 61

6. Count of Storage Locations and Operations..... 62

6.1 Specifying a Preliminary Blocking Scheme..... 62
6.2 Counting the Nonzero Coefficients and the Elementary Operations.. 65
6.2.1 First-level counts 66
6.2.2 Counts for $2^\circ \times 2^\circ$ quads..... 68
6.2.3 Remarks on the counts of larger blocks 70
6.2.4 Results of counts 72

7. Size of Coefficients during Triangular Decomposition 77

7.1 Left-Hand Side Coefficients 77
7.1.1 Station situated in interior of lowest level block 77
7.1.2 Station situated on a block barrier 81
7.1.3 Bounds on off-diagonal elements..... 82
7.1.4 Transforming away the weight singularities 84

7.2 Expected Coordinate Shifts and Right-Hand Side Coefficients 87
7.2.1 Quality of approximate coordinates 87
7.2.2 History of right-hand side coefficients 88

8. Safe Bounds on the Global Roundoff Errors 91

8.1 Roundoff Error Propagation when Weight Singularities are Removed by Transformation 91
8.1.1 Estimating the bias $E\{\zeta\}$ 91
8.1.2 Estimating the standard deviation $\sigma\{\zeta\}$ 93
8.2 Contribution of the Weight Singularities 95
8.2.1 Estimating the bias $E\{\zeta\}$ 96
8.2.2 Estimating the standard deviation $\sigma\{\zeta\}$ 96
8.3 Residual Bias on the CDC 6600 97
8.4 U.S. Network Without Doppler Stations 98

9. Attempts to Lower the Estimates 98

9.1 Review of Causes for Overestimation 98
9.2 Sign Pattern of the Coefficients $a_{ij}^{(p)}$ 101
9.3 Offsetting Bias Contributions 102
9.4 Reexamining the Row Sum Norms 103
9.5 Aiming at Realistic Bias Estimates for the IBM 360 105
9.6 Remarks on Standard Deviations of the Global Errors 106
9.7 Global Roundoff Errors of the Relative Position of Two Closely Situated Stations 107

10. Roundoff Experiments 107

10.1 Moose-Henriksen Network 107
10.1.1 Purpose and design of the roundoff experiments 108
10.1.2 A posteriori roundoff error analysis 108
10.1.3 Results of the Moose-Henriksen network experiments 110
10.1.3.1 Adjustment of the network as a whole 110
10.1.3.2 Adjustment by Helmert blocking 112
10.1.4 Extrapolation of the test results 112
10.2 Roundoff Experiments by Ebner and Mayer 113
10.3 Roundoff Experiments by Ehlert 113
10.4 Roundoff Experiments with Idealized Leveling Networks by Bartelme-Meissl 114
10.5 Roundoff Experiments Related to the Kentucky-Tennessee Test Area 114

11. Miscellaneous Complements 117

11.1 On the Choice of Norm of the Predicted Roundoff Errors in Geodetic Normal Equations 117
11.2 Asymptotic Roundoff Error Estimates 118
11.2.1 Nested dissection of homogeneous and regular networks 118
11.2.2 A general theorem 120
11.2.3 Networks without absolute position observations 122
11.2.3.1 Networks obeying the logarithmic law 123
11.2.3.2 Networks obeying the "Dutch law" 123
11.2.4 Networks with absolute position observations at regularly spaced intervals 123
11.3 Roundoff Estimates for the UNIVAC 1100/40 124
11.3.1 Double precision floating point arithmetic on the UNIVAC 1100/40 124
11.3.2 Safe bounds for the UNIVAC 1100/40 125

11.3.3 More realistic estimates 126
11.4 Recovering [pvv] from the Reduction of the Normal Equations ... 126
References 127

Figures

Figure 3.1.—Sample network 21
Figure 3.2.—Banded normal equations 21
Figure 3.3.—Profiled normal equations 22
Figure 3.4.—Sample network with stations 1 through 12 eliminated from normal equations 23
Figure 3.5.—Structure of normal equations when stations 1 through 12 are eliminated 24
Figure 3.6.—Nested dissection 25
Figure 3.7.—Sample network decomposed into two Helmert blocks 26
Figure 4.1.—Regular (unshaded) and exceptional (shaded) nodes in nested dissection 35
Figure 5.1a.—Station occupancies of 1° × 1° quads. (Numbers shown are divided by 10 and rounded.) 38
Figure 5.1b.—Station occupancies of 1° × 1° quads 39
Figure 5.1c.—Station occupancies of 2° × 2° quads 40
Figure 5.2.—A chain-type network 41
Figure 5.3.—The U.S. national geodetic network 42
Figure 5.4.—Basis figures of transcontinental traverses 43
Figure 5.5.—Sample network with superimposed finite element structure ... 43
Figure 5.6.—Quad of (a) undisturbed and (b) disturbed finite element grid . . 43
Figure 5.7.—Choice of local coordinate system in a finite element quad ... 44
Figure 5.8.—Assumed regular pattern of primary stations in a 1° × 1° finite element quad 46
Figure 5.9.—Rms point errors. Values refer to the global covariance and are listed in centimeters 48
Figure 5.10a.—Condensed covariance values for the base quad $\phi = 39°$, $\lambda = 77°$. Values shown are in units of the fourth decimal place 49
Figure 5.10b.—Pictorial representation of global covariance. Network response to latitude disturbance at $\phi = 39°$, $\lambda = 77°$ 50
Figure 5.10c.—Pictorial representation of global covariance. Network response to longitude disturbance at $\phi = 39°$, $\lambda = 77°$ 51
Figure 5.10d.—Pictorial representation of global covariance. Network response to latitude disturbance at $\phi = 47°$, $\lambda = 69°$ 52
Figure 5.10e.—Pictorial representation of global covariance. Network response to longitude disturbance at $\phi = 47°$, $\lambda = 69°$ 53
Figure 5.11.—Portion of a regular directional network 56
Figure 5.12.—Portion of a superimposed regular distance and azimuth network 57
Figure 5.13.—Finite element grid superimposed on a regular network 59
Figure 6.1.—Modified station occupancies of 1° × 1° quads. (Numbers shown are divided by 10 and rounded.) 63
Figure 6.2.—Preliminary blocking scheme 64
Figure 6.3.—Interior and boundary equations for a 1° × 1° quad and its four neighbors 65
Figure 6.4.—Distribution of interior and boundary stations in a 1° × 1° quad 65

Figure 6.5.—Idealized example of a network with invariant regional redundancy. 66
Figure 6.6.—Interior and junction equations for a 1° × 1° block. 66
Figure 6.7.—Profile of normal equations for a 1° × 1° block. 67
Figure 6.8.—Interior and junction equations for a 2° × 2° "low"-level block. 68
Figure 6.9.—Profile of normal equations for a 2° × 2° low-level block. 69
Figure 6.10.—Interior and junction equations for a 2° × 2° "medium"-level block. 70
Figure 6.11.—Profile of normal equations for a 2° × 2° medium-level block. 71
Figure 6.12.—Interior and junction equations for a 2° × 2° "modified-medium" level block. 72
Figure 6.13.—Profile of normal equations for a 2° × 2° modified medium-level block. 72
Figure 6.14a.—Π counts for 2° × 2° quads. 73
Figure 6.14b.—Π counts for 2° × 2° quads. 74
Figure 6.15a.—Γ counts for 2° × 2° quads. 75
Figure 6.15b.—Γ counts for 2° × 2° quads. 76
Figure 7.1.—Sample of a lowest level block with about 50 percent of the interior stations eliminated. 78
Figure 7.2.—Left-side normal equation coefficients of two connected stations, as discussed in the text. 80
Figure 7.3.—A chain of tightly connected stations. 86
Figure 7.4.—A nearly rigid substructure of tightly connected stations. 86
Figure 7.5a.—Pictorial representation of coordinate shifts. 89
Figure 7.5b.—Shifts used in calculations 90
Figure 8.1.—Ξ-counts for 2° × 2° quads. 94
Figure 8.2a.—Covariance for net without Doppler observations. Response of network to latitude disturbance at $\phi = 39°, \lambda = 77°$ 99
Figure 8.2b.—Covariance for net without Doppler observations. Response of network to longitude disturbance at $\phi = 39°, \lambda = 77°$. . . 100
Figure 9.1a.—Reaction forces in a poorly anchored free-distance network. . 104
Figure 9.1b.—Reaction forces in a well anchored free-distance network. . . . 104
Figure 10.1.—Test area used by Moose and Henriksen for first- and second-order networks. 109
Figure 10.2.—Distortion of a regular leveling network caused by roundoff errors. 115
Figure 11.1.—Nested dissection of a geodetic network. 119
Figure 11.2.—Simplified structure of normal equations for one level ℓ block. 120

Tables

Table 5.1.—Left-side normal equation coefficients A_{rs} for a regular directional network. (See eq. (5.15).). 56
Table 5.2.—Left-side normal equation coefficients A_{rs} for a regular distance and azimuth network. (See eq. (5.15)) 57
Table 5.3.—Normals for the idealized U.S. network accounting for directions, distances, and azimuths 58
Table 5.4.—Normals of the 8 × 8 finite element grid. 58
Table 5.5.—Sample values for kernel F_{pq} for original idealized network 61
Table 5.6.—Values for $F_{\bar{p}\bar{q}}$ for finite element network. 61
Table 5.7.—Sample values of $\Phi(\alpha)$ 61
Table 6.1.—Summary of Π and Γ counts for levels 1 to 8. 72
Table 6.2.—Summary of Π and Γ counts for "low," "medium," and "high" categories 77

Table 7.1.—Coefficients of contributions to the normals from one distance observation before transformation 84
Table 7.2.—Coefficients of the normal contributions after transformation . . 85
Table 8.1.—Bounds from eq. (8.8) on the bias $E\{\zeta_p\}$ during first iteration for the net with weight singularities removed, using an IBM-type computer 93
Table 8.2.—Bounds from eq. (8.12) on the bias $E\{\zeta_p\}$ for a network with weight singularities removed, using an IBM-type computer . . 93
Table 8.3.—Bounds from eq. (8.17) on $\sigma\{\zeta_p\}$ during first iteration and for network with weight singularities removed 95
Table 8.4.—Bounds from eq. (8.24) for contribution of weight singularities toward $E\{\zeta_p\}$ during the first iteration 96
Table 8.5.—Bounds from eq. (8.23) for contribution of weight singularities toward $E\{\zeta_p\}$ during the first iteration 96
Table 8.6.—Bounds for the CDC 6600 computer derived from eq. (8.26) showing the contribution of weight singularities toward $\sigma\{\zeta_p\}$ during the first iteration 97
Table 8.7.—Bounds for the IBM 360 computer derived from eq. (8.26) showing the contribution of weight singularities toward $\sigma\{\zeta_p\}$ during the first iteration 97
Table 8.8.—IBM 360 bias estimates for base quad ϕ=39°, λ=77°, relying on eq. (8.8) which uses Γ counts 98
Table 8.9.—IBM 360 bias estimates for base quad ϕ=39°, λ=77°, relying on eq. (8.12) which uses Ξ counts 98
Table 8.10.—Standard deviation estimates for base quad ϕ=39°, λ=77° 98
Table 9.1.—A rough estimate of size distribution for diagonal coefficients a_{ii} and the factors γ_r, δ_r by which the two row sum norms exceed a_{ii} 105
Table 9.2.—Evaluation of eq. (9.23) 106
Table 9.3.—Choice of factors ϕ_r 106
Table 9.4.—Attempted realistic estimates for global bias-type roundoff errors on the IBM 360 106
Table 10.1.—Size distribution of diagonal coefficients in the original and fully reduced normal equations 111
Table 10.2.—Additional statistics on the size of elements in the original and reduced normals 111
Table 10.3.—Size distribution of diagonal coefficients in the original and reduced normals after treating the four first-level blocks 116
Table 10.4.—Additional statistics on the size of the elements in the original and reduced normals after treating the four first-level blocks 116
Table 11.1.—Bound on bias and standard deviation of the global roundoff errors encountered on the UNIVAC 1100/40 during the first iteration of the U.S. network adjustment 126

A Priori Prediction of Roundoff Error Accumulation in the Solution of a Super-Large Geodetic Normal Equation System

Peter Meissl [1]

National Geodetic Survey
National Ocean Survey, NOAA
Rockville, Md. 20852

ABSTRACT. The theory of roundoff errors for linear equations is adapted and applied to a linear system of 350,000 unknowns, representing the normal equations of the U.S.-ground control network now being readjusted. The system is positive definite and sparse. Cholesky's algorithm is used. The equations are reordered in a way dictated by the Helmert blocking technique. The block design is based on nested dissection. A linear stochastic roundoff error propagation model is used. Two families of computers are considered that come close to representing the two extremal cases of true chopping and true rounding, the CDC 6600 (with the rounding option set into effect) and the IBM 360. Structural properties of the U.S. network relevant to roundoff error propagation are thoroughly investigated. Next to the large size of the network, weight singularities from observations of extremely high accuracy cause some concern. Bounds on bias and standard deviation of the individual components of the solution vector are derived. They indicate that the new adjustment of the North American Datum (NAD) is feasible on both types of computers.

1. INTRODUCTION AND SUMMARY OF RESULTS

1.1 Purpose

We are dealing with a symmetric and positive-definite system of linear equations

$$Ax = b \tag{1.1}$$

resulting from the least-squares adjustment of the U.S. ground control network. For an overview of the entire new adjustment project see Bossler (1976). The number of unknowns is about $n = 350{,}000$. The unknowns are latitude and longitude corrections to some 175,000 stations. Orientation unknowns are eliminated in the usual fashion when the contributions of the individual direction bundles to the normal equations are assembled. Cholesky's method is used to solve the normals. Formation and solution of the system are organized according to the Helmert blocking scheme. The prime purpose of this study is to predict the roundoff error accumulation during the solution of the normal equations.

1.2 Method

In order to keep this introductory outline simple,

[1] Permanent address: Technical University Graz, Rechbauerstr. 12, A-8010 Austria.

we pretend that the whole set of normal equations is formed before Cholesky reduction starts. Then Helmert blocking amounts to prescribing a certain ordering scheme to the stations, or equivalently, a certain ordering of the normal equations and unknowns. Helmert blocking actually does more than this. It also interferes with the process of forming the normals as well as with the process of solving them. However, since these details have a marginal impact on roundoff, their discussion will be postponed until chapter 3. Starting from the original normal equations

$$\sum_{j=1}^{n} a_{ij}\, x_j = b_i, \quad i=1,\ldots,n \tag{1.2}$$

Cholesky's method derives a triangularized system

$$\sum_{j=i}^{n} r_{ij}\, x_j = s_i, \quad i=1,\ldots,n \tag{1.3}$$

by means of the following set of formulas:

$$\left.\begin{aligned} r_{ii} &= \sqrt{a_{ii} - \sum_{k=1}^{i-1} r_{ki}^2} \\ r_{ij} &= (a_{ij} - \sum_{k=1}^{i-1} r_{ki}\, r_{kj})/r_{ii}, \; j=i+1,\ldots,n \\ s_i &= (b_i - \sum_{k=1}^{i-1} r_{ki}\, s_k)/r_{ii} \end{aligned}\right\} \; i=1,\ldots,n \tag{1.4}$$

which is then solved by back-substitution

$$x_i = (s_i - \sum_{j=i+1}^{n} r_{ij}\, x_j)/r_{ii}, \quad i=n,\ldots,1. \tag{1.5}$$

The overwhelming majority of elementary computational steps is carried out during the "triangular decomposition phase" described in eqs. (1.4). For the U.S. network it is estimated that about $0.6 * 10^{11}$ product terms $r_{ki}r_{kj}$ are calculated and accumulated in about $1.4 * 10^8$ partial sums

$$\sum_{k=1}^{i-1} r_{ki}\, r_{kj} \tag{1.6}$$

which are subsequently subtracted from the a_{ij}'s. These numbers already reflect the great saving in storage and computation time that occurs as a result of the special sparse structure of the system (1.1) and the effective exploitation of this structure by the Helmert blocking technique. There are other operations during the triangular decomposition phase. The square roots of the diagonal elements are taken, the off-diagonals are divided by the diagonals, and finally there are operations that involve the right-hand sides. The number of square roots is $n = 350{,}000$. The number of divisions of left-side coefficients is about $1.4 * 10^8$. This is also the approximate number of product terms $r_{ki}s_k$ to be evaluated in order to reduce the right-hand sides. Furthermore, there are $n = 350{,}000$ divisions involving the right-hand side. All these numbers are minuscule compared with the $0.6 * 10^{11}$ left-side product terms. The same may be said about the "back-substitution phase," as described by eq. (1.5), that involves about $1.4 * 10^8$ product term evaluations and a comparable number of subtractions. The 350,000 divisions hardly count.

Not only is the number of computational steps and, consequently, the number of roundoff errors greatest during the triangular decomposition of the matrix A, but the largest roundoff errors also occur here. This is due to certain outliers among the coefficients a_{ij} that exceed the size of other coefficient by two to three powers of 10. The outliers are caused by weight singularities, i.e., by a comparatively small number of very accurate observations. There will be no such outliers among the right-hand sides.

For these reasons, and to focus on essentials, the introduction will cover only the roundoff errors arising during the evaluation of the expressions

$$a_{ij} - \sum_{k=1}^{i-1} r_{ki}\, r_{kj}. \tag{1.7}$$

Numerical noise during the calculation of (1.7) will be the limiting factor to the accuracy of the solution of the linear normal equations.

1.2.1 Two computer families

It is assumed that all calculations are done in floating point with mantissas of a fixed length of τ digits. Let β denote the base of the number system in use. A floating point number is then represented as

$$\pm .\overbrace{xxx\ldots x}^{\tau} * \beta^{e}.$$

The signed integer exponent is denoted by e. Of course, only the signed mantissa and the signed exponent are stored. We shall assume that floating point numbers are normalized which means that the leading digit of the mantissa is different from zero. An exception is only the number zero itself, which has a special representation.

Two families of computers are considered in this study. The first is represented by the CDC 6600, whereby it is assumed that the instruction set for true rounding is consistently used. The standard CDC 6600 instruction set performs chopping rather than true rounding. If this is put into effect, the CDC 6600 is not considered a member of the first family. On the CDC 6600 we have $\beta = 2$ and $\tau = 48$. Rounding is true with some slight flaws which we presently ignore.

The second computer family is represented by the IBM 360. The base is $\beta = 16$ and the mantissa length is $\tau = 14$, assuming that double precision is consistently used. The arithmetic performs true chopping with some flaws. Infrequently the result of an addi-

tion/subtraction may differ from the truly chopped result by one unit at the last position. In this case the result is larger in magnitude than the mathematical result.

Remark: After most of the work in this report had been completed, and while the final manuscript was being compiled, the author was informed that the adjustment calculations will actually be done on a UNIVAC 1100/40. This is a binary machine with double precision arithmetic that allows for 60-bit mantissas. The machine is neither close to true rounding nor true chopping. However, the computer guarantees that the roundoff errors will not be larger than those encountered on the IBM 360. In section 11.3 we will specify the modifications of the IBM 360 roundoff error estimates that will cover the case of the UNIVAC 1100/40 machine.

1.2.2 Elementary and local roundoff errors

An elementary operation is an addition, subtraction, multiplication, or division. Assuming that the two operands a, b are correct, and that $a \diamond b$ is the mathematically correct result, while $a \circledcirc b$ is the computer's answer, the *elementary* roundoff error is defined as

$$\varepsilon = a \circledcirc b - a \diamond b. \tag{1.8}$$

We adopt a random model for the elementary roundoff errors, aiming to obtain estimates for bias (expectation, mean value) $E\{\varepsilon\}$ and standard deviation (rms.—root mean square error) $\sigma\{\varepsilon\}$. We do not assume that the operands a, b are random, at least not as far as their leading digits are concerned. Only the trailing digits of the operands are visualized as random, depending, in the case of the U.S. network, on the choice of approximate station positions and other arbitrarily chosen reference values. We are led to view elementary roundoff errors as random variables with uniform distribution in an interval whose boundaries depend on operation and operands. On the CDC 6600 the interval is centered at the result $a \diamond b$. There is no bias; $E\{\varepsilon\} = 0$, except for certain harmless outliers, e.g., those where $a \pm b$ is to be calculated and $|b| < |a|\,\beta^{-\tau}$. The length of the interval does not exceed $c * \beta^{-\tau}$, where $c = \beta^{\gamma}$ is the smallest integer power of the base β which is larger or equal to $|a|$, $|b|$, $|a \pm b|$ in the case of addition/subtraction, or larger than or equal to $|a * b|$, $|a/b|$ in the case of multiplication/division. Since the standard deviation of a uniformly distributed random variable equals the interval length divided by $\sqrt{12}$, we assume $\sigma\{\varepsilon\} = c/\sqrt{12} * 2^{-48}$ on the CDC 6600.

The IBM 360 represents nearly a truly chopping machine. On such a machine the bias $E\{\varepsilon\}$ could be as bad as $-c/2 * 16^{-14}$. The standard deviation could amount to $c/\sqrt{12} * 16^{-14}$.

The reader will find more detailed explanations in chapter 2. Here the explanation is sufficient for an intuitive understanding of elementary roundoff errors, their dependency on size of the operands *and* the result in the case of addition/subtraction; whereas in the case of multiplication/division only the size of the result is relevant. This distinction has to do with the most dreaded wiping out of leading digits that may occur during addition/subtraction but never during multiplication/division. Also, the bias that results from a machine that does not perform true rounding cannot be overemphasized.

Another essential assumption is that all elementary roundoff errors are stochastically independent. This will be important when accumulation and propagation of elementary roundoff errors are studied.

Occasionally certain batches of elementary roundoff errors will have a combined effect that is conveniently considered as an entity in the subsequent analysis. Examples are roundoff errors affecting a square root or the result of a product sum calculation. Such roundoff errors will be called *local*. In a wider sense they also include unbatched elementary roundoff errors.

1.2.3 Global roundoff errors

First let us examine the global effect of a single elementary roundoff error occurring during evaluation of (1.7). Assume then that

$$\sum_{k=1}^{p-1} r_{ki}\, r_{kj} \tag{1.9}$$

has been evaluated correctly, as well as the product term $r_{pi}r_{pj}$. But when this term was added, a roundoff error $-\varepsilon_{ij}$ occurred, causing the result to be falsified as

$$\sum_{k=1}^{p} r_{ki}\, r_{kj} - \varepsilon_{ij}. \tag{1.10}$$

Assuming that no further roundoff error was committed during all subsequent operations of triangular decomposition and back-substitution, let us try to determine the effect of the single elementary roundoff error ε_{ij} onto the final result.

One of the great merits of Wilkinson's work was pointing out that, rather than propagating the error forward through all subsequent computations, it is much simpler to trace it backwards to the original system. Let us pretend that all previous operations are performed backwards, of course, with no additional error. The surprising result is that the original system is just perturbed by ε_{ij} at the position (i,j). The coefficient a_{ij} is replaced by $a_{ij} + \varepsilon_{ij}$. Due to symmetry and because Cholesky's algorithm works with only the upper diagonal portion of the matrix A, we must also assume that a_{ji} is perturbed by ε_{ij}. Hence

the perturbed system looks like

$$\begin{bmatrix} a_{11} & \dots & a_{1i} & \dots & a_{1j} & \dots & a_{1n} \\ \vdots & & \vdots & & \vdots & & \vdots \\ a_{i1} & \dots & a_{ii} & \dots & a_{ij}+\varepsilon_{ij} & \dots & a_{in} \\ \vdots & & \vdots & & \vdots & & \vdots \\ a_{j1} & \dots & a_{ji}+\varepsilon_{ij} & \dots & a_{jj} & \dots & a_{jn} \\ \vdots & & \vdots & & \vdots & & \vdots \\ a_{n1} & \dots & a_{ni} & \dots & a_{nj} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1+\zeta_1 \\ \vdots \\ x_i+\zeta_i \\ \vdots \\ x_j+\zeta_j \\ \vdots \\ x_n+\zeta_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_i \\ \vdots \\ b_j \\ \vdots \\ b_n \end{bmatrix} \quad (1.11)$$

In a more condensed notation

$$(A + \varepsilon)(x + \zeta) = b. \quad (1.12)$$

The solution of the perturbed system is the global roundoff error which superimposes itself upon the true result x. In linear approximation we get

$$\zeta = -A^{-1}\varepsilon x. \quad (1.13)$$

The next procedure is quickly outlined. If we work in linear approximation, we may study the isolated effects of individual roundoff errors separately and afterwards superimpose their contributions to the global roundoff error ζ.

As an intermediate step in the analysis, consider all elementary roundoff errors occurring during the evaluation of (1.7). Superposition may immediately take place after tracing the individual elementary roundoff errors backward. The result will be a perturbance of a_{ij} and a_{ji}. This perturbance is again denoted by ε_{ij}; however, this time ε_{ij} is a sum of as many *as* $2\mu_{ij}$ elementary roundoff errors traced backwards. Thereby μ_{ij} denotes the number of product terms in (1.7) which are different from zero. Having reinterpreted the quantity ε_{ij} in this way it now becomes a local roundoff error, eqs. (1.11), (1.12), and (1.13) are formally unchanged.

Finally, consider all positions (i,j), $i \leqslant j$, and all local roundoff errors arising there. Then eq. (1.11) no longer applies because we have perturbances at more locations. However, eqs. (1.12), (1.13) may still be retained, provided that now we view ε as a symmetric matrix of local roundoff errors. We have an ε_{ij} different from zero at all "nonzero" locations (i,j). By a nonzero location we mean an entry (i,j), such that either $a_{ij} = 0$, or that "fill-in" occurs at (i,j) by means of a nonzero product term $r_{ki}r_{kj}$ which will make the result of eq. (1.7) nonzero, in general. Counting nonzero locations only above and including the main diagonal, their total number is clearly

$$\Pi = \sum_{i=1}^{n} \mu_{ii} + n. \quad (1.14)$$

The total number of nonzero products to be evaluated and added during the triangular decomposition phase is

$$\tfrac{1}{2}\Gamma = \sum_{i=1}^{n} \sum_{j=i}^{n} \mu_{ij}. \quad (1.15)$$

We will be very concerned with the numbers Π and Γ later when we discuss the sparse structure of A and its use by means of Helmert blocking. For the time being, our discussion applies to a sparse system as well as to a full one. We return to eq. (1.13) and write it with

$$A^{-1} = F = (f_{ij})$$

as

$$\zeta_i = -\sum_{j=1}^{n} \sum_{k=1}^{n} f_{ij} x_k \varepsilon_{jk}. \quad (1.16)$$

In order to emphasize that $\varepsilon_{ij} = \varepsilon_{ji}$ we may also write this as

$$\zeta_i = -\sum_{j=1}^{n} f_{ij} x_j \varepsilon_{jj} - \sum_{j=1}^{n} \sum_{k=j+1}^{n} (f_{ij} x_k + f_{ik} x_j) \varepsilon_{jk}. \quad (1.17)$$

Application of the conventional laws of propagation of mean and standard deviation of mutually independent random variables results in the following equations:

$$E\{\zeta_i\} = -\sum_{j=1}^{n} \sum_{k=1}^{n} f_{ij} x_k E\{\varepsilon_{jk}\}. \quad (1.18)$$

$$\sigma^2\{\zeta_i\} = \sum_{j=1}^{n} f_{ij}^2 x_j^2 \sigma^2\{\varepsilon_{jj}\} + \sum_{j=1}^{n} \sum_{k=j+1}^{n} (f_{ij} x_k + f_{ik} x_j)^2 \sigma^2\{\varepsilon_{jk}\} \quad (1.19)$$

$$Cov\{\zeta_{i_1}, \zeta_{i_2}\} = \sum_{j=1}^{n} f_{i_1 j} f_{i_2 j} x_j^2 \sigma^2\{\varepsilon_{jj}\} + \sum_{j=1}^{n} \sum_{k=j+1}^{n} (f_{i_1 j} x_k + f_{i_1 k} x_j) * (f_{i_2 j} x_k + f_{i_2 k} x_j) \sigma^2\{\varepsilon_{jk}\}. \quad (1.20)$$

These formulas are at the basis of the roundoff error estimates. Remember they account for only the local left side errors ε_{ij} that arise during triangular decomposition. More complete formulas are specified in chapter 4. In order to use these formulas, data are needed. Information must be gathered on the elements f_{ij} of the inverse, on the coordinate shifts x_j, and on the number and size of the nonzero elements a_{ij} of A, and the history of these coefficients, as well as of the fill-in coefficients during reduction.

1.3 Data

A roundoff error prediction must come early; otherwise it is useless. On the other hand, at the present early stage of the U.S. network adjustment, many data storage and retrieval facilities are not yet operational. Access to information which would lead to an accurate estimation of the quantities occurring in eqs. (1.18) to (1.20) is limited. In addition, the

boundaries of the various Helmert blocks, subdividing the U.S. network, are not laid out yet. Fortunately, roundoff error estimation means playing with orders of magnitude rather than with specific numbers. The subsequent quantifications of certain properties of the U.S. net must be viewed in this light. It really does not matter, for example, whether the elements f_{ij} of the inverse are overestimated by a factor of three, or if most coordinate shifts x_j are anticipated as being twice as large as they actually will be, or that the counts on Π and Γ will be computed very roughly.

The following properties of the U.S. ground control network are considered to be most relevant to the roundoff study.

(1) Size. As stated earlier, about 175,000 stations are anticipated in the new adjustment. About one-third of these are first- and second-order triangulation stations plus other important stations which contribute significantly to the strength of the network. About two-thirds are supplemental stations which do not significantly strengthen the network. Procedural simplicity was the reason for keeping the supplemental stations in the adjustment, rather than eliminating them prior to the adjustment, as is frequently done in classical network adjustments.

(2) Type of observations. It is estimated that 2 million to 3 million observations will be involved in the new adjustment. About 99 percent will be (unoriented) directions. About 20,000 to 30,000 will include distances and 2,000 to 3,000 azimuths. The positional fix of the network will be established by about 130 Doppler stations.

(3) Inhomogeneity. The density of stations varies from 0 to 3,000 stations per $1° \times 1°$– quad. Figures 5.1a-c illustrate how the station density varies throughout the contiguous United States.

The observational weights are another source of inhomogeneity. Many pairs and some clusters of closely situated stations are tied together by very precise measurements. The corresponding observational weights may be larger by a factor of 100 to 1,000 than the weights of the ordinary observations. As a consequence, some diagonal elements of the normal equation matrix A will be exceptionally large. There will be large off-diagonal elements, too, and they will be arranged in a pattern that results in the system having undesirable numerical properties.

(4) Structure. In some portions, the U.S. network renders the picture of a system of directional arcs. In other portions we deal with an areal directional network. Base lines are arranged at distances that usually do not exceed 100 to 200 km. A unique feature is the transcontinental traverses (TCT). Figure 5.3 shows the TCT loops. About 130 Doppler stations are distributed quite uniformly over the network.

1.3.1 Estimation of the inverse

A finite element model was set up to simulate the global features of the inverse F of the normal equation matrix A. The 350,000 station parameters were replaced by 780 coordinates of the corners of $2° \times 2°$ quads. The loss of a large number of parameters amounts mostly to a loss of local detail which must be guessed from local adjustment and from theoretical insight based on the structural properties of the net. The finite element model is described in section 5.2.

The estimated variances of the $2° \times 2°$ quad corner coordinates range typically from $(0.10\text{m})^2 = 0.01\ \text{m}^2$ in the central areas to $(0.18\text{m})^2 = 0.03\text{m}^2$ near the east and west coasts. There are outliers of $(0.35\text{m})^2 = 0.13\ \text{m}^2$ in Maine and the southern tip of Florida. Covariances taper off moderately at greater distances. Latitude-latitude covariances and longitude-longitude covariances are not as much subdued as the cross-covariances between latitude and longitude. Cross-covariances are mostly below one-tenth the size of the variances. This is important because one-half of all f_{ij}'s are cross-covariances.

The smooth features of the global covariance function resulting from the finite element model must be considered as superimposed by local peaks that account for local structure and for local weaknesses of the network. Peaks have been assumed with amplitudes of up to $(0.35\text{m})^2 = 0.13\ \text{m}^2$. The local covariances have been assumed to be zero at distances exceeding 300 km.

1.3.2 Estimation of the number of nonzero locations and elementary operation steps.

A preliminary Helmert blocking scheme, shown in figure 6.2, was designed and used to obtain a rough idea of Π, the number of nonzero left-side locations, and Γ, which is twice the number of nonzero left-side product terms during triangular decomposition. Essentially the number of stations per $1° \times 1°$-quad was used as input to the counts of Π and Γ. Missing information had to be replaced by hypothetical assumptions. It was found that $\Pi = 1.4 * 10^8$ and $\Gamma = 1.2 * 10^{11}$. These numbers may be off by a factor of three or more. However, it is expected that the true numbers will be smaller than the given ones.

It is not sufficient to count Π and Γ for the whole network. The response to an elementary roundoff error $\varepsilon_{jk}^{(el)}$ associated with a nonzero location is given by $-f_{ij}\, x_j\, \varepsilon_{jj}^{(el)}$ for $j = k$, and by $-(f_{ij}\, x_k + f_{ik}\, x_j) * \varepsilon_{jk}^{(el)}$ for $j<k$. (Recall eq. (1.16) and note that $\varepsilon_{jk} = \varepsilon_{kj}$.) The quantities f_{ij} and x_k are functions of location. Hence, it was necessary to count the number of nonzero locations and operational steps regionally. As building blocks for a regional partition, the $2° \times 2°$ quads

previously used in the simulation model for the inverse were available. It appeared preferable, however, to shift the dividing lines between the $2° \times 2°$ quads in such a way that the nodes of the finite element network became situated at the center of the new $2° \times 2°$ quads. In this way, the global covariance values became representative of all stations situated in the new quads. Section 1.4 will clarify the method for defining the quad based counts and how they were used. Details are given in chapter 6 and 8.

1.3.3 Estimation of the coordinate shifts.

Coordinate shifts which must be expected during the first iteration were estimated in a study by Vincenty (1976). By comparing Doppler positions with current coordinates, and taking into account a datum shift, Vincenty found that the coordinate shifts will mostly be below 5 m. In the Northeastern States and in Montana, larger shifts must be expected. The largest shifts will be below 15m. Vincenty's estimation procedure is given in more detail in section 7.2.1.

1.3.4 Estimation of local roundoff errors during triangular decomposition.

To estimate the mean and standard deviation of a local roundoff error ε_{ij} it is necessary to have a rough knowledge of the size of the operands of the elementary operations that cause the elementary roundoff errors whose superposition is ε_{ij}. It is also necessary to get an idea of the size of the coefficients a_{ij} in the original matrix A and of the product terms $r_{pi}\, r_{pj}$ and the partial sums (1.9) evaluated during triangular decomposition.

If one disregards the very precise measurements which cause the earlier mentioned weight singularities, the size of a typical diagonal element a_{ii} will be around 10^4 m^{-2}. (The normals will actually be scaled to arc seconds of latitude and longitude, but we prefer to scale everything to the meter during our discussion. The scale factors of f_{ij} cancel those of ε_{ij} in eqs. (1.16) to (1.20).) The off-diagonal elements a_{ij} are of similar magnitude if they are not bound to be zero.

The weight singularities cause some diagonals to be excessively large, up to 10^6–10^7. There are also off-diagonals of this size, but all large coefficients will be confined to small submatrices of A, each having only a few rows and columns.

The product terms $r_{pi}\, r_{pj}$ and the partial sums (1.9) can best be understood if their geodetic meaning is revealed. This is best done by introducing the "parially reduced" coefficients $a_{ij}^{(p)}$ as

$$a_{ij}^{(p)} = a_{ij} - \sum_{k=1}^{p} r_{ki}\, r_{kj}, \quad p<i. \tag{1.21}$$

These coefficients appear, together with $b_i^{(p)}$, if Cholesky's algorithm is organized in a different way, namely with $a_{ij}^{(0)} = a_{ij}$ and $b_i^{(0)} = b_i$ as:

$$\left.\begin{array}{ll} r_{pp} = \sqrt{a_{pp}^{(p-1)}} & \\ r_{pj} = a_{pj}^{(p-1)} / r_{pp}, & j = p+1, \ldots, n \\ s_p = b_p^{(p-1)} / r_{pp} & \\ \left.\begin{array}{l} a_{ij}^{(p)} = a_{ij}^{(p-1)} - r_{pi}\, r_{pj} \\ b_i^{(p)} = b_i^{(p-1)} - r_{pi}\, s_p \end{array}\right\} & \begin{array}{l} i = p+1, \ldots, n \\ j = i, \ldots, n \end{array} \end{array}\right\} p = 1, \ldots, n. \tag{1.22}$$

Executed in this way, Cholesky's algorithm differs from Gauss's algorithm only in the respect that the square roots of the diagonals are taken. We can now make the following statements:

$a_{ii}^{(p)}$, $i>p$, is the reciprocal of the variance of coordinate i, provided that the coordinates k, $p<k\leqslant n$, $k\neq i$ are fixed, while coordinates k, $1\leqslant k\leqslant p$, as well as coordinate i itself, are allowed to vary freely.

$-a_{ij}^{(p)}/a_{ii}^{(p)}$, $i,j>p$, $i\neq j$ is the shift, with respect to the *adjusted* position, suffered by coordinate i if coordinate j is displaced by one unit from its adjusted position, if coordinates k, $p<k\leqslant n$, $k\neq i,j$ are fixed to their adjusted position, while coordinates k, $1\leqslant k\leqslant p$ as well as coordinate i itself, are allowed to vary freely.

Based on this insight into the geodetic meaning of the partially reduced coefficients $a_{ij}^{(p)}$, it is comparatively easy to make qualitative statements about the history of these quantities during triangular decomposition. Quantitative statements, on the other hand, are more difficult to make. One has to rely on results of test calculations and on judgment. Chapter 7 deals with the estimation of the $a_{ij}^{(p)}$-coefficients. Here we only briefly summarize some essential findings.

(1) Diagonal elements $a_{ii}^{(p)}$ always decrease when p increases. This is obvious from eq. (1.21) Therefore, we may bound the diagonals $a_{ii}^{(p)}$ in terms of $a_{ii}^{(0)} = a_{ii}$, i.e., in terms of the diagonals of the original normals.

(2) If coordinate i belongs to a station which is not involved in a high-precision measurement, the history of $a_{ij}^{(p)}$, $i \leqslant j$, $0\leqslant p\leqslant i-1$ evolves without major drama. The diagonals $a_{ii}^{(p)}$ will decrease in most cases to about one-fourth to three-fourths of their original size. Only station coordinates that are eliminated at the very end of triangular decomposition are an exception. There, diagonals may drop to about 10^1. Off-diagonals $a_{ij}^{(p)}$, such that $a_{ij}\neq 0$, will not change drastically either. Fill-in coefficients $a_{ij}^{(p)}$, such that $a_{ij} = 0$, will be small in most cases.

(3) If i is a coordinate of a station involved in a high-precision measurement, then very few large coefficients are included in the equations belonging to station i. These coefficients may or may not drop sharply in size before the square root of the diagonal is taken. However, in any cluster of tightly connected stations, there will always be a few coordinates with coefficients that drop sharply. The details of how this all interacts with roundoff via the partial sums in eq. (1.21) are not so easily explained. In this context, some features of Helmert blocking which were suppressed in the introductory section come into play. We merely indicate, therefore, that the adverse effect of the large coefficients upon roundoff is very much subdued if all stations of a tightly connected cluster are treated in immediate succession. The adverse effect could be reduced to a minimum if certain modifications were made to the NGS Cholesky algorithm. However, we believe that even without these modifications, which would slow down the computation somewhat, the solution of the normals for the U.S. network is a safe procedure.

It is very important to note that the number of very large coefficients in the original normals does not increase significantly during the process of triangular decomposition. This can be inferred from the geodetic interpretation of the coefficients. Test calculations even suggest a decrease in their number. This effect is certainly due to the strong reduction in size of some large coefficients.

1.4 Evaluation

It is clear that formulas (1.18) and (1.19) for mean and standard deviation of the global roundoff errors, ζ_i, cannot be evaluated on a term-by-term basis. There are simply too many terms and there is too little information on each term. Several roads are open for simplifying the evaluation of (1.18) and (1.19). We have followed some of them in this study. To illustrate the various possibilities, focus attention again on a single elementary roundoff error, $\varepsilon_{ij}^{(el)}$, affecting a coefficient a_{ij}. Note that the local roundoff error, ε_{ij}, which is an entry of the perturbation matrix ε, is already a superposition of $2\mu_{ij}$ elementary roundoff errors. Recall that μ_{ij} is the number of nonzero product terms to be evaluated and subtracted from a_{ij}, and that any product term involves two elementary operations, one multiplication and one addition. Thus, the total number of elementary operations is about equal to Γ. The global response to one elementary roundoff error $\varepsilon_{jk}^{(el)}$ is given by eq. (1.13), or, in more detailed notation, by

$$\zeta_i = \begin{cases} -f_{ij}\,\varepsilon_{jj}^{(el)}\,x_j & \ldots \quad j=k \\ -f_{ij}\,\varepsilon_{jk}^{(el)}\,x_k & -f_{ik}\,\varepsilon_{jk}^{(el)}\,x_j \ldots j<k \end{cases} \tag{1.23}$$

Suppose that the mean and standard deviation of all elementary roundoff errors is bounded as

$$|E\{\varepsilon_{ij}^{(el)}\}| \leqslant \frac{c}{2}\beta^{-\tau} \tag{1.24}$$

$$\sigma\{\varepsilon_{ij}^{(el)}\} \leqslant \frac{c}{\sqrt{12}}\beta^{-\tau}. \tag{1.25}$$

Suppose further that $\|f\|$ is a bound on the elements f_{ij} of the inverse F, and that $\|x\|$ is a bound on the elements of the solution vector:

$$|f_{ij}| \leqslant \|f\|, \quad |x_k| \leqslant \|x\|. \tag{1.26}$$

The global response (1.23) to one elementary roundoff error is then bounded as:

$$|E\{\zeta_i\}| \leqslant c\,\|f\|\,\|x\|\,\beta^{-\tau} \tag{1.27}$$

$$\sigma\{\zeta_i\} \leqslant \frac{2c}{\sqrt{12}}\|f\|\,\|x\|\,\beta^{-\tau}. \tag{1.28}$$

Since there are Γ elementary roundoff errors, we get

$$|E\{\zeta_i\}| \leqslant c\,\|f\|\,\|x\|\,\Gamma\,\beta^{-\tau} \tag{1.29}$$

$$\sigma\{\zeta_i\} \leqslant \frac{2c}{\sqrt{12}}\|f\|\,\|x\|\,\sqrt{\Gamma}\,\beta^{-\tau}. \tag{1.30}$$

Primitive estimates based on these formulas are specified in section 4.1.4. They already indicate feasibility of the adjustment on the CDC 6600. However, the bias estimate in (1.29) comes out too large for the IBM 360. Note that (1.29) contains the factor Γ while formula (1.30) for the standard deviation has $\sqrt{\Gamma}$ as a factor. The advantage of a truly rounding machine becomes apparent!

The weakness of the primitive estimates (1.29) and (1.30) comes from the weight singularities. The constant c in (1.27) and (1.28) must be chosen in agreement with the largest elementary roundoff errors, and those are associated with the largest coefficients a_{ij} of A and its reduction states.

Much better estimates are obtained if the coefficients $a_{ij}^{(p)}$ are divided into two size classes. The first class contains the large coefficients, the second contains those of moderate and small size. Separate bounds for size and number of coefficients in these two classes must be specified. Inequalities (1.29) and (1.30) are then evaluated separately for the two size classes, and the results of both are superimposed. This procedure comes close to that used in chapter 8, where safe bounds could be obtained for the global roundoff errors, indicating feasibility of the adjust-

ment on both types of computers. The formulas in chapter 8 are more complicated because regional variations of f_{ij} and x_k have also been taken into account. Denote by ϱ,χ the individual $2° \times 2°$ quads of the partition of the U.S. net. Denote by $f_{\varrho\chi}$ the smoothed global covariance entries. Denote by $\Gamma_\varrho^{(r)}$ the number of elementary steps associated with a location (i,j), such that coordinate i refers to a station situated in quad ϱ. Denote by $\Gamma_\varrho^{(c)}$ the number of operational steps associated with a location (i,j), such that coordinate j refers to a station situated in quad ϱ. $\Gamma_\varrho^{(r)}$ and $\Gamma_\varrho^{(c)}$ are precisely the regional Γ-counts announced earlier in section 1.3. For simplicity, assume a single size class is representative of the whole network. Then

$$|E\{\zeta_\varrho\}| \leqslant \frac{c}{2} \|x\| \sum_\chi \|f_{\varrho\chi}\| \{\Gamma_\chi^{(r)} + \Gamma_\chi^{(c)}\} \beta^{-\tau} +$$

$$+ \text{ a contribution due to the local peak of } f_{ij} \qquad (1.31)$$

$$\sigma\{\zeta_\varrho\} \leqslant \frac{\sqrt{2}c}{\sqrt{12}} \|x\| \sqrt{\sum_\chi \|f_{\varrho\chi}\|^2 \{\Gamma_\chi^{(r)} + \Gamma_\chi^{(c)}\}} \beta^{-\tau} +$$

$$+ \text{ a contribution due to the local peak of } f_{ij}. \qquad (1.32)$$

Chapter 8 gives more complex equations, which also account for regional variations of x_k.

Because there is a tapering effect, the error estimates are reduced by taking into account regional variations of f_{ij}. For coordinates (i,j) that belong to two stations at a larger distance, f_{ij} tends to be smaller in size than for nearby stations. Because there are many more pairs of distant stations than there are nearby pairs, the improvement is noticeable. There is, however, another taper effect, namely one associated with the coefficients $a_{ij}^{(\rho)}$ of A and its reductions states. This taper effect implies that the local roundoff errors ε_{ij} tend to be smaller if coordinates i,j are widely spaced. In chapter 9 we have capitalized on this effect. Since judgment and plausibility considerations have been employed there that exceed the threshold of what even a practical mathematician would consider "safe," the results obtained in chapter 9 are declared to be only 95 percent safe.

It is possible to take into account the taper effect of the $a_{ij}^{(\rho)}$'s in a mathematically satisfactory way. However, it turns out that the formulas obtained do not result in improved estimates for the U.S. network. In section 11.2 we have specified asymptotic formulas for homogeneous large networks. These formulas contain unspecified constants and show how certain bounds on bias and standard deviation of the global roundoff errors ζ_i grow in proportion to the number of stations.

1.5 Results

Before we state the results of this roundoff study, let us briefly summarize some essential assumptions upon which the results depend. We include also assumptions on the right-hand-side coefficients and on other matters played down in this chapter.

(1) The results apply to the U.S. network as NGS personnel described it to me in 1977. The features listed in section 1.3 are essential. Other networks would have a different buildup of roundoff errors.

(2) The study is concerned with only those roundoff errors which arise and accumulate during the solution of the normal equations and not during their formation. To be precise, it is assumed that the normals of the lowest level Helmert blocks are without error. I do not anticipate that the roundoff errors that occur during formation of the lowest level normals will falsify the results more than those treated here.

(3) Helmert blocking is done in such a way that fill-in is effectively kept down. Our estimates of Π,Γ should by no means be surpassed by a factor of five or more. I believe that a judicious choice of block boundaries (in regions of low station density) will result in smaller Π,Γ than our estimates indicate.

(4) Seventy-five percent of the diagonals a_{ii} are below $1.8 \bullet 10^4$ m^{-2} (if normals are considered as scaled to the meter). Twenty-five percent of the diagonals may go up to $4.5 * 10^6$. These large diagonals must be associated with pairs or small clusters of tightly connected stations.

(5) The approximate coordinates of tightly connected clusters of stations must be in near agreement with the precise measurements which cause the strong ties. This will cause the right-hand sides of equations with large diagonals to be around 10^5 m^{-1} or below. In any case, approximate coordinates must be good enough that the right-hand sides do not exceed 10^5m^{-1}

Under these assumptions, the following safe bounds for the global roundoff errors ζ_i suffered by coordinate i during the *first* iteration have been derived by a procedure documented in chapter 8:

Bias:

$$|E\{\zeta_i\}| \begin{cases} = 0 & \ldots \text{ on the CDC 6600} \\ < .002 \text{ m} & \ldots \text{ on the IBM 360} \end{cases} \qquad (1.33)$$

Standard deviation:

$$\sigma\{\zeta_i\} \begin{cases} < .00012 \text{ m} & \ldots \text{ on the CDC 6600} \\ < .0000013 \text{ m} & \ldots \text{ on the IBM 360.} \end{cases} \qquad (1.34)$$

During the first iteration, maximum coordinate shifts exceeding 10 m are anticipated. Our estimates may be reformulated by stating that about three digits of the largest shift will be correct on the IBM 360, and four to five digits of the largest shift will be correct on the CDC 6600. In this formulation, the estimates also hold for the subsequent iterations in which the maximum shift will be much smaller. Accuracy does not increase indefinitely though. Eventually the errors created during formation of the normals will dominate the errors during solution.

If, as I have been assured repeatedly, the proportion of large diagonals is not above 10 percent and if a less conservative estimation procedure, documented in chapter 9, is used, the following estimates can be obtained. However, only a 95 percent probability is assigned to their validity.

Bias:

$$|E\{\zeta_i\}| \begin{cases} = 0 & \ldots \text{on the CDC 6600} \\ < .00005 & \ldots \text{on the IBM 360} \end{cases} \tag{1.35}$$

Standard deviation:

$$\sigma\{\zeta_i\} \begin{cases} < .00001 & \ldots \text{on the CDC 6600} \\ < .0000001 & \ldots \text{on the IBM 360.} \end{cases} \tag{1.36}$$

For stations situated close together the relative position, i.e., the differences in latitude and longitude, will be less perturbed by roundoff than the absolute positions to which our above estimates refer. If two stations are separated by 20 km or less, I estimate that the roundoff bias affecting the relative position is smaller by a factor of 1/10 to 1/100 than the roundoff bias affecting global positions. For standard deviations the improvement is more modest and may amount to a factor of one-third to one-tenth.

A word of caution must be expressed about $E\{\zeta_i\} = 0$ on the CDC 6600. Since rounding on the CDC 6600 is not completely true, and because $E\{\zeta_i\} = 0$ relies heavily on the linearity of the roundoff model, which hypothesis is also not completely true, one must be aware of the possibility of a small residual bias. It cannot be entirely excluded that the residual bias will be even larger than the specified standard deviations $\sigma\{\zeta_i\}$. Nevertheless it is believed that four correct digits of the largest coordinate shift will be recovered. (See section 8.3 for a discussion of residual bias.)

Remark: In section 11.3, the estimates for the IBM 360 have been modified for the UNIVAC 1100/40. An improvement of one decimal digit results for the estimates on $E\{\zeta_i\}$, $\sigma\{\zeta_i\}$. Thus, the global bias $E\{\zeta_i\}$ will not exceed 0.0002 m after the first iteration of the adjustment. Recall that the first iteration will produce coordinate shifts exceeding 10 m. Hence four correct leading decimal digits of the largest coordinate shift can be guaranteed during any iteration. There is a good chance that the actual errors will be smaller by a factor of 1/10 to 1/100. Errors in relative position of two stations at a close distance are expected to be smaller than the global errors by a factor of 1/10 to 1/100.

1.6 Verification

In order to establish confidence in the specified estimates, a number of test calculations involving a small network of about 1300 stations were run on the CDC 6600. This machine provided the advantage of switching between biased and unbiased arithmetic. By using double precision (two 60-bit computer words for one number) in one of the adjustment runs, a high-precision solution vector could be obtained. It served as an absolute basis of comparison and allowed the calculation of true roundoff errors in the case of chopping or rounding single precision arithmetic (48-bit mantissa). The NGS Cholesky algorithm was temporarily modified to produce estimates of $E\{\zeta_i\}$, $\sigma\{\zeta_i\}$ that were calculated from the actual sizes of the product terms $r_{ki}r_{kj}$ and the partial sums of these product terms as they were available during triangular decomposition. Comparison with the true roundoff errors allowed us to check the validity of our roundoff model. The outcome of the test was considered satisfactory. An additional benefit of the test calculations was obtained in the statistics on the size and number of nonzero coefficients and the number of elementary operation steps. These numbers can be compared with those predicted by the idealized counting models described in chapter 6. Chapter 10 documents the test calculations in detail. Reference is also made to some other roundoff tests which are reported in the geodetic literature.

2. ROUNDOFF ERRORS AND WAYS TO ANALYZE THEM

2.1 Elementary Roundoff Errors

A computer can perform four elementary arithmetic operations: addition, subtraction, multiplication, and division. Let the symbol $\diamond$ stand for $+$, $-$, $*$, $/$, respectively. Mathematically, any elementary operation combines two operands a, b, and yields the results $a \diamond b$, which is unique in the field of real numbers, division by zero excluded. The computer only approximates the mathematical truth. It will give $a \circledcirc b$ instead of $a \diamond b$. The difference

$$\varepsilon = a \circledcirc b - a \diamond b \tag{2.1}$$

is called *elementary roundoff error.* Note that the definition of ε does not reflect any previous roundoff errors that may have already affected the two operands a, b.

Different computers will produce different roundoff errors. We will restrict ourselves to the discussion of roundoff occurring during calculations done in normalized floating point arithmetic with fixed length mantissas. In the next subsection we will summarize the features of this mode of calculation. Subsequently, we will discuss elementary roundoff errors on two different families of computers, the first represented by the CDC 6600 and the second by the IBM 360. Sterbenz (1974) gives more details.

2.2 Normalized Floating Point Arithmetic With Fixed Length Mantissas

A normalized floating point number a is represented as

$$a = m * \beta^e. \tag{2.2}$$

Only the two components m (mantissa) and e (exponent) are stored in a computer word. Both m and e may have a sign. The symbol β stands for the base of the number system used by the computer; β equals 2 for the CDC 6600 and 16 for the IBM 360. The mantissa m, which may be viewed as an integer, has τ digits. The leftmost digit of m is nonzero (only the number zero is an exception in normalization). There are limitations on the size of exponent e which are irrelevant for the present work.

Given two operands with mantissas of length τ, it is clear that the precise calculation of $a \diamond b$ would require a mantissa of unlimited length, if $\diamond$ stands for $+$, $-$, $/$. In the case of $a * b$, a 2τ-digit mantissa would be sufficient. Since only τ digits of $a \diamond b$ can be stored, the result of an elementary operation must be modified in some way and it becomes the rounded result $a \circledcirc b$.

For the purpose of this study, roundoff errors occurring during addition and subtraction must be thoroughly understood. Floating point addition starts by making the exponents of the two operands equal. If $a = m_a * \beta^{e_a}$ and $b = m_b * \beta^{e_b}$, and if $e_a \geqslant e_b$, then the number b will be converted to the unnormalized form $b = m'_b * \beta^{e_a}$. The unnormalized mantissa m'_b differs from the normalized one, m_b, by a right-shift of $e_a - e_b$ places. If e_a is truly greater than e_b, the shift is nonzero, and is called the "preshift." The preshift may already be the reason for the loss of some digits of b. After the preshift, the sum $m_a + m'_b$ is formed in the accumulator. If overflow occurs, a right-shift of one place is performed, and the exponent e_{a+b}, which is originally assumed as e_a, is increased by one. If, on the other hand, $m_a + m'_b$ has leading zeroes, then a left-shift is performed in order to ensure normalization. The exponent is then decreased by the number of the shifted places. Shifts occurring after the formation of $m_a + m'_b$ are called "postshifts." Finally, the mantissa of the result must be shortened to τ digits before it can be transported to any of the memory locations. The final result is $a \oplus b$.

In our discussion we have excluded the case where one or both of the operands or the result is zero. Since no roundoff error occurs in these cases, they are of no interest to us. Other details, not given in the preceding paragraph, differ from machine to machine, and we shall be more specific when we discuss the CDC 6600 and the IBM 360 separately. Before we do so, we consider two ideal cases of rounding which will be called "true chopping" and "true rounding."

Suppose that the precise result of $a \diamond b$ is available and that it is normalized. True chopping simply disregards all but the leading τ digits of the mantissa. True rounding first adds a rounding digit to position $\tau + 1$, and then disregards this position and anything to the right of it. The result is rounding as it is understood in common language. For the sake of greater clarity we prefer the term "true rounding" because in the technical literature rounding stands for any procedure that replaces $a \diamond b$ by a τ-digit substitute $a \circledcirc b$.

Note that true chopping results in a number whose absolute value is less or equal to that of the unchopped one. True rounding can result in an increase or a decrease of the absolute value of the unrounded number.

2.3 Rounding on the CDC 6600

The CDC 6600 uses the binary number system. Accordingly the base is $\beta = 2$. The mantissa has $\tau = 48$ binary digits. The result of any arithmetic instruction is formed in a special register, the accumulator, which can accommodate mantissas of $2\tau = 96$ digits. This result is shortened to τ digits before it is stored away. A slight difficulty arises after addition or subtraction, because an eventual left postshift is suppressed. The result is stored away in possibly unnormalized form with τ digits and possible leading zeroes. A separate normalizing instruction is available, and we assume that it is always used immediately after any addition or subtraction. Evidently the CDC FORTRAN compiler applies this instruction automatically. Hence we may assume that all numbers in the central memory are normalized.

The standard CDC 6600 instruction set chops the result obtained in the accumulator. However, a modified instruction set is also available which comes near to truly rounding the result. The FORTRAN

compiler can be prompted to use this modified instruction set. Hence we may view the CDC 6600 as a truly rounding machine. There are some flaws. One of them is that after addition and subtraction, rounding always precedes an eventual left postshift which is done by a separate instruction as mentioned above. Hence the roundoff error is generally larger for a non-zero left postshift than it would be for true rounding without flaw.

2.4 Rounding on the IBM 360

The base of the number system is $\beta = 16$. The IBM 360 computes in the hexadecimal number system. The mantissa length is $\tau = 14$. The cells of the central memory cannot hold more than 14 hexadecimal digits which correspond to 56 binary digits. The registers which hold the operands and the result of an arithmetic operation are enlarged by one digit to the right, the so-called guard digit. If a preshift occurs, the guard digit of the operand with smaller magnitude may become nonzero. A preshift by more than one place can result in chopping off anything to the right of the guard digit. After the 15-digit result of an operation is formed, the guard digit is chopped off. In most cases the final result is equivalent to the one obtained by true chopping. In a few cases the result will differ from the truly chopped one by a 1 in the last position. The result then will be larger in magnitude than $a \diamond b$. Let us illustrate this under the assumption that $a-b$ is wanted, and that $a > b > 0$.

A right preshift of b by more than one place will make b smaller because b must be chopped (there is only one guard digit). Chopping b tends to make the result $a-b$ larger than it really is. On the other hand, $a-b$, after it has been formed in the accumulator, is chopped again, which tends to make it smaller. The combined effect of the two choppings that compete with each other may in some cases cause $a \ominus b$ to be larger than $a-b$; in some cases $a \ominus b$ will still be smaller or equal to $a-b$.

2.5 Local Roundoff Errors

The solution of a linear system of equations is done by a great number of elementary steps. For the U.S. network the number of steps is of the order 10^{11}. Frequently, the roundoff errors occurring during a number of individual steps will have a similar global effect on the final result, and the analysis may be simplified if the roundoff errors of such a batch of individual steps are treated collectively. Usually such a batch of steps will be executed either successively, or it will affect a specific important intermediate quantity. We will call the combined roundoff error of such a collection of steps a "local" roundoff error. Typical examples follow.

(1) The result of the calculation of a *square root.* The square root is usually calculated by a subroutine. It does not make much sense to analyze separately the roundoff effect of each individual step executed by this subroutine. Only the combined effect upon the quantity returned by the subroutine, i.e., the square root is of interest. Hence, we will make assumptions on the roundoff error suffered by the square root, based on experience or given in the reference manual of the subroutine.

(2) Accumulation of *inner products.* Frequently, in particular during the execution of Cholesky's algorithm, inner products of the form

$$p = \sum_{i=1}^{n} a_i * b_i \tag{2.3}$$

have to be calculated. Subsequent steps will use only p as input, but never any partial sum occurring during the calculation of p. Hence the final result of the entire calculation will depend on only the accumulated effect of roundoff errors onto p. On the other hand, this accumulated roundoff error will depend very much on the size of a_i, b_i and on the partial sums

$$p_j = \sum_{i=1}^{j} a_i * b_i. \tag{2.4}$$

Hence, case (2) differs somewhat from case (1). The concept of a local roundoff error affecting p does not save much labor. It merely makes the analysis easier to overlook. First, we have to determine how p is affected by roundoff. Once this is done, we need only to remember the roundoff error of p and forget those committed during the intermediate steps of the calculation of p. In this way, the number of quantities which must be carried along during the roundoff analysis is conveniently decreased.

Remark. Once certain roundoff errors have been accumulated into batches with a resulting local roundoff error, there is no reason to distinguish further between these local roundoff errors and the remaining unbatched elementary ones. We may call them all local roundoff errors in the wider sense, or again simply local roundoff errors. This will simplify the terminology in the discussion of global roundoff errors.

2.6 Linear Roundoff Error Propagation

Wide sense local roundoff errors affect the input to subsequent computational steps. Roundoff errors propagate. The final result is falsified by a global roundoff error which is the result of all the local ones and their propagation. The precise law of roundoff error propagation is extremely difficult to describe and to handle. Simplifying assumptions must be

made. One way to simplify the analysis is to try to establish bounds on the global roundoff errors, such as those advocated by Wilkinson and his school. Wilkinson-type estimates yield precise, deterministic bounds which in many cases turn out to be overly pessimistic. Another approach, advocated and theoretically justified by Tienari (1970), is to linearize the laws of roundoff error propagation. We will follow this line of reasoning and assume that any wide sense local roundoff error ε_i affects the final result as $c_i * \varepsilon_i$, where c_i is a vector of as many components as there are numbers representing the final result. The total effect of all elementary and local roundoff errors onto the final result is obtained as

$$\zeta = \sum_i c_i * \varepsilon_i. \tag{2.5}$$

The sum is extended over all local roundoff errors. The above formula is remarkable in some respects and deserves further discussion. At first sight it resembles a familiar formula frequently used when the error of the result of a certain formula expression due to errors in the input data is to be analyzed. The analogy is partially misleading. Roundoff errors do not depend on initial data alone; they also depend very much on the sequence of elementary steps which are executed during the evaluation of the formula. Specifying a mathematical expression such as

$$x_2 = \frac{b_2 - \dfrac{a_{12} * b_1}{a_{11}}}{a_{22} - \dfrac{a_{12}^2}{a_{11}}} \tag{2.6}$$

is not sufficient in order to analyze the roundoff errors. It must be specified in which particular sequence the individual operations are performed. A computer algorithm breaks down the formula into a sequence of elementary steps, and thus lends itself to roundoff analysis.

By the way, the above formula expresses the solution for the second unknown x_2 of the 2×2 symmetric and assumedly positive definite system of linear equations:

$$\begin{aligned} a_{11} x_1 + a_{12} x_2 &= b_1 \\ a_{21} x_1 + a_{22} x_2 &= b_2. \end{aligned} \tag{2.7}$$

If Cholesky's algorithm is used to solve the system, we should rewrite the above formula as eq. (2.8).

But even this is ambiguous. The subsequent algorithm, eq. (2.9), removes any doubt in which sequence the elementary steps are performed.

$$x_2 = \frac{\dfrac{b_2 - \dfrac{a_{12}}{\sqrt{a_{11}}} * \dfrac{b_1}{\sqrt{a_{11}}}}{\sqrt{a_{22} - \left[\dfrac{a_{12}}{\sqrt{a_{11}}}\right]^2}}}{\sqrt{a_{22} - \left[\dfrac{a_{12}}{\sqrt{a_{11}}}\right]^2}} \quad . \tag{2.8}$$

$$\begin{aligned}
&(1)\ r_{11} = \sqrt{a_{11}} \\
&(2)\ r_{12} = a_{12}/r_{11} \\
&(3)\ s_1 = b_1/r_{11} \\
&(4)\ a_{22}^{(1)} = a_{22} - r_{12} * r_{12} \\
&(5)\ b_2^{(1)} = b_2 - r_{12} * s_1 \\
&(6)\ r_{22} = \sqrt{a_{22}^{(1)}} \\
&(7)\ s_2 = b_2^{(1)}/r_{22} \\
&(8)\ x_2 = s_2/r_{22}.
\end{aligned} \tag{2.9}$$

The symbols r_{11}, r_{12}, r_{22}, s_1, s_2, $a_{22}^{(1)}$, $b_2^{(1)}$ denote intermediate results during execution of the algorithm. The notation is motivated by features of Cholesky's algorithm which transforms the system

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \tag{2.10}$$

into the triangularized system

$$\begin{bmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} \tag{2.11}$$

whereby

$$\begin{bmatrix} r_{11} & r_{12} \\ 0 & a_{22}^{(1)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} s_1 \\ b_2^{(1)} \end{bmatrix} \tag{2.12}$$

is an intermediate step. We will deal with Cholesky's algorithm more systematically later. For the moment, let us try to identify the local roundoff errors. On a real-life computer, the algorithm formulated earlier would be carried out in the following perturbed way:

$$\begin{aligned}
&(1)\ \tilde{r}_{11} = \sqrt{a_{11}} + \varepsilon_1 \\
&(2)\ \tilde{r}_{12} = a_{12}/\tilde{r}_{11} + \varepsilon_2
\end{aligned} \tag{2.13}$$

$$
\begin{aligned}
&(3)\ \tilde{s}_1 = b_1/\tilde{r}_{11} + \varepsilon_3 \\
&(4)\ \tilde{a}_{22}^{(1)} = a_{22} - \tilde{r}_{12} * \tilde{r}_{12} + \varepsilon_4 \\
&(5)\ \tilde{b}_2^{(1)} = b_2 - \tilde{r}_{12} * \tilde{s}_1 + \varepsilon_5 \\
&(6)\ \tilde{r}_{22} = \sqrt{\tilde{a}_{22}^{(1)}} + \varepsilon_6 \\
&(7)\ \tilde{s}_2 = \tilde{b}_2^{(1)}/\tilde{r}_{22} + \varepsilon_7 \\
&(8)\ \tilde{x}_2 = \tilde{s}_2/\tilde{r}_{22} + \varepsilon_8.
\end{aligned}
\tag{2.13}
$$

The roundoff errors $\varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5, \varepsilon_7, \varepsilon_8$ are elementary roundoff errors as they were introduced in section 2.1. The errors $\varepsilon_1, \varepsilon_6$ are local roundoff errors in the narrow sense. (See section 2.5.) All the ε's can be viewed as local roundoff errors in the wider sense. We have used tildes to denote quantities which are already affected by previous roundoff errors. The final, or global, roundoff error suffered by x_2 is

$$\zeta_2 = \tilde{x}_2 - x_2. \tag{2.14}$$

According to the linearity assumption, this error is represented as

$$\zeta_2 = \sum_{i=1}^{8} c_i \varepsilon_i. \tag{2.15}$$

Let us illustrate how the c_i's can be determined.

Two principles underlie any linear error analysis. These principles are (1) *isolation* of error sources, and (2) *linear superposition* of the isolated effects. The two principles allow us to analyze any ε_i separately such as if ε_i, for a particular i, is the only local roundoff error that occurs. Its effect on the final result is $c_i \varepsilon_i$. We can do this for all ε_i's in succession, and finally superimpose the $c_i \varepsilon_i$ linearly.

As an example, let us try to determine c_2. We assume that ε_2 is the only local roundoff error which occurs. Once it has occurred, we propagate it through the subsequent steps of the algorithm, using linear approximation whenever necessary. We get

$$
\begin{aligned}
&(1)\ r_{11} = \sqrt{a_{11}} \\
&(2)\ \tilde{r}_{12} = a_{12}/r_{11} + \varepsilon_2 \\
&(3)\ s_1 = b_1/r_{11} \\
&(4)\ \tilde{a}_{22}^{(1)} = a_{22} - \tilde{r}_{12}^2 = a_{22} - (r_{12} + \varepsilon_2)^2 = a_{22}^{(1)} - 2r_{12}\varepsilon_2 \\
&(5)\ \tilde{b}_2^{(1)} = b_2 - \tilde{r}_{12} s_1 = b_2 - (r_{12} + \varepsilon_2) s_1 \\
&\qquad = b_2^{(1)} - s_1\varepsilon_2 \\
&(6)\ \tilde{r}_{22} = \sqrt{\tilde{a}_{22}^{(1)}} = \sqrt{a_{22}^{(1)} - 2r_{12}\varepsilon_2} = r_{22} - (r_{12}/r_{22})\varepsilon_2 \\
&(7)\ \tilde{s}_2 = \tilde{b}_2^{(1)}/\tilde{r}_{22} = (b_2^{(1)} - s_1\varepsilon_2)/(r_{22} - (r_{12}/r_{22})\varepsilon_2) = \\
&\qquad = s_2 - (s_1/r_{22} - b_2^{(1)} r_{12}/r_{22}^3)\varepsilon_2 \\
&(8)\ \tilde{x}_2 = \tilde{s}_2/\tilde{r}_{22} = (s_2 - (s_1/r_{22} - b_2^{(1)} r_{12}/r_{22}^3)\varepsilon_2)/ \\
&\qquad (r_{22} - (r_{12}/r_{22})\varepsilon_2) \\
&\qquad = x_2 - (s_1/r_{22}^2 - b_2^{(1)} r_{12}/r_{22}^4 - s_2 r_{12}/r_{22}^3)\varepsilon_2.
\end{aligned}
\tag{2.16}
$$

Hence, we obtain

$$c_2 = -\frac{s_1}{r_{22}^2} + \frac{b_2^{(1)} r_{12}}{r_{22}^4} + \frac{s_2 r_{12}}{r_{22}^3} \tag{2.17}$$

This is a rather complicated expression which is shown here to illustrate the principle of linearization. However it is clear that the above method of deriving c_i's would not work for a system of several hundred thousand unknowns. More efficient and systematic methods must be used. They are available through the so-called principle of backward analysis, invented by Wilkinson, which is outlined in the next section. See also Wilkinson (1963).

2.7 Backward Analysis

Let us now return to the linear system

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \tag{2.18}$$

which, in matrix form, is written as

$$Ax = b. \tag{2.19}$$

The solution is expressed as

$$x = A^{-1} b. \tag{2.20}$$

The reason why the roundoff error analysis of the previous section turned out to be so complicated was that we tried, by a method called forward analysis, to propagate the roundoff error of step 2 through all the algebraically complicated steps that followed. Wilkinson pointed out that it is frequently simpler to trace local roundoff errors backwards through the previous steps of the algorithm until the original input data are reached. Then the result is a perturbed set of the original equations that is analyzed by conventional perturbation methods. In the case of our linear system, the perturbed set of equations is

$$(A + \varepsilon)(x + \zeta) = b + \eta \tag{2.21}$$

and the resulting perturbation of the solution is obtained in linear approximation as

$$\zeta = -A^{-1}\varepsilon x + A^{-1}\eta. \tag{2.22}$$

Note that ε denotes a matrix and that ζ, η denote vectors. To illustrate, let us consider again a single local roundoff error ε_2, which now will be traced backwards. Tracing ε_2 backwards is simple enough indeed, and it is immediately seen that the effect of ε_2 on the solution is the same as if the original system were perturbed in the following way:

$$\begin{bmatrix} a_{11} & a_{12} + r_{11}\varepsilon_2 \\ a_{21} + r_{11}\varepsilon_2 & a_{22} \end{bmatrix} \begin{bmatrix} x_1 + \zeta_1 \\ x_2 + \zeta_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \tag{2.23}$$

Note that the assumption of symmetry is at the basis of Cholesky's algorithm. Hence any error traced back to a_{12} must necessarily also affect $a_{21}=a_{12}$. In the notation introduced earlier we had

$$\varepsilon = \begin{bmatrix} 0 & r_{11}\varepsilon_2 \\ r_{11}\varepsilon_2 & 0 \end{bmatrix} \qquad \eta = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{2.24}$$

Consequently

$$\zeta = -\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}^{-1} \begin{bmatrix} r_{11}\varepsilon_2 x_2 \\ r_{11}\varepsilon_2 x_1 \end{bmatrix} \tag{2.25}$$

which leads to

$$\zeta_2 = -\frac{-a_{12}x_2 + a_{11}x_1}{a_{11}a_{22} - a_{12}^2} r_{11}\varepsilon_2. \tag{2.26}$$

Backward analysis is not always simpler than forward analysis, but in the case of a large linear system it is an indispensable tool.

2.8. Stochastic Assumptions on Roundoff Errors

We will treat elementary roundoff errors as mutually uncorrelated random variables. Theoretical support for this assumption is found in Feldstein (1976). Henrici (1964) gives a good presentation at an introductory level. We do not assume that the coefficients of the original equations and of the various reduction states are random variables, at least not as far as their leading digits are concerned. However, the lower order digits of these coefficients will be rather random, and so will roundoff that results from combining the coefficients by arithmetic operations. We will be interested mainly in the mean and standard deviation of the elementary and local roundoff errors, and use the familiar propagation laws to study the accumulated effect of these errors. The propagation laws follow: Let $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ be mutually independent random variables. Then

$$\begin{aligned} E\{c_1\varepsilon_1 + \ldots + c_n\varepsilon_n\} &= c_1 E\{\varepsilon_1\} + \ldots + c_n E\{\varepsilon_n\} \\ \sigma\{c_1\varepsilon_1 + \ldots + c_n\varepsilon_n\} &= \sqrt{c_1^2\sigma^2\{\varepsilon_1\} + \ldots + c_n^2\sigma^2\{\varepsilon_n\}} \end{aligned} \tag{2.27}$$

E stands for expectation (mean) and σ for standard deviation. The assumption of mutual independence is essential for the second relation but not for the first one.

Mean $E\{\varepsilon\}$ and standard deviation $\sigma\{\varepsilon\}$ of an elementary roundoff error $\varepsilon = a \circledcirc b - a \diamond b$ depend on a variety of factors, namely (1) the type $\diamond$ of the arithmetic operation, (2) the size of the numbers a, b, (3) base β and length of mantissa τ, and (4) the peculiarities of the machine arithmetic. Trying to specify precise laws for $E\{\varepsilon\}$, $\sigma\{\varepsilon\}$ would result in a discussion of a great variety of cases and in very complicated expressions. Hence, we will try to simplify these laws somewhat and arrive at workable expressions that allow us to estimate the global roundoff errors of the U.S. network with reasonable closeness and safety.

2.8.1 Addition and Subtraction

There are cases when the elementary roundoff errors are zero. If $a = m_a * \beta^{e_a}$, $b = m_b * \beta^{e_b}$, and $a \pm b = m_{a\pm b} * \beta^{e_{a\pm b}}$, and if $e_a = e_b = e_{a\pm b}$, then certainly no roundoff error occurs. If e_a and e_b differ by not more than τ, and if a left postshift of at least $|e_a - e_b|$ places occurs, ε will be zero, provided that true rounding or true chopping is done. However, in most situations where $e_a \neq e_b$, a nonzero roundoff error will occur. If the smaller exponent does not exceed the larger exponent diminished by τ, the preshift will completely wipe out the smaller operand. The roundoff error will be deterministic in this case and will be equal in magnitude to the smaller operand. The result $a \pm b$ will then be biased, even if a truly rounding machine is used. However, the bias will be very small, and we shall neglect it in the sequel, since other effects will largely dominate it. In those situations where a nonzero preshift does not completely wipe out the smaller operand, as well as in the case of a nonzero postshift, the elementary roundoff error will be due to a loss of trailing digits and may be viewed as a random variable. (Recall that leading digits are never viewed as random, only trailing ones.)

Let us now discuss the two types of computer families separately, represented by the CDC 6600 and the IBM 360. We will adopt simplified mathematical models for their rounding features and discuss the merits and shortcomings of these mathematical models.

2.8.1.1 Model for addition and subtraction on the CDC 6600

Recall that earlier we assumed that the rounding instruction set is used, and that the results are always normalized. Let c denote an integer power of the base 2 which bounds the magnitudes of the operands a, b and of the result $a \pm b$:

$$c = 2^\tau > \text{Max}\{|a|, |b|, |a \pm b|\}. \tag{2.28}$$

We assume that the elementary roundoff error

$$\varepsilon = a \oplus b - (a \pm b) \tag{2.29}$$

is a random variable having mean and standard deviation, given by

$$E\{\varepsilon\} = 0 \qquad \sigma\{\varepsilon\} = \frac{c}{\sqrt{12}} 2^{-48}. \tag{2.30}$$

The zero mean is justified by the true rounding

feature. Positive and negative roundoff errors are equally likely. The small biases that occur when the preshift wipes out the smaller operand completely are neglected. The formula for the standard deviation $\sigma\{\varepsilon\}$ results from the assumption that the roundoff errors are distributed uniformly in the interval

$$-\frac{c}{2}2^{-48} \leqslant \varepsilon \leqslant +\frac{c}{2}2^{-48}. \tag{2.31}$$

Recall that the standard deviation of a random variable which is equidistributed in the interval $[\alpha, \beta]$ has the value $[\beta-\alpha]/\sqrt{12}$.

Note that our formulas in most cases overestimate $\sigma\{\varepsilon\}$ and never underestimate it. There are situations in which $\sigma\{\varepsilon\}$ is bound to be zero. Besides this reason for overestimating $\sigma\{\varepsilon\}$, there is the problem of bounding $|a|$, $|b|$, $|a\pm b|$ in the above stated way, by the smallest power c of the base 2. In practical applications we will rarely be able to make c as small as possible, and we will overestimate it by a few powers of two. Hence our estimates will also be too large for this reason. On the other hand, they will not be overly pessimistic either, as should become clear later.

Remark: The CDC 6600 differs from true rounding in one respect that is worth mentioning again. Because the results of addition and subtraction are obtained in unnormalized form and are normalized only by a subsequent separate normalizing instruction, the roundoff error will have a larger standard deviation in all cases where a left postshift occurs. This is clear because the unnormalized result has some digits cut off that would be saved if rounding occurred after the postshift. However, note that our upper bounds on $\sigma\{\varepsilon\}$ are still valid and that the assumption of $E\{\varepsilon\} = 0$ is unaffected.

2.8.1.2 Model for addition and subtraction on the IBM 360

Since the base equals $\beta = 16$, we choose c as an integer power of 16, such that

$$c = 16^{\gamma} > \mathrm{Max}\,\{|a|, |b|, |a\pm b|\}. \tag{2.32}$$

We pretend that the IBM 360 is a truly chopping machine and assume that ε is equidistributed in the intervals

$$-c*16^{-14} \leqslant \varepsilon \leqslant 0 \quad \text{or} \quad 0 \leqslant \varepsilon \leqslant c*16^{-14} \tag{2.33}$$

depending on the sign of $a\pm b$. From these assumptions we get

$$\begin{aligned} E\{\varepsilon\} &= -\frac{c}{2}16^{-14}\,\mathrm{sign}(a\pm b) \\ \sigma\{\varepsilon\} &= \frac{c}{\sqrt{12}}16^{-14}. \end{aligned} \tag{2.34}$$

The magnitude of the mean $E\{\varepsilon\}$ as well as the standard deviation will be overestimated again. Some reasons are described in the previous section where we dealt with the CDC 6600. Another reason is that on the IBM rounding occurs *after* a possible left postshift. In addition, we know from the discussion in section 2.4 that the IBM 360 is not precisely a truly chopping machine. In the subtract magnitude case, and when the preshift exceeds one place thereby, there is a small chance that the result will be rounded upwards rather than downwards. As a consequence, the mean will be slightly decreased (in magnitude) and so will the standard deviation. We will prove this in the following paragraph.

Without losing generality we can assume $a > b > 0$, and that $a - b$ is calculated. Further, the underlying assumption is that $e_a \geqslant e_b + 2$. Rounding will deviate from true chopping whenever the guard digit of the preshifted b is zero and when nonzero digits to the right have been chopped. Otherwise true chopping will occur. If no postshift occurs, elementary roundoff errors will be equidistributed in the interval

$$-c\frac{15}{16}16^{-14} \leqslant \varepsilon \leqslant +c\frac{1}{16}16^{-14} \tag{2.35}$$

rather than in the interval

$$-c\,16^{-14} \leqslant \varepsilon \leqslant 0. \tag{2.36}$$

Hence we will have

$$E\{\varepsilon\} = -\frac{c}{2}\frac{14}{16}16^{-14} \tag{2.37}$$

while $\sigma\{\varepsilon\}$ will be unchanged as

$$\sigma\{\varepsilon\} = \frac{c}{\sqrt{12}}16^{-14}. \tag{2.38}$$

Should a postshift occur, which under the present assumption can amount to only one place to the left, $E\{\varepsilon\}$ and $\sigma\{\varepsilon\}$ would have to be divided by the base 16. Because such a postshift is unlikely, we see that the approximation is fairly close even in the pathological subtract magnitude case. Note also that under no circumstance will the magnitude of $E\{\varepsilon\}$ or of $\sigma\{\varepsilon\}$ be underestimated.

2.8.2 Multiplication and Division

In this subsection, let $\diamond$ stand for $\bullet$ or $/$. Define c as the smallest power of the base β, such that

$$c = \beta^{\gamma} > |a \diamond b|. \tag{2.39}$$

On the CDC 6600 we assume that the result $a \circledcirc b$ is the truly rounded one. Hence

$$E\{\varepsilon\} = 0$$
$$\sigma\{\varepsilon\} = \frac{c}{\sqrt{12}} 2^{-48}. \tag{2.40}$$

On the IBM 360 the result is the truly chopped one. (See Sterbenz (1974: p. 23).) Hence we have

$$E\{\varepsilon\} = -\frac{c}{2} 16^{-14} \operatorname{sign}(a \diamond b)$$
$$\sigma\{\varepsilon\} = \frac{c}{\sqrt{12}} 16^{-14}. \tag{2.41}$$

2.8.3 Local roundoff error of the square root

The square root is taken by a subroutine. If it is calculated with τ digits, its accuracy may be inferior to that of the arithmetic operations. The programmer can improve upon its accuracy either by iterative refinement or by taking the square root in higher precision. The extra computation time needed for improving the square root will be irrelevant in our case because only a few square roots are taken during Cholesky's algorithm as compared to the bulk of additions and multiplications.

We therefore assume that the accuracy of the square root is practically equivalent to that of the four arithmetic operations. Defining the local roundoff error of the square root by

$$\textcircled{2}\sqrt{a} = {}^{2}\sqrt{a} + \varepsilon \tag{2.42}$$

where the left-hand side is the result obtained by the computer, we introduce

$$c = \beta^{\tau} > \sqrt{a} \tag{2.43}$$

and assume for the CDC 6600

$$E\{\varepsilon\} = 0$$
$$\sigma\{\varepsilon\} = \frac{c}{\sqrt{12}} 2^{-48} \tag{2.44}$$

and for the IBM 360

$$E\{\varepsilon\} = -\frac{c}{2} 16^{-14}$$
$$\sigma\{\varepsilon\} = \frac{c}{\sqrt{12}} 16^{-14}. \tag{2.45}$$

3. CHOLESKY'S ALGORITHM APPLIED TO THE NORMAL EQUATIONS OF GEODETIC NETWORKS

3.1 Cholesky's Algorithm for a General Symmetric Positive Definite System

Suppose that the system is written in matrix form as

$$A\,x = b. \tag{3.1}$$

Cholesky's algorithm relies on a decomposition of the positive definite matrix A as

$$A = R^T R \tag{3.2}$$

where R is an upper triangular matrix. During the first or so-called "triangular decomposition phase" of the algorithm, the system is, in effect, multiplied by $(R^T)^{-1}$. The result is the following triangular system:

$$R\,x = s \tag{3.3}$$

with

$$s = (R^T)^{-1}\, b. \tag{3.4}$$

During the second or "back-substitution phase" of Cholesky's algorithm, the triangular system is solved for x recursively, starting with the the last component of x and proceeding to the first.

The details of Cholesky's algorithm can be best described by switching to indices notation. The original system then reads

$$\sum_{j=1}^{n} a_{ij}\, x_j = b_i,\ i = 1, \ldots, n. \tag{3.5}$$

The triangularized system is

$$\sum_{j=i}^{n} r_{ij}\, x_j = s_i,\ i = 1, \ldots, n \tag{3.6}$$

which is calculated from the original system by

$$\left.\begin{aligned} r_{ii} &= \sqrt{a_{ii} - \sum_{k=1}^{i-1} r_{ki}^2} \\ r_{ij} &= (a_{ij} - \sum_{k=1}^{i-1} r_{ki} r_{kj}) / r_{ii},\ j = i+1, \ldots, n \\ s_i &= (b_i - \sum_{k=1}^{i-1} r_{ki} s_k) / r_{ii} \end{aligned}\right\} i = 1, \ldots, n. \tag{3.7}$$

During the back substitution phase, the triangular system is solved by

$$x_i = (s_i - \sum_{j=i+1}^{n} r_{ij}\, x_j) / r_{ii},\ i = n, \ldots, 1. \tag{3.8}$$

3.2 Partial Reduction by Cholesky's Algorithm

Split the original system as

$$\begin{aligned} A_{11}x_1 + A_{12}x_2 &= b_1 \\ A_{21}x_1 + A_{22}x_2 &= b_2. \end{aligned} \tag{3.9}$$

Split R accordingly:

$$R = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix} \tag{3.10}$$

From the equation $R^TR = A$ we deduce the following identities

$$\begin{aligned} R_{11}^TR_{11} &= A_{11} \\ R_{11}^TR_{12} &= A_{12} \\ R_{12}^TR_{12} + R_{22}^TR_{22} &= A_{22}. \end{aligned} \tag{3.11}$$

Multiply the first set of the original normals by $(R_{11}^T)^{-1}$ and then eliminate the unknowns x_1 from the second set by subtracting proper multiples of the equations of the first set, the multiplying matrix factor being $A_{21}R_{11}^{-1} = R_{12}^T$. The resulting system is

$$R_{11}x_1 + R_{12}x_2 = s_1 \tag{3.12a}$$

$$A_{22}^{(p)}x_2 = b_2^{(p)}. \tag{3.12b}$$

The second set (3.12b) of these equations is called the *partially reduced* set of normal equations. Explicit and equivalent expressions for the quantities involved are

$$\begin{aligned} A_{22}^{(p)} &= A_{22} - R_{12}^TR_{12} = R_{22}^TR_{22} = \\ &= A_{22} - A_{21}A_{11}^{-1}A_{12} \\ b_2^{(p)} &= b_2 - R_{12}^Ts_1 = b_2 - A_{21}A_{11}^{-1}b_1. \end{aligned} \tag{3.13}$$

These expressions are easily checked by the identities exhibited above. The last expression in any of the two lines reveals that the reduced normal equations do not depend in any way on the peculiarities of Cholesky's algorithm. In fact, any method of elimination that removes the unknowns x_1 from the second set by subtracting proper multiples of the first set must uniquely arrive at the partially reduced normals exhibited above.

In indices notation, the partial Cholesky reduction is

$$\left.\begin{aligned} r_{ii} &= \sqrt{a_{ii} - \sum_{k=1}^{i-1} r_{ki}^2} \\ r_{ij} &= (a_{ij} - \sum_{k=1}^{i-1} r_{ki}r_{kj})/r_{ii},\ j = i+1, \ldots, n \\ s_i &= (b_i - \sum_{k=1}^{i-1} r_{ki}s_k)/r_{ii} \end{aligned}\right\} i = 1, \ldots, p \tag{3.14}$$

$$\left.\begin{aligned} a_{ij}^{(p)} &= a_{ij} - \sum_{k=1}^{p} r_{ki}\, r_{kj},\ j = i+1, \ldots, n \\ b_i^{(p)} &= b_i - \sum_{k=1}^{p} r_{ki}s_k \end{aligned}\right\} i = p+1, \ldots, n. \tag{3.14}$$

Cholesky's algorithm can be organized in many different ways. The programs used by the NGS, written by R. H. Hanson and based on earlier work of Poder and Tscherning (1973) at the Danish Geodetic Institute, execute Cholesky's algorithm in the following manner;

```
FOR j = 1 TO n+1
  FOR i = 1 TO MIN(n,j)
    SUM = 0
    FOR k = 1 TO MIN(p, i-1)
      SUM = SUM + A(k,i) * A(k,j)
    NEXT k
    A(i,j) = A(i,j) - SUM
    IF (i≤MIN(j-1,p)) A(i,j) = A(i,j)/A(i,i)
  NEXT i
  IF (j≤p) A(j,j) = √A(j,j)
NEXT j.
```

(3.15)

In this algorithm, the $A(i,j)$ are place holders. They denote storage locations for a number of quantities. In detail,

- the original coefficients a_{ij} are stored at $A(i,j)$; the original coefficients b_i are stored at $A(i,n+1)$;
- the r_{ij} are stored at $A(i,j)$, the s_i at $A(i,n+1)$;
- the $a_{ij}^{(p)}$ are stored at $A(i,j)$, the $b_i^{(p)}$ at $A(i,n+1)$.

It should be stressed that the above algorithm is still a simplification of the actual NGS programs. First, these programs make use of a more complicated data structure which allows storage and retrieval of coefficients $A(i,j)$ columnwise to and from mass storage (disks). Second, the programs allow for exploiting the sparsity of the normal equations to some extent. The normals have many zero coefficients. If the equations are ordered in a sensitive way, many of the zeroes are retained throughout the reduction. This results in a great saving of computer storage and computation time. The NGS programs store only a section of each column, excluding coefficients that will never become nonzero during the execution of the algorithm. We shall come back to the problem of ordering in section 3.5.

Remark: We briefly mention another way to execute Cholesky's algorithm, which amounts to a series

of partial reductions for p proceeding from 1 to n. In this fashion Cholesky's algorithm becomes very similar to Gauss' algorithm. Denote $a_{ij}^{(o)} = a_{ij}$ and $b_i^{(o)} = b_i$. We then have

$$\left.\begin{aligned} & r_{pp} = \sqrt{a_{pp}^{(p-1)}} \\ & r_{pj} = a_{pj}^{(p-1)}/r_{pp},\ j = p+1, \ldots, n \\ & s_p = b_p^{(p-1)}/r_{pp} \\ & a_{ij}^{(p)} = a_{ij}^{(p-1)} - r_{pi}\, r_{pj} \left\{ \begin{array}{l} i = p+1, \ldots, n \\ j = i, \ldots n \end{array} \right. \\ & b_i^{(p)} = b_i^{(p-1)} - r_{pi} s_i \end{aligned}\right\} \begin{array}{l} p = 1, \\ \ldots, n \end{array} \quad (3.16)$$

If the algorithm stops at any $p<n$, a partially reduced system results. It adds insight into the problem of equation ordering, discussed later in this section, that any equation is modified by either dividing it by the square root of the diagonal element or by subtracting proper multiples of preceding equations.

Remark. Common to all versions of Cholesky's algorithm is that they operate only on the portion above and including the main diagonal of the matrix A, as well as on the right-hand side. Hence only the upper triangular portion of the matrix A needs to be stored in computer memory. Substantially more storage is saved if the sparse structure of A is exploited, which is typical for matrices associated with network problems.

3.3 Geodetic Normal Equations

Our system of normal equations results from a geodetic ground control network. Adjustment is done on a spheroidal rotational ellipsoid. We assume that the reader is familiar with the principles of network adjustment. Our outline will mainly serve to point out peculiarities and to specify the terminology and notation used in the sequel.

The network will be adjusted by variation of parameters. The parameters, or unknowns, are the ellipsoidal coordinates of the stations (points, nodes). Any station has two parameters, namely ellipsoidal latitude and longitude. The so-called orientation unknowns of direction bundles will be eliminated before the normal equations are assembled and will not appear in the final set of equations which serves as the input in our study of roundoff errors.

Approximate coordinates must be known a priori. Denote these coordinates by the vector $p^{(o)}$. The observations ℓ, comprising distances, azimuths, bundles of directions, and Doppler positions, will not fit the approximate coordinates. There will be discrepancies $\Delta\ell$, i.e. only the set of observations $\ell - \Delta\ell$ will fit the approximate coordinates. An adjustment applies corrections v to the observations, so that they become the corrected observations $\ell + v$. It also applies shifts Δp to the approximate coordinates so that they become the adjusted coordinates $p = p^{(o)} + \Delta p$. The functional relation between the corrected observations and the adjusted coordinates is (after elimination of the orientation unknowns) in linearized form written as:

$$\Delta\ell + v = B\,\Delta p. \quad (3.17)$$

Weights are prescribed for the individual observations. They are arranged along the diagonal of the weight matrix P which has zero off-diagonal coefficients. Gauss' minimum principle, i.e.,

$$v^T P v = \text{Minimum} \quad (3.18)$$

is used to uniquely determine v and Δp satisfying the side constraints $\Delta\ell + v = B\,\Delta p$. The extremum problem leads to the normal equations

$$B^T P B\,\Delta p = B^T P\,\Delta\ell \quad (3.19)$$

which for brevity are written as

$$A\,x = b. \quad (3.20)$$

Note that the unknowns x are actually small shifts leading from the approximate coordinates to the adjusted coordinates.

An important feature of geodetic network adjustment is the local nature of the observations. Any observation involves only a small number of stations which are located close together. For distance and direction observations, direct visibility between two stations must be given. This limits the spacings between stations connected by such a line of vision to 30 km or less in most cases. The normal equation matrix will have only nonzero off-diagonal elements $a_{ij} \neq 0$, if i,j refer either to the two coordinates of one station or to coordinates of two stations connected by a measurement. Such a connection is established either by a direction, a distance, or an azimuth between the two stations, or is due to the preelimination of the orientation unknowns in case of a directional coobservation of the two stations from a third station. The Doppler position observations refer to the two coordinates of one station and will not cause any a_{ij}, $i \neq j$, to be nonzero. While the network covers a large portion of a continent and extends over several thousands of kilometers, there will only be nonzero coefficients a_{ij} if the involved stations are not farther apart than 60 km (in most cases).

Remark. In the literature on numerical linear algebra it is frequently argued that formation and

solution of a normal equation system is not a good procedure for doing a least squares adjustment. Instead one should go along with the observation equation system (3.17), subjecting it to orthogonalization, singular value decomposition, or other procedures. The argument is based on the condition number of a matrix. The condition number of the normal equation matrix is inferior to that of the observation equations. This is certainly true. On the other hand, it has been proven that storage requirement and computational labor is much less for a geodetic network if it is adjusted by the direct solution of a normal equation system as compared to any other procedure. Refer to the discussion in Avila et al. (1978, p.16). Singular value decomposition or orthogonalization appears to be very efficient for moderately large linear systems that are very ill-conditioned. In the case of very large sparse geodetic network systems which are not extremely ill-conditioned, storage requirement and computational labor are the decisive criteria for selecting a solution method. The observation equation matrix for the U.S. network is of size 3,000,000 × 350,000. To my knowledge no technique is known that preserves sparsity during orthogonalization or singular value decomposition as efficiently as that method which applies direct elimination to the normal equation system, as will be shown later in this chapter.

3.4 Geodetic Interpretation of the Partial Cholesky-Reduced System

The geodetic meaning of the quantities appearing in the system that has undergone a partial reduction by Cholesky's method is perhaps best understood in terms of a parameter transformation. The original normals are written as

$$\begin{aligned} A_{11}x_1 + A_{12}x_2 &= b_1 \\ A_{21}x_1 + A_{22}x_2 &= b_2 \end{aligned} \tag{3.21}$$

and consider a parameter transformation which changes x_1 into y_1 leaving x_2 unchanged:

$$\begin{aligned} y_1 &= R_{11}x_1 + R_{12}x_2 \\ x_2 &= \qquad\qquad x_2. \end{aligned} \tag{3.22}$$

The inverse transformation is

$$\begin{aligned} x_1 &= R_{11}^{-1}y_1 - R_{11}^{-1}R_{12}x_2 \\ x_2 &= \qquad\qquad x_2. \end{aligned} \tag{3.23}$$

The normal equations for the new parameters are

$$\begin{aligned} y_1 \qquad &= s_1 \\ A_{22}^{(p)}x_2 &= b_2^{(p)}. \end{aligned} \tag{3.24}$$

If we substitute for y_1, we get

$$\begin{aligned} R_{11}x_1 + R_{12}x_2 &= s_1 \\ A_{22}^{(p)}x_2 &= b_2^{(p)}. \end{aligned} \tag{3.25}$$

This is precisely what we get after partial Cholesky reduction. We see that hidden behind these equations is the system of normal equations involving y_1, x_2. *This system completely decomposes into two separate systems for y_1 and x_2.* It follows that the adjusted values for y_1, x_2 will be uncorrelated. The covariance matrix for x_2 will be

$$\Sigma(x_2) = (A_{22}^{(p)})^{-1}. \tag{3.26}$$

Let us go back to the original normal equations:

$$A\,x = b. \tag{3.27}$$

If a certain subset of the components of x are forced to fixed values, which amounts to fixing the corresponding coordinates at the values $p^{(o)} + x$, then the normal equations for the remaining unknowns are obtained as follows: Noting that any equation belongs to a certain coordinate, disregard all equations belonging to the fixed components. In the remaining equations, insert the prescribed values for the x's to be fixed, and move these terms toward the right. The desired system results. Note that the same procedure may be applied to the partially reduced Cholesky system

$$\begin{aligned} R_{11}x_1 + R_{12}x_2 &= s_1 \\ A_{22}^{(p)}x_2 &= b_2^{(p)}, \end{aligned} \tag{3.28}$$

provided that the fixing is restricted to coordinates out of set x_2. This observation allows us to give the coefficients r_{ij}, s_i, $a_{ij}^{(p)}$, $b_i^{(p)}$ the following geodetic interpretation.

(*) $a_{ii}^{(p)}$, $i > p$, is the reciprocal of the variance of coordinate i, provided that the coordinates k, $p < k \leqslant n$, $k \neq i$ are fixed, while the coordinates k, $1 \leqslant k \leqslant p$, as well as coordinate i itself, are allowed to vary freely.

(*) $-a_{ij}^{(p)}/a_{ii}^{(p)}$, $i,j > p$, $i \neq j$ is the shift, with respect to the *adjusted* position, suffered by coordinate i if coordinate j is displaced by one unit from the adjusted position, and if coordinates k, $p < k \leqslant n$, $k \neq i,j$ are fixed to their adjusted position, while coordinates k, $1 \leqslant k \leqslant p$ as well as coordinate i itself, are allowed to vary freely.

(*) $b_i^{(p)}/a_{ii}^{(p)}$, $i > p$ is the shift, with respect to the *approximate* position, suffered by coordinate i if coordinates k, $p < k \leqslant n$, $k \neq i$ are fixed to their approximate positions, while coordinates k, $1 \leqslant k \leqslant p$, as well as coordinate i itself, are allowed to vary freely.

(*) r_{ii}, $i \leqslant p$, is the standard deviation of coordinate i, if coordinates k, $i<k \leqslant n$, are fixed, while coordinates k, $1 \leqslant k \leqslant i$ are allowed to vary freely.

(*) $-r_{ij}/r_{ii}$, $i \leqslant p$, $j>i$ is the shift, with respect to the *adjusted* position, suffered by coordinate i, provided that coordinate j is displaced by one unit from its adjusted position, that coordinates k, $i<k \leqslant n$, $k \neq j$ are fixed to their adjusted positions while coordinates k, $1 \leqslant k \leqslant i$ are allowed to vary freely.

(*) s_i/r_{ii}, $i \leqslant p$ is the shift, with respect to the *approximate* position, suffered by coordinate i, provided that coordinates k, $i<k \leqslant n$ are fixed to their approximate positions, while coordinates k, $1 \leqslant k \leqslant i$ can vary freely.

The last three statements require an additional argument because coordinates k, $k \leqslant p$ are also held fixed, while earlier we said that fixing is restricted to the second set of unknowns, *i.e.*, those with $k>p$.

The three last statements should be clear if we set $i=p$, because then only coordinates $k>p$ are fixed. On the other hand, the r_{ij}'s are no longer subject to any change, as p moves on from i to higher values. Hence the argument also applies for $i<p$.

Remark: (Elastostatic interpretation of normal equations before and after partial reduction.) To the structural engineer the normal equations $Ax=b$ appear as equilibrium equations of an elastic system. The matrix A is called the stiffness matrix, x are coordinate shifts of the nodes, and b are external forces acting at the nodes. The coefficients of the stiffness matrix have the following physical meaning: Suppose that the system is in equilibrium with $x=0$, $b=0$. Displace coordinate j by one unit from its equilibrium position, keeping all other coordinates fixed to their equilibrium position. An elastic force will then be acting on coordinate i. This force is precisely a_{ij}. This holds also for $i=j$. The partially reduced normals $A_{22}^{(p)} x_2 = b_2^{(p)}$ refer to a so-called statically reduced system. $A_{22}^{(p)}$ is still a stiffness matrix. $a_{ij}^{(p)}$, $p<i$, $j \leqslant n$ is the force acting on coordinate i when coordinate j is displaced by one unit from its equilibrium position, when coordinates k, $p<k \leqslant n$ are fixed, while coordinates k, $1 \leqslant k \leqslant p$ are allowed to adjust freely. The right-hand coefficients $b_i^{(p)}$ have the meaning of forces. The original $b_i = b_i^{(0)}$ are nodal forces due to inconsistencies in the network. As nodes are freed during elimination, different forces $b_i^{(p)}$ must be applied to the remaining nodes such that the equilibrium position of the remaining nodes remains the same. The forces of the eliminated nodes must be transported to the uneliminated ones. Occasionally it is also advantageous to consider external forces. If the vector b is chosen as the *j-th* column of the unit matrix, the solution x of the system becomes the *j-th* column $f_{.j}$ of the inverse F of the stiffness matrix A. Hence f_{ij} is the shift of coordinate i if a unit force is applied to coordinate j. Thereby it is assumed that prior to application of the unit force a free equilibrium state had been reached. In particular, f_{ii} is the shift of coordinate i with respect to its adjusted position, if (after adjustment) a unit force is applied to coordinate i. A more lucid interpretation of the variance f_{ii} of the adjusted coordinate i can hardly be given. The elastostatic interpretation is thus somewhat simpler and of great physical significance. I personally prefer to think in terms of elastostatics, where the $a_{ij}^{(p)}$, $b_i^{(p)}$ themselves have a most simple interpretation, whereas in geodetic reasoning the ratios $a_{ij}^{(p)}/a_{ii}^{(p)}$, $b_i^{(p)}/a_{ii}^{(p)}$ are most easily understood. However, since this publication is addressed to the geodesist, elastostatic language will very rarely be used in the sequel. For further details the reader is referred to Rubinstein and Rosen (1970).

Remark: (On the near vanishing of row sums.) Another property of geodetic normal equations which will be of some importance to the roundoff study is concerned with the row sums

$$\sum_{j=1}^{n} a_{ij}^{(p)} \tag{3.29}$$

of the original as well as the partially reduced normals. If i is a coordinate whose station—call it P—is involved only in relative measurements, *i.e.* in measurements other than absolute positioning by Doppler, then the above row sum nearly vanishes for any p. The row sum vanishes precisely if the network is plane. On the ellipsoid it vanishes only approximately. The proof, for the plane network, goes back to the observational equations $Bx=\Delta l+v$. All observational equations involving station P can be thought of as being formulated in terms of differences of coordinate increments. This implies that the row sums pertaining to station P vanish. The property of station P's vanishing row sums carries over from the observational equations matrix B to the original normal equation matrix $A=B^TPB$. Note that station P's normal equations can be formed by considering only the observations that involve this station. If station P is involved in a Doppler measurement, the row sum of equation i will not vanish, even if the network is plane. However, since the Doppler observations have weights much smaller than those of the relative measurements (directions, distances, azimuths), the row sum will be appreciably smaller than the larger coefficients in the *i-th* row of A. Hence, we conclude that all row sums of the normals are small. The remark at the end of section 3.2 tells us that Cholesky's algorithm is a succession of subtractions of multiples of

rows from others. Hence the property of near vanishing of row sums is retained throughout reduction and carries over to the partially reduced normal equation matrix $A_{22}^{(p)}$.

3.5 Problem of Station Ordering

Coordinate i is associated with row and column i of the normal equations. Ordering the coordinates in a different way leads to a system of normal equations with rows and columns simultaneously permuted, *i.e.*, with diagonal elements permuted and rows and columns arranged accordingly. Mathematically, the two systems are equivalent, numerically they are not. Widely recognized in recent literature are the great differences in storage requirement and computation time that result from different orderings and when algorithms are used that take into account the sparseness of A.

In geodetic networks, nonzero off-diagonal elements result from observations between stations rather than between coordinates. The problem of ordering the unknowns becomes a problem of ordering the stations. The two coordinates of one station will always be placed together.

We will refrain from giving a thorough discussion of ordering schemes currently in fashion. We shall briefly review three ordering strategies. The first serves as an introduction to the problem, the other two will be relevant to the readjustment of the U.S. network.

3.5.1 Ordering for small bandwidth

A supposed geodetic network is depicted in figure 3.1. The solid lines indicate directions observed at both end points. Additional distances and azimuths (measured along some of the solid lines) as well as some Doppler positional observations may be available. Recall that two stations are connected by nonzero off-diagonal coefficients in the normal equations if there is a direction-, distance-, or azimuth-observation between these two points, or if the two points are directionally coobserved from a third station. In this way, station 1 is connected to stations 2,3,5,6,8,9. Station 8 is connected to 1,2,5,6,9,10,12,13,14,17,18,19. For any station i we can specify the highest numbered station s_i connected to station i. Thus $s_1=9$, $s_8=19$. We may calculate the number

$$w=2*\mathrm{MAX}(s_i-i+1) \qquad (3.30)$$

which is called the bandwidth of the system. The factor 2 has been introduced to account for the fact that we have two coordinates per station. In our above example we would have $w=2(s_8-8+1)=24$.

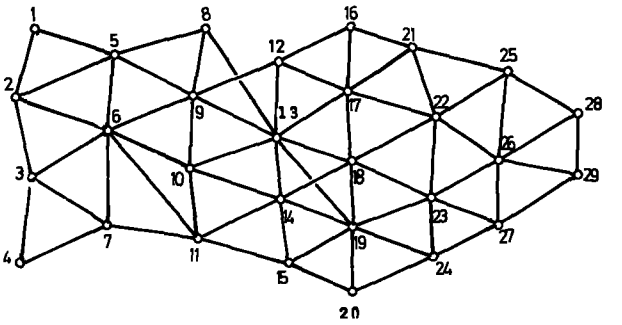

Figure 3.1.—Sample network.

It turns out that the normal equation matrix A will have nonzero coefficients restricted to a band of width w as indicated in figure 3.2. Note that w counts only lines of coefficients above and including the main diagonal. The coefficients below the main diagonal are never used.

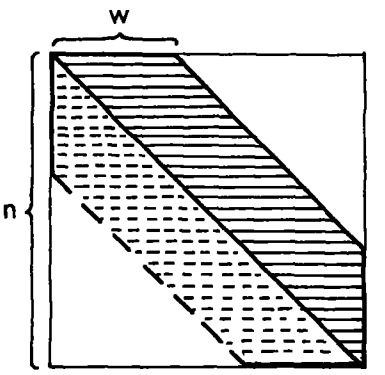

Figure 3.2.—Banded normal equations.

In general, the band will not be completely filled with nonzero coefficients of A. It will also contain some zeroes. The important thing to note, however, is that nonzero coefficients r_{ij}, $a_{ij}^{(p)}$, arising during (partial) Cholesky reduction, are also confined within the band. Some of these will appear at places where A also had nonzero coefficients, and others will take the place of original zeroes. The latter ones are called "fill-in" coefficients.

The proof that fill-in is confined to the band is most easily derived from the next to the last remark in section 3.2. There we saw that any row of any of the Cholesky reduction states results by subtracting multiples of preceding rows from it (and by dividing the row by a factor, if $i \leqslant p$). However, preceding rows k, $k<i$ can never have nonzero coefficients to the right of the rightmost eligible position for a nonzero coefficient of a row i.

A consequence of the banded structure of A is that any one of the inner products, *i.e.*, the sums of prod-

ucts appearing in Cholesky's algorithm, will have, at most, $w-1$ nonzero terms. In fact, the first version (3.7) of the full Cholesky algorithm specified in section 3.1 can be respecified as follows:

$$\left.\begin{aligned} r_{ii} &= \sqrt{a_{ii} - \sum_{k=MAX(1,i-w+1)}^{i-1} r_{ki}^2} \\ r_{ij} &= (a_{ij} - \sum_{k=MAX(1,j-w+1)}^{i-1} r_{ki}\, r_{kj})/r_{ii}, \\ j &= i+1, \ldots, MIN(n, i+w-1) \\ s_i &= (b_i - \sum_{k=MAX(1,i-w+1)}^{i-1} r_{ki} s_k)/r_{ii} \end{aligned}\right\} \quad i = 1, \ldots, n \quad (3.31)$$

and

$$x_i = (s_i - \sum_{j=i+1}^{MIN(n,i+w-1)} r_{ij} x_j)/r_{ii} \quad i = n, \ldots, 1.$$

On the one hand, a computer program for this algorithm would be more complicated; on the other hand, for $w \ll n$, it would be much faster. It would save much storage if the coefficients within the band were stored in a compacted way, for example, as the columns of an array of size $w * n$.

A different numbering of the stations would generally result in a different bandwidth w. One could try to minimize w over all possible permutations; however, this is not economical. There are computer algorithms that find near optimal orderings in a short time. Frequently, a good ordering is found by inspection. If a network is elongated, as in the example above, then numbering along the lines that cross the network at the shorter distances often leads to a good ordering. I believe the ordering specified in the figure 3.2 is near optimal.

3.5.2 Ordering for small profile

As pointed out in section 3.5.1, the saving of computer time and storage comes from the reduced number of product accumulations in Cholesky's algorithm. It is plausible, therefore, that methods have been designed which aim at minimizing the number of product accumulations in the first place rather than doing this indirectly by minimizing the bandwidth. One such method, which is still a compromise between simplicity and efficiency, is ordering for small profile.

The profile of a symmetric matrix includes all elements of a column that are located between the topmost nonzero element and the main diagonal, inclusively. Hence, an element a_{ij}, $i \leqslant j$ is within the profile, if there is an element $a_{kj} \neq 0$ for a certain $k \leqslant i$. A typical profile is shown in figure 3.3.

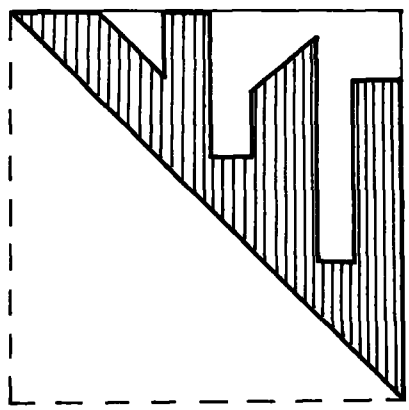

Figure 3.3.—Profiled normal equations.

The profile may include zero coefficients. Again it is important to note that fill-in is restricted to the profile. The proof relies on a similar argument as previously given for the banded structure. Subtracting a multiple of a row from a subsequent row will never cause any nonzero entry outside the profile. The Cholesky factorization $A = R^T R$ will result in a matrix R which has nonzero coefficients only within the profile. R, being upper triangular, will have zeroes below the main diagonal, whereas A will have coefficients implied by the symmetry there.

NGS computer programs which are currently being used to adjust moderately small networks (up to about 2,500 stations) rely on ordering for a small profile. The ordering algorithm, designed and described by Snay (1976), is heuristic and does not yield a minimal profile in the strict sense. It will, however, establish a fairly small profile in a short time. As will be clear later on, the algorithm will also contribute to the adjustment of the entire U.S. network.

3.5.3 Identifying nonzero coefficients for a certain reduction state

Before we proceed to still another ordering technique, we pause briefly and reflect on the problem of identifying the nonzero coefficients of A associated with a certain reduction state. Assume, for example, that the partial Cholesky reduction has "eliminated" stations 1 to 12, also marked by black circles in figure 3.4. White circles indicate stations 13 to 29 that participate in the partially reduced system $A_{22}^{(p)} x_2 = b_2^{(p)}$. The network is the same as that one in section 3.5, except that the station numbering now conforms with a changed sequence of elimination steps. From section 3.4, dealing with the geodetic interpretation of a Cholesky-reduced system, we infer that the pattern of zero and nonzero coefficients after partial Cholesky reduction up to station $p = 12$, inclusively is shown in figure 3.5.

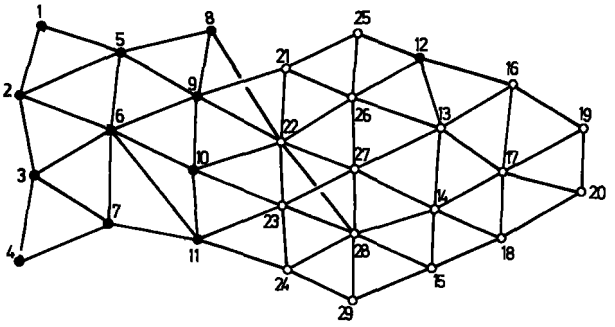

Figure 3.4.—Sample network with stations 1 through 12 eliminated from normal equations.

The numbering of rows and columns of the matrix A in figure 3.5 refers to nodes rather than to coordinates. Hence the individual entries represent actually 2×2 matrices. Heavily shaded entries represent nonzero elements of the original normals. Lightly shaded areas indicate the fill-in which occurs during partial Cholesky reduction up to and including station $p=12$. Let us give the appropriate argument for a few entries.

(*) Entry (14,21). The shading indicates fill-in. Why are nodes 14 and 21 connected at this time? According to section 3.4 (cf., the explanation of the expression $-a_{ij}^{(p)}/a_{ii}^{(p)}$ there), we assume that nodes 1 to $12(=p)$ are free, as well as node 14. We assume the other nodes fixed to their adjusted postion, except for node 21, which is displaced from its adjusted position. The displacement of node 21 will cause the direction bundles at the neighboring nodes 25, 26 to rotate. As a consequence, the free node 12 will move away from its adjusted position, causing in turn the bundle in 13 to be displaced rotationally. This bundle finally will displace station 14. Hence $a_{14,21}^{(p)}$ will be nonzero, as was to be shown. (The possibility that the resulting movement of 14 is the zero movement is neglected here, as it is in all treatises of sparse matrices.)

(*) Entry (3,8). The shading indicates fill-in again. This time we refer to the rule for $-r_{ij}/r_{ii}$ given in section 3.4. We pretend that only nodes 1,2,3 have been eliminated, i.e., we temporarily assume $p=3$. We further assume nodes 4 to 29 fixed to their adjusted positions, except for node 8 which is displaced. This causes the bundle in 5 to deviate from its adjusted position, which in turn displaces nodes 1,2. The displacement of 1 and 2 will finally displace node 3. Hence $r_{3,8}$ must be nonzero, in general.

(*) Entry (10,13). We may put $p=10$. Displacing node 13 causes movements of the bundles connected to node 13. No movement takes place to the left of the barrier formed by the double line of nodes 21 to 29. Hence the coefficient must be zero. In fact, all coefficients (i,j), $i \leqslant 11$, $12 \leqslant j \leqslant 20$, must be zero. We see that a barrier of a double line of nodes crossing the network can effectively keep down the fill-in. This observation leads us to the ordering scheme considered in the next subsection.

3.5.4 Nested dissection

We have just seen that by appropriately ordering the stations we may establish barriers which divide the network into parts such that the interior stations of one part will never become connected to interior stations of another part. The numerical analyst George (1973) fully exploited this idea. He calls his ordering scheme "nested dissection." As we shall see later, this is anticipated to some extent by what is known among geodesists as "Helmert blocking."

Figure 3.6 exemplifies the idea of nested dissection. The individual stations are not shown here. Instead, we see subsets of stations carrying labels 1 to 4. We imagine that these labels are attached to all nodes of a particular subset. Nodes carrying label 1 are eliminated first. The sequence in which this is done is not of much importance as long as the number of stations in one connected subset is small. Should this number be larger, we may imagine that an ordering for small profile is done in each individual subset. At the next step we eliminate nodes labeled 2, then 3, and finally 4.

Let us now take a look at the connections a certain node labeled i may encounter to nodes that come later in the ordering sequence. Such nodes carry either the label i or a label $j > i$. Connections to label i nodes are possible only if the other node is in the same connected label i subset. This is true, because all other label i subsets are separated by barrier subsets of higher labels. Connections of a label i node to nodes of higher labels are only possible if the higher label nodes are located at a barrier surrounding the subset of node i.

Any node will be connected to only a few nodes that come later in the ordering sequence. This is particularly true at the lower levels. It follows that matrix A will be quite sparse, although the pattern of zeroes is now rather complicated.

In order to see the power of nested dissection, we imagine a fairly homogeneous network of n stations covering a region which is shaped somewhat like a square. George (1973) shows that the number of nonzero coefficients (original A plus fill-in) is bounded by

$$\text{const}_1\ n \log n. \tag{3.32}$$

If, in contrast to this, we subject the network to ordering for small bandwidth, we can bound the nonzeroes only by

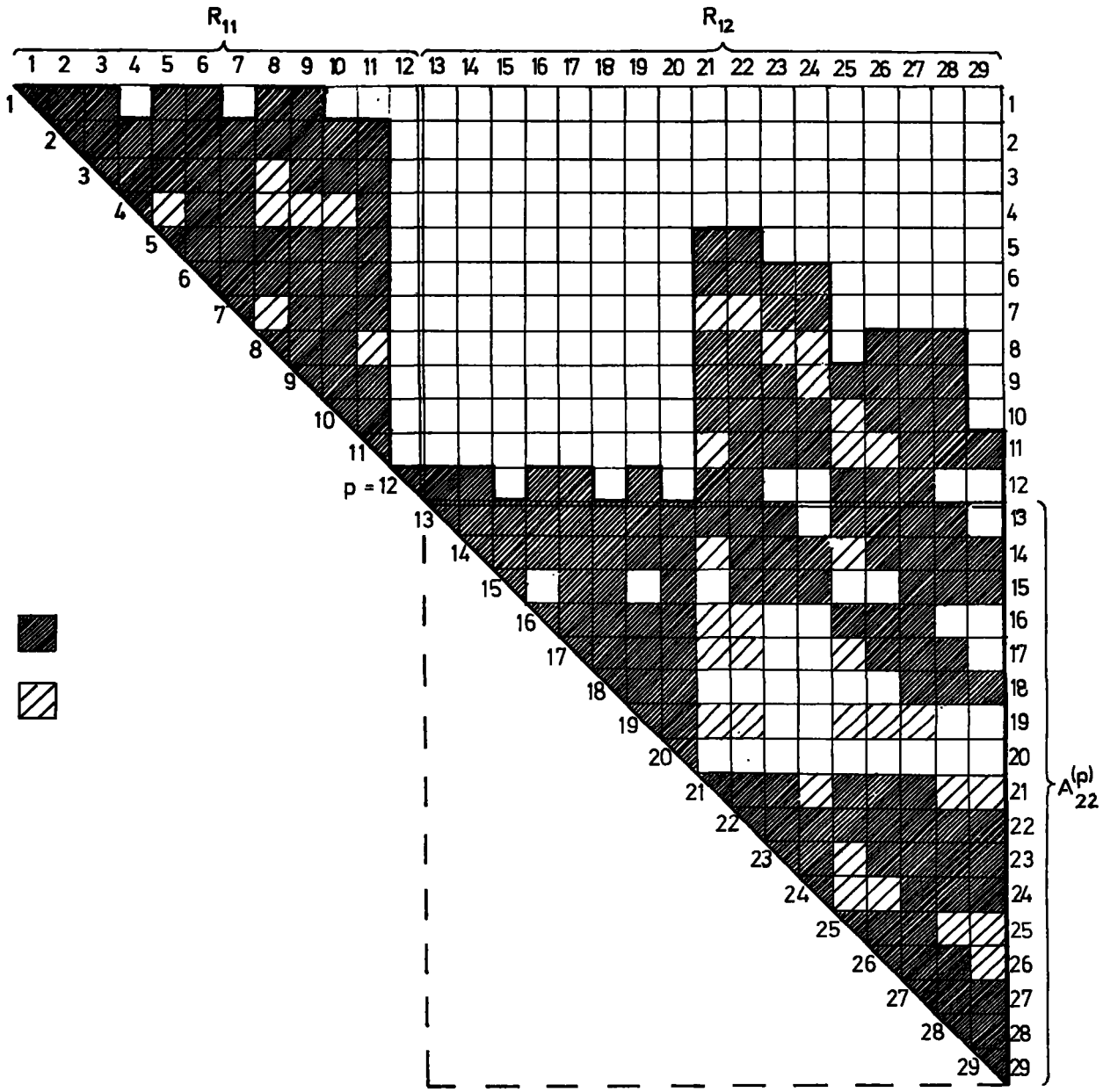

Figure 3.5.—Structure of normal equations when stations 1 through 12 are eliminated.

$$\text{const}_2\ n\sqrt{n}. \tag{3.33}$$

Also ordering for small profile could not achieve anything much better. Assuming an efficient storage scheme, the storage requirement grows roughly proportional to the number of nonzeroes. However, the factor of proportionality is different from method to method. Nested dissection, in particular, has a more complicated pattern of zeroes that necessitates the storage of additional pointers to keep track of the nonzero elements.

Despite the different proportionality factors and also the difference between const_1 and const_2 in the above formulas, it becomes clear that asymptotically, *i.e.*, as n grows on and on, nested dissection is superior. In fact, as $n \to \infty$, the ratio of storage requirement for nested dissection and bandwidth tends to zero as const $\log n / \sqrt{n}$. In this context it is interesting to note that no ordering scheme can improve upon nested dissection asymptotically by more than a constant factor.

We have argued that the number of nonzeroes is directly related to storage requirement. It is also indirectly related to the amount of computational labor. Let us take a look at the number of product accumulations necessary for the triangular decomposition of A. As it turns out, these product accumulations account for most of the computation time needed to solve the normal equations by Cholesky's method. George (1973) shows that this number is bounded by

$$\text{const}_3\, n\sqrt{n} \tag{3.34}$$

if nested dissection is done. Bandwidth ordering, on the other hand, requires

$$\text{const}_4\, n^2 \tag{3.35}$$

for a homogeneous network of the type mentioned. Again the asymptotic superiority of nested dissection becomes evident.

We conclude this subsection with a few remarks.

Remark: Asymptotic superiority of a method does not necessarily mean superiority for moderately small networks. As already indicated, the exploitation of a complicated pattern of zeroes can cause an overhead of storage and computation time. In addition to nonzero coefficients, overhead storage is needed for addressing information which must be stored and for storing a more complicated program.

Remark: Faced with a given network, the subdivision of nodes into categories of different labels is not always immediate. The network will not always be rectangularly shaped, and it will not always be possible to identify a number of first level sets equal to a power of 4. In practice, it will be necessary to compromise. Occasionally, the connected subsets of stations of the same label will deviate in number and shape from the ideal case shown in figure 3.6.

Remark: To avoid pitfalls, one must be sure that the barriers dividing the network, as indicated in figure 3.6, are virtually impenetrable. For the types of networks considered, *i.e.*, those involving bundles of directions, distances, azimuths, and absolute positions, the following rule applies. From and to a node of label i there may be lines of vision only to and from; (1) nodes of an adjacent lower label set, (2) nodes of label i which are in the same label i subset, (3) nodes of higher labeled adjacent sets. Otherwise one will try to keep the barriers as thin as possible. Roughly one will arrive at barrier sets composed of double rows of points, as already encountered in the example of figure 3.4. However, there will be exceptions, particularly in the presence of very long lines of vision.

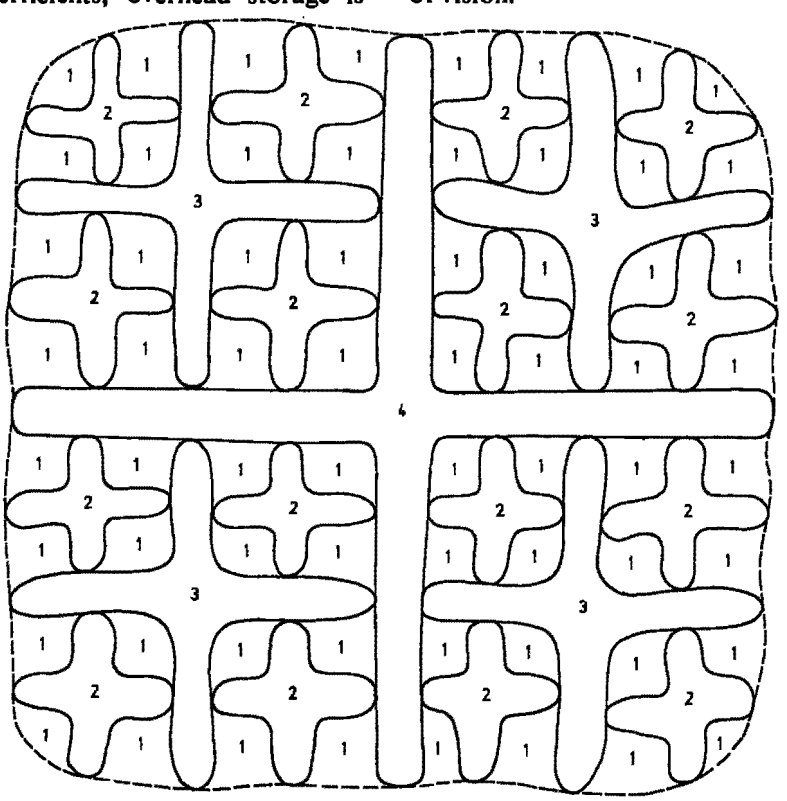

Figure 3.6.—Nested dissection.

3.5.5 Helmert blocking

Let us briefly review the basic idea of Helmert blocking for the small network shown in figure 3.4. We reproduce the network in figure 3.7. The dashed line separates two blocks. The nodes marked by simple circles are interior to the relevant block. The nodes marked by double circles are junction nodes, forming a barrier between the two blocks. The normal equations are assembled separately for each block:

$$\begin{aligned} &\text{Block 1:} \begin{bmatrix} A_{11} & B_{13} \\ B_{31} & B_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_3 \end{bmatrix} = \begin{bmatrix} a_1 \\ b_3 \end{bmatrix} \\ &\text{Block 2:} \begin{bmatrix} A_{22} & C_{23} \\ C_{32} & C_{33} \end{bmatrix} \begin{bmatrix} x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} a_2 \\ c_3 \end{bmatrix} \end{aligned} \tag{3.36}$$

Here x_1, x_2 denote the coordinates of the stations interior to blocks 1, 2, and x_3 denotes the junction station coordinates. Observations between interior stations of block 1 contribute to the block 1 equations. Observations between stations interior to block 1 and junction stations also contribute to it. A similar statement can be made for block 2. Observations between junction stations contribute to the block in which the instrument was positioned. In this context note that the dashed line attributes uniquely a block to any station.

Adding the two systems of normal equations would result in the conventional normals for the entire network. However, elimination starts for each block separately. The unknowns x_1, x_2 are eliminated from the two systems by partial Cholesky reduction:

$$\begin{aligned} &\text{Block 1:} \begin{bmatrix} R_{11} & R_{13} \\ & B_{33}^{(p)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_3 \end{bmatrix} = \begin{bmatrix} s_1 \\ b_3^{(p)} \end{bmatrix} \\ &\text{Block 2:} \begin{bmatrix} Q_{22} & Q_{23} \\ & C_{33}^{(q)} \end{bmatrix} \begin{bmatrix} x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} t_2 \\ c_3^{(q)} \end{bmatrix} \end{aligned} \tag{3.37}$$

The two partially reduced systems for the unknowns x_3 are taken out and added:

$$\{B_{33}^{(p)} + C_{33}^{(q)}\} x_3 = \{b_3^{(p)} + c_3^{(q)}\}. \tag{3.38}$$

This system is solved for x_3. Back substitution into the two above systems yields x_1, x_2.

The solution is equivalent to the solution of the normals for the entire network. The proof of equivalence is fairly simple. During the partial Cholesky reduction modifications to the coefficients pertaining to x_3, i.e., to B_{33}, b_3, C_{33}, c_3 are made only by adding to or subtracting something from them. Because the quantities added or subtracted are the same as they would be if the entire system were partially reduced, it is irrelevant whether the equations for x_3 are added before or after the partial reduction.

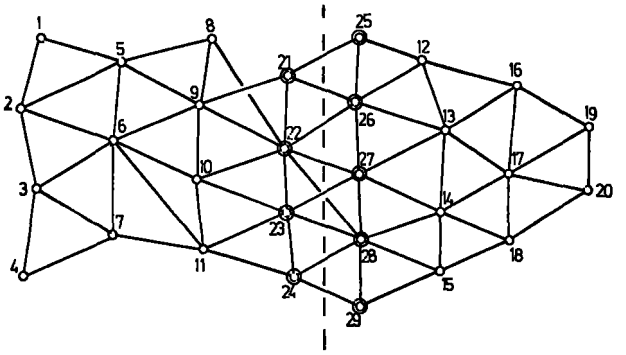

Figure 3.7.—Sample network decomposed into two Helmert blocks.

A larger network will be partitioned into more than two blocks. A hierarchy of blocks can be established that is similar to the nested dissection procedure. In fact, one can view figure 3.6 as a Helmert blocking scheme. There are as many first-level blocks as there are sets labeled 1, i.e., the number is 64. The normal equations are formed for each first level block separately. Higher labeled nodes situated in adjacent barrier sets take part in the normal equations as junction nodes. The dashed lines separating the first-level blocks have to be imagined as bisecting the barrier sets between the sets labeled 1. All observations must be used in forming the normals, and any observation must be used only once.

The interior nodes are eliminated from the first-level blocks. The partially reduced normals for the junction nodes of four adjacent earlier first-level blocks are added to form the normals of a second-level block. In such a second-level block the nodes labeled 2 now play the role of interior nodes. The junction nodes have labels higher than two. There are 16 second-level blocks. The number of blocks has been reduced by the factor of one-fourth. The interior nodes are eliminated from the second-level blocks, etc. Finally at the fourth and last level we deal with a system involving only the nodes labeled 4. We solve this system for the coordinates of these stations. Back substitution cascades down through the previous levels and successively yields the coordinates of the lower labeled stations.

What is the difference now between Helmert blocking as described here and nested dissection? Not much. In fact, Helmert blocking is slightly more sophisticated because the normals are not fully formed before reduction starts. Instead, the normals are formed separately for each first-level block. After partial reduction at any level, normals of a number

of blocks are merged by adding them. These operations have to be viewed as part of the formation of the normals rather than part of the solution process. This, by the way, conflicts somewhat with the definition of our goal, namely to analyze the roundoff errors arising and accumulating during the solution of the normal equations, assuming that the normals themselves are error-free. We will, therefore, redefine our goal so as to assume that the first-level normals are assembled error-free. Roundoff errors will be taken into account during the process of merging the partial normals of adjacent blocks. However, the influence of these errors will be marginal.

Returning to the interplay between Helmert blocking and nested dissection, George (1973) pointed out that substantial savings are realized in computer time and storage associated with the peculiar way of combining four i-level blocks to form one $i+1$-level block. Although Helmert blocking has been widely used by geodesists, I do not know of any reference where it has been done by nested dissection. Instead, in most cases, only two levels have been considered. Helmert or his geodetic followers did not appear to anticipate George's logarithmic law.

The U.S. network will be adjusted by the Helmert blocking technique. Partial reduction at the intermediate block level, as well as the reduction of the last level system will be done by Cholesky's method. First-level blocks will be ordered individually for small profile. Higher level blocks will also be ordered to some extent, but ordering becomes less significant as the systems tend to become less and less sparse.

4. ROUNDOFF ERRORS FOR A GENERAL POSITIVE DEFINITE SYSTEM

This chapter describes the general strategy of our roundoff error analysis. In order not to blur things with details, we first assume that a full Cholesky reduction is done for a positive definite system. As described in section 3.2 the NGS algorithm is assumed with the understanding that $p=n$, in which case partial reduction amounts to full reduction.

4.1 Roundoff Errors During the Triangular Decomposition Phase

The global effect of roundoff errors arising during the triangular decomposition phase will be studied by means of backward analysis. (See sec. 2.7.) The local roundoff errors encountered during triangular decomposition will be traced backward to the original equations $Ax = b$. A and b will then be perturbed by ε, η, so that the following perturbed system results:

$$(A + \varepsilon)(x + \zeta) = b + \eta. \tag{4.1}$$

The perturbation effect ζ on the solution x will be calculated in linear approximation as

$$\zeta = -A^{-1}\varepsilon x + A^{-1}\eta. \tag{4.2}$$

Mean and covariance of ε, η will be estimated based on the assumptions made in chapter 2. The mean and covariance of ζ will be estimated from some a priori knowledge of A^{-1}, x.

4.1.1 Left-hand-side local roundoff errors arising during triangular decomposition

The most serious left-hand-side roundoff errors occur during the calculation of the expressions:

$$a_{ij}^{(i-1)} = a_{ij} - \sum_{k=1}^{i-1} r_{ki} r_{kj}. \tag{4.3}$$

According to the NGS algorithm specified in section 3.2, eq. (3.15), the sequence of operations is such that the first sum

$$\sum_{k=1}^{i-1} r_{ki} r_{kj} \tag{4.4}$$

is calculated and subsequently subtracted from a_{ij}. During this calculation elementary errors and their aftereffects will cause a local error $\varepsilon_{ij}^{(p)}$ to affect $a_{ij}^{(i-1)}$. The superscript (p) stands for "product accumulation." Starting from assumedly correct values for a_{ij}, r_{ki}, r_{kj}, we will actually calculate

$$a_{ij} - \sum_{k=1}^{i-1} r_{ki} r_{kj} + \varepsilon_{ij}^{(p)}. \tag{4.5}$$

Our main problem in this section will be to estimate the mean and standard deviation of $\varepsilon_{ij}^{(p)}$ and some additional local errors. But let us first take a look beyond this goal and appreciate the benefit of backward analysis. We write the above expression as

$$(a_{ij} + \varepsilon_{ij}^{(p)}) - \sum_{k=1}^{i-1} r_{ki} r_{kj}. \tag{4.6}$$

The task of tracing the error backward is clearly trivial. We may simply replace the coefficient a_{ij} of the original normals by $a_{ij} + \varepsilon_{ij}^{(p)}$. In the case of $i \neq j$ we must take symmetry into account and also replace a_{ji} by $a_{ji} + \varepsilon_{ij}^{(p)}$, *i.e.*, we must superimpose $\varepsilon_{ij}^{(p)}$ on a_{ij} as well as on a_{ji}.

Let us now turn to the task of estimating $\varepsilon_{ij}^{(p)}$, $E\{\varepsilon_{ij}^{(p)}\}$, $\sigma\{\varepsilon_{ij}^{(p)}\}$. The local error $\varepsilon_{ij}^{(p)}$ results from the equation

$$a_{ij} - \sum_{k=1}^{i-1} r_{ki} r_{kj} + \varepsilon_{ij}^{(p)} = a_{ij} - \sum_{k=1}^{i-1}((r_{ki} r_{kj} + \varepsilon_{ijk}^{(*)}) + \varepsilon_{ijk}^{(+)}) + \varepsilon_{ij}^{(-)}. \tag{4.7}$$

Here $\varepsilon_{ijk}^{(*)}$ denote the elementary errors encountered

during the multiplications $r_{ki}r_{kj}$; $\varepsilon_{ijk}^{(*)}$ denote the elementary errors during the additions of $r_{ki}r_{kj}$ to the previously accumulated partial sum (obviously $\varepsilon_{ij1}^{(+)} = 0$); and $\varepsilon_{ij}^{(-)}$ denotes the elementary error of subtracting the fully accumulated sum from a_{ij}. We see that $\varepsilon_{ij}^{(p)}$ is represented as

$$\varepsilon_{ij}^{(p)} = \sum_{k=1}^{i-1}(-\varepsilon_{ijk}^{(*)} - \varepsilon_{ijk}^{(+)}) + \varepsilon_{ij}^{(-)}. \tag{4.8}$$

Recall from section 2.8 that we assume the elementary roundoff errors to be independent random variables. Hence $\varepsilon_{ij}^{(p)}$ is the superposition of independent random variables.

It follows at once that $E\{\varepsilon_{ij}^{(p)}\} = 0$, for a truly rounding machine and also for the CDC 6600. Hence, on such a machine $\varepsilon_{ij}^{(p)}$ is unbiased. We shall see later on that this is a great advantage.

Let us estimate the mean and standard deviation of $\varepsilon_{ij}^{(p)}$ by taking a look at the mean and standard deviation of the elementary roundoff errors composing it. In agreement with section 2.8, we have

$$\begin{aligned} E\{\varepsilon_{ijk}^{(*)}\} &= 0 && \text{on the CDC 6600} \\ |E\{\varepsilon_{ijk}^{(*)}\}| &\leqslant \frac{c_{ijk}^{(*)}}{2}\beta^{-\tau} && \text{on the IBM 360} \\ \sigma\{\varepsilon_{ijk}^{(*)}\} &\leqslant \frac{c_{ijk}^{(*)}}{\sqrt{12}}\beta^{-\tau} && \text{in any case.} \end{aligned} \tag{4.9}$$

Thereby $c_{ijk}^{(*)}$ is the smallest integer power of the base β bounding $|r_{ki}r_{kj}|$:

$$c_{ijk}^{(*)} = \beta^{\tau} > |r_{ki}r_{kj}| \tag{4.10}$$

In view of the remark at the end of section 3.2 concerning another version (3.16) of Cholesky's algorithm, we can also write

$$c_{ijk}^{(*)} = \beta^{\tau} > |a_{ij}^{(k)} - a_{ij}^{(k-1)}|. \tag{4.11}$$

In the case of $\varepsilon_{ijk}^{(+)}$, (4.9) holds with $c_{ijk}^{(*)}$ replaced by $c_{ijk}^{(+)}$. This is given by

$$\begin{aligned} c_{ijk}^{(+)} &= \beta^{\tau} > MAX\{|\textstyle\sum_{l=1}^{k-1} r_{li}\, r_{lj}|, |\sum_{l=1}^{k} r_{li}\, r_{lj}|, |r_{ki}r_{kj}|\} = \\ &= MAX\{|a_{ij} - a_{ij}^{(k-1)}|, |a_{ij} - a_{ij}^{(k)}|, |a_{ij}^{(k)} - a_{ij}^{(k-1)}|\}. \end{aligned} \tag{4.12}$$

Similarly

$$\begin{aligned} c_{ij}^{(-)} &= \beta^{\tau} > MAX\{|a_{ij}|, |a_{ij}^{(i-1)}|, |\textstyle\sum_{k=1}^{i-1} r_{ki}\, r_{kj}|\} = \\ &= MAX\{|a_{ij}|, |a_{ij}^{(i-1)}|, |a_{ij} - a_{ij}^{(i-1)}|\}. \end{aligned} \tag{4.13}$$

Since $\varepsilon_{ij}^{(p)}$ is given by (4.8) as the sum of independent random variables, we get

$$\begin{aligned} E\{\varepsilon_{ij}^{(p)}\} &= 0 && \text{on the CDC 6600} \\ |E\{\varepsilon_{ij}^{(p)}\}| &\leqslant \frac{c_{ij}^{(p)}}{2}(2\mu_{ij})\beta^{-\tau} && \text{on the IBM 360} \\ \sigma\{\varepsilon_{ij}^{(p)}\} &\leqslant \frac{c_{ij}^{(p)}}{\sqrt{12}}\sqrt{2\mu_{ij}}\,\beta^{-\tau} && \text{in any case} \end{aligned} \tag{4.14}$$

Here μ_{ij} denotes the number of *nonzero* product terms in the product sum (4.4). Twice this number gives the number of elementary operations during the calculation of (4.3) that may cause a nonzero roundoff error. The constant $c_{ij}^{(p)}$ is given by

$$\begin{aligned} c_{ij}^{(p)} &= \beta^{\tau} = \\ &= MAX\{c_{ijk}^{(*)}, c_{ijk}^{(+)}, k = 1, \ldots, i-1, c_{ij}^{(-)}\} > \\ &> MAX\{|a_{ij}|, |a_{ij}^{(i-1)}|, \\ &\qquad |r_{ki}\, r_{kj}|, k=1, \ldots, i-1, \\ &\qquad |\textstyle\sum_{k=1}^{l} r_{ki}\, r_{kj}|, l = 1, \ldots, i-1\}. \end{aligned} \tag{4.14a}$$

Note that the bound on the bias $E\{\varepsilon_{ij}^{(p)}\}$ grows in proportion to μ_{ij}, while the bound on $\sigma\{\varepsilon_{ij}^{(p)}\}$ grows in proportion only to the square root of this number.

Remark: Our formulas overestimate $E\{\varepsilon_{ij}^{(p)}\}$, $\sigma\{\varepsilon_{ij}^{(p)}\}$. Let us review the reasons for this. (1) The basic formulas of section 2.8 are already overestimates. (2) In the product-sum accumulations we bound all the elementary roundoff errors in terms of a single quantity $c_{ij}^{(p)}$, which will overestimate most of the quantities to be bounded. In particular, many of the individual product terms $r_{ki}r_{kj}$ will be much smaller than a_{ij}. We shall return to this in section 9.4 and improve the estimate of the U.S. network based on better insight into the behavior of the coefficients of the reduction states. (3) $E\{\varepsilon_{ij}^{(p)}\}$ is overestimated because the summands contributing to it will, in general, not all be of equal sign. Accordingly, there will be some offsetting of biases.

We proceed to other roundoff errors occurring during the triangular decomposition phase. There is the local roundoff error encountered while taking the square root of $a_{ii}^{(i-1)} = \hat{a}_{ii} - \Sigma\, r_{ki}\, r_{ki}$ (the sum is extended from $k=1$ to $k=i-1$). With an eye toward backward analysis we define this error in a slightly different way from section 2.8, namely such that the faulty square root is actually the precise result after taking the square root of the faulty radicand $a_{ii}^{(i-1)} + \varepsilon_{ii}^{(s)}$. We then may conveniently trace the error back by assuming an additional perturbance $\varepsilon_{ii}^{(s)}$ of the coefficient a_{ii} in the original system.

In agreement with the assumptions given in section 2.8.3, we estimate $\varepsilon_{ii}^{(s)}$ as follows:

$$\beta^{\tau} > r_{ii} = \sqrt{a_{ii}^{(i-1)}} \qquad (4.15)$$

$$c_{ii}^{(s)} = 2\, r_{ii}\, \beta^{\tau}.$$

Then

$$E\{\varepsilon_{ii}^{(s)}\} = 0 \qquad \text{on the CDC 6600}$$

$$|E\{\varepsilon_{ii}^{(s)}\}| \leqslant \frac{c_{ii}^{(s)}}{2}\beta^{-\tau} \qquad \text{on the IBM 360} \qquad (4.16)$$

$$\sigma\{\varepsilon_{ii}^{(s)}\} \leqslant \frac{c_{ii}^{(s)}}{\sqrt{12}}\beta^{-\tau} \qquad \text{in any case.}$$

Finally we deal with the errors occurring during the divisions $r_{ij} = a_{ij}^{(i-1)}/r_{ii}$. Again we shall define the errors in such a way that r_{ij} is the precise result of dividing $(a_{ij}^{(i-1)} + \varepsilon_{ij}^{(d)})$ by r_{ii}. Hence, let

$$\beta^{\tau} > |r_{ij}|$$
$$c_{ij}^{(d)} = r_{ii}\,\beta^{\tau}. \qquad (4.17)$$

Then

$$E\{\varepsilon_{ij}^{(d)}\} = 0 \qquad \text{on the CDC 6600}$$

$$|E\{\varepsilon_{ij}^{(d)}\}| \leqslant \frac{c_{ij}^{(d)}}{2}\beta^{-\tau} \qquad \text{on the IBM 360} \qquad (4.18)$$

$$\sigma\{\varepsilon_{ij}^{(d)}\} \leqslant \frac{c_{ij}^{(d)}}{\sqrt{12}}\beta^{-\tau} \qquad \text{in any case.}$$

4.1.2 Right-hand side local roundoff errors during triangular decomposition

Right-hand side local roundoff errors will be analogously defined in such a way that they may be viewed as perturbances of the right-hand side of the original normals. In view of the similarities to the previous section, we can shorten the presentation considerably.

The local errors $\eta_i^{(p)}$ occur during the product sum accumulation

$$b_i^{(i-1)} = b_i - \sum_{k=1}^{i-1} r_{ki}s_k. \qquad (4.19)$$

The right-hand side will be falsely evaluated as

$$b_i - \sum_{k=1}^{i-1} r_{ki}s_k + \eta_i^{(p)}. \qquad (4.20)$$

Define

$$d_i^{(p)} = \beta^{\tau} > \qquad (4.21)$$
$$MAX\{|b_i|, |b_i^{(i-1)}|, |r_{ki}s_k|, k=1,\ldots,i-1,$$
$$|\sum_{k=1}^{l} r_{ki}s_k|, l=1,\ldots,i-1\}.$$

Let ν_i denote the number of product terms $r_{ki}s_k$ that are different from zero. Then

$$E\{\eta_i^{(p)}\} = 0 \qquad \text{on the CDC 6600}$$

$$|E\{\eta_i^{(p)}\}| \leqslant \frac{d_i^{(p)}}{2}(2\nu_i)\beta^{-\tau} \qquad \text{on the IBM 360} \qquad (4.22)$$

$$\sigma\{\eta_i^{(p)}\} \leqslant \frac{d_i^{(p)}}{\sqrt{12}}\sqrt{2\nu_i}\,\beta^{-\tau} \qquad \text{in any case.}$$

Remark: In our quick arrival at (4.22) we have jumped over a detailed consideration of the individual elementary roundoff errors $\eta_{ik}^{(*)}$, $\eta_{ik}^{(*)}$, $\eta_i^{(-)}$, which could be introduced and analyzed completely parallel to the line of thought leading from eq. (4.8) to eq. (4.14a). Thus $d_i^{(p)}$ could be defined as

$$d_i^{(p)} = MAX\{d_{ik}^{(*)}, d_{ik}^{(*)}, k=1,\ldots,i-1, d_i^{(-)}\} \qquad (4.23)$$

with $d_{ik}^{(*)}$, $d_{ik}^{(*)}$, $d_i^{(-)}$ defined in analogy to (4.10), (4.12), (4.13).

The local errors $\eta_i^{(d)}$ occur during the *divisions* $s_i = b_i^{(i-1)}/r_{ii}$. We define them such that

$$(b_i^{(i-1)} + \eta_i^{(d)})/r_{ii} \qquad (4.24)$$

is the faultily evaluated result. Let

$$\beta^{\tau} > s_i$$
$$d_i^{(d)} = r_{ii}\,\beta^{\tau}. \qquad (4.25)$$

Then

$$E\{\eta_i^{(d)}\} = 0 \qquad \text{on the CDC 6600}$$

$$|E\{\eta_i^{(d)}\}| \leqslant \frac{d_i^{(d)}}{2}\beta^{-\tau} \qquad \text{on the IBM 360} \qquad (4.26)$$

$$\sigma\{\eta_i^{(d)}\} \leqslant \frac{d_i^{(d)}}{\sqrt{12}}\beta^{-\tau} \qquad \text{in any case.}$$

4.1.3 Global roundoff errors caused by triangular decomposition.

We combine the local roundoff errors introduced in the previous subsections as

$$\varepsilon_{ij} = \varepsilon_{ij}^{(p)} + \varepsilon_{ij}^{(s)} + \varepsilon_{ij}^{(d)}$$
$$\eta_i = \eta_i^{(p)} + \eta_i^{(d)}. \tag{4.27}$$

These errors ε_{ij}, η_i should still be considered local. Their mean and standard deviation can obviously be estimated as

$$|E\{\varepsilon_{ij}\}| \leqslant |E\{\varepsilon_{ij}^{(p)}\}| + |E\{\varepsilon_{ij}^{(s)}\}| + |E\{\varepsilon_{ij}^{(d)}\}|$$
$$\sigma\{\varepsilon_{ij}\} = \sqrt{\sigma^2\{\varepsilon_{ij}^{(p)}\} + \sigma^2\{\varepsilon_{ij}^{(s)}\} + \sigma^2\{\varepsilon_{ij}^{(d)}\}} \tag{4.28}$$

$$|E\{\eta_i\}| \leqslant |E\{\eta_i^{(p)}\}| + |E\{\eta_i^{(d)}\}|$$
$$\sigma\{\eta_i\} = \sqrt{\sigma^2\{\eta_i^{(p)}\} + \sigma^2\{\eta_i^{(d)}\}}. \tag{4.29}$$

It is important to note that the ε_{ij}, $1 \leqslant i \leqslant j \leqslant n$, as well as η_i, $1 \leqslant i \leqslant n$, are mutually independent random variables. We form the symmetric matrix $\varepsilon = (\varepsilon_{ij})$ and the vector $\eta = (\eta_i)$. The original system is then considered as being perturbed so that

$$(A + \varepsilon)(x + \zeta) = b + \eta. \tag{4.30}$$

As we have pointed out repeatedly, the perturbance ζ of x is given in linear approximation by

$$\zeta = -A^{-1}\varepsilon x + A^{-1}\eta. \tag{4.31}$$

We now denote the elements of the inverse A^{-1} by f_{ij}:

$$A^{-1} = (f_{ij}). \tag{4.32}$$

We may then expand the formula for ζ as:

$$\zeta_i = -\sum_{j=1}^{n}\sum_{k=1}^{n} f_{ij} x_k \varepsilon_{jk} + \sum_{j=1}^{n} f_{ij}\eta_j, \quad i = 1, \ldots, n. \tag{4.33}$$

Since $\varepsilon_{ij} = \varepsilon_{ji}$, we occasionally prefer to write this as

$$\zeta_i = -\sum_{j=1}^{n} f_{ij} x_j \varepsilon_{ij} - \sum_{j=1}^{n}\sum_{k=j+1}^{n} (f_{ij} x_k + f_{ik} x_j)\varepsilon_{jk} + \sum_{j=1}^{n} f_{ij}\eta_j, \qquad i = 1, \ldots, n. \tag{4.34}$$

This formula exhibits ζ_i as a sum of independent random variables. Mean and covariance of ζ_i can now be calculated as

$E\{\zeta_i\} = 0$ on an unbiased machine;
otherwise we have

$$E\{\zeta_i\} = -\sum_{j=1}^{n}\sum_{k=1}^{n} f_{ij} x_k E\{\varepsilon_{jk}\} + \sum_{j=1}^{n} f_{ij} E\{\eta_j\} \tag{4.35}$$

$$\sigma^2\{\zeta_i\} = \sum_{j=1}^{n} f_{ij}^2 x_j^2 \sigma^2(\varepsilon_{jj}\} + \sum_{j=1}^{n}\sum_{k=j+1}^{n} (f_{ij} x_k + f_{ik} x_j)^2 \sigma^2\{\varepsilon_{jk}\} + \sum_{j=1}^{n} f_{ij}^2 \sigma^2\{\eta_j\} \tag{4.36a}$$

$$Cov(\zeta_{i_1}, \zeta_{i_2}) = \sum_{j=1}^{n} f_{i_1 j} f_{i_2 j} x_j^2 \sigma^2\{\varepsilon_{jj}\} + \sum_{j=1}^{n}\sum_{k=j+1}^{n} (f_{i_1 j} x_k + f_{i_1 k} x_j) * (f_{i_2 j} x_k + f_{i_2 k} x_j)\sigma^2\{\varepsilon_{jk}\} + \sum_{j=1}^{n} (f_{i_1 j} f_{i_2 j})\sigma^2\{\eta_j\}. \tag{4.36b}$$

4.1.4 Preliminary estimates of the global roundoff errors in the U.S. network caused by triangular decomposition

Taking certain shortcuts and anticipating some results to be derived in chapters 5, 6, and 7, we will specify preliminary estimates of the global roundoff errors suffered by the coordinate shifts of the U.S. network during the first iteration of the adjustment.

The estimates derived here are rather simple. On the other hand, they are rather crude. They are sufficient to indicate the feasibility of the adjustment on the CDC 6600. More refined methods will have to be used in order to prove that the bias encountered on the IBM 360 will allow an adjustment on this machine without special precautionary measures.

Suppose that a bound $\|f\|$ is available on the elements f_{ij} of A^{-1}:

$$\|f\| = \underset{i,j}{MAX}\{|f_{ij}|\} = \underset{i}{MAX}\{f_{ii}\}. \tag{4.37}$$

Assume that a similar bound is available on the elements of x:

$$\|x\| = \underset{i}{MAX}\{|x_i|\}. \tag{4.38}$$

Suppose that c is a bound on all the quantities $c_{ijk}^{(*)}$ $c_{ijk}^{(+)}$, $c_{ij}^{(-)}$, $c_{ij}^{(s)}$, $c_{ij}^{(d)}$, involved in bounds on *elementary* roundoff errors during the transition from a_{ij} to $a_{ij}^{(i-1)}$:

$$c = \underset{1 \leqslant i \leqslant j \leqslant n}{MAX}\{c_{ijk}^{(*)}, c_{ijk}^{(+)}, k = 1, \ldots, i-1, c_{ij}^{(-)}, c_{ij}^{(s)}, c_{ij}^{(d)}\}. \tag{4.39}$$

Here we also count the square root as an elementary operation.

Then, by introducing a similar global bound for the right-hand side elementary roundoff errors:

$$d = \underset{1 \leqslant i \leqslant n}{MAX}\{d_{ik}^{(*)}, d_{ik}^{(+)}, k = 1, \ldots, i-1, d_i^{(-)}, d_i^{(d)}\}. \tag{4.40}$$

Now focus attention on one particular elementary roundoff error, call it $\varepsilon_{jk}^{(el)}$. This error enters equation (4.34) via the local error ε_{jk} of which it is a summand. It follows that the global effect $\zeta_i(\varepsilon_{jk}^{(el)})$ of $\varepsilon_{jk}^{(el)}$ is given by

$$\zeta_i(\varepsilon_{jk}^{(el)}) = \begin{cases} f_{ij}\, x_j\, \varepsilon_{jj}^{(el)} \ldots j = k \\ (f_{ij}x_k + f_{ik}x_j)\, \varepsilon_{jk}^{(el)} \ldots j \neq k \end{cases} \tag{4.41}$$

As a consequence, we can bound the bias and standard deviation of $\zeta_i(\varepsilon_{jk}^{(el)})$ as follows:

$$|E\{\zeta_i(\varepsilon_{jk}^{(el)})\}| \leqslant c\, \|f\|\, \|x\|\, \beta^{-\tau} \ldots \text{all } j,k \tag{4.42}$$

$$\sigma\{\zeta_i(\varepsilon_{jk}^{(el)})\} \leqslant \frac{2c}{\sqrt{12}} \|f\|\, \|x\|\, \beta^{-\tau} \ldots \text{all } j,k. \tag{4.43}$$

Similarly, for right-hand side errors we get

$$|E\{\zeta_i(\eta_i^{(el)})\}| \leqslant \frac{d}{2} \|f\|\, \beta^{-\tau} \tag{4.44}$$

$$\sigma\{\zeta_i(\eta_i^{(el)})\} \leqslant \frac{d}{\sqrt{12}} \|f\|\, \beta^{-\tau}. \tag{4.45}$$

All that remains is to count how many elementary operations are performed. Elementary operations with results that are known to be zero as a consequence of the sparseness of A can be excluded thereby. By Π, we denote the number of entries (i,j), $i \leqslant j$, such that $a_{ij}^{(p)} \neq 0$ for some p, $0 \leqslant p < i$. We call Π the number of nonzero locations. A nonzero location has either $a_{ij} \neq 0$, or a fill-in occurs in the location (i,j). Referring to the quantities μ_{ij}, introduced in section 4.1.1 (cf. eq. (4.14)), we also have

$$\Pi = \sum_{i=1}^{n} \mu_{ii} + n. \tag{4.46}$$

By Γ we denote the number of elementary operations needed to triangularize the matrix A. We note that

$$\Gamma = \sum_{1 \leqslant i \leqslant j \leqslant n} (2\mu_{ij} + 1) \tag{4.47}$$

with the understanding that $\mu_{ij} = 0$ for a zero location. Finally we note that the number of elementary operations needed to reduce the right hand side does not exceed 2Π.

Summing over all elementary roundoff errors during triangular decomposition, we arrive at the following estimates:

$E\{\zeta_i\} = 0 \ldots$ on the CDC 6600; otherwise:

$$|E\{\zeta_i\}| \leqslant \{c\, \|f\|\, \|x\|\, \Gamma + \frac{d}{2} \|f\|\, 2\Pi\}\, \beta^{-\tau} \tag{4.48}$$

$$\sigma\{\zeta_i\} \leqslant \{\frac{4c^2}{12} \|f\|^2 \|x\|^2 \Gamma + \frac{d^2}{12} \|f\|^2\, 2\Pi\}^{1/2}\, \beta^{-\tau}.$$

It will now be our task to specify numerical values for the various quantities occurring in these formulas. In the case of c, we will derive a slightly more general result than will be needed immediately. It will be of use in chapter 8, where more refined estimates will be obtained.

Proposition 4.1. Suppose that $\|a\|$ is a bound on the elements of the matrix A. Let $\bar{c}$ denote the smallest integer power β^{γ} bounding $\|a\|$. It then holds that

$$\begin{gathered} c_{ijk}^{(*)},\ c_{ijk}^{(+)},\ c_{ij}^{(-)} \leqslant \bar{c} \\ c_{ii}^{(s)} \leqslant 2\sqrt{\beta}\,\bar{c},\quad c_{ij}^{(d)} \leqslant \sqrt{\beta}\bar{c} \end{gathered} \tag{4.49}$$

Proof: Due to the positive definiteness of A, as well as of $A_{22}^{(p)}$, we have $|a_{ij}| \leqslant \underset{j}{\mathrm{MAX}}\, a_{ii}$, and $|a_{ij}^{(p)}| \leqslant \underset{i}{\mathrm{MAX}}\, a_{ii}^{(p)}$. Because $A_{22} - A_{22}^{(p)}$ is definite (this matrix is given by the product $A_{21} A_{11}^{-1} A_{12}$!), it follows that $a_{ii} \geqslant a_{ii}^{(p)}$. This is also clear from Cholesky's algorithm, because

$$a_{ii} - a_{ii}^{(p)} = \sum_{k=1}^{p} r_{ki}^2. \tag{4.50}$$

Examining the bounds $c_{ijk}^{(*)}$, $c_{ijk}^{(+)}$, $c_{ij}^{(-)}$, which were defined in (4.10), (4.12), (4.13), we see that the first part of the proposition is proved if we can show that

$$|a_{ij}^{(p)} - a_{ij}^{(q)}| \leqslant \|a\|. \tag{4.51}$$

By virtue of Cholesky's algorithm (cf., version (3.15) in section 3.2) we have

$$|a_{ij}^{(p)} - a_{ij}^{(q)}| \leqslant \sum_{k=q+1}^{p} |r_{ki} r_{kj}|. \tag{4.52}$$

Applying Schwarz's inequality, we further bound this as

$$|a_{ij}^{(p)} - a_{ij}^{(q)}| \leqslant \sqrt{\sum_{k=q+1}^{p} r_{ki}^2 \sum_{k=q+1}^{p} r_{kj}^2} \leqslant$$

$$MAX\{\sum_{k=q+1}^{p} r_{ki}^2, \sum_{k=q+1}^{p} r_{kj}^2\} =$$

$$= MAX\{|a_{ii}^{(p)} - a_{ii}^{(q)}|\,,\ |a_{jj}^{(p)} - a_{jj}^{(q)}|\} \leqslant \|a\|. \tag{4.53}$$

Hence the first part of the proposition is shown. The proof of the second part is shorter. We restrict it to $c_{ii}^{(s)}$. According to (4.15), we have to look for a bound of $r_{ii} = \sqrt{a_{ii}^{(i-1)}}$ which is an integer power of β. If $\bar{c} = \beta^{\gamma}$ bounds $\|a\|$, then either the expression $\beta^{\gamma/2}$ or the expression $\beta^{(\gamma+1)/2}$ bounds $\sqrt{a_{ii}^{(i-1)}}$, depending on γ being even or odd. Again by (4.15) this bound must be multiplied by $2r_{ii}$. The result is bounded by $2\beta^{(\gamma+1)/2} = 2\sqrt{\beta}\,\bar{c}$, as was to be shown.

Recall from the remarks at the end of section 3.4 that a_{ii} is the reciprocal variance of coordinate i if all the other coordinates are fixed. Hence, in a way a_{ii} measures the accuracy of coordinate i with respect to the coordinates of the neighboring stations. Faced with a certain network, one will have a priori knowledge of what this accuracy will be. In the case of the U.S. network this accuracy may be very high at stations where very precise distance measurements have been taken. There are many such stations, because the distances to the reference marks are frequently measured with an rms error of 1 mm. Hence $\bar{c}$ is very large for the U.S. network. It is about 10^6 to 10^7 m^{-2}. As one of the biggest sources of trouble, we will be concerned about it in chapter 7. To determine $\bar{c}$ properly, scaling considerations are involved, which are explained in chapter 8. As argued in section 8.2.1, we may take $\bar{c} = 5*10^6$ for both machines. Then $c = 14*10^6$ on the CDC 6600 and $c = 4*10^7$ on the IBM 360.

Our next concern is for $\|f\|$, the element bound for the inverse A^{-1}. Of course, one may take $\|f\|$ as the largest variance of any adjusted coordinate. A priori estimates of this variance are not too difficult to obtain. Much literature exists on the mathematical structure of the covariance matrix of large regular adjusted networks. There are methods to estimate the covariance to such a degree of accuracy that the requirements of a roundoff analysis are easily met. Anticipating the developments in chapter 5, we can already indicate that $\|f\|$ is close to 0.25 m^2 for the U.S. network. This corresponds to an rms accuracy of $\sqrt{0.25} = 0.5$ m.

The reader may be surprised that our formulas also require an a priori estimate of the solution vector x. Actually only a rough estimate of the largest component of x is required. For the U.S. network a few coordinate shifts are expected to exceed 10 m. Even if such a priori information were not available, our formulas would not be completely useless. The formulas still predict how many significant digits of the solution vector's largest component would be saved, i.e., remain unperturbed by roundoff.

We now turn toward d, the bound on the right-hand side quantities $d_i^{(p)}$, $d_i^{(d)}$. Part of the quantities that are to be bounded by d are the original right-hand-sides b_i and the partially reduced right hand sides $b_i^{(p)}$. From the discussion in section 3.4 we know that $b_i^{(p)}/a_{ii}^{(p)}$ is the shift of coordinate i away from its approximate position provided that coordinates k, $k>p$, $k\neq i$ are held fixed to their approximate positions and coordinates k, $k\leqslant p$, $k=i$ are allowed to adjust freely. We stay on the safe side if we take the maximum shift of 10 m together with the maximum value of a_{ii}, which gives us $5*10^6$. We arrive at $d = 5*10^7$.

Other quantities that must be bounded by d are the sums

$$\sum_{k=1}^{i} r_{ki}s_k. \tag{4.54}$$

These sums are actually the differences between b_i and $b_i^{(l)}$. Hence, we cover this case as well if we double the bound d. We arrive at $d = 10^8$. We must still take into account that d is understood as an integer power of the base β bounding all quantities in question. We stay on the safe side if we take $d = 2*10^8$ on the CDC 6600 and $d = 16*10^8$ on the IBM 360.

To obtain a rough idea of the size of Π and Γ, we imagine a network of some 200,000 stations homogeneously covering a square region. There are $\sqrt{200{,}000} = 450$ stations across the network. If we assume that minimum bandwidth ordering is performed, we find that a typical node is connected to about $2*450 = 900$ other nodes forming a barrier across the network. Taking into account that any station has two coordinates, we arrive at a bandwidth of $2*900 = 1800$. For a system of n equations and bandwidth w, Γ grows asymptotically as

$$\Gamma = n\,w^2. \tag{4.55}$$

(This is n. . .number of equations, times w. . .number of coefficients per equation, times $w/2$. . .average number of product sum accumulations per coefficient, times 2. . .since a multiplication and a summation is involved.) In this case we get $\Gamma = 400{,}000*1800^2 = 1.3*10^{12}$. In a later section we shall see that, as a result of the inhomogeneous distribution of the stations and the beneficial effect of Helmert blocking, we actually have $\Gamma = 1.2*10^{11}$. The number Π is about $n*w = 7.2*10^8$ for the homogeneous network, whereas it decreases to about $1.4*10^8$ for the real network.

Inserting these numbers into our formulas (4.48) for $E\{\zeta_i\}$, $\sigma\{\zeta_i\}$, for the CDC 6600 we get

$$E\{\zeta_i\} = 0 \tag{4.56}$$

$$\begin{aligned}\sigma\{\zeta_i\} &= [4*(14*10^6)^2/12 * .25^2 * 10^2 * 1.2*10^{11} + \\ &\quad + (2*10^8)^2/12 * .25^2 * 2 * 1.4*10^8]^{1/2} *2^{-48} = \\ &= (7.0*10^{12})^2 + (2.4*10^{11})^2]^{1/2} * 2^{-48} = \\ &= 0.025 \text{ m}\end{aligned}$$

whereas for the IBM 360 we get

$$\begin{aligned}|E\{\zeta_i\}| &\leqslant [4*10^7 * .25 * 10 * 1.2*10^{11} + \\ &\quad + 16*10^8/2 * .25 * 2 * 1.4*10^8] *16^{-14} = \\ &= [1.2*10^{19} + 5.6*10^{16}] * 16^{-14} = \\ &= 167 \text{ m}\end{aligned} \tag{4.57}$$

$$\begin{aligned}\sigma\{\zeta_i\} \leqslant\ & [4*(4*10^7)^2/12*.25^2*10^2*1.2*10^{11} + \\ & + (16*10^8)^2/12*.25^2*2*1.4*10^8]^{1/2}* \\ & 16^{-14} = \\ =\ & [(2.0*10^{13})^2 + (1.93*10^{12})^2]^{1/2}*16^{-14} = \\ =\ & 0.00028 \text{ m.}\end{aligned}$$

We know already from these crude estimates that the CDC 6600 is a safe machine on which to perform the adjustment. On the IBM 360, the bound on $\sigma\{\zeta_i\}$ is small enough, but the bound on $E\{\zeta_i\}$ is larger than the expected shift. Refined estimates of $E\{\zeta_i\}$ are therefore necessary.

Remember that the adjustment will be iterated at least once for reasons of nonlinearity. At the second iteration, the shifts will be much smaller, as will be the right-hand sides and, therefore, also the roundoff errors.

4.2 Roundoff Errors During Back Substitution

The roundoff errors arising and accumulating during the back substitution phase

$$x_i = (s_i - \sum_{j=i+1}^{n} r_{ij}x_j)/r_{ii},\ i = n, n-1, \ldots, 2, 1 \tag{4.58}$$

will also be treated by backward analysis. However, local errors will not be traced backwards all the way to the original equations $Ax = b$, but rather to the triangularized system $Rx = s$. It will be possible to trace the errors backward to the right-hand side only, so that we will be faced with the perturbed system

$$R(x + \zeta) = s + \eta. \tag{4.59}$$

We use the same symbols ζ, η for the errors of back substitution. In linear approximation, the perturbation ζ of x is given by

$$\zeta = R^{-1}\eta. \tag{4.60}$$

In the case of the U.S. network, R and R^{-1} have elements that are very different in size. This persuades us to split R to:

$$R = R_D R_G. \tag{4.61}$$

Here R_D is the diagonal matrix composed of the diagonal elements r_{ii} of R. Accordingly, R_G is an upper triangular matrix whose diagonal elements equal 1, while the off-diagonal elements are given by r_{ij}/r_{ii}. R_G and R_G^{-1} will have elements that are much more uniform in size. The above equation for ζ is rewritten as

$$\zeta = R_G^{-1} R_D^{-1} \eta. \tag{4.62}$$

This prompts us to redefine η as

$$\eta_{new} = R_D^{-1}\eta_{old}. \tag{4.63}$$

Hence we replace the equation for ζ by

$$\zeta = R_G^{-1}\eta_{new} = R_G^{-1}\eta, \tag{4.64}$$

writing η for η_{new} from now on.

4.2.1 Local errors during back substitution

As in the triangular decomposition phase, we are led to introduce local errors for the product-sum accumulations in

$$s_i - \sum_{j=i+1}^{n} r_{ij}x_j. \tag{4.65}$$

The result of calculating this by means of rounding machine arithmetic will be the same as if a precise calculation is done on the basis of s_i being perturbed by $r_{ii}\eta_i^{(p)}$:

$$s_i + r_{ii}\eta_i^{(p)} - \sum_{j=i+1}^{n} r_{ij}x_j. \tag{4.66}$$

The factor r_{ii} in front of $\eta_i^{(p)}$ results from our redefinition of η. Then $d_i^{(p)}$ is defined by

$$\begin{aligned} & r_{ii}\, d_i^{(p)} = \beta^{\gamma} > \\ & Max\{|s_i|, |r_{ij}x_j|, j = i+1, \ldots, n, \\ & |\sum_{j=i+1}^{l} r_{ij}x_j|, l = i+1, \ldots, n\}. \end{aligned} \tag{4.67}$$

Further, let ν_i denote the number of nonzero coefficients r_{ij} in row i of R. We then have

$$\begin{aligned} & E\{\eta_i^{(p)}\} = 0 && \text{on the CDC 6600} \\ & |E\{\eta_i^{(p)}\}| \leqslant \frac{d_i^{(p)}}{2}(2\nu_i)\beta^{-\tau} && \text{on the IBM 360} \\ & \sigma\{\eta_i^{(p)}\} \leqslant \frac{d_i^{(p)}}{\sqrt{12}}\sqrt{2\nu_i}\,\beta^{-\tau} && \text{in any case.} \end{aligned} \tag{4.68}$$

The remaining step is an analysis of the elementary error of the divisions by r_{ii}. Assuming that the divisor and quotient are correct, the faulty result of the division is assumed to be the correct result of

$$(s_i + r_{ii}\eta_i^{(d)} - \sum_{j=i+1}^{n} r_{ij}x_j)/r_{ii}. \tag{4.69}$$

The quotient should be x_i. We therefore define $d_i^{(d)}$ by

$$r_{ii}\, d_i^{(d)} = r_{ii}\beta^{\gamma},\ \beta^{\gamma} > |x_i|$$

i.e.,

$$d_i^{(d)} = \beta^{\tau} > |x_i|. \tag{4.70}$$

Then we get

$$\begin{aligned} E\{\eta_i^{(d)}\} &= 0 && \text{on the CDC 6600} \\ |E\{\eta_i^{(d)}\}| &\leqslant \frac{d_i^{(d)}}{2}\beta^{-\tau} && \text{on the IBM 360} \\ \sigma\{\eta_i^{(d)}\} &\leqslant \frac{d_i^{(d)}}{\sqrt{12}}\beta^{-\tau} && \text{in any case.} \end{aligned} \tag{4.71}$$

4.2.2 Global roundoff errors resulting from back substitution

We combine the local roundoff errors treated in the previous subsection

$$\eta_i = \eta_i^{(p)} + \eta_i^{(d)}. \tag{4.72}$$

Again, by applying the rules for adding two independent random variables, we will have the mean and standard deviation of η_i:

$$\begin{aligned} |E\{\eta_i\}| &\leqslant |E\{\eta_i^{(p)}\}| + |E\{\eta_i^{(d)}\}| \\ \sigma\{\eta_i\} &= \sqrt{\sigma^2\{\eta_i^{(p)}\} + \sigma^2\{\eta_i^{(d)}\}}. \end{aligned} \tag{4.73}$$

The η_i, in turn, are independent random variables. They form the vector η. As we know, the global disturbance ζ of x follows from

$$\zeta = R_G^{-1}\eta. \tag{4.74}$$

We denote by ϱ_{ij} the elements of R_G^{-1}

$$R_G^{-1} = (\varrho_{ij}). \tag{4.75}$$

Then

$$\begin{aligned} E\{\zeta_i\} &= 0 && \text{on the CDC 6600} \\ E\{\zeta_i\} &= \sum_{j=i}^{n} \varrho_{ij} E\{\eta_j\} && \text{on the IBM 360} \\ \left.\begin{aligned} \sigma\{\zeta_i\} &= \sum_{j=i}^{n} \varrho_{ij}^2 \sigma^2\{\eta_j\} \\ Cov\{\zeta_{i_1}, \zeta_{i_2}\} &= \sum_{j=\max(i_1,i_2)}^{n} \varrho_{ij_1}\varrho_{ij_2}\sigma^2\{\eta_j\} \end{aligned}\right\} && \ldots \\ && \ldots \text{ in any case.} \end{aligned} \tag{4.76}$$

4.2.3 Preliminary estimates of the global roundoff errors in the U.S. network resulting from back substitution

As discussed in section 4.1.4, we now take a preliminary look at the U.S. network and try to get a crude estimate of the global roundoff errors, caused by back substitution, and suffered by the coordinate shifts after the first iteration.

Again we take the chance of grossly overestimating things and use a single bound d on the various quantities $d_i^{(\ )}$. As in section 4.1.4, we go back to the elementary roundoff errors and argue that the contribution of any elementary roundoff error $\eta^{(el)}$ toward $E\{\zeta_i\}$ is bounded by

$$d\,\|\varrho\|\,\beta^{-\tau}/2. \tag{4.77}$$

Here $\|\varrho\|$ is a bound on the elements ϱ_{ij} of R_G^{-1}. Counting the elementary operations, we find no more than 2Π. Hence we get

$$\begin{aligned} |E\{\zeta_i\}| &< d\,\|\varrho\|\,\Pi\,\beta^{-\tau} \\ \text{and similarly} \\ \sigma\{\zeta_i\} &< d\,\|\varrho\|\,\sqrt{2\Pi}\,\beta^{-\tau}/\sqrt{12}. \end{aligned} \tag{4.78}$$

Chapter 7 will demonstrate that we can take $d = \|x\|$, $\|\varrho\| = 1$. Although these bounds are not quite strict, they will hold for most of the elementary operations counted in the number Π. The resulting errors are quite small. On the CDC 6600 we get

$$\begin{aligned} E\{\zeta_i\} &= 0 \\ \sigma\{\zeta_i\} &\leqslant 1.7*10^{-10}\ \text{m}. \end{aligned} \tag{4.79}$$

On the IBM 360 we have

$$\begin{aligned} |E\{\zeta_i\}| &\leqslant 1.9*10^{-8}\ \text{m} \\ \sigma\{\zeta_i\} &\leqslant 6.7*10^{-13}\ \text{m}. \end{aligned} \tag{4.80}$$

4.3 Taking into Account Helmert Blocking

Recall that all preceding derivations in section 4 were made under the assumption that a full Cholesky reduction is done in one sweep. What modifications will occur now if reduction proceeds according to the Helmert blocking scheme? They are rather minor. As we argued in section 3.5.5, Helmert blocking implies a certain favorable ordering of the stations. This ordering will make Π, i.e., the number of nonzero locations, small. Also Γ will be kept small; hence the favorable effect on roundoff error propagation.

Aside from its implications on station ordering, Helmert blocking introduces modifications of the computer algorithm that must be examined. First, we observe that instead of calculating the sums of products in expressions like

$$a_{ij}^{(i-1)} = a_{ij} - \sum_{k=1}^{i-1} r_{ki}r_{kj} \tag{4.81}$$

in one sweep and subtracting them afterwards from a_{ij}, the sums of the products are now computed in a number of steps. At the end of any step a certain partial reduction at a certain block level comes to its end, and the calculated segment of the inner product is subtracted from a previously partially reduced coefficient. The result is a new coefficient which, for the inner nodes, is almost fully reduced, while for the junction nodes it is again only partially reduced.

The number of operations needed to calculate the product terms $r_{ki}r_{kj}$, to accumulate them, and to subtract the accumulated sums, is the same whether the normals are solved in one sweep or whether they are subjected to Helmert blocking. Additional operations are required when blocks are combined.

After the partial reduction of a number of adjacent blocks of level i, the partially reduced normals of their junction nodes are added to form the normals of an $i+1$-level block. This causes additional operations as compared to the reduction in one sweep. Although one could argue that these additional operations are part of the process to form the normal equations, rather than part of the process of solving them, we agreed in section 3.5.5 to include the roundoff errors of these additions in our error budget.

For a large number of nodes, i.e., the first-level nodes, there are no additional operations at all. Let us now concentrate on an i-th level node. We call such a node "regular" if it is not situated at the center of a typical four-lobed set of the i-th level and if it is not situated near a spot on one of the four lobes where a lower level set touches; (see fig. 4.1 and fig. 3.6 in section 3.5.4.) Clearly, the majority of higher level nodes will be regular. It is remarkable that for any regular node only one additional operation per nonzero coefficient is susceptible to a roundoff error. This operation is an addition taking place when the partially reduced normals of the two adjacent i-1-level blocks are added. For other than regular nodes more operations may be required, but their number would never exceed three. Shaded areas in figure 4.1 indicate the locations of such nodes.

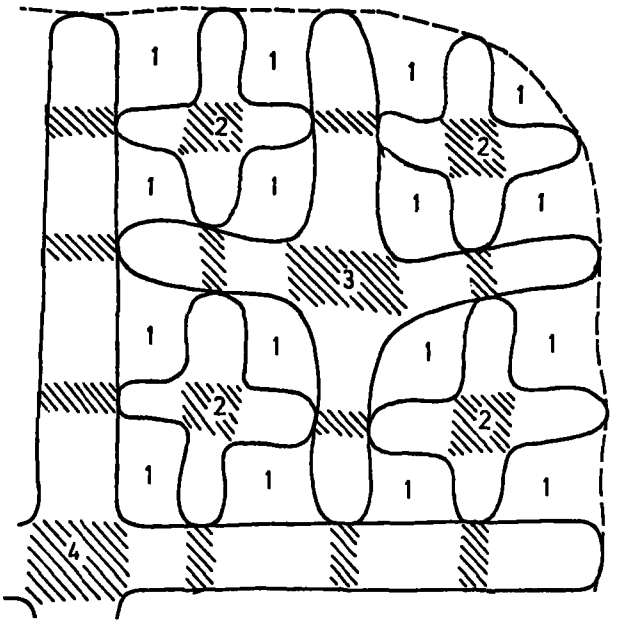

Figure 4.1.—Regular (unshaded) and exceptional (shaded) nodes in nested dissection.

4.4 Effect of Scaling the Normals on Roundoff Error Propagation

In this subsection we follow another sidetrack. Consider the problem of applying scale factors to the unknown parameters. It is occasionally argued that such scaling can improve numerical stability. This is a myth for the direct elimination method done in floating point arithmetic, as it is also clearly stated in Jennings (1977, p. 115 et seq.)

If we change the scale of the unknowns x in the normal equations

$$A\,x = b \tag{4.82}$$

we replace them by new unknowns $\bar{x}$, according to

$$x = M\bar{x}. \tag{4.83}$$

Here M is a diagonal matrix of scale factors, $M = \mathrm{diag}(m_1, \ldots, m_n)$. The scaled normals are

$$\bar{A}\,\bar{x} = \bar{b}$$

with (4.84)

$$\bar{A} = M A M, \quad \bar{b} = M b, \quad \bar{x} = M^{-1}x.$$

The matrix $\bar{A}$ has elements $\bar{a}_{ij} = m_i m_j a_{ij}$. The vector $\bar{b}$ has components $\bar{b}_i = m_i b_i$. The solution of the scaled and unscaled equations are related by $\bar{x}_i = x_i/m_i$.

First suppose that the scale factors are all integer powers of the base β. We assert that under this condition scaling has no influence on rounding. More precisely, the global roundoff errors $\zeta, \bar{\zeta}$ of the two systems are related by the same transformation as the unknowns: $\zeta = M\bar{\zeta}$, $\zeta_i = m_i\bar{\zeta}_i$. The proof for this assertion follows from the observation that all floating point numbers occurring in the original and the reduced normal equations of the two systems have identical mantissas. Only the exponents differ; but they are integers, and, therefore, not subject to roundoff. It can be shown by induction that the following identities hold: $\bar{a}_{ij}^{(p)} = m_i m_j a_{ij}^{(p)}$, $\bar{b}_i^{(p)} = m_i b_i^{(p)}$,

$\bar{r}_{ij} = m_j r_{ij}$, $\bar{s}_i = s_i$. We will refrain from giving a formal proof. The reader can verify this relation by solving a 2×2 normal equation system using decimal numbers. If the system is scaled by $m_1 = 10$, $m_2 = 100$, and solved again, the same sequences of digits appear at corresponding numbers and only the position of the decimal point changes.

Consider now the case where the scale factors are not integer powers of base β. Deviations in the roundoff will then occur. In fact, since roundoff is a random process and is excited by different trailing digits in the scaled system, the deviations may be substantial. Nevertheless, after rescaling backwards the roundoff errors in the scaled system, we can show that the two random vectors ζ and $M\bar{\zeta}$ have mean and covariance of comparable size.

The following argument suffices for the left-side roundoff errors during triangular decomposition. The other cases are treated similarly. The relevant formula is (see sec. 4.1.3, eq. (4.31)):

$$\zeta = -A^{-1} \varepsilon x. \tag{4.85}$$

Applied to the transformed equations, this gives

$$\bar{\zeta} = -\bar{A}^{-1} \bar{\varepsilon} \bar{x} = -M^{-1} A^{-1} M^{-1} \bar{\varepsilon} M^{-1} x. \tag{4.86}$$

Scaling backwards, we get

$$M\bar{\zeta} = -\bar{A}^{-1} (M^{-1} \bar{\varepsilon} M^{-1}) x. \tag{4.87}$$

It remains to be shown that the local errors ε and $M^{-1} \bar{\varepsilon} M^{-1}$ have nearly the same mean and covariance structure. Since M is diagonal, we have to show that ε_{ij} and $\bar{\varepsilon}_{ij}/(m_i m_j)$ have comparable mean and covariance. This is plausible because mean and variance of a local roundoff error ε_{ij} depend on sign and magnitude of the coefficients $a_{ij}^{(p)}$ and on changes of these coefficients for varying p.

If the mean and variance of ε_{ij} depended precisely on these numbers, then the mean and variance of ε_{ij} and $\bar{\varepsilon}_{ij}/(m_i m_j)$ would be the same. Unfortunately, it is always the next integer power of the base β bounding the operands in an elementary operation that is responsible for the roundoff during an elementary operation. Hence mean and variance of local and global roundoff errors in the unscaled and scaled system can be expected to agree within only one power of the base β, or, stated differently, agreement can be expected within only one digit of the computer-internal number system. On the CDC 6600 we have $\beta = 2$ which makes the difference marginal. On the IBM 360 we have $\beta = 16$, which is not as marginal.

I do not anticipate that the normal equations of the U.S. network adjustment are scaled for numerical purposes. Previous discussion has shown that such a scaling would be useless anyway. Occasionally we use scaling in a different way, as a theoretical tool to arrive at better estimates. The normal equations themselves will not be scaled—only the formulas that predict the roundoff error. Let us outline this in the following paragraph.

Cholesky's algorithm is applied to the unscaled normals $Ax = b$. The local roundoff errors ε_{ij} are bounded as shown in section 4.1.1, using integer powers of the base β that bound operands and results of elementary operations involving the coefficients $a_{ij}^{(p)}$. Restricting attention to the analysis of left-side triangularization roundoff errors, we deal with the formula

$$\zeta = -A^{-1} \varepsilon x. \tag{4.88}$$

It is only now when further treating this formula, that we employ the scale factors:

$$\bar{\zeta} = M^{-1} \zeta = -(M^{-1} A^{-1} M^{-1})(M \varepsilon M)(M^{-1} x). \tag{4.89}$$

This is rewritten as

$$\bar{\zeta} = -\bar{A}^{-1} \bar{\varepsilon} \bar{x}. \tag{4.90}$$

The ε are no longer roundoff errors occurring during reduction of the scaled system. They are the result of scaling the roundoff errors during reduction of the unscaled system. The advantage of using scale factors may be the following: $\bar{F} = \bar{A}^{-1}$, $\bar{A}$, $\bar{\varepsilon}$, $\bar{x}$ may be much more homogeneous than F, A, ε, x.

Hence, bounds $\|\bar{f}\|$, $\|\bar{a}\|$, $\|\bar{\varepsilon}\|$, $\|\bar{x}\|$ may be better in the sense that they come closer to the average size of the elements, whereas without scaling, bounds $\|f\|$, $\|a\|$, $\|\varepsilon\|$, $\|x\|$ must be chosen in agreement with a few outliers among the elements of these matrices and vectors.

Remark: After completing most of the manuscript, a published paper by Beresford/Parlett (1976) came to my attention. This work also points out that scaling has negligible influence. My other conclusions about widespread misconceptions were also confirmed; for example, the effect of a proper or improper choice of norms is not commonly understood. (See discussion in section 11.1.)

5. PROPERTIES OF THE U.S. NETWORK RELEVANT TO THE ROUNDOFF STUDY

A serious difficulty to be faced in our roundoff study is the limited amount of information on the U.S. network presently available for quick access. The N G S is setting up a data base for station and

observation information, but it will be several years before it is fully operational. Currently, only station files can be accessed. There is no way to perform statistical searches on the observations in a reasonable amount of time. The manner in which the observations connect the stations and the weights that will be assigned to them cannot be completely determined now. Lack of information will cause our estimates to deteriorate somewhat. We will have to rely on insight gained from: (1) searching the station files, (2) looking at various generalized diagrams of the whole network or representative portions, (3) analyzing small subnetworks, and (4) interviewing persons who are studying various special problems.

The next subsection gives a general overview of the properties of the U.S. ground control network that must be considered carefully in the roundoff study. Later in this section we will deal in more detail with some of these individual properties, arriving at estimates of the coefficients of the inverse of the network's normal equation matrix.

5.1 General Overview

5.1.1 Size of the network

The network will contain approximately 170,000 stations. About one-third of these represent first- and second-order triangulation stations, as well as transcontinental traverse stations which strengthen the network. Two-thirds of the stations will be supplementary. These include stations of local traverses, intersection and resection stations, reference marks, excentrics and others. Supplemental stations contribute little to the strength of the network. They could be eliminated prior to the adjustment, as it is done in a classical network computation. However, they will be retained to save labor costs during the adjustment of the U.S. net. Eliminating the supplemental stations would require special considerations for a large number of differently structured local systems of stations. Such considerations could be done successfully only by an expert. The computer programs would have to be much more sophisticated, and the additional programing could cause another bottleneck. Dealing with all stations simultaneously, as NGS does, shifts the burden of manual work to the computer. This causes some problems. Computation time increases to be sure. But the worst problem is the very precise and heavily weighted ties between closely spaced stations that make the normal equation matrix much more ill conditioned than it would be otherwise.

5.1.2 Type of observations

I estimate that there will be 2 million to 3 million observations. About 99 percent contain unoriented directions. About 1 percent, namely 20,000 to 30,000 involve distances. About 0.1 percent or 2,000 to 3,000, are astronomical azimuths. The positional fix of the network will be done by means of about 130 Doppler stations.

5.1.3 Inhomogeneity of the network

The density of the stations varies from 0 to 3,000 per $1° \times 1°$ quad. (A quad is a region bounded by two meridians and two parallels. Although the area of a $1° \times 1°$ quad is not constant as latitude varies, point densities are calculated with respect to these quads.) Figures 5.1a-c show how station density varies over the conterminous United States. Figure 5.1b shows that the highest densities are near the east and west coasts.

Another source of inhomogeneity is the observation weights. The program assigns rms errors to the observations according to predefined formulas and quality indicators (Schwarz 1978). To simplify our discussion, we assume that all rms errors are measured in a common unit of measure, the meter. Then it is clear what the rms error of a distance or a Doppler position means. For directions and azimuths the rms error will refer to the lateral uncertainty of the line of vision with respect to the target point. (This lateral deviation is called "Perpendikel" in the German literature.) Defined in this way, the rms error of an observation will never be less than 0.001 m. However, it may possibly increase to 1 m. As a consequence, the diagonal elements of the normal equation matrix will be in the range of perhaps 1 m^{-2} to 10^7 m^{-2}. Off-diagonal elements of comparable size will be arranged in such a way that the matrix has very unfavorable numerical properties.

5.1.4 Structure of the network

The U.S. first-order network was originally designed as a system of intersecting triangulation chains. The chains were composed of basis figures which are mainly quadrangles with diagonals. After about four or five basis figures there will be an intersection of two chains. (See figs. 5.2 and 5.3.)

Lines of vision in this chain-type first-order network typically range from 10 to 30 km, but there are outliers. Base lines are measured preferably at the intersections of the chains, but this does not mean that there is a base line at every such intersection.

	50	48	46	44	42	40	38	36	34	32	30	28	26	24
060														
062		16	38	151	46									
		31	7	8	66	2								
064		58	9	6	38	61	18							
		26	37	2	168	155	139							
066			37	13	12	137	184							
			17	18	8	98	158	38						
068			14	49	28	54	92	184						
			7	22	9	12	29	197	62	6				
070				45	18	12	19	38	96	82	66	32	32	
				14	18	12	22	38	131	192	159	111	82	87
072					38	18	28	22	48	188	81	29	71	84	9
					8	28	21	18	18	28	16	18	186	41	27
074					18	9	19	17	7	19	23	67	47	26	36	92
						3	29	38	24	14	13	48	94	28	6	77	2		8	42	14	3	5
076							26	19	17	18	13	14	48	17	18	38	61	78	58	24	5	1	
			1		7	36	21	27	11	9	18	18	18	13	6	15	9	12	11	24	58		
078			14	11	4	18	48	26	11	18	19	13	26	11	4	8	34	24	15	4			
			66	27	5	6	11	18	28	45	29	22	22	13	1	43	23	13	19	18			
080			18	12	5	4	5	38	29	14	42	21	19	12	5	12	12	14	28	11			
			8	9	5	9	8	32	35	12	33	23	44	18	4	5	3	8	55				
082		3	12	17	18	12	18	43	24	16	18	24	29	17	8	13	6	8	44				
	1	9	8	16	28	31	37	31	23	18	21	28	9	25	28	28	28	19	57	2			
084	1		8	18	7	29	38	29	28	13	24	17	24	39	29	16	15	19	35	72	4		
	2	2	13	13	18	13	18	14	11	18	15	18	19	28	27	29	37	25	19	45	1		
086	2	8	13	18	7	22	11	9	18	7	14	13	14	12	23	15	13	16	18	29	1		
	4	11	22	9	12	12	8	12	7	14	28	15	11	12	25	19	18	12	18	13			
088	4	15	16	23	38	14	8	8	8	18	28	18	18	8	15	19	24	12	26	27			
	5	18	9	27	13	12	18	8	8	28	21	29	18	9	15	16	17	13	16	29			
090	18	14	11	11	18	18	14	17	11	8	17	28	9	18	7	14	14	11	17	51	8		
	15	19	29	9	14	24	18	11	21	21	15	18	8	12	14	24	38	9	14	18	15		
092	9	12	13	3	12	9	12	13	28	16	7	26	24	33	21	16	38	13	18	5	18	45	28	3
	9	8	13	5	7	4	18	8	22	15	15	11	5	3	25	4	5	5	14	21	8	18	16	
094	12	5	2	3	18	12	9	5	18	14	18	5	4	5	14	4	15	18	6	8	5	5	2	
	12	18	18	15	14	7	6	18	13	12	2	5	12	19	5	4	8	18	8	7	4			
096	19	12	8	11	13	6	8	6	4	12	9	7	14	25	2	6	4	8	7	1				
	15	7	8	11	14	8	8	12	7	1	16	8	8	9	5	5	4	7	8	2				
098	11	2	4	8	18	4	5	14	18	3	5	8	12	18	3	4	3	7	8	5				
	3	1	2	2	6	1	8	15	6	4	8	7	6	8	4	12	6	9	7	3				
100	2		3	3	5	3	6	8	7	4	2	3	6	9	3	7	18	9	1					
	4	8	3	1	5	2	4	7	8	6	5	6	6	46	27	15	28	11						
102	2	7	3	5	7	2		8	9	5	5	4	7	9	6	8	18	2						
	1	8	4	5	7	7	4	9	2	2	3	3	2	5	3	5	9	4						
104	4	18	4	2	2	4	2	8	8	3	7	4	1	5	5	8	16	5						
	4	9	2	4	2	8	2	11	8	4	7	6	1	5	9	18	17	18						
106	7	18	1	2	4	7	6	8	14	7	3	3	6	11	9	19	19	18						
	18	7	6	6	9	18	28	8	7	8	6	2	6	7	9	27	13	4						
108	7	8	4	5	5	1	15	3	9	3	3	3	9	11	9	13	18							
	18	4	3	7	6	1	15	9	12	6	5	6	14	19	23	17	22							
110	7	9	7	6	6	11	11	8	12	8	3	4	9	18	18	19	28							
	18	17	12	21	11	21	4	3	8	7	1	28	13	14	13	21	18							
112	7	28	19	13	6	3	4	4	13	18	5	14	7	27	38	54	36							
	19	34	29	11	7	4	3	5	13	15	7	18	9	18	58	41	15							
120	7	19	31	11	14	3	2	4	1	33	18	14	16	25	39	13								
	4	12	9	19	7	1	12	12	13	19	8	14	18	37	23	5								
122	4	22	8	52	17	3	14	13	12	24	52	45	43	14										
	128	279	38	156	6	13	11	14	16	19	69	93	3											
124	29	25	113	33	29	15	18	17	12	24	18	2												
	12	16	27	1	29	55	28	18	32															
126																								

Figure 5.1a.—Station occupancies of 1° × 1° quads. (Numbers shown are divided by 10 and rounded.)

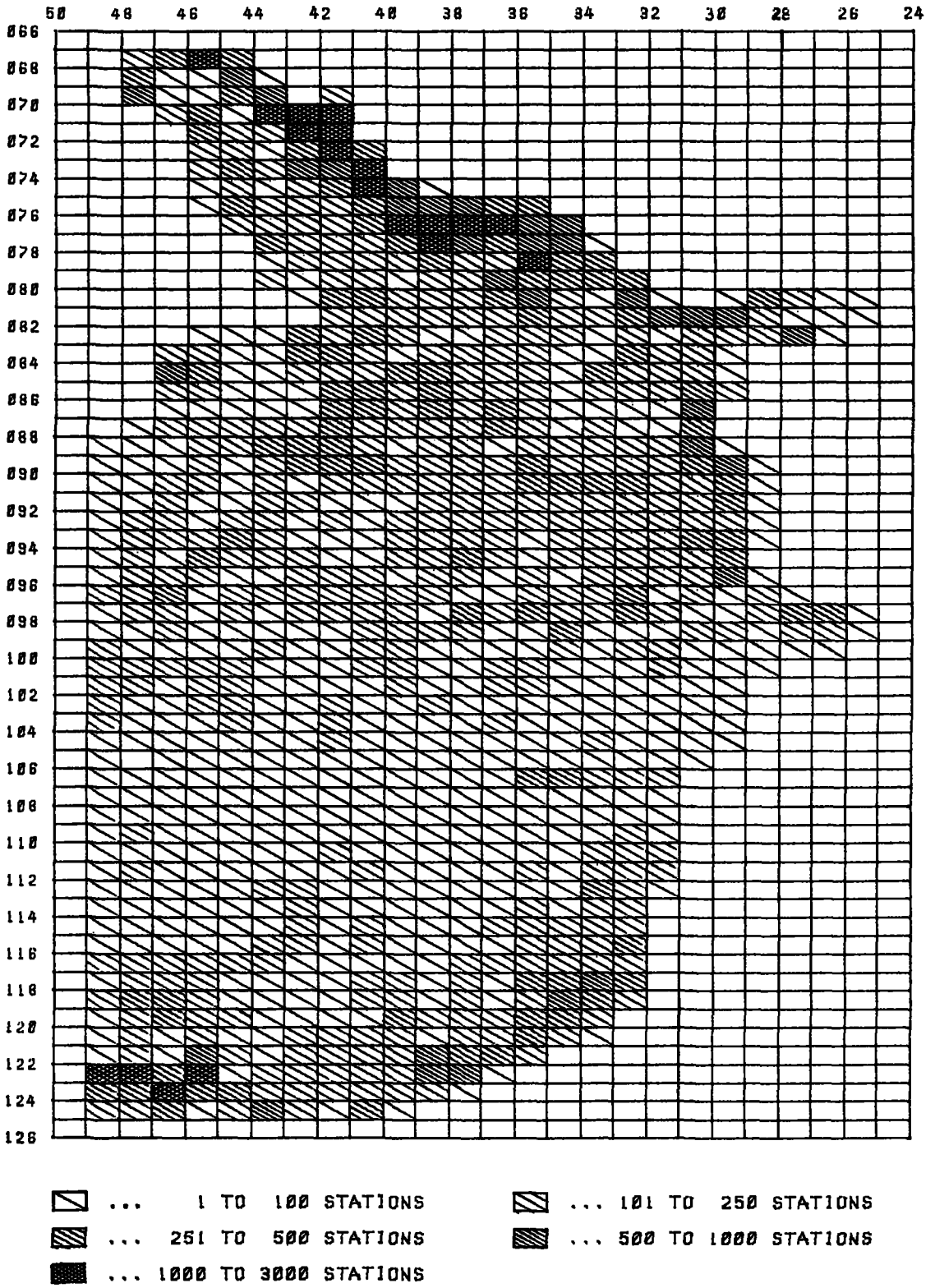

Figure 5.1b.—Station occupancies of 1° × 1° quads.

	50	48	46	44	42	40	38	36	34	32	30	28	26	24
066		447	1977											
068		965	1092	531	178									
070		250	892	4040	3234									
072			893	1154	4696									
074			740	497	2831	2143	866	322						
076			141	620	937	5392	3606	3038	90					
078				469	743	699	1438	2184	949					
080				34	1027	730	871	1795	1295	1383	1296	219	49	
082		138	160	930	845	495	661	468	626	592	389	583		
084		760	484	258	970	1297	843	309	899	750	202			
086		225	488	470	1339	792	1203	451	285	1155				
088	16	253	581	1267	1113	676	704	1062	707	1308	786			
090	39	354	392	557	444	450	612	895	931	769	749			
092	86	643	745	392	331	705	595	601	798	674	397			
094	147	524	612	460	416	658	765	414	612	568	885			
096	240	731	377	638	648	575	676	794	1061	459	465	732	29	
098	206	269	252	346	509	538	252	473	273	453	425	391	2	
100	310	381	536	285	326	345	386	508	210	284	121			
102	253	186	490	215	419	249	340	255	163	252	68			
104	55	65	160	205	368	172	263	224	346	254	31			
106	61	205	102	84	231	219	272	889	533	130				
108	56	258	156	174	164	149	106	176	378	92				
110	110	224	118	228	399	212	159	332	650	259				
112	166	245	245	521	276	183	219	359	637	37				
114	167	227	280	377	412	241	344	699	847					
116	231	677	535	311	276	239	545	892	1274					
118	262	1132	429	110	221	727	543	1402	692					
120	84	503	945	358	503	1018	1191	742	45					
122	1490	4476	2283	551	580	1294	1032							
124	123	430	242	746	416	1								
126														

Figure 5.1c.—Station occupancies of 2° × 2° quads.

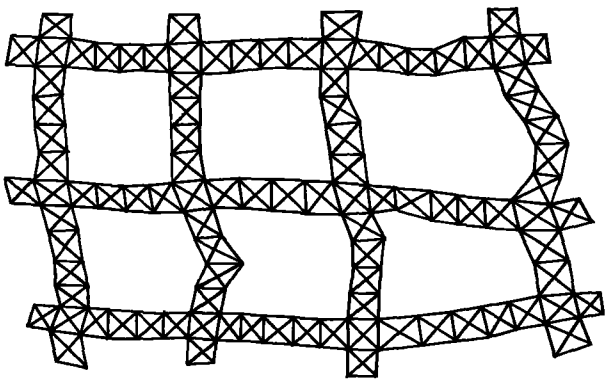

Figure 5.2.—A chain-type network.

This system of chains still prevails in many portions of the country, mainly the Rocky Mountain States of Montana, Wyoming, Colorado, and New Mexico, but also some Midwestern states, such as Iowa, Illinois, and Ohio. In some southern states, such as in Alabama, North Carolina, Texas, Florida, the first-order chains are mostly unsupported by second-order control. In large portions of the United States, not only in densely populated areas, the chains are filled in by second-order areal networks, where the lines of vision range typically from 5 to 20 km. Some metropolitan areas, for example around Washington, D.C., have high precision control networks with very short distances between stations.

A peculiar feature of the U.S. network, which adds much to its global strength, is the transcontinental traverses. These traverses are of the highest precision; their total length is about 22,000 km. Figure 5.3 clearly shows the loops. In about 80 percent of the cases the traverses are composed of basic figures that look like figure 5.4, *i.e.*, they are either narrow diamonds (60 percent of all cases) or narrow triangles (20 percent). In the diamond case the short distances between the closely situated stations are measured with an rms error of about 1 mm. The lengths of the principal distances in the traverses typically range from 7 to 20 km. The shorter lines are mostly found in the eastern region.

Astronomical azimuths are measured along the transcontinental traverses at distances that are about double those of the long lines.

Another important structural property can be called the "invariance of regional redundancy." The ratio of the number of observations in a certain region to the number of stations in that region does not depend strongly on the specific region or on the density of the stations located there. The average number of observations per station is only weakly dependent on station density. The only aspect of local structure of the network that changes with station density is scale; *i.e.*, the average length of the lines of vision. The property of invariance of regional redundancy will not hold in a strict sense. Nevertheless, we feel that it is a better approximation than the assumption that regional redundancy increases in proportion to station density.

5.2 Estimating the Inverse of the Normal Equation Matrix

From our discussions in sections 4.1.3 and 4.1.4, it became clear that we need estimates on the elements f_{ij} of the inverse of the normal equation matrix. Although only a single bound $\|f\|$ on f_{ij} was used for the preliminary estimates in 4.1.4, we will need more detailed information to arrive at refined estimates. These refined estimates will rely on the equations listed at the end of section 4.1.3. Of course, what we can do is limited by network information. On the other hand, even an error of 100 percent should not be considered too serious. It would only mean that our roundoff estimates are off by a factor of two.

We will mainly be interested in the global features of the inverse. What we can expect locally is clear from the local adjustments that are continuously performed at NGS. The desired global information will be obtained by setting up a simplified model of the network. This model will have many fewer parameters than the 400,000 unknowns of the real network. However, the parameters of the simplified model will be capable of bringing out the random global distortions of the network. The covariance matrix of the substitute parameters will be obtained by computer simulation. After presenting the numerical results, we will discuss some supporting theoretical and practical evidence for their validity.

5.2.1 Description of the model

We will assume that we are using a plane net. Any sophistication based on a curved reference surface would be out of place in the present context. Imagine a plane network, which we call the "real" one, superimposed by an artificial grid-type network, as shown in figure 5.5.

The meshes of the artificial grid network are relatively large when compared to the station spacing of the real network. For the U.S. network we imagine a grid network built of $1° \times 1°$ quads (in plane projection). Later we will even switch to $2° \times 2°$ quads. The coordinates of the grid points will be the substitute parameters and their covariance matrix will give us a

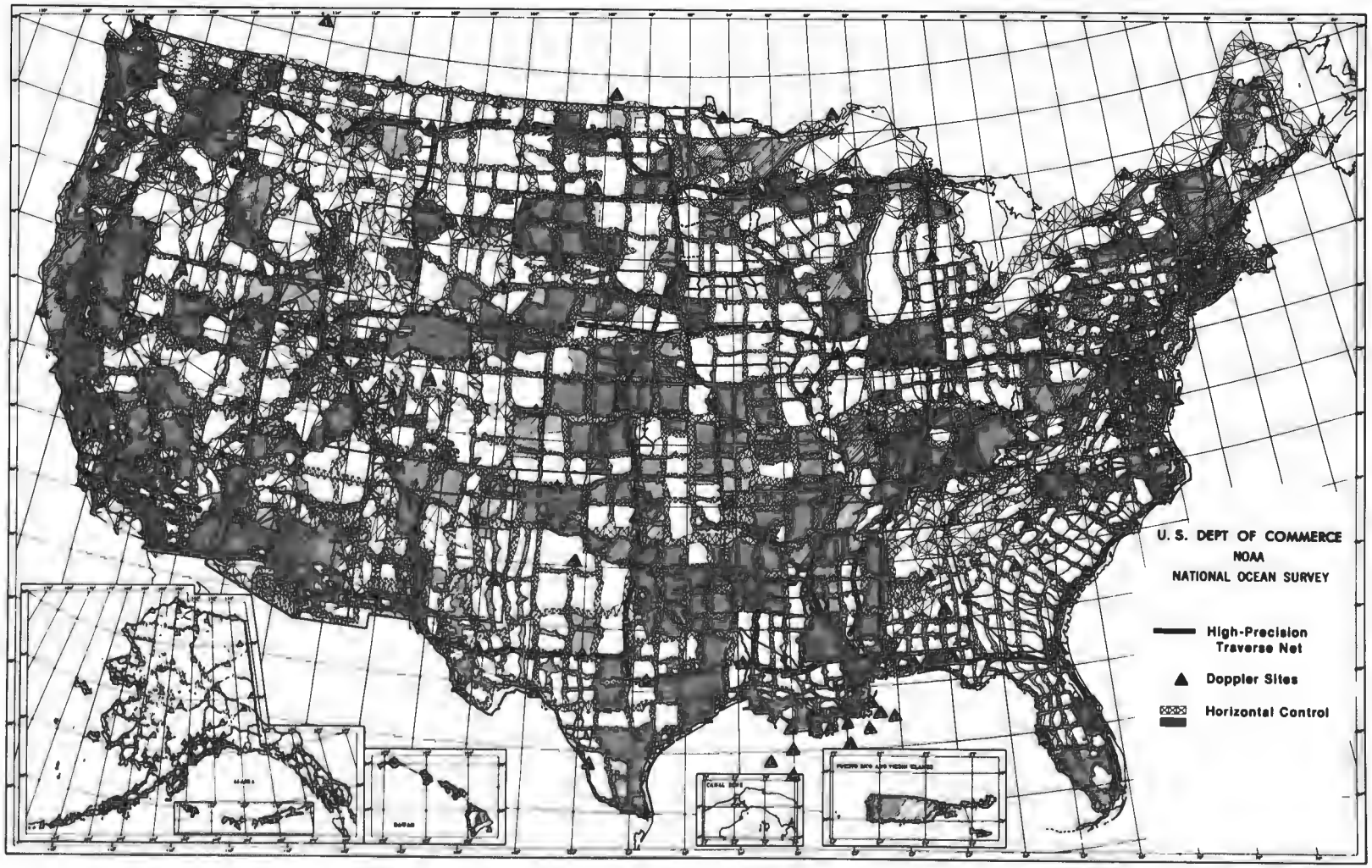

Figure 5.3.—The U.S. national geodetic network.

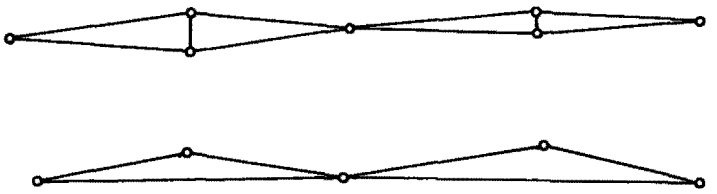

Figure 5.4.—Basis figures of transcontinental traverses.

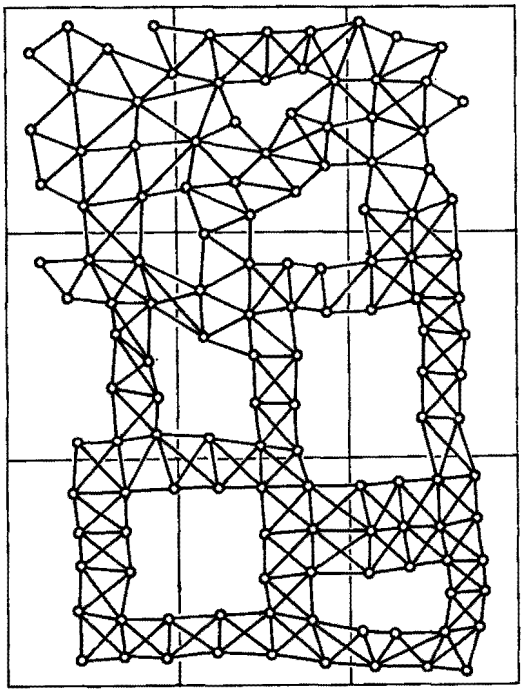

Figure 5.5.—Sample network with superimposed finite element structure.

global picture of the random distortions of the real network.

Our first problem will be to establish a functional relation between the original coordinates, which we denote by the vector p, and the substitute coordinates, which we denote by the vector q. Since there are fewer artificial coordinates than real ones, we cannot expect a one-to-one mapping between these two sets of parameters. We start by defining a mapping from q to p. For this purpose we concentrate on a single quad of the grid, as shown in figure 5.6a. Only real coordinates of stations situated in the interior or on the boundary are considered as being dependent on the four corners of the quad. The corners are numbered 1,2,3,4, as shown. If the quad is undistorted (as in fig. 5.6a), the real network nodes in its interior will have certain positions. One may assume, for definiteness, that these positions correspond to the approximate positions of the stations. Imagine now a small distortion of the quad. Figure 5.6b shows such a distortion in a grossly exaggerated way. The configuration of the interior nodes will also suffer distortion. We must define the shifts of the interior nodes in a meaningful way.

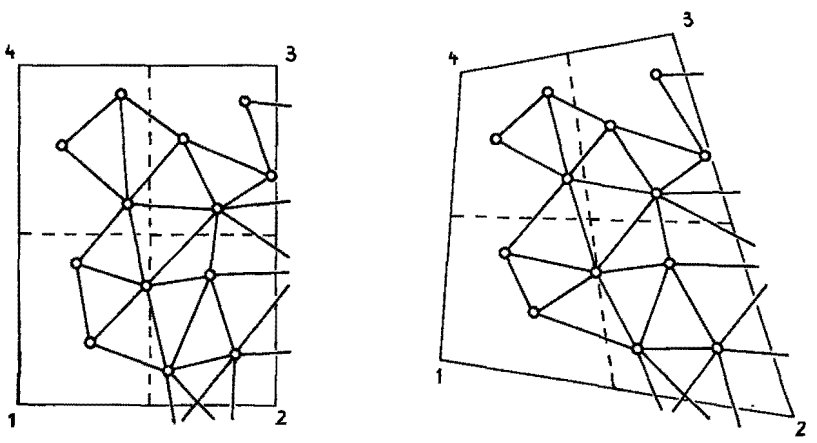

Figure 5.6.—Quad of (a) undisturbed and (b) disturbed finite element grid.

Essentially our task is to perform an interpolation. The shifts of the four corner nodes of the quad must be interpolated so that we can obtain shifts for the interior nodes that depend uniquely on those of the corner nodes.

The interpolating functions must fulfill certain requirements. They must be continuous in the interior of the quads. But an important additional requirement is consistency between the interpolating formulas of adjacent quads which share a common boundary segment. A station situated on a boundary segment must experience equal shifts if it is to be viewed as part of an adjacent quad. The interpolating functions for one quad may be called "local" functions. All local interpolating functions must combine to a global interpolating function that is continuous everywhere.

Within a certain quad we will use a bilinear interpolation formula which is one of the simplest interpolation formulas used in finite element applications (Zienkiewics 1971, sec. 7.3).

It will be necessary to switch temporarily to a more elaborate notation. Let x_j, y_j denote the coordinates of one of the real network stations situated in the interior of the quad. Let u_k, v_k, $k=1,2,3,4$ denote the coordinates of the four corners of the quad. It suffices to derive the interpolating formula for one coordinate, *i.e.*, for x as a function of the u's. The interpolating formula which gives y in terms of the v's will be formally identical.

Assume a local coordinate system with an origin at the midpoint of the quad. (See fig. 5.7.) Assume that $\pm a$, $\pm b$ are the local coordinates of the four corners. Denote by ξ,η the local coordinates of a certain interior station. Let Δu_1, Δu_2, Δu_3, Δu_4 be the shifts of the corner points. Then Δx will be the shift of the station with local coordinates ξ,η. (See fig. 5.7.) We establish an interpolation formula of the type

$$\Delta x = A + B\xi + C\eta + D\xi\eta. \tag{5.1}$$

The four coefficients A, B, C, D must be chosen

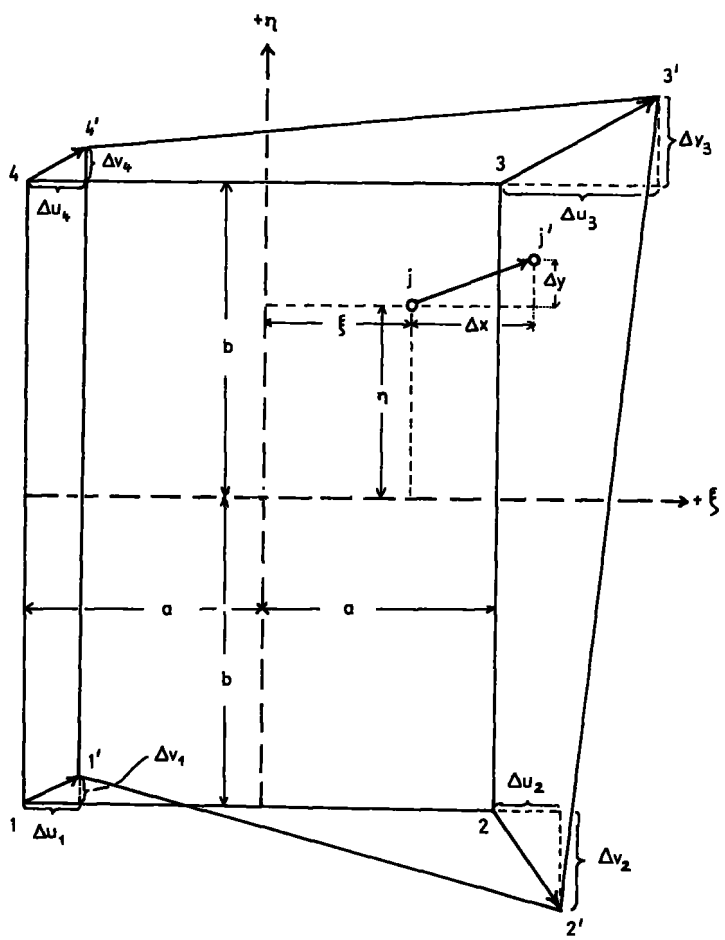

Figure 5.7.—Choice of local coordinate system in a finite element quad.

in such a way that for $(\xi,\eta) = (\pm a, \pm b)$ the shifts $\Delta u_1, \ldots, \Delta u_4$ of the four corner points result. This leads us to the following linear system:

$$\begin{bmatrix} +1 & -a & -b & +ab \\ +1 & +a & -b & -ab \\ +1 & +a & +b & +ab \\ +1 & -a & +b & -ab \end{bmatrix} \begin{bmatrix} A \\ B \\ C \\ D \end{bmatrix} = \begin{bmatrix} \Delta u_1 \\ \Delta u_2 \\ \Delta u_3 \\ \Delta u_4 \end{bmatrix}. \quad (5.2)$$

Up to the factors 2, $2a$, $2b$, $2ab$, by which the columns are multiplied, the matrix is orthogonal. Therefore inversion is easily done and yields

$$\begin{bmatrix} A \\ B \\ C \\ D \end{bmatrix} = \frac{1}{4} \underbrace{\begin{bmatrix} +1 & +1 & +1 & +1 \\ -1/a & +1/a & +1/a & -1/a \\ -1/b & -1/b & +1/b & +1/b \\ +1/(ab) & -1/(ab) & +1/(ab) & -1/(ab) \end{bmatrix}}_{S} \begin{bmatrix} \Delta u_1 \\ \Delta u_2 \\ \Delta u_3 \\ \Delta u_4 \end{bmatrix} \quad (5.3)$$

If we denote by S the matrix occurring in this formula as shown, we can write the interpolation formula for the shift Δx of an interior station with local coordinates ξ,η as

$$\Delta x = (1, \xi, \eta, \xi\eta)\, S(\Delta u_1, \Delta u_2, \Delta u_3, \Delta u_4)^T. \quad (5.4)$$

If we also want the station number j to appear in the formula, we must write

$$\Delta x_j = (1, \xi_j, \eta_j, \xi_j, \xi_j\eta_j)\, S(\Delta u_1, \Delta u_2, \Delta u_3, \Delta u_4)^T. \quad (5.5)$$

The completely analogous formula for the y coordinates is:

$$\Delta y_j = (1, \xi_j, \eta_j, \xi_j\eta_j)\, S(\Delta v_1, \Delta v_2, \Delta v_3, \Delta v_4)^T. \quad (5.6)$$

These interpolation formulas obviously have continuity in the interior of the quad. Continuity of the global interpolation formula across the quad boundary is also easily proved, either by explicit calculation or by noting that interpolation along a quad boundary is linear and, therefore, completely determined by the two corner points situated at the boundary segment.

We have now established a mapping from the quad corners to the real stations that has all the desired properties. We will now switch back to the previous notation of p, q for the vectors of the station coordinates and quad corner coordinates, respectively. We denote the increments to these coordinates by Δp, Δq. In this notation our mapping may be written symbolically as:

$$\Delta p = R\,\Delta q. \quad (5.7)$$

The matrix R appearing in this formula is sparse because the shift of a certain station depends only on the shifts of the four corners of the surrounding quad.

The next procedure is now fairly simple. Consider the linearized functional relation between corrected observations and adjusted coordinates. In section 3.3, eq. (3.17) was written

$$\Delta l + v = B\,\Delta p. \quad (5.8)$$

We now simply replace Δp by $R\,\Delta q$ and obtain

$$\Delta l + v = BR\,\Delta q = C\,\Delta q. \quad (5.9)$$

This is viewed as a new adjustment problem. The normal equations are

$$(C^T P C)\,\Delta q = C^T P\,\Delta l. \quad (5.10)$$

The solution is

$$\Delta q = (C^T P C)^{-1} C^T P\,\Delta l. \quad (5.11)$$

The covariance of the adjusted Δq is given by

$$Cov(\Delta q) = (C^T P C)^{-1}. \quad (5.12)$$

This matrix will serve us in judging the global accuracy of the network.

Remark: The transition from Δp to Δq deserves further comment. The reader may possibly miss the formula that expresses Δq in terms of Δp. We have derived Δq (eq. 5.11) as a function of Δl. In order to replace Δl by Δp, we can consider observation increments of Δl that are consistent with Δp:

$$\Delta l = B\,\Delta p. \quad (5.13)$$

We then have

$$\Delta q = (C^T P C)^{-1} C^T P B\,\Delta p. \quad (5.14)$$

Note that the matrix in front of Δp is a left inverse of R in the equation $\Delta p = R\,\Delta q$.

5.2.2 Further simplifications

The model described in subsection 5.2.1 requires a knowledge of location, type, and weight of all observations. Since this knowledge is not available, further hypothetical assumptions must be made. Information made available to me by the NGS consisted of computer-readable files which gave the number of

triangulation stations per 1°×1° quad. It also specified how this number splits into primary triangulation stations and supplemental stations of various kinds. Primary stations are first and second order triangulation stations that essentially build up the network. Supplemental stations contribute little to its overall strength. The network then is further stiffened by the transcontinental traverses. A file that listed all the stations of the transcontinental traverses, together with the latitudes and the longitudes of these stations, was also available. Another file listed those stations of the transcontinental traverses in which astronomical azimuths were taken. Finally, there was a file listing the 132 Doppler stations together with their latitudes and longitudes.

Based on the station information above, and guided by discussions with NGS staff members, notably J. G. Gergen, I have assumed a certain pattern of observations outlined in the following sections:

5.2.2.1 Assumed angular observations

Although bundles of unoriented directions are measured throughout the U.S. network, I have assumed measured angles instead, because this made the computer programs somewhat simpler. Our model will be so rough anyway that this deviation from the truth does not matter. Assume, for the moment, that primary stations in a quad are arranged in a regular pattern such that the quad decomposes into the 16 subquads shown in figure 5.8.

The size of the 1°×1° quads, by the way, is assumed to be uniform over the United States and with the following dimension: 2a × 2b with $b = 6380*\pi/360$ km, $a = b * cos(40°)$. We assume 128 angles were measured as indicated in figure 5.8.

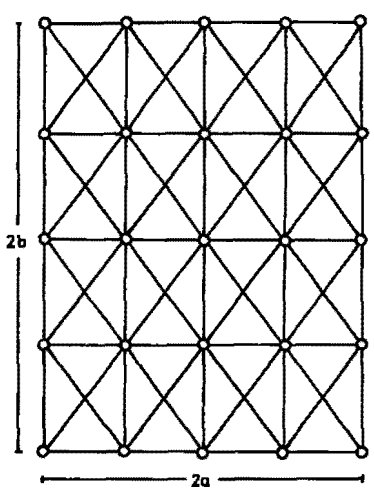

Figure 5.8.—Assumed regular pattern of primary stations in a 1° × 1° finite element quad.

There are eight angles per subquad and also eight angles per interior station. The weight of any angle is based on an assumed rms error of 1" (arc second). The contribution of the 128 angles toward the normal equation matrix C^TPC, which is the normal equation matrix for the coordinates of the 1°×1° quad corners, can be precalculated once and for all. This contribution will only involve coordinate shifts of the four corners of the quad under consideration.

The problem now is that the density of the primary stations is not uniform over all the quads. We take this into account in a rather crude way. Our precalculated contribution to the normals is based on an assumed density of 16 primary stations per quad. (Note that the stations at the boundary contribute to more than one quad). If we encounter a quad with n_q stations, we simply multiply the precalculated contribution by the weight factor $n_q/16$ before we add this contribution to the normals. An essential assumption in this approximation is that the accuracy of an angle is not dependent on the length of the lines of vision. This assumption is reasonable as long as the lines of vision are not too short, say below 2 km. However, this may safely be assumed for the primary stations.

5.2.2.2 Assumed base lines other than those in the transcontinental traverses

There are about 20,000 to 30,000 measured distances in the U.S. network. It is estimated that only a fraction of them, about 5,000, contribute significantly to the strength of the network. About 4,000 are found in the transcontinental traverses. The remaining 1,000, may be base lines that strengthen otherwise purely directional portions of the network. The other distances, about 15,000 to 20,000, are found mostly in traverses and in short connections between clustered stations, and, therefore, do not add much to the strength of the network.

In the simulation study I have assumed that there is one base line of length 10 km at the center of any quad that has more than 15 primary stations. Since the northing of this base line is uncertain, I have actually assumed two base lines, one in an east-west direction, the other in a north-south direction. The weight of the original base line, which was assumed on the basis of an rms error of 1 cm, was divided equally between the two artificial lines. Altogether, base lines were assumed in this manner for 809 1° × 1° quads.

5.2.2.3 The observations of the transcontinental traverses

Stations of the transcontinental traverses that are situated close together (fig. 5.4) have been lumped

together into one station. Having done this we can assume that there are two distances for each station. We assume an average length of 10 km for one distance and an rms. error 1 cm. Again the two distances are assumed to be east-west and north-south The weight is 1/0.01 m^{-2} per distance. This was not divided because there are actually two distances.

The stations of the transcontinental traverses in which an azimuth was observed are explicitly specified on a file. Hence these azimuths could be easily taken into account. Because the target point for an azimuth is unknown, I have assumed two azimuths, one in an east-west direction and the other in a north-south direction. I have divided up the weight between them. The weight is based on an assumed rms error of 1″ and an assumed length for the line of vision of 10 km.

5.2.2.4 The Doppler positional observations

The network has 132 Doppler stations, as shown in figure 5.3. A Doppler observation is considered equivalent to the direct observation of the two coordinates of a station. Although the NGS will assume smaller rms errors, I have used 1 m as the rms error of an observed coordinate.

5.2.3 Remarks on the computer programs for the simulation study

The computer programs used to obtain the numerical results are not the ultimate in speed and storage utilization, since I designed, coded, key-punched, and debugged them in a short period of time. Their main purpose was to obtain quick results. When it became clear that a solution for the entire network based on 1° × 1° quads was not possible in one sweep, I decided to enlarge the quads to 2° × 2° rather than lose time segmenting the programs. The procedure is best described as follows:

Imagine the normals formed by using parameters that are the coordinates of the corners of the 1° × 1° quads. The normals are precisely formed as outlined in section 5.2.2. After having the 1° × 1° quads available, a transition is made to a smaller set of parameters, i.e., to the coordinates of the corners of the 2° × 2° quads. This parameter condensation is carried out in the same spirit as the original transformation of the station coordinates to the coordinates of the 1° × 1° quad corners. The same interpolation formula specified in section 5.2.1 is used. The role of the interior stations in this section is taken over by the 1° × 1° quad corners.

In the computer the transformation to the 2° × 2° quads is not actually carried out subsequent to the formation of the 1° × 1° normals because no storage saving would be obtained. Rather, the transformation is continuously carried out during the process of assembling the normals, and only coefficients pertaining to the 2° × 2° normals are accumulated in central memory.

Having two versions of the program, one for 1° × 1° quads, the other for 2° × 2° quads, offers an opportunity of checking the model's consistency by comparing solutions obtained by means of both versions. Of course, such comparisons could be made only for a portion of the whole network. The results of such comparison, which covered the area west of the Mississippi were quite satisfactory as indicated in the following subsection.

5.2.4 Results of the simulation study

The rms point errors for the corners of the 2° × 2° quads are shown in figure 5.9. The corners have odd latitudes and longitudes. Point errors are specified in centimeters. We see that in Maine the point errors go up to nearly 50 cm. There are also larger values at the southern tip of Florida and in the northeastern corner of Washington State. Otherwise the following pattern prevails. Typical values are about 25 cm at the boundaries and 15 cm in the interior. In areas of largest station density, close to the east and west coast, these values are about 20 cm.

Sample covariances among coordinates are exhibited in figures 5.10a-e. Figure 5.10a lists the covariances of the quad corner with latitude 39° and longitude 77° with all other quad corners. The four values referring to (ϕ,ϕ), (ϕ,λ), (λ,ϕ), (λ,λ) are condensed to one number shown in the appropriate cell of figure 5.10a. This number is the average over the absolute values referring to the four coordinates. The values are listed in units of the fourth digit after the decimal point; the covariances are scaled to a meter squared.

Figures 5.10*b*,c give a pictorial representation of the covariances for the same "base quad" as in figure 5.10a, *i.e.*, $\phi = 39°$, $\lambda = 77°$. Figure 5.10b refers to the entries for (ϕ,ϕ), (λ,ϕ); figure 5.10c refers to the entries for (ϕ,λ), (λ,λ). Each figure illustrates a column of the inverse F. In other words, it gives the solution of the normals when the right-hand side is zero except for a 1 at one of the coordinates at $\phi = 39°$, $\lambda = 77°$. Note that the heavy lines in figures 5.10b-e connect points of uneven latitudes and longitudes, i.e., they connect the centers of the quads shown in the other figures. This explains the spur at the southern tip of Florida. Figures 5.10d,e refer to a different base quad, $\phi = 47°$, $\lambda = 69°$. These figures demonstrate the weakness of the network in the area of Maine.

Note that figures 5.10a-e describe only the global features of the covariance matrix and that local peaks

	50	48	46	44	42	40	38	36	34	32	30	28	26	24
066		48	40											
068		41	33	29										
070		39	29	25	25									
072			27	23	21									
074			26	22	20	20								
076				21	19	18	18	19						
078				21	18	17	17	18	19					
080				20	17	17	17	17	17	19	21	25	57	
082			21	18	16	16	16	16	17	18	20	26		
084		23	19	17	16	15	15	16	16	17	19			
086		21	18	16	15	15	15	15	15	16				
088		20	17	15	15	15	14	14	15	16	18			
090		18	16	15	15	14	14	14	15	15	18			
092	22	17	16	15	14	14	14	14	14	15	17			
094	20	17	16	15	14	14	14	14	14	15	17			
096	20	17	15	15	14	14	14	14	14	15	17	22		
098	19	17	15	15	14	14	14	14	15	15	17	22		
100	19	17	15	15	14	14	14	14	15	16	19			
102	19	17	15	15	14	14	14	15	15	16	21			
104	21	18	16	15	15	15	15	15	15	17				
106	21	18	16	16	15	15	15	15	16					
108	21	18	17	16	16	15	16	16	16	19				
110	21	18	17	16	16	16	16	16	17	20				
112	21	19	18	16	16	16	16	16	17	20				
114	22	19	18	17	17	17	17	17	18					
116	22	19	18	18	17	17	17	18	19					
118	23	20	19	19	18	18	18	20	27					
120	25	21	20	20	19	19	19	20	24					
122	26	23	22	21	21	21	21							
124	31	27	25	24	24	25								
126														

Figure 5.9.—Rms point errors. Values refer to the global covariance and are listed in centimeters.

	50	48	46	44	42	40	38	36	34	32	30	28	26	24
066		101	102											
068		101	102	102	102									
070		102	102	101	100									
072			102	099	097									
074			101	097	094	092	043	099						
076			100	095	090	086	086	091	095					
078				093	084	080	080	084	087					
080				088	079	073	073	077	080	081	082	082	079	
082		083	081	078	073	066	067	071	075	076	078	079		
084		000	075	072	067	061	061	065	069	071	075			
086		073	070	067	062	057	056	060	063	067				
088	066	067	065	061	057	053	052	056	059	062	065			
090	063	062	060	057	053	049	048	052	055	058	061			
092	060	058	055	053	049	045	045	048	051	054	056			
094	056	054	052	049	046	042	042	044	047	050	053			
096	053	051	048	045	042	039	039	041	044	047	049	051	053	
098	050	047	045	042	039	036	036	038	039	044	046	049	051	
100	046	044	042	039	036	033	033	036	038	040	043			
102	043	041	039	035	033	030	046	033	035	037	040			
104	040	038	036	033	031	028	028	030	032	035	037			
106	037	035	033	030	028	025	176	028	030	032				
108	034	032	030	028	025	023	023	025	028	030				
110	031	029	027	025	023	021	021	023	025	028				
112	029	027	025	023	021	018	019	021	023	026				
114	029	026	024	022	020	017	018	021	023					
116	029	027	025	023	021	019	020	022	025					
118	030	028	026	024	022	020	021	024	026					
120	031	029	027	026	024	022	023	026	028					
122	032	031	029	027	026	024	025							
124	034	032	031	029	028	026								
126														

Figure 5.10a.—Condensed covariance values for the base quad, $\phi = 39°$, $\lambda = 77°$. Values shown are in units of the fourth decimal place.

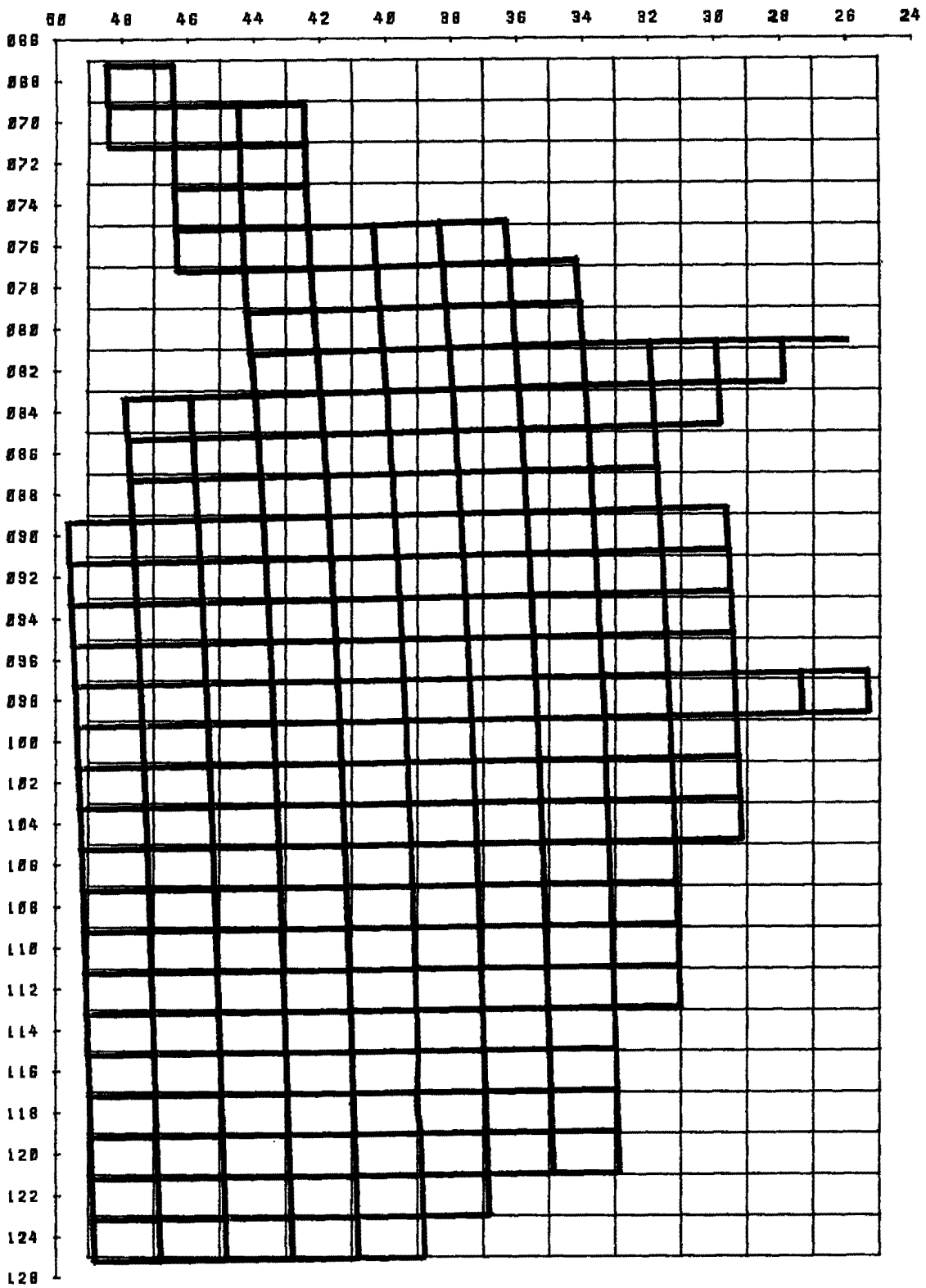

Figure 5.10b.—Pictorial representation of global covariance. Network response to latitude disturbance at $\phi = 39°$, $\lambda = 77°$.

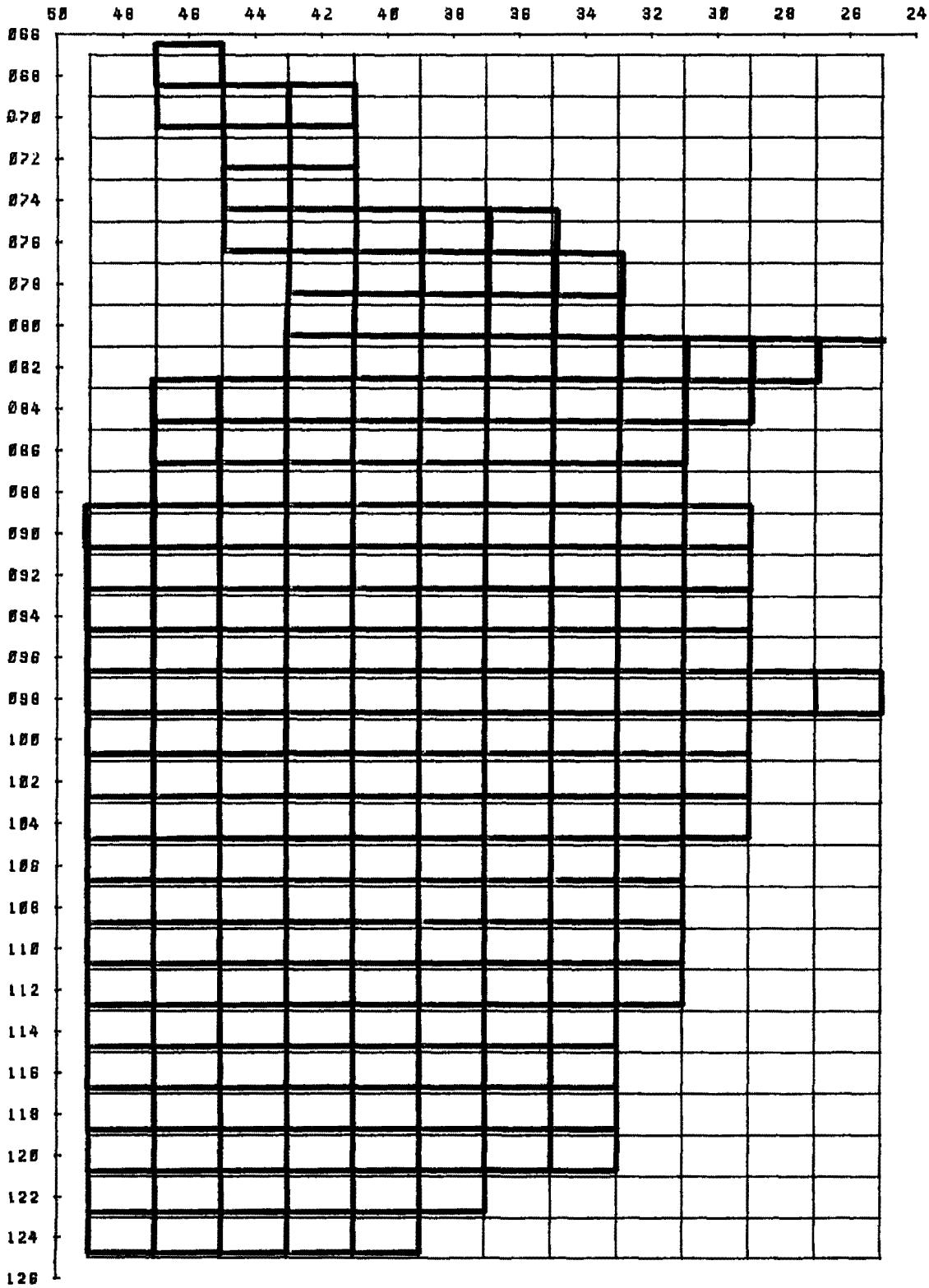

Figure 5.10c.—Pictorial representation of global covariance. Network response to longitude disturbance at $\phi = 39°$, $\lambda = 77°$.

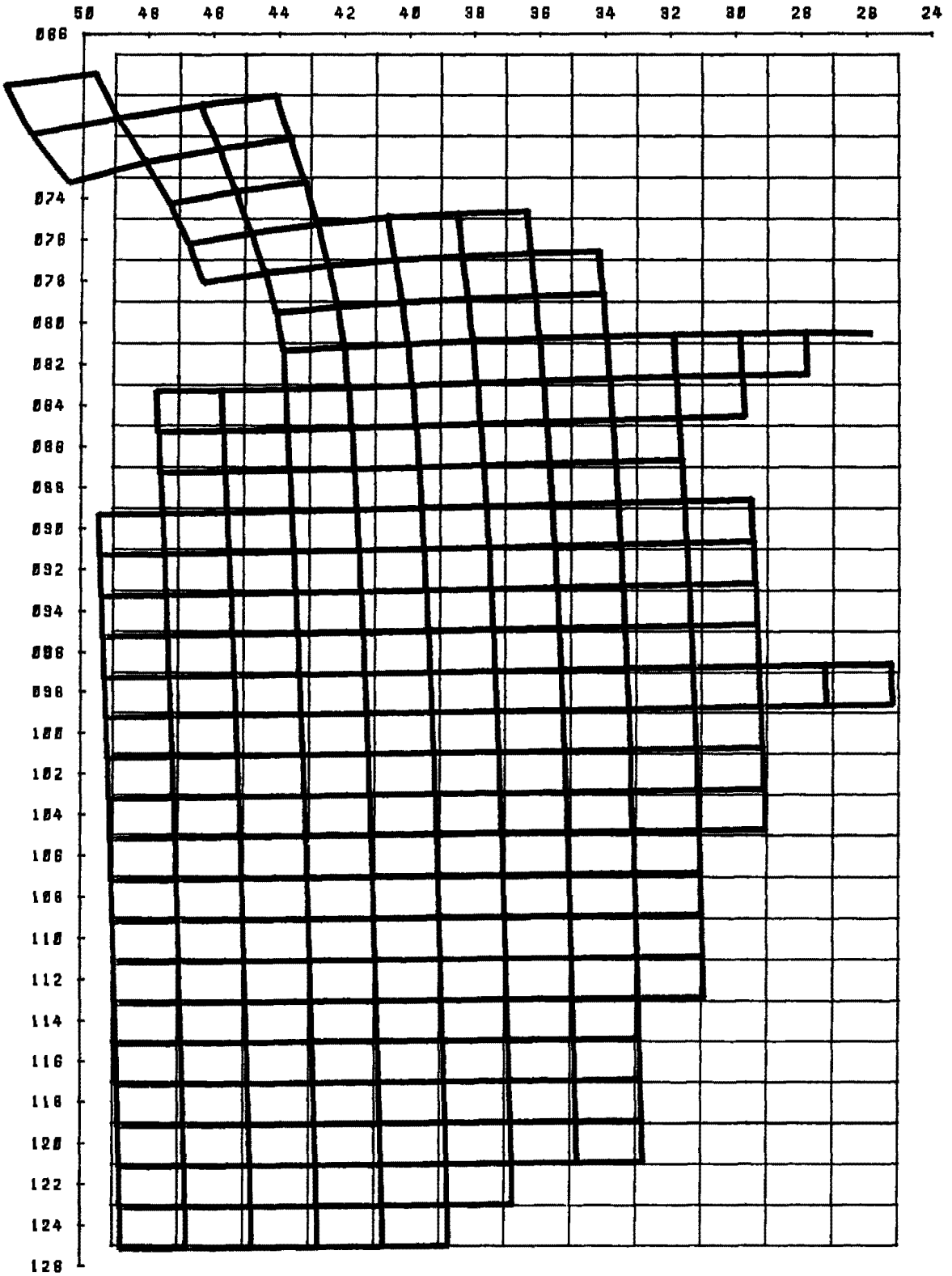

Figure 5.10d.—Pictorial representation of global covariance. Network response to latitude disturbance at $\phi = 47°$, $\lambda = 69°$.

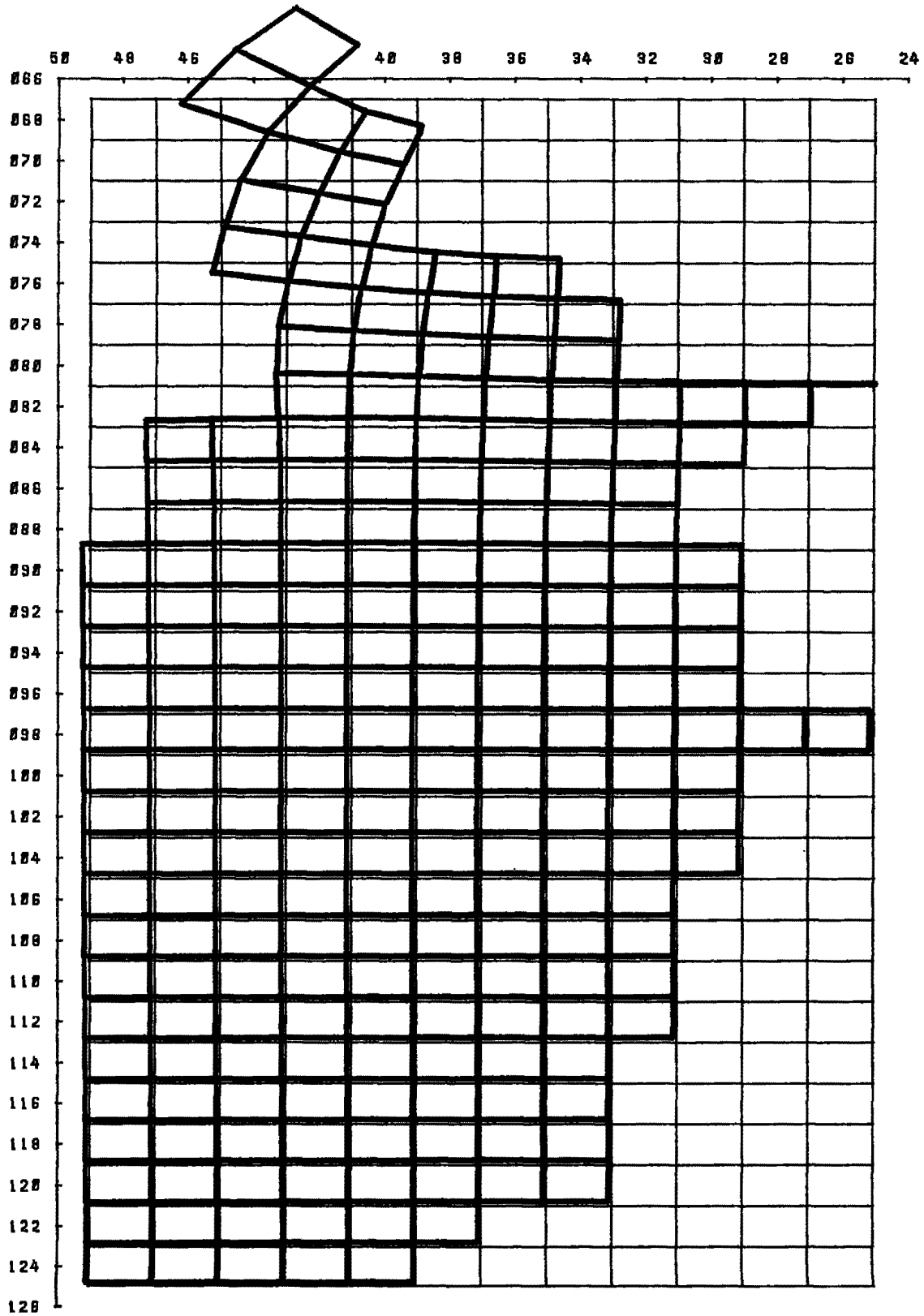

Figure 5.10e.—Pictorial representation of global covariance. Network response to longitude disturbance at $\phi = 47°$, $\lambda = 69°$.

must be imagined as being superimposed in the area of the base quads.

The results of the simulation study must be properly interpreted. What we get is a smoothed version of the covariance matrix. Imagine that a localized covariance matrix is superimposed, as mentioned above. The superimposed localized covariance matrix has entries near zero if the two stations involved are separated by, say, 300 km. The point errors of 16 cm which we obtain are not representative for the point errors of single stations. Instead they reflect the accuracy of regional means taken over the coordinates of a number of stations. To estimate the locally superimposed errors, we have to look at regional adjustments, and we can also draw some conclusions from studying large regular networks. This will be done in section 5.3.

Covariances obtained from our simulation study are considered to be quite representative, provided that the spacing between stations is more than 2 to 3 quad diameters. From various calculations, similar to the one documented in figures 5.10a-e, we see that covariances between latitudes and between longitudes hardly exceed 0.025 m². They come near to these values if both stations are close to a boundary. For central stations, covariances are mostly below 0.01 m². In any case, about 50 percent of all covariances of a certain coordinate with all others may be assumed below 0.01 m². Cross-covariances between latitudes and longitudes are generally smaller by a factor of about 0.1.

For the portion of the network west of the Mississippi, results for the 1° × 1° subdivision, as well as for the 2° × 2° subdivision, were obtained. The choice of weights and other parameters was somewhat different from that described in section 5.2.2. We therefore refrain from exhibiting numbers and simply mention that sample variances and covariances agreed within 5 to 10 percent.

5.3 Estimating the Local Features of the Covariance Matrix

Our model described in the previous section assigns only eight degrees of freedom to a 2° × 2° quad. Only the eight coordinates of the quad corners are allowed to vary. In a 2° × 2° quad there may be as many as 2 × 5,400 = 10,800 station coordinates whose degrees of freedom are almost completely suppressed. It is our task to estimate the local variations of the station coordinates within a 2° × 2° quad and also, to be sure, a little beyond the boundaries of such a quad. As mentioned earlier, the global covariance coming out of the finite element model must be superimposed by a local covariance that is mostly concentrated in 2° × 2° quad regions and is practically zero for distances exceeding, say, 300 km. To state the problem slightly differently, we are interested in those portions of relative errors for stations spaced less than 300 km apart that cannot be explained by global covariance.

Let us begin with the observation that the local covariance may have an appreciable peak. There may always be rather weak portions of the network comprising supplementary stations. However, the extension of such weak portions will not be large. Hence such a peak will always be quite narrow and after a few 10's of kilometers will have tapered off completely. Let us assume that such local peaks amount to (30 cm)², and that they taper off after 30 km.

Remark: Large station variances cause no concern if a station with a large variance is not involved in a measurement with an accuracy much higher than its coordinate errors indicate. Such a station will have small coefficients in rows and columns of the normals belonging to it. If j refers to a coordinate of such a station, then some of the f_{ij}, f_{ji}, $j=1, \ldots, n$ may be large, but a_{jk}, a_{kj}, $k=1, \ldots, n$ in the normals will be small. Eq. (4.33) will have all large f_{ij}, f_{ji} counteracted by small ε's. A similar argument applies to the right side errors η. However, the argument breaks down if there is strong coupling between a group of stations having large point errors.

Consider now the framework of the primary stations that give the network its strength. We claim, as supported with evidence below, that the individual coordinate errors of such stations will not appreciably exceed the error of the coordinates of the 2° × 2° quad corners. In areas of fill-in networks, the relative error of a station with respect to its neighboring stations approximates closely the surplus of individual errors over smoothed-out errors. This mean-square "neighborhood" error may be around (5 cm)² to (10 cm)². In areas where the triangulation chains have not been filled in, neighborhood errors may be larger, in particular between stations on different arcs. We will allow (15 cm)² to (20 cm)² in such situations.

5.4 Supporting Evidence for Estimates of Global and Local Covariances from Test Calculations

In an internal report, Vincenty (1975) describes a simultaneous adjustment of the southeastern loops of the transcontinental traverses. There are two large loops and a small one (fig. 5.3). MEADE'S RANCH is on one of the larger loops; the other large one has an arc along the east coast. The small loop includes mostly portions of Louisiana. The total number of traverse stations was 949, with 5,227 observations. The first adjustment was done with MEADE'S RANCH held fixed. A second adjustment included

19 Doppler stations along the traverses which provided the position information. By means of diagrams, Vincenty shows that adjusted latitude and longitude errors do not exceed 60 cm if Doppler stations are ignored and 31 cm if Doppler stations are included. Our finite element model, which also includes Doppler stations, predicts latitude errors of about 15 cm along the southern portion of the east coast. Locally superimposing 10 cm gives $\sqrt{15^2 + 10^2} = 18$ cm. Taking into account that Vincenty assumed smaller errors than we did for the Doppler stations (0.60 to 0.75 m vs. 1 m), and that 19 Doppler stations provide a much poorer fix than 130 Doppler stations, and finally, if the effect of fill-in with its vast number of observations is considered, there is no reason to reject the plausibility of our results.

Moose and Henriksen (1976) undertook another study of great relevance to our investigations. They performed a sequence of adjustments of a moderately large network covering portions of Mississippi, Louisiana, and Alabama. Their work is based on an earlier investigation by Dracup (1975). The primary purpose of these studies was to investigate the improvement of a network by including Doppler observations. The 1,330-network stations split into about 840 first-order stations that form a chain-type network of the kind depicted in figure 5.2, and about 490 second-order stations that provide fill-in in some portions but not in all. Doppler positions were observed at five stations. The network was strengthened by 63 distances and 22 azimuths. The first-order stations had only 18 distances. The paper by Moose and Henriksen (1976) gives partial results for 16 different adjustments classified by: inclusion or exclusion of second-order stations, inclusion or exclusion of all or some of the Doppler stations, or inclusion or exclusion of some or all of the distances. We will concentrate on only two of those adjustments. The first one, labelled B^+ in the paper, concerns only the first-order stations. The second one, labelled G^*, comprises all stations. Both adjustments use information on distances (15 vs. 60), azimuths (18 vs. 22), and Doppler stations (five in both cases).

Adjustment B^+ gives us a good opportunity to estimate the accuracy between stations on different arcs of a chain-type network. Because the network under consideration is only a subnetwork of the entire U.S. network, the accuracy will be poorer. We will overestimate the errors. If we restrict attention to the innermost arcs, reasonably close overestimates will be obtained. The test lines labelled 16.1, 16.2, 17.1, 17.2 by Moose and Henriksen (1976) span neighboring arcs. These four lines are at the center of the test network. The network boundary is two or more arc loops away. The standard errors of these lines are between 14.5 and 16.2 cm.

For adjustment G^*, only error ellipses are specified. The error ellipses are fairly uniform with semimajor axes near 57 cm and semiminor axes near 52 cm. The explanation of this phenomenon is simple. For the Doppler measurements Moose and Henriksen assumed rms errors of 0.9 m in latitude and 1.2 m in longitude. This accuracy is inferior to the relative accuracy of the network stations owing to the traditional observations of directions and distances. Hence it can be expected that the accuracy of the position fix of the network can be estimated very well by assuming that the network is rigid. Since there are five Doppler stations, this would result in $0.9/\sqrt{5}$ m $= 40$ cm latitude errors and $1.2/\sqrt{5}$ m $= 54$ cm longitude errors. The error ellipses should all be oriented in an west-east direction. This is truly the case, as shown in figure 12 of the Moose and Henriksen report. The fact that we got slightly smaller values, namely 40 cm vs. 52 and 54 cm vs. 57 accounts for the failure of the network to be fully rigid. (See also the discussion in appendix 1, Moose and Henriksen (1976).)

In our simulation study, we have assumed an error of 1 m for Doppler-observed latitudes and longitudes. Since we have about 130 Doppler stations, we would obtain $1/\sqrt{130}$ m $= 9$ cm positional errors if the network were fully rigid. Our position errors are larger. They are about 10 cm in the central areas and about 14 to 18 cm in the coastal areas. (See fig. 5.9 and recall that this figure shows point errors, i.e., the superposition of latitude and longitude errors.) We see that the network cannot be considered fully rigid, at least not as far as the coastal areas are concerned. Noting this, there is nothing to contradict the plausibility of our simulation study.

5.5 Supporting Evidence for Estimates of Global and Local Covariances from the Mathematical Theory of Regular Networks

The U.S. network can hardly be called a regular one. Nevertheless, there is no doubt that it shares some properties with networks based on completely regular design and weight structure. In cooperation with other theoretical geodesists (see Borre and Meissl (1974), Bartelme and Meissl (1974), Meissl (1976)) I have studied regular networks extensively. Without repeating detailed derivations, we will take a look at a certain regular network which shares some features with the U.S. network. The purpose is to study how relative accuracy depends on distance.

5.5.1 Regular model of the U.S. network

First imagine a purely directional network, as shown in figure 5.11. We do not make any specifications at present about the boundary of the network.

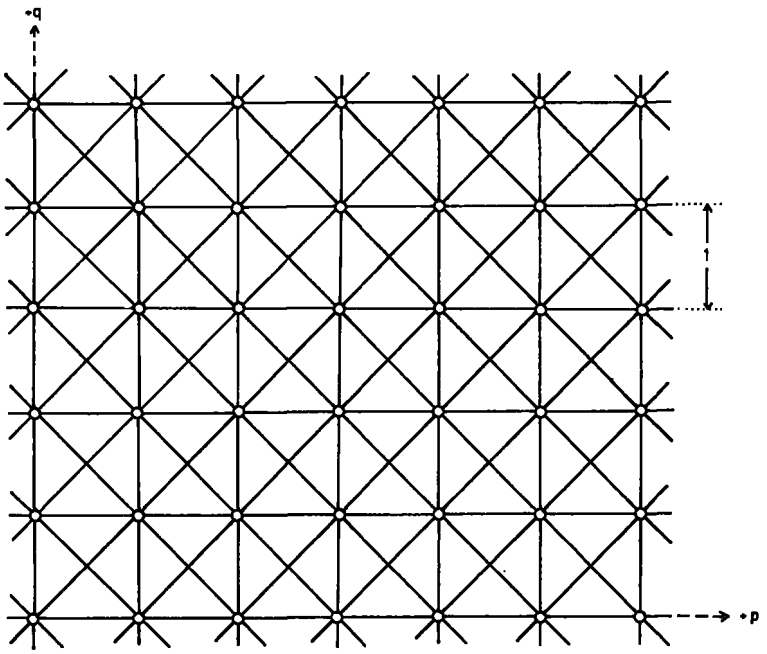

Figure 5.11.—Portion of a regular directional network.

We study the normal equations of an interior station, situated at the intersection of the grid lines p and q. We denote by $\Delta x_{pq} = (\Delta\xi_{pq}, \Delta\eta_{pq})^T$ the coordinate shifts of station (p,q). We are interested in the left side of the normals from which we imagine the direction unknowns eliminated. The normals look like

$$\sum_{r,s=-2}^{+2} A_{rs}\, \Delta x_{p+r,q+s} = \ldots \tag{5.15}$$

They may also be written as:

$$\sum_{r,s=-2}^{+2} A_{p-r,q-s}\, \Delta x_{rs} = \ldots \tag{5.15a}$$

The 2×2 coefficient matrices of the parameters Δx_{pq} do not depend on the location of (p,q). Any interior node has the same left side coefficients in its two normal equations. The normal equations are translation-invariant. The numerical values of the matrices A_{rs} can be read from table 5.1. The values are scaled to an assumed grid spacing of unity. Also the observational weights are assumed to be uniformly equal to the value $p = 1$. Obvious scale factors have to be applied if spacing and weights change.

A continuous analog of eqs. (5.15) is obtained if we let $\Delta\xi(p,q) = \Delta\xi_{pq}$, $\Delta\eta(p,q) = \Delta\eta_{pq}$, and if we view $\Delta\xi(p,q)$, $\Delta\eta(p,q)$ as smooth functions of the continuous variables p,q, interpolating the discrete values $\Delta\xi_{pq}$, $\Delta\eta_{pq}$.

TABLE 5.1.—*Left-side normal equation coefficients A_{rs} (see eq. (5.15)) for a regular directional network.*

$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{3}{16}$	0	$\frac{1}{8}$	$-\frac{1}{16}$	$\frac{1}{32}$	$-\frac{1}{32}$	+2
$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{16}$	0	0	$-\frac{1}{16}$	$-\frac{1}{16}$	0	$-\frac{1}{32}$	$\frac{1}{32}$	
0	$\frac{1}{16}$	$-\frac{1}{2}$	$-\frac{3}{8}$	-2	0	$-\frac{1}{2}$	$\frac{3}{8}$	0	$-\frac{1}{16}$	+1
$\frac{1}{16}$	$\frac{1}{8}$	$-\frac{3}{8}$	$-\frac{1}{2}$	0	$-\frac{1}{4}$	$\frac{3}{8}$	$-\frac{1}{2}$	$-\frac{1}{16}$	$\frac{1}{8}$	
$-\frac{1}{16}$	0	$-\frac{1}{4}$	0	$\frac{45}{8}$	0	$-\frac{1}{4}$	0	$-\frac{1}{16}$	0	0
0	$\frac{3}{16}$	0	-2	0	$\frac{45}{8}$	0	-2	0	$\frac{3}{16}$	
0	$-\frac{1}{16}$	$-\frac{1}{2}$	$\frac{3}{8}$	-2	0	$-\frac{1}{2}$	$-\frac{3}{8}$	0	$\frac{1}{16}$	−1
$-\frac{1}{16}$	$\frac{1}{8}$	$\frac{3}{8}$	$-\frac{1}{2}$	0	$-\frac{1}{4}$	$-\frac{3}{8}$	$-\frac{1}{2}$	$\frac{1}{16}$	$\frac{1}{8}$	
$\frac{1}{32}$	$-\frac{1}{32}$	$\frac{1}{8}$	$-\frac{1}{16}$	$\frac{3}{16}$	0	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$	−2
$-\frac{1}{32}$	$\frac{1}{32}$	$-\frac{1}{16}$	0	0	$-\frac{1}{16}$	$\frac{1}{16}$	0	$\frac{1}{32}$	$\frac{1}{32}$	
−2		−1		0		+1		+2		s, r

Taylor expansion up to second order gives

$$\Delta\xi(p + \Delta p, q + \Delta q) = \Delta\xi(p,q) + \frac{\partial \Delta\xi}{\partial p}\Delta p + \frac{\partial \Delta\xi}{\partial q}\Delta q + \frac{1}{2}\left[\frac{\partial^2\Delta\xi}{\partial p^2}\Delta p^2 + 2\frac{\partial^2 \Delta\xi}{\partial p\,\partial q}\Delta p\,\Delta q + \frac{\partial^2\Delta\xi}{\partial q^2}\Delta q^2\right] \quad (5.16)$$

together with a similar formula for $\Delta\eta(p + \Delta p, q + \Delta q)$. Inserting this into the normal equations for station (p,q), one gets (remembering that $\Delta p = \Delta q = 1$):

$$-\left\{\frac{\partial^2\Delta\xi}{\partial p^2} + \frac{\partial^2\Delta\xi}{\partial q^2}\right\} = \ldots.$$
$$-\left\{\frac{\partial^2\Delta\eta}{\partial p^2} + \frac{\partial^2\Delta\eta}{\partial q^2}\right\} = \ldots. \quad (5.17)$$

We see that the continuous approximation to the normals gives, as far as the interior stations are concerned, *Poisson's equation* for the two coordinates separately. The differential operator on the left-hand side is *Laplace's* operator.

Our next step is to add distances and azimuths to the network. In order to maintain translation invariance we will assume that distances and azimuths are measured along the lines shown in figure 5.12. Each line represents one distance and one azimuth.

As we have pointed out, the number of distances is about 0.01 of the number of directions. The number of azimuths, in turn, is about 0.1 of the number of distances. We shall account for this by weight factors p_d, p_a assigned to the individual distances and azimuths. (Recall that for directions we have assumed $p = 1$.) These weight factors account for not only the differences in number between the three kinds of measurements, but also for different individual accuracies. Whereas a direction is typically accurate to 0.5 arc seconds, *i.e.*, 2.5 ppm, distances may be as good as 1 ppm, whereas azimuths are about 5 ppm. One must also take into account that, contrary to figure 5.12, the true distance and azimuth subgraph is not connected. Connection is established via the directions, which degrades the distances somewhat. We assume

$$\begin{aligned} p &= 1 && \text{for directions} \\ p_d &= 0.225 && \text{for distances} \\ p_a &= 0.0012 && \text{for azimuths.} \end{aligned} \quad (5.18)$$

These numbers are obtained as follows: The total directional weight is the number of directions times weight factor = 2.5E6 * 1 = 2.5E6. Similarly, the total distance weight is 30,000 * 2.5^2 = 1.875E5. This is degraded by 0.75 to 1.40625E5. The total azimuth weight is 3,000 *2^{-2} = 750. The regular network has numbers of directions, distances, and azimuths in the ratio 8:2:2. Hence we solve the proportion 2.5E6 : 1.4E5 : 750 = $8p : 2p_d : 2p_a$ to obtain, with $p = 1$, the numbers in (5.18).

The normals resulting from distances and azimuths are generally structured as shown by eq. (5.15). Table 5.2 shows the numerical values for A_{rs}.

TABLE 5.2.—*Left-side normal equation coefficients A_{rs} (see eq. (5.15)) for regular distance and azimuth network.*

0	$-p_a$ 0 0 $-p_d$	0	+1
$-p_d$ 0 0 $-p_a$	$2p_d+2p_a$ 0 0 $2p_d+2p_a$	$-p_d$ 0 0 $-p_a$	0
0	$-p_a$ 0 0 $-p_d$	0	−1
−1	0	+1	⋱ s r ⋱

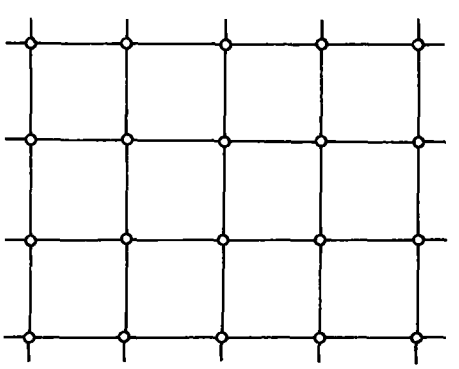

Figure 5.12.—Portion of a superimposed regular distance and azimuth network.

We may now imagine that (5.15) represents the superposition of the normals for directions, distances, and azimuths. Table 5.3 shows this superposition. We will not superimpose the Doppler measurements, at least not as long as we deal with a network covering the whole plane. There is good reason for setting aside the Doppler-derived absolute positions. The relative measurements are of superior accuracy as compared with the absolute positions. The 130 Doppler stations of the U.S. network provide a positional fix no better than $1/\sqrt{130}$ m = 9 cm. (See section 5.4.) On the other hand, in an infinite network covering the whole plane, there would be infinitely many Doppler stations. These, together with the good relative accuracy of the other measurements,

TABLE 5.3.—*Normals for the idealized U.S. network, accounting for directions, distances, and azimuths*

0.03125	0.03125	0.12500	0.06250	0.18750	0.00000	0.12500	−0.06250	0.03125	−0.03125	+2
0.03125	0.03125	0.06250	0.00000	0.00000	−0.06250	−0.06250	0.00000	−0.03125	0.03125	
0.00000	0.06250	−0.50000	−0.37500	−2.00120	0.00000	−0.50000	0.37500	0.00000	−0.06250	+1
0.06250	0.12500	−0.37500	−0.50000	0.00000	−0.47500	0.37500	−0.50000	−0.06250	0.12500	
−0.06250	0.00000	−0.47500	0.00000	6.07740	0.00000	−0.47500	0.00000	−0.06250	0.00000	0
0.00000	0.18750	0.00000	−2.00120	0.00000	6.07740	0.00000	−2.00120	0.00000	0.18750	
0.00000	−0.06250	−0.50000	0.37500	−2.00120	0.00000	−0.50000	−0.37500	0.00000	0.06250	−1
−0.06250	0.12500	0.37500	−0.50000	0.00000	−0.47500	−0.37500	−0.50000	0.06250	0.12500	
0.03125	−0.03125	0.12500	−0.06250	0.18750	0.00000	0.12500	0.06250	0.03125	0.03125	−2
−0.03125	0.03125	−0.06250	0.00000	0.00000	−0.06250	0.06250	0.00000	0.03125	0.03125	
−2		−1		0		+1		+2		s / r

TABLE 5.4.—*Normals of the 8 × 8 finite element grid*

0.00000	0.00000	0.01071	0.00193	0.04106	0.00000	0.01071	−0.00193	0.00000	0.00000	+2
0.00000	0.00000	0.00193	0.00021	0.00000	−0.00044	−0.00193	0.00021	0.00000	0.00000	
0.00021	0.00193	−0.40124	−0.00778	−0.44913	−0.00000	−0.40124	0.00778	0.00021	−0.00193	+1
0.00193	0.01071	−0.00778	−0.40124	0.00000	−0.44392	0.00778	−0.40124	−0.00193	0.01071	
−0.00044	0.00000	−0.44392	0.00000	3.26614	0.00000	−0.44392	0.00000	−0.00044	0.00000	0
0.00000	0.04106	0.00000	−0.44913	0.00000	3.26614	0.00000	−0.44913	0.00000	0.04106	
0.00021	−0.00193	−0.40124	0.00778	−0.44913	0.00000	−0.40124	−0.00778	0.00021	0.00193	−1
−0.00193	0.01071	0.00778	−0.40124	0.00000	−0.44392	−0.00778	−0.40124	0.00193	0.01071	
0.00000	0.00000	0.01071	−0.00193	0.04106	0.00000	0.01071	0.00193	0.00000	0.00000	−2
0.00000	0.00000	−0.00193	0.00021	0.00000	−0.00044	0.00193	0.00021	0.00000	0.00000	
−2		−1		0		+1		+2		$\bar{s}$ / $\bar{r}$

would provide a position fix which is unrealistically good. Hence we omit the Doppler stations altogether. Disregarding their damping effect on the covariance function of the coordinates, we may hope to obtain conservative estimates.

Next we assume a finite element grid superimposed on our regular network. Figure 5.13 indicates that any of the finite elements comprises $8 \times 8 = 64$ original squares. In addition to our earlier grid coordinates p,q, we introduce the coordinates $\bar{p},\bar{q}$ of the finite element corners as indicated in the figure. One step in $\bar{p},\bar{q}$ is equivalent to eight steps in p,q. The transformation (5.7), (5.10) is now applied to the normals, resulting in the normals for the corners of the finite element grid. Table 5.4 lists the resulting values.

Assuming that (5.15) are the normals for a regular network, we apply a Fourier transformation introducing $\Delta x(\phi,\psi)$, $A(\phi,\psi)$ by

$$\begin{aligned} \Delta x(\phi,\psi) &= \sum_{p,q=-\infty}^{+\infty} e^{i(p\phi+q\psi)}\,\Delta x_{pq} \\ A(\phi,\psi) &= \sum_{p,q=-2}^{+2} e^{i(p\phi+q\psi)}\,A_{pq} \end{aligned} \quad . \qquad (5.19)$$

The analysis applies to the original network as well as to the finite element network, with the understanding that in the finite element case, (p,q) is replaced by $(\bar{p},\bar{q})$. The normal equations then transform into

$$A(\phi,\psi)\,\Delta x(\phi,\psi) = \ldots \qquad (5.20)$$

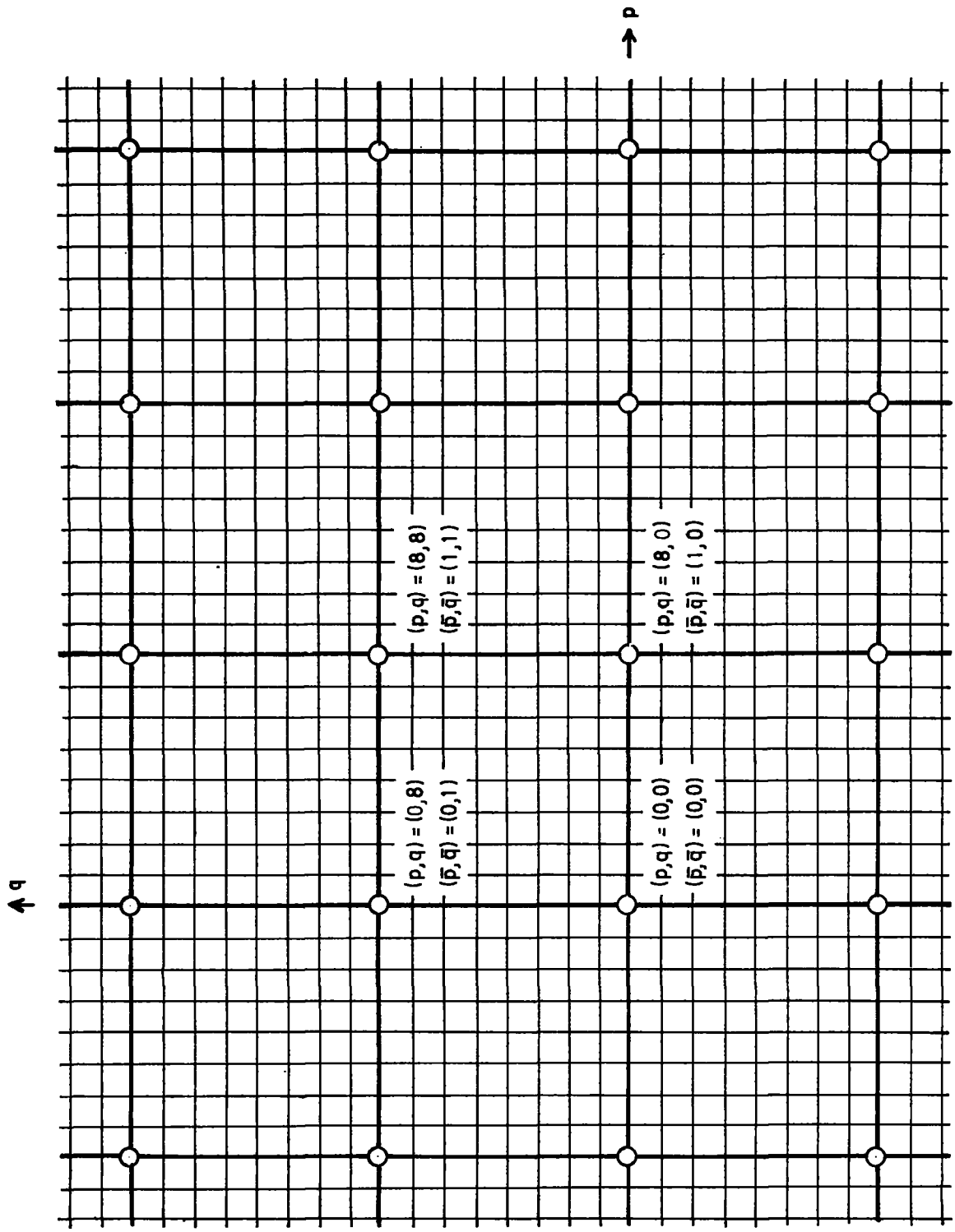

Figure 5.13.—Finite element grid superimposed on a regular network.

This is a consequence of Parseval's relation, and is explained in much detail in Bartelme and Meissl (1974). It follows that $F(\phi,\psi) = A^{-1}(\phi,\psi)$ is the Fourier transform of the inverse normal equation operator.

One is tempted to apply a backward Fourier trans-

formation to obtain the inverse normal equation operator as

$$F_{pq} = \frac{1}{(2\pi)^2} \int_{-\pi}^{+\pi}\int_{-\pi}^{+\pi} e^{-i(p\phi+q\psi)} F(\phi,\psi)\, d\phi\, d\psi. \qquad (5.21)$$

Unfortunately, this integral does not exist. The reason is that $F_{rs} \to \infty$ as $\sqrt{r^2+s^2} \to \infty$ (variances tend to infinity as the spacing between stations increases). In Bartelme and Meissl it is shown in detail that, instead of eq. (5.21), a substitute kernel F^-_{rs} can be used that is defined by

$$F^-_{pq} = \frac{1}{(2\pi)^2} \int_{-\pi}^{+\pi}\int_{-\pi}^{+\pi} (1-e^{-i(p\phi+q\psi)})\, F(\phi,\psi)\, d\phi\, d\psi. \qquad (5.22)$$

This serves its purpose as a generalized inverse of the normal equation operator in the following sense:

(*) F^-_{pq} is a solution of

$$\sum_{r,s=-2}^{+2} A_{p-r,q-s} F_{r-p',s-q'} = -\delta_{pp'}\delta_{qq'} I. \qquad (5.23)$$

Here I is the 2×2 unit matrix and $\delta_{\alpha\beta}$ is Kronecker's symbol.

(*) If a right-hand side f_{pq} is prescribed to the normals

$$\sum_{r,s=-2}^{+2} A_{p-r,q-s}\, u_{rs} = f_{pq}, \qquad (5.24)$$

and if the following consistency relation holds:

$$\sum_{p,q=-\infty}^{+\infty} f_{pq} = 0 \qquad (5.25)$$

then the appropriate solution is obtained as

$$u_{pq} = -\sum_{r,s=-\infty}^{+\infty} F^-_{p-r,q-s} f_{rs}. \qquad (5.26)$$

The Fourier transform $u(\phi,\psi)$ of u_{pq} follows from

$$u(\phi,\psi) = F(\phi,\psi) f(\phi,\psi). \qquad (5.27)$$

(*) It must be stressed that $-F^-_{pq}$ is not positive definite. The relation

$$-\sum_{p,q,r,s=-\infty}^{+\infty} f_{pq} F^-_{p-r,q-s} f_{rs} \geqslant 0 \qquad (5.28)$$

does not necessarily hold true for any f_{pq}. However (5.28) is valid if f_{pq} satisfies the consistency relation (5.25).

(*) Suppose that

$$\phi = \sum_{p,q=-\infty}^{+\infty} f_{pq}\, \Delta x_{pq} \qquad (5.29)$$

is an estimable function. The relations necessary for estimability are in (5.25). Recall that our present net is not absolutely positioned. It is sufficient that f_{pq}, aside from fulfilling (5.25), has only a finite number of nonzero coefficients. It follows that the variance of the best estimate $\tilde{\phi}$ for ϕ is given by (5.28), i.e., by

$$\sigma^2(\tilde{\phi}) = -\sum_{p,q,r,s=-\infty}^{+\infty} f_{pq} F^-_{p-r,q-s} f_{rs}. \qquad (5.30)$$

(*) The following asymptotic expansion of F^-_{pq} holds:

$$\lim_{p,q\to\infty} \{F^-_{pq} - \frac{\bar{S}}{2\pi} \log_e \sqrt{p^2+q^2}\} = C + D + \frac{\bar{S}}{2\pi}(\gamma + \log_e \pi) + \frac{1}{(2\pi)^2}\int_0^{2\pi} S(\tau) \log_e |\cos(\tau-\alpha)|\, d\tau. \qquad (5.31)$$

Thereby α is defined through

$$p = \sqrt{p^2+q^2}\cos\alpha, \qquad q = \sqrt{p^2+q^2}\sin\alpha \qquad (5.32)$$

$\gamma = 0.57721\ 56649\ldots$ is Euler's constant.

To explain the additional quantities in (5.31), the auxiliary function $P(\phi,\psi)$ is needed, which is homogeneous of degree -2 and accounts for the singularity of $F(\phi,\psi)$:

$$F(\phi,\psi) = P(\phi,\psi) + O(1), \quad \phi,\psi \to 0. \qquad (5.33)$$

From $P(\phi,\psi)$ we get

$$S(\tau) = P(\cos\tau, \sin\tau) \qquad (5.34)$$

as well as

$$\bar{S} = \frac{1}{2\pi}\int_0^{2\pi} S(\tau)\, d\tau. \qquad (5.35)$$

Furthermore

$$C = \frac{1}{(2\pi)^2}\int_{\phi=-\pi}^{+\pi}\int_{\psi=-\pi}^{+\pi} \{F(\phi,\psi) - P(\phi,\psi)\}\, d\phi\, d\psi \qquad (5.36)$$

$$D = \frac{1}{(2\pi)^2}\iint_{\substack{-\pi\leqslant\phi,\psi\leqslant+\pi \\ \sqrt{\phi^2+\psi^2}\geqslant\pi}} P(\phi,\psi)\, d\phi\, d\psi. \qquad (5.37)$$

(See Bartelme and Meissl (1974) for further details.)

Table 5.5 lists sample values of the kernel F^-_{pq} for the original idealized network (normals shown in table 5.3). Table 5.6 lists values of $F^-_{\bar{p}\bar{q}}$ for the finite element network. The tables also include relative distance and azimuth rms errors between certain pairs of stations. The pairs are identified by their difference vectors.

The close relation between the two kernels is best brought out by comparing the asymptotic expansions. For the original network we get

$$F^{-}_{pq} = \{0.14371 \log_e \sqrt{p^2+q^2} + 0.26111\} I + \Phi(\alpha) + o(1). \qquad (5.38)$$

The theoretical formula for the first constant 0.14371 is $(2\pi \sqrt{(1+p_d)(1+p_o)})^{-1}$. Confer eq. (5.18). The symbol I denotes the 2×2 identity matrix and $\Phi(\alpha)$ comprises direction-dependent terms. Recall the definition of α in eq. (5.32). Table 5.7 lists sample values of $\Phi(\alpha)$ which is a diagonal 2×2 matrix.

TABLE 5.5.—*Sample values for kernel* F^{-}_{pq} *for original idealized network*

p	q	F^{-}_{pq}		Distance error azimuth error
1	0	.19577	.00000	.6257
		.00000	.15871	.5634
1	1	.21875	.01501	.6838
		.01501	.21875	.6383
8	0	.45407	.00000	.9530
		.00000	.46613	.9655
8	8	.51019	.00050	1.0106
		.00050	.51019	1.0096
16	0	.55273	.00000	1.0514
		.00000	.56662	1.0645
16	8	.57173	.00016	1.0710
		.00016	.58012	1.0755
16	16	.60963	.00012	1.1043
		.00012	.60963	1.1041

TABLE 5.6.—*Values for* $F^{-}_{\bar{p}\bar{q}}$ *for finite element network*

$\bar{p}$	$\bar{q}$	$F^{-}_{\bar{p}\bar{q}}$		Distance error azimuth error
1	0	.29568	.00000	.7690
		.00000	.29879	.7730
1	1	.32310	.00078	.8048
		.00078	.32310	.8029
2	0	.38031	.00000	.8721
		.00000	.39208	.8855
2	1	.39573	.00007	.8912
		.00007	.40255	.8957
2	2	.43039	.00021	.9280
		.00021	.43039	.9276
3	0	.43643	.00000	.9343
		.00000	.44925	.9479
3	1	.44416	.00004	.9437
		.00004	.45479	.9526
3	2	.46443	.00007	.9655
		.00007	.46971	.9675
3	3	.49031	.00007	.9903
		.00007	.49031	.9902

TABLE 5.7.—*Sample values of* $\Phi(\alpha)$

α	$\Phi_{11}(\alpha)$	$\Phi_{22}(\alpha)$
0	−0.10704	−0.09255
$\pi/12$	−0.10598	−0.09343
$\pi/6$	−0.10314	−0.09591
$\pi/4$	−0.09943	−0.09943
$\pi/3$	−0.09591	−0.10314
$5\pi/12$	−0.09343	−0.10598
$\pi/2$	−0.09255	−0.10704

Other values of $\Phi(\alpha)$ follow from symmetry relations. The coordinate axes, *i.e.*, $\alpha = 0, \pi/2$, are axes of symmetry.

To properly compare eq. (5.38) with the corresponding formula for the finite element network, we must transform eq. (5.38) to the coarse grid $\bar{p}\bar{q}$, *i.e.*, we must substitute $(p,q) = (8\bar{p},8\bar{q})$. Calling the resulting kernel $F^{-(o)}_{\bar{p}\bar{q}}$, we get

$$F^{-(o)}_{\bar{p}\bar{q}} = \{0.14371 \log_e \sqrt{\bar{p}^2+\bar{q}^2} + 0.55984\} I + \Phi(\alpha) + o(1). \qquad (5.39)$$

Now we compare this with the formula for the finite element grid:

$$F^{-(f)}_{\bar{p}\bar{q}} = \{0.14371 \log_e \sqrt{\bar{p}^2+\bar{q}^2} + 0.38308\} I + \Phi(\alpha) + o(1). \qquad (5.40)$$

The main asymptotic term is the same in both expansions. This term brings out the logarithmic law which holds for large networks. Also the direction-dependent terms $\Phi(\alpha)$ are identical. Only the constant terms differ. They are appropriately smaller in the case of the finite elements, which apply a smoothing to the original network. It has been beautifully demonstrated that this smoothing is only local and not global.

The asymptotic difference between the two kernels is

$$F^{-(o)}_{\bar{p}\bar{q}} - F^{-(f)}_{\bar{p}\bar{q}} = 0.17676\, I + o(1). \qquad (5.41)$$

This is of the same order of magnitude as the relative mean square errors between two neighboring stations in the original network. (See table 5.5.)

5.6 Adopted Model for the Covariance

Summarizing all results and discussions of this chapter, we shall, as a basis for subsequent calculations, adopt the following model of the covariance matrix of the adjusted coordinates. For stations spaced farther than 300 km, the global covariance of section 5.2 will be used. Locally we superimpose two functions. The first one accounts for local "loose junks" of the network. It has a peak of value

(30 cm)² at zero distance and tapers off to zero within a distance of 30 km. The second one accounts for local variations of the primary stations. It has a peak value of (20 cm)² at zero distance and tapers off to zero within 300 km. As for the law of tapering off, in view of the logarithmic law we assume the following:

$$C^{(local)}(d) = \frac{p_0}{\log 4} \log \frac{4}{1 + 3\frac{d}{d_0}}. \qquad (5.42)$$

Here d is the distance between the two stations, d_0 is the distance at which the covariance decreases to zero (30 km, 300 km, respectively). p_0 is the value at $d = 0$ ((30 cm)², (20 cm)²). The constant value 4 has been chosen somewhat arbitrarily. It could be replaced by other values, say 6 or 8, without too much influence on the outcome.

Superimposing the global covariance and the two local ones, we get

$$C(d) = C^{(global)}(d) + \frac{.3^2}{\log 4} \log \frac{4}{1 + 3\frac{d_{km}}{30}} + \frac{.2^2}{\log 4} \log \frac{4}{1 + 3\frac{d_{km}}{300}}. \qquad (5.43)$$

$C(d)$ applies to latitude shifts as well as longitude shifts measured in meters.

Identical peak expressions were assumed for all four covariance entries referring to the pairings (ϕ,ϕ) (ϕ,λ), (λ,ϕ), (λ,λ) between latitudes and longitudes of two stations. Actually, I believe the cross covariances, *i.e.*, those of the type (ϕ,λ), (λ,ϕ), do not have such large peaks. Admitting such peaks for them also, will help us to stay on the safe side.

6. COUNT OF STORAGE LOCATIONS AND OPERATIONS

In chapter 5 we analyzed properties of the U.S. network that are independent of the algorithm adopted to solve the normal equations. In a mathematical sense, the inverse of the normal equation matrix does not depend on the way the normals are solved. Our next goal is to obtain estimates of the coefficients of the various reduction states that the normal equation system undergoes during its triangular decomposition. The zero reduction state is represented by the original equations. Although these coefficients are also independent on the chosen solution algorithm, it is convenient to analyze them together with those of the subsequent reduction states which depend very much on the solution algorithm.

The analysis of the size of these coefficients will be postponed until chapter 7. In this chapter we will be mainly concerned with number and pattern of the nonzero coefficients. Our general discussions on Helmert blocking and the associated fill-in which was carried out in section 3.5.5 must be specialized to the case of the U.S. network adjustment. Again we have insufficient information. Since it is not known in detail how the stations are tied together by the observations, there is no way to tell, even after prescribing the block boundaries, which nodes will be junction nodes and which ones will be interior. Also, the block boundaries that will be used during the adjustment are not yet known. It has been decided though that the blocks will be rectangular.

6.1 Specifying a Preliminary Blocking Scheme

Based on the density distribution of the stations in the 1° × 1° quads, I have designed a preliminary blocking scheme whose only purpose is to provide the necessary rough estimates for the roundoff study. It is expected that a judicious choice of block boundaries, based on insight into the actual interconnections between stations, will reduce the fill-in more than our scheme does. Hence we can view the estimates derived from our blocking scheme as being conservative.

The density distribution of stations in the 1° × 1° quads was shown in figure 5.1a; with some modifications to be explained in a moment, this illustration is repeated as figure 6.1. The density refers to all types of stations—primary triangulation, supplemental, and transcontinental traverse. Note that the values in figure 6.1 are divided by 10 and rounded. Our calculations are based on the original numbers.

Figure 6.1 emphasizes the boundaries of the 8° × 8° quads to assist the reader in easily comparing it with figure 6.2 which illustrates the preliminary blocking scheme. The hierarchy of the blocks should be recognizable from the thickness of the dividing lines. The two heaviest lines, at latitudes 25° and 49°, symbolize the Mexican and Canadian junction nodes.

Remark. Although not all these junction nodes are actually situated on the lines $\phi = 25°$ and $\phi = 49°$, we pretend that they do during most of our computer calculations, since this assumption makes the computer programs much simpler. Pseudo station occupancies were added in some of the 1° × 1° quads at latitudes 25° and 49° to account for this feature in our simulation model. In this way figure 5.1a is transformed into figure 6.1.

The whole network can be considered as one block of level 7. It is divided into two blocks of level 6

	50	48	46	44	42	40	38	36	34	32	30	28	26	24
066														
068	16	16	36	151	40									
	31	31	7	8	66	2								
070	68	68	9	6	36	61	18							
	26	26	26	37	2	100	166	139						
072	37			37	13	12	137	164						
	17			17	18	6	36	156	36					
074	14			14	49	28	54	32	184					
	7			7	22	9	12	29	197	62	6			
076					46	16	12	19	36	36	82	56	92	92
	14				14	16	12	22	36	131	192	169	111	82	67
078	38					36	18	26	22	49	160	61	29	71	64	9
	6					6	26	21	16	16	26	15	10	166	41	27
080	16					16	3	19	17	7	13	23	67	47	26	36	92									
	3						3	29	36	24	14	13	48	84	26	6	77	2		6	42	14	3	5	
082	26							26	19	17	16	13	14	46	17	16	36	61	76	58	24	6	1		
	1			1		7	36	21	27	11	9	10	16	16	13	6	15	9	12	11	24	58			
084	14		14	11	4	18	48	26	11	16	13	13	26	11	4	6	34	24	16	4					
	66		66	27	6	6	11	16	20	46	29	22	22	13	1	43	23	13	19	16					
086	16		16	12	6	4	6	36	29	14	42	21	13	12	6	12	12	14	26	11					
	6		6	3	6	3	6	32	36	12	33	23	44	16	4	6	3	6	66						
088	3	3	12	17	16	12	16	43	24	16	18	24	29	17	6	13	6	6	44						
	1	1	9	6	16	26	31	37	31	23	16	21	26	9	26	26	26	26	19	67	2				
090	1	1		6	16	7	23	36	28	26	13	24	17	24	39	23	16	16	13	36	72	4			
	2	2	2	13	13	16	13	16	14	11	16	16	16	19	26	27	29	37	26	19	46	1			
092	2	2	6	13	16	7	22	11	9	16	7	14	13	14	12	23	15	13	16	16	29	1			
	4	4	11	22	9	12	12	6	12	7	14	26	16	11	12	25	19	16	12	16	13				
094	4	4	16	16	23	36	14	6	6	6	16	26	16	16	6	16	19	24	12	26	27				
	6	6	16	9	27	13	12	16	6	6	26	21	29	16	9	16	16	17	13	16	29				
096	16	16	14	11	11	16	16	14	17	11	6	17	20	9	16	7	14	14	11	17	51	6			
	15	15	19	29	9	14	24	16	11	21	21	15	16	6	12	14	24	36	9	14	16	15			
098	9	9	12	13	9	12	9	12	13	26	16	7	26	24	33	21	16	36	13	16	5	16	46	20	3	3
	9	9	6	13	5	7	4	16	6	22	15	15	11	5	9	25	4	5	5	14	21	6	16	16		
100	12	12	5	2	3	16	12	9	5	16	14	16	5	4	5	14	4	15	16	6	6	5	5	2		2
	12	12	16	16	15	14	7	6	16	13	12	2	5	12	13	5	4	6	16	6	7	4				4
102	13	13	12	5	11	13	6	6	6	4	12	9	7	14	25	2	6	4	6	7	1					1
	15	15	7	6	11	14	6	6	12	7	1	16	6	6	9	5	5	4	7	6	2					2
104	11	11	2	4	6	16	4	5	14	16	3	5	6	12	16	3	4	3	7	6	5					5
	3	3	1	2	2	6	1	6	15	6	4	6	7	6	5	4	12	6	3	7	3					3
106	2	2		3	3	5	3	6	6	7	4	2	3	6	3	3	7	16	3	1						1
	4	4	6	3	1	5	2	4	7	6	6	5	6	6	46	27	15	26	11							11
108	2	2	7	3	5	7	2		6	3	5	5	4	7	9	6	6	16	2							2
	1	1	6	4	5	7	7	4	3	2	2	3	3	2	5	3	5	3	4							4
110	4	4	16	4	2	2	4	2	6	6	3	7	4	1	5	5	6	16	5							5
	4	4	9	2	4	2	6	2	11	6	4	7	6	1	5	3	16	17	16							16
112	7	7	16	1	2	4	7	6	6	14	7	3	3	6	11	3	19	13	16							10
	16	16	7	6	6	3	16	26	6	7	6	6	2	6	7	3	27	13	4							4
114	7	7	6	4	6	5	1	16	3	3	3	3	3	3	11	3	13	16								10
	16	16	4	3	7	6	1	16	3	12	6	6	6	14	19	29	17	22								22
116	7	7	8	7	6	6	11	11	6	12	6	3	4	3	16	16	13	26								26
	16	16	17	12	21	11	21	4	3	6	7	1	26	13	14	13	21	16								16
118	7	7	26	13	13	6	3	4	4	13	16	5	14	7	27	36	54	36								
	13	13	34	26	11	7	4	3	6	13	15	7	16	9	16	56	41	15								
120	7	7	13	31	11	14	3	2	4	1	33	16	14	16	26	39	13									
	4	4	12	3	13	7	1	12	12	13	13	6	14	16	37	23	5									
122	4	4	22	6	52	17	3	14	13	12	24	52	45	43	14											
	126	126	279	36	156	6	13	11	14	15	19	69	93	3												
124	29	29	26	113	33	29	15	18	17	12	24	18	2													
	12	12	16	27	1	23	56	26	16	32																
126																										

Figure 6.1.—Modified station occupancies of 1° × 1° quads. (Numbers shown are divided by 10 and rounded.)

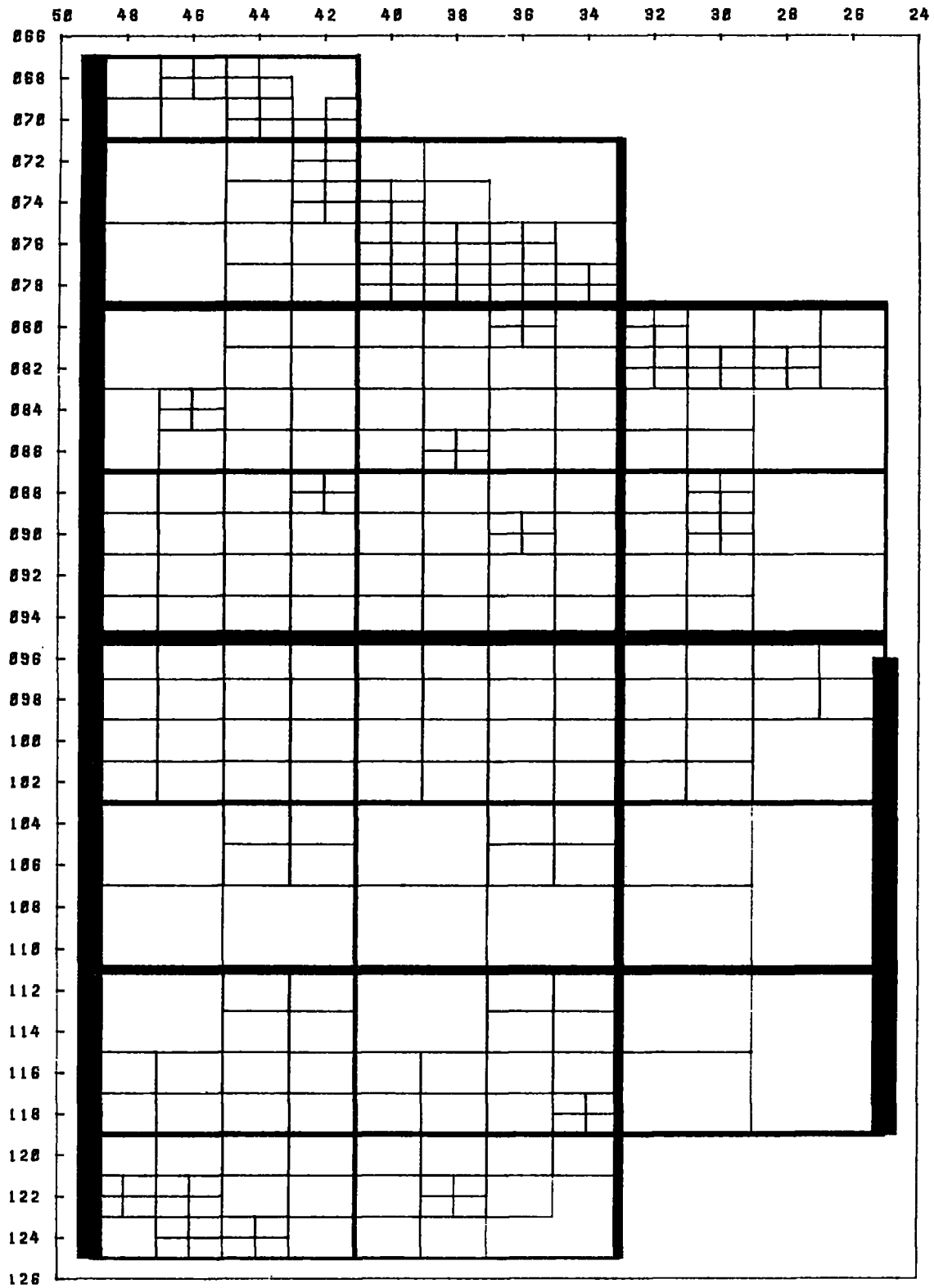

Figure 6.2.—Preliminary blocking scheme.

shown by the next to the heaviest line at longitude 95°. Each of the two blocks is subdivided into four blocks of level 5. The level 5 blocks are of varying size. The southeastern one is completely empty. The northern block of level 5 is split into four blocks—the two southern ones only into two level 4 blocks. Downwards from the fourth-level blocks (these blocks are mostly 8° × 8°), a systematic pattern prevails according to the nested dissection scheme. Subdivision can stop at levels 3, 2, or 1. Only in areas of dense control are first-level blocks of size 1° × 1° found.

Considering the pattern as being specified a priori down to the fourth level blocks of size 8° × 8°, the criteria for proceeding farther down to the third, second, and first level blocks were the following: A 2° × 2° quad was divided into four first-level 1° × 1° quads if it contained more than 1,000 stations. Similarly, a 4° × 4° quad was divided into four 2° × 2° quads if it contained more than 1,000 stations. Otherwise, a 4° × 4° quad was considered a third-level block with no subdivision. As it turned out, no 8° × 8° quads had less than 1,001 stations.

6.2 Counting the Nonzero Coefficients and the Elementary operations

The counts described in this section were derived with the assumption that the distribution of stations is uniform in any 1° × 1° quad. Consider a 1° × 1° quad together with its four neighboring quads at north, east, south, and west. (See fig. 6.3.) Assume that there are n stations in the quad and $n^{(n)}$, $n^{(e)}$, $n^{(s)}$, $n^{(w)}$ stations in the adjacent ones. We denote by i the number of interior stations of the quad multiplied by 2. By b_n, b_e, b_s, b_w we denote the number of boundary stations multiplied by 2. The reason for doubling all these numbers is that any node has two coordinates and gives rise to two equations, *i.e.*, to two rows and two columns of the normal equation matrix. We first assume that all adjacent quads at north, east, south, west are occupied by at least one station. We then calculate

$$b_n = b_s = 2\sqrt{n*\alpha}\,,\ \alpha = a/b \doteq cos(40°) \doteq 0.8$$

$$b_e = b_w = 2\sqrt{n/\alpha} = b_n/\alpha \qquad (6.1)$$

$$i = Max\{2n - b_n - b_e - b_s - b_w, 0\}$$

The underlying assumption is a regular distribution in the quad with one row of stations at the inner side of each of the four boundaries (fig. 6.4).

To be honest, our formulas do not yield quite the correct number of interior nodes (black circles) and boundary nodes (white circles) found in figure 6.4. With $n=24$, we get $b_n = b_s = 8.6$, $b_e = b_w = 11.2$. Considering that the corner nodes contribute to two boundaries each, we should actually have $b_n = b_s = 6$, $b_e = b_w = 10$. We neglect these deviations. They become less and less important as the number n of nodes increases. Our formulas are actually based on a continuous and smeared out distribution of stations.

Having calculated i, b_n to b_w under the assumption that the neighboring quads all are occupied, we must

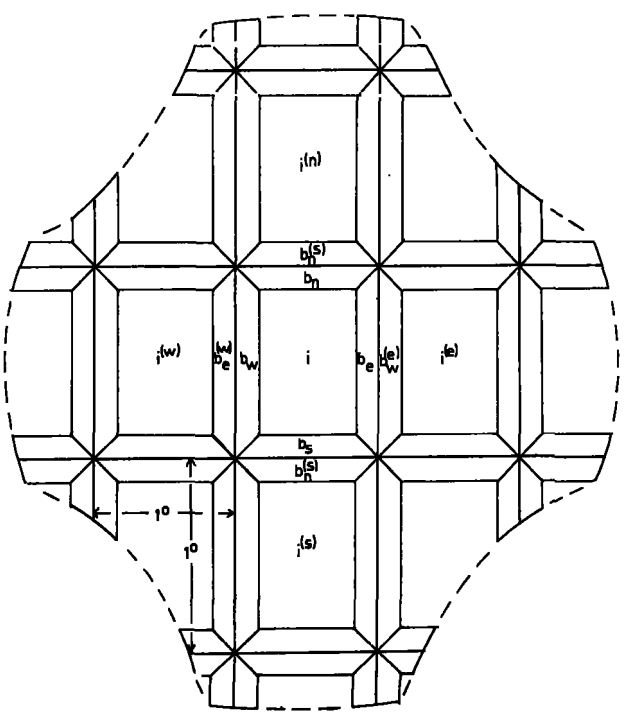

Figure 6.3.—Interior and boundary equations for a 1° × 1° quad and its four neighbors.

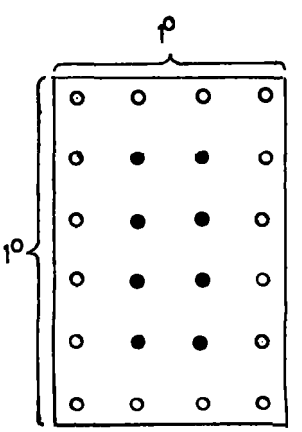

Figure 6.4.—Distribution of interior and boundary stations in a 1° × 1° quad.

consider modifications that take place if one or more of the adjacent quads are empty. It suffices to outline the case where $n^{(e)} = 0$. We then put $b_w^{(e)} = 0$, add b_e to i and finally put $b_e = 0$ afterwards. The reason for this is that if the adjacent quad is empty, there will never be a need for junction stations at the common boundary, should this boundary ever become a dividing line between blocks. If $n^{(e)} > 0$, we set

$$b_w^{(e)} = 2\sqrt{n^{(e)}/\alpha} \tag{6.2}$$

and leave i as it was before. After treating the remaining quads similarly, the numbers i, b_n, b_e, b_s, b_w, $b_s^{(n)}$, $b_w^{(e)}$, $b_n^{(s)}$, $b_e^{(w)}$ are all calculated.

Remark: Our formulas for calculating the number of boundary nodes are based on the assumption that this number increases with the square root of the density. This assumption may be questioned. However, it is consistent with our earlier assumption on "invariance of regional redundancy." (See sec. 5.1.4.) Lines of vision must decrease when density increases. The rate of decrease equals the square root of the rate of increase of the density. Figure 6.5 shows an example of a network that conforms with these assumptions.

6.2.1 First-level counts

A first-level partial block reduction is performed only at those 1° × 1° quads which qualify as outlined in section 6.1. As shown in figure 6.1, there are 124 such quads, most of them (about 60 percent) near the east coast.

In setting up the normals for a 1° × 1° block, there will be i interior equations and a number j of junction equations that is obtained as (see also fig. 6.6)

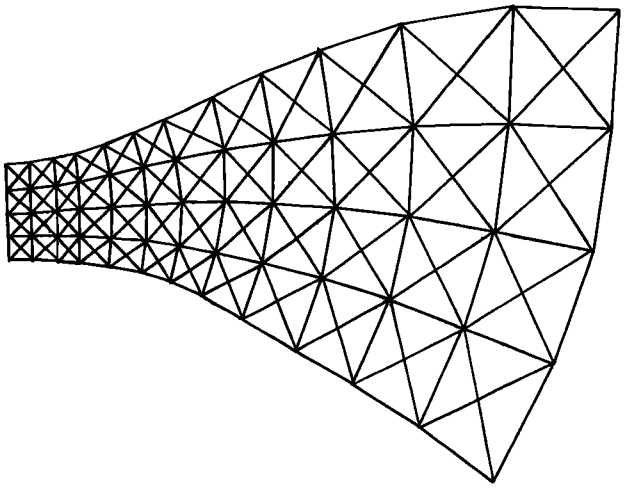

Figure 6.5.—Idealized example of a network with invariant regional redundancy.

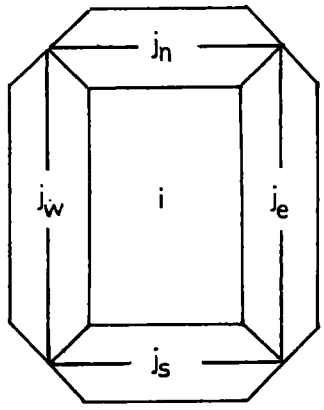

Figure 6.6.—Interior and junction equations for a 1° × 1° block.

$$j = j_n + j_e + j_s + j_w \tag{6.3}$$

with (cf. fig. 6.3)

$$\begin{aligned} j_n &= b_n + b_s^{(n)}, \; j_e = b_e + b_w^{(e)} \\ j_s &= b_s + b_n^{(s)}, \; j_w = b_w + b_e^{(w)} \end{aligned} \tag{6.4}$$

Assuming a uniform distribution of the interior nodes, there appears to be no more efficient ordering scheme than the one that produces minimum bandwidth. Referring to figure 6.6, let the interior nodes be numbered row-wise from top left (northwest) to bottom right (southeast). Then number the northern junction nodes from left to right, then the southern ones from left to right. Finally, number the union of the eastern and western junction nodes from north to south. Our subsequent formulas will be based on this numbering scheme; however, the reader is informed now that, by rotating the figure around, alternative numbering schemes will be considered for selecting the most efficient one.

Ordering the stations in the manner specified above will cause the normal equation matrix of the block to be profiled as shown in figure 6.7. After partial reduction, only a few zero coefficients will remain within the profile. Hence we can identify the profile with the set of nonzero coefficients. Another quantity depicted in figure 6.7, which has not yet been explained, is w, the "bandwidth," which equals the doubled number of interior equations in an east-west row, *i.e.*, it equals the quadruple number of stations in such a row:

$$w = 4\sqrt{\frac{i}{2} * \alpha}\,. \tag{6.5}$$

We will not count coefficients for the entire profile shown in figure 6.7, but rather only in the cross-hatched area. This is more logical because of partial reduction associated with Helmert blocking. The

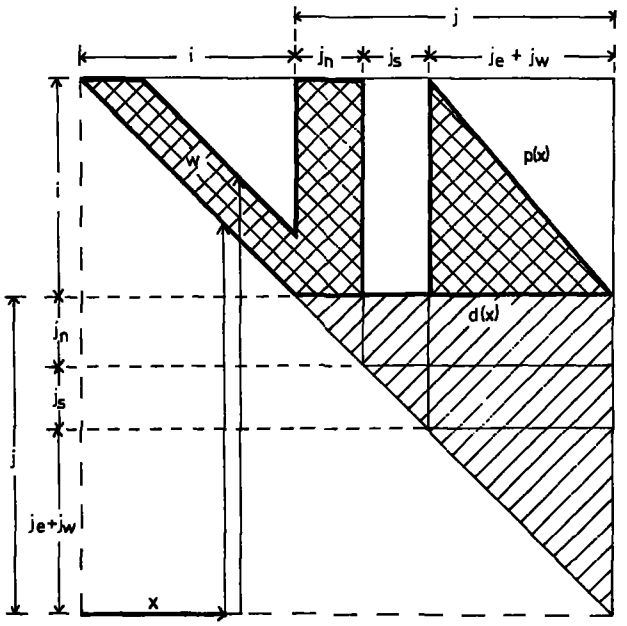

Figure 6.7.—Profile of normal equations for a 1° × 1° block.

single-hatched areas will be counted at higher block levels. Also note that the sums of products, which are so typical for the triangular decomposition phase in Cholesky's algorithm, involve only coefficients within the cross-hatched portion. Coefficients in the single-hatched area will only be modified by one single subtraction (of a precalculated product sum). Counting coefficients in the upper part of the profile is about equivalent to calculating the area of this portion of the profile. Note that the parameter x identifies a diagonal position. The profile to be counted at this diagonal position has the value

$$\Pi(x) = Max\{p(x) - d(x), 0\}. \tag{6.6}$$

(Actually $\Pi(x) = p(x) - d(x)$ in the present case. Equation (6.6) also holds good for situations arising in a different, slightly more general context).

The total profile to be counted equals the integral

$$\Pi = \int_0^{i+j} \Pi(x)\, dx = \int_0^{i+j} Max\{p(x) - d(x), 0\}\, dx. \tag{6.7}$$

Because of the simplicity of the functions $p(x)$, $d(x)$ one can specify an algebraic formula, namely

$$\Pi = w^2/2 + (i-w)w + ij_n + i(j_e+j_w)/2. \tag{6.8}$$

Proceeding to the count of the elementary operational steps needed to partially reduce the normal equation matrix of a 1° × 1° block, we note that the number of elementary steps associated with diagonal position x and with coefficients in row x is

$$\Gamma^{(row)}(x) = 2\int_x^{i+j} Min\{Max(p(x) - d(x), 0), Max(p(y) - d(x), 0)\}\, dy. \tag{6.9}$$

The factor 2 in front of the integral accounts for the fact that for any inner product term we have one multiplication and one addition.

The number of elementary steps associated with diagonal position x and with coefficients in column x is given by

$$\Gamma^{(col)}(x) = 2\int_0^{x} Min\{Max(p(y)-d(y), 0), Max(p(x)-d(y), 0)\}\, dy. \tag{6.10}$$

The total number of elementary steps is obtained in either one of the following ways:

$$\Gamma = \int_0^{i+j} \Gamma^{(row)}(x)\, dx = \int_0^{i+j} \Gamma^{(col)}(x)\, dx. \tag{6.11}$$

Because of the simplicity of $p(x)$, $d(x)$, we can specify the following algebraic expression:

$$\begin{aligned}\Gamma = 2*[&w^3/6 + j_n w^2/2 + ((j_e+j_w)/i)\, w^3/6 \\ &+ (i-w)w^2/2 \\ &+ (i-w)j_n w + ((i-w)^2/i)(j_e+j_w)w/2 + ij_n^2/2 \\ &+ ((i-w)/i)(j_e+j_w)\, w^2/2 + ij_n(j_e+j_w)/2 \\ &+ i(j_e+j_w)^2/6].\end{aligned} \tag{6.12}$$

Remark: Equations (6.6), (6.7), (6.9), (6.10), and (6.11) are general enough to extend to any partial Cholesky reduction scheme as soon as the functions $p(x)$, $d(x)$ are specified. We will also use these formulas in the subsequent sections. Equations (6.8) and (6.12) refer only to the first level counts.

Remark: One more important aspect has to be stressed. Because of the unequal number of junction stations at the north, east, south, west boundary segments, it may be advantageous to number the stations in one of four different ways. Our notation refers to row-wise numbering of the interior stations from north-west to south-east. However, row-wise numbering from southeast to northwest can occasionally result in a smaller Π and, if Π is chosen as the criterion for the efficiency of a numbering scheme, may be preferable. Two more alternatives must be considered, namely columnwise numbering from either northeast to southwest or from southwest to northeast. Columnwise numbering is associated with a different bandwidth, namely

$$w = 4\sqrt{(i/2)/\alpha}. \tag{6.13}$$

The computer algorithms used to evaluate the counts in this chapter considered the four different numbering schemes, and the one which minimized Π was chosen.

6.2.2 Counts for 2° × 2° quads

Several important new aspects enter when we proceed from 1° × 1° quads to 2° × 2° quads. On the other hand, 2° × 2° quads will be typical enough for counts of blocks having any size.

A distinction must be made whether a 2° × 2° quad splits into four first level 1° × 1° quads or not. The second case, which we refer to as a "low count" for a 2° × 2° quad, will be treated before the first one; this will later be called a "medium count" for a 2° × 2° quad. The low count is concerned with blocks that do not decompose into smaller ones. Their interior stations never become junction stations for any block. The counting procedure is similar to the one for the low counts of the 1° × 1° quads described in previous sections; however, some complications arise. Figure 6.8 shows the situation and adopted notation.

Note, for example, the interior equations of $i^{(nw)}$ also include the equations $b_i^{(nw)}$ and $b_i^{(nw)}$ of the north-western block because these equations do not play the role of junction equations. Similar conventions are adopted for the remaining three blocks. Elimination still proceeds row-wise from northeast to southwest. In the subsequent derivations it is assumed that the union of the eastern and southern junction stations is numbered from north to south.

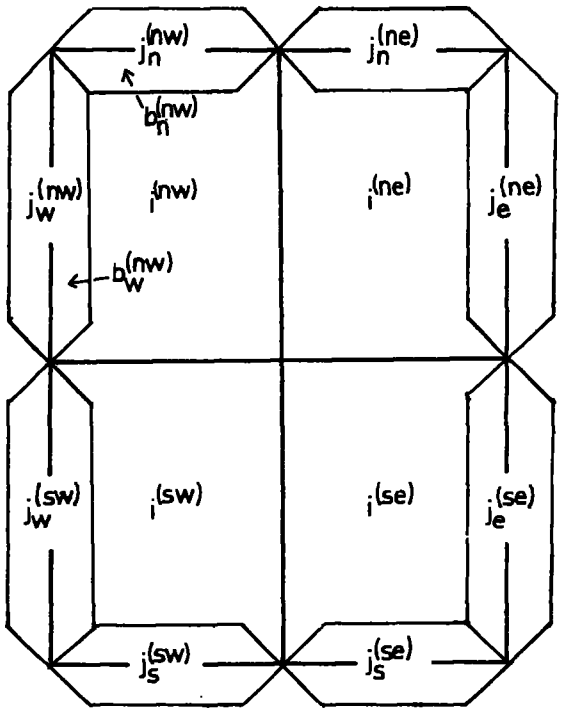

Figure 6.8.—Interior and junction equations for a 2° × 2° "low"-level block.

The bandwidth must be calculated in a more complicated way, *e.g.*, for the northern two blocks by

$$w_n = 4\sqrt{\frac{i^{(nw)}}{2}\alpha} + 4\sqrt{\frac{i^{(ne)}}{2}\alpha}\,. \tag{6.14}$$

The profile of the normals can appear as shown in figure 6.9. Here we have introduced, *e.g.*,

$$\begin{aligned} i^{(n)} &= i^{(nw)} + i^{(ne)},\ i^{(s)} = i^{(sw)} + i^{(se)} \\ j_n^{(n)} &= j_n^{(nw)} + j_n^{(ne)},\ j_s^{(s)} = j_s^{(sw)} + j_s^{(se)}. \end{aligned} \tag{6.15}$$

Formulas (6.6) and (6.7) for the Π-type count apply without change if the whole block is considered. Also, the formulas for the Γ-type counts carry over to the whole block. However, we want the counts for the individual 1° × 1° quads making up the block. For the purpose of later calculations involving Γ counts, based on eqs. (4.35) and (4.36a), we will distinguish between row counts $\Gamma^{(row)}$ and column counts $\Gamma^{(col)}$ for a certain 1° × 1° quad.

$\Gamma^{(row)}$ is the number of elementary operational steps (during triangular decomposition) involving nonzero coefficients a_{ij}, such that coordinate i is located in the 1° × 1° quad under consideration, while coordinates j can be anywhere. Similarly, $\Gamma^{(col)}$ is the number of steps involving nonzero a_{ij} such that j is located in the quad under consideration while i can be anywhere.

We denote by Γ_{pq}, $\Gamma_{pq}^{(row)}$, $\Gamma_{pq}^{(col)}$ the counts for quad (p,q), where in the present context (p,q) is one pair out of (n,w), (n,e), (s,e), (s,w). We arrive at the following equations:

$$\begin{aligned} \Pi_{pq} &= \int_{x\in Q_{pq}} \Pi(x)\,dx = \\ &\int_{x\in Q_{pq}} Max\{p(x)-d(x),0\}\,dx. \end{aligned} \tag{6.16}$$

Here the notation $x \varepsilon Q_{pq}$ indicates that the integration extends only over rows (strips of unit width) such that the diagonal position contributes to the quad under consideration. As for the Γ-counts, we have

$$\begin{aligned} \Gamma_{pq}^{(row)} &= \int_{x\in Q_{pq}} \Gamma^{(row)}(x)\,dx, \\ \Gamma_{pq}^{(col)} &= \int_{x\in Q_{pq}} \Gamma^{(col)}(x)\,dx. \end{aligned} \tag{6.17}$$

For the definition of $\Gamma^{(row)}(x)$, $\Gamma^{(col)}(x)$ see eqs. (6.9) and (6.10).

To indicate how these integrations can be orga-

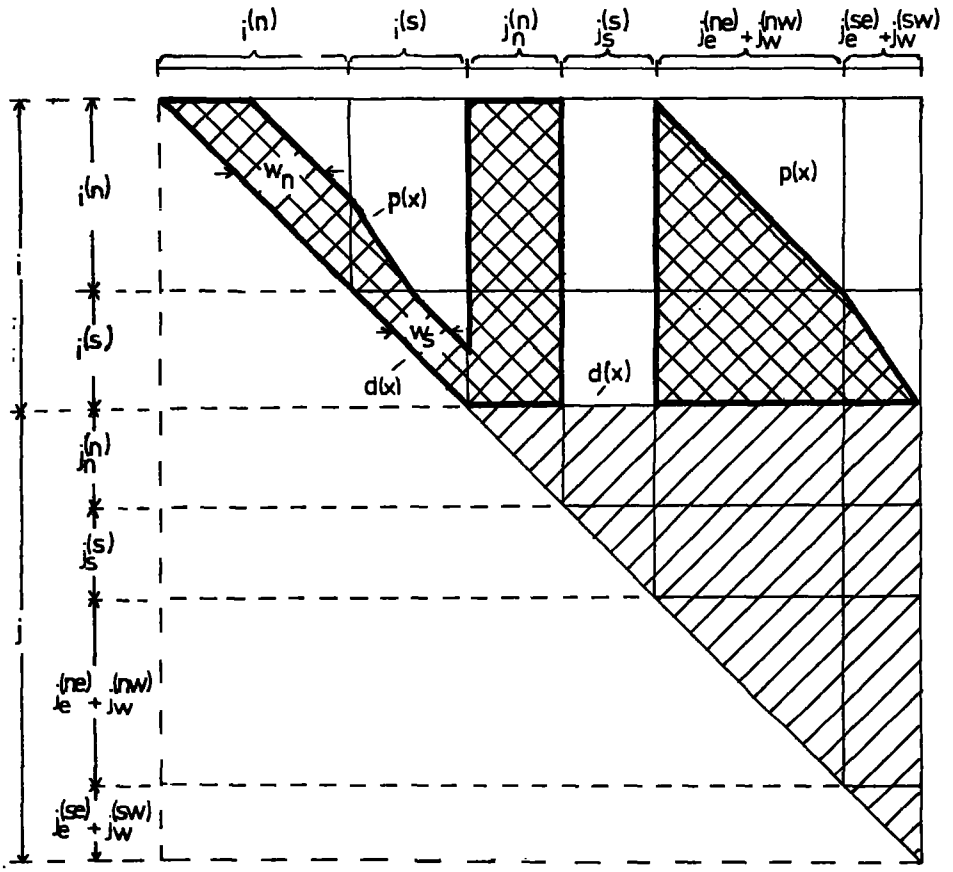

Figure 6.9.—Profile of normal equations for a 2° × 2° low-level block.

nized in practice, we give the expression for $\Gamma_{nw}^{(row)}$ as an example:

$$\Gamma_{nw}^{(row)} = \frac{j^{(nw)}}{j^{(n)}} \int_0^{i^{(n)}} \Gamma^{(row)}(x)\, dx + \frac{b_n^{(nw)}}{j_n^{(n)}} \int_i^{i+j_n^{(n)}} \Gamma^{(row)}(x)\, dx + \frac{b_w^{(nw)}}{j_e^{(ne)} + j_w^{(nw)}} \int_{i+j_n^{(n)}+j_s^{(s)}}^{i+j_n^{(n)}+j_s^{(s)}+j_e^{(ne)}+j_w^{(nw)}} \Gamma^{(row)}(x)\, dx. \tag{6.18}$$

This formula shows that we evaluate the interior contribution from the union of the two northern blocks (first integral) and then take the portion referring to the northwest block (factor in front of the first integral). We proceed in the same manner for the contribution of the northern junction equations, as well as the union of the eastern and western junction equations.

Let us mention at this point that any integral involved in a Π- or Γ-type count can be reduced to a sum of integrals whereby each summand is of the type

$$\int_0^A dx \int_0^{Bx+C} (Dy + Ex + F)\, dy. \tag{6.19}$$

Hence, algebraic expressions can be specified for all integrals. The details are elementary, but quite cumbersome. Because of the piecewise linearity of $p(x)$, $d(x)$ (see fig. 6.9), any integral of the type shown in eq. (6.18) can decompose into quite a few terms of the type shown in (6.19); therefore, we refrain from documenting these formulas here. Computer programs were written to evaluate these formulas. A good check was provided by eq. (6.11), *i.e.*, by the two independent ways to evaluate Γ.

We mention that by rotating the block by a multiple of 90 degrees, the Π count based on a somewhat

simplified approximate formula was minimized. This simplified Π count was obtained by lumping together all interior equations, as well as the northern, eastern, southern and western junction equations, and by using the 1° × 1° block formula eq. (6.8).

Let us now turn to the case where the 2° × 2° quad is composed of four first level blocks. A "medium count" of a 2° × 2° quad must be performed. Figure 6.10 explains the situation and the adopted notation.

The interior equations split into four sets i_n, i_e, i_s, i_w. These numbers in turn are composed of boundary equation counts for the participating 1° × 1° subquads, *e.g.* $i_n = b_e^{(nw)} + b_w^{(ne)}$. We form the union i_{ew} of i_e and i_w. The junction equations also split into eight sets, as shown in figure 6.10; however, it is advantageous to consider only two unions of junction sets, arriving at a northern and a southern set of junction equations:

$$\begin{aligned} j^{(n)} &= j_n^{(nw)} + j_w^{(nw)} + j_n^{(ne)} + j_e^{(ne)} \\ j^{(s)} &= j_w^{(sw)} + j_s^{(sw)} + j_s^{(se)} + j_e^{(se)}. \end{aligned} \tag{6.20}$$

We imagine the following station numbering: Set i_n precedes i_e which precedes i_w, which precedes $j^{(s)}$, so that $j^{(n)}$ is last. Within any of the sets i_n to $j^{(n)}$ we assume the station numbering as *random*.

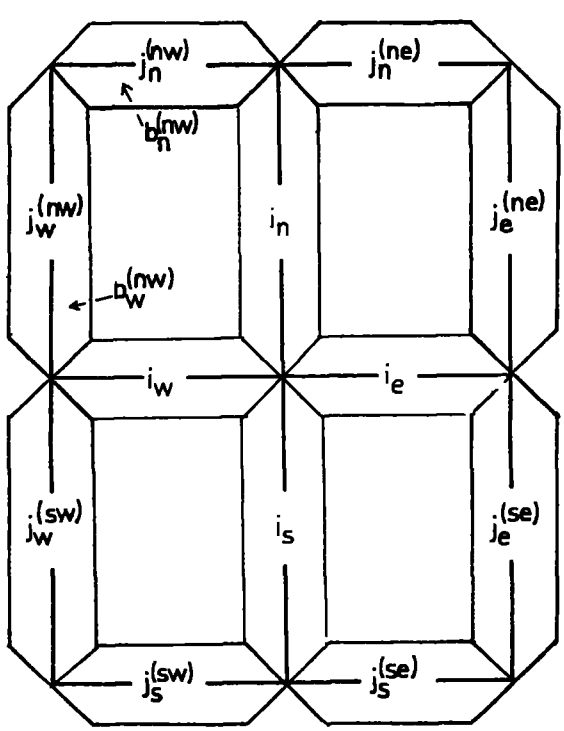

Figure 6.10.—Interior and junction equations for a 2° × 2° "medium"-level block.

The profile implied by these assumptions is shown in figure 6.11. Similar considerations, as in figure 6.9 and eq. (6.5) to (6.18) lead to the corresponding Π and Γ counts.

Because the functions $p(x)$, $d(x)$ are quite simple, the following algebraic expressions for the total counts can be specified (i. . . total number of interior equations, j. . . total number of junction equations):

$$\Pi = i^2/2 + ij - i_n(i_s + j^{(s)}) \tag{6.21}$$

$$\begin{aligned} \Gamma = 2*[&i_n^3/6 + i_n^2(i_{ew} + j^{(n)})/2 + i_s^3/6 + i_s^2(i_{ew} + j)/2 \\ &+ i_{ew}^3/6 + i_{ew}^2(i_n + i_s)/2 + (i_{ew}^2/2 + i_{ew}i_s)j^{(s)} + \\ &(i_{ew}^2/2 + i_{ew}(i_n + i_s))j^{(n)} + j^{(s)2}(i_s + i_{ew})/2 + \\ &j^{(s)}j^{(n)}(i_s + i_{ew}) + j^{(n)2}\, i/2]. \end{aligned} \tag{6.22}$$

Again we consider the individual Π, $\Gamma^{(r)}$, $\Gamma^{(c)}$ counts for the four 1° × 1° quads that comprise the block. The individual counts are affected by the assumption of random station labeling within the sets i_n, i_s, i_{ew}, $j^{(n)}$, $j^{(s)}$.

6.2.3 Remarks on the Counts of Larger Blocks

Low counts are also encountered with some 4° × 4° blocks. The procedure is similar to that described in the previous sections, only the number of 1° × 1° quads increases to 16 and the number of "block rows" increases to 4.

Medium counts for 4° × 4° (third level) and 8° × 8° (fourth level) blocks are quite analogous to those of the 2° × 2° blocks. This also holds for any larger block (fifth level and higher) that is composed of four subblocks. The only new complication arises if the subblocks are not all the same size. Individual counts for Π, $\Gamma^{(r)}$, $\Gamma^{(c)}$ for 1° × 1° subquads are nonzero only for quads that contribute to the sets of interior and junction stations.

At the fifth level and higher, there is occasionally a need to merge two subblocks instead of four. The situation and profile are shown in figures 6.12 and 6.13.

The total counts are

$$\begin{aligned} \Pi &= i^2/2 + ij \\ \Gamma &= 2[i^3/6 + i^2j/2]. \end{aligned} \tag{6.23}$$

The individual 1° × 1° block counts for Π, $\Gamma^{(r)}$, and $\Gamma^{(c)}$ rely on the assumption that the node labeling is random within the two sets i,j.

Finally, there is one "high count" for the last level, where the system is solved for the Canadian

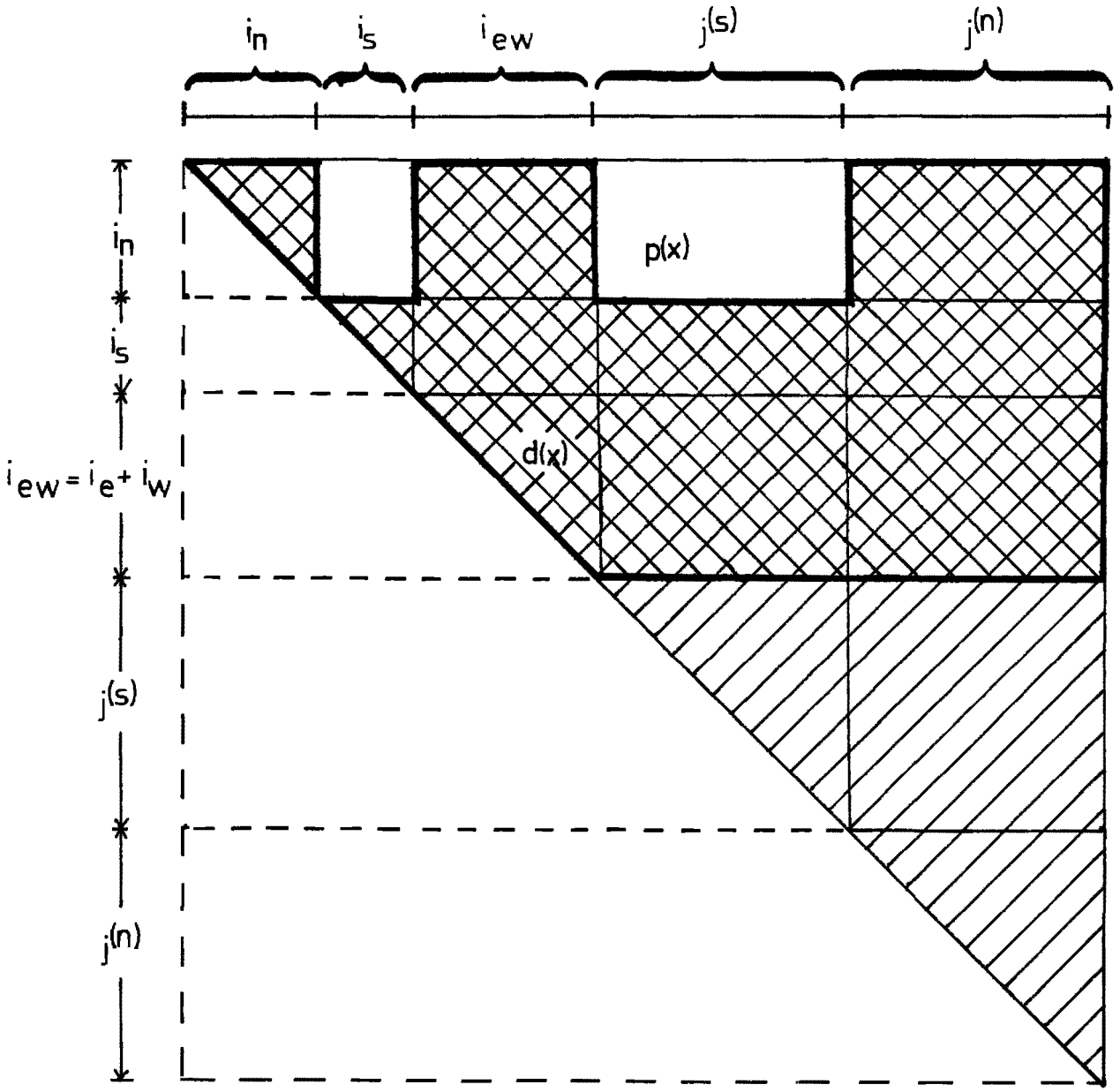

Figure 6.11.—Profile of normal equations for a 2° × 2° medium-level block.

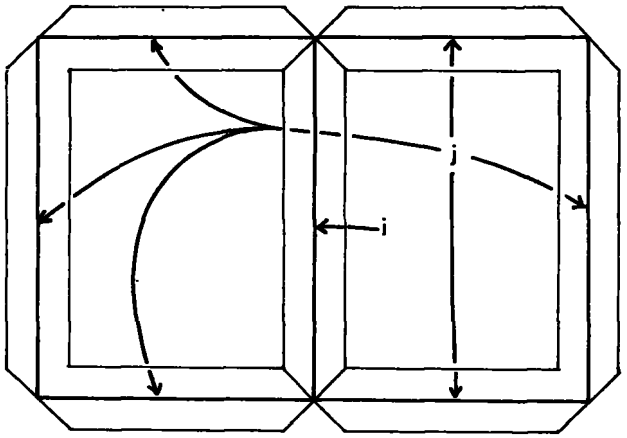

Figure 6.12.—Interior and junction equations for a 2° × 2° "modified-medium" level block.

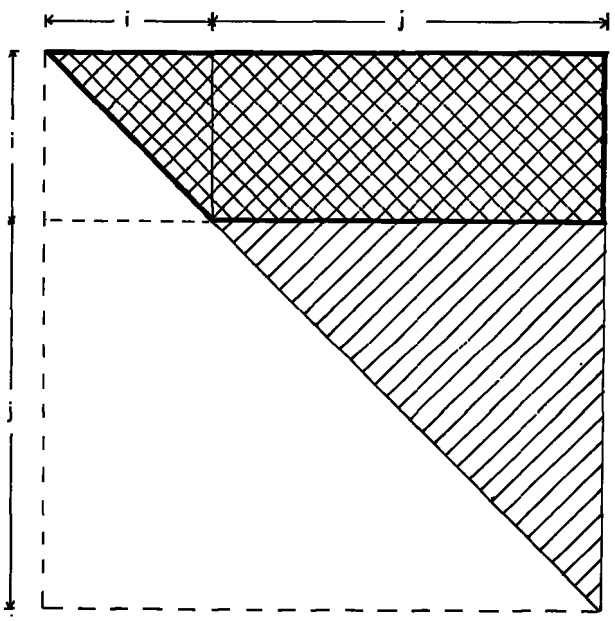

Figure 6.13.—Profile of normal equations for a 2° × 2° modified medium-level block.

and Mexican junction nodes. (The last level may contain some other stations, which are neglected here). All stations are interior now. The global counts are

$$\Pi = i^2/2$$
$$\Gamma = 2[i^3/6]. \tag{6.24}$$

The individual counts for Π, $\Gamma^{(r)}$, and $\Gamma^{(c)}$ that compose the block rely on the assumption of random station labeling.

6.2.4 Results of counts

Table 6.1 lists the results for the global counts accumulated during levels 1 to 8. (Level 8 refers to the final count which applies to the Canadian and Mexican junction station equations.)

The Π and Γ counts are listed for the various levels. The partial sums up to and including a certain level are also specified (columns headed Σ). Because it is of interest to know how many interior and junction equations participate at the different levels, these numbers are also shown in table 6.1.

Table 6.2 gives an alternative summary of the Π and Γ counts, as well as of the numbers of interior and junction equations. It splits the numbers into three categories "low," "medium," and "high." Recall that a low block is one that does not decompose into subblocks. (Such a block may be as large as 4° × 4°). The high block refers to the set of Canadian and Mexican junction nodes.

The results of the individual counts for Π, $\Gamma^{(r)}$, and $\Gamma^{(c)}$ are shown in figures 6.14a-b through 6.15a-b, where the counts were lumped into 2° × 2° quads. In figures 6.15a-b the $\Gamma^{(r)}$ and $\Gamma^{(c)}$ counts were superimposed because only the two sums will be needed later. Hence the sum of the Γ number over all quads must give twice the Γ value for the whole network, *i.e.*, $2*\Gamma = 2*1.2E11 = 2.4E11$. The counts along the northern and southern boundaries (left and right bounda-

Table 6.1.—*Summary of Π and Γ counts for levels 1 to 8.*

Level	Interior equations		Junction equations		Π		Γ	
	Count	Accum. count Σ	Count	Accum. count Σ	Count	Accum. count Σ	Count	Accum. count Σ
1	114,122	114,122	36,698	36,698	3.22E7	3.22E7	1.06E10	1.06E10
2	159,514	273,636	73,990	110,688	4.50E7	7.72E7	1.56E10	2.62E10
3	37,790	311,426	41,106	151,794	2.08E7	9.80E7	1.56E10	4.18E10
4	10,660	322,086	21,500	173,294	1.35E7	1.12E8	1.92E10	6.08E10
5	5,520	327,606	9,794	183,088	1.20E7	1.23E8	2.94E10	9.02E10
6	3,192	330,798	4,068	187,156	7.90E6	1.31E8	2.07E10	1.11E11
7	1,062	331,860	1,944	189,100	2.63E6	1.34E8	6.61E9	1.18E11
8	1,944	333,804	0	189,100	1.89E6	1.36E8	2.45E9	1.20E11

	50	48	46	44	42	40	38	36	34	32	30	28	26	24
066	11E4	19E4	95E4											
068	34E4	35E4	52E4	20E4	56E3									
070	35E4	40E3	57E4	26E5	23E5									
072	25E4		50E4	75E4	37E5									
074	95E3		58E4	35E4	27E5	15E5	59E4	98E3						
076	22E4		30E3	32E4	75E4	43E5	28E5	17E5	39E3					
078	18E4			68E4	11E5	12E5	18E5	23E5	11E5					
080	20E4			20E2	50E4	50E4	52E4	88E4	11E5	57E4	54E4	81E3	35E2	
082	13E4	34E3	94E3	56E4	68E4	39E4	73E4	54E4	68E4	44E4	20E4	18E4		
084	34E4	26E4	21E4	24E4	68E4	65E4	49E4	30E4	68E4	38E4	11E4			
086	13E4	20E4	51E4	51E4	12E5	10E5	10E5	71E4	61E4	65E4				
088	12E4	13E4	39E4	48E4	64E4	57E4	35E4	52E4	88E4	54E4	39E4			
090	20E4	16E4	36E4	34E4	50E4	34E4	50E4	42E4	12E5	47E4	53E4			
092	33E4	27E4	42E4	18E4	31E4	32E4	38E4	30E4	68E4	35E4	31E4			
094	72E4	98E4	11E5	89E4	10E5	11E5	12E5	97E4	11E5	99E4	81E4			
096	54E4	19E4	32E4	20E4	54E4	30E4	36E4	37E4	69E4	38E4	29E4	20E4	97E3	
098	46E4	15E4	20E4	22E4	39E4	41E4	26E4	28E4	39E4	25E4	29E4	12E4	60E3	
100	51E4	22E4	18E4	17E4	22E4	15E4	15E4	18E4	24E4	61E3	60E3		71E3	
102	51E4	17E4	29E4	23E4	38E4	22E4	33E4	24E4	22E4	18E4	72E3		75E3	
104	20E4	99E2	74E3	10E4	18E4	31E3	25E4	11E4	26E4	53E3	22E2		53E3	
106	28E4	13E4	16E4	58E3	15E4	13E4	42E4	36E4	41E4	10E4			91E3	
108	22E4	57E3	85E3	29E3	92E3	27E3	12E4	35E3	22E4	17E3			81E3	
110	48E4	39E4	37E4	38E4	66E4	41E4	52E4	60E4	79E4	24E4			14E4	
112	35E4	47E3	20E4	14E4	22E4	34E3	22E4	21E4	39E4	36E2			10E4	
114	38E4	22E4	29E4	19E4	29E4	18E4	38E4	43E4	66E4				20E4	
116	52E4	27E4	28E4	14E4	25E4	12E4	32E4	46E4	84E4				82E3	
118	76E4	69E4	52E4	14E4	38E4	45E4	64E4	87E4	63E4					
120	36E4	32E4	70E4	19E4	50E4	46E4	80E4	25E4	50E2					
122	20E5	35E5	15E5	30E4	49E4	61E4	61E4							
124	29E4	19E4	11E4	26E4	25E4	64E1								
126														

Figure 6.14a.—Π counts for 2° × 2° quads.

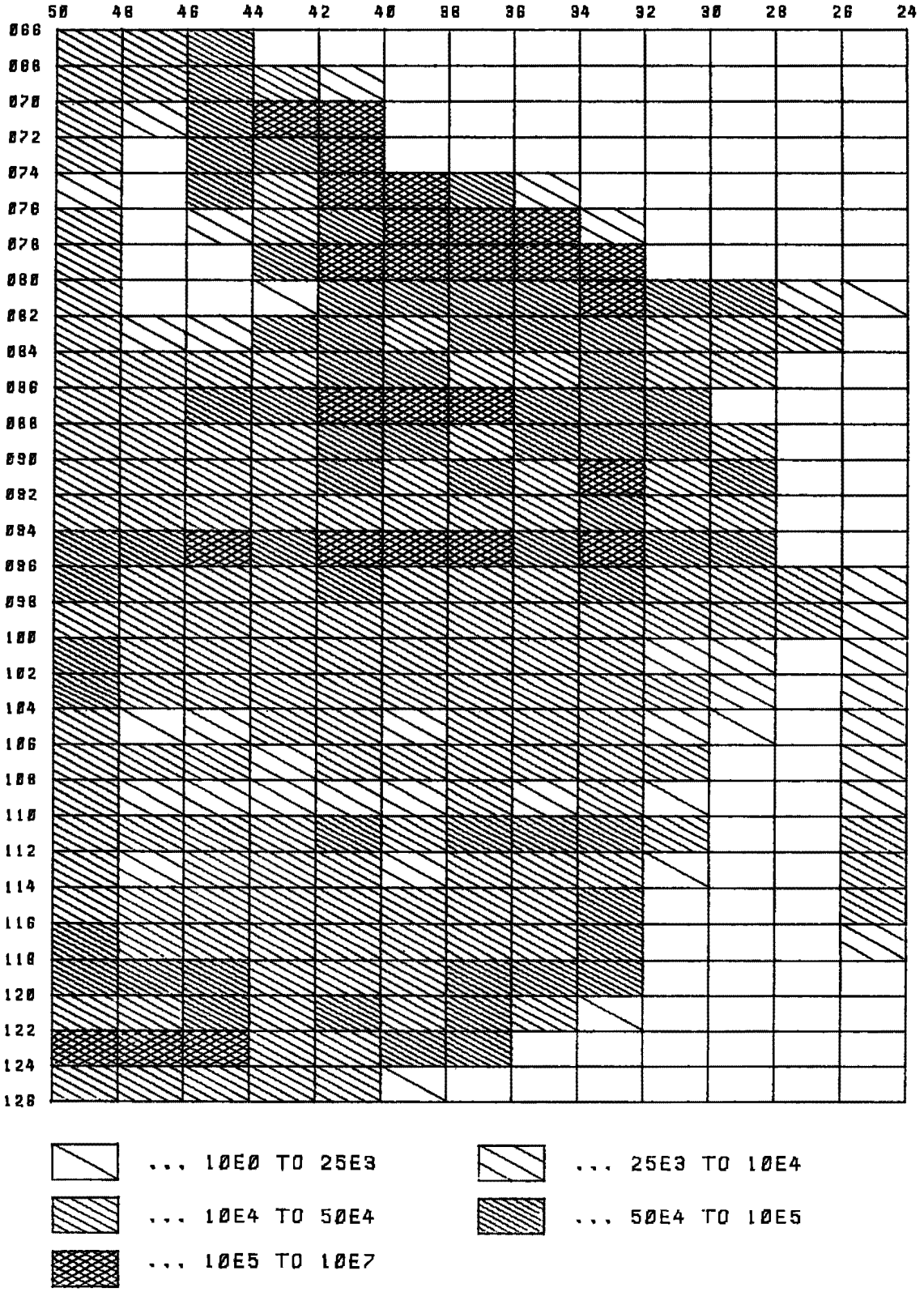

Figure 6.14b.—Π counts for 2° × 2° quads.

	50	48	46	44	42	40	38	36	34	32	30	28	26	24
066	32E7	90E6	57E7											
068	10E8	16E7	34E7	11E7	27E6									
070	98E7	12E6	72E7	30E8	23E8									
072	71E7		42E7	75E7	44E8									
074	27E7		55E7	48E7	36E8	17E8	69E7	46E6						
076	77E7		11E6	19E7	11E8	45E8	32E8	13E8	20E6					
078	78E7			19E8	34E8	32E8	46E8	56E8	31E8					
080	95E7			39E4	84E7	34E7	56E7	78E7	23E8	33E7	29E7	36E6	28E4	
082	62E7	10E6	87E6	38E7	11E8	61E7	97E7	67E7	17E8	36E7	13E7	67E6		
084	15E8	10E7	16E7	16E7	10E8	54E7	46E7	19E7	18E8	24E7	52E6			
086	61E7	47E7	12E8	12E8	32E8	22E8	28E8	16E8	17E8	63E7				
088	49E7	67E6	37E7	39E7	10E8	39E7	51E7	40E7	17E8	38E7	27E7			
090	79E7	13E7	38E7	35E7	79E7	34E7	66E7	49E7	22E8	51E7	49E7			
092	11E8	18E7	41E7	12E7	48E7	23E7	45E7	22E7	17E8	25E7	20E7			
094	28E8	36E8	41E8	34E8	37E8	41E8	46E8	33E8	42E8	24E8	20E8			
096	18E8	13E7	27E7	13E7	97E7	21E7	38E7	25E7	14E8	25E7	22E7	10E7	32E7	
098	17E8	13E7	23E7	16E7	85E7	31E7	28E7	23E7	72E7	20E7	23E7	67E6	22E7	
100	20E8	11E7	21E7	87E6	57E7	81E6	18E7	98E6	53E7	36E6	41E6		28E7	
102	19E8	24E7	48E7	29E7	91E7	32E7	53E7	34E7	58E7	14E7	53E6		32E7	
104	83E7	51E5	63E6	45E6	46E7	21E6	21E7	67E6	63E7	20E6	25E4		24E7	
106	91E7	82E6	12E7	40E6	38E7	10E7	34E7	29E7	84E7	51E6			41E7	
108	83E7	34E6	67E6	15E6	28E7	15E6	10E7	23E6	61E7	50E5			37E7	
110	17E8	12E8	10E8	12E8	22E8	13E8	12E8	17E8	23E8	45E7			64E7	
112	13E8	30E6	16E7	81E6	46E7	22E6	21E7	13E7	81E7	60E4			47E7	
114	14E8	17E7	27E7	12E7	63E7	15E7	40E7	41E7	11E8				92E7	
116	17E8	19E7	27E7	67E6	46E7	69E6	33E7	36E7	13E8				36E7	
118	21E8	11E8	80E7	21E7	68E7	65E7	82E7	10E8	98E7					
120	11E8	36E7	78E7	10E7	74E7	37E7	65E7	16E7	97E4					
122	51E8	34E8	16E8	20E7	80E7	53E7	53E7							
124	91E7	10E7	10E7	10E7	35E7	14E4								
126														

Figure 6.15a.—Γ counts for 2° × 2° quads.

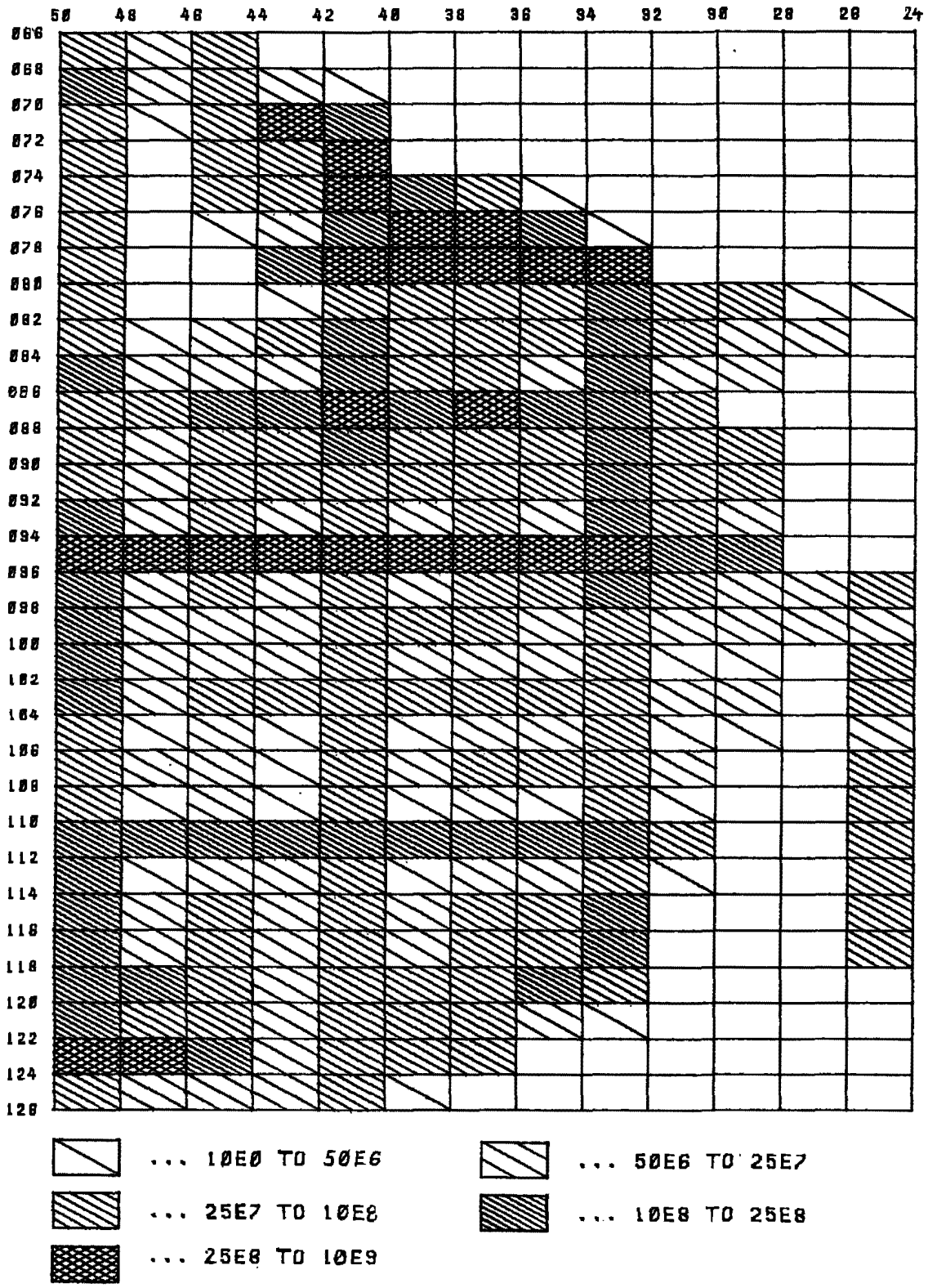

Figure 6.15b.—Γ counts for 2° × 2° quads.

Table 6.2.—*Summary of* Π *and* Γ *counts for "low," "medium," and "high" categories.*

Level	Interior equations		Junction equations		Π		Γ	
	Count	Accum. count Σ	Count	Accum. count Σ	Count	Accum. count Σ	Count	Accum. count Σ
low	282,930	282,930	100,100	100,100	7.65E7	7.65E7	2.35E10	2.35E10
med	48,930	331,860	89,000	189,100	5.75E7	1.34E8	9.41E10	1.18E11
high	1,944	333,804	0	189,100	1.89E6	1.36E8	2.45E9	1.20E11

ries) are caused by assumed pseudo-occupancies needed to account for the Canadian and Mexican junction stations. (See the discussion at the beginning of chapter 6.)

7. SIZE OF COEFFICIENTS DURING TRIANGULAR DECOMPOSITION

In this chapter we will try to gain more insight into the behavior of the coefficients $a_{ij}^{(p)}$, $b_i^{(p)}$, $0 \leqslant p < i \leqslant j \leqslant n$. These are the coefficients of the various partially reduced sets of normal equations. In chapter 3 (particularly sections 3.2 and 3.5) it became clear that the partially reduced normals are typical, not so much for Cholesky's method, but rather for any direct elimination method combined with a certain ordering strategy. A geodetic interpretation of the coefficients $a_{ij}^{(p)}$, $b_i^{(p)}$ was already given at the end of the section 3.4. This interpretation is valid for any geodetic network, however pathological it may be, and therefore it is not sufficient for estimating the size of the coefficients. Our task in this chapter is to specify the properties of the coefficients that follow from properties of the U.S. network and the chosen solution algorithm.

7.1 Left-Hand Side Coefficients

For a moment, let us pretend that Cholesky's algorithm is executed in the fashion outlined by eq. (3.16) at the end of section 3.2. We now investigate for *fixed* i,j; how $a_{ij}^{(p)}$ changes as p proceeds from 0 to $i-1$. (It cannot proceed farther because equation i is then next for pivoting.) Although Cholesky's algorithm will be executed in a different way, our procedure is logical as long as we are interested only in the size of the coefficients rather than in the way the roundoff errors accumulate. Note also that the complete history of $a_{ij}^{(p)}$ gives us information on the partial sums of the product terms $r_{ki}r_{kj}$ because

$$-\sum_{k=q}^{p} r_{ki}r_{kj} = a_{ij}^{(p)} - a_{ij}^{(q-1)}, 0<q\leqslant p\leqslant n. \qquad (7.1)$$

7.1.1 Station situated in interior of lowest level block

Suppose coordinate i belongs to a station that never contributes to any barrier between blocks. Coordinate i then will be eliminated at the lowest level. Before this happens, the coefficients $a_{ij}^{(p)}$ of equation i will undergo changes as p proceeds from 0 to $i-1$. Clearly, no real change will occur as long as p has not yet reached any coordinate that belongs to a station which is either identical with that of coordinate i or that is being connected to that station by a measurement. In section 3.3 we discussed that two stations are connected by a measurement if there is either a distance, azimuth, or direction measured between the two stations, or if the two stations are co-observed directionally from a third station. If p reaches such a coordinate for the first time, some of the coefficients $a_{ij}^{(p)}$ of equation i will change; in particular, the diagonal element $a_{ii}^{(p)}$ will be diminished by a positive amount.

Let us discuss the changes of $a_{ij}^{(p)}$ qualitatively in light of an example. Figure 7.1 shows a lowest level block. The ordering of the interior stations is assumed to be more or less row-wise from top left to bottom right. Stations with previously eliminated coordinates are indicated by black circles. Recall that the two coordinates of a station are always lumped together. For definiteness we can assume that the latitude precedes the longitude. Let us focus on the two equations associated with the station labeled Q. As long as the elimination of the interior nodes has not reached the station labeled P_o, no change occurs to the coefficients in the equations of Q.

Let i refer to the longitude of station Q. Then the coefficient $a_{ii}^{(p)}$ is the reciprocal variance of station Q's longitudinal shift, provided that coordinates i, k, $1 \leqslant k \leqslant p$, are free, while coordinates k, $p<k\leqslant n$, $k \neq i$, are fixed. Note that the fixed coordinates include station Q's latitude. The variance of Q's longitude does not change as long as Q is connected by observations to stations which are all fixed. As the latitude of P_o is eliminated, one of the coordinates to which Q is anchored ceases to be fixed. It becomes elastic, so to speak. As a consequence, station Q's coordinates will

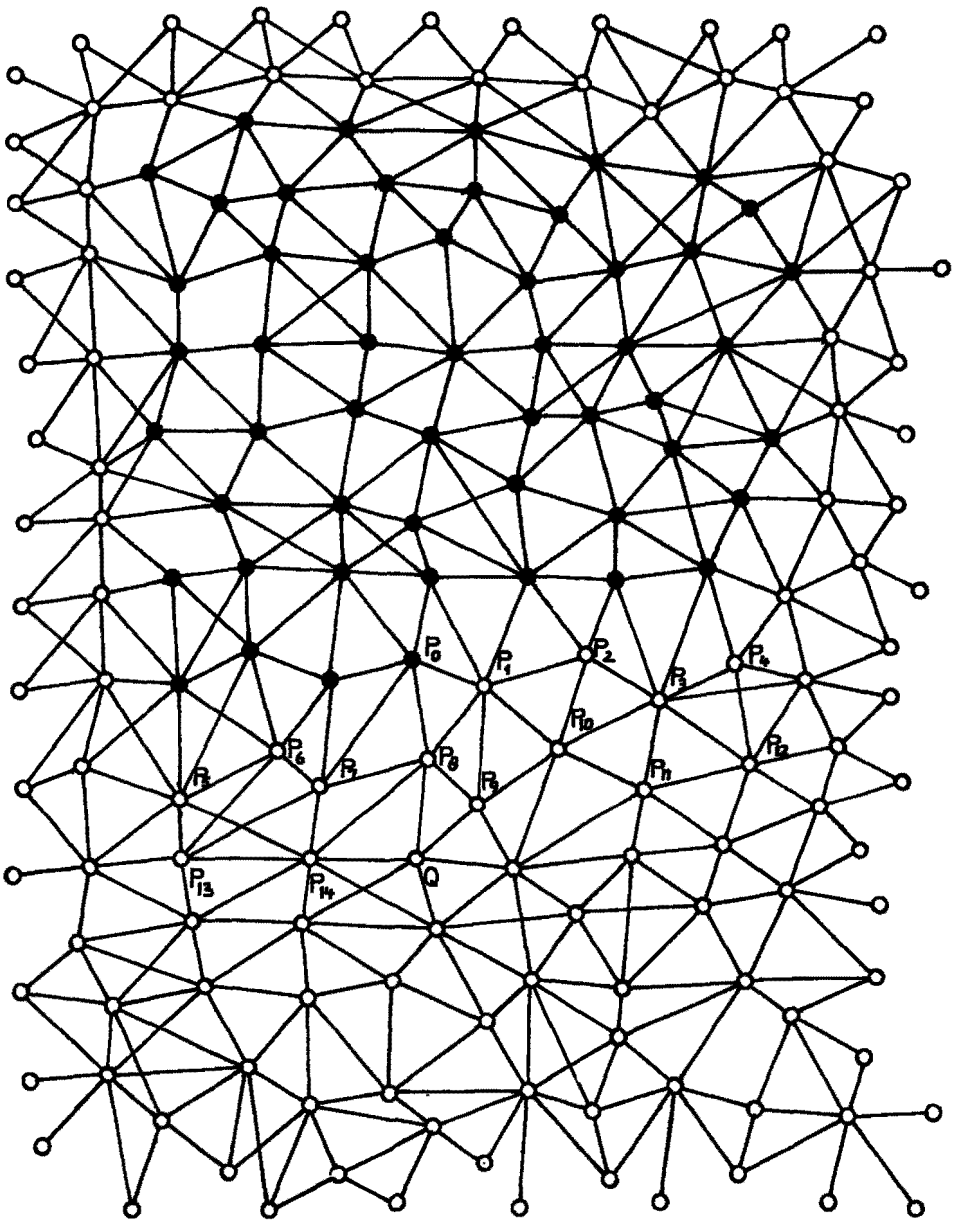

Figure 7.1.—Sample of a lowest level block with about 50 percent of the interior stations eliminated.

have somewhat larger variances. Since $a_{ii}^{(p)}$ is a reciprocal variance, it decreases. Eliminating P_0, or at least one of its coordinates, suddenly causes a number of connections to occur between station Q and other stations, namely stations P_2 through P_{14}. These connections are mostly established by fill-in coefficients. To illustrate this, imagine that P_{11} is shifted away from its adjusted position. The bundle in P_3 will transmit this movement (in general) to all the free black nodes. The bundle in P_8 will transmit this movement to Q which, in this context, is also considered free.

As the coordinates of station P_1 to P_{14} are eliminated, the $a_{ii}^{(p)}$ will change continuously. The changes will be very small if a faraway station like P_{12} is freed. The changes will be most noticeable when the immediate neighbors of Q are treated. After P_{14} is eliminated, *i.e.*, has become a free station, and after Q's latitude has been eliminated, equation i (Q's longitude) is next for pivoting. How much has $a_{ii}^{(p)}$

changed during this process? We consider two extreme cases.

(1) Station Q is not involved in any measurement of high-precision accuracy. In this case Q is connected to its neighbors by measurements of normal accuracy. Even if all the neighboring stations are fixed, as well as the latitude of Q, the longitudinal shift of Q will have an appreciable variance, which may amount to a few centimeters. Throughout this discussion we assume that all coordinate shifts are scaled to the meter. Then it follows that $a_{ij}^{(o)} = a_{ij}$ will be of the order of magnitude of 10^2 to 10^4. The variance of coordinate i will not dramatically change if some of Q's neighboring stations are freed. It will increase, it may double or triple, but it will not be multiplied by several powers of 10. To be conservative, assume that it is multiplied by 10. The coefficient $a_{ii}^{(p)}$ will correspondingly decrease to about 10^1 to 10^3. It follows that $a_{ii} = a_{ii}^{(o)}$ is representative for the size of all $a_{ii}^{(p)}$ and $0 \leqslant p < i$.

(2) Station Q is tied to another station by a very precise measurement. Let us assume that station Q is connected to station P_9 by a distance of 1 mm rms error. The coefficient $a_{ii} = a_{ii}^{(o)}$ then will be of the order 10^6. As long as P_9 is held fixed, the combined effect of the precise distance and the fixed latitude of Q will cause a very small variance of the longitude shift of Q. Hence, $a_{ii}^{(p)}$ will be of the order of magnitude 10^6 as long as station P_9 is untouched. Once the latitude of P_9 is freed, *i.e.*, is made elastic, things change drastically. A sharp drop in the accuracy of coordinate i occurs, and $a_{ii}^{(p)}$ will consequently drop to about 10^2 to 10^4. This is, of course, the most feared wiping out of leading digits. The subsequent elimination steps will not cause any further dramatic change in the size of $a_{ii}^{(p)}$.

What is the consequence of the sharp drop in the size of $a_{ii}^{(p)}$ upon the local roundoff error at position (i,i)? To answer this question we must go back to the manner in which Cholesky's algorithm is actually executed. The transition from $a_{ii}^{(o)}$ to $a_{ii}^{(i-1)}$ is done according to

$$a_{ii}^{(i-1)} = a_{ii}^{(o)} - \sum_{k=1}^{i-1} r_{ki} r_{ki}.$$

The sum is evaluated and then subtracted from $a_{ii}^{(o)}$. Let p correspond to the latitude of P_9. Then the terms $r_{ki}r_{ki}$, $1 \leqslant k < p$, will be small. The elementary roundoff errors arising from computing and summing the product terms will provide only small contributions to the local roundoff error affecting a_{ii}. The term $r_{ip}r_{ip}$ will be large. The elementary roundoff errors in evaluating and adding it will be large. Moreover, by now the partial sum of the product terms has become large. This means that all further modifications to it, caused by the elimination of stations P_{10} to P_{14}, will produce large contributions to the local roundoff error at a_{ii}. Finally, the sum is subtracted from $a_{ii}^{(o)}$, and this will cause another large elementary roundoff error.

To complete the story, we must consider the latitude of Q as well as the latitude and longitude of P_9. It is preferable to switch the indices of the equations. We let i refer to the latitude of P_9; $i+1$ consequently refers to its longitude, j, and $j+1$ refer to the latitude and longitude of Q. (See fig. 7.2.) All original coefficients a_{ii}, $a_{i,i+1}$, a_{ij}, $a_{i,j+1}$, $a_{i+1,i+1}$, $a_{i+1,j}$, $a_{i+1,j+1}$, a_{jj}, $a_{j,j+1}$, $a_{j+1,j+1}$, will be large unless, by coincidence, P_9 and Q happen to have either nearly the same latitude or the same longitude. The coefficients remain large until i is eliminated. Then they all drop sharply in size. What about the associated roundoff errors? Assuming that there are no further high precision observations in the near vicinity, and applying similar reasoning as in the preceding paragraph, we arrive at the following statements:

- All product terms $r_{ki}r_{ki}$ will be of moderate or small size. The first bad elementary roundoff error arises when the sum of these product terms is subtracted from $a_{ii}^{(o)}$ to give $a_{ii}^{(i-1)}$. Note that $a_{ii}^{(i-1)}$ is still large. At this point no leading digits have been wiped out. It is an oversimplification to associate bad roundoff errors generally with a loss of leading digits. A second bad roundoff error occurs when the square root is taken of the large coefficient $a_{ii}^{(i-1)}$.

- A similar statement can be made about the coefficients $a_{i,i+1}$, a_{ij}, $a_{i,j+1}$. The product sums subtracted from these large coefficients will be moderate, causing only two bad roundoff errors per coefficient; when the product sum is subtracted from the large coefficient and when division by $r_{ii} = \sqrt{a_{ii}^{(i-1)}}$ takes place.

- Of all product terms $r_{k,i+1}r_{k,i+1}$, whose sum will be subtracted from $a_{i+1,i+1}^{(o)}$, only the last one, *i.e.* $r_{i,i+1}r_{i,i+1}$, will be large. Hence, three bad elementary roundoff errors will occur: when $r_{i,i+1}r_{i,i+1}$ is evaluated, when this term is added to the previously accumulated partial sum, and when the sum is subtracted from $a_{i+1,i+1}^{(o)}$. A similar statement can be made about the coefficients $a_{i+1,j}$ and $a_{i+1,j+1}$. No bad roundoff errors are caused by taking the square root and by division. At that time the coefficients have already dropped in size.

- Again there will be only one large outlier among the product terms $r_{kj}r_{kj}$ whose sum is to be subtracted

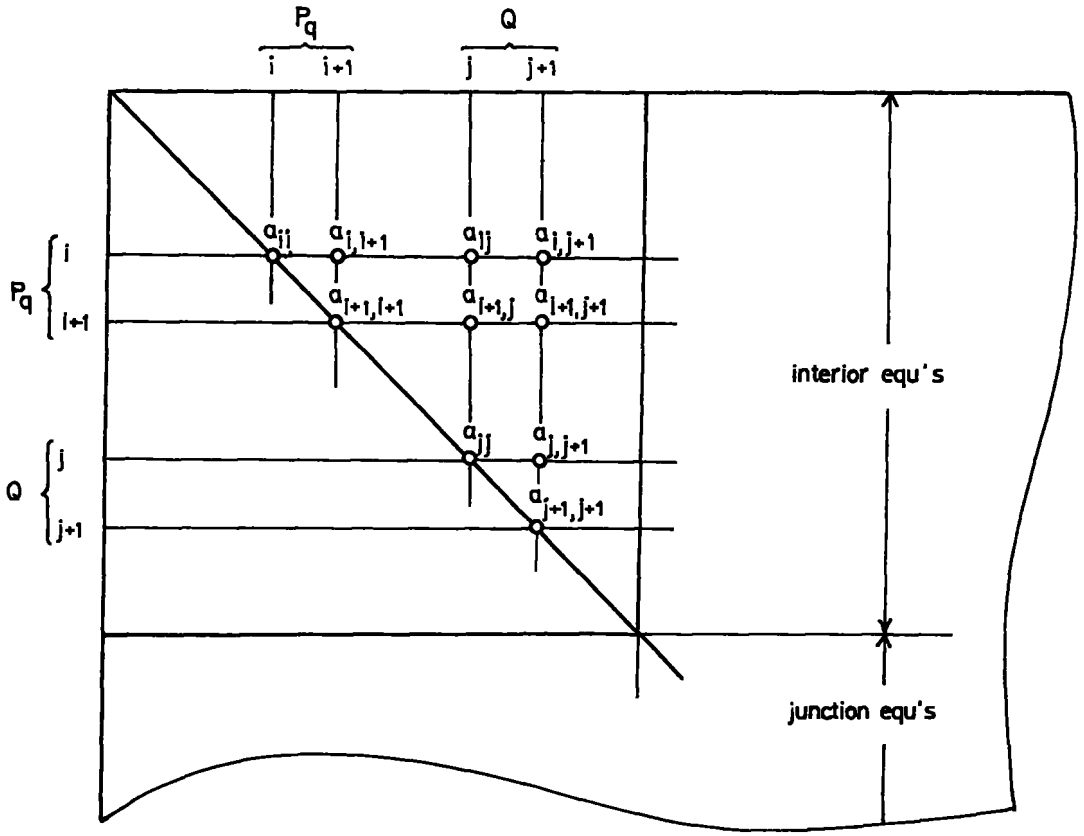

Figure 7.2.—Left-side normal equation coefficients of two connected stations, as discussed in the text.

from $a_{jj}^{(o)}$. This time, however, the outlier is not the last term. It is of course, $r_{ij}r_{ij}$, which is followed by the moderate terms $r_{kj}r_{kj}$, $i<k<j$. By being added to a large sum, these moderate terms will all cause bad elementary roundoff errors. They, in turn, will contribute unfavorably to unavoidable roundoff errors, namely those associated with $r_{ij}r_{ij}$ and with the subtraction of the product sum from $a_{jj}^{(o)}$. A similar statement can be made about the coefficients at $a_{i,j+1}$, $a_{j+1,j+1}$.

The following lesson can be learned from this experience: It is extremely disadvantageous if two stations which are strongly coupled by a very precise measurement are not assigned consecutive numbers in the ordering scheme. The more stations included between the two stations, the more product terms of the type $r_{kj}r_{kj}$, $r_{kj}r_{k,j+1}$, $r_{k,j+1}r_{k,j+1}$ will be added to an already large partial sum and, consequently, the more bad elementary roundoff errors will occur.

Remark: If one takes into account a slight complication of Cholesky's algorithm, the number of bad elementary roundoff errors can be reduced to 2 per coefficient a_{ij}. We have seen that when a station is involved in one very precise measurement, essentially only one product term $r_{ki}r_{kj}$ will be large and nearly equal to $a_{ij}^{(o)}$. Let this product term be $r_{pi}r_{pj}$. If we calculate

$$a_{ij}^{(i-1)} = a_{ij}^{(o)} - \sum_{k=1}^{i-1} r_{ki}r_{kj}$$

as (7.2)

$$a_{ij}^{(i-1)} = \{a_{ij}^{(o)} - r_{pi}r_{pj}\} - \{\sum_{\substack{k=1\\k\neq p}}^{i-1} r_{ki}\, r_{kj}\},$$

the operations being carried out, as indicated by the braces, then only two bad elementary roundoff errors occur: one when $r_{pi}r_{pj}$ is evaluated and the other when this term is subtracted from $a_{ij}^{(o)}$. The result of this subtraction will be a moderately sized quantity, and so will be all further operands. Modifying Cholesky's algorithm in the indicated way does not require much programing effort. The main objection may be that speed suffers greatly. The inner product evaluations represent the innermost loop. If a test for size is placed into the inner loop, speed may decrease as much as 30 to 50 percent. However, a slight additional complication could restore much of the origi-

nal speed. A test for size is necessary for only those coefficients a_{ij} that were originally very large. The majority of stations will not be involved in very precise measurements. The coefficients in the corresponding rows and/or columns of the normal equations will be small from the beginning. Hence the test for size of the product terms can be bypassed when these coefficients are treated. It may be worthwhile after all, to consider the following modification for partial reduction of the NGS Cholesky algorithm (which can be compared with eq. (3.15) in section 3.2):

```
  FOR j = 1 TO n+1
    FOR i = 1 TO MIN(n,j)
      SUM = 0
      IF both A(j,j) and A(i,i) are large, GOTO
        label (L1)
      FOR k = 1 TO MIN(p,i-1)
        SUM = SUM + A(k,i) * A(k,j)
      NEXT k
      GOTO label (L2)
(L1): SUM1 = 0
      FOR k = 1 TO MIN(p,i-1)                    (7.3)
        HELP = A(k,i) * A(k,j)
        IF (HELP is small) THEN SUM = SUM +
          HELP;
        ELSE SUM1 = SUM1 + HELP
      NEXT k
      A(i,j) = A(i,j) - SUM1
(L2): A(i,j) = A(i,j) - SUM
      IF (i ≤ MIN(j-1,p)) A(i,j) = A(i,j)/A(i,i)
    NEXT i
    IF (j ≤ p) A(j,j) = √A(j,j)
  NEXT j
```

This algorithm also takes care of those situations where a station is involved in more than one very precise measurement. Let us briefly review a case where coordinate i belongs to a station that takes part in a cluster of m tightly connected stations. (See figs 7.3, 7.4.) In equation i there may be more large coefficients, 2m in the worst case, when i precedes all other coordinates of the cluster in the sequence of elimination. If i is the last coordinate in this sequence, then only one large coefficient will be present. The moment when all coefficients are either past elimination or have dropped to moderate size may not occur until a coordinate of the last two remaining stations is eliminated. This is true when the cluster forms a rigid system. The situation may be slightly more favorable if the cluster consists of several rigid limbs which are pin-jointed. The modification of Cholesky's algorithm will keep the number of bad roundings down to a number which is approximately equal to $8\ m^3/3$.

7.1.2 Station situated on a block barrier

Consider the normals of a certain block for which station Q plays the role of a junction station. The coefficients in the equations of its two coordinates are not all fully assembled. Coefficients connecting station Q to interior stations of the block are fully assembled. Other coefficients may be incomplete. In any case, the diagonal coefficients are incomplete. If the block is not one of the lowest levels, the coefficients will be partially reduced. The reader is referred to section 3.5.5, which describes the interplay of assembling and reducing the normals typical in Helmert blocking.

After combining the blocks properly, station Q will eventually become an interior station of a certain higher level block. At that instant all coefficients in the Q equations will be fully assembled and at the same time partially reduced.

If station Q is not involved in a high precision measurement, the history of the coefficients $a_{ij}^{(p)}$ in the equations of its coordinates is not very dramatic. Of course, the history is no longer represented by a linear row of coefficients but, instead, by a tree branching out toward the past. This is because the station may be actively involved (*i.e.* with nonzero coefficients) in more than one block of a certain level. The tree does not have too many branches, though, because in nested dissection as shown by figure 3.6, a station will never be actively involved in more than four blocks of the same level. In fact, a regular node, situated in a level i set, will be involved in two sets within any of the levels 1 to $i-1$. (See section 4.3 and fig. 4.1.)

The diagonal elements will not undergo any large changes. Just before station Q is eliminated, these coefficients will not be much smaller than, say, 1/10 of the size of the full original normals. The size of the full original normals will be 10^2 to 10^4, just as in the case of a nonbarrier station considered in the previous section. At present, we are not very concerned with the off-diagonal elements. Because any 2×2 determinant of the normals in any reduction state must be positive, an off-diagonal coefficient will never exceed the harmonic mean of the two diagonal coefficients located in its row and column. Unfortunately this simple estimate is not of much use because it leads essentially to the unfavorable preliminary global roundoff error bounds of section 4.1.4. In section 7.1.3 we specify more useful estimates of the off-diagonal coefficients.

Meanwhile let us qualitatively discuss the case where a station Q, usually located at some barrier, is connected by a high precision measurement to a station P. Consider the partial reduction of a block in which Q serves as a junction station. It is necessary

for the block to be from a lower level than Q itself, which we denote by k. Let i, $i+1$, j, $j+1$ denote the sequence numbers of the coordinates of the stations P, Q, respectively. As long as P, Q are not eliminated, the coefficients $a_{ii}^{(p)}$, $a_{i,i+1}^{(p)}$, $a_{ij}^{(p)}$, $a_{i,j+1}^{(p)}$, $a_{i+1,i+1}^{(p)}$, $a_{i+1,j}^{(p)}$, $a_{i+1,j+1}^{(p)}$, $a_{jj}^{(p)}$, $a_{j,j+1}^{(p)}$, $a_{j+1,j+1}^{(p)}$ are all large. (Again we disregard the degenerate case where P and Q have nearly the same latitude or longitude). We discuss several subcases:

(1) Station P is an interior station. Because of the assumed regularity of Q, P must be of the lowest level. Reasoning as in the previous section, we arrive at the following conclusions:

(*) The coefficients $a_{ii}^{(p)}$, $a_{i,i+1}^{(p)}$, $a_{ij}^{(p)}$, $a_{i,j+1}^{(p)}$ never drop sharply in size before equation i is selected for pivoting.

(*) Coefficients $a_{i+1,i+1}^{(p)}$, $a_{i+1,j}^{(p)}$, $a_{i+1,j+1}^{(p)}$, $a_{jj}^{(p)}$, $a_{j,j+1}^{(p)}$, $a_{j+1,j+1}^{(p)}$ drop sharply when equation i is eliminated.

(*) If the modified algorithm for partial Cholesky reduction is used (specified in the previous section), no more than two bad roundoff errors can occur for any large coefficient.

(*) If the standard version of the algorithm is used, two bad roundoff errors will occur at a_{ii}, $a_{i,i+1}$, a_{ij}, $a_{i,j+1}$, and three bad roundoff errors at $a_{i+1,i+1}$, $a_{i+1,j}$, $a_{i+1,j+1}$. Many bad roundings can occur at any of the coefficients a_{jj}, $a_{j,j+1}$, $a_{j+1,j+1}$ if many interior stations are located subsequent to P in the ordering sequence of the current block.

(2) Q is tightly connected to another station P which is in the same barrier set. P is also assumed to be regular. Although P and Q are involved in several level l block reductions ($l<k$) the strong connection, which is caused by an observation, is always uniquely assigned to one block of a certain level. Hence, for a prespecified block level l ($l<k$) there will be only one block in which the a_{ii} to $a_{j+1,j+1}$ are large. On the other hand, such a block can be encountered for levels l, with l proceeding from 1 to $k-1$. Hence, altogether, there can be $k-1$ blocks where these coefficients are large. It follows that:

(*) There will be one bad rounding per large coefficient for any level l block in which the coefficients are large. Therefore, as long as level k is not reached, *i.e.*, as long as P, Q are junction nodes, there will be $k-1$ bad roundings.

(*) If the two stations are regular junction stations (see section 4.3.1), there will be one additional bad rounding per large coefficient at the moment when the k-level block is assembled from its $k-1$ level subblocks. If the stations are singular, more bad roundings can be expected, but their numbers would hardly exceed three.

(*) At level k, P and Q become interior nodes and essentially the same reasoning applies as for interior nodes at the lowest level. Quantitatively, one has to bear in mind that many connections exist between the stations, resulting from fill-in at the lower levels.

The following can be learned from the above statements: If the standard version of Cholesky's algorithm is used, avoid if possible junction stations that are strongly coupled. The modification takes good care of couplings between barrier stations and nonbarrier stations. Strong ties between barrier stations at high levels should be avoided under all circumstances.

7.1.3 Bounds on off-Diagonal elements

The preliminary bounds derived in section 4.1.4 on the global left-hand side roundoff errors were bad for several reasons. One reason is that a single bound $\|a\|$ was used on all nonzero elements of the normal equation matrix and its reduction states. This is a luxury one cannot afford. Even if the U.S. network did not have the much lamented high precision ties between closely situated stations, and even if it were fairly homogeneous with respect to configuration and weights, the coefficients of the partially reduced normals would differ greatly in size. Recall that $a_{ij}^{(p)}/a_{ii}^{(p)}$ is the negative shift that coordinate i suffers if coordinate j is forced away by one unit from its adjusted position, if coordinates i, k, $1 \leqslant k \leqslant p$, are free, while coordinates k, $p<k\leqslant n$, $k \neq i,j$, are fixed. The farther away coordinate j is from coordinate i, the smaller the movement of i will be, in general. Hence, one should expect that the coefficients $a_{ij}^{(p)}$, $i<j$, taper off as the distance between the two involved stations increases.

Loosely speaking, our global estimates on the left-hand side roundoff errors are weighted sums over either absolute values or squares of the coefficients $a_{ij}^{(p)}$. Although this may not be apparent from the formulas at the end of section 4.1.3, which appear to be weighted sums over $E\{\varepsilon_{ij}\}$, $\sigma^2\{\varepsilon_{ij}\}$ we draw support for our assertion because $E\{\varepsilon_{ij}\}$, $\sigma^2\{\varepsilon_{ij}\}$ are closely related to the size of the coefficients $a_{ij}^{(p)}$. We will clarify this in section 9.4. Meanwhile, we should feel sufficiently motivated to take a closer look at the following norms of the upper diagonal portions of the rows of $A_{22}^{(p)}$, the partially reduced normal equation matrix for the coordinates i, $p<i\leqslant n$:

$$\|a_{i\cdot}^{(p)}\|_1 = \sum_{j\geq i} |a_{ij}^{(p)}| \tag{7.4}$$

$$\|a_{i\cdot}^{(p)}\|_2 = \sqrt{\sum_{j\geq i} |a_{ij}^{(p)}|^2}. \tag{7.5}$$

We will try to derive bounds for these norms which rely on: (1) the positive definiteness of $A_{22}^{(p)}$, (2) the size of the diagonal element $a_{ii}^{(p)}$, and (3) the spectral norm of the original normal equation matrix.

Item (3) deserves some explanation. The spectral norm of the original normal equation matrix A equals the largest eigenvalue λ_{Max} of A. Of course, no one will go to the trouble of calculating this eigenvalue exactly. Good bounds on λ_{Max} are easily available. One of them is the row-sum norm of A. This is the largest row-wise sum of the absolute elements of A. Because A is sparse in the sense that $a_{ij} \neq 0$ only for coordinates which are connected by measurement, the row-sum norm of A will not differ too much from its largest element. It can differ from it by a factor amounting to the largest number of nonzero elements in a row.

We next introduce $\varkappa_i^{(p)}$, the number of nonzero elements $a_{ij}^{(p)}$, $j\geq i$, to the right of, and including, the diagonal element $a_{ii}^{(p)}$. We are now ready to prove the following:

Proposition 7.1.

$$\|a_{i\cdot}^{(p)}\|_1 \leq \sqrt{\varkappa_i^{(p)}\, a_{ii}^{(p)}\, \lambda_{Max}} \tag{7.6}$$

$$\|a_{i\cdot}^{(p)}\|_2 \leq \sqrt{a_{ii}^{(p)}\, \lambda_{Max}}. \tag{7.7}$$

Proof: The proof relies heavily on Schwarz's inequality which, for a positive definite matrix M and two matching vectors x, y, can be specified as follows:

$$x^T M y \leq \sqrt{x^T M x\, y^T M y}. \tag{7.8}$$

To extract the i-th row of $A_{22}^{(p)}$, we introduce the vector e which has a 1 at position i and zeroes elsewhere. In order to form the norm $\|a_{i\cdot}^{(p)}\|_1$ from the upper diagonal elements in row i, we introduce a vector w which has elements

$$w_j = 0, \qquad j<i$$

$$w_j = \operatorname{sign}(a_{ij}^{(p)}),\ i\leq j\leq n.$$

It would be logical to attach the index i to e and w, but we refrain from doing so to avoid bulky formulas. We now write

$$\|a_{i\cdot}^{(p)}\|_1 \leq \sum_{j\geq i} |a_{ij}^{(p)}| = \sum_{j\geq i} a_{ij}^{(p)} \operatorname{sign}(a_{ij}^{(p)}) = \sum_j a_{ij}^{(p)}\, w_j.$$

Hence,

$$\|a_{i\cdot}^{(p)}\|_1 = e^T A_{22}^{(p)}\, w.$$

Applying Schwarz's inequality with $M = A_{22}^{(p)}$, $x = e$, $y = w$, we deduce

$$\|a_{i\cdot}^{(p)}\|_1 \leq \sqrt{\{e^T A_{22}^{(p)} e\}\{w^T A_{22}^{(p)} w\}} =$$

$$= \sqrt{a_{ii}^{(p)}\{w^T A_{22}^{(p)} w\}}.$$

Since $A_{22} - A_{22}^{(p)}$ equals $A_{21}A_{11}^{-1}A_{12}$, and therefore is positive semidefinite, it follows that the largest eigenvalue of $A_{22}^{(p)}$ is bounded by the largest eigenvalue of A_{22}. The largest eigenvalue of A_{22}, in turn, is bounded by the largest eigenvalue λ_{Max} of A. It follows that $w^T A_{22}^{(p)} w \leq \lambda_{Max} \|w\|^2$. Noting that $\|w\|^2 = \varkappa_i^{(p)}$, we have established the first part of the proposition.

As for the second part, we start with

$$\|a_{i\cdot}^{(p)}\|_2^2 = \sum_{j\geq i} |a_{ij}^{(p)}|^2 \leq \sum_j |a_{ij}^{(p)}|^2 = \{A_{22}^{(p)} e\}^T \{A_{22}^{(p)} e\} =$$

$$= e^T A_{22}^{(p)} A_{22}^{(p)} e$$

Again we apply Schwarz's inequality, but this time with $M = A_{22}^{(p)}$, $x = e$, $y = A_{22}^{(p)} e$. The result is:

$$e^T A_{22}^{(p)} A_{22}^{(p)} e \leq \sqrt{\{e^T A_{22}^{(p)} e\}\{e^T A_{22}^{(p)} A_{22}^{(p)} A_{22}^{(p)} e\}}.$$

The first term under the square root is $a_{ii}^{(p)}$, the second one is bounded by $\lambda_{Max}\, e^T A_{22}^{(p)} A_{22}^{(p)} e$. Dividing the whole equation by $\sqrt{e^T A_{22}^{(p)} A_{22}^{(p)} e}$ yields the desired result.

Remark: Leveling networks have the remarkable property of diagonal dominance and nonpositive off-diagonal elements. It holds that

$$\sum_{p<j\leq n} a_{ij}^{(p)} \geq 0 \text{ for } p<i\leq n,\ a_{ij}^{(p)} \leq 0 \text{ for } j\neq i. \tag{7.9}$$

As these equations indicate, the properties hold not only the original normals but also for any set of partially reduced normals. Combined with the obvious fact that $a_{ii}^{(p)} > 0$, one arrives readily at the following upper bounds for $\|a_{i\cdot}^{(p)}\|_1$ and $\|a_{i\cdot}^{(p)}\|_2$, which are much better than those of proposition 7.1. (See Bartelme and Meissl 1977).

$$\|a_{i\cdot}^{(p)}\|_1 \leq 2\, a_{ii}^{(p)},\ \|a_{i\cdot}^{(p)}\|_2 \leq \sqrt{2}\, a_{ii}^{(p)}. \tag{7.10}$$

I strongly believe for reasons explained in section 9.4 that the U.S. network behaves in many ways like a leveling network. In particular, when the interior stations of a block are eliminated while the junction nodes form an impenetrable barrier, the behavior of the $a_{ij}^{(p)}$ is believed to closely resemble a leveling network. Hence I believe that in most instances proposition 7.1 overestimates the row-sum norms $\|a_{i\cdot}^{(p)}\|_1$, $\|a_{i\cdot}^{(p)}\|_2$. On the other hand, it is very difficult to make quantitative statements about how close the U.S. network comes to a leveling net. We will speculate

more on this in chapter 9. First we need to aim at safe estimates which are based on firm grounds.

If we apply the bounds of proposition 7.1 to the U.S. network with all its local weight singularities, we obtain only a marginal improvement over the earlier estimates in section 4.1.4. Therefore we will proceed along a different route and consider a hypothetical U.S. network which has the local weight singularities removed. Afterwards, we will consider modifications of the derived estimates resulting from weight singularities. These modifications will rely on theoretical considerations undertaken in the next subsection.

7.1.4 Transforming away the weight singularities

We do not anticipate that the transformations discussed in this section will actually be applied to the network during the process of adjustment. This would cause complications in the computer programs as well as much additional manual labor. The transformations are merely supposed to provide theoretical support for the prediction of the increased global roundoff resulting from the high-precision ties between closely situated stations.

Let us try to make things clear by assuming that the network has only one exceptionally accurate observation, which is a distance between the neighboring stations P,Q. Denoting, as usual, the sequence numbers of the coordinates by i, $i+1$, j, $j+1$, we first consider the normal equations for this distance observation alone. The coefficients are given in table 7.1.

TABLE 7.1.—*Coefficients of contributions to the normals from one distance observation before transformation.*

row	i	$i+1$	j	$j+1$
i	pcc	pcs	$-pcc$	$-pcs$
$i+1$	pcs	pss	$-pcs$	$-pss$
j	$-pcc$	$-pcs$	pcc	pcs
$j+1$	$-pcs$	$-pss$	pcs	pss

In table 7.1, c,s denote cosine and sine of the azimuth from P to Q; p is the weight of the distance which is assumed to be large, *i.e.*, about 10^6. Actually, these coefficients refer to a plane network. The modifications to the ellipsoid are quantitatively irrelevant in the present context, if we assume that latitude and longitude increments are scaled to the meter.

Next, imagine the normals formed for the other observations of moderate weights. The coefficients will be of the order of 10^2 to 10^4. The combined normals are obtained by addition. At the intersections of rows and columns i, $i+1$, j, $j+1$, small disturbances will be added to the large coefficients of table 7.1. At all other locations the coefficients will be small.

Let us take a look at the global effect of the 10 local roundoff errors ε_{ii}, $\varepsilon_{i,i+1}$, ε_{ij}, $\varepsilon_{i,j+1}$, $\varepsilon_{i+1,i+1}$, $\varepsilon_{i+1,j}$, $\varepsilon_{i+1,j+1}$, ε_{jj}, $\varepsilon_{j,j+1}$, $\varepsilon_{j+1,j+1}$. Since the corresponding coefficients are of magnitude 10^6, the local roundoff errors will be in the range at least $\beta^{-\tau}*10^6$. Their global effect upon coordinate k is obtained from formula (4.34) in section 4.1.3:

$$\begin{aligned}\zeta_k = {} & f_{ki}x_i\varepsilon_{ii} + (f_{ki}x_{i+1} + f_{k,i+1}x_i)\varepsilon_{i,i+1} + (f_{ki}x_j + \\ & + f_{kj}x_i)\varepsilon_{ij} + (f_{ki}x_{j+1} + f_{k,j+1}x_i)\varepsilon_{i,j+1} + \\ & + f_{k,i+1}x_{i+1}\varepsilon_{i+1,i+1} + (f_{k,i+1}x_j + \\ & + f_{kj}x_{i+1})\varepsilon_{i+1,j} + \ldots + f_{kj}x_j\varepsilon_{jj}\end{aligned} \tag{7.11}$$

Coordinates i, $i+1$, j, $j+1$ are nearby because they are connected by the precise distance. If coordinate k is also near, the elements of the inverse f_{ki}, $f_{k,i+1}$, f_{kj}, $f_{k,j+1}$ will be relatively large, with magnitude 10^{-1}. If we use $\|x\|$ as the bound for the coordinate shifts ($\|x\|$ may be as large as 10 m), we see that ζ_k will be of magnitude at least $\beta^{-\tau}\,10^5\,\|x\|$.

We shall now transform away the weight singularities by a parameter transformation. The transformation will be local; only the coordinates of the two stations P,Q will be involved. Originally we have $x_i = \Delta\xi_P$, $x_{i+1} = \Delta\eta_P$, $x_j = \Delta\xi_Q$, $x_{j+1} = \Delta\eta_Q$, where ξ_P, η_P, ξ_Q, η_Q are coordinates of the stations P and Q. We replace ξ_Q, η_Q by polar coordinates ϱ_{PQ}, ϕ_{PQ} of Q with respect to P. We have:

$$\begin{aligned}\xi_Q &= \xi_P + \varrho_{PQ}\cos\phi_{PQ} \\ \eta_Q &= \eta_P + \varrho_{PQ}\sin\phi_{PQ}.\end{aligned} \tag{7.12}$$

Linearizing and omitting the subscript PQ we get:

$$\begin{aligned}\Delta\xi_Q &= \Delta\xi_P + c\,\Delta\varrho - \varrho\, s\,\Delta\phi \\ \Delta\eta_Q &= \Delta\eta_P + s\,\Delta\varrho + \varrho\, c\,\Delta\phi\end{aligned} \tag{7.13}$$

c and s are, of course, cosine and sine of $\phi = \phi_{PQ}$. Their meaning is the same as in table 7.1.

If we transform only the normal equations due to the precise distance, the pattern of coefficients shown in table 7.1 changes over into one given in table 7.2. Only one coefficient is not zero. This is no surprise because the observation refers precisely to the new parameter $\Delta\varrho_{PQ}$ and tells us nothing about the other three.

The normals of the other observations will now be subjected to the same transformation. This will not cause the small coefficients around P,Q to increase significantly in size. After adding the two sets of normals we will have one large coefficient at a_{pp}. We will

show that the global effect of the local roundoff error ε_{pp} at this coefficient is negligibly small, even though it is of the same magnitude, $\beta^{-\tau} * 10^6$, as the earlier ε_{ii} to ε_{jj}.

First, during the triangularization process, we observe that the single large coefficient will not cause any other coefficient to become large. The large diagonal coefficient is never subtracted from anything. This is an important observation which will be exploited below. Meanwhile, we accept the fact that $a_{pp}^{(p)}$ undergoes changes during the elimination of the preceding equations. It will slightly decrease in size and, as indicated, a local roundoff error ε_{pp} of magnitude $\beta^{-\tau} * 10^6$ must be taken into account. This is a slight improvement over the earlier situation; we must now deal with only one large local roundoff error instead of 10. Its global effect is

$$\zeta_k = f_{kp}\, x_p\, \varepsilon_{pp} \tag{7.14}$$

The most substantial improvement comes from the fact that f_{kp}, x_p are extremely small. The change $x_p = \Delta Q_{PQ}$ of the precise distance will hardly surpass 3 mm. This compares with about 10 m of the coordinate shifts and accounts for an improvement of 10^{-3}. The coefficients of the inverse f_{kp} will be shown to be of magnitude $1/p \doteq 10^{-6}$. This is very clear for $k = \varrho$. The f_{pp} is the variance of the adjusted distance between P and Q and this a posteriori variance is certainly not larger than the a priori variance 10^{-6} of the measured distance. For $k \neq \varrho$ the argument is more complicated. Consider the 2×2 matrix

$$\begin{bmatrix} f_{kk} & f_{kp} \\ f_{pk} & f_{pp} \end{bmatrix}$$

TABLE 7.2.—*Coefficients of the normal contributions after transformation.*

row	i	$i+1$	ϱ	ϕ
i	0	0	0	0
$i+1$	0	0	0	0
ϱ	0	0	p	0
ϕ	0	0	0	0

This is a 2×2 submatrix of the inverse referring to the two parameters k,ϱ. Under the hypothetical assumption that these two parameters are the last to be eliminated, the inverse of the 2×2 submatrix would be

$$\begin{bmatrix} a_{kk}^{(k-1)} & a_{kp}^{(k-1)} \\ a_{pk}^{(k-1)} & a_{pp}^{(k-1)} \end{bmatrix} = \frac{1}{f_{kk}f_{pp}-f_{kp}^2} \begin{bmatrix} f_{pp} & -f_{kp} \\ -f_{pk} & f_{kk} \end{bmatrix}. \tag{7.15}$$

This is the partially reduced normal equation matrix for x_k and ΔQ_{PQ} after all other parameters have been eliminated. Hence $a_{pp}^{(k-1)}$ is the reciprocal variance of the distance P-Q when coordinate k is held fixed. Because we do not expect fixing coordinate k to contribute significantly to the accuracy of the adjusted distance P-Q, $a_{pp}^{(k-1)}$ will be about 10^6. $a_{kk}^{(k-1)}$ is the reciprocal variance of coordinate k when the distance P-Q is fixed. Fixing the distance will not prevent the net from floating in position; therefore, we expect the variance to be about 0.1. Hence $a_{kk}^{(k-1)}$ is about 10. $a_{kp}^{(k-1)}/a_{kk}^{(k-1)}$ is the displacement of coordinate k when the distance P-Q is expanded by one unit from its adjusted length. We expect that this displacement is not larger than 1. Hence $a_{kp}^{(k-1)} \doteq 10$. The equation

$$a_{kp}^{(k-1)} = \frac{-f_{kp}}{f_{kk}f_{pp}-f_{kp}^2} \tag{7.16}$$

shows that f_{kp} must indeed be of the order 10^{-6} Q.E.D.

It follows that the global roundoff error due to ε_{pp} is of magnitude $\beta^{-\tau} * 10^2$. It is smaller than the global roundoff error caused by any other ε_{ii}. The beneficial effect of our transformation becomes obvious. The procedure of transforming away weight singularities is readily generalized to more complicated situations. This will be demonstrated by two examples:

(1) A linked chain of very precise distances. Suppose that three distances P-Q, Q-R, R-S, have been measured with very small observation error. The other observations in the vicinity of station P, Q, R are of moderate accuracy. (See fig. 7.3.) An appropriate set of substitute parameters would be: The two coordinates of Q, the polar coordinates of P,R with respect to Q, and the polar coordinates of S with respect to R. The transformed normal equation matrix would have three large diagonal elements corresponding to the three precise distances. There would be no large off-diagonal elements. Global roundoff errors would not be influenced heavily by the local errors at the large elements.

(2) A cluster of stations with very accurate, and possibly redundant, mutual ties. Suppose that stations P, Q, R, S, T are tied together by a set of measurements of very high accuracy. The accurate measurements may even be redundant among themselves, in the sense that four sides plus two diagonals in the quadrilateral P, Q, R, S, are of high accuracy. (See fig. 7.4.) An appropriate way to transform away the weight singularities would be the following. Introduce a local coordinate system with origin in R, and with local ξ-axis (line R-P), local η-coordinate of P (distance R-P), local ξ,η-coordinates Q, S, T. In the

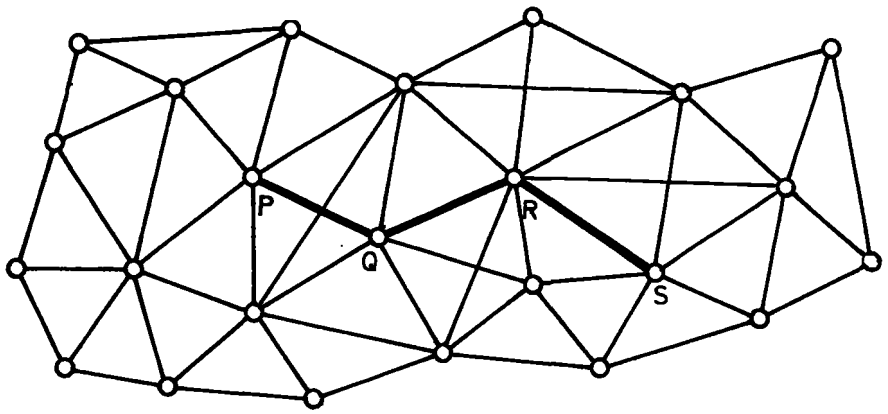

Figure 7.3.—A chain of tightly connected stations.

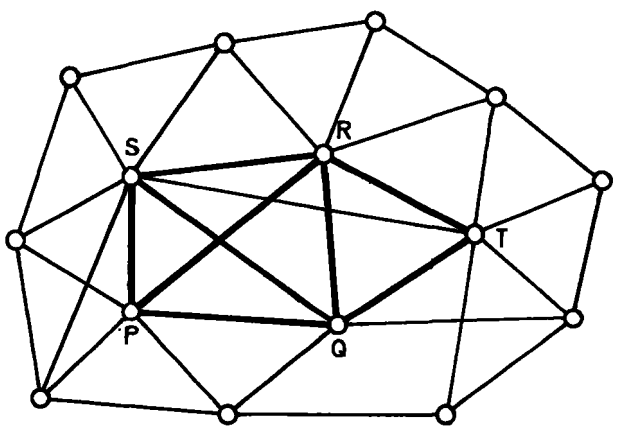

Figure 7.4.—A nearly rigid substructure of tightly connected stations.

vicinity of the five stations, the new normal equation matrix will have large elements at the intersection of rows and columns belonging to the local coordinates. There will also be large off-diagonal elements. This new feature is due to the redundancy among the very precise measurements. The inverse of the normal equation matrix will have very small elements in rows and in columns belonging to the local coordinates. This, together with the smallness of the increments to the local parameters, will effectively keep down any adverse global effect of the local roundoff errors at the large elements of the normals.

Remark. It is really not too disturbing that redundancies among strong ties between clustered stations also cause some off-diagonal elements of the normals, namely those coupling the local coordinates, to be large (even after transformation.) By an additional effort, one could get rid of these large coefficients. The procedure would be to transform the submatrix of the normals (referring to the local coordinates) to a system of eigenvectors. (It would be sufficient to do this for the normals of the precise measurements only!) Such a transformation would replace the parameters $\Delta\eta_P$, $\Delta\xi_R$, $\Delta\eta_R$, $\Delta\xi_S$, $\Delta\eta_S$, $\Delta\xi_T$, $\Delta\eta_T$ by linear functions of these parameters. The parameters $\Delta\xi_Q$, $\Delta\eta_Q$, $\Delta\phi_{QP}$ would be unaffected. The resulting normal equation matrix would have large diagonal elements at the locations of the new local parameters only. The analogy to the previous case would be complete. Such an additional spectral transformation does not offer an additional numerical advantage. Hence, even if local transformations would be used in practice, it is not necessary to avoid large off-diagonals under all circumstances. The crucial requirement is merely that any large local submatrix of the normal equation matrix A has a small inverse and small elements are included in all rows and columns of A^{-1} that touch the submatrix.

The outlined procedure to transform away weight singularities is theoretically straighforward and unsophisticated, but its practical application is cumbersome because it does not lend itself easily to automated treatment. For a configuration of strongly tied stations, a careful human inspection and a choice of substitute parameters appears to be inevitable. Hence it is not anticipated that the procedure will be applied to the U.S. network. Why then have we gone to so much trouble to present it here? Because the possibility of transforming away weight singularities by *local* changes of parameters gives us important insight into the behavior of the large coefficients during triangular decomposition. Without this insight we could not be sure that the few large elements in the original normals would not multiply during the elimination procedure.

Suppose that $x = Vy$ is a parameter transforma-

tion that removes the weight singularities for a set of strongly coupled parameters. V will be sparse. Rows and columns referring to parameters outside the set will have zeroes, except for the diagonal positions where there are 1's. The inverse $W = V^{-1}$ will have the same sparse structure. The relation between the original normal matrix A and the transformed one $\bar{A}$ is

$$\bar{A} = V^T A V \quad \text{and} \quad A = W^T \bar{A} W. \tag{7.17}$$

From our previous discussion we know that $\bar{A}_{22}^{(p)}$ will have large elements only at the intersections of rows and columns referring to strongly coupled parameters. Our purpose is to show that the same holds for $A_{22}^{(p)}$. Partition the above relation $A = W^T \bar{A} W$ as

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}^T \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \\ \bar{A}_{21} & \bar{A}_{22} \end{bmatrix} \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \tag{7.18}$$

Assume first that all strongly coupled parameters are associated with rows and columns of A_{22}. We then have $W_{11} = I$, $W_{12} = 0$, $W_{21} = 0$. Hence

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} W_{22} \\ W_{22}^T \bar{A}_{21} & W_{22}^T \bar{A}_{22} W_{22} \end{bmatrix} \tag{7.19}$$

Applying partial reduction we find

$$\begin{aligned} A_{22}^{(p)} &= A_{22} - A_{21} A_{11}^{-1} A_{12} = W_{22}^T \bar{A}_{22} W_{22} - \\ &\quad - W_{22}^T \bar{A}_{21} \bar{A}_{11} \bar{A}_{12} W_{22} = \\ &= W_{22}^T (\bar{A}_{22} - \bar{A}_{21} \bar{A}_{11} \bar{A}_{12}) W_{22} = W_{22}^T \bar{A}_{22}^{(p)} W_{22}. \end{aligned} \tag{7.20}$$

It follows that $A_{22}^{(p)}$ and $\bar{A}_{22}^{(p)}$ are transformed into each other by the sparse matrix W_{22} which, being a submatrix of W, ensures that $A_{22}^{(p)}$ will have large elements at only the intersections of rows and columns of the strongly tied parameters.

We still have to consider the case where not all strongly tied parameters are associated with A_{22}. This case can be reduced to the earlier one by the following argument. View those stations whose strongly coupled parameters are associated with A_{11} as temporary or auxiliary stations, whose coordinates are of no interest after adjustment. Hence these coordinates can be eliminated immediately after forming the normal equations yielding a modified set of normals containing only parameters of interest. In these modified normals there will be strong ties between the remaining coordinates of the strongly coupled set. These ties are partly due to direct observations between the corresponding stations and partly to the eliminated stations. There will be no strong ties to other stations in the modified normals. Our previous argument can now be applied to the modified normals, proving also that in the general case $A_{22}^{(p)}$ will have large elements only at the intersections of rows and columns of the strongly coupled stations.

Remark: Although our argument applied only to $A_{22}^{(p)}$, we infer easily that it carries over to the coefficients of (R_{11}, R_{12}) which comprise the already eliminated equations. Eliminated equation i equals the top row of $A_{22}^{(i-1)}$ divided by the square root of the diagonal element!

7.2 Expected Coordinate Shifts and Right-Hand Side Coefficients.

According to section 3.4, the right-hand side coefficients $b_i^{(p)}$ are best understood by considering the ratios $b_i^{(p)}/a_{ii}^{(p)}$. Such a ratio is the shift of coordinate i with respect to its approximate positions while coordinates k, $p<k\leqslant n$, $k\neq i$ are fixed to their approximate positions, while coordinates k, $1\leqslant k\leqslant p$, $k=i$ are allowed to adjust freely. To estimate the right hand side coefficients during the various reduction states, we must rely on (a) estimates of the diagonals $a_{ii}^{(p)}$, which are available from our previous discussion, and (b) estimated shifts of the stations with respect to their approximate positions, assuming that a certain subset of stations is fixed to their approximate positions. The coordinate shifts enter the roundoff estimates not only indirectly via the right hand sides, but also directly through the global roundoff formulas. (See sec. 4.1.3.) It is now opportune for us to take a detour and survey all the information available on coordinate shifts, *i.e.*, on the quality of the approximate coordinates.

7.2.1 Quality of approximate coordinates

As usual, our entire discussion will be based on the assumption that coordinate shifts are scaled to the meter and that the normals are formed in agreement with this. The fact that shifts and normals actually are scaled differently, namely to arc seconds of latitude and longitude, will have only a marginal effect on the estimates to be considered later.

Let us consider these factors: Factor no. 1 is the assertion by NGS that the approximate coordinates of stations which are connected by one of the high precision measurements will be in near agreement with the values of this precise measurement. If, for example, such a measurement is a distance between

P, Q, the approximate coordinates of P, Q will match this distance in a way that the residual deviation is below two or three times the rms error of the precise distance. Shifts resulting from the subsequent adjustment will therefore essentially amount to a common translation and rotation of the two stations P, Q.

This consistency of approximate coordinates with high precision ties is extremely important. It implies that the right hand sides of equations belonging to strongly coupled stations will never become so large that they threaten the numerical validity of the results in the way the left-hand side coefficients do in such equations. Think of the original normals and let i, $i+1$, j, $j+1$ refer to stations P, Q. Then b_i/a_{ii} is the shift that the latitude of P suffers when all other coordinates (including the longitude of P) are fixed to their approximate values. Since the approximate positions of P, Q match the precise distance, Q will not exert a strong pull onto P away from its approximate position. Hence the shift b_i/a_{ii} will be small, amounting to a few millimeters. On the other hand, a_{ii} being of magnitude 10^6, is large. Hence, b_i is expected to be around 10^3. A similar reasoning applies to b_{i+1}, b_j, b_{j+1}. The argument also extends to $b_i^{(p)}$, $b_{i+1}^{(p)}$ and so on, and for $p > 0$.

Factor no. 2 is the predicted global shifts of the coordinates as taken from a most valuable study by Vincenty (1976). For cartographic purposes, Vincenty predicted the change in map corners after the new adjustment. The change occurs for two reasons: (1) the choice of a new datum and (2) distortions in the old network. The choice of the new datum is not yet definite. It is clear, however, that it will be based on an ellipsoid which will be centered at the Earth's mass center as good as possible, and fit well to the global geoid. The ellipsoid parameters will not be far from

$$a = 6\ 378\ 135 \text{ m} \qquad f = 1/298.26. \quad (7.21)$$

Vincenty's calculations were based on the ellipsoid that underlies the Naval Weapons Laboratory's (NWL) NWL10F datum. This datum is one in a series developed at NWL by sophisticated filtering of Doppler data. Vincenty's procedure compared the three-dimensional coordinates of the Doppler stations as calculated from the old North American Datum of 1927 (NAD27) with those calculated from the NWL10F datum. The map corners were interpolated from the shifts and afterwards were reduced to the NWL10F datum. The mean value of the three-dimensional coordinate shifts indicate a datum shift of about

$$x = -10.0 \text{ m} \quad y = 153.7 \text{ m} \quad z = 178.1 \text{ m} \qquad (7.22)$$

This datum shift is not relevant to the roundoff study because approximate coordinates will be corrected for this datum shift. More relevant are the residual deviations of the individual station shifts from the mean shift (datum shift). These residual deviations indicate distortions within the network, and may also be partly attributed to Doppler noise. The residual shifts again reduced down to the NWL10F ellipsoid, are roughly sketched in figures 7.5a-b. Figure 7.5a is a pictorial representation of the shifts of locations of odd latitude and longitude. Scale is provided by figure 7.5b which shows the mean of latitude and longitude shifts, superimposed by a random disturbance of 2 m. The distortions are below 5 m in most cases. Exceptionally large distortions, up to and exceeding 10 m, are found in Maine and Montana.

The distortions calculated by Vincenty provide a global picture of the anticipated coordinate changes. They will most strongly contribute to the prediction of the left-hand side global roundoff errors because they enter these formulas, directly. (See section 4.1.3.) But will they also give us indications of the size of the right-hand sides? Hardly! A right-hand side coefficient b_i of the original normals reflects only the *relative* shift of coordinate i with respect to the neighboring stations fixed to their approximate position, and also with respect to the second coordinate of the station to which coordinate i belongs. The relative accuracy of the approximate coordinates of neighboring stations is expected to be much higher than 5 m (the amount of the smooth global distortions). Hence, in estimating the right-hand sides of the original normals we must rely on estimates of the local inconsistencies between the approximate station positions.

We may also rely on factor no. 3 which consists of the computer outprints of local adjustments. Coordinate shifts and right-hand sides of such adjustments give us "snapshots" of some of the local situations in the U.S. network.

7.2.2 History of right-hand side coefficients

Because of the previously mentioned consistency of approximate coordinates and high precision measurements, it is not necessary, as it was for left side coefficients, to distinguish between stations involving such high precision measurements and between other stations. For the original normal equations of a lowest level block we take the right sides of sample calculations as representative. They are of the order of magnitude of 10^1 to 10^3. How will these right sides change as elimination proceeds according to the Helmert blocking scheme? Not much, as we shall see.

Consider first a station interior to a lowest level block. It is likely that such a station is still tied to some fixed (*i.e.*, not yet eliminated) neighboring stations at the moment when it is next in turn for piv-

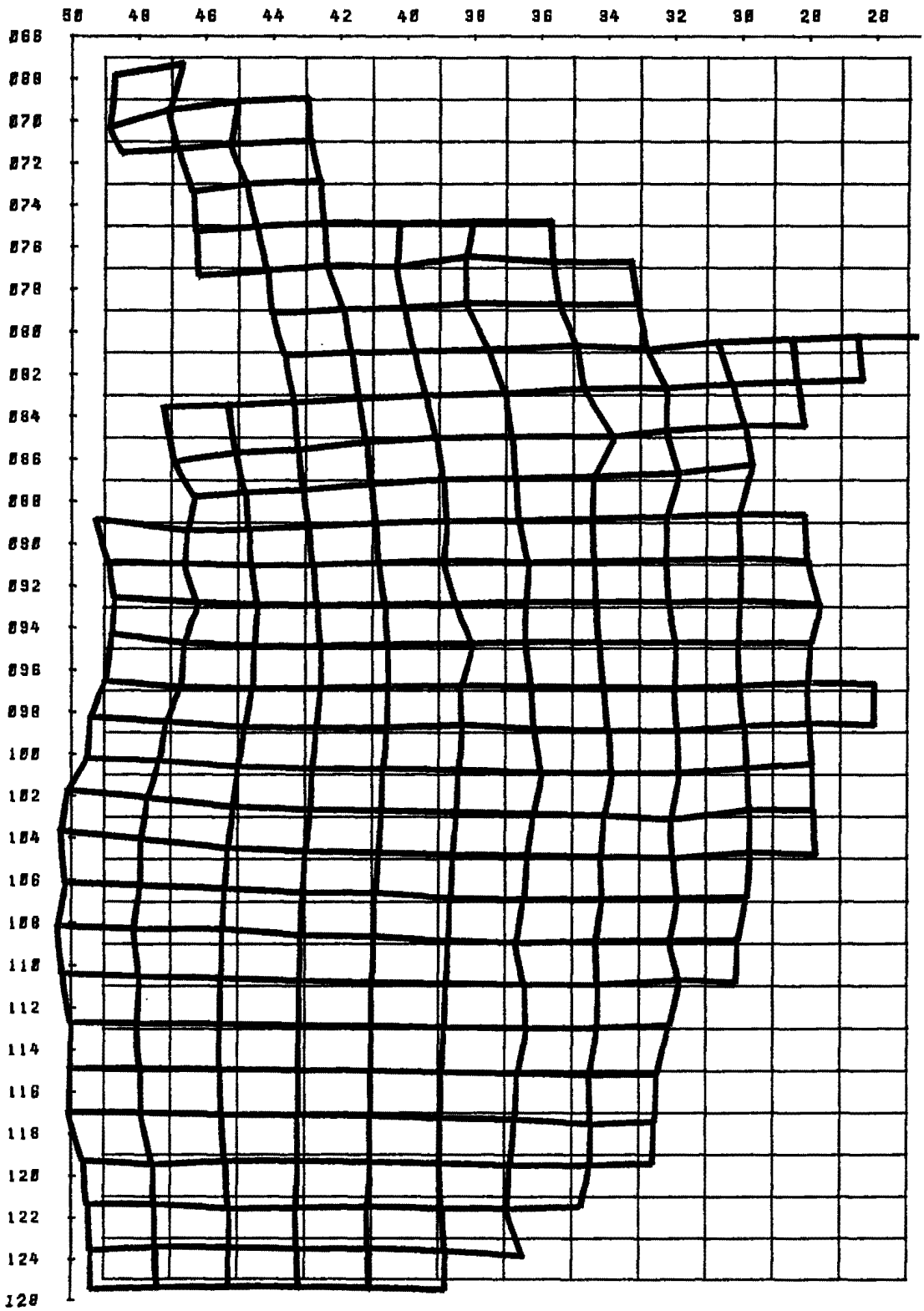

Figure 7.5a.—Pictorial representation of coordinate shifts.

	50	48	46	44	42	40	38	36	34	32	30	28	26	24
066	7.2	7.2	5.5											
068	9.1	9.1	7.1	5.7	5.5									
070	6.0	5.7	6.0	6.4	5.2									
072	5.1		5.1	4.9	4.9									
074	4.5		4.5	4.2	4.6	3.9	3.7	3.1						
076	4.6		4.6	3.8	4.3	4.0	5.2	3.1	2.5					
078	3.6			3.6	3.1	3.6	4.5	2.7	2.2					
080	2.8			2.8	2.5	2.9	2.6	2.3	2.2	3.2	3.9	4.2	3.6	
082	3.0	3.0	3.0	2.8	2.5	2.2	2.0	2.6	3.6	4.1	4.2	4.4		
084	4.3	4.3	3.2	3.1	2.3	2.0	2.0	3.6	3.6	5.1	4.4			
086	4.8	4.8	3.1	2.6	2.1	2.0	2.2	2.9	4.4	6.1				
088	2.2	3.1	2.6	2.1	2.0	2.1	2.3	2.8	3.5	4.1	3.9			
090	2.1	2.2	2.2	2.0	2.1	2.2	2.9	2.9	3.3	4.1	3.6			
092	2.9	3.3	2.6	2.2	2.3	2.6	2.9	3.2	3.5	3.9	4.3			
094	3.3	2.6	2.5	2.5	2.6	3.7	3.1	3.5	4.1	3.6	4.2			
096	2.5	2.3	2.3	2.5	2.6	3.1	3.2	3.4	3.6	3.7	4.2	4.0	4.0	
098	4.0	2.9	2.3	2.4	2.6	3.5	3.2	3.6	3.6	4.2	4.5	4.1	4.1	
100	4.3	3.6	2.3	2.4	2.5	2.9	3.5	3.6	3.8	4.5	5.0		5.0	
102	7.1	5.1	2.7	2.3	2.5	2.7	3.1	3.1	3.4	4.6	4.3		4.3	
104	7.7	5.6	3.1	2.3	2.3	2.5	2.8	3.1	3.2	4.6	4.2		4.6	
106	6.1	5.2	3.6	2.5	2.5	2.3	2.5	3.2	3.9	4.1			4.1	
108	6.5	5.4	4.0	2.6	2.3	2.2	2.1	2.9	3.3	3.5			3.5	
110	5.4	4.2	3.2	2.3	2.0	2.1	2.5	2.9	4.4	3.3			3.3	
112	3.9	3.9	2.9	2.1	2.0	2.0	2.5	2.7	3.2	3.1			3.2	
114	3.5	3.2	2.5	2.0	2.0	2.0	2.4	2.7	2.6				2.6	
116	3.4	3.0	2.5	2.2	2.1	2.2	2.6	3.8	3.2				3.2	
118	3.2	3.5	2.8	2.5	2.4	2.4	2.8	3.5	3.1					
120	3.1	3.2	3.2	3.2	2.8	2.9	2.9	2.9	3.0					
122	3.5	3.2	2.9	3.0	2.7	3.2	4.6							
124	3.1	2.9	2.6	2.5	2.4	2.8								
126														

Figure 7.5b.—Shifts used in calculations.

oting. Here, as in section 7.1.1, we temporarily use the idea that Cholesky's algorithm is executed in the fashion of successively eliminating equations. (See the remarks at the end of section 3.2.) We study the changes of $b_i^{(p)}$ for fixed i, and p varying from 0 to $i-1$. As long as a station is tied to fixed, *i.e.*, uneliminated stations, we can expect that the $b_i^{(p)}$ will reflect mainly the discrepancies between the observations to the neighboring stations and the approximate positions. This means that $b_i^{(p)}$ will not differ dramatically from $b_i^{(o)} = b_i$ in the original normals. Stations that are farther away are not likely to transmit a strong influence through the maze of the eliminated stations. In those exceptional cases where a station stands like an island in a sea of eliminated stations (with junction stations on its shore line), we do not expect either that the behavior of the right sides will be significantly different. It is true that for such a station we expect a somewhat larger shift when it is finally freed. But the fixed junction stations will prevent a shift at the global scale of some 5 m. The shift may be perhaps 1 to 2 m. The variance of the station coordinates may be of the order 10^{-1}. Hence $a_{ii}^{(p)} \doteq 10^1$. This gives us $b_i^{(p)} = a_{ii}^{(p)} x_i \doteq 10^1 * 2 = 20$ which is well within the assumed range of the $b_i^{(o)}$.

Our argument can be extended to stations that are eliminated at a medium level. Ties to neighboring stations are always likely. But, also, the right-hand side of the very last coordinate to be eliminated will not deviate too much. The diagonal element is the reciprocal variance of this coordinate when it comes out of the adjustment, hence $a_{nn}^{(n-1)} \doteq 10^1$. The shift of the coordinate with respect to the approximate position can be up to 10 m. This gives a $b_n^{(n-1)}$ that is of magnitude 10^2.

Until now we have been concerned exclusively with the coefficients $b_i^{(p)}$, *i.e.*, with the right-hand sides of equations whose time has not yet come to be eliminated. The coefficients s_i of the eliminated equations differ from $b_i^{(i-1)}$ by division through the square root r_{ii} of the diagonal element $a_{ii}^{(i-1)}$. Because the diagonal elements differ greatly in size, so can s_i. From all we have said, there is no reason to fear that s_i will become unduly large. Possibly only a certain s_i may become very small because of a large r_{ii}. We must then deal with a station that is tied strongly to another station which comes later in the elimination sequence. The small s_i reflects the fact that coordinate i moves slightly when the neighboring stations are held fixed.

8. SAFE BOUNDS ON GLOBAL ROUNDOFF ERRORS

In sections 4.1.4 and 4.2.3 we specified preliminary bounds on the global roundoff errors in the U.S. network. Details that were anticipated in these earlier sections concerning bounds on elements of A, $A_{22}^{(p)}$, $F = A^{-1}$, b, $b_2^{(p)}$, x as well as on the numbers of elementary operations have been filled in. It is remarkable that the preliminary bounds indicate feasibility of the adjustment with only one exception. This is the global bias $E\{\zeta\}$ caused by left-side triangularization roundoff errors on an IBM 360 computer. In this chapter we will be mainly concerned about remedying this situation. Insight gained through the discussions in chapter 7 will be beneficial. We will also reconsider the other bounds and improve them.

8.1 Roundoff Error Propagation When Weight Singularities Are Transformed Away

We assume hypothetically that all weight singularities of the U.S. network are transformed away by the method given in section 7.1.4. Because the transformations are not actually carried out, we can even indulge in the luxury of assuming that for clusters containing more than two tightly connected stations, spectral parameters have been introduced, as discussed in the remarks of 7.1.4. We further apply a scale transformation to the local parameters of high accuracy such that their variances become comparable in size to the variances of the ordinary coordinate shifts. The changes to these local parameters then will no longer be measured in meters, but rather in units close to a decimeter. This results in larger numerical values for the local parameter changes. Transforming the normal equations in this way will result in a system in which no element a_{ij} of the left side will be larger than about 10^4. The units of a diagonal a_{ii} are still m^{-2} if i refers to a global coordinate. For i referring to a local parameter, different units are in effect, but it is not necessary in the sequel to make constantly explicit reference to these different units. A magnitude of 10^4 m^{-2} for a_{ii} corresponds to a "neighborhood accuracy" of the remaining global stations, which is in the centimeter range. The inverse F of the normal equation matrix A will only be slightly affected by the transformations. We may retain our earlier assumptions on the size of the elements f_{ij}.

8.1.1 Estimating the bias $E\{\zeta\}$

Concentrating on the left-side roundoff errors during triangular decomposition, our starting point will be eq. (4.34), with the right-side contribution omitted. This is

$$\zeta_i = -\sum_{j=1}^{n} f_{ij} x_j \varepsilon_{ij} - \sum_{j=1}^{n} \sum_{k=j+1}^{n} (f_{ij} x_k + f_{ik} x_j) \varepsilon_{jk}. \quad (8.1)$$

Replacing ζ_i, ε_{jk} by $E\{\zeta_i\}$, $E\{\varepsilon_{jk}\}$, taking absolute values and counting the contribution of the diagonal terms ε_{jj} twice, we arrive at:

$$|E\{\zeta_i\}| \leqslant \sum_{j=1}^{n} \sum_{k=j}^{n} \{|f_{ij}|\,|x_k| + |f_{ik}|\,|x_j|\}\,|E\{\varepsilon_{jk}\}|. \qquad (8.2)$$

This can be rearranged to yield:

$$|E\{\zeta_i\}| \leqslant \sum_{j=1}^{n} |f_{ij}| \{\sum_{k=j}^{n} |x_k|\,|E\{\varepsilon_{jk}\}| + \sum_{k=1}^{j} |x_k|\,|E\{\varepsilon_{kj}\}|\}. \qquad (8.3)$$

Recall from section 4.1.1 that ε_{jk} is the sum of $2\mu_{jk} + 1 \doteq 2\mu_{jk}$ elementary roundoff errors $\varepsilon_{jk}^{(el)}$. Here the symbol $\varepsilon_{jk}^{(el)}$ stands for $\varepsilon_{jkl}^{(*)}$, $\varepsilon_{jkl}^{(+)}$, $\varepsilon_{jk}^{(s)}$, $\varepsilon_{jk}^{(d)}$, and μ_{jk} is the number of nonzero product terms needed to reduce a_{jk} ($\mu_{jk} = 0$ if a_{jk} is a zero location). We will work with a single bound on $E\{\varepsilon_{jk}^{(el)}\}$ which is obtained in the following manner:

Because weight singularities have been removed, no left-side coefficient a_{ij} will be larger in absolute value than 10^4. In agreement with proposition 4.1, see also eq. (4.49), we need an integer power $\bar{c}$ of the base $\beta = 16$ that will bound all $|a_{ij}|$. To specify this bound properly, we must temporarily switch to a set of normals scaled in agreement with coordinate shifts measured in arc seconds. A maximum element $\|a\| = 10^4 m^{-2}$ corresponds to a maximum of $10^4 * |R\pi/(180*3600)^2| = 10^4 * 30.9^2$ arc sec^{-2}. This gives about $9.6* 10^6$ arc sec^{-2}. The smallest power of 16 bounding this is $16^6 \doteq 1.7*10^7$. Scaling this backwards to meters, we arrive at $1.7*10^7/30.9^2 = 1.8 *10^4$. Hence our scaling value is now $\bar{c} = 1.8* 10^4$. Multiplying by $2\sqrt{\beta} = 8$, we get $c \doteq 14*10^4$. With this we bound the nonzero biases as

$$|E\{\varepsilon_{ii}^{(el)}\}| \leqslant \frac{c}{2} 16^{-14} \text{ on the IBM 360.} \qquad (8.4)$$

We immediately lower this bound by the following arguments: To obtain c, multiplication of $\bar{c}$ by 8 is necessary only because of elementary operations involving the square root of the diagonal elements ($r_{ii} = \sqrt{a_{ii}}$). Because diagonals are counted twice, we may lower the factor from 8 to 4. In the present context, it can even be lowered to 1. The factor is indeed 1, if $\varepsilon_{jk}^{(el)}$ refers to an elementary roundoff error $\varepsilon_{jkl}^{(*)}$, $\varepsilon_{kl}^{(+)}$ occurring during the evaluation and summation of the product terms $a_{jl}a_{kl}$. Only the above mentioned square roots and the division that establishes $r_{jk} = a_{jk}/a_{jj}$ still necessitate a factor of 4. However, about a 1,000 times more operations are involved in the product sum accumulations than there are square roots and divisions. Because we are presently working with a uniform bound on all roundoff errors, we may indeed lower the factor to 1. Hence in this subsection we use

$$|E\{\varepsilon_{ij}^{(el)}\}| \leqslant \frac{\bar{c}}{2} 16^{-14} \doteq 10^4*16^{-14} \text{ on the IBM 360.} \qquad (8.5)$$

After these preparations we can derive from (8.3) the following estimate:

$$|E\{\zeta_i\}| \leqslant 10^4 * 16^{-14} \sum_{j=1}^{n} |f_{ij}| \{\sum_{k=j}^{n} 2\mu_{jk}|x_k| + \sum_{k=1}^{j} 2\mu_{kj}|x_k|\}. \qquad (8.6)$$

Using a single bound $\|x\|$ on $|x_k|$, we can also use the following inequality:

$$|E\{\zeta_i\}| \leqslant 10^4 *16^{-14} * \|x\| \sum_{j=1}^{n} |f_{ij}| \{\sum_{k=j}^{n} 2\mu_{jk} + \sum_{k=1}^{j} 2\mu_{kj}\}. \qquad (8.7)$$

We will use both eqs. (8.6) and (8.7) in the sequel. Using first (8.7), which is simpler, we proceed by approximating the sum and by assuming constant values f_{ij} as long as any of the subscripts refers to a certain $2°\times 2°$ quad. Stated more precisely: Let all $2°\times 2°$ quads be labelled by Greek indices ϱ,χ. All coefficients f_{ij} will have the same value $f_{\rho\chi}$ as long as i is a coordinate out of a certain $2° \times 2°$ quad Q_ρ and j is a coordinate out of a different $2° \times 2°$ quad Q_χ. Also, ζ_i is replaced by the piecewise constant function ζ_ρ. The replacement of f_{ij} by $f_{\rho\chi}$ is justifiable for only the global part $f_{ij}^{(global)}$ of the covariance. The global part goes over into $f_{\rho\chi}^{(global)}$. If $\varrho = \chi$, or if quad Q_ρ is a neighbor to Q_χ, a local part (peak) of f_{ij}, denoted by $f_{ij}^{(local)}$, must be taken into account. This is done by means of a correction to the result obtained from $f_{\rho\chi}^{(global)}$.

In addition, global roundoff errors ζ_i will be replaced by ζ_ρ, as mentioned above. From (8.7) we get:

$$|E\{\zeta_\rho\}| \leqslant 10^4 * 16^{-14} * \|x\| * \sum_{\chi} f_{\rho\chi}^{(global)} \{\Gamma_\chi^{(r)} + \Gamma_\chi^{(c)}\} + \text{peak contribution from } f_{ij}^{(local)}. \qquad (8.8)$$

The quantities $\Gamma_\chi^{(r)}$, $\Gamma_\chi^{(c)}$ are precisely the $2°\times 2°$ quad-based individual Γ row and column counts (see chapter 6, especially figs. 6.15a-b):

$$\Gamma_\chi^{(r)} = \sum_{j\in Q_\chi} \sum_{k=j}^{n} 2\mu_{jk}, \quad \Gamma_\chi^{(c)} = \sum_{j\in Q_\chi} \sum_{k=1}^{j} 2\mu_{kj}. \qquad (8.9)$$

Assuming $\|x\| = 10$, we evaluate (8.8) for certain $2°\times 2°$ quads. Table 8.1 gives the results. The quads ϱ are identified by latitude and longitude of their midpoint. The quads were selected as being representative for certain regions of the U.S. network. From some of these quads maximum error contributions can be expected.

TABLE 8.1.—*Bounds from eq. (8.8) on the bias $E\{\zeta_\rho\}$ during first iteration, for the net with weight singularities removed, using an IBM-type computer*

Quad ϱ		Type of contribution		Bound on
ϕ	λ	Global m	Local m	$E\{\zeta_\rho\}$ m
39	77	0.00186	0.00040	0.00226
47	69	0.00245	0.00009	0.00254
47	121	0.00185	0.00044	0.00230
41	97	0.00144	0.00036	0.00180
35	111	0.00144	0.00021	0.00165

The local peak contribution is based on formula (5.43) given in section 5.6, where it was assumed that we were dealing with two local peaks of the type

$$p(d) = \frac{p_o}{\log 4} \log \frac{4}{1+\frac{3d}{d_o}} . \tag{8.10}$$

If our quad Q_ρ shows a Γ count of Γ_ρ, we must assume a worst case, namely, that all roundings occur along a linear set of stations. Such a linear set of stations refers to a major block boundary; as we have seen, the majority of roundoff errors arises in connection with block boundaries. The minimum width of a $2°\times 2°$ block is about 180 km. Hence a cautious way in which to bound the local peak contribution from (8.10) is

$$10^4 * 16^{-14} \|x\| \frac{\Gamma_\rho}{180} 2 \int_0^{d_o} p(x)\, dx \doteq$$

$$10^4 * 16^{-14} \|x\| p_o * 0.78 \frac{d_o}{180} \Gamma_\rho . \tag{8.11}$$

If $d_o > 180$, one should replace Γ_ρ by the largest Γ count of Q_ρ or of any adjacent quad.

Because shifts are not 10 m at every point, we can improve these estimates somewhat by using (8.6). With similar reasoning, we arrive at the following formula:

$$|E\{\zeta_\rho\}| \leqslant 10^4 * 16^{-14} \sum_\chi f_{\chi\rho}^{(global)} \{\Xi_\chi^{(r)} + \Xi_\chi^{(c)}\}.$$

$$+ \text{ peak contribution from } f_{ij}^{(local)} \tag{8.12}$$

The quantities $\Xi_\chi^{(r)}$, $\Xi_\chi^{(c)}$ are "shift weighted" Γ counts, defined as follows:

$$\Xi_\chi^{(r)} = \sum_{j\in Q_\chi} \sum_{k=j}^{n} \|x\|_{\omega(k)} 2\mu_{jk}$$

$$\Xi_\chi^{(c)} = \sum_{j\in Q_\chi} \sum_{k=1}^{j} \|x\|_{\omega(k)} 2\mu_{kj}. \tag{8.13}$$

Here $\omega(k)$ refers to the $2°\times 2°$ quad Q_ω in which coordinate k is located. $\|x\|_\omega$ is a bound on the shift in quad Q_ω. Because the shifts are not equal in latitude and longitude, a bound on the average shift is used. Recall the shifts in the $2°\times 2°$ quads shown in figures 7.5a-b. To account for local variations of the coordinate shifts, any bound $\|x\|_\omega$ on the average is replaced by $\sqrt{\|x\|_\omega^2 + 2^2}$. This means that 2-m local shifts are randomly superimposed. Figure 7.5b shows the values of $\sqrt{\|x_\omega\|^2+2^2}$ for the various quads.

A moderate modification of the computer programs used to evaluate $\Gamma_\chi^{(r)}$, $\Gamma_\chi^{(c)}$ yielded the shift-weighted Γ-counts $\Xi_\chi^{(r)}$, $\Xi_\chi^{(c)}$, which we call "Ξ-counts." The superposition $\Xi_\chi^{(r)} + \Xi_\chi^{(c)}$ is shown in figure 8.1. Table 8.2 lists the bounds on $E\{\zeta_\rho\}$ resulting from this procedure.

TABLE 8.2.—*Bounds from eq. (8.12) on the bias $E\{\zeta_\rho\}$ for a network with weight singularities removed, using an IBM-type computer*

Quad ϱ		Type of contribution		Bound on
ϕ	λ	Global m	Local m	$E\{\zeta_\rho\}$ m
39	77	0.00067	0.00016	0.00083
47	69	0.00098	0.00004	0.00102
47	121	0.00063	0.00016	0.00079
41	97	0.00050	0.00012	0.00062
35	111	0.00049	0.00007	0.00056

8.1.2 Estimating the standard deviation $\sigma\{\zeta_i\}$

Beginning with eq. (8.1) and taking variances, we obtain

$$\sigma^2\{\zeta_i\} = \sum_{j=1}^{n} f_{ij}^2 x_i^2 \sigma^2\{\varepsilon_{jj}\} + \sum_{j=1}^{n} \sum_{k=j+1}^{n} \{f_{ij}x_k + f_{ik}x_j\}^2 \sigma^2\{\varepsilon_{jk}\}. \tag{8.14}$$

Using the inequality $(a+b)^2 \leqslant 2(a^2+b^2)$, we see that

$$\sigma^2\{\zeta_i\} \leqslant 2 \sum_{j=1}^{n} \sum_{k=j}^{n} \{f_{ij}^2 x_k^2 + f_{ik}^2 x_j^2\} \sigma^2\{\varepsilon_{jk}\}. \tag{8.15}$$

Imagining ε_{jk} as a superposition of elementary roundoff errors $\varepsilon_{jk}^{(el)}$ and bounding $\sigma\{\varepsilon_{jk}^{(el)}\}$ by

$$\sigma\{\varepsilon_{jk}^{(el)}\} \leqslant \frac{c}{\sqrt{12}} \beta^{-\tau}$$

whereby $c = 2\sqrt{\beta}\,\bar{c}$, $\bar{c} = 1.8 * 10^4$, we find that $c = 5 * 10^4$ for the CDC 6600 and $c = 14 * 10^4$ for the IBM 360. It follows that

$$\sigma\{\varepsilon_{jk}^{(el)}\} \leqslant \begin{matrix} 1.5 \bullet 10^4 * 2^{-48} \ldots \text{CDC 6600} \\ 4.2 \bullet 10^4 \bullet 16^{-14} \ldots \text{IBM 360} \end{matrix} . \tag{8.16}$$

By applying similar reasoning to that in section 8.1.1, we obtain the counterpart of (8.8):

	50	48	46	44	42	40	38	36	34	32	30	28	26	24
066	14E8	62E7	33E8											
068	45E8	12E8	23E8	70E7	14E7									
070	42E8	99E6	40E8	17E9	12E9									
072	30E8		22E8	38E8	21E9									
074	11E8		26E8	23E8	17E9	74E8	29E8	15E7						
076	32E8		53E6	74E7	49E8	18E9	14E9	46E8	58E6					
078	28E8			68E8	12E9	11E9	15E9	17E9	96E8					
080	33E8			13E5	23E8	10E8	16E8	20E8	70E8	10E8	11E8	14E7	10E5	
082	21E8	40E6	25E7	10E8	31E8	15E8	24E8	18E8	56E8	14E8	54E7	29E7		
084	55E8	41E7	50E7	43E7	28E8	11E8	10E8	55E7	59E8	11E8	27E7			
086	21E8	14E8	35E8	34E8	88E8	60E8	80E8	47E8	53E8	28E8				
088	17E8	20E7	96E7	87E7	27E8	83E7	13E8	11E8	58E8	16E8	11E8			
090	27E8	34E7	93E7	79E7	20E8	83E7	18E8	15E8	75E8	20E8	19E8			
092	40E8	51E7	10E8	28E7	13E8	68E7	13E8	72E7	55E8	99E7	82E7			
094	96E8	12E9	13E9	11E9	12E9	13E9	15E9	11E9	14E9	91E8	77E8			
096	67E8	35E7	71E7	33E7	30E8	66E7	12E8	87E7	49E8	96E7	92E7	43E7	12E8	
098	62E8	42E7	64E7	40E7	26E8	98E7	91E7	81E7	25E8	80E7	97E7	27E7	87E7	
100	75E8	44E7	63E7	21E7	18E8	24E7	57E7	33E7	18E8	15E7	17E7		10E8	
102	73E8	87E7	15E8	86E7	28E8	98E7	16E8	10E8	19E8	59E7	21E7		12E8	
104	31E8	26E6	25E7	11E7	15E8	57E6	60E7	20E7	21E8	92E6	10E5		94E7	
106	35E8	42E7	45E7	11E7	12E8	27E7	95E7	88E7	28E8	20E7			16E8	
108	31E8	18E7	24E7	47E6	93E7	36E6	28E7	65E6	21E8	17E6			14E8	
110	64E8	41E8	33E8	42E8	70E8	41E8	40E8	57E8	78E8	15E8			24E8	
112	50E8	10E7	48E7	17E7	12E8	51E6	55E7	39E7	24E8	16E5			18E8	
114	52E8	55E7	80E7	28E7	17E8	35E7	10E8	12E8	34E8				34E8	
116	61E8	60E7	76E7	15E7	12E8	16E7	94E7	11E8	41E8				13E8	
118	76E8	35E8	24E8	62E7	19E8	18E8	23E8	33E8	29E8					
120	41E8	11E8	24E8	31E7	21E8	11E8	20E8	48E7	25E5					
122	18E9	11E9	49E8	60E7	23E8	16E8	19E8							
124	32E8	31E7	31E7	28E7	10E8	34E4								
126														

Figure 8.1.—Ξ counts for 2° × 2° quads.

$$\sigma\{\zeta_\rho\} \leqslant \frac{c}{\sqrt{12}}\sqrt{2}\,\|x\|\sqrt{\sum_{\chi}\{f_{\rho\chi}^{(global)}\}^2\{\Gamma_\chi^{(r)}+\Gamma_\chi^{(c)}\}}\,\beta^{-\tau}+$$

$$+\text{ peak contribution from } f_{ij}^{(local)}. \qquad (8.17)$$

The peak contribution must be superimposed in the most pessimistic way. By reasoning similar to the transition from (8.14) to (8.15), we conclude that the superposition is bounded by $\sqrt{2}$ times the mean square superposition of the global and local parts.

The counterpart of (8.12) is obtained as

$$\sigma\{\zeta_\rho\} \leqslant \frac{c}{\sqrt{12}}\sqrt{2}\,\sqrt{\sum\{f_{\rho\chi}^{(global)}\}^2\{\overset{2}{\Xi}{}_\chi^{(r)}+\overset{2}{\Xi}{}_\chi^{(c)}\}}\,\beta^{-\tau}+$$

$$+\text{ peak contribution from } f_{ij}^{(local)}. \qquad (8.18)$$

Here we have introduced the "squared shift-weighted" Γ counts:

$$\begin{aligned}\overset{2}{\Xi}{}_\tau^{(r)} &= \sum_{j\in Q_\tau}\sum_{k=j}^{n}\|x\|^2_{\omega(k)}\,2\mu_{jk}\\ \overset{2}{\Xi}{}_\tau^{(c)} &= \sum_{j\in Q_\tau}\sum_{k=1}^{j}\|x\|^2_{\omega(k)}\,2\mu_{kj}.\end{aligned} \qquad (8.19)$$

The results of calculations based on formulas (8.17) are shown in table 8.3. Because the errors are quite small, formula (8.18) was not used.

The local contribution of a peak (8.10) to $\sigma\{\zeta_\rho\}$ was evaluated from

$$\frac{c}{\sqrt{12}}\sqrt{2}\,\|x\|\sqrt{\frac{\Gamma_\rho}{180}\,2\int_0^{d_0}p^2(x)\,dx}\,\beta^{-\tau} =$$

$$= \frac{c}{\sqrt{12}}\sqrt{2}\,\|x\|\,p_0 * 0.662 * \sqrt{\frac{d_0}{180}\,\Gamma_\rho}\,\beta^{-\tau}. \qquad (8.20)$$

8.2 Contribution of the Weight Singularities

As we argued at length in chapter 7, the weight singularities will cause large coefficients to be superimposed on the left-hand sides of the normal equations. These coefficients can go up to $5*10^6$. They are restricted to the intersections of rows and columns of stations which form strongly connected clusters. There will be no coupling between different clusters. As elimination proceeds, no additional large coefficients will arise at other locations. Large coefficients may or may not drop to moderate size before their equation is eliminated.

To my knowledge, the number of stations involved in high precision measurements does not exceed 25 percent of the total number, which is below 200,000. Hence we expect to have fewer than 100,000 equations in which such large coefficients occur. The number of clusters, or the average number of stations per cluster is not known. Because the number of large coefficients increases if there are fewer but larger clusters, I assume there are about 10,000 clusters, each having 5 stations. This is a conservative figure, because many clusters involve only two stations. The number of large coefficients then would be around $10{,}000 * (10*9/2) = 450{,}000$.

As argued in chapter 7, there will be many elementary operations involving a large coefficient which will not cause a large local roundoff error. Hence it is extremely pessimistic to assume that every elementary operation causes a roundoff error with magnitude of $5*10^6 * \beta^{-\tau}$. Further we assume that no provisions have been made to get rid of the strongly coupled stations at the lowest block level. Hence the distribution of "singular" stations among the various block levels may be the same as the distribution among the whole population of stations. At a diagonal coefficient there will be $2\mu_{ii} + 1$ elementary roundoff errors. The number of roundoff errors at an off-diagonal location (i,j) is $2\mu_{ij} + 1$, which, in view of $\mu_{ij} \leqslant \mu_{ii}$, is smaller than the number of errors in the corresponding location (i,i). Hence it is conservative to use the number $2\mu_{ii} + 1$ also for the off-diagonals. Tightly connected stations will be situated close together. Therefore, the shift values x_i, x_j for two such stations i,j are assumed to be identical.

Summarizing, we pretend that for 25 percent of the equations $2 * 2.5 * (2\,\mu_{ii} + 1)$ bad roundings will occur close to the diagonal position; therefore, we can put $f_{ij}\,x_k = f_{ik}\,x_j = f_{ij}\,x_j$.

TABLE 8.3.—*Bounds from eq. (8.17) on $\sigma\{\zeta_\rho\}$ during first iteration and for network with weight singularities removed.*

Quad ϱ		Type of contribution				Bound on $\sigma\{\zeta_\rho\}$	
		global		local			
ϕ	λ	CDC 6600	IBM 360	CDC 6600	IBM 360	CDC 6600	IBM 360
39	77	2.8E-6	3.0E-8	2.1E-6	2.3E-8	4.9E-6	5.4E-8
47	69	4.9E-6	5.3E-8	9.8E-7	1.1E-8	7.1E-6	7.6E-8
47	121	2.8E-6	3.0E-8	2.2E-6	2.4E-8	4.9E-6	5.4E-8
41	97	2.0E-6	2.2E-8	1.9E-6	2.1E-8	4.0E-6	4.2E-8
35	111	2.0E-6	2.2E-8	1.5E-6	1.6E-8	3.5E-6	4.0E-8

8.2.1 Estimating the bias $E\{\zeta_\rho\}$

Starting with eq. (8.2) and using the simplification just discussed, we arrive at the following estimate for the bias contribution resulting from the large coefficients:

$$|E\{\zeta_i\}| \leqslant 0.25 * 2 * 2.5 * \sum_{j=1}^{n} 2|f_{ij}|\,|x_j|\,|E\{\varepsilon_{jj}\}|. \quad (8.21)$$

As a bound for the elementary roundoff errors, we can use (8.5) with $\bar{c}$ replaced by $16^8/30.9^2 = 5*10^6$. Because ε_{jj} is composed of $2\mu_{jj}+1 \doteq 2\mu_{jj}$ elementary roundoff errors, we get

$$|E\{\zeta_i\}| \leqslant 0.25 * 2 * 2.5 * 5 * 10^6 * \frac{1}{2} * 16^{-14} *$$

$$* \sum_{j=1}^{n} 2|f_{ij}|\,|x_j|\,2\mu_{jj} \doteq$$

$$1.25 * 10^7 * 16^{-14} * \sum_{j=1}^{n} |f_{ij}|\,|x_j|\mu_{jj}. \quad (8.22)$$

Applying the same procedure as in section 8.1.1, we get

$$|E\{\zeta_\rho\}| \leqslant 1.25 * 10^7 * 16^{-14} * \sum_{\chi} |f_{\rho\chi}^{(global)}|\,\|x\|_\chi\,\Pi_\chi$$

$$+ \text{ peak contribution due to } f_{ij}^{(local)}. \quad (8.23)$$

We see that this time the Π-counts Π_χ for the various quads Q_χ come in. Using a uniform bound $\|x\|$ on $\|x\|_\chi$ we get

$$|E\{\zeta_\rho\}| \leqslant 1.25 * 10^7 * 16^{-14} * \|x\| * \sum_{\chi} |f_{\rho\chi}^{(global)}|\,\Pi_\chi$$

$$+ \text{ peak contribution due to } f_{ij}^{(local)}. \quad (8.24)$$

The estimates resulting from (8.24) are shown in table 8.4. The table also shows earlier bounds (from table 8.1) that are applied to the network with weight singularities removed. The last column shows the total bounds obtained by superposition.

TABLE 8.4.—*Bounds from eq. (8.24) for contribution of weight singularities toward $E\{\zeta_\rho\}$ during the first iteration*

Quad ϱ		Type of contribution		Bound on $E\{\zeta_\rho\}$ from eq. (8.24)	Earlier bound from table 8.1	Total bound on $E\{\zeta_\rho\}$
ϕ	λ	Global	Local			
39	77	0.00138	0.00048	0.00185	0.00226	0.00411
47	69	.00198	.00010	.00208	.00254	0.00462
47	121	.00130	.00039	.00169	.00230	0.00399
41	97	.00100	.00012	.00112	.00180	0.00292
35	111	.00100	.00009	.00109	.00165	0.00274

Based on eq. (8.23), slightly improved estimates are calculated, as shown in table 8.5. The table also shows earlier bounds (from table 8.2) that are applied to the network with weight singularities removed. The last column shows the total bounds obtained by superposition.

TABLE 8.5.—*Bounds from eq. (8.23) for contribution of weight singularities toward $E\{\zeta_\rho\}$ during the first iteration*

Quad ϱ		Type of contribution		Bound on $E\{\zeta_\rho\}$ from eq. (8.23)	Earlier bound from table 8.2	Total bound on $E\{\zeta_\rho\}$
ϕ	λ	Global	Local			
39	77	0.00052	0.00019	0.00071	0.00083	0.00154
47	69	.00090	.00006	.00096	.00102	0.00198
47	121	.00044	.00012	.00057	.00079	0.00136
41	97	.00035	.00005	.00040	.00062	0.00102
35	111	.00034	.00004	.00038	.00056	0.00094

8.2.2 Estimating the standard deviation $\sigma\{\zeta_i\}$

Just as (8.21) is the counterpart of (8.2), the following equation is the counterpart of (8.15):

$$\sigma^2\{\zeta_i\} \leqslant 0.25 * 2 * 2.5 * 2 \sum_{j=1}^{n} 2 f_{ij}^2 x_j^2 \sigma^2\{\varepsilon_{jj}\}. \quad (8.25)$$

By reasoning as described previously, we arrive at:

$$\sigma\{\zeta_\rho\} \leqslant \frac{c}{\sqrt{12}} * 2.23 * \beta^{-\tau}\,\|x\|\,\sqrt{\sum_{\chi} \{f_{\rho\chi}^{(global)}\}^2\,\Pi_\chi}$$

$$+ \text{ peak contribution due to } f_{ij}^{(local)}. \quad (8.26)$$

The values of c are cautiously taken to be $2\sqrt{\beta} * \bar{c}, \bar{c} = 5*10^6$. We get $c = 14*10^6$ for the CDC 6600 and $4*10^7$ for the IBM 360. This leads to tables 8.6 and 8.7, which contain estimates for the standard deviation component resulting from weight singularities. The tables also show earlier bounds (from table 8.3) that apply to the network with weight singularities removed. The last column of each table shows the total bounds obtained by superposition according to the rule explained in the text following eq. (8.17).

TABLE 8.6.—*Bounds of the CDC 6600 computer derived from eq. (8.26) showing the contribution of weight singularities toward $\sigma\{\zeta_\rho\}$ during the first iteration*

Quad ϱ		Type of contribution		Bound on $\sigma\{\zeta_\rho\}$ from eq. (8.26)	Earlier bound from table 8.3	Total bound on $\sigma\{\zeta_\rho\}$
ϕ	λ	Global	Local			
39	77	3.1E-5	2.8E-5	5.9E-5	4.9E-6	8.4E-5
47	69	5.6E-5	1.3E-5	8.2E-5	7.1E-6	1.2E-4
47	121	2.9E-5	2.5E-5	5.4E-5	4.9E-6	7.6E-5
41	97	2.0E-5	1.4E-5	3.5E-5	4.0E-6	5.0E-5
35	111	2.1E-5	1.2E-5	3.4E-5	3.5E-6	4.8E-5

TABLE 8.7.—*Bounds on the IBM 360 computer derived from eq. (8.26) showing the contribution of weight singularities toward $\sigma\{\zeta_\rho\}$ during the first iteration*

Quad ϱ		Type of contribution		Bound on $\sigma\{\zeta_\rho\}$ from eq. (8.26)	Earlier bound from table 8.3	Total bound on $\sigma\{\zeta_\rho\}$
ϕ	λ	Global	Local			
39	77	3.5E-7	3.1E-7	6.6E-7	5.4E-8	9.4E-7
47	69	6.3E-7	1.5E-7	9.1E-7	7.6E-8	1.3E-6
47	121	3.3E-7	2.8E-7	6.1E-7	5.4E-8	8.6E-7
41	97	2.3E-7	1.6E-7	4.0E-7	4.2E-8	5.6E-7
35	111	2.4E-7	1.3E-7	3.8E-7	4.0E-8	5.4E-7

8.3 Residual Bias on the CDC 6600

The roundoff estimates determined for the CDC 6600 were quite small. It was estimated that the first adjustment run, where some coordinate shifts were expected to exceed 10 m, will yield a solution vector with an error of about 0.0001 m. This implies that any further iteration will give about five correct digits of the largest coordinate shift. Nevertheless some words of caution are appropriate:

Remember that the CDC 6600 estimates rely on two idealized assumptions.

(1) It was assumed that the elementary roundoff errors are completely unbiased. This is not entirely true. The CDC is not a truly rounding machine in the mathematically strict sense. But even on a truly rounding machine, an elementary roundoff error may be biased on some rare occasions, namely when the two operands are of very different magnitude. (Refer to the discussion in sec. 2.8.1.)

(2) The propagation of local roundoff errors to the global ones was done according to a linear model. For example, it was assumed that, if $Ax = b$, then $\zeta = -A^{-1}\varepsilon\, x$ is the solution of $(A+\varepsilon)(x+\zeta) = b$. However, this is only true to the first degree of approximation. The neglected higher order terms could cause some small bias of ζ.

Let us assume hypothetically that the standard instruction set is used on the CDC 6600. Then chopping occurs instead of (nearly) true rounding, and a bias-analysis, similar to the one done for the IBM 360, could be performed. It would result in global roundoff error estimates approaching 0.5m. Hence only one correct significant digit of the largest coordinate shift can be guaranteed from the linear model. Nonlinearity effects then can no longer be neglected and the results may possibly be entirely useless. Although we believe that our estimation procedure, detailed in section 8.1 and 8.2, overshoots the errors somewhat, I do not recommend doing the adjustment on the "chopping" CDC 6600.

Fortunately, any residual bias encountered on the "truly rounding" CDC 6600 will be much smaller. I am not able to state precisely how much smaller it will be. From test calculations, which will be documented in chapter 10, I conclude that the residual bias will be smaller by at least three powers of 10. Hence at least four correct digits of the largest coordinate shift will be recovered correctly.

Let us summarize by stating that the CDC 6600 is a safe machine on which to do the adjustment, although one cannot entirely exclude the fact that the estimates in the previous section are too optimistic by

a power of 10. Even if this were not the case, for larger and larger networks one must be aware that the residual bias effects will eventually outgrow the effect of the standard deviations of the elementary roundoff errors.

8.4 U.S. Network Without Doppler Stations

If we discard the 130 Doppler stations and fix the absolute position by constraining one station, MEADE'S RANCH, to a fixed position, we are dealing with a different problem and the roundoff estimates must be revised. The coefficient matrix A changes and so does the inverse F, the right-hand side vector b and the solution vector x. It turns out that our rather crude estimates of the magnitude of the coefficients of A, b, x can be retained. Only F has to be reevaluated because the elastic properties of a network that has a single fixed point are quite different from those of a network with many absolute position observations.

We therefore reevaluate the global part of F; a pictorial representation of the two columns referring to the base quad $\phi = 39°$, $\lambda = 77°$ is given in figures 8.2a-b. These figures must be compared with figures 5.10b-c. As for the local part of F, we can retain the expression implied by eq. (5.43). However it is important that the local peaks be applied twice, once at the base quad (as before) and again at the fixed station, MEADE'S RANCH. From the peak at MEADE'S RANCH we omit the first part which accounts for ill-determined stations. The local peak at MEADE'S RANCH will cause some distortions there. Otherwise, it will result in two constants, one for each coordinate direction, which have to be added to the covariances throughout the remainder of the network. Note that figures 8.2a-b do not show the distortions caused by the local peaks. However, the common shifts caused by the additive constants are shown because we decided to consider them part of the global features of the inverse F.

Based on this modified inverse, the calculations in sections 8.1 and 8.2 have been repeated, but only for one base quad, $\phi = 39°$, $\lambda = 77°$. The results are shown in tables 8.8, 8.9, and 8.10. Compared to the estimates for the Doppler-equipped network, we see that about one power of 10 is lost for the bias as well as for the standard deviation.

TABLE 8.8.—*IBM 360 bias estimates for base quad $\phi = 39°$, $\lambda = 77°$, relying on eq. (8.8) which uses Γ counts.*

Without weight singularities	global	0.0079
	local	.0004
	Σ	.0083
Weight singularity contribution	global	.0057
	local	.0005
	Σ	.0062
Total bias		.0144

TABLE 8.9.—IBM 360 bias estimates for base quad $\phi = 39°$, $\lambda = 77°$, *relying on eq. (8.12) which uses Ξ counts.*

Without weight singularities	global	0.0028
	local	.0002
	Σ	.0030
Weight singularity contribution	global	.0021
	local	.0002
	Σ	.0023
Total bias		.0053

TABLE 8.10.—*Standard deviation estimates for base quad $\phi = 39°$, $\lambda = 77°$.*

		CDC 6600	IBM 360
Without weight singularities	global	1.1E-5	1.2E-7
	local	2.1E-6	2.3E-8
	rms mean	1.6E-5	1.7E-7
Weight singularity contribution	global	1.2E-4	1.3E-6
	local	2.8E-5	3.1E-7
	rms mean	1.7E-4	1.8E-6
Total standard deviation		2.4E-4	2.8E-6

9. ATTEMPTS TO LOWER THE ESTIMATES

Having established safe bounds for roundoff error accumulation during the solution of the normal equations of the U.S. network by using a computer similar to the CDC 6600 or the IBM 360 we feel free to speculate on how much we may have overestimated the errors. In this chapter we will take a different and less rigorous approach, one which will rely on insight, judgment, plausibility considerations, and educated guesses. No 100 percent warranty for these estimates will be given.

9.1 Review of Causes for Overestimation

In earlier chapters we repeatedly pointed out the causes for overestimating roundoff errors. Let us briefly review them:

Overestimating starts early, namely at the estimates for bias and standard deviation of the elementary roundoff errors. (Refer to the discussion in section 2.8.1.1 and 2.8.1.2.) Overestimation of elementary roundoff errors was not considered too serious, and we will not try to improve things on the elementary level.

Overestimation took place when the local roundoff errors ε_{ij} were analyzed. These are the left-side local roundoff errors of the triangular decomposition phase. The ε_{ij}'s affect the coefficients a_{ij} by means of

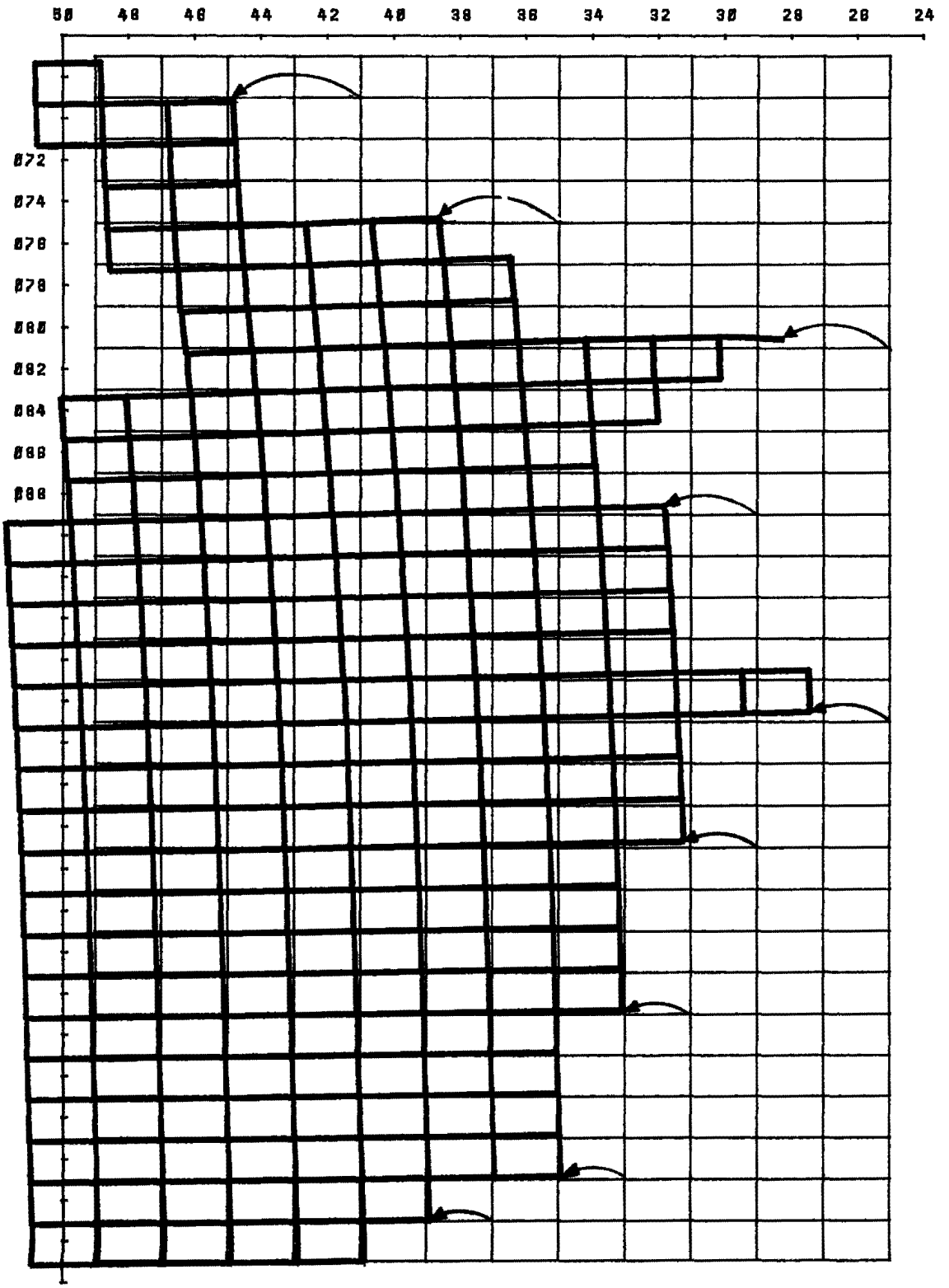

Figure 8.2a.—Covariance for net without Doppler observations. Response of network to latitude disturbance at $\varphi = 39°$, $\lambda = 77°$.

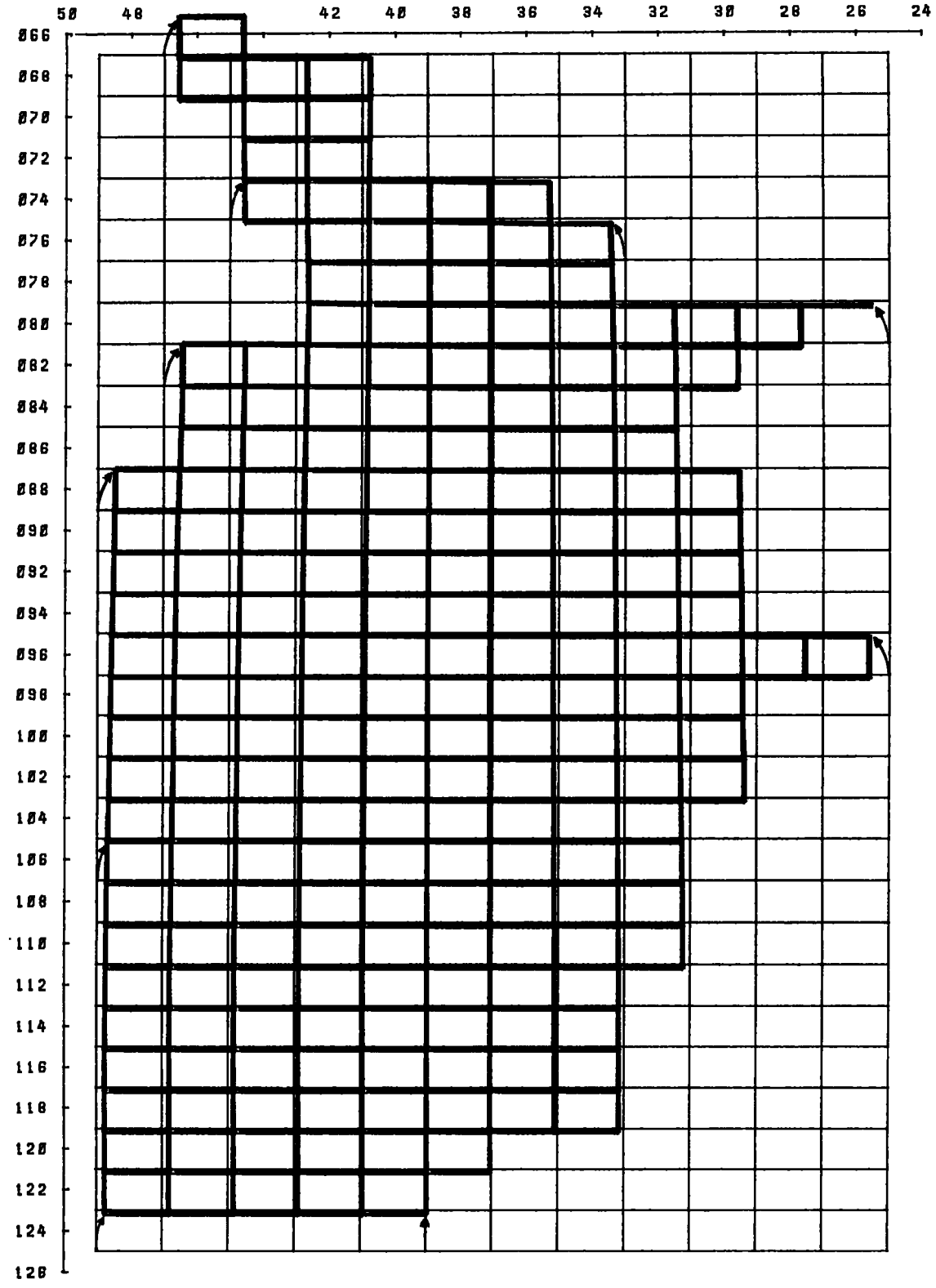

Figure 8.2b.—Covariance for net without Doppler observations. Response of network to longitude disturbance at $\varphi = 39°$, $\lambda = 77°$.

backward analysis. There are three main reasons for overestimating the ε_{ij}.

(1) We allowed for one elementary error $\varepsilon_{ijk}^{(*)}$ for the calculation of a product term $r_{ki}r_{kj}$ and another error, $\varepsilon_{ijk}^{(+)}$, for adding the product term to a previously accumulated sum. In many cases, in particular when a station is not involved in a weight singularity, the accumulated sum will dominate the product term in absolute size. $\varepsilon_{ijk}^{(*)}$ will then be dominated by $\varepsilon_{ijk}^{(+)}$ and therefore may be neglected. This means that our estimates are too large by a factor of about 2, as far as the bias $E\{\zeta_i\}$ is concerned, and by a factor of about $\sqrt{2}$, as far as the standard deviation $\sigma\{\zeta_i\}$ is concerned.

(2) Means and standard deviations of the ε_{ij}'s have been grossly overestimated in many cases. $E\{\varepsilon_{ij}\}$, $\sigma\{\varepsilon_{ij}\}$ depend on the $a_{ij} = a_{ij}^{(0)}$ and the history of the $a_{ij}^{(p)}$'s as p proceeds from 0 to $i-1$. We have overestimated the size of the $a_{ij}^{(p)}$'s particularly for most of the cases where i,j refer to coordinates of stations located at an appreciable distance. The reader is referred to section 8.1, where we tried to bound the bias for the case of removed weight singularities. A single constant c was used to bound the smallest power of the base β, which power in turn bounds $\|a\|$, the largest element of the normal equation matrix A. Hence all elements of A are essentially bounded by a single constant. Means and standard deviations of all elementary roundoff errors $\varepsilon_{ijk}^{(*)}$, $\varepsilon_{ijk}^{(+)}$, $\varepsilon_{ijk}^{(-)}$ were bounded in terms of c. To anyone who has insight into the elastic properties of a geodetic network this must appear as a waste of magnitudes. We shall try to convey such insight in subsequent sections.

(3) There is a third reason for overestimating the local roundoff errors ε_{ij}. In the case of a weight singularity, some coefficients $a_{ij}^{(p)}$ are very large. We have assumed a large roundoff error at *any* elementary operation during the transition from a_{ij} to $a_{ij}^{(i-1)}$. However, as argued in section 7.1.1, the number of bad roundings involving such a coefficient is smaller in many cases, even if the standard version of Cholesky's algorithm is used.

Finally, let us discuss the simplifying assumptions that were made when the local roundoff errors were propagated to the global ones. In particular these assumptions made the bias estimates too large.

(1) The elements f_{ij} of the inverse F and the coordinate shifts x_k were overestimated slightly. I believe that the local peaks of F were assumed to be too large. Also, the superposition of a 2m random noise level upon the x_k appears to be quite pessimistic.

(2) Alternating signs of various quantities that contributed to the error budget were neglected. In estimating the global bias, only absolute values of individual contributions were summed up. The inverse F has some elements f_{ij} that are negative. However, it is not expected that much cancellation comes from the negative f_{ij}'s. The negative elements of F are certainly dominated by the positive ones. However, negative shifts will be just as likely as positive ones. Vincenty (1976) based his datum shift computation on the average shift of a selected number of Doppler stations. This implies that the sum of residual shifts for these Doppler stations must be zero. The sum of shifts for all stations, Doppler and others, will not be exactly zero, but will be close to zero.

The most remarkable offsetting of error contributions can be expected to come from the alternating signs of the coefficients $a_{ij}^{(p)}$. The row sums of A, $A_{22}^{(p)}$ will be shown to have a tendency to cancel out. This phenomenon, together with the smooth variation of f_{ij}, x_k, causes an offsetting of bias contributions. This offsetting is most pronounced if the modified Cholesky version is used.

(3) After careful study the correlation pattern of the global roundoff errors will reveal that the roundoff errors of closely spaced stations will be strongly correlated. This implies that the relative accuracy of closely spaced stations is less perturbed by roundoff than the relative accuracy of stations spaced farther apart.

9.2 Sign Pattern of the Coefficients $a_{ij}^{(p)}$

Let us assume temporarily that the networks reference surface is a plane. Assume further that no Doppler measurements are performed and that no station is fixed by constraint. All row sums of the normal equation matrix A would vanish. Even the row sums of the partially reduced matrices $A_{22}^{(p)}$ would vanish:

$$\sum_{j=p+1}^{n} a_{ij}^{(p)} = 0\,, p = 0, \ldots, n-1. \qquad (9.1)$$

The reason for the vanishing row sums of A is that only relative measurements were performed. A common shift of all stations of the network must go undetected by the normals. The following must be true:

$$\sum_{j=1}^{n} a_{ij}x_j = \sum_{j=1}^{n} a_{ij}\ (x_j + c).$$

This is equivalent to eq. (9.1) for $p=0$. The reader should notice that the argument extends even to the case of different shifts in the two coordinate directions. To extend the argument to the partially reduced matrices $A_{22}^{(p)}$, one simply has to keep in mind that any row of $A_{22}^{(p)}$ is a linear combination of the rows of A.

Now let us consider the case of absolute station positions as obtained by Doppler measurements. Equation (9.1) then fails to hold true in general. It

fails when i refers to a Doppler station or to a station connected to a Doppler station. Recall that connections are established by either direct observation, by directional coobservation, or fill-in. It is remarkable that eq. (9.1) continues to hold true for all coordinates i referring to stations that are neither Doppler stations nor connected to Doppler stations. Because there are about 130 Doppler stations, the number of original stations for which eq. (9.1) is true (*i.e.*, for $p=0$) is originally quite large. It decreases as p increases. However, for stations situated at greater distances from a Doppler station, eq. (9.1) will be approximately true.

The validity of eq. (9.1) is further impaired by the fact that the reference surface of our network is not a plane. Again we expect that the deviations of the ellipsoid from a plane will not cause large discrepancies. Locally, the ellipsoid is well approximated by a plane, and the larger coefficients refer to pairs of stations at a close distance.

The preceding discussion can be summarized by stating that there is a tendency for the row sums (9.1) to cancel out. In the next subsection we will examine how this can affect the accumulation of bias-type roundoff errors.

9.3 Offsetting Bias Contributions

Recall the basic formula (4.35) for propagating local biases to the global bias. Restricted to left-side errors, the formula states

$$E\{\zeta_i\} = -\sum_{j=1}^{n} \sum_{k=1}^{n} f_{ij}\, x_k\, E\{\varepsilon_{jk}\}. \tag{9.2}$$

The local errors ε_{jk}, having the symmetry property $\varepsilon_{jk} = \varepsilon_{kj}$, depend on the history of the coefficients $a_{ij}^{(p)}$, as p goes from 0 to $i-1$. A diagonal coefficient $a_{ii}^{(p)}$ is always positive. It decreases as p increases. Because most roundoff errors occur during the accumulation of the product sum over r_{ki}^2, which is subtracted from a_{ii} to yield $a_{ii}^{(i-1)}$, and because true chopping makes the sum smaller, $a_{ii}^{(i-1)}$ will actually be too large in many cases. The local roundoff error ε_{ii} tends to be positive. If the net were a leveling net, all off-diagonal elements $a_{ij}^{(p)}$ would be negative or zero. By reasoning as in the case for the diagonals, we conclude that ε_{ij}, $i \neq j$, tends to be negative. The U.S. network is not a leveling net. Hence not all $a_{ij}^{(p)}$ can be negative. However, as shown in the previous section, the row sums (9.1) tend to cancel out. Can we conclude from this that the row sums of the bias matrix ε show a tendency to cancel? Not directly, but we will try to find supporting evidence.

First, we observe that of all the coefficients $a_{ij}^{(p)}$ in row i of the original or partially reduced normals, only those coefficients which connect i's station to a station in the vicinity will be of an appreciable size. This will be discussed in more detail in a later section. Meanwhile we will take it for granted. Next we try to analyze the bias

$$E\{\varepsilon_{ijk}\} = E\{\varepsilon_{ijk}^{(\circ)}\} + E\{\varepsilon_{ijk}^{(+)}\}$$

occurring during the evaluation of a product term $r_{ki}r_{kj}$ and its addition to the previously accumulated sum. Recall in this context

$$\sum_{l=1}^{k} r_{li}\, r_{lj} = a_{ij} - a_{ij}^{(k)}. \tag{9.3}$$

The bias of the local roundoff error ε_{ijk} depends on the absolute and relative size of the three quantities $a_{ij} - a_{ij}^{(k-1)}$ (product sum before), $r_{ki}r_{kj}$ (product term), and $a_{ij} - a_{ij}^{(k)}$ (product sum afterwards). Unfortunately, the dependency is nonlinear, involving the next larger integer powers of the base β. Let us study two important cases:

(1) Here i is the coordinate of a station not involved in a high precision measurement. As argued in section 7.1.1, the coefficients $a_{ij}^{(p)}$ will not undergo dramatic changes, and a smooth transition will occur from a_{ij} to $a_{ij}^{(i-1)}$. This implies that most of the time the product term $r_{ki}r_{kj}$ will be small compared to the sum. The elementary roundoff errors $\varepsilon_{ijk}^{(\circ)}$ then can be neglected. $E\{\varepsilon_{ijk}^{(+)}\}$ will be about $c_{ij}/2$, where c_{ij} is the smallest signed power of β bounding $a_{ij} - a_{ij}^{(k)}$. The sum

$$\sum_{j=p+1}^{n} (a_{ij} - a_{ij}^{(k)}) \tag{9.4}$$

is near zero. But can we expect that the corresponding sum

$$\sum_{j=p+1}^{n} c_{ij} \tag{9.5}$$

will be near zero? If the base is 2, we can do so with some justification. But when the base is 16, we generally expect that the positive c_{ii} caused by the diagonal element is more than offset by the sum over the negative c_{ij}'s, some of which can be as large as c_{ii} itself. In other words, we can say that whatever the value of the sum (9.4), the sum (9.5) can be larger by a factor amounting to several multiples of the base β. The sum (9.5) can still be small enough to make the idea of bias offsetting attractive.

Small row sums will cause small global roundoff errors if the quantities $f_{ij}x_k$ in (9.2) are uniform, but overall uniformity cannot be expected. However, if i is at a larger distance from j (*i.e.* i's station is at a larger distance from j's station), f_{ij} will be fairly uniform. As long as k is near j, x_k will not vary greatly. Hence, for a large number of important contribu-

tions to the sum (9.2), namely those resulting from j farther away from i, and k in the vicinity of j, uniformity will prevail. Note that k near j implies that $a_{ij}^{(p)}$ and hence $E\{\varepsilon_{jk}\}$ will be relatively large. The size of these quantities decreases as k goes farther away from j.

(2) Suppose now that i's station, call it P, is tied to j's station, which we call Q, by a high precision measurement. Let P precede Q in the elimination. This case was discussed in great detail in section 7.1.1. We take the notation from there and let $i, i+1, j, j+1$ refer to latitude and longitude of P, Q respectively. Assume first that the standard version of the N.G.S. Cholesky's algorithm is in effect. The large coefficients $a_{ii}^{(p)}$, $a_{i,i+1}^{(p)}$, $a_{ij}^{(p)}$, $a_{i,j+1}^{(p)}$ of eq. i will never drop sharply in size before elimination reaches this equation. As argued in section 7.1.1, there will be two bad roundings per coefficient. The large coefficients come in pairs; $a_{ii}^{(p)}$ is nearly equal to $-a_{ij}^{(p)}$, and $a_{i,i+1}^{(p)}$ is nearly equal to $-a_{i,j+1}^{(p)}$. As a consequence, the biases in equation i will offset themselves nicely. There are no qualms about considerations such as the "next integer power of the base β." The three large coefficients of row $i+1$, *i.e.*, $a_{i+1,i+1}^{(p)}$, $a_{i+1,j}^{(p)}$, $a_{i+1,j+1}^{(p)}$ will each suffer three bad roundings. The biases of the first and third offset each other nicely. However, the bias of $a_{i+1,j}^{(p)}$, caused by three bad roundings, is offset only by the bias caused by two bad roundings at $a_{i+1,i}^{(p)} = a_{i,i+1}^{(p)}$. Hence the offsetting is incomplete. Things get worse when we come to eqs. $j, j+1$. Offsetting is incomplete even if Q follows P immediately, *i.e.* if $j=i+2$. If other stations are to be eliminated between P and Q, this is unfortunate, and there is no hope for offsetting the biases of many bad roundoff errors.

On the other hand, if the improved version of the NGS Cholesky's algorithm is used, there will be exactly two bad roundings per large coefficient. Offsetting the biases then will be ideal.

9.4 Reexamining the Row Sum Norms.

Recall the definition of the row sum norm $\|a_{i\cdot}^{(p)}\|_1$ in eq. (7.4), *i.e.:*

$$\|a_{i\cdot}^{(p)}\|_1 = \sum_{j\geq i} |a_{ij}^{(p)}|. \tag{9.6}$$

A bound for these norms was derived in proposition 7.1. The reader may have wondered why we did not use this bound in section 8.1 to improve the estimate of the global bias of the homogeneous network. Instead, we have been working with a single bound $\|a\|$ for all elements of A. The reason for not using proposition 7.1 was that for the U.S. net no improvement was attainable over the other method. To safely apply proposition 7.1 on the IBM 360, one must allow for another safety factor of $\beta = 16$. This spoils the estimate so much that it becomes inferior to the one obtained by the more primitive method based on a single bound. Only in the case of a much larger network would the asymptotic superiority of the other method result in an improvement. (See sec. 11.2.)

However, I firmly believe that the bound on $\|a_{i\cdot}^{(p)}\|_1$ obtained in proposition 7.1 does not reflect the true asymptotic behavior of these row sum norms. I believe that the true asymptotic behavior is more favorable and that, at least in most cases, the following is true for some constant γ:

$$\|a_{i\cdot}^{(p)}\|_1 \leq \gamma\, a_{ii}. \tag{9.7}$$

Exceptions to this rule occur, particularly in a few instances at the very end of the triangular decomposition phase.

Intuitive support for the assumption that $\|a_{i\cdot}^{(p)}\|_1$ can never get too large in comparison to a_{ii}, in fact, even in comparison to $a_{ii}^{(p)}$, which is smaller than a_{ii}, comes from the geodetic interpretation of the $a_{ij}^{(p)}$'s. We write

$$\|a_{i\cdot}^{(p)}\|_1 = \sum_{j\geq i} a_{ij}^{(p)} = a_{ii}^{(p)} + \sum_{j>i} a_{ij}^{(p)} \operatorname{sign}(a_{ij}^{(p)}). \tag{9.8}$$

The ratio

$$-\sum_{j=i+1}^{n} a_{ij}^{(p)} \operatorname{sign}(a_{ij}^{(p)}) / a_{ii}^{(p)} \tag{9.9}$$

is, according to the geodetic interpretation of the coefficients $a_{ij}^{(p)}$, the shift coordinate i suffers when coordinates j, $j>i$, are displaced by $\operatorname{sign}(a_{ij}^{(p)})$, while coordinates k, $p<k<i$ are fixed, and coordinates k, $1\leq k\leq p$, $k=i$ are allowed to vary freely. All fixing and shifting are done with respect to the adjusted position of the original network. Displacing some coordinates j by amounts of $+1$ or -1, while keeping some coordinates fixed and allowing others to vary freely, should not cause coordinate i to wander too far. Unless the network has very funny elastic properties and is poorly anchored, I do not expect that coordinate i will move more than a few units. This implies that the ratio $\|a_{i\cdot}^{(p)}\|_1/a_{ii}^{(p)}$ will not exceed a few units.

Remark: For a leveling net the ratio is restricted to the range

$$0 \leq \|a_{i\cdot}^{(p)}\|_1 / a_{ii}^{(p)} \leq 2. \tag{9.10}$$

This is a consequence of the maximum principle valid for leveling networks. Refer to Bartelme and Meissl (1977) where roundoff error propagation is treated for a leveling network.

The analogy of a geodetic network to a mechanically elastic system suggests even more strongly that

(9.7) will be valid in most cases. As mentioned in the remarks of section 3.4, the $a_{ij}^{(p)}$, $p<i,j\leqslant n$, are the coefficients of a stiffness matrix $A_{22}^{(p)}$. This stiffness matrix refers to an elastic system in which coordinates k, $p<k\leqslant n$ are fixed while coordinates k, $1\leqslant k\leqslant p$ are allowed to vary freely. The coefficient $a_{ij}^{(p)}$ is the force exerted on coordinate i when coordinate j is displaced by one unit. Because of symmetry, the roles of i,j can be interchanged. The interpretation is also valid for $i=j$. Let us displace coordinate i by one unit. The $a_{ij}^{(p)}$ are then the reactional forces felt by i itself and by the other fixed coordinates j, $j>i$. Then $\|a_i^{(p)}\|_1$, as given by (9.2), is the sum of the magnitudes of all these reactional forces. The force at i is $a_{ii}^{(p)}$. Again, one should not expect that the sum of the magnitudes of all the reactional forces $a_{ij}^{(p)}$, $j>i$ to exceed the force $a_{ii}^{(p)}$ at i by a considerable amount. If it did, the network would either be poorly anchored by fixing the coordinates j, $j>i$, or the network would otherwise act like a strange mechanical machine.

Let us take a look at figures 9.1a-b which illustrate our line of reasoning. To make things more clear, both figures refer to a free distance network. Such a network can be viewed as a discrete analogue to a (not necessarily homogeneous) rubber disk which slides on a smooth plane surface. The circles denote fixed stations. (The rubber disk is pinned down to the supporting surface there.) Fixing of stations is done with respect to the equilibrium position, except for station coordinate i which is displaced by one unit from its equilibrium position. Figures 9.1a-b illustrate the reactional forces that may arise at the fixed stations. The reactional forces must have a vanishing resultant force and a vanishing resultant moment. Otherwise they will adjust themselves so that the distortional energy of the elastic system is minimized. (Geodetically, the distortions cause residuals to the observables. The weighted sum of the squares of the residuals is minimized.)

Figure 9.1(a) shows a network with a poorly anchored station, belonging to coordinate i, with respect to the other stations that belong to the other fixed coordinates. The requirement of vanishing resultant moment causes some reaction forces to be unduly large. Hence the sum of absolute values of the reaction forces considerably exceeds the force acting at the displaced coordinate. Consequently $\|a_i^{(p)}\|_1$ is much larger than $a_{ii}^{(p)}$, and can even be much larger than a_{ii}. The network in figure 9.1(b) is well anchored. The sum of the absolute reactional forces does not exceed the reactional force at the displaced coordinate i by a large amount. $\|a_i^{(p)}\|_1$ will not be much larger than $a_{ii}^{(p)}$.

The most typical situation during the elimination

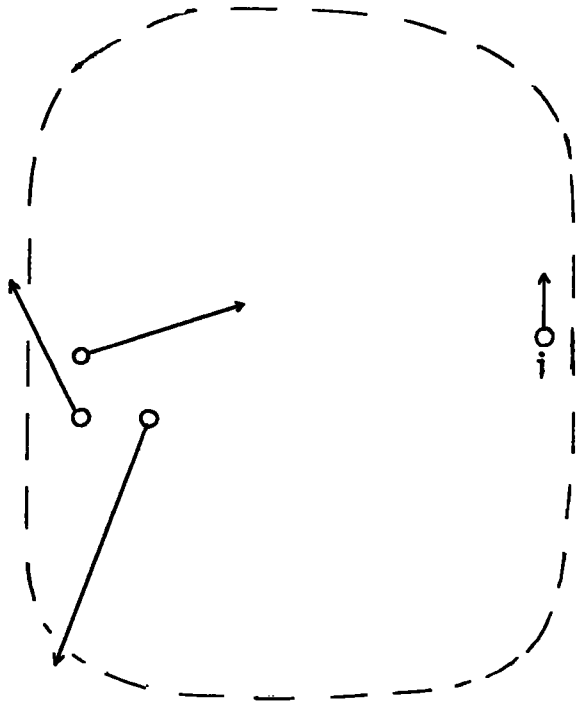

Figure 9.1(a).—Reaction forces in a poorly anchored free-distance network.

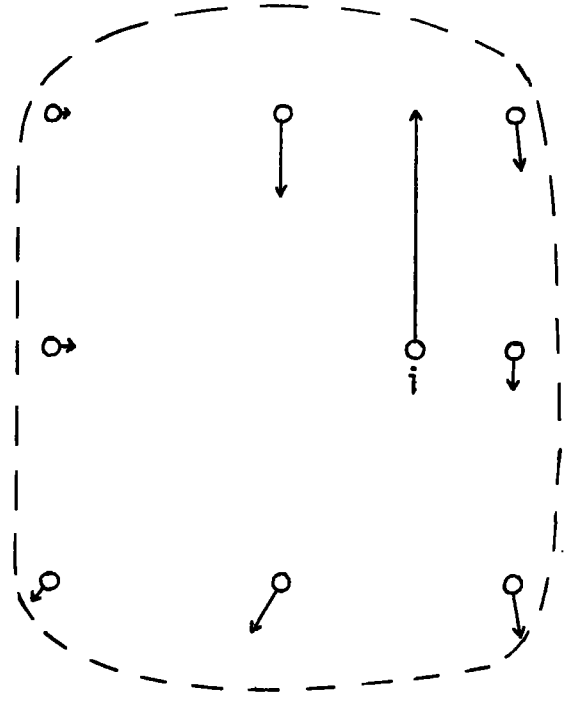

Figure 9.1(b).—Reaction forces in a well anchored free-distance network.

procedure of the U.S. network will be the partial reduction within a block. The block may be at the lowest level or at an intermediate level. In both cases, the outer junction nodes act as fixed stations (circles). The anchoring of such a configuration clearly resembles the case illustrated in figure 9.1(b), which supports the reasoning that $\|a_i^{(p)}\|_1$ will barely exceed $a_{ii}^{(p)}$.

Situations similar to the one illustrated in figure 9.1(a) may occur only at the very end of the triangular decomposition phase. Actually, the situation is not as bad as shown in figure 9.1(a), which refers to a pure distance network. In a distance network, all reaction moments must be taken by the fixed stations. However, in the presence of azimuths and Doppler stations, a large part of the reactional moments will be taken by these measurements. (Azimuths counteract rotational movements of portions of the rubber disk; Doppler stations counteract positional displacements.)

Let us briefly turn to the row sums $\|a_i^{(p)}\|_2$ defined by eq. (7.5):

$$\|a_i^{(p)}\|_2 = \sqrt{\sum_{j \geq i} |a_{ij}^{(p)}|^2}. \tag{9.11}$$

The corresponding bound derived in proposition 7.1 is much better than in the case of $\|a_i^{(p)}\|_1$. This bound certainly reflects the true asymptotic behavior of $\|a_i^{(p)}\|_2$. The only reason for concern is the constant λ_{MAX}, which must be chosen in agreement with the largest coefficients of A. This means, it must be chosen in agreement with the most heavily weighted observations. For a coordinate i, referring to a station not involved in a weight singularity, the constant λ_{MAX} is replaced by something much smaller.

Having exhibited qualitative reasons for my belief in the favorable behavior of the two row sum norms, I find it much harder to make quantitative statements. I feel much like a meteorologist who is supposed to predict tomorrow's weather from today's weather maps, based on his experience and on computer models which are highly simplified and rely on our present incomplete knowledge of the physics of the atmosphere.

Table 9.1 is the result of some statistics which I derived from test adjustments of small portions of the network. I admit that table 9.1 is also the result of some guesswork. I have tried to partition the diagonals a_{ii} into four different size classes. The boundaries are denoted by $\|a\|_r$, $r = 1,2,3,4$. The fractions of diagonals falling into the various size classes are denoted by ψ_r. For any size class, we have allowed for separate constants by which the two row sum norms $\|a_i^{(p)}\|_1$, $\|a_i^{(p)}\|_2$ may exceed the diagonals a_{ii}. These two constants are denoted by γ_r, δ_r.

TABLE 9.1.—*A rough estimate of size distribution for diagonal coefficients a_{ii} and the factors γ_r, δ_r by which the two row sum norms exceed a_{ii}.*

r	$\|a\|_r$ m^{-2}	$\|a\|_{r+1}$ m^{-2}	ψ_r	γ_r	δ_r
1	4.5E6	2.8E5	0.10	2.5	2
2	2.8E5	1.8E4	.15	5	2.5
3	1.8E4	1.1E3	.65	10	5
4	1.1E3	0	.10	50	10

For instance, row 1 indicates that 10 percent of the diagonals a_{ii} are assumed within the range of $4.5E6 \geq a_{ii} \geq 2.8E5$, and that $\|a_i^{(p)}\|_1 \leq 2.5\ a_{ii}$, and $\|a_i^{(p)}\|_2 \leq 2\ a_{ii}$ in these cases. (Coefficients a_{ii} are measured in units of m^{-2}.)

9.5 Aiming at Realistic Bias Estimates for the IBM 360

Starting with eq. (4.35), with the right-hand side contribution omitted, we use

$$E\{\zeta_i\} = -\sum_{j=1}^{n} \sum_{k=1}^{n} f_{ij}\, x_k\, E\{\varepsilon_{jk}\}. \tag{9.12}$$

Using the bound $\|x\| = 10$ on the x_k, we arrive at:

$$E\{\zeta_i\} \leq \|x\| \sum_{j=1}^{n} |f_{ij}| \sum_{k=1}^{n} |E\{\varepsilon_{jk}\}|. \tag{9.13}$$

Our next step is to estimate the sum

$$\sum_{k=1}^{n} |E\{\varepsilon_{jk}\}|. \tag{9.14}$$

We recall that the local roundoff error ε_{jk} is the superposition of $2\,\mu_{jk} + 1 \cong 2\,\mu_{jk}$ elementary roundoff errors $\varepsilon_{jk}^{(\sigma l)}$. We assume that $|E\{\varepsilon_{jk}^{(\sigma l)}\}|$ is bounded in terms of $|a_{jk}^{(p)}|$ as

$$|E\{\varepsilon_{jk}^{(\sigma l)}\}| \leq \phi_j\, |a_{jk}^{(p)}| * 16^{-14}/2. \tag{9.15}$$

To the factor ϕ_j, a value will be assigned later. We get

$$\sum_{k=1}^{n} |E\{\varepsilon_{jk}\}| \leq \phi_j \sum_{k=1}^{n} |a_{jk}^{(p)}| * 2\,\mu_{jk} * 16^{-14}/2.$$

Remembering that $\mu_{jk} \leq \mu_{jj}$, we get

$$\sum_{k=1}^{n} |E\{\varepsilon_{jk}\}| \leq \phi_j\, \mu_{jj} \sum_{k=1}^{n} |a_{jk}^{(p)}| * 16^{-14}. \tag{9.16}$$

Note that p is still unspecified in this expression. We introduce

$$\gamma_j = \underset{p}{Max}\ \{\sum_{k=j}^{n} |a_{jk}^{(p)}|/a_{jj}\}. \tag{9.17}$$

This allows us to write (9.16) as

$$\sum_{k=1}^{n} |E\{\varepsilon_{jk}\}| \leq \phi_j\, \mu_{jj}\, \gamma_j\, a_{jj} * 16^{-14}. \tag{9.18}$$

Inserting this into (9.13), we arrive at

$$|E\{\zeta_i\}| \leq \|x\| \sum_{j=1}^{n} \phi_j\, f_{ij}\, \mu_{jj}\, \gamma_j\, a_{jj} * 16^{-14}. \tag{9.19}$$

Here we have used $\gamma_j a_{jj}$ as a bound on $\sum_{k\geq 1} |a_{jk}^{(p)}|$ rather than as a bound on $\sum_{k\geq j} |a_{jk}^{(p)}|$. This is admissible because for $j \leq p$, $a_{ij}^{(p)} = 0$, and because (9.17) must also hold for $j = p+1$.

Referring to table 9.1, we evaluate (9.19) separately for the four size classes. We get

$$|E\{\zeta_i\}| \leq \|x\| \sum_{r=1}^{4} \phi_r \gamma_r \|a\|_r * \sum_{\substack{a_{jj} \text{ in} \\ \text{class } r}} f_{ij}\mu_{jj} * 16^{-14}. \quad (9.20)$$

or

$$|E\{\zeta_i\}| \leq \|x\| \sum_{r=1}^{4} \phi_r \|a\|_r \gamma_r \psi_r \sum_{j=1}^{n} f_{ij}\mu_{jj} * 16^{-14}. \quad (9.21)$$

We evaluate the following sums which we call Ψ_i:

$$\sum_{j=1}^{n} f_{ij}\mu_{jj} = \Psi_i. \quad (9.22)$$

This was actually done before during the evaluation of formulas (8.21) and (8.22). For all i's referring to coordinates situated in a 2° × 2° quad denoted by ϱ, we found that

$$\sum_{j=1}^{n} f_{ij}\mu_{jj} \cong \sum_{\chi} f_{\rho\chi}^{(global)} \Pi_\chi + \text{peak contribution due to } f_{ij}^{(local)}. \quad (9.23)$$

No new computations are necessary. We use the intermediate results obtained during the evaluation of the entries in table 8.4. The values are listed in table 9.2 and are denoted by Ψ_ρ.

TABLE 9.2.—*Evaluation of eq. (9.23)*

ϕ	λ	Global m^2	Local m^2	Ψ_ρ m^2
39	77	0.792E6	0.275E6	0.107E7
47	69	.114E7	.605E5	.120E7
47	121	.752E6	.225E6	.977E6
41	97	.576E6	.715E5	.647E6
35	111	.576E6	.509E5	.627E6

It remains for us to assign values to the factors ϕ_r. These values are supposed to be a compromise between all the various effects that were reviewed and discussed in sections 9.1 to 9.4. Our chosen values are listed in table 9.3. We motivate their choice as follows:

TABLE 9.3.—Choice of factors ϕ_r

r	ϕ_r
1	0.125
2	.25
3	1.00
4	1.00

Using such considerations as the "next power of the base $\beta = 16$," begin with a factor of 16. The factor 0.5 is applied because, in most cases, one elementary error out of $\varepsilon_{ijk}^{(*)}$, $\varepsilon_{ijk}^{(*)}$ is effective. Another factor of 0.5 is applied to compensate for the overestimation of the x_k by $\|x\| = 10$. The factor 0.25 is supposed to take care of cancellations resulting from alternating signs. This already explains the values for $r = 3$ and 4. For $r = 1$ and 2, *i.e.*, for the two largest size classes where weight singularities are contributed, we allow another factor 0.25 to account for the large coefficients where only a fraction of all the roundings is bad. Finally, we apply another factor of 0.5 at $r = 1$ because not all the large coefficients are equal to the upper bound of this size class.

Based on these assumptions, we arrive at the following formula:

$$|E\{\zeta_\rho\}| \leq 10 \sum_{r=1}^{4} \phi_r \|a\|_r \gamma_r \psi_r \Psi_\rho * 16^{-14}. \quad (9.24)$$

The evaluation results are given in table 9.4. The bounds refer to the first iteration where coordinate shifts exceeding 10 meters are anticipated.

TABLE 9.4.—*Attempted realistic estimates for global bias-type roundoff errors on the IBM 360.*

ϕ	λ	Bound m
39	77	4.8E-5
47	69	5.2E-5
47	121	4.2E-5
41	97	2.8E-5
35	111	2.8E-5

It is seen that the bounds are smaller by about two powers of 10 than the safe bounds obtained in chapter 8. They indicate that about 4 to 5 correct digits can be recovered during any iteration.

9.6 Remarks on Standard Deviations of the Global Errors

A similar calculation could be performed to obtain improved estimates of $\sigma\{\zeta_i\}$. We will not document such a calculation here, but merely mention that the improvement would hardly exceed one power of 10. It is interesting to ask for the reason for such a meager result.

First of all, there are no offsetting effects. Second, one has to keep in mind that $\sigma\{\zeta_i\}$, being a mean square average, enhances the contribution of the larger roundoff errors relatively more than $E\{\zeta_i\}$ does. This is theoretically obvious; and it is also indicated in chapter 8, where the various tables show, that the contributions of the local peaks of the covariance are relatively larger in the case of the standard deviations. Therefore, we should try to improve these local contributions, and it is here that a factor of about 1/10 could be gained. Simply by being less pessimistic about the number and size of the large

elements of A, one could lower the estimates of $\sigma\{\zeta_i\}$. All other considerations which, as in the case of the bias-type estimates of the previous section, lead to a replacement of the Γ-type counts by Π-type counts, have little effect on improving σ-type estimates.

9.7 Global Roundoff Errors of the Relative Position of Two Closely Situated Stations

To illustrate let P, Q denote two stations which are located at a distance below 20 km. The relative position of the two stations, *i.e.*, the difference vector between P and Q, will generally be less perturbed by roundoff than the absolute positions of P, Q. This can be inferred from eqs. (4.33), (4.35), and (4.36a), and from the properties of the inverse. Let i_1, i_2 refer to the latitudes of the two stations P, Q. Let us investigate the global roundoff error $\zeta_{i_1 i_2}$ suffered by the latitude difference. We have

$$\zeta_{i_1 i_2} = \zeta_{i_2} - \zeta_{i_1}. \tag{9.25}$$

Again, we restrict our attention to the left-side errors during triangular decomposition. Equation (4.33) then implies that

$$\zeta_{i_1 i_2} = -\sum_{j=1}^{n} \sum_{k=1}^{n} (f_{i_2 j} - f_{i_1 j})\, x_k \varepsilon_{jk}. \tag{9.26}$$

For a coordinate j whose station R is far away from P, Q it will hold that $f_{i_1 j}$ nearly equals $f_{i_2 j}$. The geodetically minded reader may infer this from f_{ij}, which is composed of a smoothly varying global component and a quickly decaying local one. The elastostatic interpretation of f_{ij} supports our reasoning even more strongly. In this interpretation, f_{ij} is the shift suffered by coordinate j if a unit force is applied to coordinate i. Therefore, it is assumed that before the application of the unit force, a state of equilibrium has been reached. If R is far away from P, Q, the principle of St. Venant guarantees that shifts of j due to i_1 and i_2 will be nearly equal.

Quantitative statements are more difficult to make. From equations analogous to (4.35) and (4.36a) we deduce

$$E\{\zeta_{i_1 i_2}\} = -\sum_{j=1}^{n} \sum_{k=1}^{n} (f_{i_2 j} - f_{i_1 j})\, x_k \varepsilon_{jk}, \quad \varepsilon_{jk} = \varepsilon_{kj} \tag{9.27}$$

$$\begin{aligned} \sigma\{\zeta_{i_1 i_2}\} = {} & \sum_{j=1}^{n} (f_{i_2 j} - f_{i_1 j})^2 x_j^2 \sigma^2\{\varepsilon_{jj}\} + \\ & + \sum_{j=1}^{n} \sum_{k=j+1}^{n} [(f_{i_2 j} - f_{i_1 j}) x_k + \\ & + (f_{i_2 k} - f_{i_1 k}) x_j]^2 \sigma^2\{\varepsilon_{jk}\}. \end{aligned} \tag{9.28}$$

Only the differences $f_{i_2 j} - f_{i_1 j}$ enter these formulas. As argued above, these differences will be small if j is farther away from i_1, i_2. Referring to table 9.2, we can neglect the global contributions.

The resulting improvement is not yet impressive because the local contributions are of the same magnitude as the global ones. There are three reasons why the local contributions will also be subdued.

(1) $f_{i_2 j} - f_{i_1 j}$ will also be smaller than f_{ij} for many j's in a close vicinity. Local adjustments indicate that the relative accuracy is better than the global one by about one power of 10. Let us cautiously assume that an improvement of 1/2.5 is obtained in this way for bias type estimates.

(2) The local estimates given in table 9.2 are still pessimistic. We may disregard the loose chunks of the network whose relative accuracy is of little interest. The remaining local peaks of f_{ij} are certainly smaller in areas of dense control. However, it is in these areas where most calculations are done. Let us assume that as a result of this phenomenon another improvement of 1/2.5 occurs for bias type estimates. (One should bear in mind that this improvement also occurs to the global accuracy of stations. However, here the contribution of the global part of f_{ij} dominates the contribution of the local part, whereas in the case of relative accuracy the contribution of the global part of f_{ij} is missing.)

(3) The local contributions were estimated with the assumption that we are dealing with a station situated on or near a high-level boundary. Such a station will be eliminated during the later stages of triangular decomposition, or it will have stations in its vicinity which are eliminated at a later stage. This will make the local contribution of the local roundoff error large since the Π and Γ counts for the surrounding local area are large. It follows that for stations which are not close to a higher block level boundary, an additional improvement of relative accuracy will occur which may amount to one or two powers of 10 for the bias.

Let us summarize the preceding discussion by stating that the bias of the relative roundoff error of two closely situated stations may be smaller than the global bias by one to three powers of 10. The improvement of the standard deviation will not be as significant and is estimated not to exceed one power of 10.

10 ROUNDOFF EXPERIMENTS

10.1 Moose-Henriksen Network

Data from Moose and Henriksen (1976) were used to perform some detailed roundoff experiments. Section 5.4 already gives some statistics on this portion of the U.S. network which covers areas of Mississippi, Louisiana, and Alabama. The network version

used in our experiments had 1336 triangulation stations, of which five were Doppler stations. A total of 73 distances and 25 azimuths were included. A sketch of the network is shown in figure 10.1.

In chapter 5 we previously discussed some of the computational results obtained by Moose and Henriksen (1976). Now we will refer to additional calculations which used only input data of the Moose-Henriksen network adjustment. These additional calculations were done by W. H. Dillinger. Some temporary changes to the NGS Cholesky subroutine were made by R. H. Hanson to provide statistical information relevant to the accumulation of roundoff errors. Programing support was also given by J. F. Isner.

10.1.1 Purpose and design of the roundoff experiments

The experiments were carried out on the CDC 6600. This computer offered the advantage of simulating unbiased as well as biased arithmetic in an easy way. A further advantage was that the precision could be easily extended to allow one number to be represented by two computer words, each having 60 bits. The results of this double precision run could be used as an absolute basis of comparison. They were practically as good as the mathematically exact solution.

The purpose of the experiments was to check the following items:

(1) The size distribution of the coefficients of A, b of the original and of the partially reduced normal equations. As pointed out in chapters 7,8, and 9, it is essential for our estimates that not too many coefficients of A be very large and that the number of large coefficients does not significantly increase during the elimination procedure. Furthermore, it could be verified that the right hand sides b are actually well behaved and do not constitute a critical factor in the roundoff analysis.

(2) The number of operational steps. By counting the number of product terms $r_{ki}r_{kj}$ evaluated during the triangularization phase and the percentage of these terms which resulted in a nonzero value, we obtained a realistic picture of the number of nontrivial operational steps needed to partially reduce a subnetwork of some 1,300 stations. The counts could be compared with those predicted by our idealized model described in chapter 6.

(3) Check of the validity of the roundoff model. At least in the case of a small realistic example of the U.S. network it could be shown that our statistical assumptions about the behavior of the elementary roundoff errors were sound and that the linear model used to propagate these roundoff errors was reasonable.

The following computer runs were made.

(A) Adjustment of the network as a whole in double precision. The shifts obtained provided the desired absolute basis for comparison.

(B) Adjustment of the network as a whole with standard (*i.e.*, not truly rounding) arithmetic. Comparison with the results of (A) yielded the true roundoff errors for the biased CDC 6600.

(C) Adjustment of the network as a whole with truly rounding arithmetic. Comparison with the results of (A) yielded the true roundoff errors for the unbiased CDC 6600.

(D) Adjustment by the Helmert blocking technique. A partition was used which had 37 first-level blocks, 13 second-level blocks, 4 third-level blocks, and 1 fourth-level block. This adjustment was done only once, with the truly rounding instructions set into effect.

During adjustments (B) through (D), the above mentioned statistics on size and distribution of nonzero elements were performed. In addition, we attempted to obtain a bound on the bias and on the standard deviation of the global roundoff errors by calculating mean and standard deviation of the left-side local roundoff errors ε_{ij} from the actual numerical values of the partial sums

$$\sum_{l=1}^{k} r_{li}\, r_{lj} \tag{10.1}$$

as they were available during the Cholesky reduction. The details of this a posteriori roundoff analysis are described in the next section.

10.1.2 A posteriori roundoff error analysis

The reader is reminded of our discussion of elementary roundoff errors $\varepsilon_{ijk}^{(*)}$, $\varepsilon_{ijk}^{(+)}$ as they arise and accumulate during the evaluation of the product sums

$$\sum_{k=1}^{i-1} r_{ki}r_{kj}. \tag{10.2}$$

When a product term $r_{ki}r_{kj}$ is evaluated, an elementary roundoff error $\varepsilon_{ijk}^{(*)}$ occurs. When the product term is added to the partial sum (10.1), an error $\varepsilon_{ijk}^{(+)}$ occurs. In most cases the partial sum will be larger than the term added. Hence $\varepsilon_{ijk}^{(+)}$ will dominate $\varepsilon_{ijk}^{(*)}$.

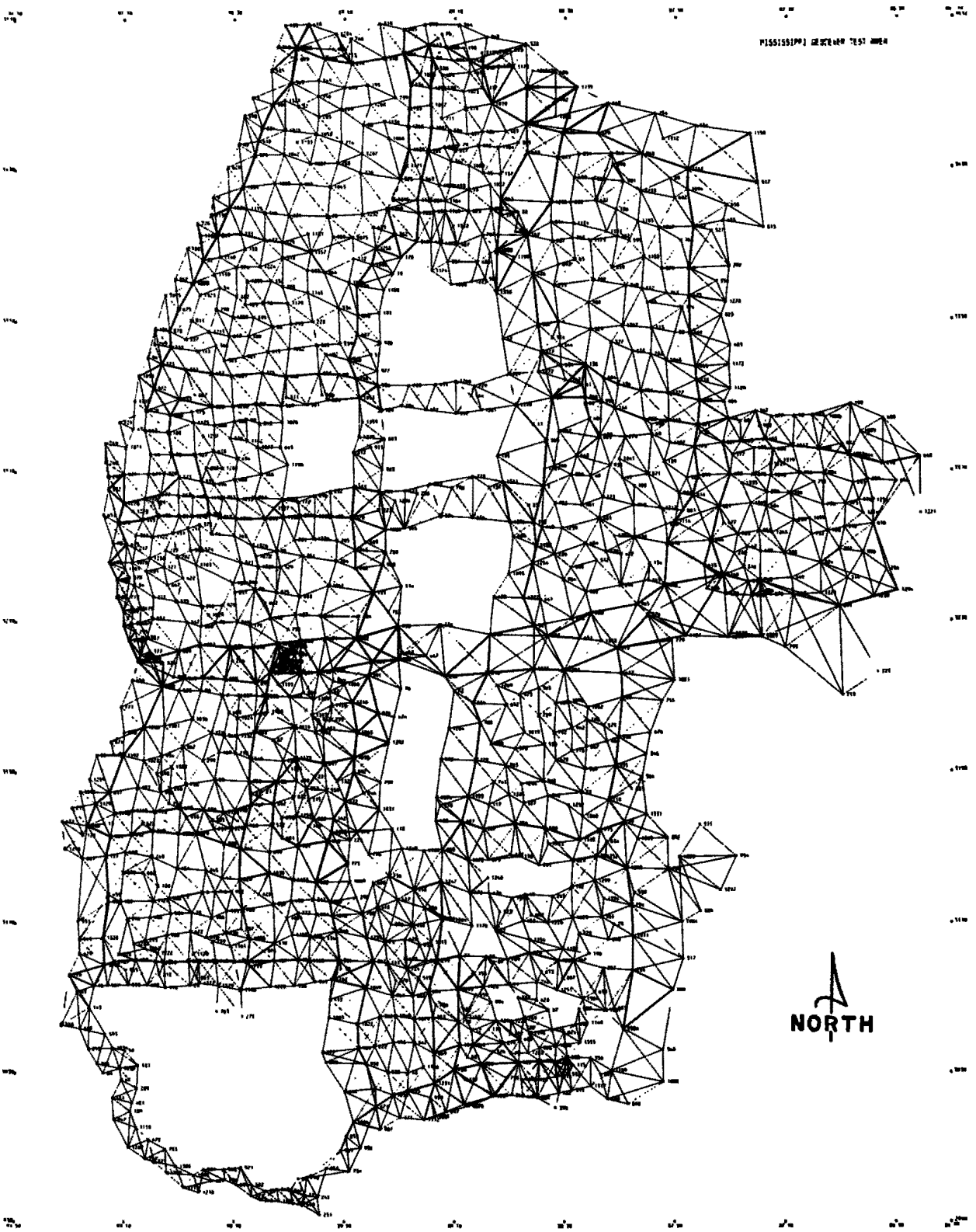

Figure 10.1.—Test area used by Moose and Henriksen for first- and second-order networks.

The magnitude of an eventual bias and of the standard deviation will be

$$|E\{\varepsilon_{ijk}^{(+)}\}| \leqslant \frac{1}{2} |\sum_{l=1}^{k} r_{li} r_{lj}| \beta^{-\tau} \tag{10.3}$$

$$\sigma\{\varepsilon_{ijk}^{(+)}\} \leqslant \frac{1}{\sqrt{12}} |\sum_{l=1}^{k} r_{li} r_{lj}| \beta^{-\tau}. \tag{10.4}$$

Actually we should have used the next integer powers of the base $\beta = 2$ to properly bound the numerators, but we will be satisfied with the above approximations.

Introducing the following sums

$$A_{ij} = \sum_{k=1}^{i-1} |\sum_{l=1}^{k} r_{li} r_{lj}| \tag{10.5}$$

$$S_{ij}^2 = \sum_{k=1}^{i-1} |\sum_{l=1}^{k} r_{li} r_{lj}|^2 \tag{10.6}$$

it becomes clear that bias and standard deviation of the local roundoff errors ε_{ij} are approximated by the following expressions:

$$|E\{\varepsilon_{ij}\}| \leqslant \frac{1}{2} A_{ij} \beta^{-\tau} \tag{10.7}$$

$$\sigma\{\varepsilon_{ij}\} \leqslant \frac{1}{\sqrt{12}} S_{ij} \beta^{-\tau}. \tag{10.8}$$

In order to arrive at global estimates for $E\{\zeta_i\}$, $\sigma\{\zeta_i\}$ we take (4.35) and (4.36a) restricted to the left-side error contributions. Thus we take

$$E\{\zeta_i\} = -\sum_{j=1}^{n} \sum_{k=1}^{n} f_{ij} x_k E\{\varepsilon_{jk}\} \tag{10.9}$$

$$\sigma\{\zeta_i\} = [\sum_{j=1}^{n} f_{ij}^2 x_j^2 \sigma^2\{\varepsilon_{jj}\} + \\ + \sum_{j=1}^{n} \sum_{k=j+1}^{n} (f_{ij} x_k + f_{ik} x_j)^2 \sigma^2\{\varepsilon_{jk}\}]^{\frac{1}{2}}. \tag{10.10}$$

A straightforward procedure would have been to evaluate these formulas using numbers f_{ij}, x_k as they were obtained from the adjustment. The solution vector x occurring in these formulas is known and causes no problem. However, to calculate all the elements f_{ij} of the inverse appeared to be too laborious, even for a small network. Hence a simplified procedure was used which was based on element-wise bounds $\|f\|$ and $\|x\|$ on F and x. By summing (10.5) and (10.6) we obtained

$$A = \sum_{i=1}^{n} \sum_{j=i}^{n} A_{ij} \tag{10.11}$$

$$S^2 = \sum_{i=1}^{n} \sum_{j=i}^{n} S_{ij}^2 \tag{10.12}$$

They give rise to the following estimates of bias and standard deviation of the global roundoff errors:

$$\frac{|E\{\zeta_i\}|}{\|x\|} \leqslant A \, \|f\| \, 2^{-48} \tag{10.13}$$

$$\frac{\sigma\{\zeta_i\}}{\|x\|} \leqslant \frac{2}{\sqrt{12}} S \, \|f\| \, 2^{-48}. \tag{10.14}$$

These formulas are obtained by bounding f_{ij}, x_k in (4.35) and (4.36a) in terms of $\|f\|$ and $\|x\|$ and carrying out the summations in a straightforward way, bearing in mind the definitions and relations of (10.3) to (10.4). We preferred to divide the formulas by $\|x\|$ in order to obtain errors relative to the largest coordinate shift. In the present example the largest shift will be about 5 m, not 10 m as in the case of the entire network.

Remark. The adjustments were done in measuring units corresponding to arc seconds of latitude and longitude. In this study we scaled all values to the meter. This caused a small complication because the scale factors for the two coordinate directions differ by a factor equal to the cosine of the latitude. We ignored this difference, and consequently scaling is strictly correct for only the latitude. Some of the numbers presented below will be slightly off. Referring to (10.13), (10.14) we will specify values for A, S which are too small by a factor between 1 and 0.5. Such a factor is not significant in roundoff estimates, and we will go along with these slightly wrong scale factors.

As a bound $\|f\|$ for the elements f_{ij} we took

$$\|f\| = (0.45\text{m})^2 \doteq 0.20\text{m}^2. \tag{10.15}$$

This estimate was obtained by looking at the diagonal elements f_{ij} of the inverse. These diagonal elements were quite uniform in size, suggesting, what was also plausible from other reasons, that the main contribution to the inverse F came from the uncertainty of the Doppler positions for which the rms errors of 0.75 m in each coordinate direction were assumed. The network behaved like a fairly stiff disk which was fixed by weak elastic forces to the five positions corresponding to the Doppler stations. The bound in (10.15) is not rigorous and is exceeded by some outliers. However, most elements are smaller.

10.1.3 Results of the Moose-Henriksen network experiments

10.1.3.1 Adjustment of the network as a whole

The following values for A, S were obtained:

$$A = 5.0\text{E}9 \text{ m}^{-2} \quad S = 5.5\text{E}7 \text{ m}^{-2}.$$

This leads to the following estimates:

$$\frac{|E\{\zeta_i\}|}{\|x\|} \leqslant 3.6\text{E}-6$$

. . . in the case of chopping arithmetic (10.16)

$$\frac{\sigma\{\zeta_i\}}{\|x\|} \leqslant 2.3\text{E}-8$$

. . . in the case of rounding arithmetic. (10.17)

The true relative roundoff errors obtained by comparison with the double length arithmetic results were the following:

6E−7 . . . in case of chopping arithmetic (10.18)

8E−9 . . . in case of rounding arithmetic. (10.19)

The agreement with the values predicted by (10.16) and (10.17) is considered satisfactory, even if we admit that the values in (10.16) and (10.17) came out a little bit too small as a result of our "cheating" on the scale factors. The σ-type estimates are too large by a factor of three. Without "cheating," this factor would have been four to five. Overestimation may be due partly to the overestimation of the inverse which has many elements smaller than those implied by (10.15). One should also note that longitude shifts were consistently smaller than latitude shifts. The bias estimates are too large by a factor of six (which should actually be nine to ten). Here one should note that (10.16) does not account for any offsetting of biases caused by alternating sign.

We proceed to specify the statistics obtained on the size of the coefficients of the normal equations. Table 10.1 shows how the diagonals of the original and the fully reduced normals decompose into size classes. By comparing these entries with table 9.1, the assumptions made in chapter 9 are conservative for this test network. The columns titled "Relative" in table 10.1 contain occupancies of the size classes and should be compared to the corresponding numbers ψ in table 9.1.

TABLE 10.1—*Size distribution of diagonal coefficients in the original and fully reduced normal equations*

Size - class From m^{-2}	To m^{-2}	Counts for original normals Absolute	Relative	Counts for reduced normals Absolute	Relative
4.5E6 -	1.8E4	161	0.060	70	0.026
1.8E4 -	1.1E3	2378	.890	2162	.809
1.1E3 -	0	133	.050	440	.165
Total		2672	1.000	2672	1.000

The last two columns in table 10.1 refer to the reduced normal equation system. The reduced system is $Rx = s$, where R is a triangular matrix corresponding to the Cholesky decomposition $A = R^tR$. For ease in comparison with the numbers of the original normals we multiplied all rows of the reduced system by its diagonals r_{ii}. This multiplication transforms r_{ii} into $a_{ii}^{(i-1)}$. We recognize from table 10.1 that the number of large coefficients does not increase. Rather, we notice a decrease in the number of large and medium-sized coefficients during the reduction. This completely confirms our reasoning in chapter 7.

Additional statistics on the size of the coefficients are shown in table 10.2. Here we find the maximum and average values for the diagonals a_{ii}, the two row sum norms $\|a_{i\cdot}\|_1$ and $\|a_{i\cdot}\|_2$, and the right-hand side coefficients of the original and the reduced normals. The values of the reduced normals result from multiplication by the diagonals, as pointed out above.

TABLE 10.2.—*Additional statistics on the size of elements in the original and reduced normals.*

Quantity	Original normals Maximum m^{-2}	Average m^{-2}	Reduced normals Maximum m^{-2}	Average m^{-2}
a_{ii}	1.5E6	2.3E4	1.4E6	1.2E4
$\|a_{i\cdot}\|_1$	3.1E6	4.2E4	2.8E6	2.8E4
$\|a_{i\cdot}\|_2$	2.1E6	1.4E5	2.0E6	1.1E5
$\|b_i\|$	3.5E5	1.0E3	3.5E6	5.0E2

Finally, we consider the number of product term evaluations. It was found that 1.13E7 times a product $r_{ki}r_{kj}$ had to be evaluated. Only in 4 percent of these cases was the result zero. This indicates that only a small portion of the computational effort was wasted. Multiplying 1.13E7 by 2, we obtain

$$\Gamma \cong 2.3\text{E}7. \qquad (10.20)$$

Returning to eq. (6.12) with the evaluation now simplified because there are no boundary stations, we find that by applying this equation a value of 4.6E7 is predicted. (We have used $i = 2672$ and a value of $w = 130$ which follows from eq. (6.5) with $\alpha = 0.8$.) We are satisfied to see the magnitude of the Γ-count predicted correctly by a formula which was based on very idealized assumptions.

10.1.3.2 Adjustment by Helmert blocking

It is pointed out that the decomposition of a subnetwork of less than 1,500 stations into 37 first-level Helmert blocks is not a very realistic procedure in view of the anticipated Helmert blocking design for the U.S. network. Subnetworks of this size will be decomposed, at most, into four blocks. The reason for partitioning the Moose-Henriksen network into so many blocks was unrelated to the roundoff experiments. It resulted from a desire to have a thorough check on the validity of the computer programs for higher level block design. The results obtained are nonetheless interesting. The values for A and S were:

$$A = 6.9\text{E}8\ \text{m}^{-2} \quad S = 1.9\text{E}7\ \text{m}^{-2}. \qquad (10.21)$$

This leads to the estimates:

$$\frac{|E\{\xi_i\}|}{\|x\|} \leqslant 8.6\text{E}{-7}$$

. . . in the case of chopping arithmetic (10.22)

$$\frac{\sigma\{\xi_i\}}{\|x\|} \leqslant 1.4\text{E}{-8}$$

. . . in the case of rounding arithmetic. (10.23)

The true relative roundoff errors were available for only the case of rounding arithmetic. They amounted to approximately

$$2.5\text{E}{-9}$$

. . . in the case of rounding arithmetic. (10.24)

The total count of product evaluations resulted in 2.5E7. It is rather surprising that only about half of these evaluations resulted in a nonzero product. The reason for this may be that too many Helmert blocks were used. It may also be that the programs for reducing the fill-in at a higher level were not yet fully developed. Anyway, we obtained 1.3E7 nonzero product terms $r_{ki}r_{kj}$, which is about the same number as before. The reason for smaller values of A and S and, consequently, for smaller roundoff errors, must be explained by the smaller coefficients $a_{ij}^{(p)}$ encountered during the reduction of the normals. The smaller coefficients must be explained by a better anchoring of the stations already eliminated by those not yet eliminated. Recall that stations which are eliminated last tend to be arranged along block boundaries that criss-cross the network. These frames of noneliminated stations give the network much strength throughout most of the reduction process.

This again illustrates the interesting fact that in roundoff studies not only is the strength of the entire network relevant, but also the strength of the subnetworks which comprise, at any stage of the triangularization procedure, the total number of stations eliminated as free ones and the total number of noneliminated stations as fixed ones.

10.1.4 Extrapolation of the test results

It is interesting to extrapolate the test results of (10.18) and (10.19) to the whole network. We adjust the estimates of (10.18) and (10.19) by taking into account two effects:

(1) The inverse of the whole network is smaller than the inverse of the test network. The whole network is positioned by 120 Doppler stations, the test network by only five. This will decrease the diagonals f_{ii} by a factor of one-half to one-third. The decrease of the off-diagonals will be much greater, and may amount to a factor of one-tenth. Hence we multiply the bias estimate (10.18) by 0.1. The σ estimate, which is based on a weighted square mean of the f_{ij}, does not react as strongly to the smaller off-diagonals. Hence we allow a factor of only one-fifth in the case of (10.19).

(2) The number of operations is much larger in the case of the entire network. The test network yielded a Γ count of 2.3E7. For the whole network $\Gamma = 1.2\text{E}11$ was the estimated value. This accounts for a factor of $1.2\text{E}11/2.3\text{E}7 \doteq 5{,}000$ for bias, and a factor of $\sqrt{5{,}000} \doteq 70$ for standard deviation.

The extrapolated figures then are given by:

$$\frac{|E\{\xi_i\}|}{\|x\|} \leqslant 3.0\text{E}{-4}$$

. . . in the case of chopping arithmetic (10.25)

$$\frac{|\sigma\{\zeta_i\}|}{\|x\|} \leq 1.0\text{E}-7$$

. . . in the case of rounding arithmetic. (10.26)

Eq. (10.26) suggests that about six or seven correct decimal digits are obtained for rounding on the CDC 6600. This surpasses our safe estimates, given in chapter 8, by about two decimal digits. The less conservative estimates of chapter 9 are surpassed by about one digit.

Because the structure of the entire network is not quite the same as that of the small network, the extrapolated values should be regarded with reservation.

10.2 Roundoff Experiments by Ebner and Mayer

Ebner and Mayer (1976) reported on extensive roundoff experiments that were done for photogrammetric block networks. Incidentally they also used a CDC 6600, and their calculations were also done with mantissas of 48 bits. However, they employed the standard instruction set which is chopping. A portion of their experiment was concerned with planimetric blocks. These blocks have a geometric strength similar to purely angular geodetic networks. Absolute positioning and scale of the photogrammetric blocks resulted from a number of fixed control points at the block perimeter. Both the number of control points as well as the size of the networks were varied to study the effect of these design parameters onto the global roundoff errors.

In Meissl (1972a) and (1976) the geometric strength of a planimetric block with dense perimeter control was shown to be qualitatively equivalent to that of a traditional geodetic network involving directions, distances, and azimuths. This means that the elements of the inverse F of such a network show a tendency to grow proportional to the logarithm of the number of stations. Because the logarithm grows very slowly, the elements f_{ij} of F can be viewed as being bounded for our present purpose.

For a block of 2,500 stations with dense perimeter control, the roundoff errors amounted to a loss of three decimal digits. The relative errors $|\zeta|/\|x\|$ were about 8E−11. This is astonishingly small. How can this figure be reconciled with the relative errors of about 8E−7, which we obtained for our network of 1,300 stations? The answer lies in the local structure. The photogrammetric blocks are very homogeneous. Therefore the diagonals of the original normals must be quite uniform. Our network has weight singularities that make some of the coefficients of the original normals larger than most of the other coefficients by a factor of about 1E4. This factor explains the difference nicely. In fact, it explains it almost too well because we should stress that there are also other differences that should be considered. The photogrammetric block has about twice as many stations. This may increase the number of computational steps by a factor of four. Ebner and Mayer also used Cholesky's method, but it was organized differently than the NGS programs. I do not expect that this had a large effect on roundoff. Finally, we must consider differences in local topology other than those already discussed. Balancing all these factors, we conclude that the Ebner-Mayer experiments can be viewed as a confirmation of our results, where we allow for an uncertainty of about one to two decimal places.

Remark. Ebner and Mayer also gave results for blocks without redundant perimeter control. Here relative errors of about 1E−7 were observed. The additional loss of about four decimal digits must be due mainly to a much larger inverse F associated with an unconstrained angular network. Referring once more to Meissl (1976), the elements of the inverse showed a tendency to grow in proportion to the number of stations. For a network of 2500 stations (this corresponds to a choice of $n = 50 = \sqrt{2500}$ in the referenced paper, where the number of stations is about equal to n^2), the difference between a constrained perimeter and an unconstrained perimeter can be about $n^2 = 2500$. (See eqs. (3.10) and (8.26) in Meissl (1976).) This explains strikingly a phenomenon encountered in the Ebner-Mayer experiments. At the same time, it confirms that we have identified correctly the main sources of the global roundoff errors and that our model for propagating roundoff errors from their origin to the final results is sound.

10.3 Roundoff Experiments by Ehlert

To test the validity of the network adjustment programs for different computers and to study the roundoff errors experimentally, a computer program generating nearly regular and homogeneous test networks was designed for the members of the International Association of Geodesy (IAG) Special Study Group 4:38, "Computer Techniques in Geodesy." A program list was circulated as an appendix to Circular Letter No. 2/1973. Together with my coworker K. Stubenvoll, I used it for some tests on a UNIVAC 494. The roundoff errors of the station coordinates were so strikingly correlated that I was prompted to do further theoretical research. Together with my coworker N. Bartelme, a thorough study was undertaken on roundoff error propagation in homogeneous leveling networks. Supporting test calculations are reviewed in the next section.

Although a few members of SSG 4:38 reported test

results for small networks (See Meissl (1975a)), no one made a systematic comparison of roundoff errors for larger networks. However, a calculation done by D. Ehlert offered a check by itself. Ehlert adjusted a network of 1,141 stations by Helmert blocking on a TR 440 computer. Because the adjustment was done in a plane projection and the original normals were formed for the unconstrained, freely floating network, the reduction of the normals offered a check on their rank deficiency. The check indicated that about nine decimal digits of the triangularized normals were correct.

The calculations were done in single precision. The TR 440 has a base of $\beta = 16$ and a mantissa length of 38 bits. This is unusual because it is not divisible by 4. Its accuracy corresponds to 10 or 11 decimal digits. The TR 440 is a "beautiful" machine because it truly rounds the result of an arithmetic operation.

It would be a fallacy to equate the perturbations observed in the singularity check to the local roundoff errors, ε_{ij}. The ε_{ij}'s are errors traced backward to the original system. The singularity check does not refer to the original system but rather to the triangularized system. One way to analyze the errors of the triangularized system is to propagate the backward traced errors ε_{ij} forward again by the following steps:

The perturbed original system is (4.30), *i.e.*:

$$(A + \varepsilon)(x + \zeta) = b + \eta$$

This system must be triangularized without further error. If R^TR is the Cholesky decomposition of A, and $(R+\delta)^T(R+\delta)$ is that of $A+\varepsilon$, it follows that in linear approximation

$$R^T\delta + \delta^TR = \varepsilon. \qquad (10.27)$$

These relations can be used to calculate δ recursively.

In the present case, a shortcut is available. Ehlert's block has about as many stations as the test network treated in section 10.1.3.2. Hence the Γ counts should be comparable. Assuming $\Gamma \cong 1E7$, and taking into account the homogeneity of the network, as well as its strength, which should not make $\|a\| \bullet \|f\|$ much larger than one, we conclude that a pessimistic upper bound for the global roundoff errors suffered by the coordinate shifts is given by a loss of $log_{10}\sqrt{\Gamma}$ decimal digits. This amounts to three or four decimal digits. Because of the tapering effect of the coefficients $a_{ij}^{(\rho)}$ we can expect a better result, a loss of about two digits. Recall that Ebner-Mayer adjusted a homogeneous network which had twice as many stations. They used a chopping arithmetic and lost only three decimal digits. The TR 440 is a rounding machine and the present network is smaller. Hence a loss of two digits appears to be a reasonable estimate.

The coordinate shifts calculated during the early stages of back substitution should have as many correct digits as the number of coefficients at the bottom of the triangularized system. These shifts are obtained from the coefficients at the bottom by a few arithmetic operations. We conclude that the coefficients at the bottom of the triangularized system have only the last two decimal digits perturbed. This is in fair agreement with the perturbations resulting from the singularity check.

10.4 Roundoff Experiments with Idealized Leveling Networks by Bartelme-Meissl

Bartelme and Meissl (1975, 1977) reported on the theoretical investigations of roundoff errors during the direct solution of normal equations of large homogeneous leveling networks. Supporting test calculations have been carried out. True rounding was simulated by means of specially written subroutines. The simulated computer had a base of $\beta = 2$ and a mantissa length of $\tau = 16$.

A square shaped leveling net of 15*15 = 225 stations was adjusted. Gauss-Jordan elimination was used to solve the normals. The theoretical formulas predicted a loss of eight binary digits. The actually observed roundoff errors came very close to this number.

The correlation pattern of the roundoff errors is beautifully illustrated by figure 10.2. What looks like a peak is actually the fixed station at the center. All other roundoff errors were negative. It was a picture such as this one that started the theoretical research into roundoff errors in geodetic network computation.

10.5 Roundoff Experiments Related to the Kentucky-Tennessee Test Area.

About a year after this report was completed and while the manuscript was being reviewed, W. H. Dillinger provided me with additional test results from a network covering most of Kentucky and Tennessee. This network, comprised of 3,380 stations, is bounded by the 35° and 38° parallels and by the 83° and 88° meridians. It was divided into four first-level Helmert blocks to be combined subsequently into one second-level block.

The network was processed primarily as a test of NGS programs and procedures for the new adjustment of the North American Datum. A detailed account of this pilot test can be found in Timmerman (1978). The roundoff experiments represented only a marginal feature. Their design is similar to that described in section 10.1 for the Moose-Henriksen

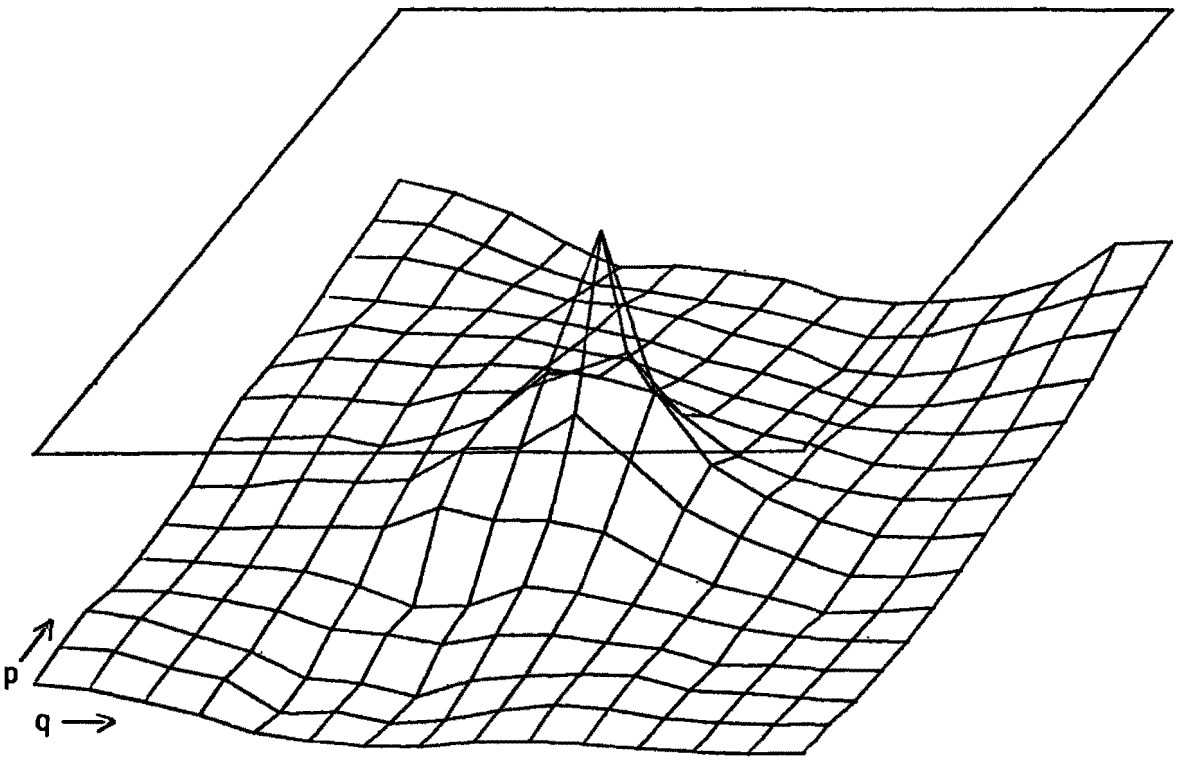

Figure 10.2.—Distortion of a regular leveling network caused by roundoff errors.

network. However, there are a few shortcomings in the present case, namely:

(1) Because of the larger size of the network, it was considered too expensive to calculate a double-precision solution. Hence only two single-precision solutions obtained on the CDC 6600 could be compared. They refer to the cases of chopping and rounding arithmetic.

(2) No information was available on the inverse because the newly developed routines for Helmert blocking were still incomplete.

(3) No Doppler observations were used. The absolute position of the network was arbitrarily fixed by constraining one station to its approximate position. In addition, about 120 stations along the perimeter were not rigidly connected to the network. The positions of these stations were also constrained to their approximate positions.

A comparison of the chopped and rounded solutions showed agreement to about six leading decimal digits for the larger coordinate shifts. Since the chopped solution is inferior to the rounded solution, we derive essentially the error of the former one. The relative error of the chopped solution, *i.e.*, largest absolute error divided by the largest coordinate shift, came out to about 1E−6.

Another interesting feature was the great uniformity of the observed absolute errors, which is in full agreement with the strong correlation of the global roundoff errors previously inferred theoretically. Recall that we have repeatedly pointed out that the accuracy of the relative position between stations in a vicinity will be superior to the global accuracy. (See section 9.7.)

The remaining features of the Kentucky-Tennessee roundoff experiment are the same as those described in section 10.1.2 for the Moose-Henriksen network. There is no need to duplicate the explanations of the a posteriori analysis, which is based on statistical information on number and size of coefficients and partial sums arising during Cholesky decomposition. Using the same symbolism as in section 10.1, we merely state the results.

The numbers A and S (see eqs. 10.11 and 10.12) resulted in

$$A = 3.7\text{E}10 \tag{10.28}$$

$$S = 1.7\text{E}8 \tag{10.29}$$

Making certain assumptions on the bound $\|f\|$ on the elements of the inverse, which are described below, we obtain:

$$\frac{|E\{\xi_i\}|}{\|x\|} \leqslant 6.6\text{E}{-6} \quad \text{... in the case of chopping} \tag{10.30}$$

$$\frac{\sigma\{\xi_i\}}{\|x\|} \leqslant 1.7\text{E}{-8} \quad \text{... in the case of rounding} \tag{10.31}$$

These results are analogous to those of eqs. (10.16) and (10.17). To derive them according to (10.13–14) we need $\|f\|$. Since no information was available on the inverse, we were forced to guess. We have assumed $\|f\| = 0.05$ m $= (0.22\text{ m})^2$ which is smaller by a factor of one-fourth in comparison to the Moose-Henriksen net. As a guideline for our guess we used the fact that constraining at least one station tends to make the errors smaller than those of a network whose absolute position is derived from five Doppler observations with rms errors of about 1 m.

The number in eq. (10.30) must be compared with the experimentally observed error of the chopped solution which was about 1E−6. The agreement is good; however, we must bear in mind the uncertainties in estimating $\|f\|$.

Let us also extrapolate the errors for the entire network in the same way as described in section 10.1.4. The Kentucky-Tennessee network gave a Γ count of 3.5E8. Using the Γ count of 1.2E11, predicted in table 6.1, for the entire network, we arrive at extrapolated errors of 2E−3 for the chopped solution and 3E−7 for the rounded solution. We did not apply a correction as in the Moose-Henriksen network because a change occurred to the inverse. The extrapolations compare reasonably well to those obtained from the Moose-Henriksen network. Again the results show that even a chopping CDC 6600 would do the job satisfactorily, although our more conservative theoretical predictions do not suggest that such a procedure is safe.

Now let us also try to extrapolate the Γ count to the entire network. If we reduce only the four first-level blocks of our test network, the corresponding Γ count amounts to 1.7E8. Upgrading this by a factor of 60, in order to account for about 200,000 stations instead of 3,400, we obtain 1E10, a number whose order of magnitude should agree with the low-level count for the entire network obtained by the procedures described in chapter 6. The top right corner of table 6.2 gives a value of 2.35E10. The agreement is considered satisfactory.

It is also interesting to compare the statistics on the size of the coefficients of the original and the reduced normals with those obtained for the Moose-Henriksen network. Tables 10.3 and 10.4 correspond to tables 10.1 and 10.2 in section 10.1.3. The present tables have an extra column because the normals were not reduced in one sweep. Therefore, the statistics of the reduced first level blocks are listed separately for interior and junction stations. We refrain from exhibiting the numbers for the second-level block reduction. As pointed out earlier, there were some loosely connected stations which the Cholesky routine took care of by automatically fixing them to their approximate position. This was done by placing a large number at the appropriate diagonal positions. Consequently our statistics on the size of the coefficients were falsified.

TABLE 10.3.—*Size distribution of diagonal coefficients in the original and reduced normals after treating the four first-level blocks.*

Size - class		Original normals		Reduced normals Interior equations		Reduced normals Junction equations	
From	To	Absolute	Relative	Absolute	Relative	Absolute	Relative
4.6E6	1.8E4	2314	0.32	1044	0.19	75	0.04
1.8E4	1.1E3	3013	.42	2858	.52	606	.36
1.1E3	0	1905	.26	1632	.29	1017	.60
Total		7232	1.00	5534	1.00	1698	1.00

TABLE 10.4.—*Additional statistics on the size of the elements in the original and reduced normals after treating the four first-level blocks.*

Quantity	Original normals		Reduced normals Interior equations		Reduced normals Junction equations	
	Maximum	Average	Maximum	Average	Maximum	Average
a_{ii}	2.9E6	1.2E5	2.4E6	7.0E4	1.8E6	1.7E4
$\|a_{i\cdot}\|_1$	7.0E6	2.1E5	6.3E6	1.6E5	4.0E6	2.9E4
$\|a_{i\cdot}\|_2$	3.5E6	3.5E5	3.1E6	3.1E5	2.2E6	1.3E5
$\|b_i\|$	1.5E6	2.7E3	1.5E6	1.7E3	9.0E3	1.4E2

Some concern is caused in table 10.3 by the number 0.32 in the first row of the column titled "original normals, relative" count. This number gives the ratio of large diagonal elements to the total number of diagonal elements in the original normals. In the most conservative estimates of chapter 8, a bound of 0.25 was assumed for this ratio. On the other hand, it is comforting to see this ratio drop sharply during reduction. For the partially reduced normals of the junction stations, this amounts to only 0.04. Such a drop was not considered in chapter 8. Hence there is no need yet to revise the estimates made there.

11. MISCELLANEOUS COMPLEMENTS

11.1 On the Choice of Norm of the Predicted Roundoff Errors in Geodetic Normal Equations

If our nonzero bounds on $|E\{\zeta_i\}|$ are doubled, they constitute rigorous bounds on the global roundoff errors. This is true since for any elementary roundoff error $\varepsilon_{ij}^{(e)}$ we have always used a bound on $|E\{\varepsilon_{ij}^{(e)}\}|$ which was not smaller than one-half the maximum size of $|\varepsilon_{ij}^{(e)}|$. In the literature on roundoff errors one finds a preference for roundoff estimates that have rigorous upper bounds (Wilkinson 1963, Bunch 1974, and Gear 1975). In this section we will try to explain why formulas found in the literature could not be immediately applied to the problem of predicting the roundoff errors in the U.S. network adjustment.

The usual procedure for analyzing the effect of the left-side triangulation roundoff errors is to specify elementwise bounds on the perturbation matrix ε which fulfills:

$$(A+\varepsilon)(x+\zeta) = b. \tag{11.1}$$

The elementwise bounds usually rely on the maximum modulus of any partially reduced coefficient $a_{ij}^{(p)}$. They also rely on the number of elementary operation steps. In this respect, the procedure in sections 8.1 and 8.2 for obtaining bias estimates did not deviate very much. The main difference there was that we separated the effect of a small number of large coefficients $a_{ij}^{(p)}$ from the effect of a large number of small coefficients. In chapter 9, we also made use of the fact that the row sums of $a_{ij}^{(p)}$ are actually much smaller than a_{ii} times the number of coefficients per row.

There is, however, another reason why the formulas which I have found in the literature would grossly overestimate the errors suffered by the adjusted coordinate shifts. This reason is more subtle and we will try to explain it without going into geodetic network theory in depth.

Solving (11.1) to the first degree of accuracy, we obtain

$$\zeta = -A^{-1}\varepsilon\, x. \tag{11.2}$$

Let $\|v\|_2$ denote the Euclidean norm of a vector v and let $\|M\|_2$ be the spectral norm of a matrix M. Then we have

$$\|Mv\|_2 \leqslant \|M\|_2\|v\|_2. \tag{11.3}$$

If v is the eigenvector corresponding to the largest eigenvalue of M^TM, then

$$\|Mv\|_2 = \|M\|_2\|v\|_2.$$

The square roots of the eigenvalues of M^TM are also called the singular values of M. If M is positive definite, the singular values are identical to the eigenvalues. From (11.2) and (11.3) it follows that

$$\|\zeta\|_2 \leqslant \|A^{-1}\|_2\|\varepsilon\|_2\|x\|_2. \tag{11.4}$$

From the unperturbed normal equations $Ax = b$, we infer

$$\|b\|_2 \leqslant \|A\|_2\|x\|_2. \tag{11.5}$$

Combining (11.4) and (11.5) and introducing the condition number $\mathrm{cond}_2(A)$ by

$$cond_2(A) = \|A\|_2\|A^{-1}\|_2 \tag{11.6}$$

we obtain

$$\frac{\|\zeta\|_2}{\|x\|_2} \leqslant cond_2(A)\,\frac{\|\varepsilon\|_2\|x\|_2}{\|b\|_2}. \tag{11.7}$$

This is the type of estimate we obtain if we use the roundoff literature as a guideline. Frequently, more precise formulas are preferred that do not rely on first-order approximation.

One of the remarkable features of geodetic networks is that the elementwise bound $\|f\|$ on $F = A^{-1}$ turns out to be much smaller than the spectral norm bound $\|F\|_2 = \|A^{-1}\|_2$. For large and homogeneous geodetic networks it has been shown that $\|f\|/\|F\|_2$ decreases asymptotically like a constant times $log(n)/n$. Here n is the number of stations. See Meissl (1974) for a review of papers dealing with network theory. Actually, the stated result refers to networks with no absolute position measurements. In the presence of Doppler stations it is no longer valid theoretically. However, it is still a reasonable guideline, as long as the accuracy of the Doppler stations is low compared with the accuracy of the relative observations.

Intuitively, the reason for the discrepancy between $\|f\|$ and $\|F\|_2$ is the following: Representing (11.2) as

$$\zeta = -A^{-1}(\varepsilon x) \tag{11.8}$$

one infers that $\|\zeta\|_2$ will come close to

$$\|\zeta\|_2 = \|A^{-1}\|_2 \|\varepsilon x\|_2. \qquad (11.9)$$

if ζ and εx are expressible in terms of eigenvectors of A, which belong to small eigenvalues of A. (Small eigenvalues of A correspond to large eigenvalues of A^{-1}.) Such eigenvectors are associated with *smooth* distortions of the whole network. The network corresponds to an elastic system with elastic elements between neighboring stations. A smooth distortion of a certain amplitude will result in a small amount of associated elastic energy. A high frequent distortion of comparable amplitude will give a much larger energy value. For a smooth network distortion ζ, the ratio $\|\zeta\|_2/\|\zeta\|$ will be large, *i.e.*, the spectral norm of ζ will be much larger than the magnitude of the largest component of ζ. This is one of the main reasons why eq. (11.7) will never give good estimates if one is interested in positional errors, *i.e.*, in errors ζ of the components x relative to the magnitude of the largest component.

To further illustrate our line of reasoning, suppose that ζ corresponds to a smooth distortion of the network with an amplitude of 0.1 mm. Because ζ has about 350,000 components, the magnitude of the Euclidean norm $\|\zeta\|_2$ can be the same as that of $0.1*\sqrt{350{,}000} = 59$ mm. Equation (11.7) must take care of such situations. Hence it overshoots the error norm $\|\zeta\| = Max\ (|\zeta_i|)_i$ by a factor of 600. The estimates of chapters 8 and 9 are based on $\|\zeta\|$ and not on $\|\zeta\|_2$. Hence more favorable results could be obtained.

11.2 Asymptotic Roundoff Error Estimates

In this section we look beyond the particular case of the U.S. network. We are concerned with networks whose number of stations grows beyond all limits. For simplicity, we exclude a number of subcases that require separate and elaborate arguments. We assume that our networks are homogeneous with respect to the distribution of observational weights. We also assume that the reference surface is a plane and that the networks extend in both coordinate directions. Thus we exclude strip-like networks. We further exclude pathologically shaped boundaries, and deal with square-shaped networks only. Nevertheless our results will be representative for a fairly general class of networks.

We assume that the normals are solved by Cholesky's method and that the solution process is organized according to the nested dissection scheme.

The estimates will be qualitative in the sense that they will contain unspecified constants. The presentation in this section will not be entirely self-contained. We shall use some results available from the theory of geodetic networks.

11.2.1 Nested dissection of homogeneous and regular networks

Figure 11.1 is a repeat of figure 3.6 except for a slight modification. The modification involves a reverse numbering of the block levels. We denote the block levels by $\ell = 0,1,2, \ldots$, starting from the highest level and proceeding downward. It is important to note that the bars of the cross-like sets which form the impenetrable barriers are rather narrow. In chapter 6 we assumed double rows of stations along these barriers.

Let n denote the number of stations in the entire network. Then the width of the network comprises $\sqrt{n}$ stations. At level ℓ, the network decomposes into 4^ℓ blocks. A barrier set crossing one of these blocks is comprised of stations whose number must be bounded by $\sqrt{n}/(2^\ell)$ multiplied by a constant. It follows that the total number of stations inside the block must be bounded by a constant times $\sqrt{n}/(2^\ell)$. We must be aware that the equations belonging to stations inside a block split into interior equations and junction equations. The junction equations participating in the partial reduction of a block involve not only stations inside the block but also stations of neighboring blocks. We introduce

$$m_\ell = O(\sqrt{n}/2^\ell) \qquad (11.10)$$

as a common bound of the interior as well as the junction equations in a block of level ℓ. It is now easy to derive total Π and Γ counts for the entire network. Although the normals of the block may be structured as shown in figure 6.11, we assume nonzero coefficients over the entire block, as shown in figure 11.2. Applying the methods given in chapter 6 for the problem of counting nonzero coefficients and elementary computational steps, we see that the contribution of a row of interior equations labelled x to the total Π count is the same as x itself. The contribution to the Γ count is $2x*(2m_\ell - x)$. For a row x of junction equations the corresponding numbers are m_ℓ and $2m_\ell*x$. Hence the Π and Γ counts for the block are given by

$$\Pi = \int_0^{m_\ell} x\,dx + \int_0^{m_\ell} m_\ell\,dx = O(m_\ell^2) \qquad (11.11)$$

$$\Gamma = \int_0^{m_\ell} 2x(2m_\ell - x)dx + \int_0^{m_\ell} 2m_\ell\,x\,dx = O(m_\ell^3). \qquad (11.12)$$

Combining this with (11.10) we get

$$\Pi = O(n/4^\ell), \quad \Gamma = O(n\sqrt{n}/8^\ell)$$

$$\ldots \text{for one level } \ell \text{ block.} \qquad (11.13)$$

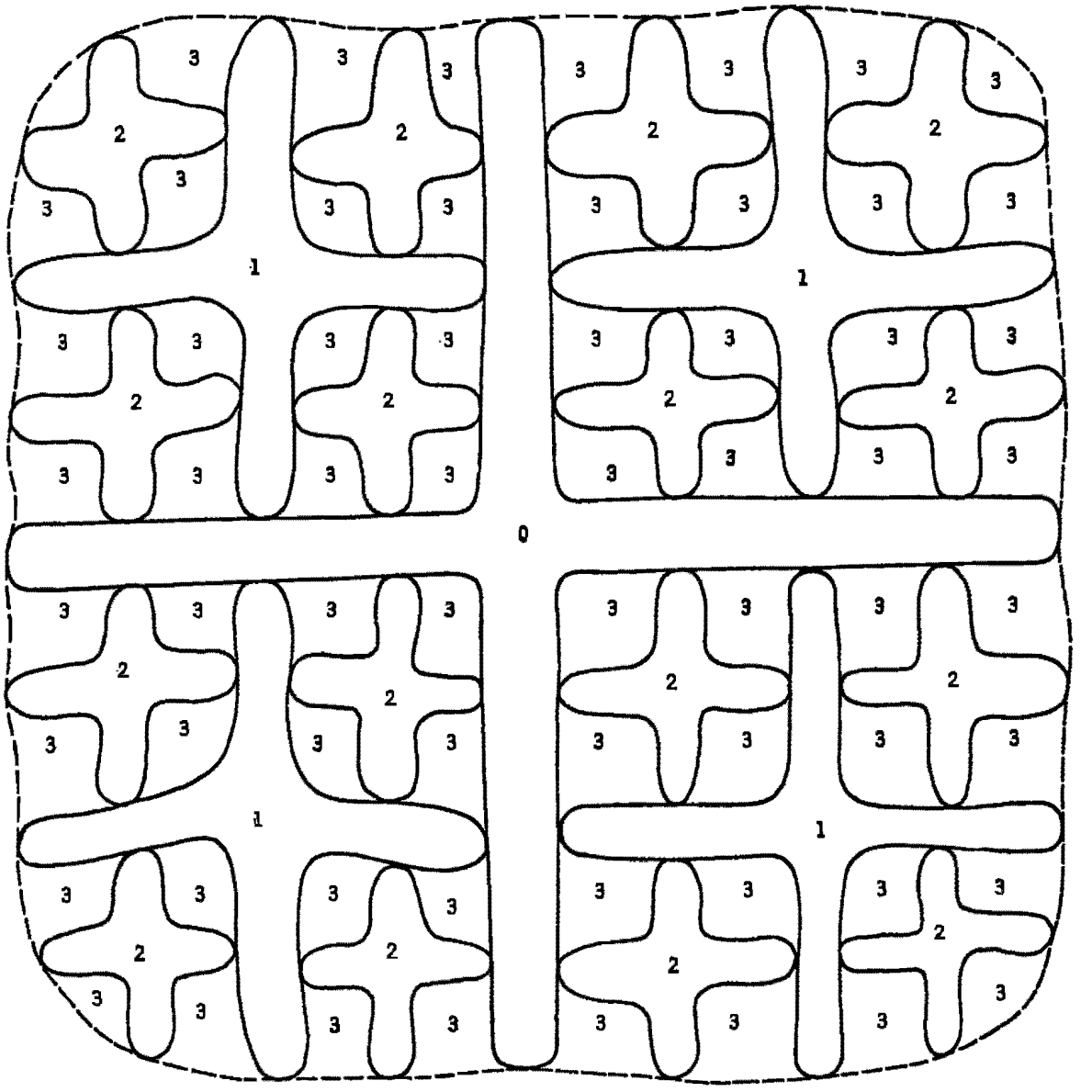

Figure 11.1.—Nested dissection of a geodetic network.

Multiplying these numbers by the number of blocks at level l, which is 4^l, we obtain the Π and Γ counts for level l

$$\Pi = O(n), \Gamma = O(n\sqrt{n}/2^l)$$

. . . for all level l blocks (11.14)

Now we sum over the levels. Clearly the number of levels is bounded by a constant plus $log_2\sqrt{n}$. This is what nested dissection is all about. We note that $log_2\sqrt{n} = O(log\ n)$. Hence the total Π and Γ counts are obtained as

$$\Pi = \sum_{l=0}^{O(log\ n)} O(n) = O(n\ log\ n)$$

$$\Gamma = \sum_{l=0}^{O(log\ n)} O(n\sqrt{n}/2^l) = O(n\sqrt{n})$$

. . . for the whole network. (11.15)

These results are due to George (1973). It is pointed out that our symbol n corresponds to n^2 in George's notation.

Recall from section 3.5.4 that the Π and Γ counts for minimum bandwidth ordering are given by

$$\Pi = O(n\sqrt{n}), \quad \Gamma = O(n^2). \qquad (11.16)$$

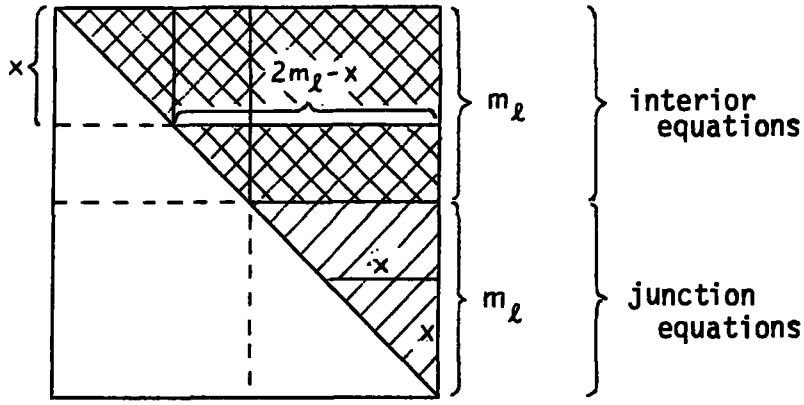

Figure 11.2.—Simplified structure of normal equations for a one level-ℓ block.

11.2.2 A general theorem

In this subsection we will bound the bias and standard deviation of the global roundoff error suffered by a coordinate shift in terms of the following quantities:

- n . . . number of stations
- $\|a\|$ . . . bound on the largest element of A
- $\|f\|$ . . . bound on the largest element of F, the inverse of A
- $\|x\|$ . . . bound on the largest element of the solution vector x
- $\|b^{(\cdot)}\|$ . . . bound on the largest right hand side $b_i^{(p)}$, encountered in the original normals or in any of the partially reduced normals $A_{22}^{(p)}x = b_2^{(p)}$.

Remark: A quantity like $\|b^{(\cdot)}\|$ should actually not appear in an a priori roundoff estimate. For any estimate involving $\|b^{(\cdot)}\|$ we will also specify one that is free of this quantity. The reason for including $\|b^{(\cdot)}\|$ in some estimates is that we believe this quantity to be extremely well behaved in practical applications. (See the discussion in section 7.2.2.)

The following theorem is only based on the assumption that A is a positive definite matrix referring to a sparse system which may be subjected to nested dissection as outlined in the previous section. Such systems are not only typical for geodetic networks, but also for a wide range of structural analysis problems treated by the finite element method.

Theorem 11.1. Suppose that the normals of a large and homogeneous geodetic network are solved by Cholesky's method and that the solution process is organized according to the nested dissection scheme. Assume that the calculations are done in floating point arithmetic, that β is the base of the number system, and that the mantissas have a fixed length of τ digits. Rounding may be biased or unbiased. In the latter case the bias estimates specified below are irrelevant. We split the global roundoff errors ζ_i suffered by coordinate x_i as

$$\zeta_i = \zeta_i^{(lt)} + \zeta_i^{(rt)} + \zeta_i^{(bs)} \tag{11.17}$$

Here the superscripts lt, rt and bs refer to components of ζ_i that are propagated from different types of local roundoff errors, namely: left-side errors during the triangular decomposition, right-side errors during triangular decomposition, and back-substitution errors.

The following estimates are true in linear approximation:

$$|E\{\zeta_i^{(lt)}\}| \leqslant \|a\|\ \|f\|\ \|x\|\ O(n^{\frac{5}{4}})\ \beta^{-\tau+1} \tag{11.18}$$

$$|E\{\zeta_i^{(rt)}\}| \leqslant \|f\|\ \|b^{(\cdot)}\|\ O(n \log n)\ \beta^{-\tau+1} \tag{11.19}$$

$$-''- \leqslant \|a\|\ \|f\|\ \|x\|\ O(n^{\frac{5}{4}})\ \beta^{-\tau+1} \tag{11.20}$$

$$|E\{\zeta_i^{(bs)}\}| \leqslant \sqrt{\|a\|\ \|f\|}\ \|x\|\ O(n)\ \beta^{-\tau+1} \tag{11.21}$$

$$\sigma\{\zeta_i^{(lt)}\} \leqslant \|a\|\ \|f\|\ \|x\|\ O(\sqrt{n \log n})\ \beta^{-\tau+1} \tag{11.22}$$

$$\sigma\{\zeta_i^{(rt)}\} \leqslant \|f\|\ \|b^{(\cdot)}\|\ O(\sqrt{n \log n})\ \beta^{-\tau+1} \tag{11.23}$$

$$-''- \leqslant \|a\|\ \|f\|\ \|x\|\ O(n^{\frac{3}{4}})\ \beta^{-\tau+1} \tag{11.24}$$

$$\sigma\{\zeta_i^{(bs)}\} \leqslant \sqrt{\|a\|\ \|f\|}\ \|x\|\ O(\sqrt{n \log n})\ \beta^{-\tau+1}. \tag{11.25}$$

We note that the constants hidden in the O symbols depend on only local structure, *i.e.*, on the pattern of observations and their weights.

Proof: Consider partial reduction of a level ℓ block. Focus attention on a local lt type error $\varepsilon_{ijk}^{(pr)} = \varepsilon_{ijk}^{(s)} + \varepsilon_{ijk}^{(t)}$ occurring during the evaluation and summation of a product term $r_{ki}r_{kj}$. Recalling (7.1), we have the representation $r_{ki}r_{kj} = a_{ij}^{(k)} - a_{ij}^{(k-1)}$. Bias and standard deviation of $\varepsilon_{ijk}^{(pr)}$ may be bounded as

$$\begin{aligned} |E\{\varepsilon_{ijk}^{(pr)}\}| &= O(\|a_{ij}^{(k)}\| + \|a_{ij}^{(k-1)}\|)\,\beta^{-\tau+1} \\ \sigma\{\varepsilon_{ijk}^{(pr)}\} &= O(\|a_{ij}^{(k)}\| + \|a_{ij}^{(k-1)}\|)\,\beta^{-\tau+1} \end{aligned} \tag{11.26}$$

(See section 4.1.1.) Summing over all locations (i,j), $j \geqslant i$, and using the definition of the row sum norms $\|a_{i\cdot}^{(k)}\|_1$ and $\|a_{i\cdot}^{(k)}\|_2$, found in (7.4) and (7.5), we get

$$\begin{aligned} \sum_{j\geqslant i} |E\{\varepsilon_{ijk}^{(pr)}\}| &= O(\|a_{i\cdot}^{(k)}\|_1 + \|a_{i\cdot}^{(k-1)}\|_1)\,\beta^{-\tau+1} \\ \sum_{j\geqslant i} \sigma^2\{\varepsilon_{ijk}^{(pr)}\} &= O(\|a_{i\cdot}^{(k)}\|_2^2 + \|a_{i\cdot}^{(k-1)}\|_2^2)\,\beta^{-2\tau+2}. \end{aligned} \tag{11.27}$$

Now we use the bound on the row sum norms provided by proposition (7.1). In these bounds we can replace λ_{MAX} by $O(\|a\|)$, as argued in section 7.1.3. We obtain

$$\|a_{i\cdot}^{(k)}\|_1 = \|a\|O(\sqrt{x_i}),\ \|a_{i\cdot}^{(k)}\|_2 = \|a\|O(1).$$

Here x_i is the number of nonzero coefficients $a_{ij}^{(i-1)}$ to the right and including the diagonal position in equation i. It follows that

$$\begin{aligned} \sum_{j\geqslant i} |E\{\varepsilon_{ijk}^{(pr)}\}| &= \|a\|\, O(\sqrt{x_i})\,\beta^{-\tau+1} \\ \sum_{j\geqslant i} \sigma^2\{\varepsilon_{ijk}^{(pr)}\} &= \|a\|^2\, O(1)\,\beta^{-2\tau+2}. \end{aligned} \tag{11.29}$$

Our next step is to sum these relations over k. Only roundoff errors associated with nonzero product terms contribute nontrivially to the sum. We recall that μ_{ij} is the number of nonzero product terms which are subtracted from a_{ij} during triangular decomposition, and that $\mu_{ij} \leqslant \mu_{ii}$. It follows that the row sums of the local roundoff errors ε_{ij}, which are the superposition of the $\varepsilon_{ijk}^{(pr)}$, are bounded by

$$\begin{aligned} \sum_{j\geqslant i} |E\{\varepsilon_{ij}\}| &\leqslant \|a\|\, \mu_{ii}\, O(\sqrt{x_i})\,\beta^{-\tau+1} \\ \sum_{j\geqslant i} \sigma^2\{\varepsilon_{ij}\} &\leqslant \|a\|^2\, \mu_{ii}\, O(1)\,\beta^{-2\tau+2}. \end{aligned} \tag{11.30}$$

We now replace the summation over rows i by integration. Dealing with one ℓ level block, we label an interior row by x, as shown in figure 11.2. We then can take $\mu_{xx} = x$ and $x_x = 2m_\ell - x$. Now we integrate the interior rows from $x=0$ to $x=m_\ell$. Proceeding in a similar manner for the junction equations, where, in agreement with figure 11.2, we take $\mu_{xx} = m_\ell$ and $x_x = x$, we obtain

$$\begin{aligned} \sum_i \sum_{j\geqslant i} |E\{\varepsilon_{ij}\}| &\leqslant \|a\| \{\int_0^{m_\ell} x\sqrt{2m_\ell - x}\,dx + \\ &\quad + \int_0^{m_\ell} m_\ell\sqrt{x}\,dx\}\, O(1)\,\beta^{-\tau+1} \\ &= \|a\|\, O(m_\ell^{\frac{5}{2}})\,\beta^{-\tau+1} \\ \sum_i \sum_{j\geqslant i} \sigma^2\{\varepsilon_{ij}\} &\leqslant \|a\|^2 \{\int_0^{m_\ell} x\,dx + \\ &\quad + \int_0^{m_\ell} m_\ell\,dx\}\, O(1)\,\beta^{-2\tau+2} \\ &= \|a\|^2\, O(m_\ell^2)\,\beta^{-2\tau+2}. \end{aligned} \tag{11.31}$$

Inserting from eq. (11.10) we get

$$\begin{aligned} \sum_i \sum_{j\geqslant i} |E\{\varepsilon_{ij}\}| &\leqslant \|a\|\, O(n^{\frac{5}{4}}/2^{\frac{5\ell}{2}})\,\beta^{-\tau+1} \\ \sum_i \sum_{j\geqslant i} \sigma^2\{\varepsilon_{ij}\} &\leqslant \|a\|^2\, O(n/4^\ell)\,\beta^{-2\tau+2} \end{aligned}$$

$$\ldots \text{ for one block of level } \ell. \tag{11.32}$$

Summing over all 4^ℓ blocks at level ℓ, we obtain

$$\begin{aligned} \sum_i \sum_{j\geqslant i} |E\{\varepsilon_{ij}\}| &\leqslant \|a\|\, O(n^{\frac{5}{4}}/2^{\frac{\ell}{2}})\,\beta^{-\tau+1} \\ \sum_i \sum_{j\geqslant i} \sigma^2\{\varepsilon_{ij}\} &\leqslant \|a\|^2\, O(n)\,\beta^{-2\tau+2} \end{aligned}$$

$$\ldots \text{ for all blocks of level } \ell. \tag{11.33}$$

Finally, summing over all levels whose number is $O(\log n)$, we find

$$\begin{aligned} \sum_i \sum_{j\geqslant i} |E\{\varepsilon_{ij}\}| &\leqslant \|a\|\, O(n^{\frac{5}{4}})\,\beta^{-\tau+1} \\ \sum_i \sum_{j\geqslant i} \sigma^2\{\varepsilon_{ij}\} &\leqslant \|a\|^2\, O(n \log n)\,\beta^{-2\tau+2} \end{aligned}$$

$$\ldots \text{ for all levels.} \tag{11.34}$$

Referring to (4.35) and (4.36a), we take the square root of the second expression; we then multiply both expressions by 2, $\|f\|$, $\|x\|$ to get the global bounds

$$\begin{aligned} |E\{\zeta_i\}| &\leqslant \|a\|\, \|f\|\, \|x\|\, O(n^{\frac{5}{4}})\,\beta^{-\tau+1} \\ \sigma\{\zeta_i\} &\leqslant \|a\|\, \|f\|\, \|x\|\, O(\sqrt{n \log n})\,\beta^{-\tau+1}. \end{aligned} \tag{11.35}$$

Thus we have proved the assertions (11.18), (11.22) on the lt type errors. Let us outline the proofs for the remaining assertions. A right-hand side $b_i^{(k)}$ occurring during partial reduction of a block at level ℓ can be bounded by either $\|b^{(1)}\|$ or by

$$\|b_i^{(k)}\| \leqslant \|a\|\, \|x\|\, O(\sqrt{m_\ell}). \tag{11.36}$$

The last bound follows from $b_i^{(k)} = \sum_j a_{ij}^{(k)} x_j$ and also from modifying the proof of the first part of proposition (7.1) in such a way that it applies to the entire row instead of the upper diagonal portion of the row.

The number of product terms subtracted from a right-side coefficient is bounded by μ_{ii}. Hence for one block of level ℓ we get

$$\sum_i |E\{\eta_i\}| \leqslant \|b^{(\cdot)}\| \{\int_0^{m_l} x\,dx + \int_0^{m_l} m_l\,dx\}\, O(1)\, \beta^{-\tau+1}$$

$$= \|b^{(\cdot)}\|\, O(m_l^2)\, \beta^{-\tau+1} = \|b^{(\cdot)}\|\, O(n/4^l)\, \beta^{-\tau+1}. \quad (11.37)$$

This is alternatively bounded by

$$\sum_i |E\{\eta_i\}| \leqslant \|a\|\, \|x\| \{\int_0^{m_l} x\sqrt{m_l}\,dx + \int_0^{m_l} m_l\sqrt{m_l}\,dx\}\, O(1)\, \beta^{-\tau+1}$$

$$= \|a\|\, \|x\|\, O(m_l^{\frac{5}{2}})\, \beta^{-\tau+1}$$

$$= \|a\|\, \|x\|\, O(n^{\frac{5}{4}}/2^{\frac{5l}{2}})\, \beta^{-\tau+1}. \quad (11.38)$$

The blockwise sum for $\sigma^2\{\eta_i\}$ is bounded by

$$\sum_i \sigma^2\{\eta_i\} \leqslant \|b^{(\cdot)}\|^2 \{\int_0^{m_l} x\,dx + \int_0^{m_l} m_l\,dx\}\, O(1)\, \beta^{-2\tau+2}$$

$$= \|b^{(\cdot)}\|^2\, O(m_l^2)\, \beta^{-2\tau+2}$$

$$= \|b^{(\cdot)}\|^2\, O(n/4^l)\, \beta^{-2\tau+2} \quad (11.39)$$

or alternatively by

$$\sum_i \sigma^2\{\eta_i\} \leqslant \|a\|^2\, \|x\|^2 \{\int_0^{m_l} x\, m_l\,dx + \int_0^{m_l} m_l m_l\,dx\}\, O(1)\, \beta^{-2\tau+2}$$

$$= \|a\|^2\, \|x\|^2\, O(m_l^3)\, \beta^{-2\tau+2}$$

$$= \|a\|^2\, \|x\|^2\, O(n^{\frac{3}{2}}/8^l)\, \beta^{-2\tau+2}. \quad (11.40)$$

Eqs. (11.37-40) lead in a straightforward way to eqs. (11.19-20) and (11.23-24). This proves the assertions on the rt type errors.

Turning to back-substitution errors, we are faced with the problem of specifying elementwise bounds on the left and right side coefficients and on the row sum norms of the triangularized system $Rx = s$. From $A = R^T R$ and $F = R^{-1}(R^T)^{-1}$ it is easy to conclude that elementwise bounds on R, R^{-1} are given by

$$\|r\| = O(\sqrt{\|a\|}), \quad \|r^{(-1)}\| = O(\sqrt{\|f\|}). \quad (11.41)$$

Again using $R^T R = A$, and modifying the proofs of proposition 7.1, we show that

$$\|r_{i\cdot}\|_1 = \sqrt{\|a\|}\, O(\sqrt{x_i}), \quad \|r_{i\cdot}\|_2 = \sqrt{\|a\|}\, O(1) \quad (11.42)$$

are valid bounds on the row sum norms of R. It follows that $\sqrt{\|a\|}\, \|x\|\, O(\sqrt{m_l})$ is a bound on the right side s_i of equation i in the triangular system $Rx = s$ when the equation is associated with a block of level l. Note that the triangular system comprises only interior equations. There are no junction equations. The local errors occurring during the treatment of an interior equation i in a block at level l are therefore bounded by

$$|E\{\eta_i\}| \leqslant \sqrt{\|a\|}\, \|x\|\, O(\sqrt{m_l})\, \beta^{-\tau+1}$$

$$\sigma^2\{\eta_i\} \leqslant \|a\|\, \|x\|^2\, O(m_l)\, \beta^{-2\tau+2}. \quad (11.43)$$

We recall the general discussion in section 4.2.2 and note that $x_i \leqslant 2m_l$. Hence the superpositions of all these errors per block are obtained by the now familiar integration procedure as:

$$\sum_i |E\{\eta_i\}| \leqslant \sqrt{\|a\|}\, \|x\| \int_0^{m_l} \sqrt{m_l}\,dx\, O(1)\, \beta^{-\tau+1}$$

$$= \sqrt{\|a\|}\, \|x\|\, O(m_l^{\frac{3}{2}})\, \beta^{-\tau+1}$$

$$= \sqrt{\|a\|}\, \|x\|\, O(n^{\frac{3}{4}}/2^{\frac{3l}{2}})\, \beta^{-\tau+1}$$

$$\sum_i \sigma^2\{\eta_i\} \leqslant \|a\|\, \|x\|^2 \int_0^{m_l} m_l\,dx\, O(1)\, \beta^{-2\tau+2}$$

$$= \|a\|\, \|x\|^2\, O(m_l^2)\, \beta^{-2\tau+2}$$

$$= \|a\|\, \|x\|^2\, O(n/4^l)\, \beta^{-2\tau+2}$$

$$\ldots \text{ for one block of level } l. \quad (11.44)$$

We multiply by 4^l to sum over the blocks per level. When we subsequently sum over the levels, it is important to keep in mind that their number is bounded by $\log_2\sqrt{n} + O(1)$. We finally get

$$\sum_i |E\{\eta_i\}| = \sqrt{\|a\|}\, \|x\|\, O(n)\, \beta^{-\tau+1}$$

$$\sum_i \sigma^2\{\eta_i\} = \|a\|\, \|x\|^2\, O(n \log n)\, \beta^{-2\tau+2}$$

$$\ldots \text{ for the whole network.} \quad (11.45)$$

Multiplying by the norms $\|r^{(-1)}\| = O(\sqrt{\|f\|})$, as given by (11.41), we finally get (11.21) and (1.25). This concludes the proof of the theorem.

11.2.3 Networks without absolute position observations

We now deal with networks whose measurements are only relative. Of course, the measurements are also restricted to local connecting pairs, or small sets of stations situated in close vicinity. The locality of the observations is necessary for applying the nested dissection. Our networks are also assumed to be "locally stable." By this we mean that they are composed of many interlocking small subnetworks which are stable by themselves. Thus we exclude pathological networks which may not be decomposed into local solvable subnetworks. Because the absolute position is not observed, it must be fixed by constraint. We will further distinguish between two practically important classes of networks.

11.2.3.1 Networks obeying the logarithmic law

For these networks the following is asymptotically true:

$$\|a\| \, \|f\| = O(\log n). \tag{11.46}$$

The following types of networks are included in this general class:

(*) Networks with regularly distributed directions, distances, and azimuths, fixed by a small number of constrained stations containing at least one fixed station. We call these networks "direction, distance, and azimuth networks that are weakly constrained."

(*) Direction and distance networks that are weakly constrained by widely spaced fixed stations.

(*) Pure distance networks that are weakly constrained by widely spaced fixed stations.

(*) Pure direction (*i.e.*, purely angular) networks that are constrained by fixing all stations along the perimeter. It is also sufficient if the fixed stations along the perimeter are closely spaced. Such networks include planimetric photogrammetric blocks with dense perimeter control.

All these networks obey the logarithmic law specified in (11.46). For more details refer to Meissl (1969, 1972, and 1974), Borre and Meissl (1974), and Bartelme and Meissl (1974). Combining (11.46) with theorem (11.1), we conclude that the global roundoff errors are asymptotically bounded as

$$|E\{\zeta_i\}| \leqslant \|x\| \, O(n^{\frac{5}{4}} \log n) \, \beta^{-\tau+1} \tag{11.47}$$

$$\sigma\{\zeta_i\} \leqslant (\|x\| + \frac{\|b^{()}\|}{\|a\|}) \, O(\sqrt{n} (\log n)^{\frac{3}{2}}) \, \beta^{-\tau+1} \tag{11.48}$$

$$\sigma\{\zeta_i\} \leqslant \|x\| \, O(n^{\frac{3}{4}} \log n) \, \beta^{-\tau+1}. \tag{11.49}$$

As previously explained in chapters 7 and 9, I believe *that in many practical situations the buildup of* roundoff errors is more closely reflected by the following estimates that have a smaller asymptotic growth rate:

$$|E\{\zeta_i\}| \leqslant \|x\| \, O(n (\log n)^2) \, \beta^{-\tau+1} \tag{11.50}$$

$$\sigma\{\zeta_i\} \leqslant \|x\| \, O(\sqrt{n} (\log n)^{\frac{3}{2}}) \, \beta^{-\tau+1}. \tag{11.51}$$

However, this can be strictly guaranteed only for leveling networks and for special distance and azimuth networks that are equivalent to them. (See Borre and Meissl (1974) for a discussion of such networks.)

11.2.3.2. Networks obeying the "Dutch law"

Such networks are characterized by the following asymptotic law:

$$\|a\| \, \|f\| = O(n). \tag{11.52}$$

Prominent examples of such networks are purely angular networks weakly constrained by widely spaced fixed points, and that contain at least two fixed points. Also planimetric photogrammetric blocks are included if they are constrained in the same way. Refer to Meissl (1976) for a proof of this law, which was postulated by the Dutch geodesists Tienstra, Baarda, and Alberda. Independently, the law was experimentally discovered by the photogrammetrists Ackermann and Ebner.

Combining this law with theorem (11.1), we arrive at the following rigorous asymptotic bounds:

$$|E\{\zeta_i\}| \leqslant \|x\| \, O(n^{\frac{9}{4}}) \, \beta^{-\tau+1} \tag{11.53}$$

$$\sigma\{\zeta_i\} \leqslant (\|x\| + \frac{\|b^{()}\|}{\|a\|}) \, O(n^{\frac{3}{2}} \sqrt{\log n}) \, \beta^{-\tau+1} \tag{11.54}$$

$$\sigma\{\zeta_i\} \leqslant \|x\| \, O(n^{\frac{7}{4}}) \, \beta^{-\tau+1}. \tag{11.55}$$

Again, I believe that the following estimates reflect the true situation in a better way:

$$|E\{\zeta_i\}| \leqslant \|x\| \, O(n^2 \log n) \, \beta^{-\tau+1} \tag{11.56}$$

$$\sigma\{\zeta_i\} \leqslant \|x\| \, O(n^{\frac{3}{2}} \log n) \, \beta^{-\tau+1}. \tag{11.57}$$

11.2.4 Networks with absolute position observations at regularly spaced intervals

If absolute positions are observed at a subset of stations that covers the entire area of the network and is not too widely spaced, the strength of the network, which is reflected by $\|a\|$ and the size of the elements f_{ij} of the inverse, improves drastically. The type of the relative observations therefore is immaterial. We may even deal with a purely angular network, superimposed, for example, by Doppler observations.

Suppose that we deal with two coordinates i,j which belong to two different stations. We denote the distance between these two stations by d_{ij}. The following is asserted:

$$\|a\| \, \|f\| = O(1) \tag{11.58}$$

$$f_{ij} = O(\frac{1}{d_{ij}^k}) \text{ for any } k > 0. \tag{11.59}$$

Thus it turns out that, given the size of $\|a\|$, the elements f_{ij} of the inverse are asymptotically bounded. Furthermore, the f_{ij} decay more rapidly than any negative power of the distance between the two stations involving coordinates i and j.

Statement (11.58) is rather trivial. The position of any station can be derived by nonrigorous adjustment from observations in a subregion that includes one or two stations with observed absolute position. Hence the variance of the position of any station is bounded after nonrigorous adjustment. Rigorous adjustment will yield even smaller variances f_{ii}. Hence $\|f\|$ must be bounded.

A proof of statement (11.59) is not yet available in the geodetic literature. The statement can be proved easily by Fourier analysis for a regular infinite network covering the whole plane. The stations with absolute position observations must form a regular subgrid similar to the one shown in figure 5.13. The network is partitioned into clusters of stations such that any cluster contains one station with observed position. These clusters take over the role of a single station of the Fourier analysis outlined in section 5.5.1. The analog of the matrix function $A(\phi,\psi)$ introduced by eq. (5.19) will be a matrix function which has twice as many rows and columns as there are stations in one cluster. We find that the analog to the kernel function A_{pq} introduced by (5.15) has a true inverse F_{pq} and that its Fourier transform $F(\phi,\psi) = A(\phi,\psi)^{-1}$ is an analytic periodic function. Its Fourier coefficients are the f_{ij} for which the property (11.59) then can be inferred.

Based on (11.58-59) roundoff error estimates can be specified that show a smaller asymptotic growth rate than those of the previous section. The improvement is not very dramatic. However, I believe it is possible to prove that a statement analogous to (11.59) is also valid for the coefficients $a_{ij}^{(p)}$, namely

$$a_{ij}^{(p)} = O\left(\frac{1}{d_{ij}^{k}}\right), \text{ for any } k > 0. \qquad (11.60)$$

Since many details of the proof are yet to be worked out, the remaining statements in this subsection are formulated only as conjectures.

(*) Although the number of elementary operational steps involving one left-side location (i,j) or one right-side location (i) during the reduction of the normals increases indefinitely as the number n of stations increases, mean $E\{\varepsilon_{ij}\}$, $E\{\eta_i\}$, and standard deviation $\sigma\{\varepsilon_{ij}\}$, $\sigma\{\eta_i\}$ of the local roundoff errors ε_{ij}, η_i remain bounded because of (11.60).

(*) Although the number of local errors ε_{ij}, η_i during triangular decomposition and back-substitution increases indefinitely as n increases, the total influence of all these errors upon the global position of a particular station remains bounded because of (11.59).

Hence the following should be true:

$$|E\{\zeta_i\}| = \|x\|\, O(1)\, \beta^{-\tau+1} \qquad (11.61)$$

$$\sigma\{\zeta_i\} = \|x\|\, O(1)\, \beta^{-\tau+1}. \qquad (11.62)$$

The relations (11.59 and 60) express a nearly complete lack of coupling between distant portions of the network. The network is practically only as strong as a collection of separately adjusted subnetworks. By adjusting the whole network in one piece, an enormous number of operational steps are carried out that involve very small operands, whose effect on the final result is practically zero.

11.3 Roundoff Estimates for the UNIVAC 1100/40

Because I received the news that the adjustment of the U.S. network would be done on the UNIVAC 1100/40 after this study had been nearly completed, it was not possible to give this machine proper emphasis in my discussions. (Refer also to the remarks at the end of section 1.5.) I do not consider it a waste of time that I concentrated on the CDC 6600 and the IBM 360 because these machines come close to representing the two extreme cases of true rounding and true chopping. The UNIVAC 1100/40 performs something in between. Its arithmetic is biased, but the bias is difficult to describe. It is easier to specify bounds on the bias that overestimate it somewhat. Once this is done, it is easy to modify our so-called "safe" estimates for the IBM 360 and apply them to the UNIVAC 1100/40.

11.3.1 Double precision floating point arithmetic on the UNIVAC 1100/40

Here we are dealing with a binary machine, *i.e.*, the base $\beta = 2$. Double precision floating point numbers allow for mantissas of $\tau = 60$ bits. Mantissas are considered as being normalized for our purposes. Mantissas of negative numbers are 1's complemented. Normalization implies that the leftmost digit of a positive mantissa is a 1, while for a negative mantissa it is a 0. If a number is used as an operand during an elementary arithmetic operation, the mantissa is placed into the 60 rightmost positions of one of the 72-bit registers that participate in the arithmetic operations. The 12 positions to the left are filled with signbits that are 0's in the case of positive numbers and 1's in the case of negative numbers.

It is sufficient to understand the processes of double precision addition/subtraction and multiplication. The number of divisions and square roots is negligible during Cholesky reduction of the normals in the U.S. network. Let us start with the simple case involving the addition of two positive numbers a, b.

Assume that the exponent of a is not smaller than that of b. The mantissa of a is placed into the 72-bit augend register as described above. The mantissa of b is placed into the addend register. Exponents are placed into separate registers. If the exponent of b is smaller than that of a, then the mantissa of b is preshifted to the right by an appropriate number of positions. Thereby all trailing digits that leave the addend register are lost. There is no guard digit as in the case of the IBM 360. Thus a preshift can decrease the addend b. Now addition is performed in the augend register. Although in reality the adder works subtractively the manufacturer's manual assures us that we may imagine addition in a traditional way. See the reference on the Sperry-Univac 1100 series. If the sum has 61 significant digits, a postshift occurs by one place. It may diminish the sum. We recognize that the result $a \oplus b$ is the one obtained by truly chopping $a+b$.

Let us consider the case where a and b are negative. The complementary mantissas are then placed into the 72-bit registers. A preshift of b can decrease the magnitude of b. Because the sign bits are equal to 1 in the leading positions of the two registers, the addition causes a carry at the leftmost position of the 72-bit adder. The carry is added to the rightmost position. This "end around-carry" ensures that the 1's complement of the result is correctly obtained. A postshift may again reduce the magnitude of the sum. We again recognize that a truly chopped result is obtained.

Things change when we turn to the "subtract magnitude" case. It suffices to outline the case of a subtraction $a-b$, assuming that $a>b>0$. Before addition takes place, the augend register holds a and the addend register holds the complement of b. A preshift of b may decrease its magnitude. Hence the sum will be either correct or too large, but it will never be too small. Normalization of the sum may necessitate a postshift to the left. This will not change the sum. We recognize that $a \ominus b$ is the upward rounded difference $a-b$.

We summarize our findings by stating that the "add magnitude" case of addition/subtraction produces a downward rounded (chopped) result, while the "subtract magnitude" case causes upward rounding.

Therefore, let c be an integer power of the base $\beta = 2$ such that c bounds $|a|$, $|b|$, and $|a \pm b|$. It holds that

$$|E\{\varepsilon^{(\pm)}\}| \leqslant \frac{c}{2} 2^{-60} \tag{11.63}$$

$$\sigma\{\varepsilon^{(\pm)}\} \leqslant \frac{c}{\sqrt{12}} 2^{-60}. \tag{11.64}$$

The case of double precision floating point multiplication is more involved. Mantissas of the operands are made positive before multiplication starts. The signs are treated separately. A multiplicand register holds a, a multiplier register holds b. The multiplicand is repeatedly added to a properly shifted partial product which is accumulated in a 72-bit accumulator. These additions take place in agreement with the digits of the multiplier. Of course, the process begins by looking at the least significant rightmost digits of the multiplier and proceeds to the most significant leftmost digits. Things are speeded up by treating the digits in pairs. In the case of a pair "00," a right shift by only two digits takes place in the accumulator. Any right shift causes a loss of digits shifted out of the accumulator. A pair "01" requires addition of the multiplicand followed by a right shift of two places. A pair "10" causes a right shift of one digit, followed by addition of the multiplicand, followed by another right shift of one digit. A pair "11" will cause the multiplicand to be subtracted from the partial product which is then right shifted by two places. A borrow of one must then be added to the portion of b to the left of the pair of digits just treated. The procedure replaces the addition of three times the number a by subtracting a and adding it four times. In this way the leftmost 60 digits of the product of the 2 mantissas are correctly obtained. The 60 rightmost positions are chopped off. The leading digit to the left can be zero. A left postshift by one place is then necessary. In this case, only 59 significant digits are recovered.

We summarize by stating that $a \odot b$ equals $a * b$, chopped to either 60 or 59 digits depending on whether the full product of the two mantissas has 120 or 119 nontrivial digits. In order to bound the elementary roundoff error of multiplication, we let c be an integer power of the base 2 that bounds $a * b$. It follows that

$$|E\{\varepsilon^{(*)}\}| \leqslant c\, 2^{-60} \tag{11.65}$$

$$\sigma\{\varepsilon^{(*)}\} \leqslant \frac{2c}{\sqrt{12}} 2^{-60} \tag{11.66}$$

Note that these bounds are twice as large as those related to true chopping.

11.3.2 Safe bounds for the UNIVAC 1100/40

All that is needed now is to repeat the calculations in chapter 8 that apply to the IBM 360 using modified bounds on mean $E\{\varepsilon_{ij}^{(r)}\}$ and standard deviation $\sigma\{\varepsilon_{ij}^{(r)}\}$ of the elementary roundoff errors. The new bounds are obtained by applying factors to the corresponding bounds in chapter 8. The factors are obtained as follows:

(*) The overwhelming majority of elementary roundoff errors occur in pairs $\varepsilon_{ijk}^{(-)}$ and $\varepsilon_{ijk}^{(+)}$ during evaluation and addition of the product terms $r_{ki}r_{kj}$. As compared to the IBM 360, the bounds for the UNIVAC 1100/40 on the elementary roundoff errors during multiplication are multiplied by a factor of 2, if we temporarily disregard the difference due to different β and τ. Hence we apply a factor of $(1+2)/2 = 1.5$ in the case of the bias, and a factor of $\sqrt{(1^2+2^2)/2} \doteq 1.58$ in the case of the standard deviation.

(*) Because of the difference between β and τ on the two computers, we apply a factor of $2^{-60}/16^{-14} = 1/16$ to all estimates.

(*) We must reexamine the choice of the factors c and $\bar{c}$ given in chapter 8. These factors bound elements a_{ii} in the two different size classes in terms of integer powers of the base β. Occasionally a factor of $\sqrt{\beta}$ or of $2\sqrt{\beta}$ is also involved. We find that there is no reason to further adjust the bias estimate. In the case of the standard deviation we see that another factor of $\sqrt{2}/\sqrt{16} \doteq 0.35$ must be taken into account.

Summarizing, we conclude that a factor of 0.094 must be applied to the bias type estimates and a factor of 0.035 must be applied to the standard deviation estimates. Treating the last column of tables 8.5 and table 8.7 in the indicated way, we arrive at the numbers listed in table 11.1.

TABLE 11.1.—*Bound on bias and standard deviation of the global roundoff errors encountered on the UNIVAC 1100/40 during the first iteration of the U.S. network adjustment.*

Quad ϕ	Quad λ	Bound on $E\{\zeta_p\}$ m	Bound on $\sigma\{\zeta_p\}$ m
39	77	0.00014	3.2E-8
47	69	0.00019	4.4E-8
47	121	0.00013	2.6E-8
41	97	0.00010	2.0E-8
35	111	0.00009	2.0E-8

About one more correct decimal digit is gained on the UNIVAC 1100/40 as compared with the IBM 360. This is, of course, already indicated by the factor 0.094 which is applied to the bias type estimates.

11.3.3 More realistic estimates

Our description will be very brief for adapting the estimates of chapter 9 to the UNIVAC 1100/40. When the NGS algorithm accumulates the product sums, the "add magnitude" case will prevail. This is strictly true for the product sums which refer to diagonal positions. For the off-diagonals it is not strictly true; however, I believe the transition of a_{ij} to $a_{ij}^{(i-1)}$ proceeds monotonically in many cases. Since the "add magnitude" case is associated with chopping, most arguments used in chapter 9, in the case of the IBM 360, carry over to the UNIVAC 1100/40. Hence, I believe that the estimates can be lowered by about one to two decimal places, if one allows for a small probability of error.

11.4. Recovering $[pvv]$ from the Reduction of the Normal Equations.

In textbooks on geodetic least-squares adjustment, *e.g.*, Jordan/Eggert/Kneissl (1961), the symbol $[pvv]$ denotes the weighted square sum of the residuals of the observations after adjustment. (v are the corrections to the observations after the adjustment.) The symbol $[p\ell\ell]$ stands for the weighted square sum of residuals before adjustment. (ℓ are the discrepancies between observations and corresponding values calculated from approximate coordinates.) If the right-side vector b of the normal equations is augmented by $[p\ell\ell]$ and if the reduction of the normals is carried out one step farther, then $[p\ell\ell]$ is transformed into $[pvv]$. The relevant formula is particularly simple, if Cholesky reduction is used. It reads as

$$[pvv] = [p\ell\ell] - \sum_{i=1}^{n} s_i^2. \qquad (11.67)$$

Refer to eq. (3.7) for the definition of s_i. The question is: to what extent is $[pvv]$ perturbed by roundoff errors during the solution of the normals?

The quantity $[pvv]$, divided by the difference of the number of observations minus the number of unknowns, gives the squared unit weight error. The unit weight error should not differ appreciably from 1. The number of observations is 2-3,000,000. Hence we expect $[pvv] < 10^7$.

Next we estimate how much $[p\ell\ell]$ can exceed $[pvv]$. The ratio $[p\ell\ell]/[pvv]$ depends on the local quality of the approximate coordinates. I believe that discrepancy ℓ_i, calculated from the difference of an observed quantity minus the quantity computed from the approximate station coordinates, does not exceed the rms error of the measurement by a factor of 10 to 100. Note that the observations connect only stations within a close vicinity. Although we expect coordinate shifts to exceed 10 m during the first iteration, I do not believe that the relative position of neighboring stations calculated from the approximate coordinates will deviate by more than a few decimeters from the relative position calculated from the adjusted coordinates. Hence during the first iteration, I expect that the ratio of $[p\ell\ell]/[pvv]$ will be below $10^3 - 10^4$. The ratio will be much closer to 1 during subsequent iterations. Then we will observe a ratio $[p\ell\ell]/[pvv] \cong 1$. In the first iteration, we can have

$[pll] \cong 10^{11}$, and consequently $\Sigma\, s_i^2 \cong 10^{11}$. The average size of an s_i is therefore $\sqrt{10^{11}/350{,}000} \cong 500$. Our problem is to estimate the roundoff errors suffered by the right-hand sides s_i of the triangular system. Because errors are encountered during back substitution, we can assume that the s_i are not less accurate than the solution vector is. The relative accuracy of s_i and x_i will be about equal if i approaches n, the number of equations. If the solution vector is good to about four to five leading decimal digits, then $\Sigma\, s_i^2$ is good to the same number of leading digits. (The 350,000 summations can be safely ignored). If $[pll] \cong 10^{11}$, $\Sigma\, s_i^2 \cong 10^{11}$, and $[pvv] \cong 10^7$, then $[pvv]$ may be totally wrong, because of a wiping out of the leading digits.

It should not be viewed as a catastrophe if $[pvv]$ is wrong after the first iteration. Subsequent iterations will give $a\,[pvv]$ which is good to as many digits as the solution vector. If $[pvv]$ were to have some correct digits even after the first iteration, then the number of correct digits of the solution vector can be estimated from those of $[pvv]$, plus the difference of decimal digits of $[pll]$ and $[pvv]$.

REFERENCES

Avila, J., Malloy, B., and Tomlin, J., 1978: Use of the ILLIAC IV for the readjustment of the North American Datum. *T.M.* 5732, Institute for Advanced Computation, Sunnyvale, Calif. 94085, 87 pp.

Bartelme, N. and Meissl, P., 1974: Strength analysis of distance networks. *Reports of the Geodetic Institutes of the Technical University Graz,* No. 15, Graz, Austria, 57 pp.

Bartelme, N. and Meissl, P., 1975: Theoretical analysis of rounding error propagation during the direct solution of geodetic normal equations of the leveling type. *In* Contributions of the Graz Group to the XVI General Assembly of the International Association of Geodesy/International Union of Geodesy and Geophysics (IAG/IUGG), Grenoble, France, Aug. 1975, edited by P. Meissl, H. Moritz, and K. Rinner. *Reports of the Geodetic Institutes of the Technical University Graz,* No. 20, Austria, 1–16.

Bartelme, N. and Meissl, P., 1977: Rounding error propagation in a general leveling network. *Proceedings of the International Symposium on Optimization of Design and Computation of Control Networks.* International Association of Geodesy, Sopron, Hungary, 1977, Akademiai Kiado, Budapest, 437–458. (Invited paper)

Beresford/Parlett, 1976: Roundoff error in the solution of finite element systems. Address to the Symposium on Formulations and Computational Algorithms in Finite Element Analysis, Massachusetts Institute of Technology, sponsored by the United States and Germany, August 1976.

Borre, K. and Meissl, P., 1974: Strength analysis of leveling type networks. An application of random walk theory. *Report of the Danish Geodetic Institute* No. 50, Copenhagen, Denmark, 80 pp.

Bossler, J. D., 1976: The new adjustment of the North American horizontal datum. *EOS Transactions of the American Geophysical Union,* 57 (8), 557–562.

Dracup, J. F. 1975: Use of Doppler positions to control classical geodetic networks. Paper presented to the XVI General Assembly of IUGG/IAG, Grenoble, France, 1975, 12 pp. (Available from National Oceanic and Atmospheric Administration, Rockville, Md. 20852.)

Ebner, H. and Mayer, R., 1976: Numerical accuracy of block adjustments. *Photogrammetria,* 32, 101–109.

Erismann, A. M. and Reid, J. K., 1973: Monitoring the stability of triangular factorization of a sparse matrix, *Numerische Mathematik,* 22, 183–186.

Feldstein, A., 1976: Convergence estimates for the distribution of trailing digits. *Journal of Association for Computing Machinery,* 23 (2), 287–297.

Gear, C. W., 1975: Numerical errors in sparse linear systems. *Reports of the Department of Computer Science,* University of Illinois, Urbana-Champaign.

George, A., 1973: Nested dissection of a regular finite element mesh. *SIAM Journal of Numerical Analysis,* 10 (2), 345–363.

George, A., 1977: Numerical experiments using nested dissection methods to solve n by n grid problems. *SIAM Journal of Numerical Analysis,* 14 (2), 161–179.

Henrici, P., 1964: *Elements of Numerical Analysis.* Wiley, New York/London/Sydney, 291–321.

Jennings, A., 1977: *Matrix Computation for Engineers and Scientists.* Wiley, New York, N.Y., 115–116.

Jordan/Eggert/Kneissl, 1961: *Handbuch der Vermessungskunde, Band 1: Mathematische Grundlagen, Ausgleichsrechnung und Rechenhilfsmittel.* 10. Auflage. Metzler'sche Verlagsbuchhandlung, Stuttgart, 391–503.

Meissl, P., 1969: A note on error propagation in leveling networks. *Proceedings of the IV Symposium on Mathematical Geodesy,* Trieste, 1969.

Meissl, P., 1972: A theoretical error propagation law for Anblock networks with constrained boundary. *Oesterreichische Zeitschrift fuer Vermessungswesen,* 60, 61–65.

Meissl, P., 1973: On the random error propagation in relatively large networks, *Studia geophysica et geodetica,* 17, 7–21.

Meissl, P., 1974: Strength of continental terrestrial networks. Paper presented to International Symposium on Problems Related to the Redefinition of the North American Geodetic Networks, University of New Brunswick, Fredicton, N.B., Canada, May 20–25, 1974. *Canadian Surveyor,* 28 (5), 582–589.

Meissl, P., 1975: Die Vermeidung von Rechenunschaerfen infolge von Gewichtsinhomogenitaeten bei einem Netzausgleich, *Vermessung, Photogrammetrie, Kulturtechnik,* 1975, 253–256.

Meissl, P., 1975a: Report of Special Study Group No. 4:38 of IAG, "Computer Techniques in Geodesy." Presented to the XVI General Assembly of the IAG/IUGG, Grenoble, France, Aug., 1975. *In* Contributions of the Graz Group to the XVI General Assembly of IUGG/IAG, Grenoble, France, edited by P. Meissl, H. Moritz, and K. Rinner , *Reports of the Geodetic Institutes of the Technical University Graz,* No. 20, Graz, Austria, 91–110.

Meissl, P., 1976: Strength analysis of two-dimensional angular Anblock networks, *Manuscripta Geodaetica,* 1 (4), 293–333.

Moose, R. E. and Henriksen, S. W., 1976: Effect of Geoceiver observations upon the classical triangulation network. *NOAA Technical Report* NOS 66 NGS 2, National Oceanic and Atmospheric Administration, Rockville, Md., 65 pp. (Available from National Technical Information Service (NTIS), Springfield, Va. 22161, accession no. PB260921.)

Peters, G. and Wilkinson, J. H., 1975: On the stability of Gauss-Jordan elimination with pivoting. *Journal of Association for Computing Machinery,* 18, 20–24.

Poder, K. and Tscherning, C. C., 1973: Cholesky's method on a computer. Internal Report No. 8 of the Danish Geodetic Institute, Copenhagen, 22 pp.

Rubinstein, M. and Rosen, R., 1970: Error analysis in structural computation. *Journal of the Franklin Institute,* 290, 37–48.

Schwarz, C. S., 1978: TRAV10 horizontal network adjustment program. *NOAA Technical Memorandum* NOS NGS-12, National Oceanic and Atmospheric Administration, Rockville, Md., 52 pp. (Available from NTIS, Springfield, Va., accession no. PB283087.)

Snay, R. A., 1976: Reducing the profile of sparse symmetric matrices. *NOAA Technical Memorandum* NOS NGS-4, National Oceanic and Atmospheric Administration, Rockville, Md., 24 pp. (Available from NTIS, Springfield, Va., accession no. PB258476.)

SPERRY UNIVAC: Chapter 4, CAU arithmetic section, *SPERRY UNIVAC 1110 Series, 1110 and 1100/40 Systems Processor and Storage.* Sperry Rand Corp., St. Paul, Minn. 55165.

Spitzer, F., 1964: *Principles of Random Walk.* Van Nostrand, Toronto/New York/London, 121–128.

Sterbenz, P. H., 1974: *Floating Point Computation.* Prentice Hall, New York, N.Y., 316 pp.

Stoehr, A., 1950: Ueber einige partielle Differenzengleichungen mit konstanten Koeffizienten. *Mathematische Nachrichten,* 3, 208–242, 295–315, 330–357.

Timmerman, E., 1978: Pilot test of block validation. National Geodetic Survey, internal report, National Oceanic and Atmospheric Administration, Rockville, Md., 74 pp.

Vincenty, T., 1975: Experiments with adjustments of geodetic networks and related subjects. National Geodetic Survey internal report, National Oceanic and Atmospheric Administration, Rockville, Md., 46 pp.

Vincenty, T., 1976: Determination of North American Datum 1983 coordinates of map corners. *NOAA Technical Memorandum* NOS NGS-6, 8 pp. (Available from NTIS, Springfield, Va., accession no. PB262442.)

Wilkinson, J. H., 1961: Error analysis of direct methods of matrix inversion. *Journal of Association for Computing Machinery,* 8, 281–335.

Wilkinson, J. H., 1963: *Rounding Errors in Algebraic Processes.* Prentice Hall, New York, N.Y., 161 pp.

Zienkiewics, O. C., 1971: *The Finite Element Method in Engineering Science.* McGraw-Hill, New York, N.Y., 107–110.

☆ U.S. GOVERNMENT PRINTING OFFICE : 1980 O—297-468

U.S. DEPARTMENT OF COMMERCE
National Oceanic and Atmospheric Administration
National Ocean Survey
National Geodetic Survey, OA/C18x2
Rockville, Md. 20852

OFFICIAL BUSINESS

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

PRINTED MATTER