Large Scale Matrix Problems

Edited by

Åke Björck
Robert J. Plemmons
Hans Schneider

Large Scale Matrix Problems

Edited by

Åke Björck
Linköping University
Linköping, Sweden

Robert J. Plemmons
University of Tennessee
Knoxville, Tennessee

Hans Schneider
University of Wisconsin—Madison
Madison, Wisconsin

NORTH HOLLAND
New York • Oxford

Elsevier North Holland, Inc.
52 Vanderbilt Avenue, New York, New York 10017

Distributors outside the United States and Canada:

Elsevier/North-Holland
335 Jan van Galenstraat, P.O. Box 211
Amsterdam, The Netherlands

The material in this book also appeared as Volume 34, 1980 of *Linear Algebra and Its Applications*. ©Elsevier North Holland, Inc.

The paper by Alan George and Michael T. Heath (pp. 69–83) is in the public domain.

Library of Congress Cataloging in Publication Data

Main entry under title:

Large scale matrix problems.

Bibliography: p.
Includes index.
1. Matrices. I. Björck, Ake, 1934-
II. Plemmons, Robert J. III. Schneider, Hans, 1927 (Jan. 24)-
QA188.L37 512.9′434 80-22058
ISBN 0-444-00563-3

Desk Editor Valerie DeBenedette
Design Series
Production Managers John Mickelbank and Joanne Jay
Compositor Science Typographers, Inc.
Printer Capital City Press

Manufactured in the United States of America

Contents

I. LEAST SQUARES AND APPLICATIONS

GENE H. GOLUB (Stanford, CA) AND ROBERT J. PLEMMONS (Knoxville, TN)
Large-Scale Geodetic Least-Squares Adjustment by Dissection and Orthogonal Decomposition 3

ERIC GROSSE (Stanford, CA)
Tensor Spline Approximation 29

Å. BJÖRCK (LINKÖPING, SWEDEN) AND I. S. DUFF (Didcot, U.K.)
A Direct Method for the Solution of Sparse Linear Least Squares Problems 43

ALAN GEORGE (Waterloo, Ont., Canada) AND MICHAEL T. HEATH (Oak Ridge, TN)
Solution of Sparse Linear Least Squares Problems Using Givens Rotations 69

II. SYSTEMS OF LINEAR EQUATIONS AND APPLICATIONS

M. NEUMANN (Nottingham, U.K.)
A Combined Direct-Iterative Approach for Solving Large Scale Singular and Rectangular Consistent Systems of Linear Equations 85

ROBERT R. BITMEAD (Queensland, Australia) AND BRIAN D. O. ANDERSON (Newcastle, Australia)
Asymptotically Fast Solution of Toeplitz and Related Systems of Linear Equations 103

BENGT ASPVALL (Stanford, CA) AND YOSSI SHILOACH (Haifa, Israel)
A Fast Algorithm for Solving Systems of Linear Equations with Two Variables per Equation 117

JOHN DE PILLIS (Riverside, CA)
How to Embrace Your Spectrum for Faster Iterative Results 125

ALAN GEORGE AND HAMZA RASHWAN (Waterloo, Ont., Canada)
On Symbolic Factorization of Partitioned Sparse Symmetric Matrices 145

David M. Young and Kang C. Jea (Austin, TX)
Generalized Conjugate-Gradient Acceleration of Nonsymmetrizable Iterative Methods 159

III. EIGENVALUE PROBLEMS

D. E. Longsine and S. F. McCormick (Fort Collins, CO)
Simultaneous Rayleigh-Quotient Minimization Methods for $Ax=\lambda Bx$ 195

C. C. Paige (Montreal, Que., Canada)
Accuracy and Effectiveness of the Lanczos Algorithm for the Symmetric Eigenproblem 235

E. Seneta (Sydney, Australia)
Computing the Stationary Distribution for Infinite Markov Chains 259

Y. Saad (Grenoble, France)
Variations on Arnoldi's Method for Computing Eigenelements of Large Unsymmetric Matrices 269

IV. OPTIMIZATION PROBLEMS

Ikuyo Kaneko (Madison, WI) and Jong-Shi Pang (Pittsburgh, PA)
Some *n* by *dn* Linear Complementarity Problems 297

Tommy Elfving (Linköping, Sweden)
On Some Methods for Entropy Maximization and Matrix Scaling 321

James R. Bunch (La Jolla, CA) and Linda Kaufman (Murray Hill, NJ)
A Computational Method for the Indefinite Quadratic Programming Problem 341

Dianne Prost O'Leary (College Park, MD)
A Generalized Conjugate Gradient Algorithm for Solving a Class of Quadratic Programming Problems 371

Author Index 401

Subject Index 403

Preface

The purpose of this special issue is to present a collection of papers in numerical linear algebra which involve the development and analysis of rigorous mathematical models or algorithms for solving problems involving large scale matrices. In recent years, large scale matrix problems of ever-increasing size have arisen. One reason for this is that modern acquisition technology allows the collection of massive amounts of data. Another factor is the tendency of engineers and scientists to formulate more and more complex and comprehensive models in order to obtain fine resolution and realistic detail in describing physical systems. It is important to note that the ability of mathematicians and computer scientists to handle increasingly larger problems is at least as much due to the improvement of existing algorithms and software and the development of new and more elaborate methods as to the increase in the computing power of modern machines. Particular areas in which such large scale matrix problems occur include the least squares adjustment of geodetic data, the least squares fitting of multivariate data by splines, the computation of stationary distribution vectors of infinite Markov decision chains, the computation of eigenelements of large symmetric and unsymmetric matrices, the solution of large scale quadratic programming problems and the solution of maximum entropy problems in image reconstruction and transportation planning. Each such application area is represented in one or more papers in this issue.

We have loosely organized the papers into general categories which deal, respectively, with (1) Least squares and applications, (2) Systems of linear equations and applications, (3) Eigenvalue problems, and (4) Optimization problems. Within each category we have chosen to arrange the papers in the order in which they were recorded as received by the Editors. The general category in which a specific paper is placed is determined partly by its area of application and partly by its mathematical character.

The first category contains papers which describe new or improved algorithms for solving large sparse linear least squares problems. Applications are given here to the adjustment of massive amounts of geodetic data and to the fitting of multivariate data by tensor spline approximations.

The second category is concerned with methods for solving systems of linear equations. The methods involve both iterative and sparse matrix direct techniques, together with combinations of the two. Particular attention is paid here to the speed of the algorithms in question.

Category three contains papers on computing eigenelements of large sparse matrices. The methods described here involve Raleigh quotient minimization, the Lanczos algorithm, and variations of Arnoldi's method.

The fourth and last category is concerned with selected problems involving optimization techniques. Linear complementarity problems and entropy maximization problems and applications are discussed, as well as new techniques for solving certain classes of quadratic programming problems.

In a more general context, this special issue presents research papers in numerical linear algebra, but with considerable influence from computer science. In addition, many of the papers are permeated with the application of large scale matrix algorithms and software to the solution of current and relevant engineering and scientific problems.

ÅKE BJÖRCK
ROBERT J. PLEMMONS
HANS SCHNEIDER

Large-Scale Geodetic Least-Squares Adjustment by Dissection and Orthogonal Decomposition

Gene H. Golub*
Computer Science Department
Stanford University
Stanford, California 94305

and

Robert J. Plemmons**
Departments of Computer Science and Mathematics
University of Tennessee
Knoxville, Tennessee 37916

Submitted by Å. Björck

ABSTRACT

Very large-scale matrix problems currently arise in the context of accurately computing the coordinates of points on the surface of the earth. Here geodesists adjust the approximate values of these coordinates by computing least-squares solutions to large sparse systems of equations which result from relating the coordinates to certain observations such as distances or angles between points. The purpose of this paper is to suggest an alternative to the formation and solution of the normal equations for these least-squares adjustment problems. In particular, it is shown how a block-orthogonal decomposition method can be used in conjunction with a nested dissection scheme to produce an algorithm for solving such problems which combines efficient data management with numerical stability. The approach given here parallels somewhat the development of the natural factor formulation, by Argyris et al., for the use of orthogonal decomposition procedures in the finite-element analysis of structures. As an indication of the magnitude that these least-squares adjustment problems can sometimes attain, the forthcoming readjustment of the North American Datum in 1983 by the National Geodetic Survey is discussed. Here it becomes necessary to linearize and solve an overdetermined system of approximately 6,000,000 equations in 400,000 unknowns—a truly large scale matrix problem.

*Research supported in part by U.S. Army Grant DAAG29-77-G0179 and by National Science Foundation Grant MCS78-11985.

**Research supported in part by U.S. Army Grant DAAG29-77-G0166.

1. INTRODUCTION

Recent technological advances have made possible the collection of vast amounts of raw data describing certain physical phenomena. As a result, the sheer volume of the data has necessitated the development of new elaborate schemes for processing and interpreting it in detail. An example is in the adjustment of geodetic data.

Geodesy is the branch of applied mathematics which is concerned with the determination of the size and shape of the earth, and the directions of lines and the coordinates of stations or points on the earth's surface. Applications of this science include mapping and charting, missile and space operations, earthquake prediction, and navigation. The current use of electronic distance-measuring equipment and one-second theodolites for angle measurements by almost all surveyors necessitates modern adjustment procedures to guard against the possibility of blunders as well as to obtain a better estimate of the unknown quantities being measured. The number of observations is always larger than the minimum required to determine the unknowns. The relationships among the unknown quantities and the observations lead to an overdetermined system of nonlinear equations. The measurements are then usually adjusted in the sense of least squares by computing the least-squares solution to a linearized form of the system that is not rank-deficient.

In general, a *geodetical position network* is a mathematical model consisting of several mesh points or geodetic stations, with unknown positions over a reference surface or in three-dimensional space. These stations are normally connected by lines, each representing one or more observations involving the two stations terminating the line. The observations may be angles or distances, and thus they lead to nonlinear equations involving, for example, trigonometric identities and distance formulas relating the unknown coordinates. Each equation typically involves only a small number of unknowns.

As an illustration of the sheer magnitude that some of these problems can attain, we mention the readjustment of the North American Datum—a network of reference points on the North American continent whose longitudes, latitudes and, in some cases, altitudes must be known to an accuracy of a few centimeters. This ten-year project by the U.S. National Geodetic Survey is expected to be completed by 1983. The readjusted network with very accurate coordinates is necessary to regional planners, engineers, and surveyors, who need accurate reference points to make maps and specify boundary lines; to navigators; to road builders; and to energy-resource developers and distributors. Very briefly, the problem is to use some 6,000,000 observations relating the positions of approximately 200,000 stations (400,000 unknowns) in order to readjust the tabulated values for their latitudes and longitudes. This leads to one of the largest single computational efforts ever attempted—that of computing a least-squares solution of a very sparse system of 6,000,000 nonlinear equations in 400,000 unknowns. This problem is described in detail by Meissl [22], by Avila and Tomlin [4], and from a layman's point of view by Kolata [20] in *Science*.

In general then, geodetical network adjustment problems can lead (after linearization) to a very large sparse overdetermined system of m linear equations in n unknowns

$$Ax \approx b, \tag{1.1}$$

where the matrix A, called the *observation matrix*, has full column rank. The *least-squares solution* to (1.1) is then the unique solution to the problem

$$\min_x \| b - Ax \|_2 .$$

An equivalent formulation of the problem is the following: one seeks to determine vectors y and r such that $r + Ay = b$ and $A^t r = 0$. The least-squares solution to (1.1) is then the unique solution y to the nonsingular system of *normal equations*

$$A^t A y = A^t b. \tag{1.2}$$

The linear system of equations (1.2) is usually solved by computing the *Cholesky factorization*

$$A^t A = R^t R, \qquad R =$$

and then solving $R^t w = A^t b$ by forward substitution and $Ry = w$ by back substitution. The upper triangular matrix R is usually called the Cholesky factor of $A^t A$, but we will use the term *Cholesky factor of* A.

Most algorithms for solving geodetic least-squares adjustment problems (see [3], [7], [22], or [4]) typically involve the formation and solution of some (weighted) form of the normal equations (1.2). But because of the size of these problems and the high degree of accuracy desired in the coordinates, it is important that particular attention be paid to sparsity considerations when forming $A^t A$ as well as to the numerical stability of the algorithm being used. It is generally agreed in modern numerical analysis theory (see [15], [21], or [25]) that orthogonal decomposition methods applied directly to the matrix A

in (1.1) are preferable to the calculation of the normal equations whenever numerical stability is important. Since A has full column rank, the Cholesky factor R of A can be computed by

$$Q^tA=\begin{bmatrix} R \\ 0 \end{bmatrix}, \qquad Q^tQ=I, \qquad R=\begin{bmatrix} \diagdown & \\ 0 & \diagdown \end{bmatrix} \tag{1.3}$$

where the orthogonal matrix Q results from a finite sequence of orthogonal transformations, such as Householder reflections or Givens rotations, chosen to reduce A to upper triangular form. It should be mentioned that the use of Givens rotations is normally preferable to the use of Householder reflections for orthogonal decompositions involving sparse matrices (see [8]). In addition, A should normally be preordered by some scheme in order to reduce the fill-in in R. This could be accomplished, for example, by a symbolic formation of A^tA, followed by an ordering scheme given in [12], or by permuting the rows and columns of A directly. Further research on this topic is needed.

Since A has the *orthogonal decomposition*

$$A=Q\begin{bmatrix} R \\ 0 \end{bmatrix},$$

then defining

$$Q^tb=\begin{bmatrix} c \\ d \end{bmatrix},$$

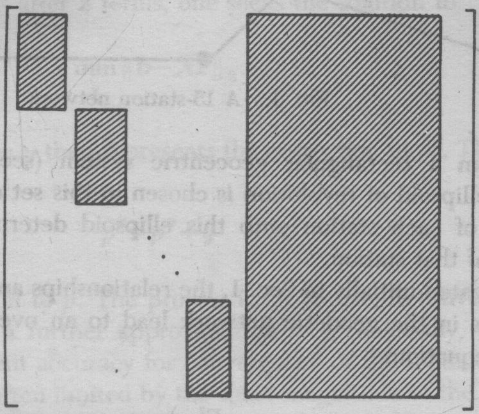
where c is an n-vector, the least-squares solution y to (1.1) is obtained by solving $Ry=c$ by back substitution. The greater numerical stability of the orthogonal-decomposition method results from the fact that the spectral condition number of A^tA in the normal equations (1.2) is the square of the spectral condition number of A. The orthogonal decomposition method (1.3) has other advantages, including the ease with which updating and downdating of the system (1.1) can be accomplished, and the fact that possible fill-in in forming the normal equations is avoided (see, for example, [5]). However, orthogonal decomposition techniques for solving large sparse least-squares problems such as those in geodesy have generally been avoided, in part because of tradition and in part because of the lack of effective means for preserving sparsity and for managing the data.

Modern techniques for solving large-scale geodetic adjustment problems have involved the use of a natural form of nested dissection, called Helmert blocking by geodesists, to partition and solve the normal equations (1.2). Such techniques are described in detail in [4], in [18], and in [22], where error analyses are given.

The purpose of this paper is to develop an alternative to the formation and solution of the normal equations in geodetic adjustments. We show how the orthogonal decomposition method can be combined with a nested dissection scheme to produce an algorithm for solving such problems that combines efficient data management with numerical stability.

In subsequent sections the adjustment problem is formulated, and it is shown how nested dissection leads to an observation matrix A in (1.1) of the special partitioned form

$A =$



(1.4)

where the diagonal blocks are normally rectangular and may be dense or sparse, and where the large block on the right-hand side is normally sparse with a very special structure. The form (1.4) is analyzed and a block-orthogonal decomposition scheme is described. The final section contains some remarks on the advantages of the approach given in the paper and relates the concepts mentioned here to further applications. Numerical experiments and comparisons will be given elsewhere, e.g., in [16].

2. GEODETIC ADJUSTMENTS

In this paper we consider geodetical position networks consisting of mesh points, called stations, on a two-dimensional reference surface. Associated with each station there are two coordinates. A line connecting two stations is roughly used to indicate that the coordinates are coupled by one or more physical observations. Thus the coordinates are related in some equation that may involve, for example, distance formulas or trigonometric identities relating angle observations. An example of such a network appears in Fig. 1.

More precisely, one considers a coordinate system for the earth and seeks to locate the stations exactly, relative to that system. Usually coordinates are

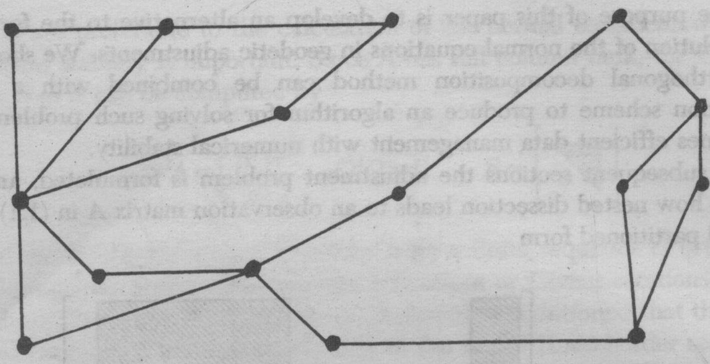


FIG. 1. A 15-station network.

chosen from a rectangular geocentric system (see [7]). Furthermore, a reference ellipsoid of revolution is chosen in this set of coordinates, and the projection of each station onto this ellipsoid determines the latitude and longitude of that station.

As indicated initially in Sec. 1, the relationships among the coordinates of the stations in the geodetic network lead to an overdetermined system of nonlinear equations

$$F(p)=q, \tag{2.1}$$

where

$$p=\text{vector of unknown coordinates},$$

$$q=\text{vector of observations}.$$

The components of $F(p)$ represent the equations that express the relationships among the unknown parameters and the observations or measurements made, for example, by surveyors.

A common procedure for solving the overdetermined system (2.1) is the method of *variation of parameters*. (This is generally called the Gauss-Newton nonlinear least-squares algorithm in the mathematical literature.) Approximate coordinates are known *a priori*. Let

$$p^0=\text{current vector of approximate coordinates}.$$

Then if F has a Taylor's series expansion about p^0, there follows the relationship

$$F(p)=F(p^0)+F'(p^0)(p-p^0)+\cdots,$$

where $F'(p^0)$ denotes the Jacobian of F at p^0. Then taking

$$A = F'(p^0),$$

$$x = p - p^0,$$

$$b = q - F(p^0)$$

and truncating the series after 2 terms, one seeks the solution to

$$\min_x \| b - Ax \|_2. \tag{2.2}$$

The least-squares solution y then represents the correction to p^0. That is, one takes

$$p^1 = p^0 + y$$

as the next approximation to p. The process is, of course, iterative, and one can use p^1 to compute a further approximation to p. Normally, the initial coordinates have sufficient accuracy for convergence of the method, but the number of iterations is often limited by the sheer magnitude of the computations. Thus a very accurate approximation to y is desired.

Actually, the equations are usually weighted by use of some positive diagonal matrix W, where the weights are chosen to reflect the confidence in the observations: thus (2.2) becomes

$$\min_x \| W^{1/2}b - W^{1/2}Ax \|_2.$$

For simplicity, we will use (2.2) in the analysis to follow. The procedure we discuss, however, will not be complicated by the weights.

Due to the sheer volume of the data to be processed in many adjustment problems, it is imperative to organize the data in such a way that the problem can be broken down into meaningful mathematical subproblems which are connected in a well-defined way. The total problem is then attacked by "solving" the subproblems in a topological sequence. This "substructuring" or "dissection" process has been used by geodesists for almost a century. The method they have employed dates back to Helmert (1880) [19] and is known as Helmert blocking (see [27] for a historical discussion).

In Helmert blocking, geographical boundaries for the region in question are chosen to partition it into regional blocks. This technique orders the stations appropriately in order to establish barriers which divide the network

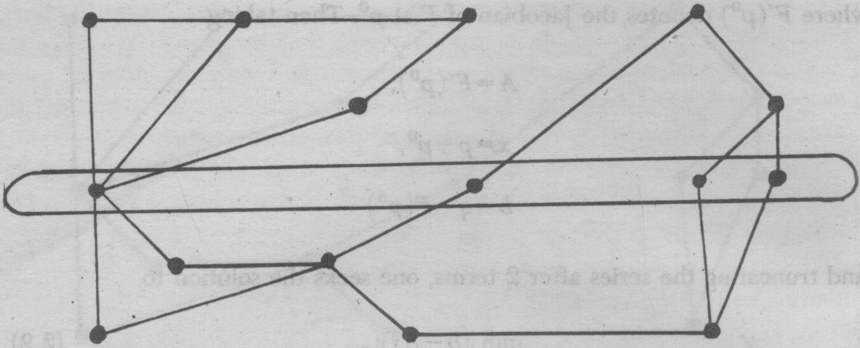


FIG. 2. One level of Helmert blocking.

into blocks. The barriers are chosen so that the interior stations in one block are not coupled by observations to interior stations in any other block. These interior blocks are separated by sets of junction stations which are coupled by observations to stations in more than one block. An example of such a partitioning of the geodetic network in Fig. 1 to one level of Helmert blocking is provided in Fig. 2. Here the circled nodes represent the junction stations chosen for this example.

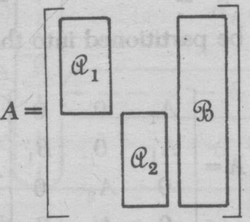
The particular form of Helmert blocking we will use here is the same as that used by Avila and Tomlin [4] for partitioning the normal equations. That procedure, in certain respects, is a variation of the nested dissection method developed by George [9, 10], George and Liu [12], and George, Poole, and Voigt [13]. The primary emphasis of the nested dissection strategy has been on solving symmetric positive-definite systems of linear equations associated with finite-element schemes for partial differential equations. There, the finite-element nodes are ordered in such a way that the element matrix B is permuted into the block partitioned form

$$B=\left[\begin{array}{ccccc|c} B_1 & 0 & \cdots & 0 & \\ 0 & B_2 & \cdots & 0 & \\ \vdots & \vdots & \ddots & \vdots & & C \\ 0 & 0 & \cdots & B_k & \\ \hline & & C^t & & & D \end{array}\right],$$

where the diagonal blocks are square.

In our case we use the following dissection strategy in order to permute the observation matrix A into the partitioned form (1.4). Our procedure will be called *nested bisection*.

Given a geodetical position network on a geographical region $\mathbb{R}$, first pick a latitude so that approximately one-half of all the stations lie south of this latitude. This forms two blocks of interior stations and one block of separator or junction stations, and contributes one level of nested bisection (see Fig. 3). Now order the stations in $\mathbb{R}$ so that those in the interior regions $\mathcal{A}_1$ appear first, those in the interior region $\mathcal{A}_2$ appear second, and those in the junction region $\mathcal{B}$ appear last; order the observations (i.e., order the equations), so that those involving stations in $\mathcal{A}_1$ come first, followed by those involving stations in $\mathcal{A}_2$; then the observation matrix A can be assembled into the block-partitioned form



Thus A can be expressed in the block-partitioned form

$$A = \begin{bmatrix} A_1 & 0 & B_1 \\ 0 & A_2 & B_2 \end{bmatrix},$$

where the A_i contain nonzero components of equations corresponding to cordinates of the interior stations in $\mathcal{A}_i$ and where the B_i contain the nonzero components of equations corresponding to the coordinates of the stations in the junction region $\mathcal{B}$.

The referee has pointed out that a finer matrix partition is possible. In region $\mathcal{A}_1$ there are equations involving only stations in $\mathcal{A}_1$ and none in

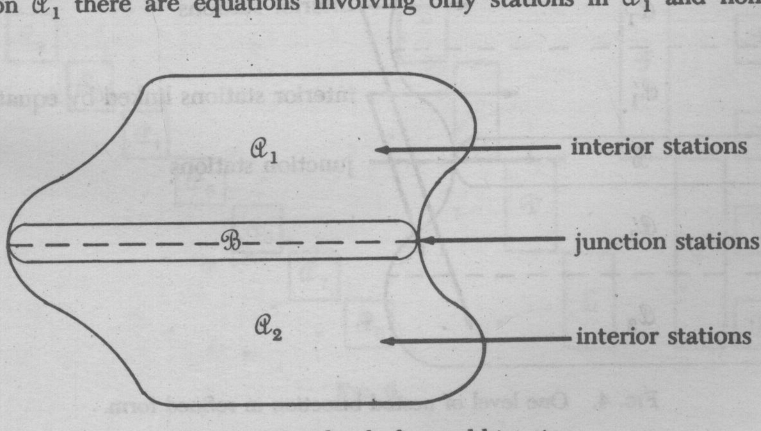
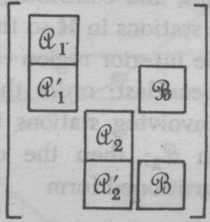


FIG. 3. One level of nested bisection.

region $\mathcal{B}$, and some involving stations in both $\mathcal{A}_1$ and $\mathcal{B}$. Similarly for $\mathcal{A}_2$ and $\mathcal{B}$. Figure 3 can then be refined to Fig. 4.

In this case, the matrix A can be further assembled into the block form

$A =$ 

and the matrix A can thus be partitioned into the finer block form

$$A = \begin{bmatrix} A_1 & 0 & 0 \\ A_1' & 0 & B_1 \\ 0 & A_2 & 0 \\ 0 & A_2' & B_2 \end{bmatrix}.$$

In the actual implementation of the algorithm to follow, this finer structure should probably be exploited. However, for simplicity we will ignore this finer partitioning in the discussion.

Next, in *each* of these halves we pick a longitude so that approximately one-half of the stations in that region lie to the east of that longitude. This constitutes level 2 of nested bisection. The process can then be continued by

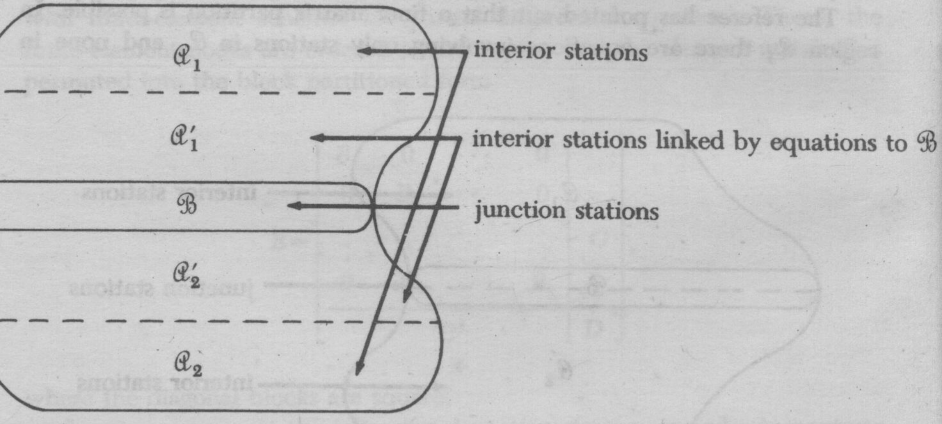


FIG. 4. One level of nested bisection in refined form.