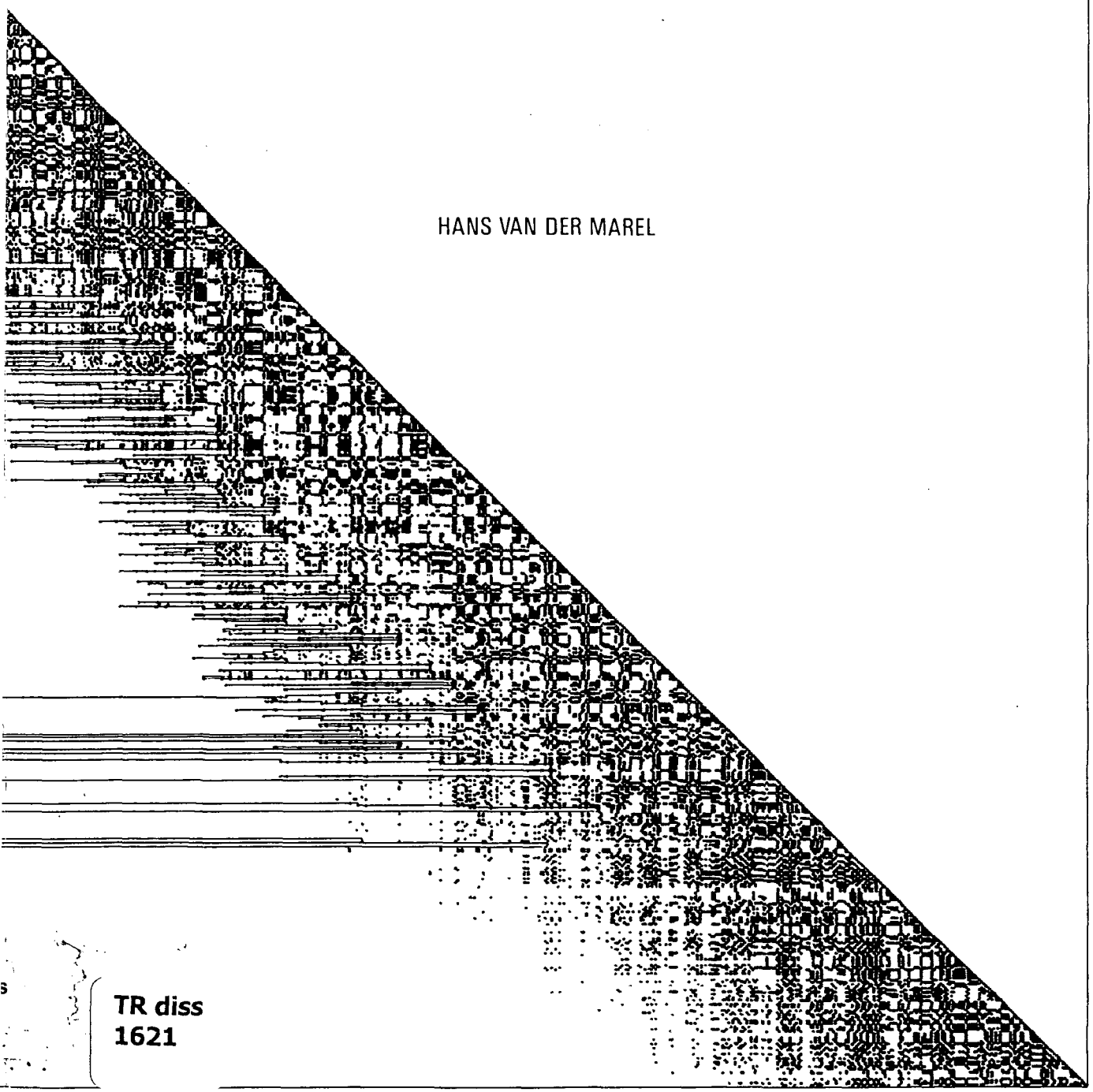
On the "Great circle reduction" in the data analysis for the astrometric satellite hipparcos

1 author:

Hans Van Der Marel
Delft University of Technology
85 PUBLICATIONS **553** CITATIONS

ON THE "GREAT CIRCLE REDUCTION" IN THE DATA ANALYSIS FOR THE ASTROMETRIC SATELLITE HIPPARCOS

HANS VAN DER MAREL



ON THE "GREAT CIRCLE REDUCTION" IN THE DATA ANALYSIS FOR THE ASTROMETRIC SATELLITE HIPPARCOS

PROEFSCHRIFT

TER VERKRIJGING VAN DE GRAAD VAN DOCTOR AAN DE TECHNISCHE UNIVERSITEIT DELFT, OP GEZAG VAN DE RECTOR MAGNIFICUS, PROF. DR. J.M. DIRKEN, IN HET OPENBAAR TE VERDEDIGEN TEN OVERSTAAN VAN EEN COMMISSIE DOOR HET COLLEGE VAN DEKANEN DAARTOE AANGEWEZEN, OP DINSDAG 29 MAART 1988, TE 14.00 UUR

DOOR

HANS VAN DER MAREL

GEODETISCH INGENIEUR
GEBOREN TE WASSENAAR

Dit proefschrift is goedgekeurd door de

promotoren prof.dr.-ing. R. Rummel en prof.dr. W.N. Brouw

On the cover:

The non-zero structure of the attitude part of the normal matrix, which occurs during attitude smoothing in the great circle reduction, after elimination of the star parameters and after the reordering of the attitude parameters (B-splines) modulo 60^0.

Lit.: This thesis, chapter 8.

Stellingen behorende bij het proefschrift:

On the "Great Circle Reduction" in the data analysis for the Astrometric Satellite Hipparcos

Promotoren: prof.dr.-ing. R. Rummel & prof.dr. W.N. Brouw

Hans van der Marel, 29 maart 1988.

---

1. Het is typerend voor de Delftse aanpak van de Hipparcos gegevensverwerking dat wij het stelsel vergelijkingen van de Reductie op Cirkels, althans voor een deel van de sterren, exact zijn gaan oplossen. Oorspronkelijk was dit niet de bedoeling, temeer daar het mathematische model zelf al een benadering is. Zo worden bijvoorbeeld de coëfficiënten van de ster- en standonbekenden in de gelineariseerde vergelijkingen exact uitgerekend, terwijl een benadering met +1 en -1 zou volstaan. Voor de exacte coëfficiënten geldt echter, tenminste zolang hun berekening niet veel extra rekentijd kost: baat het niet, het schaadt ook niet.

   Lit.: Dit proefschrift, sec. 5.2

2. De zogenaamde drie-stap-procedure die voor de Hipparcos gegevensverwerking gebruikt wordt is gebaseerd op een benaderde vereffening in fasen. De benaderingen zijn of relatief onschadelijk of worden ongedaan gemaakt door de gehele vereffening op een (blok) Gauss-Seidel achtige manier te itereren, waarbij echter geen rekening gehouden wordt met de correlatie tussen de sterabscissen op een Referentie Grote Cirkel. In verband hiermee is het aan te bevelen de door Sansò voorgestelde alternatieve methode verder te ontwikkelen.

   Lit.: Betti et al., A rigorous approach to attitude and sphere reconstitution in Hipparcos project. In: Proc. 3rd FAST Thinkshop, Bari, Nov. 1986

   Dit proefschrift, sec. 4.4

3. De "invloed van de attitude" op de variantie van de sterabscissen is een betere maat voor de sterkte van het netwerk van sterren tijdens de Reductie op Cirkels dan de rigidity factor die in verschillende MATRA en ESA studies gebruikt wordt.

   Lit.: MATRA, Hipparcos Overall System Technical Report, Chap. 3: Accuracy Analysis, E21-HIP-767, 1982

   Dit proefschrift, sec. 5.5

4. De zogenaamde "attitude smoothing" gedurende de Reductie op Cirkels behelst het gladstrijken van de standparameters in de scan-richting met behulp van bijvoorbeeld B-splines. Dit geeft tevens, afhankelijk van de stermagnitude en met uitzondering van magnitude 11 en 12 sterren, een 20-40% lagere rms fout in de sterabscissen en daarmee eenzelfde verbetering in de sterposities, eigenbewegingen en parallaxen.

Lit.: Dit proefschrift, sec. 6.7

5. Ondanks het feit dat de Hipparcos catalogus een kleine verbetering kan geven in astro-geodetische metingen op aarde zal zij toch nauwelijks echte toepassingen in de geodesie en geofysica hebben. Astronomische metingen voor geodetische en geofysische toepassingen zijn, t.g.v. de invloed van de atmosfeer, nu en in de naaste toekomst niet voldoende nauwkeurig om volledig van de Hipparcos catalogus te kunnen profiteren. Andere "ruimte" meettechnieken, zoals VLBI, laser afstandsmeting naar satellieten en GPS, zijn nauwkeuriger en/of geschikter.

Lit.: Dit proefschrift, sec. 2.3 & 2.5

6. De grote betrokkenheid van geodeten bij de voorbereiding van de wetenschappelijke gegevensverwerking voor de astrometrische satelliet Hipparcos onderstreept eens te meer, gezien de in de vorige stelling gesignaleerde geringe toepassingsmogelijkheden, het dienstverlenende karakter van de geodesie.

7. De meetgegevens van Hipparcos zullen op een zgn. mainframe-computer of een aantal mini-computers verwerkt gaan worden. De gegevens zouden echter in principe ook, na een aantal kleine aanpassingen in de programmatuur, op een (krachtige) personal computer met een daaraan gekoppelde, relatief goedkope, multiprocessor machine verwerkt kunnen worden. Het is te verwachten dat deze ontwikkeling zich zal voortzetten in tal van andere geodetische rekenprojecten.

8. Een aangepaste sequentiële vereffening, vergelijkbaar met de in de Reductie op Cirkels gebruikte methode voor het corrigeren van de zgn. gridstap fouten, is ook geschikt voor het berekenen van benaderde waarden in geodetische vereffeningsvraagstukken. Wanneer dan ook een aantal in de dagelijkse praktijk veel voorkomende grove fouten in aanmerking genomen worden ontstaat zo een simpel geodetisch expertsysteem.

9. Het schrijven van een manuscript achter een tekstverwerker, behoudens het overtikken van een reeds uitgewerkte tekst, is in twee opzichten verschillend van de traditionele "pen en papier" methode:
   1) slechts een beperkt aantal regels is tegelijkertijd zichtbaar,
   2) het aantal wijzigingen per teksteenheid is onbeperkt.
   Het tweede punt heeft een voor- en een nadeel; het voordeel is dat stukken tekst snel verbeterd kunnen worden. Het nadeel is echter dat stukken tekst welke aan aanzienlijke veranderingen onderhevig zijn nooit tussentijds "klaar" komen en zelden opnieuw opgezet worden, zoals dat anders gebeurd wanneer het papier vol is. Door het ontbreken van dit natuurlijke regelmechanisme concentreert men zich vaak in een te vroeg stadium op onnodige details en niet meer op het verhaal zelf. In verband hiermee is het aan te bevelen alle boekjes "Hoe schrijf ik een scriptie..." en dergelijke te herschrijven.

10. De invoering van de zogenaamde uniforme jaar indeling, die voor de studierichting geodesie neerkwam op een overgang van een semester systeem naar een kwartaal systeem met bijbehorende tentamenperioden, heeft eerder tot een versnippering van de aangeboden stof geleid dan de beoogde concentratie. Dit werd mede veroorzaakt door de toename van de totale lengte van de college perioden tesamen met de gelijktijdige overgang van een vijf- naar een vierjarige opleiding.

11. Gezien het lage salaris van assistenten en onderzoekers in opleiding, de problemen bij de werving van kapabele kandidaten en het algemeen maatschappelijk belang, is het aan te bevelen hen na voltooing van hun post-doctorale opleiding vrijstelling van militaire dienstplicht te verlenen.

Abstract

In this thesis several aspects of the scientific data reduction for the astronomical satellite Hipparcos are discussed. The Faculty of Geodesy of the Delft University of Technology participates in the data reduction in the framework of the international FAST consortium. Hipparcos (an acronym for **HI**gh Precision **PAR**allax **CO**llecting Satellite) is scheduled for launch in the spring of 1989 under supervision of the European Space Agency (ESA). During its operational life time of 2.5 years the satellite will scan the celestial sky in a slowly precessing motion and measure the angles between stars which are 60^0 apart. The observations will be done in the visible part of the electromagnetic spectrum. The Hipparcos data reduction aims at the construction of a precise star catalogue: The catalogue will contain the position, annual proper motion and annual parallax of about 110,000 stars, up to visual magnitude 12-13. The accuracy will be a few milliarcseconds and a few milliarcseconds per year respectively.

Besides a short introduction of the Hipparcos mission, the scientific objectives and the measurement principle, and a brief analysis of the data reduction as a whole, three topics are discussed in this thesis:
- model assumptions, estimability and accuracy of the great circle reduction,
- attitude smoothing, which improves the results of the great circle reduction,
- the numerical methods for the great circle reduction.

These subjects all concern one phase of the data reduction: the so-called great circle reduction. The great circle reduction comprises a half-daily least squares solution of some 80,000 observations with 2,000 unknown star abscissae and some 50 instrumental parameters. Depending on the solution method chosen, also some 18,000, or in case of attitude smoothing 600, attitude parameters have to be solved. The great circle reduction is a relatively modest adjustment problem in the complete data reduction, but one which must be solved several times per day over a period of several years.

The first four chapters are of an introductory nature. In chapter 2, which is more or less self contained, the scientific objectives and possible -geodetic- applications of the Hipparcos catalogue are sketched. In chapter 3 the Hipparcos measurement principle and raw data treatment are described and in chapter 4 a start is made with the description of the data reduction. It is in this chapter that the great circle reduction, the main subject of this thesis, is introduced and placed within the total data reduction.

The model assumptions, estimability and accuracy of the great circle reduction results are investigated in chapter 5. The great circle reduction processes only observations of stars within a small band (2^0) on the celestial sphere. Therefore, only one coordinate can be improved, *viz.* the abscissa on a reference great circle chosen somewhere in the middle of the band. The ordinates are not improved, *i.e.* they are fixed on their approximate values, which results in errors in the estimated star abscissae. By iterating the complete data reduction several times, in order to obtain better approximate values for the ordinates, the modelling error finally becomes very small and can be neglected. In chapter 5 analytical formulae for the magnitude of this error are derived. Further we investigate, one by one, the estimability of the instrumental parameters. They appear generally to be estimable. At the end of this chapter the covariance function of the star abscissae is computed for a regular star network using Fourier analysis. Throughout this chapter analytical results are compared with test computations on simulated data.

Chapter 6 is devoted to attitude smoothing. Smoothing of the attitude improves not only the quality of the attitude parameters, but also the quality of the star abscissae. We will consider in particular numerical smoothing with B-splines; the attitude is modelled by a series expansion using the above mentioned B-splines as base functions. The number of attitude parameters is reduced considerably; instead of the 18,000 geometric attitude parameters now only 600 are needed. But if the degree of smoothing is too high, systematic errors are introduced. The number of parameters have been chosen in such a way that the extra error introduced by smoothing is negligible.

Chapters 7 and 8 deal with the numerical methods for solving the sparse systems of equations which arise during the great circle reduction. Choleski factorization of the normal equations has been chosen as solution method. Optimization of the calculations is worthwhile, since such a system has to be solved several times per day. Computing time and memory requirements depend on the order in which the unknowns are eliminated. The best order appears to be: first the attitude unknowns, then the star unknowns and finally the instrumental unknowns. However, in the case of attitude smoothing it is better to eliminate the star unknowns first, and then the attitude and instrumental unknowns. Also the order in which the star parameters are eliminated, or in the case of attitude smoothing the attitude parameters, is important. Therefore, in chapter 8 several reordering procedures are evaluated. It turns out that the so-called banker's algorithm, which operates on the graph of the system, gives the best results in both cases. But also a synthetic ordering, which orders the star abscissae modulo 60^0, gives good results. The same algorithm, but then modulo 360^0, can be applied to the attitude unknowns for smoothing.

Finally, in chapter 9, methods are given for handling certain ambiguities in the data. Although the Hipparcos instrument is able to measure phases very accurately, the integer number of periods must follow from approximate data. This results in a large number of so-called grid step errors of about 1" (100 times the precision of measurement). These errors must be detected and corrected during the great circle reduction. Some strategies are discussed in chapter 9. The most successful strategy is based on an approximate sequential adjustment, which can be applied before and after the least squares adjustment.

In the appendices descriptions are given of the FAST great circle reduction software (appendix A) and of the simulated data used in simulation experiments with the great circle reduction software (appendix B). The results of these simulation experiments are used throughout this thesis for illustration. Finally, appendix C contains some background material on the numerical methods for solving large sparse systems of linear equations having a positive definite matrix.

Colophon

Illustrations:	M.G.G.J. Jutte (fig. 2.4, 4.1, 5.2) A.B. Smits (reproductions)
Print:	Meinema B.V., Delft

De Reductie op Cirkels in de gegevensverwerking voor de Astrometrische Satelliet Hipparcos

Samenvatting:

In dit proefschrift komen een aantal aspecten aan de orde van de wetenschappelijke gegevensverwerking ten behoeve van de astronomische satelliet Hipparcos, waar de faculteit der geodesie van de Technische Universiteit Delft, in het kader van het internationale FAST consortium, aan deel neemt. Hipparcos (**High Precision Parallax Collecting Satellite**) wordt naar verwachting in het voorjaar 1989 door de Europese Ruimtevaart Organisatie E.S.A. gelanceerd. De satelliet zal daarna gedurende 2.5 jaar in een langzaam roterende beweging de hemel aftasten en hierbij, door middel van een spiegel, hoeken meten tussen sterren die ongeveer 60^{o} van elkaar afstaan. De metingen vinden in het zichtbare licht plaats. Het doel van de gegevensverwerking is om uit de ruwe meetgegevens een stercatalogus te berekenen, die van ca. 110.000 sterren, tot aan magnitude 12-13, de positie, de jaarlijkse eigenbeweging en de parallax met een nauwkeurigheid van enige milliboogseconden (resp. milliboogseconden per jaar) zal bevatten.

Na een korte introductie van de Hipparcos missie, het wetenschappelijke belang en het meetprincipe, alsmede na een beknopte analyse van de gegevensverwerking in zijn geheel, komen in dit proefschrift een drietal onderwerpen aan de orde:

(1) analyse van de precisie, schatbaarheid en modelfouten van de zgn. reductie op cirkels,

(2) methoden voor het "glad maken" van de standgegevens van de satelliet (de zgn. "attitude smoothing"), die tot doel heeft de uitkomsten van de bovengenoemde reductie op cirkels nog eens te verbeteren,

(3) numerieke methoden voor het oplossen van de grote stelsels -ijlevergelijkingen die tijdens de reductie op cirkels voorkomen.

Deze drie onderwerpen hebben alle betrekking op één fase van de gegevensverwerking: de zgn. reductie op cirkels (great circle reduction). De reductie op cirkels betreft een kleinste-kwadraten vereffening van zo'n 80.000 waarnemingen (een halve dag aan metingen) met 2.000 steronbekenden en 50 instrumentele onbekenden. Afhankelijk van de gekozen oplossingsmethodiek moeten ook nog eens zo'n 18.000 of, in het geval van "attitude smoothing", 600 standonbekenden opgelost worden. De reductie op cirkels is nog een relatief bescheiden vereffeningsprobleem in de gehele gegevensverwerking, maar wel één dat gedurende meerdere jaren enige malen per dag opgelost moet worden.

De hoofdstukken 1 t/m 4 hebben een inleidend karakter. In hoofdstuk 2, dat min of meer op zichzelf staat, worden de doelstellingen van de Hipparcos missie en de mogelijke astronomische en geodetische toepassingen van de Hipparcos stercatalogus besproken. In hoofdstuk 3 wordt het meetprincipe beschreven en in hoofdstuk 4 wordt een beschrijving en een beknopte analyse van de gegevensverwerking gegeven.

In hoofdstuk 5 worden precisie, schatbaarheid en modelfouten van de in de reductie op cirkels opgeloste onbekenden onderzocht. Gedurende de reductie op cirkels worden slechts waarnemingen verwerkt naar sterren gelegen in een smalle band ($\sim 2^{0}$) op de hemel. Daarom kan maar één coordinaat verbeterd worden, nl. de abscissen op een referentie cirkel -gekozen- ergens in het

midden van die band. De ordinaten krijgen geen correctie, waardoor foutjes in de berekende sterabscissen onstaan. Door de gehele gegevensverwerking een aantal malen te itereren, om verbeterde benaderde waarden voor de ordinaten te verkrijgen, wordt de uiteindelijke fout verwaarloosbaar klein. In hoofdstuk 5 worden analytische formules voor de grootte van deze fout afgeleid. Voorts is in hoofdstuk 5, stuk voor stuk, de schatbaarheid van de diverse instrumentele parameters onderzocht, welke, op een paar na, goed schatbaar blijken te zijn. Aan het eind van het hoofdstuk wordt voor regelmatige gevallen de covariantiefunctie van de sterabscissen afgeleid met behulp van Fourier methoden. Door het hele hoofdstuk heen worden de analytische resultaten vergeleken (en aangevuld) met proefberekeningen op gesimuleerde gegevens.

Hoofdstuk 6 is geheel gewijd aan het glad maken van de standgegevens, wat we in het vervolg "attitude smoothing" zullen noemen, en de daarmee gepaard gaande verbetering van de sterabscissen. De nadruk ligt in dit hoofdstuk op smoothing met behulp van zgn. B-splines, waarbij de stand van de satelliet gemodelleerd wordt door een tijdreeks met bovengenoemde B-splines als basisfuncties. Een geweldige reductie in het aantal standonbekenden is het gevolg: in plaats van de 18.000 die we eerst hadden blijken er nu maar ongeveer 600 nodig te zijn. Het gevolg is wel dat er extra modelfouten ontstaan, maar het aantal B-splines is zo gekozen dat deze fout verwaarloosbaar klein is.

In hoofdstuk 7 en 8 wordt ingegaan op de numerieke methoden die gebruikt worden voor het oplossen van de ijle stelsels vergelijkingen, zoals die bij de reductie op cirkels optreden. Als oplossingsmethode is Choleski-factorizatie van de normaal vergelijkingen gekozen. Optimalisatie van de berekeningen is van belang, daar een dergelijk stelsel een aantal malen per dag opgelost moet worden. Snelheid en geheugengebruik hangen o.a. van de volgorde waarin de onbekenden worden geëlimineerd. De beste volgorde blijkt te zijn: eerst de standonbekenden, dan de steronbekenden en vervolgens de instrumentele onbekenden. Echter, in het geval van attitude smoothing is het beter eerst de steronbekenden te elimineren, en vervolgens pas de stand- en instrumentele onbekenden. Ook de volgorde waarin de steronbekenden, en in het geval van smoothing de standonbekenden, berekend worden is belangrijk. Daartoe worden in hoofdstuk 8 verschillende ordenings procedures onderzocht. Het blijkt dat de zgn. "bankiers" algorithme, die op de graaf van het stelsel werkt, in beide gevallen het beste resultaat geeft. Maar ook een synthetische ordening, die de sterren op volgorde van hun abscissen modulo 60^0 zet, blijkt uitstekend te voldoen. Hetzelfde algorithme kan ook toegepast worden op de standonbekenden, maar dan blijkt dat het beter is modulo 360^0 te ordenen.

In het laatste hoofdstuk (9) komt een toetsings probleem aan de orde. Het Hipparcos meetinstrument voert namelijk wel heel nauwkeurige fasemetingen uit, maar het gehele aantal perioden moet uit benaderde waarden volgen. Een groot aantal fouten, van ongeveer 1 boogseconde (100 maal de meetprecisie), is het gevolg: de zgn. grid-stap fouten. De grid-stap fouten moeten o.a. tijdens de reductie op cirkels opgespoord en verbeterd worden. Daartoe worden in hoofdstuk 9 een aantal strategiën besproken. De meest succesvolle strategie blijkt te bestaan uit een aangepaste sequentiële vereffening, zowel voor als na de eigenlijke kleinste-kwadraten vereffening toe te passen.

In de appendices A en B wordt vervolgens een beschrijving gegeven van de software voor de reductie op cirkels en de gesimuleerde data, die gebruikt zijn in de diverse simulatie experimenten. Appendix C bevat achtergrond-informatie betreffende de numerieke methoden die gebruikt worden voor het oplossen van de kleinste-kwadraten problemen.

Curriculum Vitae

Hans van der Marel was born on the 2nd of August 1959 in Wassenaar, the Netherlands. He attended secondary school in Wassenaar, where he obtained the certificate Atheneum B. In September 1977 he started to study Geodesy at the Delft University of Technology, where he graduated in August 1983 (cum laude) under supervision of prof.dr.ir. W. Baarda, on the subject of the Astrometry satellite Hipparcos. His thesis received the 1983 University award for excellent graduate work. Part of his final work was done at CERGA (November '82 - February '83), under the direction of prof. J. Kovalevsky. Before his stay at CERGA Van der Marel worked (February '82 - August '82) for the "Rijkswaterstaat" on the precise positioning of a special submersible robot, which has been used during the construction of the 9 km long Oosterschelde storm surge barrier. Apart from that, he was a student assistant in the Photogrammetry section for one year, and held several positions within the department of Geodesy and the society of Geodesy students "Snellius", including the membership of the governing body of the department.

From September 1983 until July 1987 Van der Marel was employed by the Netherlands Organization for the Advancement of Pure Research (ZWO). During this period he worked at the department of Geodesy on the development of methods for the scientific data reduction of the astrometric satellite Hipparcos, the account of which forms the subject of this thesis. The work is done in close cooperation with other research groups in the framework of the scientific data reduction consortium FAST. Van der Marel is a member of the FAST software advisory group, and since January 1987 he is task leader for the great circle reduction in FAST and a member of ESA's Hipparcos Science Team.

Since December 1987 he is working on a research fellowship of the Netherlands Academy of Sciences at the Delft University of Technology.

Acknowledgements

The author wishes to express his thanks to the following organizations and institutes:
- The Netherlands Organization for the Advancement of Pure Research (ZWO) for their financial support in the form of a research fellowship,
- The Faculty of Geodesy of the Delft University of Technology for their kindness in providing all facilities needed and for their support in the form of a five month temporary position,
- Centro di Studi sui Sistemi (CSS), Torino, for their hospitality and support during his visit in April 1986, and
- Centre National des Recherches Scientifiques (CNRS), Paris, and Centre d'Etudes et Recherches Géodynamiques et Astronomiques (CERGA), Grasse, for their travel grants.

The author wishes to thank Diederik van Daalen in particular, who supervised the Hipparcos project at the Faculty of Geodesy until December 1986, for his inspiring enthousiasm and guidance. The author is much indebted to prof.dr.ir. W. Baarda, who was the author's promotor until he finally had to retire. The work done by Frank van den Heuvel, Johan Kok, Paul de Jonge, Ruud Verwaal, Peter Joosten, Luc Amoureus and Arjen Bax, who participated in the Hipparcos project at Delft, and all Hipparcos colleagues, is gratefully acknowledged.

Abbreviations

ABM	apogee booster motor
APE	astrometric parameter extraction
AR	attitude reconstitution
ARI	Astronomisches Rechen-Institut, Heidelberg (Germany)
BDL	Bureau des Longitudes, Paris (France)
CDS	Centre des Donnees Stellaires, Strassbourg (France)
CERGA	Centre d'Etudes et des Recherches Géodynamiques et Astronomiques, Grasse (France)
CNES	Centre National d'Etudes Spatiales, Toulouse (France)
CSS	Centro di Studi sui Sistemi, Torino (Italy)
ESA	European Space Agency
ESOC	European Space Operations Center, Darmstadt (Germany)
FAST	Fundamental Astronomy by Space Techniques consortium, one of the scientific consortia in charge of the Hipparcos data reduction
FOV	field of view
GCR	great circle reduction
GPS	global positioning system
HIPPARCOS	high precision parallax collecting satellite
IDT	image dissector tube
IFOV	instantaneous field of view
INCA	input catalogue consortium, scientific consortium in charge of the compilation of the "input" catalogue
JPL	Jet Propulsion Laboratory, Pasadena (USA)
MESH	industrial consortium responsible for building the satellite
NDAC	Northern Data Analysis Consortium, one of the scientific consortia in charge of the Hipparcos data reduction
RGC	reference great circle
SC	scan circle
SLR	satellite laser ranging
SM	star mapper
SR	sphere reconstitution
TDAC	Tycho Data Analysis Consortium, scientific consortium in charge of the data reduction of the complementary Tycho mission
TUD	Delft University of Technology, Delft (Netherlands)
TYCHO	name of the complementary mission flown on board of the Hipparcos satellite
VLBI	very long base-line interferometry

ON THE "GREAT CIRCLE REDUCTION" IN THE DATA ANALYSIS FOR THE ASTROMETRIC SATELLITE HIPPARCOS

7. NUMERICAL TECHNIQUES FOR THE GREAT CIRCLE REDUCTION

8. ORDERING OF THE UNKNOWNS DURING THE GREAT CIRCLE REDUCTION

9. GRID STEP AMBIGUITY HANDLING

APPENDICES:

A. DELFT GREAT CIRCLE REDUCTION SOFTWARE

B. SIMULATED DATA FOR THE GREAT CIRCLE REDUCTION

C. COMPUTER SOLUTION OF LEAST SQUARES PROBLEMS

CHAPTER 1

INTRODUCTION

In this chapter the Hipparcos astrometry satellite mission is introduced. The goals and the scientific, especially the geodetic, involvement are sketched.

1.1 The Hipparcos Mission

Hipparcos is the name of an astronomical satellite observing at visual wavelengths and being built by the European Space Agency (ESA). It is the first satellite devoted entirely to astrometry. The launch is scheduled for the spring of 1989 by the European Ariane 4 launcher from Kourou in French Guyana. The satellite will be stationed, during its operational lifetime of 2.5 years, in a geostationary orbit (36,000 km altitude). The Hipparcos mission aims at constructing two large and very precise stellar catalogues, the so-called *Hipparcos* and *Tycho* catalogue. The Tycho and Hipparcos catalogue form a drastic improvement of existing catalogues, both with respect to positional accuracy and with respect to the size of the catalogue.

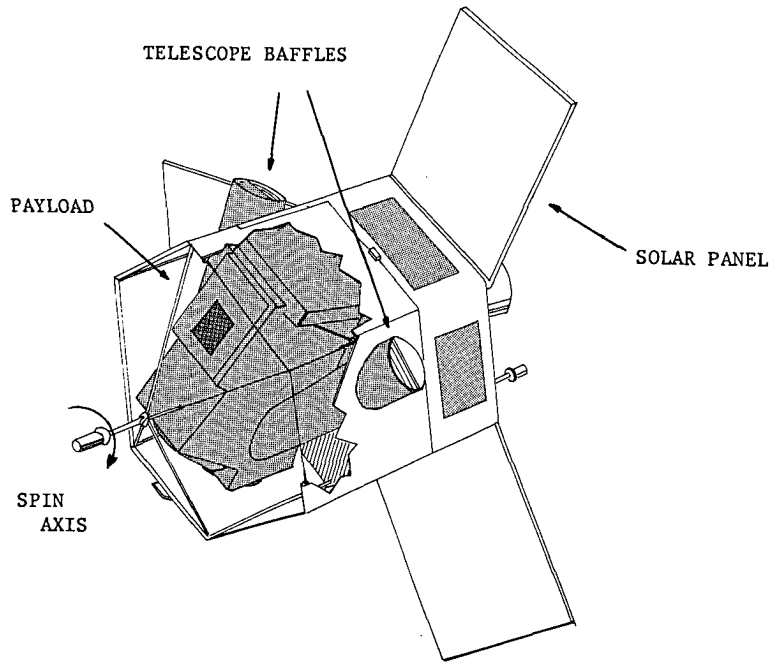


Figure 1.1 - The Hipparcos satellite

The primary aim of the Hipparcos mission is the construction of the Hipparcos catalogue, a precise star catalogue, containing the 5 astrometric parameters (position, proper motion and annual parallax) of some 110,000 stars up to visual magnitude 12-13. The accuracy of about 60,000-80,000 relatively bright stars, the so-called "survey", will be 1-2 mas (milliarc-seconds). The survey stars are evenly distributed over the celestial sphere, *i.e.* there are 1.5 à 2 stars per square degree. The accuracy of the, generally

fainter, 30,000-50,000 non-survey stars, chosen because of their individual interest within specific astronomical research proposals, is somewhat less than that of the survey, about 3-4 mas, depending on their magnitude. Due to the relative short duration of the mission, the precision of the proper motions and parallaxes will be of the same order of magnitude as those for the star positions, *viz.* 1-2 mas/year for each component of proper motion of survey stars and 1-2 mas for the parallax of survey stars. In order to obtain this precision 2.5 years of data is needed. When less than one year of data were available, the proper motions and parallaxes cannot be determined at all, but half a year of data is sufficient to compute the positions only.

The secondary aim of the mission is the construction of the Tycho catalogue, containing the positions, magnitudes and colours of some 400,000 to 1,000,000 stars. This catalogue is computed from the star mapper data. The star mapper is primarily used for the attitude reconstruction of the satellite, but reprocessing of its data with the attitude obtained from the main reduction will give positions with a typical accuracy of 30 mas.

Hipparcos is an acronym for **HI**gh **P**recision **PAR**allax **CO**llecting Satellite, but its name has also been chosen as a tribute to the ancient Greek astronomer *Hipparchus* (190-120 BC), who constructed one of the first known stellar catalogues and discovered astronomical precession by comparing his results with those of his predecessors. Hipparchus also determined the Moon's parallax, and thus its distance from Earth, something the Hipparcos satellite will do for the stars by measuring their annual parallaxes. The annual parallax of a star is the apparent angular displacement of its position as the Earth moves in its orbit around the Sun. It is a very small effect (< 1") which was discovered, long after it was predicted, by Thomas Henderson in 1832-33. Before that, in 1718, Edmund Halley discovered that some stars have proper motions by comparing his own measurements with those of Hipparchus.

For *astronomers* the parallaxes, together with proper motions, magnitudes and colours, form the main goal of the Hipparcos mission. These data are the raw material from which stellar luminosities, distances, masses, etc. are computed. At present, only a few thousand parallaxes, of nearby stars, are known. The results of the Hipparcos mission, compared to existing data, are impressive: 125 times more significant parallaxes at the 0"002 level, more precise and more consistent proper motions and the extension of the 4,000 FK5 stars into a consistent celestial reference frame consisting of more than 100,000 stars. A similar achievement from the ground is simply impossible, because it would not only require a breakthrough in instrumentation, notably for systematic errors (*e.g.* due to tube flexure and local refraction), but would also require a very extensive program of ground based observations, involving many observatories during several decades. With Hipparcos a global coverage of the sky will be obtained using a single instrument, which is impossible from the ground.

To *geodesy* and *geodynamics* the positions, together with proper motions, are of more interest. They form a formidable extraterrestrial reference system: precise, well materialized and free of systematic influences. However, the Hipparcos reference system cannot be accessed with the same precision by optical instruments from Earth. Therefore, for scientific applications (Earth rotation, polar motion, global deformations), optical astrometry cannot compete fully with "new" geodetic techniques like satellite laser ranging (SLR) and very long base-line interferometry (VLBI). Other geodetic techniques, like "Doppler" satellite positioning and the global positioning system (GPS), which have precisions comparable to good astrometric observations, are considered more practical and have replaced

astrometry already. Despite this, there may be two or three possible applications of the Hipparcos catalogue in geodesy and geodynamics, which will be discussed in chapter 2.

1.2 Scientific Involvement

ESA heavily relies on the scientific community in order to process the satellite data. Two scientific data reduction consortia, called NDAC and FAST, are both going to process the data from the main instrument in order to compute the Hipparcos catalogue. The two data analysis consortia are each going to produce a stellar catalogue, following slightly different procedures. Two parallel data reduction chains will increase the confidence in the final results. At the end of the mission, when it has been verified that the results agree sufficiently, the two catalogues will be merged, but already during the data analysis regular comparisons will be made.

The Hipparcos data reduction is an adjustment process, raising many interesting geometric and computational questions, fitting well in current geodetic research. Therefore geodesists from Copenhagen, Milano and Delft are participating in the data reduction consortia. The geodesists from Copenhagen participate in the Northern Data Analysis Consortium (NDAC), which includes scientific groups from Denmark, Sweden and the United Kingdom. The chief responsibility of the Copenhagen geodesists is the so-called great circle reduction, which comprises a half-daily solution of some 70,000 equations with 2,000 unknown abscissae on a chosen Reference Great Circle. The actual computations will be carried out in several places: at the Royal Greenwhich Observatory (raw data treatment), Copenhagen University Observatory (great circle reduction) and Lund Observatory, Sweden (final catalogue).

Geodesists from Delft and Milano participate in the FAST (Fundamental Astronomy by Space Techniques) consortium. FAST consists of research groups from France, Italy, Germany, the Netherlands and the United States. The faculty of Geodesy from the Delft University is responsible for the FAST great circle reduction, and has developed a large software package for this task. The Milano geodesists are more concerned with the next step of the data reduction, namely the construction of the final catalogue. The main body of computations for FAST will be done at CNES (Centre National d'Etudes Spatiales) in Toulouse, France, and at the Astronomisches Rechen Institut (ARI), Heidelberg, Germany. The Space Research Laboratory in Utrecht will, once a week, carry out a first check of the data.

The Hipparcos reference frame, by itself, has no reference to inertial space. The data reduction consortia, however, intend to establish a link between the Hipparcos catalogue and the VLBI and FK5 quasi-inertial reference frames. Therefore, the Hipparcos catalogue becomes a very dense and precise realization of the VLBI and FK5 quasi-inertial reference systems. Almost all FK5 stars are observed by Hipparcos, so few problems in linking the two systems are expected. The link to the extra galactic VLBI reference frame is realized through additional observations. The Jet Propulsion Laboratory (JPL, United States) has scheduled a number of VLBI observations to point-like radio stars, of which the optical component (hopefully coincident with the radio component) will be observed by Hipparcos. Other links to the extragalactic reference frame can be obtained through observations with the Hubble Space Telescope.

There are two other scientific consortia involved in the Hipparcos mission: the Tycho data reduction consortium (TDAC) and the input catalogue consortium (INCA). The input catalogue consortium is responsible for creating

a stellar catalogue which on its own is already of great value. The INCA catalogue contains the positions, proper motions, parallaxes (if known), magnitudes, colours (if known) and some other indices of the so-called program stars, the stars which are going to be observed during the Hipparcos mission. The program stars have been selected on the basis of proposals by the astronomical community. In order to get all the necessary data a large number of additional astrometric and photometric measurements (from Earth) have been carried out.

The Tycho data analysis consortium (TDAC) is going to reprocess the star mapper data, which is primarily used for the attitude determination of the satellite, to produce the Tycho catalogue with 400,000 - 1,000,000 stars, up to visual magnitude 10. The positional accuracy of this catalogue is expected to be of the order of 30 mas, but also very valuable photometric information (magnitude and colour) will be collected. There is no preplanned observing program for the Tycho experiment, but the data analysis task is greatly helped when there are reasonable a-priori positions. For this purpose the Strassbourg Stellar Data Base (CDS) and the Space Telescope Guidance Star Catalogue, a very dense catalogue constructed for the guiding system of the Hubble Space Telescope, will be used.

1.3 Guide to the Reader

The author's main research contribution to the Hipparcos data reduction concerns the great circle reduction and, more particularly,
- model assumptions and accuracy of the great circle reduction,
- attitude smoothing,
- numerical methods for the great circle reduction.

The great circle reduction comprises a half-daily solution of some 70,000 equations with 2,000 unknown star abscissae on a chosen Reference Great Circle. Attitude smoothing improves the results of the great circle reduction. These topics form the main body of this thesis, contained in chapters 5-9.

Chapters 2, 3 and 4 are of an introductory nature. In chapter 2, which is more or less self contained, the scientific objectives and possible -geodetic- applications of the Hipparcos catalogue are sketched. In chapter 3 the Hipparcos measurement principle and raw data treatment are described and in chapter 4 a start is made with the description of the data reduction. It is in this chapter that the great circle reduction, the main subject of this thesis, is introduced and placed within the context of the total data reduction.

The model assumptions and the accuracy of the great circle reduction are discussed in chapter 5. Chapter 6 is devoted to the attitude smoothing. Chapters 7 and 8 deal with the numerical methods used for the large scale least squares adjustment carried out during the great circle reduction. The ordering of the unknowns, which has a large influence on the efficiency of the great circle reduction, is treated in chapter 8. Finally, in chapter 9, methods are given for recovering from certain ambiguities in the data, the so-called grid step ambiguities.

In the appendices descriptions are given of the FAST great circle reduction software (appendix A) and of the simulated data used in simulation experiments with the great circle reduction software (appendix B). The results of these simulation experiments are used throughout this thesis to illustrate matters. Finally, appendix C contains some background material on the numerical methods for solving large sparse systems of positive definite equations.

CHAPTER 2

SCIENTIFIC OBJECTIVES OF THE HIPPARCOS CATALOGUE

In this chapter the scientific objectives of the Hipparcos mission are discussed. Some historical background is presented as an introduction. Special attention is given to the Hipparcos reference system, and its role in the unification and "inertialisation" of two existing celestial reference frames. In particular the proposed connection of the Hipparcos reference frame with the extra-galactic VLBI reference system is of interest. Finally a few possible geodynamical and geodetic applications are given.

2.1 Historical Background of Hipparcos

Astrometry, or positional astronomy, is the oldest branch of astronomy. Until the invention of the optical telescope, by 1609, all observations were done with the naked eye. Therefore, the upper bound for the positional accuracy used to be set by the resolution power of the naked eye, which is about one minute of arc. Two famous astronomers of this pre-telescopic era, Hipparcos and Tycho Brahe, need mentioning, since their names have been given to the two star catalogues which will be produced by the Hipparcos mission. The Greek astronomer Hipparchus (190-120 BC) already calculated the distance of the Moon from Earth by measuring the Moon's parallax. Hipparchus also made a star map, which led, when it was compared with the work of his predecessors, to the discovery of the precession of equinoxes. Seventeen centuries later, after Copernicus had introduced the heliocentric concept, Tycho Brahe, with the help of his brass azimuth quadrant, carried out a long series of observations during the second half of the sixteenth century. His observations, which had an accuracy better than 1', provided the basis for Keppler's laws of planetary motion.

After the invention of the optical telescope the angular error fell to several seconds of arc at the first half of the eighteenth century, and to better than one second of arc in the middle of the nineteenth century (figure 2.1). Some of the landmarks in astrometry, in chronological order, are the discovery of stellar aberration and nutation around 1700, of stellar proper motion by Halley in 1718, of the constant parallax of stars due to the motion of the Sun by Herschel in 1783 and finally the long expected discovery of the annual parallax by Henderson in 1831-1832 and Bessel in 1837-1838.

Another major step forward was the invention of the photographic camera at the end of the nineteenth century. The technique is to measure the position of the selected stars relative to a few reference stars surrounding it. This invention greatly economized the determination of proper motions and parallaxes, which are determined by measuring the shift in the star position from a large number of plates taken over a number of years. Several thousands of parallaxes have now been measured, although not always with a satisfactory accuracy. In this line of work some stars have to be used as a reference, and the positions, proper motion and parallax of these reference stars have to be known precisely. In this century several catalogues of reference stars have been compiled from meridian circle and astrolabe observations.

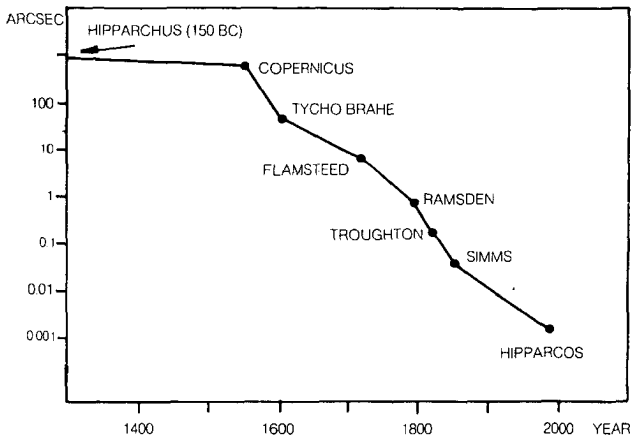


Figure 2.1 - The evolution of the error in astrometric measurements (courtesy of D. Hughes [*Perryman, 1985*])

Ground based measurements must be made through the atmosphere, so they are affected by atmospheric turbulence and refraction. Local atmospheric circumstances, mechanical deformations of the telescope under gravity and thermal effects, and seasonal variations give systematic errors in individual instruments. These systematic errors make it very difficult to establish a satisfactorily accurate reference catalogue covering the whole sky, with homogeneous errors in the astrometric parameters. Therefore, in 1966 a proposal for a space astrometry mission was submitted by prof. P. Lacroûte. Two major advantages of a space mission can be indicated: 1) the disturbing influence of the *atmosphere* is eliminated, 2) one *single* instrument will be able to cover the whole sky.

This preliminary proposal was soon followed by a series of more elaborate and ambitious proposals. Finally, a feasibility study by the European Space Agency (ESA) was initiated in 1977, the so-called phase A study [*ESA, 1979*]. This led to the adoption of the project by ESA in March 1980. The detailed design study (phase B) was completed in December 1983, after which the hardware phase began (phase C). The launch is scheduled for April 1989. In the meantime several scientific data analysis consortia have been set up. In 1981 the Fundamental Astronomy by Space Techniques (FAST) consortium was founded [*FAST, 1981*], of which the faculty of Geodesy of the Delft University of Technology became a member. It is one of the three scientific data reduction consortia which are going to process the Hipparcos data. The other consortia are the Northern Data Analysis Consortium (NDAC), and the Tycho Data Analysis Consortium (TDAC), responsible for the Tycho catalogue. A fourth consortium, the Input Catalogue Consortium (INCA), has just finished the task of compiling an input catalogue for Hipparcos, which contains a-priori data about the selected program stars.

2.2 Astrometry from Earth

2.2.1 Astrometric Techniques

Astrometry is concerned with the position, distance, motion, dimension and geometry of celestial bodies. The instantaneous location of a star in three dimensional space can be given in spherical coordinates, *e.g.* the

distance to the barycentre of our solar system and two angles, which give the position on a two dimensional manifold, a sphere of unit radius around the barycentre called the *celestial sphere*. The distance of nearby stars is computed from the -observed- parallax of stars. The *parallax* is the apparent displacement in position of celestial objects due to a change in the position of the observer. The parallactic displacement caused by the annual motion of the Earth around the Sun is called the *annual* or *trigonometric parallax*. The motion of celestial objects are also given in a radial component, the so-called *radial velocity*, and a component projected on the celestial sphere, the *proper motion*. The radial velocity is determined by measuring the Doppler shift of the stellar light. The proper motion is determined from two or more position measurements at different epochs (figure 2.2).

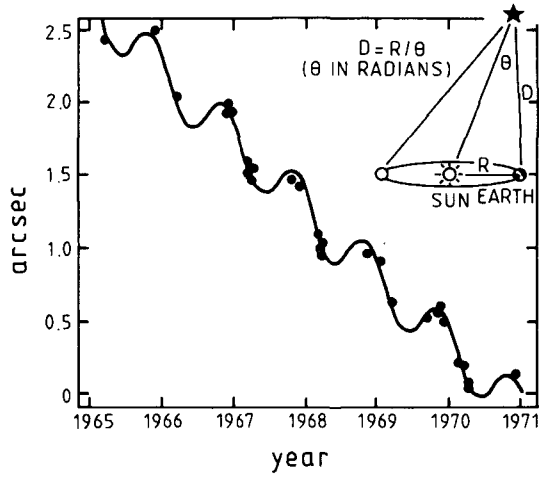


Figure 2.2 - The effect of parallax and proper motion on the observed star position (Courtesy of New Scientist [*Perryman, 1985*])

Astrometric techniques are classified according to the size of the field of view, ϑ, which is needed to obtain the desired result. Kovalevsky distinguishes 5 classes [*Kovalevsky, 1984*]: very narrow field ($\vartheta \leq 10''$), narrow field ($\vartheta \leq 0.^{\circ}5$), wide field ($\vartheta \leq 5^{\circ}$), semi global (a part of a hemisphere) and global astrometry.

Very narrow field astrometry ($\vartheta \leq 10''$) is devoted to the study of multiple stars without reference to nearby stellar surroundings. The main instrumental tools are long focus telescopes. In combination with a technique known as speckle interferometry, which eliminates the effect of atmospheric turbulence (seeing), the resolving power is reduced to the theoretical diffraction limit of the telescope. The errors are now in the order of 0"005 and 0"002. *Narrow field astrometry* ($\vartheta \leq 0.^{\circ}5$) is used when it is necessary to link the position of a star to neighbouring ones. This is the case for parallax and proper motion determination relative to a set of reference stars with known parallaxes and or proper motions. The derived parallaxes and proper motions are obviously very sensitive to systematic errors. The best results with this method are of the order of 0"004 for parallaxes (using many photographic plates over a period of several years), but currently less than thousand parallaxes are known with this precision and this number increases by not more than fifty per year.

Wide field astrometry ($\vartheta \leq 5^{\circ}$) is essentially relative astrometry using photographic plates taken by astrographs and Schmidt telescopes. This technique aims at determining star positions with respect to some reference

stars whose celestial coordinates are known. The actual measurement error is of the order of 0"1 for most modern equipment, but the computed positions are seriously affected by uncertainties in the global positions of the reference stars, which may range between 0"4 and 1" (see table 2.1).

Semi global and *global astrometry* are concerned with the determination of positions of stars far apart from each other. The best available instruments are -automatic- meridian circles and astrolabes, which have a typical precision of 0"2 in a single observation of stars brighter than magnitude 11. With the present automatic instruments stellar coordinates to better than 0"1 are produced operationally, and with a production rate significantly better (up to 20,000 observations per year) than the classical visual instruments. However, connections can only be established within a certain portion of the sky, and, therefore, catalogues obtained by a single instrument cover only part of the sky. The major problem is formed by systematic errors, caused by the telescope (tube flexure), site (local refraction) and by the change of seasons, because some stars are only observable in winter, some in summer. Therefore it is not easy to combine the individual catalogues, which is the aim, and method, of *global astrometry*. Hipparcos is a global astrometry mission, but the quality of its results is comparable to those of (very) narrow field astrometry from ground based observations.

2.2.2 Global Astrometry

The objective of global astrometry is to establish a single consistent reference frame, possibly non-rotating (inertial), materialized in many star positions and proper motions, and with regional errors reduced to a minimum. An inertial frame can essentially only be constructed from the analysis of the motion of celestial bodies (Moon and planets), under the assumption of a dynamical model of their motion in inertial space (*i.e.* not containing any inertial rotational term). The choice for a specific dynamic model defines the reference system. Another possibility to define a non-rotational reference system is to assume that some distant objects (galaxies, quasars) have no detectable apparent motion. Once a reference system is defined, it must be materialized, *i.e.* coordinates, and possibly motions, must be assigned to a sufficiently dense network of celestial bodies. Such a materialization is called a reference frame or Fundamental Catalogue: the set of coordinates associated with a reference system.

Presently a new fundamental catalogue is coming into use, the FK5. The FK5 contains as many as 4500 stars (magnitude $V < 9$) with random errors of the order of 0"03 in position and 0"002 per year in proper motion [*Fricke, 1980*]. This is a considerable improvement compared to the FK4, which had at epoch 1980 random errors of ~0"12 and regional errors of up to 0"2 especially in the Southern hemisphere. The FK5 is constructed from the old FK4 data and 150 new catalogues, each based on observations by one single instrument. In the FK5 also a new dynamic model was used, based on new data and revised astronomical constants.

The FK4 and FK5 do not have sufficient stars to be used as reference for wide field photographic astrometry. For the reduction of a plate of 2 x 2 degrees about 15 reference stars are needed, *i.e.* 4 stars per square degree, whereas the FK4 or FK5 contains only one star per 9 square degrees. Therefore the FK4 has been extended by, mainly, meridian observations, leading to a system of International Reference Stars (IRS). The IRS stars are given in the AGK3R and SRS catalogues (respectively in the Northern hemisphere, observed around 1959, and Southern hemisphere, observed around 1968) which contain together 38,000 stars. Other, but less homogeneous, catalogues are the

photographic AGK3 catalogue (Northern hemisphere, observed between 1930 and 1960) with 180,000 stars and the SAO catalogue with 500,000 stars. After applying systematic corrections FK5 - FK4 the IRS stars have positions known to about 0"3 at the present epoch. The AGK3 stars have mean errors of the order of 0"4 for bright stars, and above 0"5 for the faint stars, while the positions of stars in the SAO catalogue have often errors of more than 1".

Table 2.1 - Typical accuracy of existing vs. Hipparcos catalogues. (fundamental: FK4, FK5; reference: AGK3R, SRS; photographic: AGK3)

catalogue	typical no. of stars	rms position error (1990)	rms proper motion error
fundamental	5,000	30 mas	2 mas/year
reference	38,000	300 mas	4 mas/year
photographic	180,000	>500 mas	10 mas/year
Hipparcos	110,000	2 mas	2 mas/year
Tycho	1,000,000	30 mas	--

The Jet Propulsion Laboratory VLBI (Very Long Base-line Interferometry) reference frame is the best available quasi-inertial frame at the moment. It is composed of more than 100 sources, quasars, with mean errors of 0"005 in their positions. However, the extension of this system to stellar positions is quite difficult. Generally the optical counter parts of the quasars are very faint, and are not directly accessible to semi global astrometry, but a link can be established through large field photographic plates. The errors in these links are of the order of 0"1. The precision of the links could be increased by narrow field astrometry, but then the problem is to find a bright enough star (preferably FK5) close by. So, presently, it does not seem that this link is very significant. More important is the link of the VLBI reference system with the future Hipparcos catalogue, which will also contain the FK5 stars.

2.2.3 Limitations of Earth based Observations

The accuracy of astrometric observations from the surface of the Earth is degraded by atmospheric, gravitational and geodynamical effects. These effects are absent in measurements from space. The atmospheric influences are the most fundamental. Firstly, seeing (turbulence) and refraction cause random errors in the observations. Secondly, refraction, due to site and seasonal variations, gives a significant systematic error which is difficult to detect and which averages out only very slowly with more measurements.

The *atmospheric refraction* is caused by the spherical atmospheric layers. It is a large effect which increases with the zenith distance. The normal part can be modelled as a function of the zenith distance and a few other parameters [*Tengstrom and Teleki, 1978*]. The auxiliary parameters are either measured locally, *e.g.* temperature, air pressure and humidity, follow from the regional weather situation or are determined from the measurements itself. The refraction depends also on the star colour. This, in principle, could be used to eliminate the parallax by measuring in two different

colours, but the effect is quite small, and therefore it is difficult to get usable results.

The anomalous refraction error is still largely systematic, and can reach hundreds of mas, although correction to 10-20 mas may be possible [*Sugawa and Naito, 1982*]. The systematic part of the anomalous refraction error depends on the site, time and season of observation. Therefore, it averages out only slowly. Høg [*Kovalevsky, 1984*] found from empirical data that the error decreases as $T^{-1/4}$, with averaging over increasing observation time T. But, in particular, stars which can be observed only during a certain part of the night or in one of the seasons, can have large systematic errors in their positions which average out even slower.

Seeing is caused by turbulence in the atmosphere. It results in intensity variations (scintillation) and in short periodic ray bending, both spatially and temporally [*Tengstrom & Teleki, 1978*]. The turbulent cells are typically 10-30 cm in size. The so-called atmospheric coherence time, the period during which a certain optical situation remains stable, is not long (typically 0.01-0.5 s.), since the turbulent cells move with the winds through the light path. In telescopes with an aperture smaller than the width of the cells the turbulence results in image motion. In telescopes with a larger aperture the various atmospheric cells through which the light passes, give different images: the speckles. These speckles are randomly moving in the field. For observing times larger than the atmospheric coherence time this results in blurred images. The size of this effect can be large, several seconds of arc, and down to slightly less than one second of arc in good nights. The photocentre cannot be determined to better than 5%-10% of the blurred image, resulting in an error of a few hundred mas, maybe 50 mas at the best. Fortunately this error averages out faster than the refraction error. Theoretical work by Lindegren [*Kovalevsky, 1984*] showed that the error in a measured angle ϑ between two stars near the zenith decreases as $\vartheta^{1/4}T^{-1/2}$, with averaging over increasing observation time T.

The pull of the Earth's gravity affects the stability of the instrument (*e.g.* tube flexure) and this gives small systematic errors. Also geodynamical effects, by which we mean the anomalous part of Earth rotation, polar motion, tides and Earth crust deformations, introduce errors. So far, we assumed that diffraction, photon noise and detector noise are not significant. This holds only for good instruments, *viz.* the diffraction limited image varies from 20 mas for large telescopes and large zenith tubes to 2" for ordinary geodetic instrumentation.

These limitations work in two directions. Firstly, the site and seasonal effects on the refraction are the limiting factor in global astrometry from Earth. These effects make it almost impossible to establish a satisfactory reference frame, with homogeneous errors in position and proper motions. Therefore, the only certain way to get away from these limitations is to go into space. Secondly, for geodynamical applications (*e.g.* Earth rotation parameters) any homogeneous reference system, such as the Hipparcos one, has to be accessed from Earth by astrometrical observations. Again, the atmosphere is, and stays, the limiting factor.

2.3 The Scientific Objectives of the Mission

Hipparcos is essentially a global astrometry mission, but with an accuracy comparable to (very) narrow field astrometry. It is the first satellite mainly devoted to global optical astronomy in the visual wavelengths. Two precise stellar catalogues, containing the positions, proper

motions, parallaxes, magnitudes and colours of stars, will be constructed: the so-called Hipparcos and Tycho catalogues. The applications of the Hipparcos and Tycho catalogues in astronomy will discussed briefly. More applications are discussed in the proceedings of several colloquiums on the scientific aspects of the Hipparcos astrometry mission [*Barbiery and Bernacca, 1979, Perryman and Guyenne, 1982, Guyenne and Hunt, 1985*]

2.3.1 The Hipparcos and Tycho Catalogues

The *Hipparcos catalogue* is the primary aim of the mission. It will be computed from the main grid data and it will contain the positions, proper motions, parallaxes and magnitudes of some 110,000 stars up to magnitude 12-13. The accuracy of about 60,000-80,000 relatively bright stars, the so-called *survey*, will be 1.5 to 2 mas for each component of the position, as well as for the yearly proper motions and parallax. Their systematic -regional- errors will be not more than a fraction of a mas. The survey stars are evenly distributed over the celestial sphere, *i.e.* 1.5 a 2 per square degree. A large fraction of the magnitude nine stars, and almost all stars brighter than magnitude eight, will be survey stars. The accuracy of the generally fainter non-survey stars is about 3 to 4 mas (depending on their magnitude) for positions, yearly proper motion and parallax. These fainter stars are chosen because of their astronomical or astrophysical interest. More than 200 research projects have been submitted to ESA, requesting the observation of much more stars than can be observed by Hipparcos.

A secondary aim of the mission is the *Tycho catalogue*. The Tycho catalogue will contain the positions, magnitudes and colours of some 400,000 to 1,000,000 stars. This catalogue is computed from the star mapper data. The star mapper is primarily used for the attitude reconstruction of the satellite, but reprocessing of this data with the attitude obtained from the main reduction will give positions with a typical accuracy of 30 mas.

Furthermore, a substantial fraction of the program stars are double or even multiple. It is possible to compute some of the orbital parameters and magnitudes of double star components from the Hipparcos data, which is another aim of the mission. Also a number of minor planets (asteroids) is included in the observing program for solar system reference frame purposes.

2.3.2 Global Astrometry with Hipparcos

The Hipparcos mission offers great advantages over classical global astrometry; the major advantage is that a single instrument, outside the disturbing influence of the atmosphere and able to observe large angles ($\sim 60^\circ$), is used for the complete sky. Therefore, regional errors in the final catalogues are believed to be absent, which is of great benefit to statistical kinematic studies of our galaxy. In addition, the Hipparcos and Tycho catalogues are dense enough to be used directly in wide field photometric astrometry. The Tycho catalogue, which will contain 10-20 stars per square degree, could even be used in narrow field astrometry.

The Tycho and Hipparcos catalogues drastically improve the positional accuracy of existing catalogues. But in order to preserve the catalogue precision throughout time precise proper motions are needed. The rms error in the position at an epoch different from the central epoch is (law of error propagation)

$$\sigma_{p(t)} = \sqrt{\sigma^2_{p(t_0)} + \sigma^2_{\mu} (t-t_0)^2}$$

with $\sigma_{p(t)}$ the rms error at an epoch t, σ_{μ} the rms error in the proper motion and t_o the central epoch of the observations to a certain star (at the central epoch of observation the star position is not correlated with the proper motion). The rms error as a function of the epoch is given in figure 2.3
for different catalogues. The Hipparcos data alone does not bring a similar improvement to the rms error of the proper motions as it does to the positions, although a combination of the Hipparcos data with existing data, or better, with a second Hipparcos in ten years time, will give an additional improvement to the proper motions and hence, to the future quality of the catalogue (Figure 2.3). However, the systematic error (not given in figure 2.3) in the Hipparcos proper motions will be much smaller than in existing catalogues, and this is just what makes the Hipparcos catalogue so worthwhile. Also the Tycho proper motions can be improved, even down to the Hipparcos accuracy, by combining the Tycho catalogue with existing catalogues.

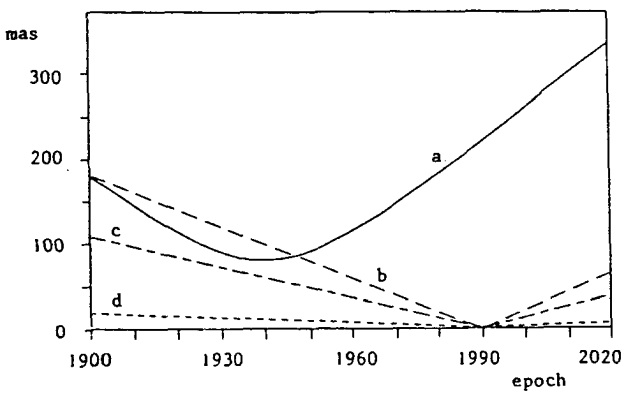


Figure 2.3 - The accuracy evolution (rms) of various catalogues: a) IRS (38,000 stars), b) Hipparcos (110,000 stars), c) IRS + Hipparcos and d) Hipparcos plus a second Hipparcos in ten years.

An important application of the precise Hipparcos star positions, at other epochs than the central epoch, is a new reduction of existing photographic plates, which exist from the beginning of this century. A new reduction of this old material may be useful for the determination of proper motions of fainter stars (down to magnitude 13-14) for the purpose of studies on galactic kinematics [*De Vegt, 1982*]. Similarly, a new reduction of the old latitude determinations could give an important improvement in the Earth rotation parameters from 1900 onward. For most applications it is necessary that the proper motions are given in an inertial frame. Therefore, a link between the Hipparcos system and the FK5 and VLBI reference systems is foreseen.

2.3.3 Astrophysical Applications

The catalogue will not only be used as a reference for other astrometric work, but the proper motion and parallax of the 110,000 Hipparcos stars will also be used directly for astrophysical work. For astronomy the parallaxes and the proper motions of stars, together with their magnitudes and colours, form the main content of the mission. The accuracy of Hipparcos' proper motions and parallaxes is comparable to the internal accuracy of (very) narrow field astrometry. The expected rms errors for the approximately 80,000

stars brighter than magnitude 9.5 are:
- 1.5 to 2 mas/year for each component of the proper motion,
- 1.5 to 2 mas for the parallax.
The systematic errors are expected to be a fraction of a mas. For magnitude 12 stars the r.m.s. error is of the order of 4 mas/year for the proper motion and 4 mas for the parallax.

The annual parallaxes are the basis of all distance measurements in the universe. Actually, the distance of only a small fraction of the stars (0.1% of our galaxy) can be determined from the parallax directly, but all other methods are, in one or more steps, calibrated on the basis of these parallaxes. Thus, any improvement in the parallax situation leads to better distances and an improved accuracy of the cosmic distance scale. From the distance, and the magnitude, which is also determined by the Hipparcos mission, the actual luminosity of stars can be computed. This will lead to an improved calibration of the Hertzsprung-Russell diagram, which gives the relation between the luminosity and colour of stars[1]. Distances are also needed to compute the masses of the components of a double star from the orbital parameters. At present, only 500 stars have parallaxes measured with a precision better than 2 mas, but, according to Hanson [*Kovalevsky, 1986*], some may also have systematic errors of a few mas. In the Hipparcos catalogue one hundred times more parallaxes, with a precision to better than 2 mas and without significant systematic errors, will be given.

In numbers, the situation with proper motions is better than for parallaxes. Many stars have proper motions known with an internal precision (*i.e.* relative to their neighbours) of the order of several mas/year, determined from observations over more than 50 years. The systematic -regional- errors, however, are in the order of 5 to 15 mas/year, except for the 5000 FK4 and FK5 stars, which have errors of a few mas/year. The proper motions play an important role in kinematic studies of our galaxy.

2.4 Link to the FK5 and VLBI Inertial Reference Systems

The celestial reference system defined by the Hipparcos and Tycho catalogues is precise, reasonably stable (which depends on the quality of the proper motions), well materialized and conceptually simple. Unfortunately, this reference system[2] has no direct reference to inertial space, not by theory and not by direct observation. Eventually a link between the Hipparcos system and the two presently available inertial systems, based on solar system dynamics and on the positions of extragalactic objects respectively, will be made. So for the first time the two principally different inertial systems now available will be compared.

The Hipparcos and Tycho catalogues are computed from angular measurements. This causes an indeterminacy of the system: the angles are invariant under a rotation of the coordinates for positions and proper motions, which results in a rank defect of 6 during the data reduction. The rank defect is solved by imposing some additional constraints. There is a certain arbitrariness in the choice of these constraints, and consequently in

[1] It turns out that there is a strict relation between the star type, *i.e.* the stage of its evolution, and its position in the Hertzsprung-Russell diagram.

[2] Since the computation of the Tycho catalogue is based on the attitude data computed in the Hipparcos system it is safe to assume that both catalogues refer to the same reference system, the so-called Hipparcos reference system.

the definition of the reference system, as is often stressed by Baarda [*Baarda, 1973*]. In the data reduction the rank defect is solved by fixing the position and proper motion of one and a half star to zero, although the rank defect disappears during the data reduction because some additional information is supplied through the approximate values for the star positions (see chapter 4). For practical reasons (*e.g.* for studies of galactic kinematics) the Hipparcos reference system should be close to an inertial reference system. Therefore, the coordinates of proper motion should be rotated, to form a quasi-inertial system (for the instantaneous positions it is often just desirable that the zero parallel of latitude is close to the ecliptic or equator). The operators that transform different coordinate representations of the same, generally geometric, quantities into each other are known in geodesy as *S-transformations* (see appendix C).

The link to solar system dynamics is made indirectly, through the existing FK5 star catalogue, and directly by Hipparcos observations of some asteroids. About 20-30 asteroids of magnitudes 8 to 12 will be observed by Hipparcos. These observations will give some, but no decisive, information on the orientation of the Hipparcos reference frame. More is expected of the link to the FK5 system [*Röser, 1983*], while the comparison of the FK5 with Hipparcos will show the regional errors of the FK5. The FK5 is currently the best approximation of a truly inertial, dynamically defined system, based on many observations of solar system objects (see section 2.2.2). The orientation of the FK5 system is accurate to some 1.5 mas/year [*Kovalevsky, 1984*], the same as for Hipparcos.

The link with the VLBI reference system is made indirectly, because the optical counterparts, except one, of the distant radio sources are too faint to be observed by Hipparcos directly. Instead, two types of indirect connections are foreseen: a space and a ground tie. The first scheme uses optical ties by NASA/ESA's Hubble Space Telescope between quasars and their close optical neighbours seen by Hipparcos [*Froeschlé & Kovalevsky, 1982*]. In the second, and most promising, scheme the ties will be made by VLBI observations between quasars and point-like radio-optical stars within the Hipparcos program [*Preston et al., 1983, Lestrade et al., 1985*]. It is expected that about 20-30 of these point-like radio-optical stars will be in the measurement program. In both schemes the links are expected to give the rotation to better than 1 mas/year.

Finally, all the presented methods combined are believed to give a quasi-inertial system down to a level below 0.5 mas/year. So, for the first time, the two existing quasi-inertial systems, based on solar system dynamics and extra-galactic VLBI sources respectively, are to be compared. Thus, the Hipparcos system plays an important role in the unification of the celestial reference systems.

2.5 Geodynamical Applications of the Hipparcos Reference Frame

The connection between the Hipparcos reference system and the terrestrial reference system cannot be established directly by Hipparcos, but should be established by optical astrometry from Earth[3]. The accuracy of optical astrometric measurements from the Earth is dominated by catalogue and observational errors. The catalogue error will decrease to a few mas once the Hipparcos catalogue becomes available. This is well below the present observational errors, due to atmospheric seeing, atmospheric refraction and

[3] Although an interesting proposal for observing laser beacons on Earth by Hipparcos was done in [*Bertotti et al., 1983*].

instrumental imperfections. For geodetic observations, using portable equipment like zenith cameras and theodolites, the observational error is in the order of 300-700 mas. For fundamental observations, using for instance meridian circles, zenith tubes and astrolabes, an observational error of 70-100 mas is attainable at present, and in future an error of 10-20 mas may even be reached. A survey of possible geodetic and geodynamical applications is given in [*Groten, 1982*], [*Van Daalen, 1985b*] and [*Van Daalen & Van der Marel, 1986a*].

Typical geodetic applications, like local or regional geoid determination, orientation of 2-D networks and vertical directions in 3-D networks, require for their astronomical measurements a relative precision of 10^{-6} (700 mas), which is of the same order of magnitude as the present catalogue errors (see table 2.1 and figure 2.3). Therefore geodetic observations may profit somewhat from the Hipparcos catalogue, especially because of the absence of systematic errors. But the current decrease in the use of optical methods in geodesy cannot be stopped, other methods are more practical. *I.e.* satellite Doppler measurements and the global positioning system are more practical for positioning, and gravimetric measurements are more practical for geoid determination. However, astronomical geodesy might remain attractive for the determination of vertical directions in 3-D networks, especially in mountainous areas, using transportable zenith cameras [*Bürki et al., 1983*].

The obvious disadvantages of optical astrometry are its dependence on a clear sky and nighttime, the time consuming observations, requiring experienced observers (somewhat less for zenith cameras), and the vulnerable and expensive instruments. But there are advantages as well: optical astronomy requires not a high organization level and no satellite orbits and the model is, except for the atmosphere, well established and simple. Sometimes it is advocated that astronomical geodesy might remain attractive for developing countries [*Birardi, 1982*].

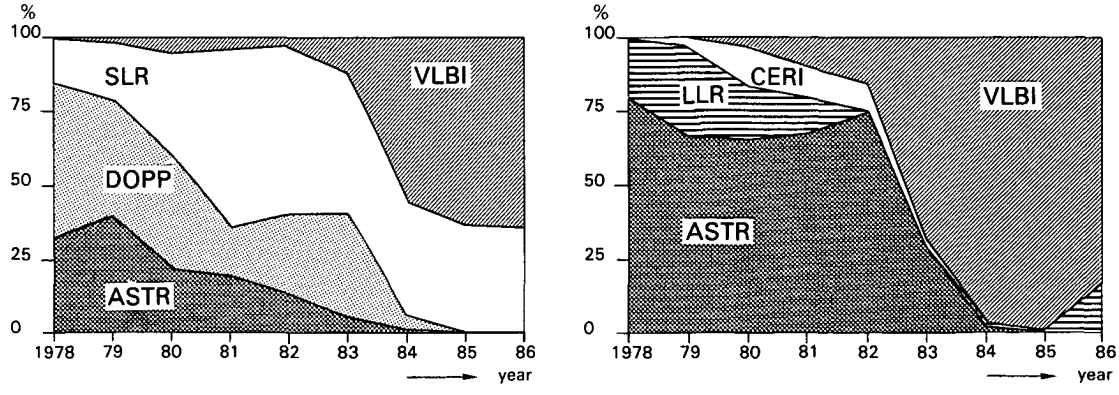


Figure 2.4 - Relative weights of the different observing techniques for the 5 day BIH values of polar motion(left) and Universal Time (right). VLBI: Very Long Baseline Interferometry, CERI: Connected Element Radio Interferometry, SLR: Satellite Laser Ranging, LLR: Lunar Laser Ranging, ASTR: Astrometry, DOPP: Doppler Satellite Solution.

For scientific applications, like the determination of Earth rotation and polar motion, tectonic motions and satellite orbits, the best possible accuracy is desired. At present astronomical geodesy is being replaced by

other techniques because these are more accurate. This tendency is clearly illustrated by the relative weights of the different observing techniques for the 5 day values of polar motion and Universal Time in figure 2.4 [*BIH, 1981-1986*]. With laser ranging and VLBI an accuracy of a few centimeters (3 cm ≃ 1 mas) can be achieved. This is well below even our optimistic estimate for the optical error in astrometry. Therefore, this type of applications will probably hardly profit from the Hipparcos reference system.

Unlike the new techniques, astrometry has a long record of good observations, even back to 1900. Laser and VLBI only started to give useful results 10 years ago. Since the Hipparcos mission will improve stellar catalogues for several decades backwards, recomputation of historical astronomical data may give valuable results on Earth rotation theory and tectonic motions. Satellite orbit determination by direction measurements to satellites, for the purpose of gravity field studies, might profit somewhat from Hipparcos, although even our optimistic estimate of the optical error corresponds to ~1 m at Lageos height. However, according to Smith and Marsh [*Smith and Marsh, 1986*] old camera data still provide valuable information on the zonal coefficients in the spherical harmonics development of the gravity field. Recomputation of this old data may give small improvements, but a more up to date and accurate direction measurement system would be more interesting. Finally, the Hipparcos catalogue, which is practically errorless, will allow studies of the other error sources such as refraction.

The conclusion is that, although the Hipparcos catalogue can give some improvement in geodetic and geodynamical applications, the accuracy of astronomical measurements is not sufficient, not now or in the near future, to be able to have a large impact in geodesy and geodynamics.

CHAPTER 3

HIPPARCOS MEASUREMENT PRINCIPLE

In this chapter the optical configuration of the satellite, the scanning motion during its 2.5 years of mission, the measurements and their preprocessing are described.

3.1 A primer on Hipparcos

In order to reach the goal of the Hipparcos mission, *i.e.* the determination of the astrometric parameters of stars on the entire sky with *uniform* precision (global astrometry), the basic measurement must allow to determine large angles with a very high precision. Therefore, the Hipparcos telescope simultaneously observes, by means of a special - beam combining - mirror, two small patches of the celestial sphere, thus reducing large angles between stars in different fields of view to small angles between their images in the focal plane. The fields of view (FOV), which are 54'× 54' each, are located 58°, the so-called basic angle, apart (figure 3.1).

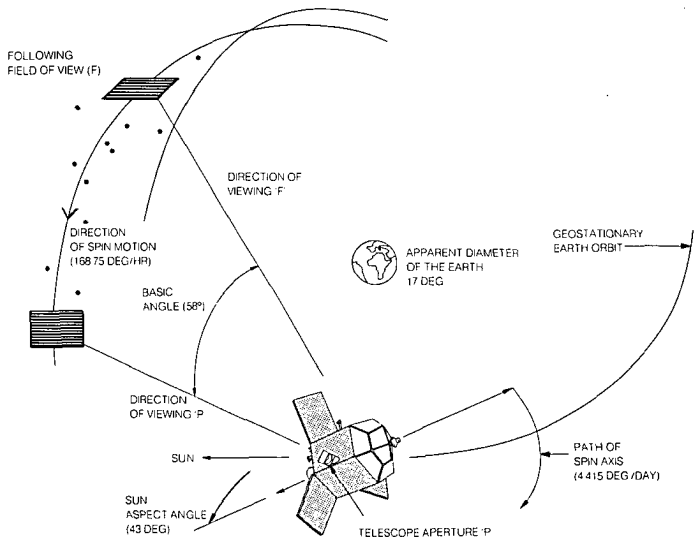


Figure 3.1 - Hipparcos measurement principle
(courtesy ESA)

The satellite is rotating slowly (11.25 rev/day) around an axis perpendicular to the two viewing directions. Thus the star images move first slowly through the first, so-called "preceding", field of view, in about 18 seconds, and reappear 20 minutes later in the "following" field of view. A grid of transparent and opaque bands, mounted in the focal plane with its bands perpendicular to the scanning direction, modulates the light of the (moving) star images (figure 3.2). The modulated light is sampled by a

detector known as an image dissector tube (IDT) at a frequency of 1200 Hz. The detector has a small sensitive area (38"), the so-called instantaneous field of view (IFOV), that can be directed through the field of view in order to select a star to be observed.

On the average four or five program stars are simultaneously present in the two fields of view. The detector is able to track - under computer control - the program stars one at a time during their passage across the field. By observing the program stars many times in turn, over short intervals of time, quasi-simultaneous observations are obtained (figure 3.5). The amount of observing time to be spent on a star, and the way in which time slices are allocated to stars, are controlled by a *star observing strategy* implemented in an on-board computer program.

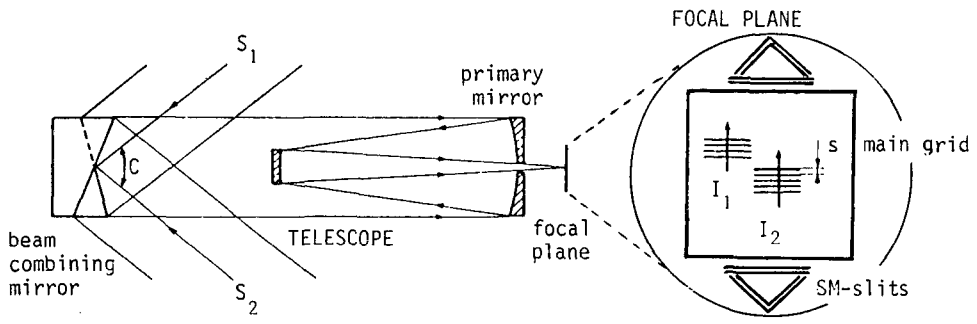


Figure 3.2- Schematic view of the telescope, with beam combiner and modulating grid.

The satellite axis of rotation is slowly precessing, with about 11.5 arcminute per hour, such that in the next revolution a new - partly overlapping - strip on the celestial sphere is scanned. To be more precise, the spin axis will be kept at a constant inclination of approximately 43° to the direction of the Sun, and will revolve around the sun in about 57 days. In this way the complete celestial sphere will be scanned several times during the mission, and a dense net of -one dimensional- measurements (about eighty per star) on well inclined great circles is obtained (Figure 3.3).

The image dissector tube (IDT) data are transmitted to the ESOC (European Space Operations Centre) ground station in Odenwald (West Germany), and will be given by ESOC to the two data reduction consortia for further processing. The data reduction consortia are going to calculate from the IDT data, per observation *frame* of 2.13 seconds, the amplitude and phase - at mid-frame time - of each of the modulated star signals. The phase of the modulated star signal at mid-frame time depends on the position of the star image on the grid. Consider the projection of a perfect grid on the celestial sphere, with slits at regular distances s, then the angular distance between a chosen reference slit and the image of a star i is

$$x_{ki} = (n_{ki} + \varphi_{ki}) \cdot s \qquad (3.1)$$

with φ the grid phase computed for the k'th frame, n the integer slit number and s the nominal separation of the slits (1"208). The integer slit number is not observed, but has to be computed somehow from the a-priori star position on the grid.

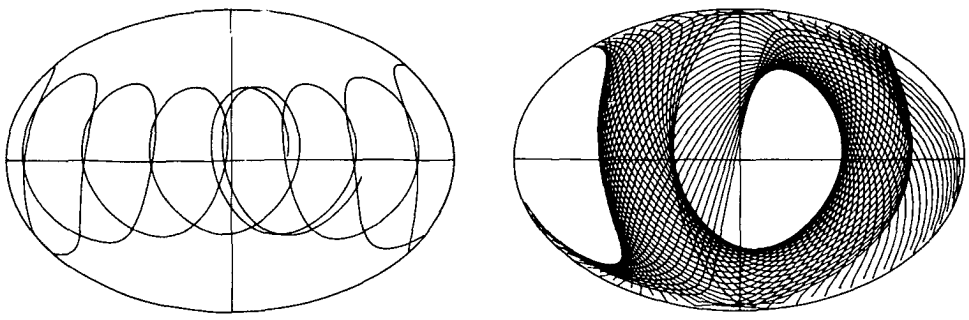


Figure 3.3 - The yearly path of the spin axis (left) and the path of the field of view for a two months period (right). The celestial sphere is projected on the $(\lambda\cos\vartheta,\vartheta)$ plane, with (λ,ϑ) ecliptic coordinates, such that the Sun moves along the horizontal line. In the right figure only 1 out of 5 scans is drawn.

The angular distance x is defined along an - unknown - great circle, the scan circle. The scan circle has orthogonal intersections with the slits. The pole and origin (the location of the central slit) of the scan circle, which are directly related to the attitude of the satellite, are not known very accurately. In order to compute the two-dimensional position of a star in a celestial reference frame, observations from different scan circles have to be combined. However, since the poles and origins are undefined, such a computation is not very straightforward. The indeterminacy in the origins (along scan attitude) can be avoided simply by considering angles between simultaneously observed stars in an observation frame, and then solving positions relative to the other stars, or, the origin of the scan circles can be solved for each observation frame as an additional unknown. The scan circle poles, *i.e.* the location of the scan circles (transversal attitude), cannot be resolved from the same data, so here additional measurements are needed. These measurements can be provided by the so-called *star mapper*. The star mapper data has not to be very precise, because the angles along a scan circle, determined by the *main* instrument, are rather insensitive to small variations in the scan circle poles.

The star mapper grids (one for redundancy) are located on both sides of the modulating grid (figures 3.2 and 3.4). Each star mapper consists of 4 slits inclined and 4 slits perpendicular to the scanning direction, and a pair of photomultipliers associated with a dichroic filter providing photometric information on the star in red and blue bands. The combination of vertical and inclined slits allows for two dimensional measurements of the star position, from which the attitude (especially scanning circle poles) can be resolved, though with a restricted accuracy. Numerical filters are used to detect the transit time of a star through each slit system, the time lag between a crossing through the vertical and the inclined slit system is a measure for the vertical position of the star on the modulating grid (*i.e.* along the bands). Each set of slits is arranged a-periodically to optimize the analysis of the detector signal.

The star mapper signal is first analyzed on-board the spacecraft for the purpose of attitude determination. A rather precise attitude (1") is needed by the on-board software for the piloting of the image dissector tube and for maintaining the prescribed scanning motion. The three-axis attitude of the

spacecraft is computed by combining the transit times associated with the passages of stars across both slit systems, in the two fields of view, with their (approximate) catalogue positions. The same kind of analysis is repeated on-ground by both data reduction consortia, but using more precise computations, more stars and better catalogue positions, in order to compute an even more precise attitude (0"1). This precision is sufficient for the approximate values in the linearized measurement equations. This computation scheme is essentially followed by both consortia. However, it may be conceptually better to consider the star mapper transit times as a second group of - two-dimensional - measurements (see chapter 4).

In order to linearize the equations and to compute the integer slit numbers, approximate values for the attitude and positions are needed. The desired 0"1 accuracy is not achieved in one step, but by an iteration process. Approximate star parameters are at first provided from the INCA catalogue. Later on during the reduction the intermediate Hipparcos catalogues are used. The attitude data is computed from the star mapper measurements, using the approximate star parameters as a starting point.

The star mapper data is also used by the complementary Tycho mission. The Tycho and Hipparcos missions cannot be flown independently. Tycho makes full use of the Hipparcos results to compute from the star mapper data star positions and photometric parameters of some 400,000 - 1,000,000 stars. On the other hand Hipparcos possibly may profit from the photometric data provided by Tycho, to correct for colour dependent errors.

3.2 Hipparcos Scanning Motion

The Hipparcos satellite will be controlled to scan a predefined path on the celestial sphere. The scanning law is based on the following rules:
- the satellite revolves 11.25 times per day around its spin axis,
- the spin axis will be kept at a constant inclination of 43° to the direction of the Sun, and will revolve around the Sun in 57 days (6.4 rev/year).

Therefore the nominal scanning motion has principal components of 11.25 rev/day, 6.4 rev/year and, due to the yearly motion of the Sun along the ecliptic, 1 rev/year.

The scanning law results in many intersecting scan circles, with 1) sufficient variation in azimuth in order to be able to get to know both coordinates and 2) sufficiently spread over the year and mission in order to get the parallaxes and proper motions. The maximum angles at which the scanning circles intersect is somewhere between 47° and 90°, depending on the position on the celestial sphere. After half a year the celestial sphere is completely scanned, and star positions can be computed. The parallaxes and proper motions[1] of stars can only be computed after one year (*i.e.* two full scans) of data. The scanning law does not take into account the positions of the Earth and Moon. Whenever one of the fields of view comes in the neighbourhood of the Earth's or Moon's disk the detectors must be switched off. This results in some loss of observing time (6%).

The attitude will be controlled by means of simultaneous cold gas jet firings on all three axes at irregular intervals, namely as far apart as possible, to get the smoothest attitude possible. A smooth attitude motion is necessary for attitude smoothing. Therefore, intermittent attitude control by cold gas jet firings, instead of the more usual - continuous - reaction wheel

[1] Although proper motions alone can be computed from less than one year of data.

control, is used, because it gives much less attitude jitter. Deviations from the nominal position of up to 10 arcmin are permitted. When one of the attitude axes becomes out of bounds cold gas jets are fired on all three axes; the duration of firing (50 - 500 ms) is computed for each axis separately. The strategy is optimized for obtaining long intervals without any firings; the computations use a model for the perturbing torques, so that actually the natural torques will help to follow the scanning law. Various computer simulations have shown that the nominal interval between two successive firings is of the order of 600 seconds, and at worst 100 seconds.

3.3 The Optical Configuration

The telescope is of a fully reflective Schmidt corrected design, with a focal length of 1400 mm. Besides a 290 mm spherical primary mirror and a flat folding mirror (needed to keep the telescope compact), it contains a mirror which combines the two fields of view into one image. The angular distance between the two fields of view is 58°. The Schmidt correction for spherical aberration has been applied in the beam combiner. The focal plane of the telescope is curved; this implies that the grid must also be written on a curved substrate.

One would expect the Hipparcos telescope, being all reflective, to be free of chromatic effects. This is not the case, because the diffraction pattern of the star image depends on the star colour. The position of the centre of gravity of the image is, therefore, also colour dependent. This effect can be as large as 5 mas and should be calibrated very carefully, or, solved during the data reduction. The colours of many stars are not known a-priori with sufficient precision, so in many cases the correction cannot be calculated even if the effect is well calibrated. The star colours can be determined by the Tycho experiment, which uses the star mapper, but it is not yet clear if these are available in time to be of any use to the Hipparcos data reduction.

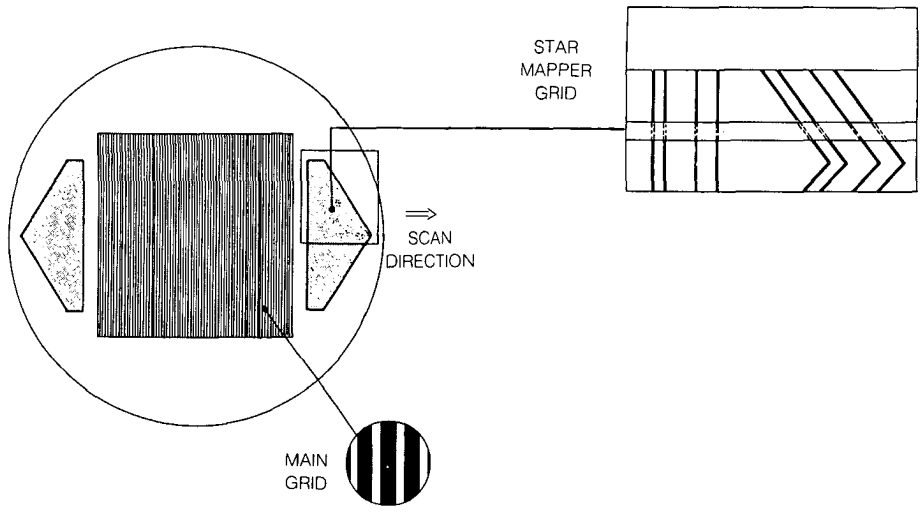


Figure 3.4 - The focal plane with the main and star mapper grids.

The modulating grid consists of 2688 pairs of transparent and opaque bands with a period corresponding to 1"208 on the celestial sphere. The grid, with dimensions of 22 mm x 22 mm, covers an area of 0.9° x 0.9° on the sky.

The ratio of the widths of the opaque and transparent bands has been optimized in order to obtain the largest possible modulation depth and efficiency. The error in the phase determination caused by statistical photon noise is inversely proportional to the modulation depth and the number of photons that are received by the detector. The grid is written on a curved substratum in 7800 identical patches, which are aligned as good as possible. Alignment errors, the so-called medium scale distortion of the grid, must be determined and corrected during the data reduction.

It is not sufficient to sample the modulated signal with a simple photomultiplier, which is sensitive for photons coming from all over the field of view. Instead a more advanced detector, a so-called image dissector tube (IDT), is used. An image dissector tube is only sensitive for photons coming from a small - steerable - region on the sky, the so-called instantaneous field of view (IFOV), which is about 35" in diameter. Therefore individual star images will generally not be mixed, although the modulated signal of some stars will be perturbed by the light of other, usually bright, stars in their neighbourhood. This effect, called veiling glare, can be corrected during the data reduction once the magnitude and position (on the grid) of perturbing stars are known. The perturbing star may be observed in the other field of view, so the stars are not necessarily neighbours on the celestial sphere. Double stars with large separations will be treated in the same way as veiling glare stars, *i.e.* the position of the individual components will be estimated separately. However, double stars with small separations (<30") will be treated as one, using special algorithms, because these small separations produce different modulation patterns.

The star mapper grids, consisting of 4 inclined and 4 vertical slits each, are located on both sides of the modulating grid. Only one of the star mapper grids is used at a time, the other is provided for redundancy. Two photomultipliers, each associated to a different colour band, are used as detectors. The photomultipliers are, in contrast to the image dissector tube, sensitive for photons from all over the star mapper grid.

3.4 The Star Observing Strategy

Quasi-simultaneous observations between stars present in the same observation frame are obtained by frequently switching the sensitive area of the image dissector tube between stars. In this way the attitude jitter along the scanning direction is partly filtered out. The way in which the stars are followed, and the time which is available to each star, is directed by the *Star Observing Strategy*.

Hipparcos observes only *selected* stars, and obviously the star observing strategy must be told which stars to observe and how to find them on the grid. The Input Catalogue Consortium, which selected the 110,000 program stars from among a total of about 200,000 proposed by the astronomical community, has compiled a star catalogue with the position, apparent magnitude and other information of these stars. The European Space Operations Centre (ESOC) at Darmstadt, Germany, is going to uplink every five minutes a subset of this information, pertaining to the stars which are expected to be visible in the next five minutes. The uplinked positions are used for two purposes:

- computation of the star mapper attitude, using the positions of bright stars visible in the star mapper field of view, and
- to compute, from the star mapper attitude and uplinked catalogue positions, the location of the star images on the grid.

The star locations are used to pilot the sensitive area of the detector

(figure 3.5). The *target observing time*, and some other indices, are used to determine how much time should be allocated to each visible star.

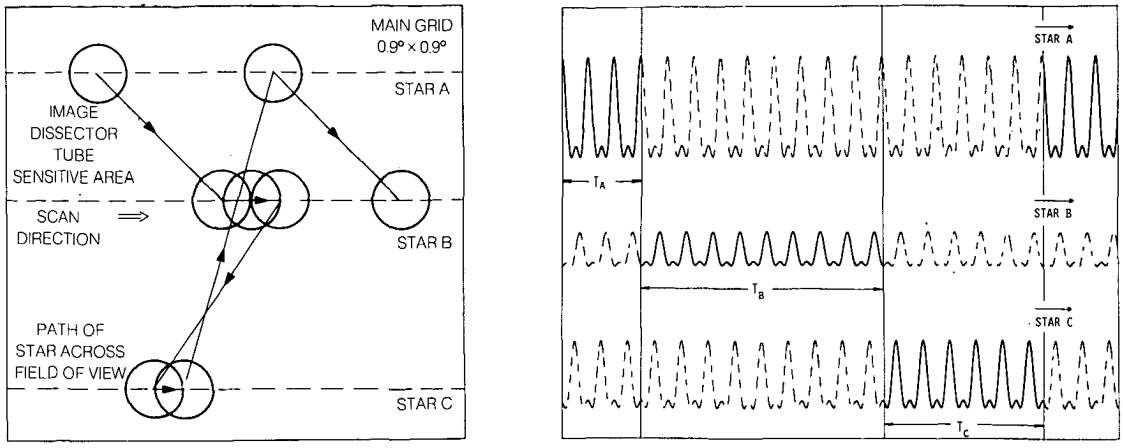


Figure 3.5 - The motions of the sensitive area of the image dissector tube (IFOV) over the field (left) and the resulting, observed, signals (right). (Courtesy ESA)

The observing time is distributed in multiples of an elementary interval, duration *T2*, a so-called *slot*. Each slot consists of 8 sampling periods. The sampling period, *T1*, is 1/1200 s. Stars are always observed for an integer number of slots. Groups of up to 10 stars are observed during a period *T3* = 20•*T2*. In one observation frame, duration *T4*, the same group of stars are observed 16 times in the same order, so the frame period *T4* = 16•*T3* = 320•*T2* ≃ 2.13 s. (table 3.1). The star observing strategy decides how many of the 20 available slots should be allocated to each star. The decision is based on the following data:

- a selection index, which gives the priority with respect to other program stars.
- the minimum observing time in number of slots (*T2*). This is a simple function of the magnitude of the star.
- the target observing time, which depends on the desired observing time, the expected number of scans over the whole mission and the observing time already spend on the star.

These parameters may vary during the mission.

Table 3.1 - Periods in the observing strategy

	name	duration	
T1	sampling period	1/1200 s.	
T2	observation slot	1/150 s.	(8 *T1*)
T3	interlacing period	2/15 s.	(20 *T2*)
T4	observation frame	32/15 s.	(16 *T3*)

3.5 Phase Estimation from IDT Data

The modulated star signal can be written as the convolution of the telescope two-dimensional diffraction pattern, the star image and the transmission function of the grid. The geometry of the telescope is such that for most stars only the first two harmonics will be present in the modulated signal. Then the modulated intensity $I(t)$ of the star can be written as

$$I(t) = I_0 + B + I_0 M_1 \cos(g_t + \varphi_1) + I_0 M_2 \cos(2(g_t + \varphi_2)) \qquad (3.2)$$

where B is the observed intensity due to the sky background, straylight and IDT dark count, I_0 the mean observed star intensity, depending on its apparent magnitude and the Hipparcos telescope characteristics: M_1 and M_2 are the modulation coefficients, the amplitude of the first and second harmonic, φ_1 and φ_2 are the phase shift of the first and second harmonic at mid-frame time t=0, and g_t the incremental displacement on the grid as a function of time t [*Canuto et al., 1983b, Kovalevsky et al., 1985b*]. The incremental displacement, assuming a uniform scanning speed ω and a regular grid, is simply ωt.

The incremental displacement g_t of a star on the grid is actually

$$g_t = g(t,\mathbf{p}) + \text{jitter} + \text{grid irregularities} \qquad (3.3)$$

where $g(t,\mathbf{p})$ is a simple parametric model, with parameters $\mathbf{p}$, which are either known from the attitude reconstruction or have to be estimated. Jitter is defined as the attitude deviations from the parametric model, that act as additional noise. The grid irregularities are assumed to be calibrated before launch, during the in orbit commissioning phase and as well during the data reduction itself. For short intervals of time (*e.g.* the frame period *T4*) the parametric model can be very simple, $g_t = \omega \cdot t$, where ω is the scanning speed which is assumed to be known from the attitude reconstitution or can be estimated from the IDT data.

In one observation frame of 2.13 s. each star will be observed 16 times during a number of consecutive slots *T2*. The expected value of the *photon count* of the k'th sample is given by

$$I_k = Round\left\{ \int_{T1} I(t)\, dt \right\} \qquad (3.4)$$

The actual counts are Poisson distributed. The basic modulation frequency, given by the scanning speed and grid period, is approximately 140 Hz, corresponding to a period of a little more than one slot *T2*.

The modulating coefficients M_1 and M_2 depend on the diffraction pattern of the star image. For single stars M_1 and M_2 differ, due to the star colour influence, by not more than 7% from some nominal values M_1^0 and M_2^0. However, for double stars M_1 and M_2 can take any value between zero and the single star value [*Kovalevsky et al., 1985b*]. In addition, the images corresponding to both harmonics do not coincide, *i.e.* $\varphi_1 \neq 2\varphi_2$. So, essentially five

parameters have to be estimated from the modulated signal, *e.g.* I_0+B, I_0M_1, I_0M_2, φ_1 and φ_2, the so-called *five parameter model*. For single stars, in view of the small deviation of M_1 and M_2 from their nominal values and the small phase difference $2\varphi_2-\varphi_1$, a single weighted phase may be estimated. Now only 3 parameters, *e.g.* B, I_0 and the *weighted* phase φ have to be estimated, the so-called *three parameter model*. The weighted phase φ can be computed as

$$\varphi = c_1 \varphi_1 + c_2 \varphi_2 \quad , \quad c_1 + c_2 = 1 \tag{3.5}$$

where the weighting factors c_1 and c_2 ($c_1 \simeq 0.56$ and $c_2 \simeq 0.44$) depend on the estimated modulation coefficients, and thus on the star colour [*Kovalevsky, 1984*].

In the case of double, or multiple, stars both the five and three parameter solutions are computed. The five parameter solution is used in the further analysis of double star systems. The three parameter solution, with a single weighted phase, is processed by the standard data analysis chain starting with the great circle reduction. The weighted phase refers to some "*mean*" position on the grid, and does not correspond to a physical point in the multiple star system. In addition the modulation coefficients and the observed phase difference change with the angle from which the multiple system is scanned.

Uncertainties in the computed modulation coefficients, which affect the weighting factors, make the definition of the weighted phase in case of double stars even more disputable. Therefore, at least within FAST, the weighting factors will be fixed throughout the mission on a value close to the single star case. Double stars must be flagged at this stage, so that they will get the proper treatment with extra parameters in subsequent stages of the reduction. The weighting factors will also be fixed for the single star case, partly because it is the simplest solution, but also in order to have a consistent definition of the weighted phase over a reference great circle. It has been shown that the degradation due to the choice of fixed coefficients is at most 2% for single stars [*Kovalevsky et al., 1985b*]. However, since the modulation coefficients and phase difference are slightly colour dependent, the weighted phase depends on the star colour index too. This chromatic effect must be calibrated during the further data analysis.

The three and five parameter solutions can be computed by a maximum likelihood estimation, although this is not a very efficient approach with respect to computing time. Therefore both consortia compute an approximate maximum likelihood solution. The five parameter solution is computed by the following procedure:

- "binning" of the photon counts (*i.e.* partitioning into classes),
- Fourier analysis of the binned counts, which gives approximate values for the 5 parameters,
- Gauss-Markov estimation of the 5 parameters, where the normal matrix is computed from the binned counts and the weights are taken from the Fourier analysis.

This procedure is not directly used on equation 3.2, but on a slightly modified equation (see *e.g.* [*Fassino, 1986*]). The covariance matrix of the computed parameters is an important byproduct of the Gauss-Markov estimation.

An approximate maximum likelihood solution for the three parameter solution can be obtained by the following procedure:
- least squares estimation of B, I and φ from the five parameter solution,
- Gauss-Markov estimation of the three parameter solution from the photon counts, using the provisional least squares solution in order to determine the weights and using the binned data to compute the normal matrix.

Statistical tests are used to test for model deviations from the five and three parameter solution (is the signal according to the model), and for significance (is it modulation or noise) [*Fassino, 1986*]. During the phase estimation the alignment errors of the grid patches will be corrected. These medium scale grid distortions will be calibrated on the ground and during the in-orbit commissioning phase of the satellite.

The main source of errors in the estimated phase is the, Poisson distributed, photon noise. Other errors, notably due to attitude jitter, not calibrated small and medium scale distortions of the grid and variability in the scan speed, stay below 1 mas. So the error in the phase estimate is dominated by photon noise, and can be described very well by the variances computed during the approximate maximum likelihood estimation. Moreover, the correlation between the phase estimates for different stars is negligible.

CHAPTER 4

GEOMETRIC ASPECTS OF THE HIPPARCOS DATA REDUCTION

The astrometric parameters of the program stars are computed by an iterative adjustment in steps from the "observed" main grid (IDT) phases and the star mapper transit times. In particular the so-called three step procedure, the base-line for both consortia, is discussed.

4.1 Introduction

The Hipparcos data reduction consists of two stages. In the first stage the relative position of the star images on the grid at a given instant of time are computed from the image dissector tube (IDT) and the star mapper (SM) photon counts. More specific, the IDT phases of the program stars visible in the field of view are given at regular intervals of ~2.13 s. On the other hand, the transit times of bright program stars through the inclined and vertical reference slits, are computed from the SM data. The second stage consists mainly of a large scale adjustment, during which the five astrometric parameters of all program stars are computed, starting from the IDT phases and SM transit times.

Both NDAC and FAST, the scientific consortia in charge of the Hipparcos data reduction, have organized the second stage as an *iterative adjustment in steps*, following roughly the *three step procedure* proposed by Lindegren [*Lindegren, 1979*]. The three steps are:

(1) attitude reconstitution (AR) and great circle reduction (GCR),
(2) sphere reconstitution (SR),
(3) astrometric parameter extraction (APE)

In the first step ~10 hours of measurements, covering about five successive scan circles, are collected and processed together. The star mapper transit times and main grid phases are treated separately. During the *attitude reconstitution* the three-axis attitude of the satellite is computed from the SM transit times, using a given star catalogue. During the *great circle reduction* the star abscissae on a reference great circle (RGC), chosen somewhere in the middle of the scanning circles, are computed from the main grid phases by a weighted least squares adjustment. At the same time an improved along scan attitude and some instrumental parameters are computed. The star ordinates are not solved, and also the abscissae are determined with an arbitrary zero point. The unknown zero points are solved in the second step, using only the abscissae of primary stars collected over many well inclined RGC's. The astrometric parameters of the primary stars are determined at the same time. Finally, in the third step, the astrometric parameters of the remaining, secondary, stars are computed. At this point it should be mentioned that we implicitly deal with the FAST approach. The NDAC approach is more or less similar, and only major deviations will be mentioned.

The adjustment in steps is iterated to overcome the approximate character of the solution. The iterations are needed mainly because in several stages of the reduction only part of the unknowns is actually solved, the rest is kept fixed. A process of this type can be formulated in terms of

a block Gauss-Seidel solution of a positive definite system. In our case, the three step procedure, the convergence of this process is very good: just 2 or 3 iterations will do. The three step procedure is very practical because it follows closely the data acquisition process. The first step, consisting of the attitude reconstitution and great circle reduction, can be done immediately after the data has been collected. The sphere reconstitution and astrometric parameter extraction will be done after each half a year of data have been processed by the great circle reduction, resulting in a series of provisional star catalogues. Each time a new catalogue has been computed part of the RGC sets are re-computed, using the new catalogue. Nevertheless, within FAST several alternative schemes have been proposed [*Betti et al., 1983a, 1985b, 1986b, Galligani, 1986*]. A final iteration with one of these methods is foreseen as an independent check and in order to avoid certain systematic effects.

The Hipparcos data reduction poses some interesting estimability questions. It is generally accepted that a free network of angular measurements, between uniformly moving stars on the celestial sphere, has a rank defect of six. Six constraints are needed to get a solution: *e.g.* the position and proper motion of one point, and the azimuth and its proper motion component to an other point, can be fixed. This also holds for the Hipparcos network, although upon closer investigation it turns out that this is not immediate. During the three step procedure the RGC poles are fixed, which amounts to a weak over-constraining of the solution which, however, can be neglected after the final iteration. Theoretically, due to the parallax and annual proper motion of stars, the rank defect should disappear at all. However, the rank defect remains 6, because parallaxes and proper motions are not large enough to give a significant contribution to the unknown orientation of the reference frame [*Betti and Sanso, 1983b, Donati, 1986b, Van Daalen et al., 1986c*].

4.2 The Geometric Relations

Preprocessing of the IDT and star mapper photon counts results in two types of observables:

- The IDT phases, or along scan grid coordinates, computed every 2.13 seconds from the modulated light signals of, on the average, 4-5 quasi simultaneously visible program stars.
- Star Mapper transit times, the epoch at which a relatively bright program star crosses one of the slit systems.

These observations can be expressed, using non-linear equations, in terms of the astrometric parameters, which are the main objective of the mission, and some auxiliary parameters. In this section we will derive these equations, *viz.* our mathematical model, beginning with the astrometric parameters, and finishing with the above mentioned observations.

4.2.1 Catalogue Positions

The primary aim of the Hipparcos satellite is to compute a star catalogue of some 112,000 stars. The positions of these celestial objects are specified by a set of coordinates in a specific celestial reference system. The celestial reference system is defined by its origin and the direction of its axis. The origin of the celestial reference frame is chosen at the *barycentre* of our Solar system. The first axis is chosen in the direction to the vernal equinox, that is where the path of a fictious mean Sun intersects the equator. The third axis is chosen orthogonal to either the mean *ecliptic* or the mean *equatorial* plane, resulting respectively in an ecliptic or equatorial reference system. The second axis completes the system. The

ecliptic and equator are not fixed in inertial space, therefore such a definition is only valid at a certain epoch, currently J2000.0.

The positions of celestial objects are given by the distance R from the origin of the system and two spherical angles l and b, called *longitude* and *latitude*. The longitude and latitude define a unit vector $\mathbf{r}$,

$$\mathbf{r} = \begin{bmatrix} \cos l \cos b \\ \sin l \cos b \\ \sin b \end{bmatrix} \qquad (4.1.a)$$

which gives the fictitious position on a two-dimensional manifold, called *celestial sphere*. The position vector of the celestial objects is then

$$\mathbf{x} = R \cdot \mathbf{r} \qquad (4.1.b)$$

The distance R is usually given in Astronomical Units (AU), the "average" distance of the Earth from the Sun [*IAU, 1977*]. Then

$$R = 1 \,/\, \sin \omega \qquad (4.2)$$

with ω the *annual parallax*. The annual parallax is the maximum -apparent- angular displacement of the position of the star due to the eccentricity of the measurement platform, usually Earth, with respect to the barycentre of our solar system. The parallax is an observable during the Hipparcos mission.

Most celestial objects are moving in time with respect to each other. Therefore catalogue positions also refer to a common epoch. For Hipparcos an epoch will be chosen close to the mean time of observation (*e.g.* J1990.0). Stars, or multiple star systems, have no detectable deviations from a rectilinear motion during the lifetime of Hipparcos, and therefore positions at other epochs are given by

$$\mathbf{x} = \mathbf{x}_{t_0} + (t - t_0)\,\dot{\mathbf{x}} \qquad (4.3)$$

Let us decompose $\dot{\mathbf{x}}$ into components along and at right angles to the direction vector $\mathbf{r}$ at the common epoch t_0, then

$$\dot{\mathbf{x}} = R\,\upsilon\,\mathbf{r} + R\,\mu_l \cos b\,\mathbf{e}_l + R\,\mu_b\,\mathbf{e}_b \qquad (4.4)$$

with $R \cdot \upsilon$ the *radial velocity* and μ the two components of *proper motion* of a star in radians. $\mathbf{r}$, $\mathbf{e}_l$ and $\mathbf{e}_b$ are unit vectors which form a orthonormal system, with

$$\mathbf{e}_l = \begin{bmatrix} -\sin l \\ \cos l \\ 0 \end{bmatrix} \quad \text{and} \quad \mathbf{e}_b = \begin{bmatrix} -\cos l \sin b \\ -\sin l \sin b \\ \cos b \end{bmatrix} \qquad (4.5)$$

Due to the radial velocity component the stars do not move with an uniform velocity over a great circle on the celestial sphere, *i.e.* despite the rectilinear motion of stars the proper motion may be different at other epochs (This effect is, of course, caused by the different distance to stars at other epochs). The distance R_t at epoch t is

$$\|\mathbf{x}_t\| = R \sqrt{1 + 2\upsilon(t-t_0) + (\mu^2+\upsilon^2)(t-t_0)^2} \qquad (4.6)$$

with $\mu^2 = \mu_l^2 \cos^2 b + \mu_b^2$. When we neglect second order effects, caused by the change in distance due to the non negligible radial velocity, the direction

vector **r** at time t becomes

$$\mathbf{r}_t = \{ (1+\upsilon(t-t_0))\ \mathbf{r} + \mu_l \cos b\ (t-t_0)\ \mathbf{e}_l + \mu_b (t-t_0)\ \mathbf{e}_b \} \cdot \frac{R}{R_t} \qquad (4.7)$$

The radial velocity component is, however, very small compared to the distance itself, and therefore it will not change the observed parallaxes and proper motions of stars significantly during the Hipparcos mission. This also implies that the radial velocities are not estimable from the Hipparcos data. The second order effects, due to the radial velocity, on the proper motions and parallaxes will be applied as corrections to the observations.

4.2.2 Star Positions as seen by Hipparcos

In an instrumental reference frame the stars are seen at positions different from those of the barycentre of the solar system. Three effects play an role. Firstly, the instrumental reference frame has not the same orientation as the adopted celestial reference frame. Secondly, the Hipparcos measurement platform is not at the barycentre of the solar system, which results in the parallactic displacement of objects, of which the annual parallax is a part. Thirdly, the platform is moving with respect to the barycentre and the actual space-time metric is curved, resulting, respectively, in stellar aberration and relativistic light deflection. Obviously the first effect, the orientation of the instrumental frame, is the largest of the three. In fact, either the complete three-axis attitude of the satellite must be solved, giving rise to many auxiliary unknowns in the mission, or we must switch from directions to angles as observations, which, in fact, comes down to the same thing. The effects mentioned under the second and third point are much smaller. They can be associated with small displacements of the barycentric star position, resulting in respectively the geometric and apparent position of stars.

The *geometric position* of a star (at time t) refers to the centre of the Earth. It follows from the barycentric position when displacements caused by proper motion and *annual* parallax are taken into account, *i.e.* it follows directly from the five unknown astrometric parameters. The *apparent position* at time t is defined as the point on the celestial sphere where the object is *seen* from the centre of the Hipparcos instrument. It follows from the geometric position when displacements due to stellar aberration, relativistic light deflection, daily parallax and second order effects due to a non-zero radial velocity are taken into account. In fact, the Hipparcos measurements have to be corrected for the difference between apparent and geometric positions. But the apparent position cannot be computed from Hipparcos data alone: the radial velocity and orbital parameters of the spacecraft are needed in order to compute differential (daily) parallax, the ephemerides of the Sun, Earth, Moon and the large planets are needed in order to compute relativistic light deflection, and the spacecraft's velocity vector is needed in order to compute stellar aberration [*Walter et al., 1986*]. The orbital parameters of the spacecraft are not very critical, but an error in the spacecraft's velocity of 1 m/s already gives an aberration error in the star positions of 1 mas. It is expected that the velocity is determined to better than 30 cm/s, so the error remains acceptable.

The three-axis attitude of the satellite is introduced as an auxiliary unknown in the mission. The situation differs from observations done from the Earth, for which the orientation in space is fairly well known, *e.g.* from universal time and precession, nutation and polar motion series. Let **p** be the unit direction vector of star i in the instrument frame and **r** the unit direction vector referring to the geometric position in the celestial

reference frame at the time of observation, then

$$\mathbf{p}_{ki} = \mathbf{A}_k \cdot (\mathbf{r}_{ki} + \Delta\mathbf{r}_{ki}) \qquad (4.8)$$

where **A** is the so-called attitude matrix, an orthogonal rotation matrix which gives the orientation of the instrument frame in the celestial reference frame, and **Δr** the correction for apparent places. The index k refers to the observation frame. The attitude matrix **A** is defined by three rotations, *e.g.* two rotations which define the position of the scanning circle pole (the Z-axis of the instrument frame) and a rotation around the Z-axis which defines the position -along scan- of the X-axis. The geometric position can be expressed in the 5 unknown astrometric parameters. For an ecliptic reference system, using equation (4.7), but neglecting the radial velocity and changes in the vector norm, and introducing the effect of annual parallax, we get:

$$\mathbf{r}_{ki} = \mathbf{r} + (\mu_l \cos b\ (t-t_0) + \omega \cos \vartheta_t)\ \mathbf{e}_l + \mu_b (t-t_0)\ \mathbf{e}_b \qquad (4.9)$$

with ϑ the angle between the star and the Sun as seen from the instrument.

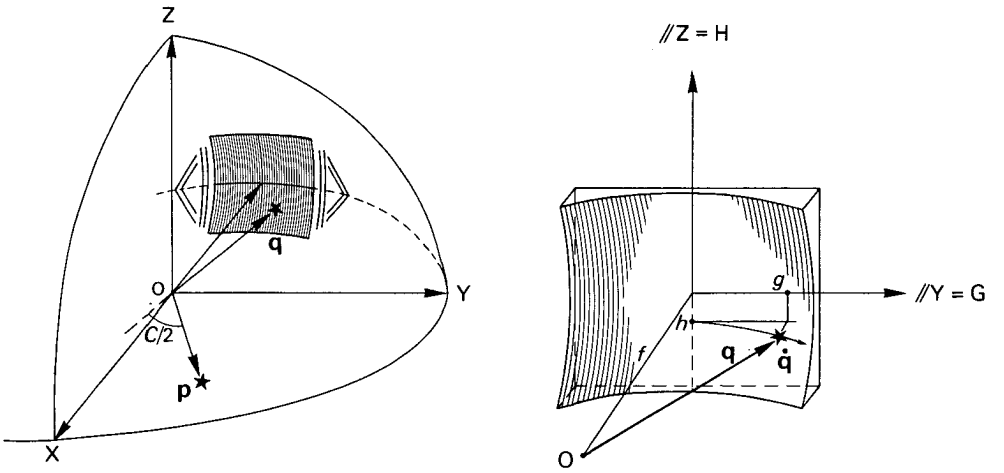


Figure 4.1 - geometry of the Hipparcos telescope

Let us define an -instrumental- reference frame, where the X-axis is defined as the bisector of the optical axes for the viewing directions of the telescope and where the Z-axis is perpendicular to the optical axis and has its positive direction corresponding to the rotation vector of the satellite. The Y-axis completes the triad. Now let us assume a perfect telescope and beam combining mirror, then the Cartesian coordinate vector of the star image on the -curved- focal plane is

$$\mathbf{q}_{ki} = -f \cdot \begin{bmatrix} \cos C/2 & f_i \sin C/2 & 0 \\ -f_i \sin C/2 & \cos C/2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \mathbf{p}_{ki} \qquad (4.10)$$

with f the focal length of the telescope (~1400 mm), $C/2$ half the basic angle (~29°) and f_i the field of view index, f_i=+1 for the preceding field of view

and f_i=-1 for the following field of view. The index i refers to the star and the index k refers to the observation frame. Let $(x_{ki} + f_i \cdot C/2)$ and y_{ki} be the spherical angles of a star on the celestial sphere in the above mentioned instrumental reference system, and let **p'** be the corresponding Cartesian coordinate vector

$$\mathbf{p}'_{ki} = \begin{bmatrix} \cos x_{ki} \cos y_{ki} \\ \sin x_{ki} \cos y_{ki} \\ \sin y_{ki} \end{bmatrix} \qquad (4.11)$$

then equation (4.10) becomes simply $\mathbf{q}=-f\,\mathbf{p}'$. The spherical angles x and y are the so-called *field coordinates* (see figure 4.1).

Despite the fact that the actual grid, which is located in the -curved- focal plane, is written on a curved substrate, the computations are carried out on a flat surface [*Kovalevsky et al., 1986c*]. Parallel projection of the grid on a flat surface leaves the second and third component of the image coordinate vector **q** unchanged, while the first component becomes $-f$. The second and third components of **q**, after division by the focal length f, are the so-called *grid coordinates* g and h:

$$\begin{aligned} g_{ki} &= -\sin x_{ki} \cos y_{ki} \\ h_{ki} &= -\sin y_{ki} \end{aligned} \qquad (4.12)$$

The grid coordinates, which are actually direction cosines, are dimensionless quantities. The positive axis of g is in the direction of the moving star images.

4.2.3 Observations on the Main Grid

Let us assume a grid, which after projection on a flat surface, gives perfectly parallel slits at regular intervals s' (8.2 μm). Assume as well that the grid is perfectly aligned with the telescope, *i.e.* the projection of the slits on the flat surface are all parallel to the Z-axis and there is a -reference- slit with g=0 at the centre of the slit. Then the along scan grid coordinate of star i, observed in the k'th frame, is

$$g_{ki} = (n_{ki} + \varphi_{ki}) \cdot s \qquad (4.13)$$

with φ the observed IDT phase, n the integer slit number and s the grid period. The grid period s is, like the grid coordinates g and h, divided by the focal length f, with

$$s = s'/f \simeq \frac{8.2\ \mu\text{m}}{1400\ \text{mm}} \ (\simeq 1.208\ \text{arcsec}) \qquad (4.14)$$

So, s, g and h are dimensionless quantities. The integer slit number is computed from approximate values for the star position and satellite attitude. Approximate star parameters are at first provided by the INCA catalogue. Later on during the reduction intermediate Hipparcos catalogues are used. The attitude is initially provided from the star mapper measurements, but the along scan attitude will be improved later on.

The actual instrument and grid are, of course, not perfect. Firstly, the width and separation of the individual slits may be slightly different, and there may be blemishes on the grid. This is the *small scale* distortion of the

grid. Secondly, the grid is not written in one go, but by repeating individual patches. These patches may have small alignment errors, causing *medium scale* distortions of the grid. Finally, the substrate on which the grid is written and the two mirrors, the beam combining and primary mirror, are not perfectly aligned with the adopted instrument system and they may deform due to temperature variations and ageing. But also chromaticity effects in the optics give small displacements of the star image in the focal plane, which depend on the star colour and position in the field of view. These effects are comprised in the *large scale* distortion of the grid.

The observed grid phase will be corrected for the small and medium scale distortion of the grid on basis of calibration measurements, obtained on ground and during the in-orbit commissioning of the satellite. By convention, the grid coordinate g will be computed from formula (4.13), using the grid phase φ after the described correction. On the other hand, the field coordinates are associated with the apparent position of the observed star. Therefore, the nominal transform of grid to field coordinates, given in equation (4.12), does not hold in general due to the large scale distortion of the instrument. Instead we have

$$\begin{aligned} g_{ki} &= g(x_{ki}, y_{ki}, (B\text{-}V)_i, f_i, t_k) \\ h_{ki} &= h(x_{ki}, y_{ki}, (B\text{-}V)_i, f_i, t_k) \end{aligned} \qquad (4.15)$$

which is the so-called *field to grid* transform, or vice-versa, *grid to field* transform. Equation (4.15) can also be written as (4.12) *plus* the large scale distortion. It turns out that both the field to grid transform and the large scale distortion, can be described with sufficient precision as a polynomial function of time, star colour and place on the grid for each field of view separately. The coefficients of this polynomial will be estimated in the data reduction, mainly during the great circle reduction. It has been shown that a third order polynomial in the grid or field coordinates is sufficient [*Bertani et al., 1986, Badiali et al, 1986*]. The precise form of the polynomial is discussed in chapter 5.

4.2.4 Star Mapper Observations

The star mapper transit times form a second group of observables. Although they are less precise than the observations from the main grid, they are indispensable. The star mapper observations are in fact complementary to the main grid observations since they provide information on both g and h, whereas the main grid observations give only information on the along scan component g.

Let us assume a perfect star mapper grid. After projection on a flat surface the reference lines of the vertical and inclined slits systems respectively have the coordinates representation (g_v, z) and $(g_i \pm z, z)$, with z a free variable and with $|z| \leq 40'$ (figure 4.2). Assume that the star images are moving with uniform velocity ω over the grid. Let t_v be the time the star crosses the vertical slit and t_i the time the star crosses the inclined slit, then at t_v the position of the star image on the grid is

$$\begin{aligned} g_{(t_v)i} &= g_v \\ h_{(t_v)i} &= \pm \left((g_v - g_i) - \omega \cdot (t_v - t_i) \right) \end{aligned} \qquad (4.16)$$

The star mapper grid coordinates can be related to the field coordinates by

the same kind of transformation as used for the main grid, except that in case of the star mapper the medium scale distortion is absent. The large scale distortion is in principle the same as for the main grid, and it is currently investigated if the large scale distortion estimated from the main grid data can also be applied to the star mapper data.

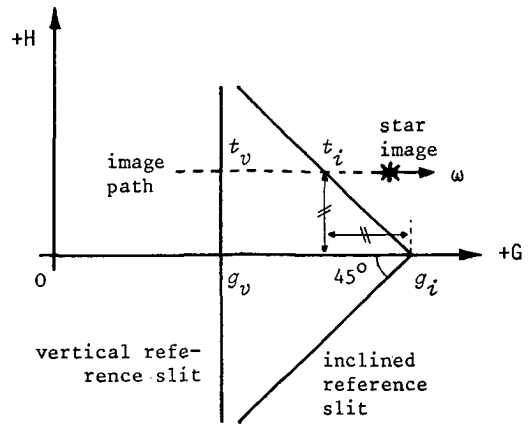


Figure 4.2 - Star Mapper measurements

4.3 The three step procedure

4.3.1 The Principles

In principle the five astrometric parameters - the barycentric position, proper motion and parallax of stars - can be computed by a single weighted least squares adjustment from the main grid phases and star mapper transit times. Over the 2.5 years of the mission some 150×10^6 main grid phases, and some 14×10^6 star mapper transit times, are collected, and some 600,000 astrometric parameters, a large number of auxiliary attitude (1×10^6 in smoothing mode, 90×10^6 in geometric mode) and instrumental parameters (100,000) have to be computed. Undoubtedly, it is impossible with present day techniques to solve such a system in a single step.

Therefore, both data reduction consortia have organized the geometric part of their reduction process in the form of an *iterative adjustment in steps*, following roughly the *three step procedure* of [*Lindegren*, 1979]. The steps are: 1) attitude reconstitution and great circle reduction, 2) sphere reconstitution and 3) astrometric parameter extraction. "Iterative adjustment in steps" is not a pleonasm: the problem is set up as an approximate adjustment in steps which is to be iterated several times to overcome the approximate character of the solution (figure 4.3). Essential in the three step procedure is that:

a. the processing, in the first step, does not wait till both coordinates of a star can be estimated,
b. the star mapper data and IDT are treated separately by respectively the attitude reconstitution and great circle reduction,
c. two groups of stars (primary and secondary) are distinguished, which get a different treatment in the sphere reconstitution and astrometric parameter extraction.

Actually, the problem is split in a large number of smaller adjustments, especially in the first step.

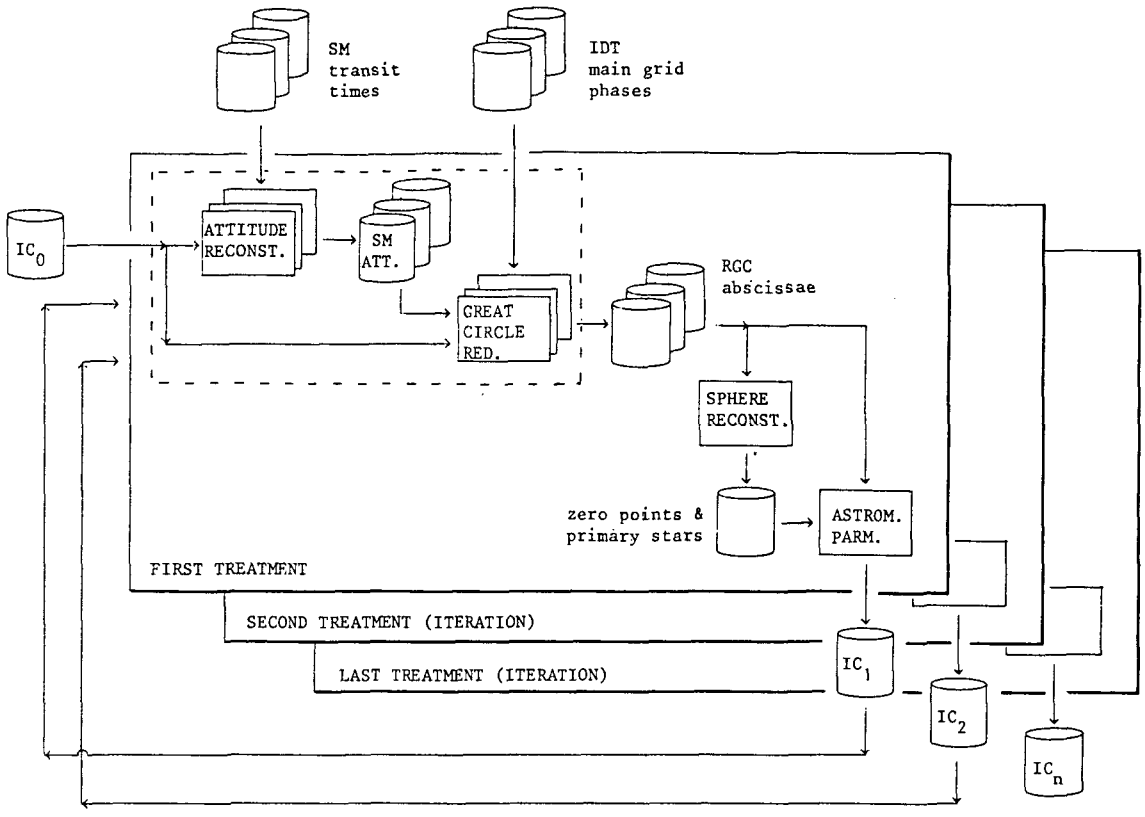


Figure 4.3 - The three step procedure (IC_0 = INCA input catalogue, IC_n = intermediate Hipparcos catalogues)

In the first step of the three step procedure the processing is carried out within so-called *RGC-sets*, contiguous batches of data collected during about 5 revolutions of the satellite (10.7 hours of data). Thus, there will be approximately 1800 different RGC-sets over the mission (2.5 year). A *reference great circle* (RGC) is chosen in the middle of the scanning circles which form the RGC-set. The precession of the scanning circles is $\sim 0.^\circ 4$/revolution, hence the maximum inclination of the scanning circles with respect to the RGC is $\sim 1^\circ$. As a result, in each RGC-set only a small band on the celestial sphere is scanned. The average width of this band is 2°, with a minimum of $\sim 1^\circ$ near the nodes of the scanning circles with the RGC and a maximum of $\sim 3^\circ$ at a distance of 90° from the nodes. Some of the properties of a typical RGC-set are summarized in table 4.1.

The star positions and satellite attitude are solved from the data collected during an RGC set in an intermediate reference frame, defined by the chosen reference great circle (RGC). The relative motions of the stars due to proper motion and annual parallax within an RGC set are very small. Therefore, the proper motion and annual parallax are not solved in the first step, but are kept fixed. The star mapper and main grid data are processed separately in a weighted least squares sense by respectively the attitude reconstitution and great circle reduction. It is, however, characteristic for Hipparcos that its main grid provides along scan information only. Therefore,

and because of the small inclination of the scanning circles with respect to the RGC, the great circle reduction can only determine the along scan attitude component and star abscissae in the RGC reference frame. Furthermore, it is typical for the attitude reconstitution that, except for the first treatment of the data, the star positions and along scan attitude are kept fixed and only the two transversal attitude components are estimated.

Table 4.1 - Properties of a typical RGC (assuming an input catalogue of 112,000 stars and an even distribution of the stars over the celestial sphere).

duration:	10.7 hours
number of revolutions:	5
average scan speed:	168.75 arcsec/s
precession of the scanning circles:	0.4 ^{0}/revolution
overlap between two consecutive frames:	89 %
width of the band with observed stars:	~2^0 (min. 1^0, max. 3^0)
average number of:	
stars (~1.8% of the program stars):	2000
passages of a star through both fields of view:	
main grid:	2.2
star mapper slit system:	1.7
program stars simultaneously visible on the grid:	4.5
observation frames:	18,000
grid coordinates:	80,000

The results of several great circle reductions, corresponding to at least half a year of data, are further processed in the sphere reconstitution and astrometric parameter extraction. Two types of stars are distinguished: in general bright, unproblematic, stars are called *primary stars* and faint, doubtful stars are called *secondary stars*. During the sphere reconstitution the unknown zero points of the RGC's and the astrometric parameters of primary stars are solved using only the observations to primary stars. Finally in the astrometric parameter extraction the astrometric parameters of the secondary stars are solved. It is assumed that at least 40% of the stars will survive the reduction process as primary stars.

The data reduction, in particular the great circle reduction and attitude reconstitution, should be iterated several times mainly because in several stages of the reduction only part of the unknowns are actually solved, the rest is kept fixed. This iteration process is very effective: 2-3 iterations are sufficient. In table 4.2 a conservative estimate of the expected accuracy after the last iteration, for each processing step is given.

A number of data reduction tasks, in addition to the ones mentioned before, are dedicated to *photometry* (magnitude determination), *calibration*, *double stars* and *minor planets*. The magnitude of the program stars is determined in two steps. In the first step, which coincides with the first step of the three step procedure, the star magnitudes are computed from the

IDT data collected over the RGC-set. In the second step the individual magnitude estimates obtained for one star are combined. This is done independently from the three step procedure, and is a so-called off-line task. The calibration, minor planets and double star tasks are also off-line data reduction tasks. Other tasks, closely related to the FAST data reduction procedure are the so-called *preparation & reception of data* and *preparation of iteration* tasks.

Table 4.2 - Expected accuracy (conservative) for different phases of the data reduction [*Perryman, 1985*]

Task	Parameter	Expected Accuracy
raw data treatment	star mapper transit	100 mas (B=9)
	IDT grid phase	15 mas for 1 sec obser-tion (B=9)
attitude reconstitution	three-axis attitude	100 mas on all three axis
great circle reduction	star abscissa along a fixed reference great circle	6 mas (B=9, geometric)
sphere reconstitution	positions, parallax proper motion	2 mas (B=9) 2 mas/year (B=9)

4.3.2 Attitude Reconstitution

During the *attitude reconstitution* the *three-axis* attitude of the satellite is computed from the star mapper transit times. In FAST each of the attitude components is represented as a trigonometric series up to the 15th order for *each* revolution, plus some functions which model the effects of gas jet actuations. The unknown coefficients of these functions, in total about 200 for each angle, are estimated in weighted least squares sense from the star mapper transit times of some 1000 bright stars. Each passage of a star through the star mapper field of view results in two star mapper observations, one for the vertical and one for the inclined slit system. There are ~7 star mapper observations per star, because a star is observed on the average in 1.7 scans (see table 4.1) and in both fields of view. Therefore, the total number of star mapper transit times is about 7000.

The star positions could be determined at the same time, but of course, this is only useful if the star positions can really profit from this determination. This is in general not the case, except when the data is treated for the first time with the original INCA input catalog (the first year of data). These improved star positions simplify the slit number determination during the great circle reduction (see sec. 4.3.3). In all other cases the star positions are fixed. More precisely, after one iteration of the data reduction process an improved star catalogue, with errors of 10 mas or better, and a better along scan attitude, computed during the great circle reduction with an accuracy of the order of 2-15 mas, are available. These cannot be improved any further by the attitude reconstitution, and are therefore not adjusted. The expected accuracy of the attitude reconstitution is given in table 4.3. In cases where the catalogue errors are large (first treatment) a 10th order instead of a 15th order trigonometric series is used.

Table 4.3 - Expected accuracy of the attitude reconstitution from [*Donati et al., 1986a, Belforte, 1986b*]

Parameter	Expected Accuracy
Star mapper transit (input)	100 mas (B=9)
Accuracy with negligible catalogue errors:	
Along scan attitude component	29 mas rms
Transversal attitude components	37 and 63 mas rms
Accuracy with catalogue errors of 1":	
Along scan attitude component	100 mas rms
Transversal attitude components	400 and 600 mas rms
Star positions (if improved)	140 mas rms (abscissae) 420 mas rms (ordinates)

In NDAC a slightly different procedure is adopted. First, NDAC uses also the gyroscope readings, and secondly in NDAC the attitude is modelled by a B-spline series. It can be shown that the gyroscope readings do not improve the accuracy of the attitude, and therefore they are not used within FAST [*Donati et al., 1986a*]. The modelling of the attitude by B-splines is somewhat similar to the approach used for the attitude smoothing during the great circle reduction. The two attitude models which are used within FAST (one based on trigonometric functions and the other on B-splines) are discussed in detail in chapter 6.

4.3.3 Great Circle Reduction

The *great circle reduction* forms a geometric adjustment problem on the sphere, with grid coordinates as observations and with three types of unknowns: attitude and star abscissae along the RGC, and some instrumental coefficients. In fact two types of attitude are produced. At first a *geometric* attitude is estimated, consisting of attitude abscissae at mid frame times (one per 2 seconds). Later these abscissae are *smoothed* to form a continuous representation using B-splines (one per 2 minutes). Smoothing of the attitude improves also the quality of the star abscissae, but if the degree of smoothing is too high systematic errors are introduced. The adjustment problem is solved in a weighted least squares sense. The unknowns consist of ~2000 star abscissae, ~18000 (geometric) respectively ~600 (smoothed) attitude parameters per RGC, and ~50 instrumental parameters. On the average 4.5 stars are simultaneously visible in the field of view, so there are about 80,000 grid coordinates (observations). The error in the grid coordinates is dominated by the photon noise on the phase estimates, so the observations are assumed to be uncorrelated. The expected accuracy of the unknowns is given in table 4.4.

Table 4.4 - Expected accuracy of the parameters computed during the great circle reduction

Parameter	Expected Accuracy
IDT grid phase (input):	10 mas for 1 s observation (B=9)
Star abscissa along a fixed reference great circle:	4 mas in geometric mode (B=9) 3 mas in smoothing mode (B=9)
Along scan attitude:	7 mas in geometric mode 3 mas in smoothing mode
Large scale instrumental distortion:	0.5 mas (corner of the field of view)

The great circle reduction results suffer from two types of indeterminacies:
- the star ordinates and transversal attitude components cannot be improved from the one-dimensional main grid data because of the small inclination of the scanning circles with respect to the RGC,
- the abscissae are determined up to an arbitrary zero-point only, *i.e.* the system to be solved has a rank deficiency of one.

These indeterminacies are overcome in the next stages of the data reduction, when the one-dimensional information collected on many RGC's is combined and processed together. During the great circle reduction the rank deficiency is solved provisionally by forcing the abscissa correction of one star, the so-called *base-star*, to zero. Afterwards the solution is transformed into the *minimum norm*, *i.e.* minimum variance or, barycentric solution, by shifting the solution such that the sum of the corrections to the star abscissae becomes zero.

Apart from the two indeterminacies discussed earlier there is yet another one: the main instrument is only able to observe phases and cannot determine the integer slit numbers. The integer slit numbers are computed from the approximate values for the star positions and the star mapper attitude, which are essentially based on the same positions. Given a slit period of 1"2 and uncertainties in the approximate values of 0"4 - 1"0 at the start of the reduction, there will be many errors in the computed slit numbers, resulting in inconsistencies throughout the reduction. Special algorithms, both during the great circle reduction and astrometric parameter extraction, are designed in order to correct these grid step inconsistencies (see chapter 9).

4.3.4 Sphere Reconstitution

In the sphere reconstitution, the second step of the three step procedure, the unknown zero-points on the RGC's and the astrometric parameters of some 40,000 primary stars are computed by a weighted least squares adjustment. More specific, from the first year of data a *static* solution is computed, with positions only. Later follow *dynamic*, complete solutions with proper motions and parallaxes as well. In general bright, unproblematic, stars are primary and faint, doubtful, stars are secondary.

The abscissae of primary stars, computed during several great circle reductions on well inclined RGC's, are the main input. Over the whole mission (2.5 years) there are about 10^6 primary star abscissae, and at least 200,000 star unknowns (for each primary star 5) and ~1800 unknown RGC zero-points. At the same time approximately 10-20 global parameters are solved. The global parameters model certain instrumental effects, like periodic basic angle variations and constant chromaticity, which cannot be estimated during the great circle reduction (see chapter 5). The accuracy of the estimated astrometric parameters and zero-points are given in table 4.5.

Table 4.5 - Expected accuracy of the parameters computed during the sphere reconstitution: 40,000 stars, 5 mas rms on the abscissae (input) [*Bucciarelli et al., 1986*]

Parameter		Expected Accuracy
position:	β	1.5 mas
	$\lambda \cos\beta$	2.2 mas
proper motion:	μ_β	1.9 mas/year
	$\mu_\lambda \cos\beta$	2.7 mas/year
parallax:		2.1 mas
zero point:		1.2 mas

The RGC-poles are not solved, but kept fixed. This certainly introduces a bias in the first catalogues, but this bias will disappear after a few iterations. Actually, the RGC-poles cannot be estimated with a better accuracy from the RGC-abscissae than their approximate values "determined" during the attitude reconstitution. In fact, the *values* of the poles are fixed from the very start, but the *actual* poles shift slightly over the celestial sphere with new iterations as better star positions become available for the attitude reconstitution. So, in this sense the position of the RGC poles are determined mainly by the attitude reconstruction. A rough guess is that the error in the final RGC-poles is of the order of a few mas, on the other hand, if they are estimated during the sphere reconstitution the error is ~15 mas [*Van Daalen et al., 1986c, Donati, 1986b*].

The sphere reconstitution problem is solved in weighted least squares sense using the LSQR iterative solution method [*Tommasini et al., 1983, 1985a*]. The observation weights are determined from the variances of the star abscissae computed during the great circle reduction. The correlation between the abscissae is neglected, *i.e.* the weight matrix is diagonal. This is maybe the worst approximation of all in the three step procedure, because this cannot be recovered by iterating. Fortunately, it does not introduce a bias or systematic error. It makes a difference wether the abscissae variances are computed from the minimum norm solution *or* the one base star solution. We believe that the minimum norm variances should be used, because there the correlations are the weakest. Ordinarily, if the inverse of the co-variance matrix is used as the weight matrix, there is no difference.

The rank defect in the sphere reconstitution is an interesting problem. Arcwise measurement on the sphere is invariant under rotations. Therefore, one would suspect that the rank defect of the static solution is 3 and that the dynamic solution has a rank defect of 6. This rank defect can be overcome by imposing additional constraints, just sufficient to provide the missing information. A possible choice of constraints is to fix the position and proper motion of "one and a half" stars at zero: *i.e.* one star is constrained in both directions, the other in just one, but one must be somewhat careful here. Other choices of constraints are also possible, *e.g.* 6 zero-points, 3 at the beginning and 3 at the end of the mission, can be fixed. However, simulation experiments show that the sphere reconstitution is non-singular and can be solved well without imposing extra constraints [*Lindegren et al., 1985c, Bucciarelli et al., 1986*]. Theoretically, in the case of non-infinitesimal proper motions and parallaxes the whole rank deficiency disappears, but the proper motions and parallaxes are so small that they introduce the orientation quite weakly [*Van Daalen et al., 1986c*]. Apparently, this effect is not responsible for the non-singular simulation experiments. In the sphere reconstitution the RGC-poles were fixed, and it turns out that they determine the reference system and are responsible for the non-singularity. The cumulated orientation information from 1800 RGC's, although the orientation information contained in a single RGC pole is not very accurate (~15 mas), gives an overall determination of the orientation of the order of 0.3 mas. There is a serious danger of over-constraining which may lead to distortions, especially if additionally "one and a half" stars are fixed.

4.3.5 Astrometric Parameter Extraction

The secondary stars are treated after the sphere reconstitution by the *astrometric parameter extraction*. Star by star the, on the average ~70, star abscissae are collected and the 5 astrometric parameters are solved. The main input are the abscissae of secondary stars computed during the great circle reduction and RGC zero-points computed during the sphere reconstitution. The secondary stars do not influence each other and they do not affect the primary stars and the zero-points. The accuracy of the estimated astrometric parameters are given in table 4.5.

Special attention is given to the problem of grid step ambiguities, which already occurred during the great circle reduction. In the great circle reduction the data belonging to one RGC-set is made -internally- consistent, but the abscissae may still have an error of one or more slit periods. These remaining errors have to be removed during the astrometric parameter extraction. During the sphere reconstitution, in principle, nothing has to be done to the grid step ambiguities, because 1) the abscissae of primary stars will have fewer grid step ambiguities, and 2) these ambiguities will disappear when the data reduction is iterated. The number of grid step ambiguities in the primary stars is smaller because these stars have better initial (INCA) catalogue positions, and because most of them were already used, and possibly improved, in the attitude reconstitution. Of course, the sphere reconstitution converges only in the presence of a few remaining grid step ambiguities.

4.4 Discussion of the Three Step Procedure

4.4.1 Introduction

The complexity of the three step procedure may seem puzzling at first. In fact, the three step procedure hinges on two different principles:

(1) approximate adjustment in steps,
(2) block (Gauss-Seidel) iterative solution method.

In the Hipparcos data reduction the observations are partitioned into RGC sets, which are processed separately in a local -intermediate- reference frame, resulting in many parallel adjustment steps. Each RGC set is processed in two steps, *viz.* the attitude reconstitution and great circle reduction. The outcomes of the parallel RGC adjustments are combined in the sphere' reconstitution and astrometric parameter extraction. This procedure is an adjustment in steps, but it is not done in an rigorous manner:
1) the non-significant observation material in a step is neglected,
2) not all the unknown parameters in a step are actually solved, but some are fixed on their approximate values,
3) the outcomes of a preceding step are treated as uncorrelated quantities by the next steps, *i.e.* the correlation introduced by each step is neglected.

In each step some of the unknowns are fixed on their approximate values. Therefore, it is necessary to iterate the adjustment, so that the different groups of unknowns are solved alternately. In fact, this corresponds to a block (Gauss-Seidel) iterative solution method, the second principle on which the reduction hinges. In our case the block Gauss-Seidel solution converges fast, just two or three iterations are sufficient.

In an adjustment in steps the outcomes of each step are treated as observations in the next step. In the *approximate* adjustment in steps the correlation of the outcomes, which is usually introduced by a preceding step, is not taken into account. This results in some loss of information, which cannot be recovered by an iteration process. Fortunately, due to the Gauss-Seidel iteration process, most of the outcomes enter the next steps as approximate values and not as observations, hence, the correlations play no role. The abscissae computed by the great circle reduction are an exception; they are used in the sphere reconstitution and astrometric parameter extraction as observations. Therefore, by not taking into account the correlation of the abscissae, some loss of information occurs.

In each of the steps the non-significant observation material is neglected. This is done very carefully, so that no noticeable loss in accuracy for the astrometric parameters occurs. Actually, the complete treatment of the IDT and star mapper data can be separated: The main instrument gives only information about the coordinates in the along scan direction, and not in the other direction: Hence, the transversal attitude components and the star ordinates in the intermediate RGC reference frame have to be determined from the star mapper measurements. On the other hand, the accuracy of the star mapper is not sufficient to contribute significantly in the determination of the along scan coordinates. During the attitude reconstitution and great circle reduction it is, because of the short time span (10 hours), not necessary or possible to solve the proper motions or parallaxes.

In the sphere reconstitution and astrometric parameter extraction also some data is neglected: The abscissae of secondary stars do not participate in the determination of the unknown zero points on the RGC and the astrometric parameters of the primary stars, and the RGC poles are not solved. Certainly some loss of information occurs. It is expected that about

40% of the stars survive the reduction as primary stars. Primary stars are generally bright and unproblematic stars, while, on the other hand, secondary stars are usually faint, doubtful, stars, *e.g.* double stars. Therefore, about 80% - 90% of the observation weight will be contained in the primary star abscissae, and so the loss of information remains acceptable. On the other hand, secondary stars are sometimes very doubtful objects, and therefore it is fortunate that they do not influence each other and they do not influence the primary stars.

Below we will take a look at the complete system. First we consider an approximate system of equations, from which we can separate the treatment of IDT and star mapper data. Then we will have an look at the block Gauss-Seidel solution of this system. Finally, an intermediate reference frame is introduced, and, when combined with the block Gauss-Seidel solution method, the three step procedure emerges. The discussion in this section does not concentrate on how it is done, but more on how it could be done, in such an way as to clarify certain aspects of the three step procedure.

4.4.2 Separation of IDT and Star Mapper data

The linear observation, or correction, equations are computed in the usual way from the truncated Taylor expansion of the non-linear equations (4.8), (4.9), (4.10) and (4.15), around some approximate values for the unknowns. The linearized observation equations, in matrix notation, for the complete reduction are

$$\begin{bmatrix} \Delta y^I \\ \Delta y^S \end{bmatrix} = \begin{bmatrix} A_b^I & A_a^I & A_s^I \\ A_b^S & A_a^S & A_s^S \end{bmatrix} \begin{bmatrix} \Delta x_b \\ \Delta x_a \\ \Delta x_s \end{bmatrix} \qquad (4.17)$$

where Δy^I and Δy^S are two vectors with the observed value of the grid coordinate (IDT phases) and star mapper transit time respectively, minus a value computed from approximate values for the unknowns. Δx_a and Δx_b contain the unknown corrections to the along scan attitude parameters and the two attitude components which give the scan circle pole. Δx_s is the vector with corrections to the longitude, latitude, proper motion and parallax of the stars. The design matrix, A, contains the partial derivatives $\partial y/\partial x$ as usual. The design matrix blocks A_a, A_b and A_s have respectively 1, 2 and 5 (2 if proper motions and parallaxes are not considered) non-zero elements per row. The instrumental unknowns have been omitted in these equations, since they are not very relevant for the discussion of the three step procedure.

The attitude is represented by numerical functions, so that the star mapper and IDT measurements, which refer to different times, can be linked. The most simple representation of the attitude parameters is a first order B-spline centered at the mid-frame times, which is equivalent to the geometric attitude representation. More advanced attitude representations are introduced for the purpose of attitude "smoothing". The attitude of the satellite, *i.e.* the orientation of the instrumental frame, is defined by three rotations. The two transversal components define the position of the scanning plane pole, the Z-axis of the instrument frame. The third rotation is around the Z-axis, and it defines the position of the X-axis along the scanning circle. Actually, this is just one possible choice of attitude angles. But what is important, is that with this choice the the coefficients in A_b^I are close to zero (Later, after we have introduced an intermediate

reference frame, we shall see that the choice of transversal attitude parameters is not very critical).

There are two types of star mapper transit times, namely for the vertical and inclined slit system. Therefore, $\mathbf{y}^s$ can be partitioned in a part with observations from the vertical slit, $\mathbf{y}^{sv}$, and, a part with observations from the inclined slit, $\mathbf{y}^{si}$, and the design matrices can be partitioned correspondingly. The magnitude of the non-zero coefficient in the design matrix blocks are given in table 4.6.

Table 4.6 - Design matrix coefficients

	$\mathbf{x}_b$	$\mathbf{x}_a$	$\mathbf{x}_s$
$\mathbf{y}^I$	$<10^{-2}$	~ -1	$[-1,+1]$
$\mathbf{y}^{sv}$	$<10^{-2}$	~ -1	$[-1,+1]$
$\mathbf{y}^{si}$	$\sim \pm 0.5$	$\sim \pm 0.5$	$[-1,+1]$

The normal equations, when partitioned correspondingly to the linearized observation equations (4.17), can be written in matrix notation as

$$\begin{bmatrix} \mathbf{N}^I_{bb}+\mathbf{N}^S_{bb} & \cdots & \cdots \\ \mathbf{N}^I_{ab}+\mathbf{N}^S_{ab} & \mathbf{N}^I_{aa}+\mathbf{N}^S_{aa} & \cdots \\ \mathbf{N}^I_{sb}+\mathbf{N}^S_{sb} & \mathbf{N}^I_{sa}+\mathbf{N}^S_{sa} & \mathbf{N}^I_{ss}+\mathbf{N}^S_{ss} \end{bmatrix} \begin{bmatrix} \mathbf{x}_b \\ \mathbf{x}_a \\ \mathbf{x}_s \end{bmatrix} = \begin{bmatrix} \mathbf{b}^I_b+\mathbf{b}^S_b \\ \mathbf{b}^I_a+\mathbf{b}^S_a \\ \mathbf{b}^I_s+\mathbf{b}^S_s \end{bmatrix} \qquad (4.18)$$

where $\mathbf{N}_{pq}=\mathbf{A}^T_p\mathbf{W}\mathbf{A}_q$ and $\mathbf{b}_p=\mathbf{A}^T\mathbf{W}\mathbf{y}_p$ for $p,q=a,b,s$, and $\mathbf{W}$ the weight matrix of either the grid coordinates or star mapper transit times. The weights of the observations are inversely proportional to the variances of the observations. The variances of the star mapper transit times are roughly a factor 100 larger than the variances of the grid coordinates. So,

$$(\mathbf{W}^I)_{ii} \gg (\mathbf{W}^S)_{jj} \qquad (4.19)$$

for IDT and star mapper observations of the same star. The normal equations (4.18) may be approximated by

$$\begin{bmatrix} \mathbf{N}^{II}_{bb} & \cdot & \cdot \\ \mathbf{N}^{II}_{ab} & \mathbf{N}^I_{aa} & \cdot \\ \mathbf{N}^{II}_{sb} & \mathbf{N}^I_{sa} & \mathbf{N}^I_{ss} \end{bmatrix} \begin{bmatrix} \mathbf{x}_b \\ \mathbf{x}_a \\ \mathbf{x}_s \end{bmatrix} = \begin{bmatrix} \mathbf{b}^{II}_b \\ \mathbf{b}^I_a \\ \mathbf{b}^I_s \end{bmatrix} \qquad (4.20)$$

It seems plausible, taking into account the results of equation (4.19) and of table 4.6, that the loss of information in the approximate normal equations system is negligible.

The treatment of IDT and star mapper data can be separated. The approximate system can be solved with the block Gauss-Seidel method; the iteration formulae, for $m=0,1,2,\ldots$, are

step 1: solve $\mathbf{x}_b^{(m+1)}$ in least squares sense from

$$\mathbf{A}_b^S \Delta\mathbf{x}_b^{(m+1)} = \mathbf{y}^S - \mathbf{y}^S(\mathbf{x}_b^{(m)}, \mathbf{x}_a^{(m)}, \mathbf{x}_s^{(m)}) \qquad (4.21.a)$$

step 2: solve $\mathbf{x}_a^{(m+1)}$ and $\mathbf{x}_s^{(m+1)}$ in least squares sense from

$$\mathbf{A}_a^I \Delta\mathbf{x}_a^{(m+1)} + \mathbf{A}_s^I \Delta\mathbf{x}_s^{(m+1)} = \mathbf{y}^I - \mathbf{y}^I(\mathbf{x}_b^{(m+1)}, \mathbf{x}_a^{(m)}, \mathbf{x}_s^{(m)}) \qquad (4.21.b)$$

with $\mathbf{y}(.)$ the non-linear relations and $\mathbf{x}_a^{(0)}, \mathbf{x}_b^{(0)}$ and $\mathbf{x}_s^{(0)}$ approximate values for the unknowns. In terms of the normal matrices the iteration formulae are

step 1: solve $\Delta\mathbf{x}_b^{(m+1)}$ from

$$\mathbf{N}_{bb}^S \Delta\mathbf{x}_b^{(m+1)} = \mathbf{b}^S - \begin{pmatrix} \mathbf{N}_{ba}^S & \mathbf{N}_{bs}^S \end{pmatrix} \begin{pmatrix} \Delta\mathbf{x}_a \\ \Delta\mathbf{x}_s \end{pmatrix}^{(m)} \qquad (4.22.a)$$

step 2: solve $\Delta\mathbf{x}_a^{(m+1)}$ and $\Delta\mathbf{x}_s^{(m+1)}$ from

$$\begin{pmatrix} \mathbf{N}_{aa}^I & \cdot \\ \mathbf{N}_{sa}^I & \mathbf{N}_{ss}^I \end{pmatrix} \begin{pmatrix} \Delta\mathbf{x}_a \\ \Delta\mathbf{x}_s \end{pmatrix}^{(m+1)} = \begin{pmatrix} \mathbf{b}_a^I \\ \mathbf{b}_s^I \end{pmatrix} - \begin{pmatrix} \mathbf{N}_{ab}^S \\ \mathbf{N}_{sb}^S \end{pmatrix} \Delta\mathbf{x}_b^{(m+1)} \qquad (4.22.b)$$

with $\Delta\mathbf{x}_a^{(0)}=0$ and $\Delta\mathbf{x}_s^{(0)}=0$.

A block Gauss-Seidel process converges for every positive definite matrix [*Varga*, 1962, p. 77-78]. Fast convergence, however, requires that, during each step, the errors in the uncorrected part only have minor influence on the outcomes. This is the case (i) when the errors are anyhow small with respect to the observation precision, but also (ii) when they have just a small effect on the observations (small coefficient in the observation equations), or (iii) when they average out during the adjustment (small coefficients in the normal equations). In particular during later stages of the reduction the errors in the uncorrected part are small with respect to the observation precision: In the *first step* the errors in $\mathbf{x}_a$ and $\mathbf{x}_s$ are very small, ~7 mas and ~2 mas respectively, compared to the standard deviation of the observations $\mathbf{y}^S$, typically 100 mas. In the *second step* the errors are larger than the observation precision, *i.e.* the error in $\mathbf{x}_b$ is typically 50-100 mas, compared to the observation precision of 10 mas. Fortunately the error in $\mathbf{x}_b$ has only a small influence on the observations because of the small coefficients in $\mathbf{N}_{ab}^S$ and $\mathbf{N}_{sb}^S$. This effect is investigated in chapter 5. So in both cases the contribution of the second part of the right hand sides, which depends on the normal matrix blocks pertaining to the star mapper data, is small. Therefore, the above mentioned iteration process for Hipparcos converges very good. Just 2 or 3 iterations are sufficient.

The second iteration equation (4.21.b), which is also a two by two block partitioned system, can be solved similarly by a block Gauss-Seidel approach. This is actually identical to the approach proposed by Sansò [*Betti et al, 1986b*]. A final iteration with this method is foreseen as an independent check of the three step procedure. In this approach certain approximations and systematic effects are avoided which remain in the three step procedure. Especially the problem with the correlation between the RGC abscissae, which is neglected in the three step procedure, is avoided.

In the three step procedure the above mentioned block Gauss-Seidel process is not carried out integrally. Both the first, 4.21.a, and second, 4.21.b, iteration equation are solved in steps.

4.4.3 Effect of an Intermediate Reference Frame

During the three step procedure the observation vector **y** is partitioned in RGC sets, and the attitude and star unknowns are solved in an intermediate reference frame, defined by the chosen reference great circle (RGC) and an arbitrary origin on the RGC. Let us introduce an intermediate reference frame for an RGC set labelled j, then the linearized relations between the vector of corrections to the attitude and star parameters in the intermediate reference frame and Hipparcos reference frame are

$$\Delta \mathbf{x}_{*|j} = \mathbf{A}_{*,j}\Delta \mathbf{x}_{*} + \mathbf{C}_{*,j}\Delta \mathbf{c}_{j}$$

with

$\Delta \mathbf{c}_j$ vector of corrections to the three parameters describing the orientation of the jth RGC reference frame,

$\Delta \mathbf{x}_{*|j}$ vector of corrections to the unknown parameters in the reference frame belonging to the jth RGC,

and with $\mathbf{A}_{*,j}$ and $\mathbf{C}_{*,j}$ the corresponding matrices with partial derivatives for $*=a,b,s$. Let

$\Delta \mathbf{y}_j$ the vector of the observed value minus computed value for the jth RGC,

$\Delta \mathbf{x}_s$ vector of corrections to the astrometric parameters in the Hipparcos reference frame.

Then we may rewrite the iteration equations (4.21.a) and (4.21.b) as:

step 1: solve $\mathbf{x}_{b|j}^{(m+1)}$ for each RGC in least squares sense from

$$\mathbf{A}_{b|j}^{S}\Delta \mathbf{x}_{b|j}^{(m+1)} = \Delta \mathbf{y}_{j}^{S} - \mathbf{A}_{a|j}^{S}\Delta \mathbf{x}_{a|j}^{(m)} - \mathbf{A}_{s|j}^{S}\Delta \mathbf{x}_{s|j}^{(m)} \qquad (4.23.a)$$

step 2: solve $\mathbf{x}_{a|j}^{(m+1)}$ and $\mathbf{x}_{s|j}^{(m+1)}$ for each RGC in least squares sense from

$$\mathbf{A}_{a|j}^{I}\Delta \mathbf{x}_{a|j}^{(m+1)} + \mathbf{A}_{s|j}^{I}\Delta \mathbf{x}_{s|j}^{(m+1)} = \Delta \mathbf{y}_{j}^{I} - \mathbf{A}_{b|j}^{I}\Delta \mathbf{x}_{b|j}^{(m+1)} \qquad (4.23.b)$$

step 3: solve $\mathbf{x}_{s}^{(m+1)}$ and $\mathbf{c}^{(m+1)}$ in least squares sense from all equations

$$\mathbf{A}_{s,j}\Delta \mathbf{x}_{s}^{(m+1)} + \mathbf{C}_{j}\Delta \mathbf{c}_{j}^{(m+1)} = \Delta \mathbf{x}_{s|j}^{(m+1)} \qquad (4.23.c)$$

for m=0,1,2,.. . Here, $\mathbf{A}_{*|j}\mathbf{A}_{*,j}=(\mathbf{A}_*)_j$ is a small part of the design matrices of equation (4.21), corresponding to the jth RGC set. In the first two steps a relatively small system of equations has to be solved several times. The third step has to be solved only once per iteration. Contrary to eq. 4.21, we use in each iteration the same approximate values; therefore it is necessary to correct the linearized observations (in the right hand sides of the equations) for the improvements brought about by previous iterations. Now let us partition $\mathbf{x}_{s|j}$ as

$$\mathbf{x}_{s|j} \longrightarrow (\mathbf{x}'_{s|j}, \mathbf{x}''_{s|j})$$

where $\mathbf{x}'_{s|j}$ consists of the RGC abscissae and $\mathbf{x}''_{s|j}$ of the RGC ordinates, proper motions and parallax. Let the design matrix $\mathbf{A}_{s|j}$ be partitioned correspondingly. Because of the limited duration of the RGC and the small inclination of the scanning circles with respect to the RGC the non-zero elements in $\mathbf{A}''_{s|j}$ will be very small compared to the non-zeroes in $\mathbf{A}'_{s|j}$, which are close to one (see chapter 5). Therefore we can rewrite (4.23) as

$$\begin{aligned}
\mathbf{A}^S_{b|j}\Delta\mathbf{x}^{(m+1)}_{b|j} &= \Delta\mathbf{y}^S_j - \mathbf{A}^S_{a|j}\Delta\mathbf{x}^{(m)}_{a|j} - \mathbf{A}^S_{s|j}\Delta\mathbf{x}^{(m)}_{s|j} \\
\mathbf{A}^I_{a|j}\Delta\mathbf{x}^{(m+1)}_{a|j} + \mathbf{A}'_{s|j}\Delta\mathbf{x}'^{(m+1)}_{s|j} &= \Delta\mathbf{y}^I_j - \mathbf{A}^I_{b|j}\Delta\mathbf{x}^{(m+1)}_{b|j} - \mathbf{A}''_{s|j}\Delta\mathbf{x}''^{(m)}_{s|j} \qquad (4.24) \\
\mathbf{A}'_{s,j}\Delta\mathbf{x}^{(m+1)}_s + \mathbf{C}'_j\,\Delta\mathbf{c}^{(m+1)}_j &= \Delta\mathbf{x}'^{(m+1)}_{s|j}
\end{aligned}$$

where we have omitted the equations $\Delta\mathbf{x}''_{s|j} = \mathbf{A}''_{s,j}\Delta\mathbf{x}_s + \mathbf{C}''_j\Delta\mathbf{c}_j$. Similarly, let us partition $\mathbf{c}_j$ as

$$\mathbf{c}_j \longrightarrow (\mathbf{c}^-_j, \mathbf{c}^\perp_j)$$

with $\mathbf{c}^-_j$ the correction to the assumed origin of the jth RGC and $\mathbf{c}^\perp_j$ the corrections to the assumed positions of the jth RGC pole. Let the design matrix $\mathbf{C}'$ be partitioned correspondingly as

$$\mathbf{C}'_j \longrightarrow (\mathbf{C}^-_j, \mathbf{C}^\perp_j)$$

The coefficients of the design matrix $\mathbf{C}^\perp$ are much smaller than those of $\mathbf{C}^-$. It can be shown that $\mathbf{c}^\perp_j$ cannot be estimated with sufficient precision [*Van Daalen, 1986c*]. Therefore, corrections to the pole are fixed to zero. Therefore the third equation in (4.24) becomes

$$\mathbf{A}_{s,j}\Delta\mathbf{x}^{(m+1)}_s + \mathbf{C}^-_j\Delta\mathbf{c}^{-(m+1)}_j = \Delta\mathbf{x}'^{(m+1)}_{s|j} - \mathbf{C}^\perp_j\Delta\mathbf{c}^{\perp(m)}_j$$

The parameters $\Delta\mathbf{c}^\perp_j$ are not estimated in the adjustment. The value of $\Delta\mathbf{c}^\perp_j$ is forced to zero by the iteration process, *i.e.* any physical change in the RGC poles is absorbed in the next iteration by a systematic rotation of the attitude parameters.

In the three step procedure the observations equations are linearized in each iteration, using the results of a previous iteration as approximate data. Then, the iteration equations are for $m=0,1,2,\ldots$

AR: solve $\mathbf{x}_{b|j}^{(m+1)}$ for each RGC in least squares sense from

$$\mathbf{A}_{b|j}^{s}\Delta\mathbf{x}_{b|j}^{(m+1)} = \mathbf{y}_j^s - \mathbf{y}_j^s(\mathbf{x}_a^{(m)},\mathbf{x}_b^{(m)},\mathbf{x}_s^{(m)}) \qquad (4.25.a)$$

GCR: solve $\mathbf{x}_{a|j}^{(m+1)}$ and $\mathbf{x}_{s|j}^{(m+1)}$ for each RGC in least squares sense from

$$\mathbf{A}_{a|j}^{I}\Delta\mathbf{x}_{a|j}^{(m+1)} + \mathbf{A}'_{s|j}\Delta\mathbf{x}'^{(m+1)}_{s|j} = \mathbf{y}_j^I - \mathbf{y}_j^I(\mathbf{x}_a^{(m)},\mathbf{x}_b^{(m+1)},\mathbf{x}_s^{(m)}) \qquad (4.25.b)$$

SR: solve $\mathbf{x}_s^{(m+1)}$ and $\mathbf{c}^{(m+1)}$ in least squares sense from all equations

$$\mathbf{A}'_{s,j}\Delta\mathbf{x}_s^{(m+1)} + \mathbf{C}_j^{-}\Delta\mathbf{c}_j^{-(m+1)} = \Delta\mathbf{x}'^{(m+1)}_{s|j} \qquad (4.25.c)$$

We are now left with a step wise adjustment embedded in an iterative solution scheme. In the step wise adjustment the correlations created in one step should be taken into account in the other step. *I.e.* in the three step procedure the correlation of $\Delta\mathbf{x}'_{s|j}$, introduced by the great circle reduction, should be taken into account in the sphere reconstitution. This is, however, not done in the three step procedure.

CHAPTER 5

GREAT CIRCLE REDUCTION

During the great circle reduction star abscissae are computed, by a single least squares adjustment, in an intermediate reference frame from contiguous batches of data of about 10 hours duration. Three types of unknowns are solved: star abscissae, along-scan attitude parameters and instrumental parameters. In this chapter the estimability, precision and systematic errors of the unknowns are discussed.

5.1 Introduction

In the great circle reduction contiguous batches of data, gathered during about 5 revolutions (10 hours), are processed together. The star abscissae and the attitude component along a reference great circle (RGC), chosen "somewhere in the middle" of the five scanning circles, are computed by a weighted least squares adjustment. No effort is made to estimate the star ordinates and the other two attitude components. They are badly estimable due to the small inclination of the scanning circles w.r.t. the RGC. Not estimable at all is the zero point of the abscissae. Both indeterminacies are overcome in the next steps of the data reduction, when the abscissae on many intersecting RGC's are combined and processed together.

Three types of unknowns have to be solved in a weighted least squares sense from the ~70.000 grid coordinates observed per RGC: the ~1800 star abscissae, forming our prime objective, the along scan attitude and some instrumental parameters. Two types of attitude unknowns are computed: First ~18,000 geometric attitude parameters, one parameter per observation frame of 2.13 s. Later this geometric attitude is smoothed to form a continuous representation of the attitude using only ~600 parameters. Smoothing of the attitude also improves the quality of the star abscissae, although excessive smoothing may introduce systematic errors (see chapter 6). Statistical tests will validate the degree of smoothing. Some 50 instrumental coefficients are included in the adjustment. Although this is only a relatively small number, their estimation is time consuming since they enter into each observation equation.

The main inputs into the great circle reduction task come from the phase extraction and attitude reconstruction tasks, and from an (external) star catalogue. The attitude and star catalogue data, which are used in the linearization of the great circle reduction equations and to replace non-estimable star and attitude components, are treated as approximate values in the adjustment. The grid coordinates, which are computed by the phase extraction task (see chapter 3) and which are referring to the mean instant of the frame, are treated as observations in the least squares adjustment. The error in the grid coordinates is dominated by photon noise. Hence the errors depend strongly on the magnitude of the star and may be assumed to be uncorrelated. Therefore, in the least squares adjustment a diagonal weight matrix may be used.

The star abscissae are the main outcome of the great circle reduction, they are further processed in subsequent stages (sphere reconstruction and astrometric parameter extraction) of the data reduction. Since one RGC refers to data of at most half a day, only instantaneous positions, for a reference epoch somewhere in the middle of the RGC, are solved and no proper motions or parallaxes. Although for most stars the geometric positions, *i.e.* their position, proper motion and parallax, are constant during a RGC period, this is not true for the apparent, or observed, positions. Aberration and relativistic effects, which depend on the velocity and position of the satellite and on the observed star direction, cannot be ignored. They can, however, be computed with sufficient accuracy from a-priori data and corrections can be applied. For some very near and fast stars the geometric positions do vary during one RGC. The necessary corrections again can be calculated from a-priori data. Some minor planets are on the observing list; their geometric positions are certainly not constant during one RGC and they will be solved with independent positions, every time they appear in a frame.

During the great circle reduction we do not distinguish between primary and secondary stars, as explained in chapter 4, but rather between *active* and *passive* stars. The abscissae of primary stars form the basis for the sphere reconstruction: the abscissae of secondary stars are only treated by the astrometric parameter extraction. The distinction between *active* and *passive* stars is internal to the great circle reduction (NDAC does not distinguish between active and passive stars, instead they do an iterative reweighting of observations). Grid coordinates of active stars participate in the rigorous least squares adjustment which computes the abscissae of active stars, along-scan attitude and instrumental parameters. The passive stars are added in later, using the previously computed active star abscissae, attitude and instrumental parameters, without modifying them. In general, passive stars are "problem" stars, *viz.* stars with a high probability of erroneous measurements, or very faint stars, which anyhow do not contribute very much to the attitude and instrumental solution. Passive stars, in combination with *internal* and *external* iterations, play an important role in the testing and validation of our observations.

Two types of iterations are needed: *external* iterations of the complete great circle reduction with an improved attitude description and star catalogue after a preliminary sphere reconstruction and astrometric parameter extraction, and *internal* iterations within the great circle reduction itself. Internal iterations are needed for the testing and validation of our observations; skipping doubtful observations, correction of slit number errors, generating diagnostics, etc.. Depending on the outcome of statistical tests data is validated, *i.e.* accepted as sufficiently conforming to the model, or rejected. In the latter case not only proper diagnostics must be generated, but also a new solution without the rejected, and possibly erroneous, observations has to be computed. Like in any weighted least squares adjustment with a substantial fraction of erroneous measurements this procedure has to be iterated several times, each time involving a new least squares adjustment. This is, of course, not a very attractive prospect for a large adjustment problem such as the great circle reduction. However, the burden is lightened considerably by a-priori selection of suspected problem stars (our passive stars), combining internal iterations with necessary (for other reasons) external iterations, and special procedures for correcting slit number errors (grid step inconsistency testing).

Grid step ambiguities and inconsistencies are a consequence of the fact that the main instrument is only able to observe phases, *i.e.* grid coordinates modulo the slit period. The integer slit numbers are computed from the a-priori values of the star catalogue positions and the star mapper

attitude, which in itself is based on the same catalogue data. Given a slit period of 1.2" and uncertainties in the a-priori values of 0.4"-1.0" in the initial star catalogue, there will be a substantial number of *grid step errors*, *i.e.* errors in the computed slit numbers, resulting in ambiguities and inconsistencies throughout the reduction. The great circle reduction notices a grid step error only if some of the grid coordinates in the RGC set have a grid step error different from those for other grid coordinates pertaining to the same star. If all observations of a star in the RGC set have the same grid step error this will not be noticed during the great circle reduction. The known magnitude of the grid step errors and the large number of grid step inconsistencies expected during the first treatment of the data (in the external iteration process) justify special algorithms. An advantage is that the magnitude of the errors is known, therefore we may correct our observations rather than reject them, which is not only more efficient, but gives savings in computing time as well.

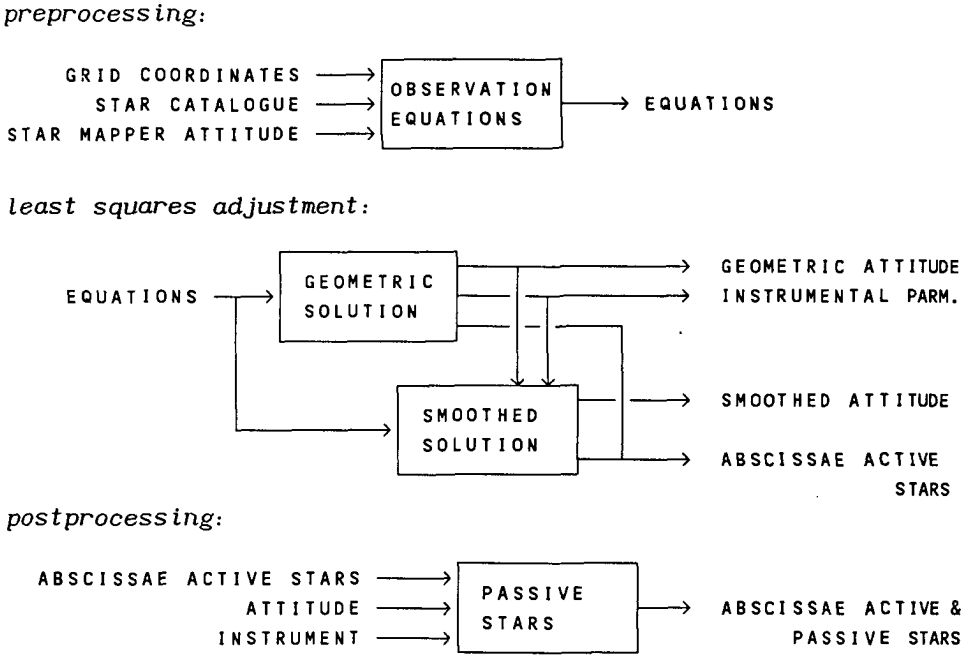


Figure 5.1 - Structure of the FAST great circle reduction

In figure 5.1 the global structure of the FAST great circle reduction software is given. The boxes denote the main tasks to be performed, while the lines indicate the data-streams. Three parts can be clearly distinguished: *preprocessing*, *least squares adjustment* and *postprocessing*. During the preprocessing the active and passive stars are selected, the observation equations are computed and a first grid step inconsistency correction takes place. The abscissae and variances of the active stars are computed during the least squares adjustment, based on the geometric or the smoothed attitude, which is computed simultaneously with the star and instrument solution. During postprocessing the abscissae of the passive stars are computed, final testing takes place and a report is made. Grid step inconsistency correction and testing are not limited to a single process: throughout the great circle reduction inconsistencies will be detected and corrected, and observations will be tested.

This chapter deals with the geometric equations for the great circle reduction, the large scale instrumental parameters and the expected accuracy of the results. In chapter 6 the methods for attitude smoothing, and the resulting improvement in accuracy, are discussed in detail, while the actual solution methods are treated in chapters 7 and 8. Finally chapter 9 deals with the grid step inconsistency problem. In this chapter, and the remaining chapters, we will refer to results obtained by the FAST great circle reduction software on simulated data. In particular results will be given for two datasets: one simulated by the group of prof. Kovalevsky at CERGA, Grasse [*Falin et al., 1986a, 1986b*], and the other one simulated by L. Lindegren of Lund observatory [*Lindegren, 1986*] for the NDAC consortium. Details about the software and the simulations can be found in appendix A and B respectively. For more details about the results of the great circle reduction software on these datasets we refer to [*van der Marel et al., 1986c, 1986d, van der Marel, 1987a*].

5.2 Observation Equations for the Great Circle Reduction

5.2.1 Non-linear Equations

The observation equations for the great circle reduction can be derived from a simple coordinate transformation, at mid-frame time, between the instantaneous measurement frame and the RGC reference frame. Let $\mathbf{u}(\alpha,\beta)$ and $\mathbf{v}(\alpha,\beta)$ be two unit vectors which give the apparent direction to a star in the measurement and RGC reference frame respectively, with

$$\mathbf{u}(\alpha,\beta) = \begin{bmatrix} \cos\alpha\cos\beta \\ \sin\alpha\cos\beta \\ \sin\beta \end{bmatrix}_{instr}, \quad \mathbf{v}(\alpha,\beta) = \begin{bmatrix} \cos\alpha\cos\beta \\ \sin\alpha\cos\beta \\ \sin\beta \end{bmatrix}_{RGC}$$

Let $\mathbf{A}$ be the rotation matrix defined by a 3-2-1 sequence of Euler rotations:

$$\mathbf{A}(\psi,\zeta,\xi) = \mathbf{R}_1(\xi)\cdot\mathbf{R}_2(\zeta)\cdot\mathbf{R}_3(\psi)$$

Here, $\mathbf{R}_i(\alpha)$ is an elementary Euler rotation over α around the i'th axis. The non-linear equations for star i observed in frame k are then

$$\mathbf{u}(x_{ki}\pm\tfrac{1}{2}C, y_{ki}) = \mathbf{A}(\psi_k, -\zeta_k, \xi_k)\ \mathbf{v}(\psi_i^g+\Delta\psi_{ki}^a, \zeta_i^g+\Delta\zeta_{ki}^a) \qquad (5.1)$$

with ψ_i^g and ζ_i^g the abscissa and ordinate of the geometric star position in the RGC frame at the reference epoch, C the basic angle and x_{ki} and y_{ki} are the field coordinates in the measurement frame. The subscript i always refers to a star, and the subscript k refers to a measurement frame (at mid-frame time). The angles $\Delta\psi_{ki}^a$ and $\Delta\zeta_{ki}^a$ are small corrections to the geometric positions, for apparent places (aberration and relativistic effects), for residual proper motion and for parallax. The two attitude angles ψ_k and ζ_k can be viewed as the abscissa and ordinate of the telescope x-axis. Note that the sign of ζ_k is reversed in order to have a consequent definition of ζ_i and ζ_k. See figure 5.2.

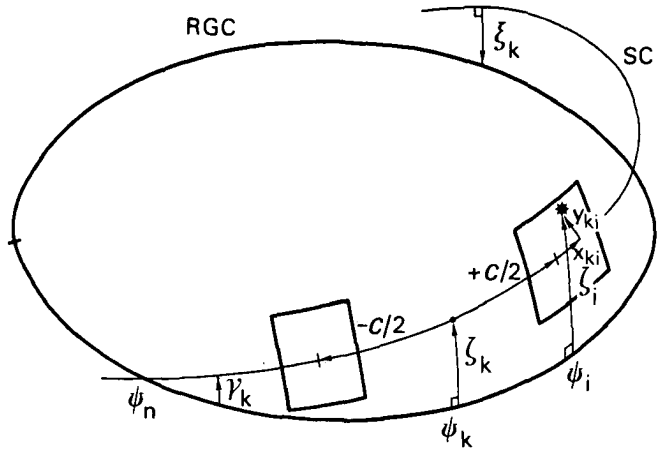


Figure 5.2 - Geometry of the great circle reduction

Only the field coordinate x_{ki} is related to measurements by the main instrument (see section 4.2.3), *i.e.*

$$x_{ki} = -g_{ki} + d_{ki}$$

with g_{ki} the grid coordinate, which is counted in the direction opposite to x_{ki}, and d_{ki} the -unknown- large scale instrumental distortion which is defined in section 5.4. The grid coordinate is computed from the observed phase φ_{ki} ($0 \le \varphi_{ki} < 1$) by equation 4.13:

$$g_{ki} = (n_{ki} + \varphi_{ki})\, s$$

with n_{ki} the integer slit number and s the nominal slit period. It is assumed that the observed phase is already corrected for small and medium scale grid distortions.

5.2.2 Linearization

The linearized observation equation is obtained by taking the truncated Taylor expansion of equation 5.1 around approximate values for the unknowns. The linearized equation is

$$\Delta\underline{x}_{ki} = \frac{\partial x}{\partial \psi_i}\Delta\underline{\psi}_i + \frac{\partial x}{\partial \zeta_i}\Delta\underline{\zeta}_i + \frac{\partial x}{\partial \psi_k}\Delta\underline{\psi}_k + \frac{\partial x}{\partial \zeta_k}\Delta\underline{\zeta}_k + \frac{\partial x}{\partial \xi_k}\Delta\underline{\xi}_k + \qquad (5.2)$$
$$+ \mathbf{c}^T\Delta\underline{\mathbf{d}} + O(\varepsilon^2)$$

with $\Delta\underline{x}_{ki} = \underline{x}_{ki} - x^{\circ}_{ki}$ the observed value of the along scan field coordinate, $\underline{x}_{ki}$, minus the value, x°_{ki}, computed from approximate values for the unknowns, using the non-linear model. The Δ-quantities on the right hand side are the unknown corrections to the approximate values for the unknowns: *viz.* $\Delta\underline{\psi}_i$ and $\Delta\underline{\zeta}_i$ are the unknown correction to the approximate values for the geometric position, $\Delta\underline{\psi}_k$, $\Delta\underline{\zeta}_k$ and $\Delta\underline{\xi}_k$ are the unknown correction to the approximate values for the attitude, and $\Delta\underline{\mathbf{d}}$ is the vector with the unknown correction to

the approximate values for the instrumental parameters. **c** is the vector with partial derivatives $\partial x/\partial \cdot$ for the instrumental parameters. The underscores in equation 5.2 denote stochastic variables.

The partial derivatives for the attitude and star part are computed directly from the formulae **u=Av** rather than by explicit goniometric formulae. The partial derivatives of **u** are computed from

$$\frac{\partial \mathbf{u}}{\partial \cdot} = \frac{\partial \mathbf{A}}{\partial \cdot}\,\mathbf{v} + \mathbf{A}\,\frac{\partial \mathbf{v}}{\partial \cdot} \tag{5.3}$$

The partial derivatives of the spherical angles α and β are computed in turn from the cartesian vector $\mathbf{u}(\alpha,\beta)=(u_1,u_2,u_3)^T$ by

$$\begin{aligned} \frac{\partial \alpha}{\partial \cdot} &= \frac{u_1\,\frac{\partial u}{\partial \cdot}_2 - u_2\,\frac{\partial u}{\partial \cdot}_1}{1-u_3^2} \\ \frac{\partial \beta}{\partial \cdot} &= (1-u_3^2)^{-\frac{1}{2}}\,\frac{\partial u}{\partial \cdot}_3 \end{aligned} \tag{5.4}$$

In case of a 3-2-1 sequence of Euler rotations the computations can be simplified if **A** and **v** are redefined as $\mathbf{R}_1(\xi)\mathbf{R}_2(-\zeta)$ and $\mathbf{v}(\psi_i-\psi_k,\zeta_i)$. Computer evaluation of the above formulae turns out to be much faster than the corresponding evaluation of goniometric formulae or quaternion relations [*Amoureus, 1984*].

In our analysis of the estimability and modelling error we will however frequently need explicit formulae. In table 5.1 the partial derivatives and the approximations up to the second order quantities are given. The angle γ is the inclination of the scanning circle with respect to the RGC (see figure 5.2). Also new abscissae have been introduced to simplify the formulae, *viz.*

$$\psi_{ki}=\psi_i-\psi_k,\quad \psi_{nk}=\psi_k-\psi_n \text{ and } \psi_{ni}=\psi_i-\psi_n,$$

with ψ_n the abscissa of the ascending node of the scanning circle. Some useful relations are

$$\begin{aligned} \cos(\gamma_k) &= \cos(\zeta_k)\cos(\xi_k) \\ \sin(\xi_k) &= -\sin(\gamma_k)\cos(\psi_{nk}) \\ \sin(\zeta_k) &= \sin(\gamma_k)\sin(\psi_{nk})/\cos(\xi_k) \\ \sin(y_{ki}) &= -\sin(\gamma_k)\sin(\psi_{ni})\cos(\zeta_i) + \cos(\gamma_k)\sin(\zeta_i) \end{aligned} \tag{5.5.a}$$

with approximations up to the second order of the small quantities

$$\gamma_k \simeq \sqrt{\zeta_k^2+\xi_k^2} \qquad \xi_k \simeq -\gamma_k\cos(\psi_{nk}) \qquad \zeta_k \simeq \gamma_k\sin(\psi_{nk})$$

$$y_{ki} \simeq \zeta_i - \gamma_k\sin(\psi_{ni}) \tag{5.5.b}$$

These relations, and the partial derivatives given in table 5.1, will be used frequently in this chapter.

Table 5.1 - Goniometric formulae for the partial derivatives

$$\frac{\partial x}{\partial \psi_i} = \frac{1}{\cos^2 y_{ki}} \{ \cos \gamma_k \cos^2 \zeta_i + \sin \gamma_k \sin \zeta_i \cos \zeta_i \sin \psi_{ni} \}$$

$$\simeq 1 - \gamma_k \zeta_i \sin(\psi_{ni}) - \frac{1}{2} \gamma_k^2 \cos(2\psi_{ni})$$

$$\simeq 1 - \frac{1}{2} \gamma_k^2 - \gamma_k y_{ki} \sin(\psi_{ni})$$

$$\frac{\partial x}{\partial \psi_k} = - \frac{\partial x}{\partial \psi_i}$$

$$\frac{\partial x}{\partial \zeta_i} = \frac{1}{\cos^2 y_{ki}} \{ \sin \gamma_k \cos \psi_{ni} \} \simeq \gamma_k \cos \psi_{ni}$$

$$\frac{\partial x}{\partial \zeta_k} = \frac{1}{\cos^2 y_{ki}} \{ \sin \gamma_k \cos \psi_{nk} - \cos \zeta_i \sin y_{ki} \sin \psi_{ki} \}$$

$$\simeq \gamma_k \cos \psi_{nk} - y_{ki} \sin \psi_{ki}$$

$$\frac{\partial x}{\partial \xi_k} = \cos(x_{ki} \pm \tfrac{1}{2}C) \tan y_{ki} \simeq y_{ki} \cos \psi_{ki}$$

5.2.3 Partial Observation Equations

Since the measurements only refer to the along scan component, information on the star ordinate and the transversal attitude components is only available through the inclination γ of the scan circles with respect to the RGC. Because of the small inclination ($|\gamma| \leq 1^\circ$) the transversal components ζ_i, ζ_k and ξ_k are only weakly estimable, and, as we will show in section 5.3, should not be estimated during the great circle reduction.

So, finally, the correction equation (5.2) for Δx_{ki}, the "observed minus computed" value of the grid coordinate of star i in frame k, after linearization around approximate values for the attitude and star abscissae, ψ_k^0 and ψ_i^0, and a -nominal- instrument $\mathbf{d}^0$, becomes

$$\Delta x_{ki} = a_k \Delta \psi_k + b_{ki} \Delta \psi_i + \mathbf{c}_{ki}^T \cdot \Delta \mathbf{d} \qquad (5.6)$$

with $a_k = \frac{\partial x}{\partial \psi_k} \simeq -1$, $b_{ik} = \frac{\partial x}{\partial \psi_i} \simeq 1$ and $|(\mathbf{c}_{ki})_m| \lesssim 1$.

In our software a_k, *i.e.* the partial derivative $\frac{\partial x}{\partial \psi_k}$, will be computed in a way different from that indicated in table 5.1. Namely, we prefer to fix

modelling error in the field coordinates is given by

$$\sigma_{x_{ki}} \sigma_{x_{lj}} = C_s(i-j)\ \sigma_s^2 + C_a(k-l)\ \sigma_a^2 \qquad (5.9.a)$$

with C_s and C_a functions of star and attitude coordinates, computed by the law of propagation of variances on equation 5.8, using the approximate formulae of table 5.1 for the partial derivatives, *i.e.*

$$C(i-j) = \begin{cases} \gamma_k \gamma_l \cos^2(\psi_{ni}) & \text{for } i=j \\ 0 & \text{for } i \neq j \end{cases}$$

$$C(k-l) = \begin{cases} \{\ \gamma_k^2 \cos^2(\psi_{nk}) + y_{ki} y_{kj} \cos^2(\psi_i - \psi_j) + \\ \quad \gamma_k \cos(\psi_{nk})(y_{ki}\sin(\psi_{ki}) + y_{kj}\sin(\psi_{kj}))\ \} & \text{for } |\psi_k - \psi_l| \leq 2f \\ 0 \quad \text{for } |\psi_k - \psi_l| \gtrsim 30^0 \\ \text{unspecified elsewhere} \end{cases} \qquad (5.9.b)$$

with $\sin(\psi_k - \psi_i) \simeq \pm 0.5$ (depending on the field of view) and f half the size of the field of view, *i.e.* f=.45 deg.

The rms and maximum modelling error in the field coordinates, along-scan attitude and star abscissae over an RGC set, as functions of the number of scan circles in the RGC set, are computed below. Thereby the modelling errors are expressed in the form

$$\begin{aligned} rms_{me} &= f \sqrt{a_a \sigma_a^2 + a_s \sigma_s^2} \\ max_{me} &= 3f \sqrt{b_a \sigma_a^2 + b_s \sigma_s^2} \end{aligned} \qquad (5.10)$$

with a and b the error factors. The maximum modelling error is defined as follows: first take the maximum variance, then the square root and finally multiply by three. The error factors a and b for the rms and maximum error, which will be computed in the next paragraphs, are given in table 5.4 as functions of the number of scan circles $2m+1$. In our formulae we assume that the RGC consists of $2m+1$ scans and that the scanning circle precesses by about f per scan, so that the inclination of the scan circles $\gamma = (j+\delta)f$, with $j = -m, \ldots, +m$ and $|\delta| < f/2$.

5.3.1.2 The Modelling Error in the Field Coordinates

In order to get the rms error in the *field coordinates* we take the square root of the average variance of the field coordinates. What we call the maximum modelling error is defined as follows: we first take the maximum variance of the field coordinates, then the square root and finally multiply by three. The variance of the field coordinates (from eq. 5.9) is

$$\gamma_k^2 \cos^2(\psi_{ni})\ \sigma_s^2 + \{\gamma_k^2 \cos^2(\psi_{nk}) \pm \gamma_k y_{ki} \cos(\psi_{nk}) + y_{ki}^2\ \}\ \sigma_a^2$$

Assuming a star in the preceding field of view, the maximum of table 5.4 is clearly realized for $y_{ki} \simeq f/2$ and $\psi_{nk} \simeq -15^\circ$ (so $\psi_{ni} \simeq 15^\circ$). The rms error is

obtained by averaging over the RGC, *i.e.* over k and i. Let <..> denote averages, then from

$$\langle\gamma_k^2\rangle = \frac{1}{3}m^2f^2 \qquad \langle\cos^2\psi\rangle = \frac{1}{2} \qquad \langle y_{ki}\rangle = 0$$
$$\langle y_{ki}^2\rangle = \frac{1}{3}f^2 \qquad \langle y_{ki}\gamma_k\cos\psi\rangle = 0 \qquad \langle\gamma_k^2\cos^2\psi\rangle = f^2\frac{(2m+1)^2}{24} \qquad (5.11)$$

follow the formulae of table 5.4.

5.3.1.3 The Modelling Error in the Along Scan Attitude

An impression of the rms modelling error in the *a-posteriori attitude*, ψ_k, is obtained by evaluating the variance of the mean observation in the frame. Let n be the average number of stars in a frame, and assume that the same number of stars is observed in the preceding and following field of view. Assuming furthermore that the averages $\langle y_{ki}\rangle$ over the preceding and following field of view are zero, then the variance of the mean observation in frame k becomes

$$\gamma_k^2\cos^2(\psi_{ni})\ (\ \sigma_a^2 + \sigma_s^2/n\ )$$

A more elaborate expression for the variance of the mean observation, *e.g.* for different number of stars observed in the preceding and following field, can be found in [*van Daalen, van der Marel, 1987*]. Averaging over ψ and γ gives the formulae from table 5.4. When there is only one observation the error of the mean observation reduces to the error in the field coordinates itself, in which case the maximum error is realized.

5.3.1.4 The Modelling Error in the Star Abscissae

The modelling error in the *star abscissae*, is computed in a way similar to that for the attitude: first we evaluate the variance of the mean observation to a star, then we average over the scans. The covariance of two field coordinates in frame k and l (from eq. 5.9) is

$$\gamma_k\gamma_l\cos^2(\psi_{ni})\ \sigma_s^2 + \delta_{kl}\ \{\gamma_k^2\cos^2(\psi_{nk}) \pm \gamma_k y_{ki}\cos(\psi_{nk}) + y_{ki}^2\}\ \sigma_a^2$$

with $\delta_{kl}=1$ if $\psi_k-\psi_l\leq 2f$, else 0. Here we suppose that the attitude error during one passage of a star is fully correlated, *i.e.* $\delta_{kl}=1$ if the two field coordinates are observed during the same passage. So k and l refer here to passages rather than frames.

Let $\{\gamma\}_i$ be the average inclination on which star i is observed, then the contribution of the error in the *star part*, *i.e.* the influence of ε_{ζ_i}, to the variance of the mean observation becomes

$$\{\gamma\}_i^2\cos^2(\psi_{ni})\ \sigma_s^2 \qquad (5.12)$$

The value of $\{\gamma\}_i$ depends on ψ_{ni} and ζ_i. *E.g.* at the node $\{\gamma\}_i\simeq 0$, because

star i is generally observed in all scans. But away from the node $\{\gamma\}_i$ can be as large as the maximum inclination mf, when star i is at the edge of the RGC and observed in only one passage. So the maximum of this part is simply $m^2f^2\sigma_s^2$, for a star i close to, but not at the node and only observed on a few (one) scan circles.

Now consider the contribution of the error in the *attitude part* to the variance of the mean observation. Assume that the inclination γ and field coordinate y remain constant during two consecutive passages of the field of view. Then, using $\psi_{nk}=\psi_{ni}\pm30^0$, the variance of the mean of two consecutive passages is

$$\left(\frac{1}{2}\gamma^2(\cos^2(\psi_{ni}) + \frac{1}{2} \right) + \frac{1}{2}\gamma\, y \sin(\psi_{ni}) + y^2 \Big)\, \sigma_a^2$$

Averaging over the different scans gives for the contribution of the attitude to the variance of the mean observation

$$\frac{1}{n_i}\left(\frac{1}{2}\{\gamma^2\}_i\cos^2(\psi_{ni}) + \frac{1}{4}\{\gamma^2\}_i + \frac{1}{2}\{\gamma\}_i\{y\}_i\sin(\psi_{ni})+\{y^2\}_i \right)\sigma_a^2 \qquad (5.13)$$

with n_i the number of scans on which star i is observed (here $\{\}$ denotes averaging over the scans). The maximum (see table 5.4) is reached for a star observed on only one scan circle, with $\gamma=mf$, $y=f$ and $\sin(\psi)=1/2m$ ($\sim15^\circ$ for $m=2$, *i.e.* 5 scans per RGC).

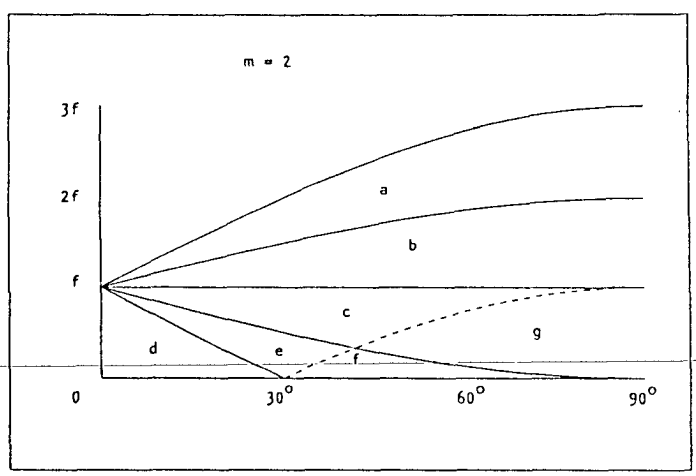


Figure 5.3 - The sectors of the RGC with different "scan patterns"

The rms modelling error in the star abscissae is computed by averaging the contribution of the star and attitude part over the RGC. In order to calculate n_i and functions of $\{\gamma\}_i$, which depend on ψ_{ni} and ζ_i, we partition the RGC in "sectors" for which the stars have the same "scan pattern", *i.e.* the same number of scans n and the same average scan number $\{j\}$. In figure 5.3 the sectors (only the first quadrant is considered) for an RGC consisting of 5 scan circles ($m=2$) are given. The values of n and functions of $\{\gamma\}$ are given in table 5.2, $\int 1$ gives the area of each sector.

Table 5.2 - The observation characteristics of the sectors ($j = \gamma/f$)

Sector	n	$\{j\}$	$\{j\}^2$	$\{j^2\}$	$\int 1$	$\int \cos^2(\psi)$
a	1	2.	4.	4.	1.	.33
b	2	1.5	2.25	2.5	1.	.33
c	3	1.	1.	1.67	.46	.25
d	5	0.	0.	2.	.26	.26
e	4	.5	.25	1.5	.17	.15
f	3	0.	0.	.67	.14	.05
g	2	.5	.25	.5	.54	.08

Careful, sectorwise summation, assuming $\langle\{\gamma\}_i\{y\}_i \sin(\psi_{ni})\rangle=0$ and $\langle\{y^2\}_i\rangle=.33f^2$, finally yields the desired -averaged- coefficients for m=2, given in table 5.3. The coefficients for m=1 and m=3 in table 5.4 have been obtained by a similar approach. The formula in table 5.4 is obtained by fitting a polynomial to the results of table 5.3. In figure 5.4 the error factors a_a and a_s, computed from eq. 5.12 and 5.13, are given over the first quadrant of the RGC for m=2.

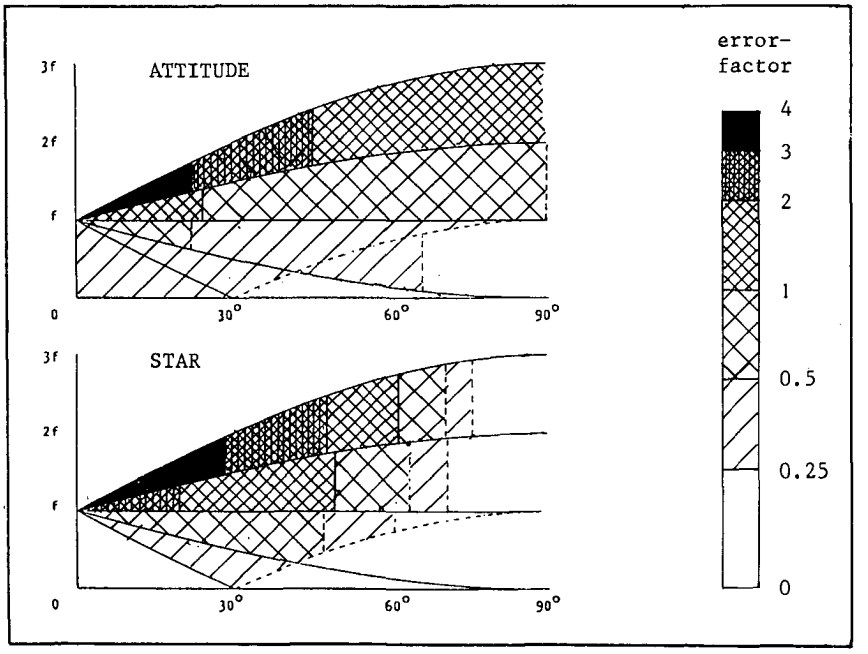


Figure 5.4 - The error factors over the first quadrant of the RGC (m=2)

Table 5.3 - Error factor star abscissae for m=1,2,3

	a_a	a_s
m=1	.45	.19
m=2	.89	.67
m=3	1.5	1.5

5.3.1.5 Conclusions

The error factors a and b for the rms and maximum error computed in the preceding paragraphs are given in table 5.4 as functions of the number of scan circles $2m+1$. The rms and maximum modelling errors follow from equation (5.10):

$$\text{rms}_{me} = f \sqrt{a_a \sigma_a^2 + a_s \sigma_s^2}$$

$$\max_{me} = 3f \sqrt{b_a \sigma_a^2 + b_s \sigma_s^2}$$

If σ_a and σ_b are expressed in tenths of an arc second (0"1) and if the modelling error is expressed in mas; then we may take for $f \simeq .8$ mas/0"1.

Table 5.4 - Maximum and rms modelling error factors

	rms error factor		max error factor	
	attitude a_a	star a_s	attitude b_a	star b_s
field coord.	$\frac{(2m+1)^2}{24} + \frac{1}{3}$	$\frac{(2m+1)^2}{24}$	m^2+m+1	m^2
attitude	$\frac{(2m+1)^2}{24}$	$\frac{(2m+1)^2}{24\,n}$	m^2+m+1	m^2
star	$\frac{(m+1.1)^2+1}{12}$	$\frac{1}{6}m^2$	$\frac{9}{8} + \frac{3}{4}m^2$	m^2

Typical values of σ_a are 0.2" to 0.4" for the first treatment and 0.1" after iterations; σ_b is typically in the range 0.2" to 0.8" for the first treatment and practically zero for the final iteration. So the rms modelling error in the star abscissae will be of the order of 0.6 mas in the final iteration, while it can be as large as 10 mas during the first treatment. These, and similar analytic formulae for the maximum and the rms modelling error in the angle between two stars in a frame, have also been derived in [*van Daalen & van der Marel, 1987*].

5.3.2 Experimental Results on the Modelling Error

The formulae of table 5.4 have been independently verified by an Monte-Carlo simulation and by testruns with the great circle reduction software on three externally simulated datasets [*van der Marel et al., 1986c, van der Marel & van Daalen, 1986d, van der Marel 1987a*]. In table 5.5 the theoretical modelling errors applicable to the Lund data (see section 5.1 and appendix B) are given, with in brackets the experimentally obtained errors. For CERGA dataset II similar results are obtained: for the final iteration we found for the modelling error in the star abscissae 0.77 mas rms and a maximum of 5 mas. In figure 5.5 a scatter diagram of the modelling error in the star abscissae is given for the same data (first treatment). Clearly visible are the predicted maxima near the nodes (situated at ~80° and ~260°), but on the nodes the errors are small.

Table 5.5 - Predicted and experimental (in brackets) modelling errors for the Lund dataset (in mas)

	iteration (σ_a=0"1, σ_b=0)		first treatment (σ_a=0"1, σ_b=1"5)	
	rms	max	rms	max
field coordinate	0.9	6.3	12.0 (11.6)	72.3 (73.1)
attitude	0.8	6.3	6.0	72.3
star	0.6 (.51)	3.3	10.1 (10.7)	72.3 (60.1)

The rms modelling error in the least squares residuals is smaller than the modelling error in the field coordinates: for the Lund data we found 0.36 and 7.8 mas for respectively iteration and first treatment, and for CERGA dataset II 0.54 mas in iteration mode. This error is, as expected, roughly a factor $\sqrt{M/N}$, with M the number of observations and N the number of unknowns, smaller than the modelling error in the field coordinates. The factor is for both datasets ~2. The modelling error in the least squares residuals is large enough to be identified in statistical tests.

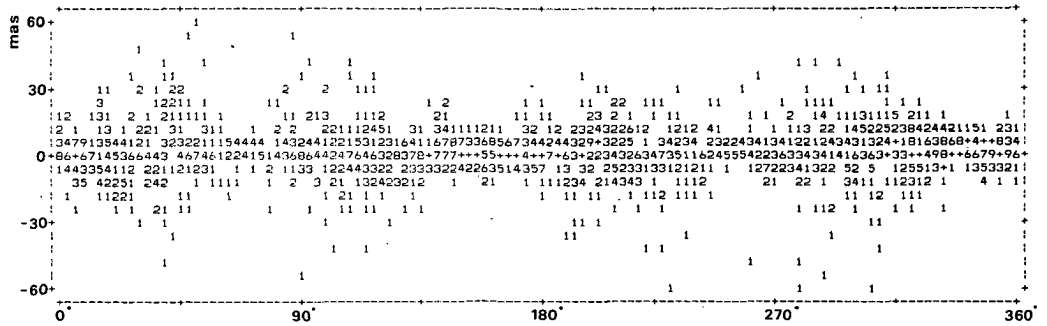


Figure 5.5 - Scatter plot of the modelling error (y-axis) in the star abscissae (x-axis) for the Lund data (first treatment), the numbers give an indication of the number of data points which fall in a character pixel.

5.3.3 Estimability of the Transversal Components

Consider the normal matrix resulting from the approach where both $\{\psi_i, \psi_k\}$ and $\{\zeta_i, \zeta_k, \xi_k\}$ are estimated. Let the normal matrix be block partitioned corresponding to the different types of unknowns. The average diagonal element is, we assume, equal to the average of the squared partial derivatives, multiplied by the number of observations per star or frame. Assume, as before, that the RGC consist of $2m+1$ scans and that the scanning circle precesses by about f per scan, with $f \sim .45^0$ (half the size of the field of view). Integration over the RGC gives for the averages $\langle\rangle$ of the partial derivatives

$$\left\langle \frac{\partial x}{\partial \psi_i} \right\rangle = -\left\langle \frac{\partial x}{\partial \psi_k} \right\rangle \simeq 1 \quad , \qquad \left\langle \frac{\partial x}{\partial \zeta_i} \right\rangle = \left\langle \frac{\partial x}{\partial \zeta_k} \right\rangle = \left\langle \frac{\partial x}{\partial \xi_k} \right\rangle \simeq 0$$

and

$$\left\langle \frac{\partial x}{\partial \psi_i}^2 \right\rangle = \left\langle \frac{\partial x}{\partial \psi_k}^2 \right\rangle \simeq 1 \quad , \qquad \left\langle \frac{\partial x}{\partial \zeta_i}^2 \right\rangle \simeq f^2 \frac{(2m+1)^2}{24} \tag{5.14}$$

$$\left\langle \frac{\partial x}{\partial \zeta_k}^2 \right\rangle \simeq f^2 \frac{(2m+1)^2}{24} + \frac{1}{6} f^2 \qquad \left\langle \frac{\partial x}{\partial \xi_i}^2 \right\rangle \simeq \frac{1}{6} f^2$$

Putting $m=2$ (*i.e.* 5 scanning circles) and $f=.45^\circ$, we get $|\zeta| \leq 1^0$, $|\xi| \leq 1^0$ and $|\zeta - \xi| \leq 0.5^0$. So, the squared partial derivatives are approximately

$$\left\langle \frac{\partial x}{\partial \psi_i}^2 \right\rangle = \left\langle \frac{\partial x}{\partial \psi_k}^2 \right\rangle \simeq 1, \quad \left\langle \frac{\partial x}{\partial \zeta_i}^2 \right\rangle \simeq 6\ 10^{-5}, \quad \left\langle \frac{\partial x}{\partial \zeta_k}^2 \right\rangle \simeq 7\ 10^{-5}, \quad \left\langle \frac{\partial x}{\partial \xi_k}^2 \right\rangle \simeq 10^{-5}$$

The averages belonging to ζ_i, ζ_k and ξ_k are roughly a factor $\lambda^2 = 2\ 10^4$ smaller than the averages belonging to ψ_i and ψ_k. Therefore, on the average, the (diagonal) elements in the corresponding normal matrix blocks are smaller by the same factor.

If we assume that the variances are roughly inversely proportional to the diagonal of the normal matrix, it follows immediately that the variances of the transversal components are $\sim 2\ 10^4$ times larger than the variances of the abscissae. *I.e.* the standard deviations of the transversal components will be of the order of several hundreds of mas. Considering that the transversal components are known a-priori (in the final iteration) with an accuracy better than 100 mas, we can conclude that the transversal components should not be estimated during the great circle reduction. In addition, the condition number of the system of equations becomes very large (roughly a factor $2\ 10^4$ is involved), which is another reason not to pursue this approach.

From equation 5.14 follows, along the same lines as above, that the standard deviation of the transversal components is roughly

$$\sigma_t \simeq \frac{5}{(2m+1)f}\, \sigma_l$$

with σ_l the standard deviation of the along scan components. The turn over point, for a possible use of this approach, is the number of scan circles $(2m+1)$ for which the standard deviation in the transversal components becomes smaller than the rms error in the approximate data for the transversal components. Let us assume that the rms error in the approximate data is 100

mas, and that the standard deviation of the along scan components is 4 mas, then, a very rough estimate for the turn over point, is 25 scan circles per RGC.

5.4 Large Scale Calibration during the Great Circle Reduction

5.4.1 Mathematical Model for the Large Scale Distortion

In the great circle reduction software the large scale field to grid transform,

$$g = G(\ x\ ;\ y, B\text{-}V, \bar{f}, t)$$

equation (4.15) of section 4.2.3, is modelled as a distortion of the grid coordinate, *i.e.*

$$x_{ki} = -g_{ki} + d_{ki}$$

It is assumed that the grid coordinates g, which are computed from the observed IDT (main grid) phase, are already corrected for small and medium scale distortions. The medium scale distortion gives the relative distortion of grid patches, which were written at one go of the grid manufacturing process. During an observation frame, the period over which photon counts are collected for a single phase estimate, the star image has passed several grid patches, each with his own alignment errors. Therefore, it is better that the medium scale corrections are applied before the phase estimate is computed. The same holds, of course, for the small scale distortion of the grid.

The large scale instrumental distortion d must take care of the large scale distortion of the grid, basic angle variations, chromaticity effects and deformation of the optics. Some of these effects are variable over the mission, and/or depend on star colour and temperature gradients of parts of the spacecraft. Fortunately, over one RGC set, these effects can be described by a few (≈50) parameters [*Kovalevsky et al., 1986c*], which can be determined during each great circle reduction. The large scale distortion d is computed from the -general- polynomial model

$$d = \sum_{k=0}^{n_a}\left[\ (B\text{-}V\ -0.5)^k \cdot \sum_{i+j=1}^{n_k}\sum a_{kij(\bar{f})}\bar{x}^i\bar{y}^j\ \right] + \qquad (5.15.a)$$

$$+\ \bar{f}\ \sum_{l=0}^{n_b}\left[\ (B\text{-}V\ -0.5)^l \cdot \sum_{m=0}^{n_l} b_{lm}\bar{t}^m\right]$$

with $\bar{f}$ the field index ($\bar{f}$=+1 for the preceding field, $\bar{f}$=-1 for the following field), $B\text{-}V$ the star colour index, $\bar{x} = x/f$ and $\bar{y} = y/f$ the normalized field coordinates, where f is half the size of the field of view, and $\bar{t} = (t-T)/T$ the normalized RGC time, where T is half the nominal length of the RGC set (*e.g.* 18,000 frames). The normalized coordinates are dimensionless quantities in the interval [-1,+1]. The first part of the distortion polynomial describes deformations over the field of view as function of the star colour. This distortion does not change significantly during one RGC, but it may be different for the two fields of view. The second part, which describes the basic angle deformation, depends on star colour and time of observation. Contrary to the field of view deformation, the basic angle deformation may vary significantly during one RGC (*e.g.* due to thermal effects). Actually, the basic angle deformation cannot be described completely by a polynomial; there are also some sinussoïdal terms, which will be estimated during the sphere reconstitution.

The parameters a_{kij} and b_{lm} in the large scale distortion polynomial are estimated *each* great circle reduction, and for the preceding and following field of view different sets of coefficients a_{kij} are used. Some of the coefficients have a clear physical interpretation: a_{0ij} gives the distortion in the upper right corner of the field pertaining to the power $x^i y^j$, a_{1ij} gives the additional distortion for a star with colour index B-V=1.5, and the coefficients b give the basic angle distortion at the end of the RGC set. In printed form the coefficients are usually expressed in mas.

The choice for a power series is a little arbitrary. Other type of functions are -technically- possible. However, there were no numerical or functional reasons for choosing different types of functions, while power series have some computational advantages. A disadvantage of power series is, however, the possible numerical instability. An indication for the instability of the power series are the large correlations between the estimated coefficients (table 5.6). The stability of orthogonal (Legrende) polynomials is better, but, here, they are not necessary because of the low degrees involved. The degree of the polynomial is dictated by the instrumental specifications and actual behaviour, and can in the software easily be changed by setting some parameters. The present choices for the polynomial degrees are: n_a=n_b=1, $n_{k=0}$=3, $n_{k=1}$=1, $n_{l=0}$=2 and $n_{l=1}$=0 [*Kovalevsky et al., 1986c*].

In the present definition the large scale distortion contains the nominal field to grid transform, which was defined in equation 4.12 as

$$g = - \sin x \cos y$$

This results in some large instrumental parameters, but the nominal field to grid transform can be modelled with sufficient accuracy by a third order polynomial. In our least squares adjustment we will first correct the observed grid coordinates by an approximate instrumental distortion, which also includes the nominal field to grid transform. Therefore, the instrumental unknowns are actually corrections to approximate values for the instrumental parameters.

5.4.2 Vector Notation and Alternative Representations

The large scale distortion d can be written more compactly in vector notation. Define a vector **p**, with elements

$$p_{kij} := (B\text{-}V - 0.5)^k \, \bar{x}^i \bar{y}^j$$

for k=0,...,n_a , i=0,...,n_k , j=0,...,n_k-i and i+j≠0, and let **a** be the vector with the unknown parameters a_{kij} corresponding to the powers in **p**. Define in a similar way the vector **t**, with elements

$$t_{lm} := (B\text{-}V - 0.5)^l \, \bar{t}^m$$

for l=0,...,n_b and m=0,...,n_l and let **b** be the vector with the corresponding parameters b_{lm}. Then the large scale distortion d is

$$D = (\mathbf{p}^T , \bar{f}\, t^T) \begin{bmatrix} \mathbf{a}_{(\bar{f})} \\ \mathbf{b} \end{bmatrix} \qquad (5.15.b)$$

with one set of coefficients **a** for the preceding field and one for the following field, depending on the field index $\bar{f}$. The number of parameters $|\mathbf{p}_{n_0 n_1}|$ for the field of view distortion is computed from

$$|\mathbf{p}_{n_0, n_1}| = \sum_{i=0}^{1} \left\{ \frac{(n_i+1)(n_i+2)}{2} - 1 \right\} \qquad (5.16)$$

The large scale distortion of the field of view, $\mathbf{a}^T\mathbf{p}$, can be developed into a part common to both fields of view, $\mathbf{g}^T\mathbf{p}$, and a part containing half the difference between the two fields of view, $\mathbf{h}^T\mathbf{p}$. The distortion is then

$$d = (\mathbf{p}^T , \bar{f}\, \mathbf{p}^T , \bar{f}\, t^T) \begin{bmatrix} \mathbf{g} \\ \mathbf{h} \\ \mathbf{b} \end{bmatrix} \qquad (5.17)$$

This is the so-called G/H representation, the first representation is called the P/F representation. The parameters g_{kij} *and* h_{kij} from **g** and **h** can be computed from

$$g_{kij} = \frac{(a^p_{kij} + a^f_{kij})}{2} \qquad h_{kij} = \frac{(a^p_{kij} - a^f_{kij})}{2} \qquad (5.18.a)$$

where the upper indices *p* and *f* denote the coefficients for the preceding and following field of view respectively. The coefficients a_{kij} can be computed from g_{kij} and h_{kij} by

$$a^p_{kij} = g_{kij} + h_{kij} \qquad a^f_{kij} = g_{kij} - h_{kij} \qquad (5.18.b)$$

Due to the physics of the distortion the degree of the $\mathbf{h}^T\mathbf{p}$ polynomial may be less than the degree of the $\mathbf{g}^T\mathbf{p}$ polynomial. Therefore, the total amount of parameters needed for the G/H representation may be less than for the P/F representation, *i.e.* for some coefficients we have $a^p = a^f$.

5.4.3 Estimability of the Instrumental Parameters

In this section the estimability of four groups of instrumental parameters is discussed:

1) the *constant terms*: g_{000} and g_{100},
2) the *only-y-terms*: g_{00i}, g_{10i}, h_{00i} and h_{10i}, with $i \geq 1$,
3) the *scale factors*: g_{010}, g_{110}, h_{010} and h_{110},
4) the *basic angle terms*: b_{00} (h_{000}), b_{01},.., b_{10} (h_{100}),.., etc.

It turns out that the first group of parameters is not estimable, therefore the constant terms are not included in the instrumental deformation polynomial **g**. On the other hand h_{000} and h_{100}, which are very well estimable, are not included in **h**, because they appear already in the basic angle deformation polynomial. The other groups are generally estimable, except in some specific situations, *viz.* when the variation in the inclination of the

scan circles is not sufficient or when only a partial scan is observed. In other cases they are generally even estimable on a local scale, provided there are sufficient observations.

Let $\mathbf{A}_a$, $\mathbf{A}_s$ and $\mathbf{A}_i$ be the design matrix blocks belonging to respectively the attitude, star and instrument part, and let $\mathbf{a}_i$ be some column of the design matrix block $\mathbf{A}_i$ belonging to the instrumental unknown x_i of which we want to verify the estimability. Furthermore, assume that the columns in the design matrix block $(\mathbf{A}_a,\mathbf{A}_s)$ are linearly independent, which implies that we have already removed some column in order to cope with the rank deficiency in the problem. We will now investigate the estimability of the instrumental parameters one by one, neglecting the other instrumental parameters at the same time. In the definition of the instrumental parameters, in the preceding section, we have already prevented dependencies among the instrumental parameters themselves. The instrumental parameter x_i is not estimable if the columns of in the design matrix $(\mathbf{A}_a,\mathbf{A}_s,\mathbf{a}_i)$ are -nearly- linearly dependent. Geometrically this means that $\mathbf{a}_i$ is -almost- in the subspace, $R(\mathbf{A})$, spanned by the columns of $\mathbf{A}=(\mathbf{A}_a,\mathbf{A}_s)$. Let δ be the angle between $\mathbf{a}_i$ and $R(\mathbf{A})$. Then we have near dependency if δ is small, *i.e.* the length of the residual vector of $\mathbf{a}_i$ after orthogonal projection on $R(\mathbf{A})$,

$$\bar{\mathbf{a}}_i = (\mathbf{I} - \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T)\mathbf{a}_i$$

is small. Teunissen gives

$$\bar{\mathbf{a}}_i^T\bar{\mathbf{a}}_i = \mathbf{a}_i^T\mathbf{a}_i \sin^2\delta$$

[*Teunissen, 1985, pp. 44-45*]. The variance $\sigma^2_{x_i}$ of x_i is given by

$$\sigma^2_{x_i} = \frac{\sigma^2}{\bar{\mathbf{a}}_i^T\bar{\mathbf{a}}_i}$$

whereby we simply assumed that the observations $\mathbf{y}$ have a covariance matrix $\sigma^2\mathbf{I}$. So, when $\bar{\mathbf{a}}_i$ is short, $\sigma^2_{x_i}$ becomes large. But, when $\bar{\mathbf{a}}_i$ is short, also one of the eigenvalues of the design matrix becomes small, and therefore the numerical condition of the system deteriorates. For very short $\bar{\mathbf{a}}_i$ the design matrix will be ill-conditioned, resulting in numerical instability during the inversion.

Assume there exists an $n_a\times 1$ column vector $\mathbf{f}_a$ and an $n_s\times 1$ column vector $\mathbf{f}_s$ such that

$$\bar{\mathbf{a}}'_i = \mathbf{a}_i - \mathbf{A}_a\mathbf{f}_a - \mathbf{A}_s\mathbf{f}_s \qquad (5.19)$$

is small. Then a lower bound for the variance $\sigma^2_{x_i}$ of x_i becomes

$$\sigma^2_{x_i} \geq \frac{\sigma^2}{\bar{\mathbf{a}}_i'^T\bar{\mathbf{a}}_i'}$$

[*Teunissen, 1985*]. The orthogonal projector $\bar{\mathbf{a}}_i$ follows from minimizing

$$\min_{\mathbf{f}_a,\mathbf{f}_s} \| \mathbf{a}_i - \mathbf{A}_a\mathbf{f}_a - \mathbf{A}_s\mathbf{f}_s \| = \min_{\mathbf{f}_a,\mathbf{f}_s} \| \bar{\mathbf{a}}'_i \|$$

Below we will not always evaluate this minimum. We will be already satisfied with an $\bar{\mathbf{a}}'_i$ close to the minimum, which gives a sufficiently sharp lower bound for the standard deviation of the instrumental parameters. This standard deviation will be used as a criterion for the estimability of the instrumental parameters. An instrumental parameter is not estimable: (i) if the precision by which it can be estimated is worse than our a-priori knowledge of this parameter, or (ii) if the effect of this parameter on the observations is larger than a fraction of a mas. The influence of the error in the instrumental parameter on the observations is simply

$$-\mathbf{a}_i \varepsilon_{x_i}$$

Therefore, the second criterium becomes $(\mathbf{a}_i^T\mathbf{a}_i\sigma_{x_i}^2) \ll \sigma_y^2$, or, $\dfrac{\mathbf{a}_i^T\mathbf{a}_i}{\bar{\mathbf{a}}_i^T\bar{\mathbf{a}}_i} \ll 1$.

Usually we only prove that the above does not hold, using $\bar{\mathbf{a}}'_i$ instead of $\bar{\mathbf{a}}_i$.

The instrumental parameter g_{000}, which represents a constant term common to all measurements, is clearly not estimable. Let $\mathbf{a}_i$ be the column vector with coefficients belonging to g_{000} (all one's), then

$$\mathbf{a}_i - \mathbf{A}_a\mathbf{f}_a = \mathbf{0} \quad , \quad (\mathbf{f}_a)_k = 1/a_k$$

with a_k the coefficient for the attitude from $\mathbf{A}_a$. So the column vector $\mathbf{a}_i$ is a linear combination of a part of the columns of $\mathbf{A}$, and evidently g_{000} is not estimable. The instrumental parameter g_{000} rather defines the internal frame in which the grid coordinates are measured, which by itself is not observable.

The parameter g_{100}, the so-called *constant chromaticity* (a kind of colour dependent offset), is estimable in theory through the variation in inclination of the scan circles. However, considering the small variation in inclination for a regular RGC, it turns out that this term is only badly estimable. This can be shown as follows: Let $\mathbf{a}_i$ be a column vector with the coefficients from the linearized observation equations corresponding to g_{100}, consisting of the colour indices $B\text{-}V$, then a -near- minimum for equation 5.19 is obtained for $\mathbf{f}_a=\mathbf{0}$ and $(\mathbf{f}_s)_j=(B\text{-}V)_j$, with $(B\text{-}V)_j$ the colour index of star j. For a star j observed in frame k the value of $(\bar{\mathbf{a}}'_i)_{kj}$ is

$$(\bar{\mathbf{a}}'_i)_{kj} = (B\text{-}V)_j\,(1 - b_{kj}) \simeq (B\text{-}V)_j \left\{ \frac{1}{2}\gamma_k^2 + \gamma_k\, y_{kj} \sin(\psi_{nj}) \right\}$$

Integration over the RGC gives

$$\bar{\mathbf{a}}'^T_i\bar{\mathbf{a}}'_i = M \langle (B\text{-}V)^2 \rangle \langle (1-b_{kj})^2 \rangle \simeq M \left(\frac{m^4}{20} + \frac{(2m+1)^2}{72}\right) f^4$$

where $2m+1$ is the number of scan circles, f half the size of the field of

view, M the number of observations and $\langle(B-V)^2\rangle \simeq 1$. For $m=2$ and $f=.45^0$, we have $\bar{\mathbf{a}}_i'^T\bar{\mathbf{a}}_i' \simeq .7\ 10^{-8}M$. Assuming 10^5 observations with typical standard deviation of 10 mas, then a lower bound for the standard deviation of g_{100}, neglecting correlations, is ~0.4". The influence on the observations is of the same order of magnitude, this means that estimation of g_{100} in the normal scanning mode is out of the question. The constant chromaticity is, however, estimable during the sphere reconstruction, but also a special calibration device is installed on-board the spacecraft.

The *only-y-terms* g_{00*}, as well, become estimable through the variation in inclination among the scanning circles in the RGC. Let $\mathbf{a}_i$ be the column of the design matrix corresponding to the only-y-term g_{001}. If we take for $(\mathbf{f}_a)_k=\zeta_k/a_k$ and $(\mathbf{f}_s)_j=\zeta_j$ a -near- minimum for equation 5.19 is obtained, with

$$(\bar{\mathbf{a}}_i')_{kj} = y_{kj} - \zeta_k - b_{kj}\zeta_j \simeq \gamma_k(\sin\psi_{nk} - \sin\psi_{nj}) + O(\gamma_k^2\,\zeta_j) + O(\gamma_k\,\zeta_j^2)$$

Since $\psi_{kj} \simeq \pm 30^\circ$ the minimum becomes roughly $\gamma_k(.13\sin\psi_{nk} \pm .5\cos\psi_{nk})$. Integration over the RGC gives

$$\bar{\mathbf{a}}_i^T\bar{\mathbf{a}}_i \simeq f^2\,\frac{(2m+1)^2}{96}\,M$$

Typically for $m=2$, $f=.45^\circ$ and $M=10^5$ $\bar{\mathbf{a}}_i^T\bar{\mathbf{a}}_i$ becomes ~2. Then, assuming a standard deviation of 10 mas for the observations, the lower bound for the standard deviation of the unnormalised coefficient g_{001} becomes ~7 mas. Normalization gives a lower bound for the standard deviation of ~.06 mas (the standard deviation in the upper right hand corner of the field of view), which is at the same time, except for a factor $1/\sqrt{2}$, the influence on the observations. This value corresponds reasonably well to the actually computed value of 0.17 mas for CERGA dataset II (table 5.7). The same type of reasoning applies to the other only-y-terms. So the only-y-terms appear to be reasonably estimable on a normal RGC. However, if the variation in inclination becomes too small, *e.g.* when only one scan is treated, it is better not to estimate the only-y-terms, but to use, or interpolate, the values of previously computed RGC's.

The scale factors g_{010} and g_{110} become generally estimable when a basic angle has been spanned. Then two stars, approximately one basic angle apart, are connected both by direct observation and a chain of small angles between stars in the same field of view. The differential scale factors h_{010} and h_{110} become estimable after the same stretch is crossed by both the preceding and following field of view. Similarly the constant part of the basic angle becomes estimable when a chain of (large) angles has been closed around the circle. If the RGC consists of more than one scan the basic angle can be estimated as a slowly varying function of time. Since each star observed in different scans contributes to these closure conditions, the basic angle deformation is in general very well estimable.

All the other parameters, corresponding to powers of xy, x^2, x^2y, xy^2, x^3, etc., are already estimable on a local scale, *i.e.* from observation within the same frame. The chromaticity terms become only estimable if there is sufficient variation in the star colour indices. Since each star, not

observed in the center of the field, contributes, these parameters will generally be well estimable. The only possible problem that still could occur, because we have, above, considered the terms one by one, is the large correlation between the individual parameters. However, as will be shown by simulation experiments, correlations will be in general smaller than 0.1, except for a few terms given in table 5.6. The influence of the fact that the instrumental parameters have to be estimated on the standard deviations of the star abscissae is almost negligible (0.27 mas for CERGA dataset II; see the next section).

Table 5.6 - Correlations (> 0.1) between the instrumental parameters for CERGA dataset II

	g001	*g002*	*g003*	*g010*	*g011*	*g012*	*g020*	*g021*	*g003*	*b00*	*b10*
g001	1		.75					.32			
g002		1				.13					
g003	.75		1								
g010				1		.42	.12		.81		
g011					1			.19			
g012		.13		.42		1					
g020				.13			1		.24		
g021	.32				.19			1			
g030				.81			.24		1		
b00										1	.15
b10										.15	1

Table 5.7 - Normalized instrumental parameters, standard deviation (σ) and true errors (ε) in mas for CERGA dataset II. Errors exceeding the 2σ bound are marked by an * (For I and PO see text).

	expected value	standard dev. (σ)	ε_I	ε_{PO}	units
g001	0.56	0.17	-1.12 *	-0.72 *	mas
g002	-47.24	0.12	0.03	0.03	:
g003	-0.91	0.22	0.05	0.07	.
g010	4.55	0.13	-0.28 *	-0.02	
g011	81.06	0.09	0.09	0.	
g012	3.12	0.17	0.	0.01	
g020	47.53	0.09	-0.01	-0.01	
g021	-1.38	0.17	-0.11	0.	
g030	1.60	0.17	0.30	0.01	
g101	0.94	0.14	0.11	0.05	.
g110	0.	0.10	0.06	-0.	:
b00	35.17	0.04	0.	-0.	mas
b01	-0.02	0.06	0.	0.01	mas/(6 hours)
b10	-0.07	0.06	0.	-0.	mas/mag

The powers in x, y and *B-V* must be computed from a-priori data. This data may not always be very good, especially during first treatment. This results in errors in the estimated parameters. After some iterations of the complete data reduction chain the errors in x and y, which are computed from the star catalogue and star mapper attitude, will be negligible. However the error in the star colour indices, which are not determined by the Hipparcos data analysis consortia, will not be improved. It is hoped that improved colour indices, computed by the Tycho data analysis consortium using the star mapper data, become available before the final iteration of the Hipparcos data reduction.

The standard deviations and true errors computed from CERGA dataset II are given in table 5.7. The true errors are computed for two cases of iteration data, once with noisy observations (I) and once with perfect observations (PO). The true errors in the run with perfect observations represent the influence of the errors in the approximate data on the instrumental coefficients. In table 5.8 the modelling errors in the only-y-terms for CERGA and Lund data, for the first treatment (F) and for the final iteration (I), are summarized.

Table 5.8 - Modelling errors (in mas) for the only-y-terms (for I and F see text)

	CERGA II (I) σ_a=0"1, σ_s=0"1	Lund (I) σ_a=0"1, σ_s=0"0	Lund (F) σ_a=0"1, σ_s=1"5
g001	-.72	.17	1.77
g002	.03	.04	-.62
g003	.07	-.15	-.72
g101	.05	-.02	-1.18

The modelling error in the *g001* term estimated from CERGA dataset II is much larger than the standard deviation and also much larger than the modelling error for the corresponding term of the Lund data. Statistical tests on the least squares residuals indicate that the observations do conform to the model. Therefore, the reported modelling error in *g001*, is almost certainly not harmful. There are two possible explanations: There has been an error in the conversion of the simulated data from tape to disk, which resulted in a constant bias of ~100 mas on the attitude parameters for CERGA dataset II (this error was only made in the iteration dataset I). Another possibility is that an error has been made in the computation of the true (expected) value, which involves several instrumental transformations, or that a (typing) error has been made in one of the reported true (expected) values. We think that the first explanation is the most likely, but this has not yet been verified by test runs.

5.5 Analysis of the Variances

5.5.1 Results from Simulation Experiments

In our simulation experiments several measures for the accuracy of the star, attitude and instrumental parameters are used, namely the *formal variance*, *true error* and *modelling error*. The true and modelling errors give the difference between the adjusted values and the simulated true values (mathematical expectation). The modelling error, which is a special case of the true error, does not contain the effect of noisy measurements. In particular, it gives the contribution of the error in the approximate values to the error in the final outcomes. The modelling error is computed from simulation experiments with perfect, noiseless, observations. The variances, and more generally the complete covariance matrix, follow from the inverse of the normal matrix. These, formal, variances do not account for model deviations. The variance is one of the characteristic parameters which specify the expected (normal) distribution of errors. The true error, minus the modelling error, should be unbiased and fit this distribution (see appendix B).

Table 5.9 - Square root of the mean variances in mas, per magnitude class, for CERGA dataset II

				--geometric--		--smoothing--	
B	n_B	σ_{obs}	σ_{ins}	σ_{att}	σ_{star}	σ_{att}	σ_{star}
2	4	.11	.28	3.17	3.19	2.04	2.06
3	4	.16	.32	3.25	3.27	2.00	2.03
4	14	.33	.26	3.13	3.16	2.00	2.05
5	46	.41	.25	3.11	3.15	2.00	2.05
6	117	.64	.27	3.19	3.27	2.02	2.13
7	292	.98	.26	3.19	3.34	2.01	2.25
8	714	1.52	.26	3.24	3.59	2.01	2.54
9	532	2.25	.27	3.24	3.97	2.02	3.04
10	120	3.23	.25	3.32	4.64	2.02	3.82
	1843	1.80	.27	3.24	3.71	2.02	2.71

The formal star variances can be separated into three components; 1. the variance σ^2_{obs} when only photon noise, attitude jitter, etc. is taken into account, assuming a perfect attitude and instrument, 2. the influence of the along scan attitude determination σ^2_{att}, and 3. the influence of the determination of the instrumental parameters σ^2_{ins}. The variance of the stars after adjustment is

$$\sigma^2_{star} = \sigma^2_{obs} + \sigma^2_{att} + \sigma^2_{ins} \qquad (5.20)$$

The σ^2_{obs} of a star is computed from the cumulated a priori observation weights to this star and σ^2_{ins} is the difference of the computed star variances with and without solving for instrumental parameters. Finally σ^2_{att} is a derived quantity, computed from the above mentioned variances. In table 5.9 and figure 5.6 the square root of the mean variance, per magnitude class, computed from CERGA dataset II is given.

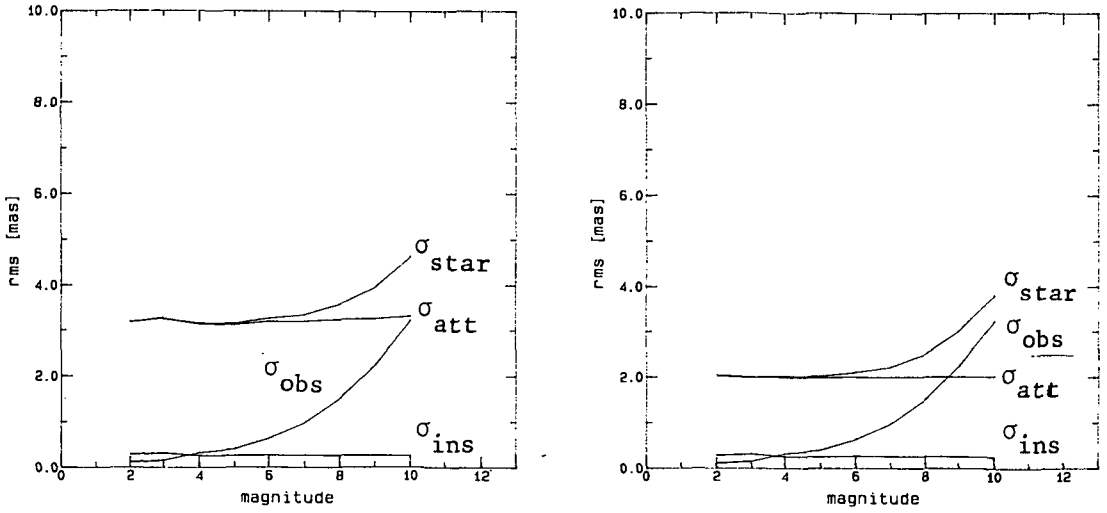


Figure 5.6 - Standard deviation for CERGA dataset II.
left: geometric solution, right: smoothed solution

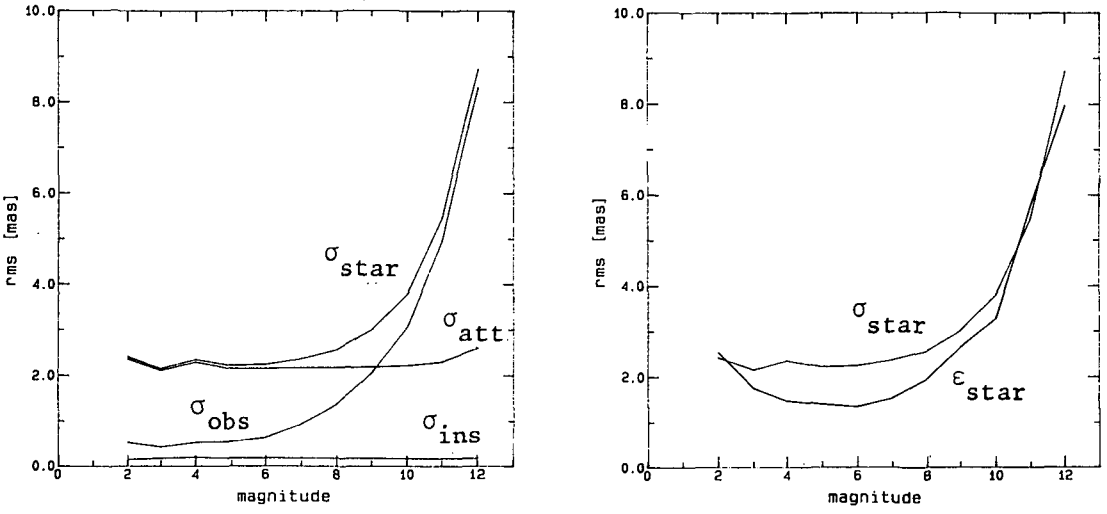


Figure 5.7 - Standard deviations (σ) and true error (ε) for the Lund data (geometric solution)

Observe that the influence of the attitude, 3.2 mas for the geometric solution and 2.0 mas in the smoothed solution, is the same for each magnitude class. The influence of the instrumental parameters, 0.27 mas, which is very small compared to the influence of the attitude, is also the same for each magnitude class. On the other hand the σ^2_{obs} varies between 0.1 mas for very bright stars and 3.2 mas for magnitude 10 stars, and even larger for magnitude 12-13 stars, as can be observed from figure 5.7, where the variances for the Lund dataset are given. In figure 5.7 also the rms of the true error per magnitude class is given.

5.5.2 The Inverse and the Eigenvalues of a Cyclic Symmetric Matrix

For regular star fields an analytic expression for the variances and covariances can be computed. In case of a regular network of stars the normal matrix, after elimination of the attitude parameters, becomes circularly symmetric and may be inverted by Fourier methods.

A cyclic matrix is a square matrix whose rows are, except for a cyclic permutation, identical. Let **A** be a N by N cyclic matrix, then **A** can be written as

$$(\mathbf{A})_{ij} = a(i-j) \qquad (5.21.a)$$

with $a(k)$ the *real* sequence of data $a(0),..,a(N-1)$. The difference $(i-j)$ must be taken modulo N. Now the system of equations

$$\mathbf{A}\,\mathbf{x} = \mathbf{y}$$

can be rewritten as a convolution

$$\sum_{j=1}^{N} a(i-j)\; x(j) = y(i) \qquad (5.21.b)$$

with $x(j)=(\mathbf{x})_j$ and $y(i)=(\mathbf{y})_i$. According to the convolution theorem the discrete Fourier transform of a convolution is the product of the two discrete Fourier transforms, *i.e.*

$$A(j)\cdot X(j) = Y(j) \qquad (5.21.c)$$

with $A(j)$, $X(j)$ and $Y(j)$ the coefficients of the discrete Fourier transform. Hence the solution $x(i)$ of the normal equations can be computed from $X(j)$ by the inverse transform, where $X(j)=Y(j)/A(j)$.

The inverse $\mathbf{A}^{-1}$ of a cyclic matrix **A** is also cyclic. Let

$$(\mathbf{A}^{-1})_{ij} = c(i-j)$$

, then it follows immediately that the coefficients of the discrete Fourier transform of $c(.)$ are equal to $1/A(j)$. Furthermore it is easy to show that the coefficients $A(j)$ of the discrete Fourier transform of $a(i)$ are equal to the eigenvalues λ_j of **A**.

The discrete Fourier transform pair for a *real* sequence of data $x(0),\ldots,x(N-1)$ is defined as

$$\begin{aligned} X(j) &= \sum_{k=0}^{N-1} x(k)\; e^{-i\omega_j k} \\ x(k) &= \frac{1}{N} \sum_{j=0}^{N-1} X(j)\; e^{i\omega_j k} \end{aligned} \qquad (5.22)$$

with $\omega_j=2\pi j/N$. Although $x(k)$ is real, $X(j)$ are generally complex numbers, however $X(j)$ is identical to the complex conjugate $X^*(N-j)$, so only $N/2$ complex numbers are unique. For *cyclic symmetric* matrices **A** an even sequence $a(k)$ is obtained, *i.e.* $a(k)=a(N-k)$. Then the so-called cosine transform

$$X(j) = 2 \sum_{k=1}^{N/2} x(k) \cos\left(\frac{-2\pi jk}{N}\right)$$
$$x(k) = \frac{2}{N} \sum_{j=1}^{N/2} X(j) \cos\left(\frac{2\pi jk}{N}\right) \qquad (5.23)$$

may be used instead of the exponential series and the numbers $X(j)$ are real.

5.5.3 Covariance Function for a Regular Star Network of Uniform Magnitude

A theoretical estimate for the variance of the star abscissae has been derived in [*Høyer et al., 1981*] and [*Burrows, 1982*] for a regular network of stars of the same magnitude. In both approaches the attitude unknowns are eliminated, which results in a relatively small system of normal equations. Because of the regularity of the star network and the uniform star magnitude the normal matrix is circularly symmetric. The normal equations can be written now as a convolution and may be inverted analytically by Fourier methods.

Considering that the coefficients $a_k \simeq -1$ and $b_{ki} \simeq +1$, the observation equations can be simplified to

$$\Delta x_{ki} = \Delta\psi_i - \Delta\psi_k$$

where the instrumental parameters have been neglected, which anyhow have a very small influence on the standard deviations. Let us assume that the observation errors are dominated by photon noise, so they are uncorrelated and the observations weights, w_{ki}, are inversely proportional to the variance of the observations. The normal equations are then

$$\begin{bmatrix} \mathbf{N}_{aa} & \mathbf{N}_{as} \\ \mathbf{N}_{sa} & \mathbf{N}_{ss} \end{bmatrix} \begin{bmatrix} \mathbf{x}_a \\ \mathbf{x}_s \end{bmatrix} = \begin{bmatrix} \mathbf{b}_a \\ \mathbf{b}_s \end{bmatrix} \qquad (5.24)$$

with $(\mathbf{x}_a)_k = \Delta\psi_k$, $(\mathbf{x}_s)_i = \Delta\psi_i$, $\mathbf{b}_a = -\sum_{i \in P_k} w_{ki} \Delta x_{ki}$, $\mathbf{b}_s = \sum_{k \in P_i} w_{ki} \Delta x_{ki}$, $(\mathbf{N}_{aa})_{kk} = \sum_{i \in P_k} w_{ki} \dot{=} w_k$, $(\mathbf{N}_{ss})_{ii} = \sum_{k \in P_i} w_{ki} \dot{=} w_i$, $(\mathbf{N}_{as})_{ki} = -w_{ki}$ iff $(k,i) \in P$ or else zero. The set P is defined by

$$P = \{ (k,i) \mid \text{star } i \text{ observed in frame } k \} \qquad (5.25.a)$$

and the sets P_k and P_i, are defined by

$$P_k = \{ \{i\} \mid (k,i) \in P \}$$
$$P_i = \{ \{k\} \mid (k,i) \in P \} \qquad (5.25.b)$$

give respectively the stars observed in frame k and the frames in which star i is observed. Furthermore, let $m_k = |P_k|$ be the number of stars observed in frame k, $n_i = |P_i|$ be the number of observations to star i and $M = \sum_k m_k = \sum_i n_i$ the total number of observations and N the number of stars.

Elimination of the attitude gives the reduced normal equations

$$\bar{\mathbf{N}}_{ss}\,\mathbf{x}_s = \bar{\mathbf{b}}_s \tag{5.26}$$

with

$$(\bar{\mathbf{N}}_{ss})_{ij} = w_i\delta_{ij} - \sum_{k\in P_i\cap P_j}\frac{w_{ki}w_{kj}}{w_k}$$

$$(\bar{\mathbf{b}}_s)_i = \sum_{k\in P_i} w_{ki}\Delta\bar{x}_{ki}$$

where $\delta_{ij}=1$ iff $i=j$, else $\delta_{ij}=0$ and $\Delta\bar{x}_{ki} = \Delta x_{ki} - \Delta x_k$, with

$$\Delta x_k = \sum_{j\in P_k}\frac{w_{kj}\Delta x_{kj}}{w_k}$$

the average observation in a frame.

The reduced normal matrix for the star part is negative for every element except the diagonal. Let us define a diagonal matrix **D**, with $(\mathbf{D})_{ii}=w_i$, then $\mathbf{N}_{ss}$ can be decomposed as **D-AD**, with $\mathbf{A}=\mathbf{I}-\mathbf{D}^{-1}\mathbf{N}$. Obviously $(\mathbf{A})_{ij}\geq 0$ and so the spectral radius of **A** is bounded by one [*Burrows, 1982*]. Then the star solution may be computed from the following -Jacobi- iteration formula

$$\mathbf{x}_s^{(i)} = \mathbf{D}^{-1}\bar{\mathbf{b}}_s + \mathbf{A}\,\mathbf{x}_s^{(i-1)} \tag{5.27}$$

with

$$(\mathbf{A})_{ij} = \frac{1}{w_i}\sum_{k\in P_i\cap P_j}\frac{w_{ki}w_{kj}}{w_k}$$

Written out in full

$$\Delta\psi_i = \frac{1}{w_i}\sum_{k\in P_i} w_{ki}\Delta\bar{x}_{ki} + \frac{1}{w_i}\sum_{j=1}^{N_s}\sum_{k\in P_i\cap P_j}\frac{w_{ki}w_{kj}}{w_k}\Delta\psi_j \tag{5.28}$$

Equation 5.28 is equivalent to Burrows' iteration formula [*Burrows, 1982, eq. 2.1.11, p. 5*].

Now assume a regular star network, with $2m$ stars observed per frame, of which m stars fall in the preceding field of view and m in the following field of view and assume equal observation weights $w_{ki}=\bar{w}$. So $w_k=2m\bar{w}$ and $w_i=n\bar{w}$, with n the number of observations per star. Assume further that the basic angle C corresponds to an integer g, with $C = 2\pi N/g$. In this particular case the normal matrix $\mathbf{N}_{ss}$ and iteration matrix **A** are cyclic symmetric, with for $(\mathbf{A})_{ij}=a(i-j)$

$$a(i) = \begin{cases} \frac{1}{2m}\left(1 - \frac{|i|}{m}\right) & |i| \leq m \\ \frac{1}{4m}\left(1 - \frac{|i\pm g|}{m}\right) & |i\pm g| \leq m \\ 0 & \text{else} \end{cases} \tag{5.29}$$

The Fourier coefficients $A(j)$ of the sequence $a(i)$ may be computed in the following way: let $F_1(j)$ be the Fourier coefficients of an elementary "hat"

function $f_1(i)$, with $f_1(0)=1$ and $f_1(i)=0$ for $|i|>m$. Obviously $a(i)$ consists of three scaled, shifted, hats. So, using the property that the Fourier transform of a shifted function $f'(t)=f(t-t_0)$ is equal to $F'(\omega)=e^{-it_0\omega}F(\omega)$, the transform of $a(i)$ becomes, after some rewriting,

$$A(j) = \frac{1}{m}\cos^2(\pi jg/N)\ F_1(j) \tag{5.30}$$

The hat function may be computed as the convolution of two sampled periodic block functions. The discrete Fourier transform of the sampled block function, defined by

$$f_0(i) = \begin{cases} 1 & \text{for } |i \bmod(N)| \le m/2 \\ 0 & \text{else} \end{cases}$$

is

$$F_0(j) = \frac{\sin(\pi jm/N)}{\sin(\pi j/N)}$$

Then, according to the convolution theorem, the discrete Fourier transform of $f_1(i)$, when properly scaled, is

$$F_1(j) = \frac{1}{m}F_0(j)\cdot F_0(j) = \frac{\sin^2(\pi jm/N)}{m\sin^2(\pi j/N)} \tag{5.31}$$

Combining the results of eq. 5.30 and 5.31 gives for the eigenvalues $\lambda(j)=A(j)$ of **A**

$$\lambda(j) = \frac{\sin^2(\pi jm/N)\cos^2(\pi jg/N)}{m^2\sin^2(\pi j/N)} \tag{5.32}$$

which is, except for different definitions of m, identical to the result obtained in [*Burrows, 1982, Høyer et al., 1981*]. The eigenvalues $\lambda(j)$, $j=1,..,N/2$, of **A** are plotted in figure 5.8.

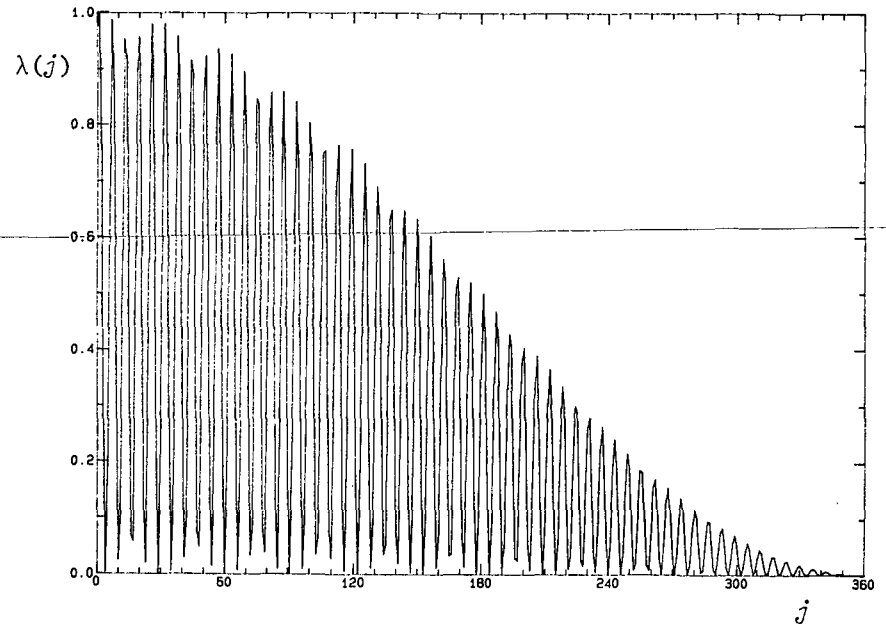


Figure 5.8 - Eigenvalues of the iteration matrix **A**, for $N=720$, $m=2$ and $C=58^\circ$.

Actually, we are not interested in the eigenvalues of **A** themselves, but they are only a instrument in the process of calculating the covariances of the star parameters. Let $\mathbf{C}= \sigma_0^2 \mathbf{N}_{ss}^{-1}$ be the covariance matrix of the star parameters, with σ_0 the standard deviation of unit weight. **C** is also cyclic symmetric, so $\mathbf{C}_{ij} = c(i-j)$, with $c(i-j)$ the so-called covariance function. The covariance function $c(k)$ is computed from the inverse transform of the eigenvalues $1/\lambda_n$ of $\mathbf{N}^{-1}$, with λ_n the eigenvalues of $\mathbf{N}_{ss}$ and λ_n equal to $n\bar{w}(1-\lambda_a)$ where λ_a are the eigenvalues of **A**. Then

$$c(k) = \bar{\sigma}^2 \frac{2}{N} \sum_{j=1}^{N/2} \frac{1}{(1-\lambda(j))} \cos\left(\frac{2\pi jk}{N}\right) \qquad (5.33)$$

is the covariance function of the star parameters, with $\lambda(j)=A(j)$ the j'th eigenvalue of the iteration matrix **A** and $\bar{\sigma}^2=\sigma^2/n\bar{w}$. The summation starts at 1, because $\lambda(0)=1$ corresponds to the rank defect in **A** (*viz.* the corresponding eigenvalue in the normal matrix is zero). Omitting $\lambda(0)$ from the inverse transform gives the same results as if we would have computed the minimum norm solution. The covariance function is plotted in figure 5.9 for $\bar{\sigma}^2=1$.

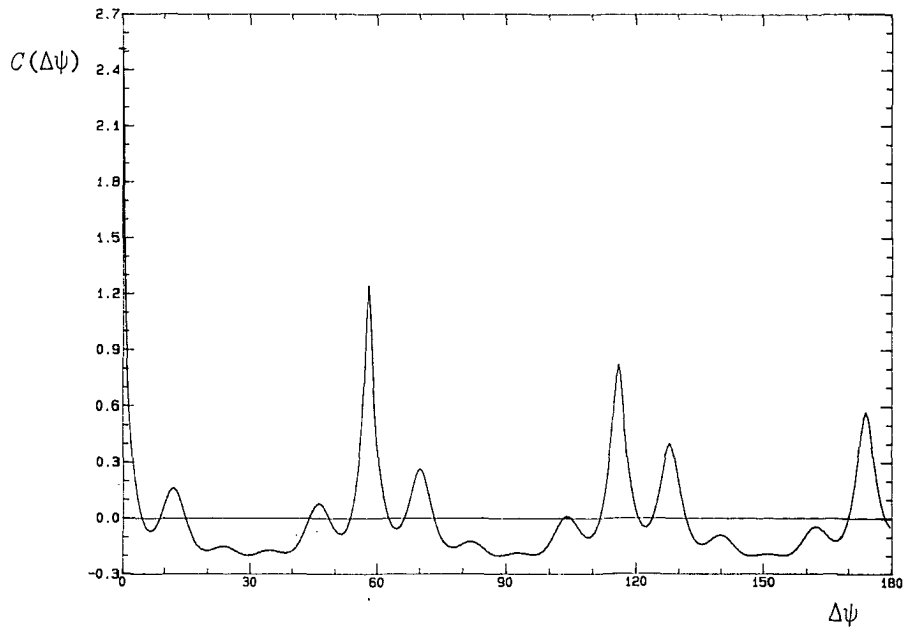


Figure 5.9 - Covariance function of the star parameters, for N=720, m=2 and C=58°.

The variance of the abscissae, σ^2_{star}, is given by $c(0)$,

$$c(0) = \bar{\sigma}^2 \frac{2}{N} \sum_{j=1}^{N/2} \frac{1}{(1-\lambda(j))} = \bar{\sigma}^2 R \qquad (5.34)$$

where N is the number of stars and $\bar{\sigma}^2=\sigma^2_0/n\bar{w}$, the hypothetical variance of a star, assuming a perfect attitude knowledge (σ^2_{obs}). The second part of the expression is the so-called rigidity factor R, which is defined as

$$R = \frac{\sigma^2_{star}}{\sigma^2_{obs}} = \frac{\text{variance after adjustment}}{\text{variance assuming a perfect attitude and instrument}}$$

The rigidity factor can be considered as a measure for the geometric strength of the network of measurements. The rigidity factor as a function of N is plotted in figure 5.10, for different values of the basic angle C and size of the field of view f.

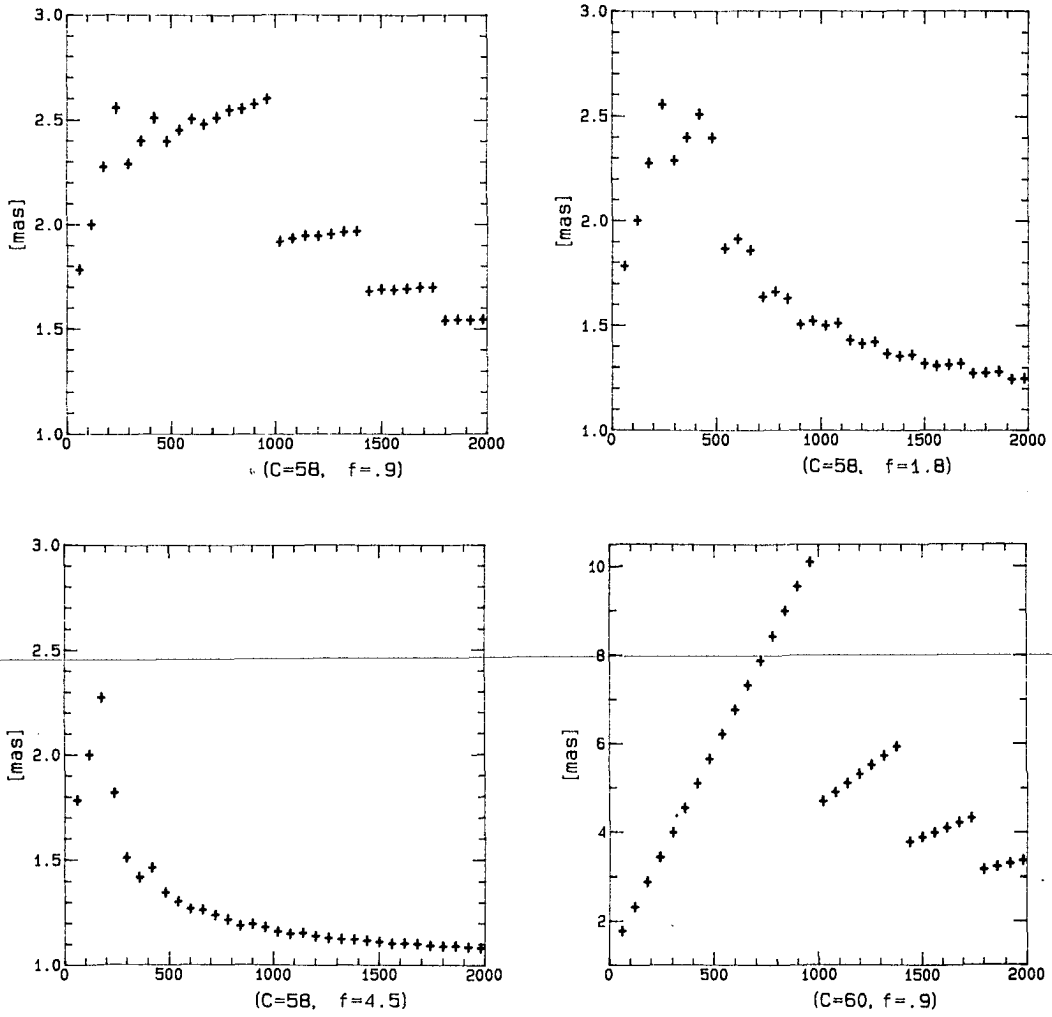


Figure 5.10 - Rigidity (R) versus number of stars (N) for different values of C and f

The values for m are computed as follows,

$$m = Round(\frac{Nf}{2\pi}), \quad m \geq 2 \tag{5.35}$$

which accounts for the jumps in the rigidity in figure 5.10 (*e.g.* around N=1000, 1400 and 1800 for f=.9°). Figure 5.10, as a measure for $c(0)$, *i.e.* σ^2_{star}, can be misleading, because the rigidity should be multiplied by $\bar{\sigma}^2$, which is increasing with N:

$$\bar{\sigma}^2 = \frac{2\ m}{T4\ n} <\sigma^2_{t=1}> = \frac{Nf}{T4\ \pi n} <\sigma^2_{t=1}> \simeq \frac{N}{852\ n} <\sigma^2_{t=1}> \tag{5.36}$$

with $<\sigma^2_{t=1}>$ the mean variance of the field coordinate for one second observing time and $T4$ the frame period (~2.13 s.). The results in figure 5.10 for f=1.8° and f=4.5° are more or less representative for the improvement in the star abscissae due to attitude smoothing (the value of $\bar{\sigma}$ should be computed with f=0.9°, since the actual field is not changed). Further, it will be clear from figure 5.10 that C=58° is a better choice for the basic angle than C=60°.

Table 5.10 - The, mean observed and theoretical (regular), variance and rigidity for the CERGA and Lund data.

	CERGA dataset II		Lund dataset	
	observed	regular	observed	regular
N	1843	880	1964	960
n	40.7	18	40.0	18
m	2.2		2.4	
$\sigma_{t=1}$ (mas)	8.0		10.5	
$\bar{\sigma}$ (mas)	1.8	1.9	2.5	2.6
R	4.2	2.5	1.8	2.6
σ_{star} (mas)	3.7	3.0	3.3	4.2

We should be very careful in comparing these theoretical results with the results obtained from simulation experiments. Two, not very realistic, assumptions have been made in the derivation of these theoretical results:

a. a regular scan pattern, *i.e.* each star must be observed the same number of times n,
b. equal observation weights, *i.e.* the magnitude of stars must be the same and stars must be distributed regularly over the RGC band.

In the typical scan pattern of the RGC-set the stars near the scan nodes are observed more frequently (n~90) than the stars at 90° from the nodes (n~18). Actually, the first assumption holds only in case of a RGC set consisting of one scan circle. Therefore, in interpreting the theoretical results we must do as if the RGC consists of one scan circle. *I.e.* instead of N, the number of stars in the RGC set, and n, the average number of times a star is observed, we will use

$$N' = \frac{2\pi}{f}\ m \quad , \quad n' = 18$$

The value of m, the average number of stars visible in one of the fields of view, is not changed. In table 5.10 the theoretical and observed values, for the rigidity and variance are given for the CERGA and Lund data.

The regular rigidity does not fit the experimental data very well. For CERGA dataset II, which contains in this case stars brighter than magnitude 10 only, the theoretical values for the rigidity and star variance are too low. For the Lund data, which contains stars up to magnitude 13, the theoretical values are too high. The rigidity depends very much on the star magnitude: *viz.* in case of the Lund data the rigidity varies from 15 for stars brighter than magnitude 6, to almost 1 for magnitude 13 stars. Evidently, the regular rigidity is not a good measure for the strength for a network of stars with different magnitude.

5.5.4 Variance for a Regular Star Network of Different Magnitudes

In the preceding section we assumed that all observation have equal weights, and therefore both the normal matrix and iteration matrix **A** were circularly symmetric. In this section we will derive an analytical expression for the case where the observation weights are not equal. However, we still assume that stars are observed the same number of times, *i.e.* the scan pattern must be regular.

Again let **D** be the diagonal matrix with $(\mathbf{D})_{ii}=w_i$. Then **N** can be decomposed into

$$\mathbf{N} = \mathbf{D} - \mathbf{DAD} \tag{5.37}$$

with $\mathbf{A} = \mathbf{D}^{-1}-\mathbf{D}^{-1}\mathbf{N}\,\mathbf{D}^{-1}$. Let **A'** be the iteration matrix from the preceding section, then $\mathbf{A}=\mathbf{A}'\mathbf{D}^{-1}$. Obviously $(\mathbf{A})_{ij}\geq 0$, with

$$(\mathbf{A})_{ij} = \frac{1}{w_i w_j} \sum_{k\in P_i\cap P_j} \frac{w_{ki}w_{kj}}{w_k} \tag{5.38.a}$$

Now assume $w_{ki}= w_i/n$ and n, the number of observations per star, is the same for each star. So

$$(\mathbf{A})_{ij} = \frac{1}{n^2} \sum_{k\in P_i\cap P_j} \frac{1}{w_k} \tag{5.38.b}$$

Further assume that the basic angle corresponds to an integer number of stars g and that $w_k= w_a$ is the same for each frame. Then

$$w_a = 2\,\frac{m}{n}\,\langle w_i\rangle$$

with $\langle w_i\rangle$ the mean weight per star. In this particular case **A** is circularly symmetric, with $(\mathbf{A})_{ij}=a(i-j)$ identical to equation (5.29) except for a constant factor $\langle w_i\rangle$. Thus the eigenvalues $\lambda(j)=A(j)$ of **A** are precisely a factor $\frac{2m}{nw_a} = \frac{1}{\langle w_i\rangle}$ smaller than the eigenvalues of eq. (5.32) for Burrow's iteration matrix, which was used in the preceding section.

The inverse of the normal matrix **N** is equal to

$$\mathbf{N}^{-1} = \mathbf{D}^{-1} + \mathbf{A} \sum_{k=0}^{\infty} (\mathbf{DA})^k \qquad (5.39)$$

Now, let us consider the *second part* of this equation. Generally, the product of two cyclic symmetric matrices, which results again in a cyclic symmetric matrix, can be written in terms of a convolution, and the same can be done for the powers of the cyclic symmetric matrix. But, unfortunately, the matrix product **DA**, or more specific, **D**, is generally not cyclic symmetric. **D** is only cyclic symmetric when its diagonal elements are equal. Now, we will use in the second part of the equation **D'** instead of **D**, with $(\mathbf{D'})_{ii}=\langle w_i \rangle \ \forall i$, where $\langle w_i \rangle$ is the average weight per star, and we will use only in the first part of the equation the original weights. Then, the discrete Fourier coefficients $\lambda_i(j)$ for row i of the second part of equation (5.39) are

$$\lambda_i(j) = \lambda_a(j) \sum_{k=0}^{\infty} (\langle w_i \rangle \lambda_a(j))^k = \frac{\lambda_a(j)}{1 - \langle w_i \rangle \lambda_a(j)} \qquad (5.40)$$

with $\lambda_a(j)$ the eigenvalues of the cyclic symmetric iteration matrix **A** and assuming $|\langle w_i \rangle \lambda_a(j)| < 1$ for the second part of the expression. Since $\lambda_a(j) < \frac{2m}{nw_a} = \frac{1}{\langle w_i \rangle}$, our assumption $|\langle w_i \rangle \lambda_a(j)| < 1$ is always fulfilled.

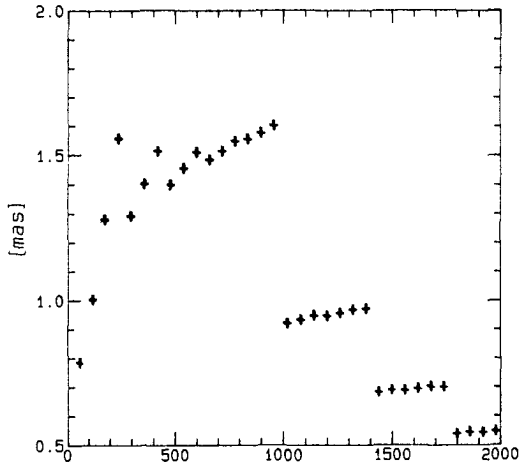


Figure 5.11 - Influence of the attitude for $C=58^0$ and $f=0.9^0$.

For the i'th row of the covariance matrix of the star part we may write $c_i(i-j)=C_{ij}$, with c_i the covariance function for the i'th unknown. Then c_i follows from the inverse Fourier transform of the eigenvalues $\lambda_i(j)$, with

$$c_i(k) = \sigma_0^2 \frac{2}{N} \sum_{j=1}^{N/2} \frac{\lambda_a(j)}{(1-\langle w_i \rangle \lambda_a(j))} \cos\left(\frac{2\pi jk}{N}\right) \qquad k > 0 \qquad (5.41.a)$$

and $c_i(0) = \sigma_{obs}^2 + \sigma_{att}^2$, with $\sigma_{obs}^2=\sigma_0^2/w_i$ the variance assuming perfect

attitude knowledge and σ^2_{att} the so-called *influence of the attitude*, given by

$$\sigma^2_{att} = \sigma^2_0 \frac{2}{N} \sum_{j=1}^{N/2} \frac{\lambda_a(j)}{(1-<w_i>\lambda_a(j))} \equiv \bar{\sigma}^2 R_{att} \qquad (5.41.b)$$

The weights w_i only plays a minor role in σ^2_{att}, therefore the influence of the attitude σ^2_{att} is almost independent from the star under consideration. This effect is also observed in simulation experiments. The influence of the attitude is given in figure 5.11 as function of N, for f=0.9°, C=58° and $\bar{\sigma}^2=\sigma^2_0/<w_i>=1$, so that actually the rigidity factor $R_{att}=\sigma^2_{att}/\bar{\sigma}^2$ is plotted. These theoretical results still do not agree very well with our experimental results: $R_{att}\simeq 2.9$ for CERGA dataset II and $R_{att}\simeq .74$ for the Lund data.

Similar results are obtained if the star parameters are eliminated first. Elimination of the star parameters gives a new normal matrix $\bar{\mathbf{N}}_{aa} = \mathbf{N}_{aa} - \mathbf{N}_{as}\mathbf{N}^{-1}_{ss}\mathbf{N}_{sa}$ for the attitude part. Then, after inversion of $\bar{\mathbf{N}}_{aa}$, which, under certain assumptions, may be done analytically, the covariance matrix of the star part can be computed from (see appendix C)

$$\bar{\mathbf{N}}^{-1}_{ss} = \mathbf{N}^{-1}_{ss} + \mathbf{N}^{-1}_{ss}\mathbf{N}_{sa}\bar{\mathbf{N}}^{-1}_{aa}\mathbf{N}_{as}\mathbf{N}^{-1}_{ss} \qquad (5.42)$$

So

$$(\bar{\mathbf{N}}^{-1}_{ss})_{ij} = \frac{1}{w_i}\delta_{ij} + \frac{1}{w_i w_j}\sum_{k\in P_i}\sum_{l\in P_j} w_{ki}w_{lj}(\bar{\mathbf{N}}^{-1}_{aa})_{kl} \qquad (5.43)$$

where δ_{ij}=1 iff i=j, else δ_{ij}=0. Assume as before that every star is observed in n frames and $w_{ki}=w_i/n$, then with $c_i(i-j)=(\bar{\mathbf{N}}^{-1}_{ss})_{ij}$,

$$c_i(i-j) = \frac{1}{w_i}\delta_{ij} + \sum_{k\in P_i}\sum_{l\in P_j}(\bar{\mathbf{N}}^{-1}_{aa})_{kl} \qquad (5.44)$$

which is more or less the same result as before, *i.e.* the part with the double summation gives the influence of the attitude on the star covariances. Clearly, this is an average over a large number of frames, and it is, therefore, less sensitive for fluctuations in the star magnitude. This holds especially in case of attitude smoothing. So, we believe that σ_{att} is a better measure for the strength of the network than the rigidity. Equation 5.44 can be inverted analytically under certain assumptions; *i.e.* each star must be observed on the same number of scans and the sum of the observation weights over an observation frame must be the same for all frames. This approach is less restrictive on the scan pattern than before, but the scan pattern must still be regular. Therefore, we will not pursue this approach any further.

Our conclusion is that, although analytical formulae give some insight in the covariance structure, they can not yet be used to compute the variances and covariances analytically with sufficient precision to be of any use in the data reduction. Therefore, the (co)variance must still be computed by (partial) inversion of the normal matrix (see chapter 7), or other types of quality indicators, generally functions of the least square residuals, have to be used. Further, we believe that the influence of the attitude is a better measure for the strength of the network than the rigidity.

CHAPTER 6

ATTITUDE SMOOTHING

Smoothing of the along scan attitude during the great circle reduction improves the precision of both the along scan attitude parameters and the star abscissae. In this chapter a number of models for the Hipparcos attitude are evaluated, with emphasis on numerical smoothing with B-splines, and the corresponding improvement of the star abscissae is discussed.

6.1 Introduction

The attitude of the Hipparcos spacecraft is, except for small vibrations (jitter) a smooth function of time. Thus the along scan attitude, which is initially computed per observing frame of 2.13 s., can be further improved by introducing relations between the attitude values of neighbouring frames. In fact an additional adjustment of the along scan attitude, the so-called *smoothing* step, is carried out using a model for the attitude which requires relatively few parameters. The improvement of the attitude leads also to improved star abscissae and instrumental parameters, and hence to an improved final star catalogue.

Two ways of smoothing are distinguished, *dynamical* and *numerical* smoothing. In case of dynamical smoothing we think of an actual dynamical model of the spacecraft, with as unknowns just the initial conditions and some physical parameters describing the torques. In case of numerical smoothing the attitude will be developed in the form of some time series, with as unknowns the coefficients of the series. Smoothing reduces the number of attitude unknowns; hence, it effectively amounts to enlarging the field of view. Within FAST two different approaches for numerical smoothing exist. For the three-axis attitude reconstruction from Star Mapper transit times [*Donati et al, 1986a*] the so-called "semi-dynamical model"[1], developed by Centro di Studi sui Sistemi (CSS) in Torino, is used. In the great circle reduction numerical smoothing of the along scan attitude with *B-splines* has been chosen [*van Daalen & van der Marel, 1986a*].

The B-spline model aims at reconstituting the along scan attitude angle for a complete reference great circle with an accuracy of 2-3 mas, which is a considerable improvement over the so-called "*geometric*" attitude with one parameter per frame. The B-spline model consists of a finite series of shifted B(ase)-splines of fixed degree. The shape of each B-spline depends on the location of the so-called knots. Multiple, overlapping, knots have to be inserted to account for attitude rate jumps due to control pulses (gas jet actuations), and single knots have to be inserted to account for attitude perturbations between the gas jet actuations.

The semi-dynamical model aims at reconstituting the attitude (over one revolution) from star mapper data with a typical accuracy of 100 mas. It consists of a finite series of base functions which gives the - theoretical - response of a rigid body to two types of torques: namely the response to the

[1] Actually CSS calls it "dynamical model". We prefer to call it "semi-dynamical" in order to prevent confusion with true dynamical smoothing.

external and internal disturbing torques (disturbance torque response) and the response to a train of impulses due to the control torques (gas jet response). The measurements from the star mapper (SM), which are used by the attitude reconstruction, are fewer and less precise than the measurements from the image dissector tube (IDT), which form the basis of the along scan attitude computed during the great circle reduction. This accounts for the different results of the attitude models used in the attitude reconstitution and great circle reduction, but it does not tell us anything about the precision of the two models themselves. Also star mapper measurements are only available at irregular times, whereas the IDT measurements are available at regularly spaced times.

In this chapter both the B-spline and semi-dynamical model are evaluated in the context of the great circle reduction. True dynamical smoothing is only briefly touched upon.

6.2 The Hipparcos attitude

6.2.1 The Hipparcos Attitude Motion

The Hipparcos satellite is a three-axes controlled body which follows a nominal scanning law composed of three motions (figure 6.1):

1. spin motion around the body z-axis, perpendicular to the two viewing directions, at a rate ω_d=168.56 arcsec/s (11.25 rev/day),
2. precessing motion of the body z-axis around the satellite-Sun line at a rate of ω_p=0.263 arcsec/s (6.4 rev/year) and with a constant inclination of $\sigma=43^0$
3. yearly motion of the Sun, *i.e.* the satellite-Sun line moves in the ecliptic plane at an average rate ω_s=0.041 arcsec/s (1 rev/year).

The spin and precession rates are actually slightly modulated to maintain a constant scan rate of ω_0=168.75 arcsec/s along the so-called *scanning circle*, the intersection of the viewing plane with the celestial sphere.

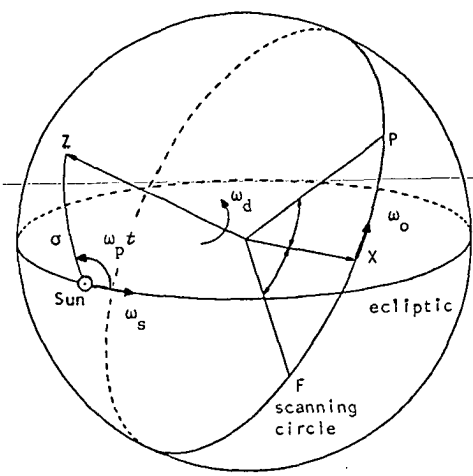


Figure 6.1 - Attitude rates of the nominal Hipparcos scanning law

The maximum permissible deviation of the three satellite axes from the scanning law is 10 arcmin. When one of the axis deviates by more than 10 arcmin., or when the scan rate deviates by more than 2%, gas jets will be fired which will generate control torques on all three axes. The duration of each firing (50-500 ms) is computed for each axis separately, using a control strategy which takes the expected disturbance torques into account and is designed to maximize the time interval between gas jet firings. On the average gas jets will be fired once per 600 s, but at least 400 s apart [*British Aerospace, 1983*]. The effect of such a control strategy is a non uniform sequence of variable length control pulses on all three axes, with a value of 0.02 Nm.

Internal and external disturbance torques perturb the nominal attitude motion. The most severe disturbances are internal torques due to the rate gyro's and mass unbalance (non-diagonal tensor of inertia), and external torques due to the solar radiation pressure, gravity gradient torque and the spacecraft electric dipole in the Earth's magnetic field. Less severe disturbances are torques due to the Earth's albedo, Earth's infrared emission and solar wind. An estimate of the peak values is given in table 6.1.

Table 6.1 - Expected peak values of the disturbance torques [μNm] from [*Belforte et al., 1986a*]

Solar radiation	11.
Gyro moments	10.6
Gravity gradient	1.5
Unbalanced masses	1.2
Earth magnetic field	.65
Earth infrared emission	.06
Earth albedo	.03
Solar wind	.01

The internal torques are expected to be constant during a spin period, and are mainly perpendicular to the spin axis. The solar radiation and solar wind torques are expected to be periodic with the spin motion (due to the spin motion and quasi-constant Sun aspect angle). The other torques are periodic over 12 or 24 hours due to the geostationary orbit and are modulated by the spin motion.

A detailed knowledge of the properties of the Hipparcos attitude prior to the mission is essential for both numerical and dynamical smoothing. For numerical smoothing in the attitude reconstitution and great circle reduction phase semi-analytical representations of the attitude parameters in time will be used. The choice of a representation, but also the number of parameters, are affected by the properties of the Hipparcos attitude. For the dynamical smoothing a model based on the spacecraft's equations of motion will be used.

The most straightforward way to study the behaviour of the Hipparcos attitude is to carry out a number of numerical integrations of a system that represents as well as possible the dynamical properties of the satellite. The simulated attitude produced can be analyzed or used directly in tests of the smoothing procedures. However, in order to prevent optimistic conclusions, the assumptions underlying the simulation process should be completely independent from the smoothing procedures.

The rotation of a *rigid* body is given by the *kinematic*- and *dynamic* equations of motion in the body frame

$$\frac{dA}{dt} = \Omega A$$
$$I \frac{d\vec{\omega}}{dt} = \vec{N}(A) - \vec{\omega} \times I\vec{\omega} \qquad (6.1)$$

with:

A the orthogonal *attitude matrix*, which gives the orientation of the body frame with respect to an inertial system,

$\Omega = \begin{bmatrix} 0 & \omega_3 & -\omega_2 \\ -\omega_3 & 0 & \omega_1 \\ \omega_2 & -\omega_1 & 0 \end{bmatrix}$, where ω_1, ω_2 and ω_3 are the three components of the angular velocity vector,

I the 3x3 symmetric *moment of inertia* tensor, and,

$\vec{N}(A)$ the *torque* on the spacecraft, given in the body frame.

Numerical integration with the appropriate values for I and $\vec{N}(A)$ gives the attitude and angular velocity vector of the spacecraft. The actual satellite, due to the rate gyros and the somewhat flexible solar arrays, is not really a rigid body, but these effects will be neglected. Table 6.2 gives an overview of the attitude simulators, used in the various simulation experiments which are discussed in this chapter.

Table 6.2: Overview of the attitude simulators and simulated control- and disturbance torques.

	CERGA	Great Circle Red.	CSS
Inertial System	Ecliptic	RGC	RGC
Attitude Rep. Representation	3-1-3 Euler angles, transformation to 3-2-1 angles wrt. the RGC system	3-2-1 Euler angles	3-2-1 Euler angles
Scanning law	realistic	simplified	realistic
Num. Integration	Bulirsch-Stoer	Runge-Kutta	?
Tensor of Inertia	diagonal	diagonal	full
Torques:			
Solar Radiation	1. 1st 6 harmonics 2. Computation (see section 6.2.3): a. tabulated b. computed	1st 6 harmonics	variable number of harmonics, computed from CERGA solar radiation torque (2a) (max. 100)
Gyro induced	yes	yes	yes
Gravity grad.	yes	no	yes
Control strat.	1)British Aerospac. 2)Pinard et al.	ad hoc	British Aerospace

The most important torques are undoubtedly the control, solar radiation and gyro induced torques. The control and solar radiation torque, which have a large influence on the attitude modelling, are discussed below. The gyro induced torque is expected to be constant over the mission (except when the configuration of the rate-gyros is changed) and therefore hardly influences the attitude modelling. The gravity gradient torque is small because of the high orbit of the spacecraft and its symmetric and compact construction.

6.2.2 Control Torques

An intermittent attitude control by gas jet firings, which minimizes jitter and is optimized for smoothing, is adopted for Hipparcos. The thrust of the gas jets are fixed (0.02 N), but the duration can be varied between 50 and 500 ms, in steps of .13 ms. Assume for the moment that the tensor of inertia is diagonal, with $I_{11}=I_{22}$ and $I_{33}\simeq 350$ [Nms^2], then the effect of a gas jet firing of t_{on} seconds on the along scan attitude is given by

$$\ddot{\phi} = N_3 \,/\, I_{33} \simeq 6\ t_{on}\ 10^{-5}\ [s^{-2}] \qquad (6.2)$$

with $N_3 = 0.02 \cdot t_{on} \cdot l$ [Nm] and $l \simeq 1$ [m]. So a maximum firing of 500 ms gives a scan rate change of 6 arcsec/s ($\simeq$4% of the scan rate).

Between gas jet firings the attitude is driven by the smaller disturbance torques, which result in a smooth attitude motion. However a considerable amount of high frequency attitude jitter is generated by the gas jet actuations due to structural resonances, especially influencing the observations in the first few frames after the actuation. On the whole the attitude jitter is approximately five times as low as in a satellite driven by reaction wheels [*Kovalevsky, 1984*].

6.2.3 Solar Radiation Torque

The torque induced by the solar radiation is computed by modelling the satellite surface using 16 simple geometric elements: 3 rectangular solar arrays, 6 rectangular walls including 6 rectangular radiators and an hexagonal bottom. The shadows cast by the solar arrays on the walls are taken into account: the illuminated part of each wall is divided into rectangles and triangles with uniform surface structures [*van der Marel, 1983a, Pinard et al., 1983, British Aerospace, 1983*]. For each surface the specular and diffuse reflection, as well as the absorption are taken into account. Only radiation pressure by the Sun is considered, the contribution by the Earth's albedo can be neglected (see table 6.1).

For each elementary surface the position $\mathbf{r}_i(t)$ of the center of pressure and the illuminated area $A_i(t)$ of the surface are computed. Let $\mathbf{s}$ be the unit body vector in the direction of the Sun, $\mathbf{n}_i$ the outer normal vector of each surface element and C_a, C_s, C_d the coefficients of absorption, specular and diffuse reflectivity of each elementary surface, with for an opaque surface $C_a+C_s+C_d=1$. Then the torque is given by

$$\vec{N}(t) = \sum_{i=1}^{16} \left(-p\ A(t)\ \langle \mathbf{n},\mathbf{s}\rangle\ [(1-C_s)\mathbf{s}+2(C_s\langle \mathbf{n},\mathbf{s}\rangle+\tfrac{1}{3}C_d)\mathbf{n}] \right)_i \times \mathbf{r}_i(t) \qquad (6.3)$$

where p is the mean flux (radiation per unit area) divided by the speed of

light, p depends on the distance of the Earth to the Sun. According to [*Wertz, 1978*] p=1358 $Js^{-1}m^{-2}$ at 1 AU.

Let the unit body vector **s** in the direction of the Sun be given by s^T= (cos α cos σ , sin α cos σ , sin σ), with σ the almost constant *Sun aspect angle* ($43^0 \pm 10'$) and α the *Sun aspect azimuth*. The amplitude and time derivatives of the solar pressure torque, assuming a point like Sun, have been plotted in figure 6.2 as function of the Sun aspect azimuth.

Discontinuities in the first derivative of the torque occur when a side panel enters and another leaves the illuminated side of the spacecraft, which happens at $\alpha = 30^0 + i \times 60^0$ (i=0,..,5). The corresponding torque goes through zero, but the center of pressure has a discontinuity in its position; it jumps from one face to another. This discontinuity influences the third derivative of the attitude. There are also discontinuities in the second derivative of the torque when the shadow of the solar arrays enters or leaves one of the surface elements: the center of pressure then changes suddenly its path without a discontinuity in position.

The solar radiation torque is periodic with the spin frequency (Sun aspect azimuth). Therefore its components along the body axes can be developed into a Fourier series with fundamental frequency $f_0=2\pi/\omega_0$, with f_0=1/7680 Hz for the nominal scan rate ($\omega_0 \simeq 168.75$ arcsec/s). In figure 6.2 the amplitude spectrum of the computed solar radiation torque is given. The amplitude decreases exponentially with the frequency, by about an exponent of two (40 dB Nm/decade).

In reality the solar radiation pressure torque will be smoother. The Sun, which so far was modelled as a point source, has an apparent diameter D of approximately $0.^{\circ}53$, and therefore the previously computed discontinuities are in reality smoothed over 11.3 s ($.53/f_0$) in time. In addition the geometric figure of the spacecraft is more complex: the basic shapes are present, but telescope baffles, apogee booster, gas-jet trusters, antenna's and surface irregularities have not been modelled. On the one hand these irregularities will smooth the existing discontinuities, but on the other hand they will make the torque more random. Other deviations in the computed torque may be caused by variations in the solar flux, mismodelling and aging of the coefficients of absorption and reflection.

Let $n'_x(\alpha,\sigma)$ be any component of the previously computed solar pressure torque, then a more realistic torque $n_x(\alpha,\sigma)$ follows from the convolution

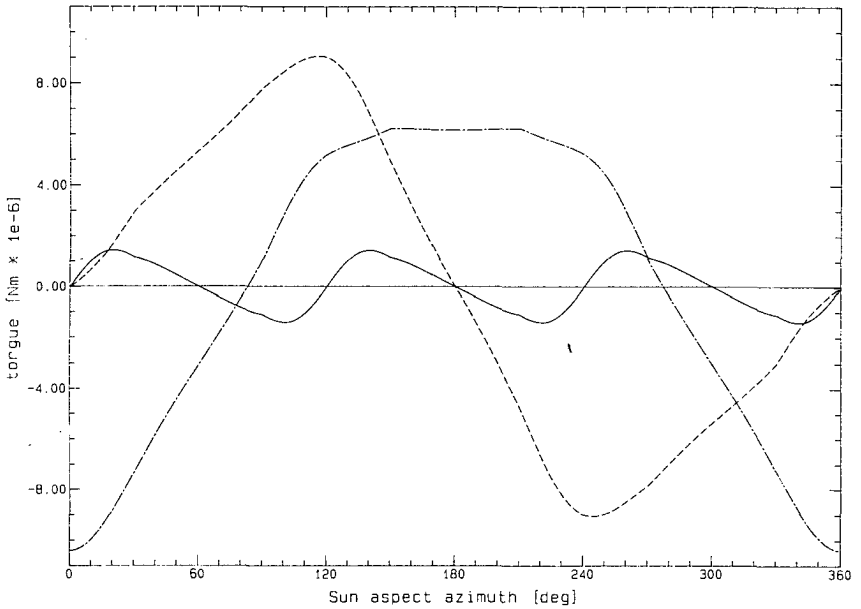
$$n_x(\alpha,\sigma) = \iint n'_x(\alpha-\delta,\sigma-\varepsilon)\ h(\delta,\varepsilon)\ d\delta d\varepsilon \qquad (6.4.a)$$

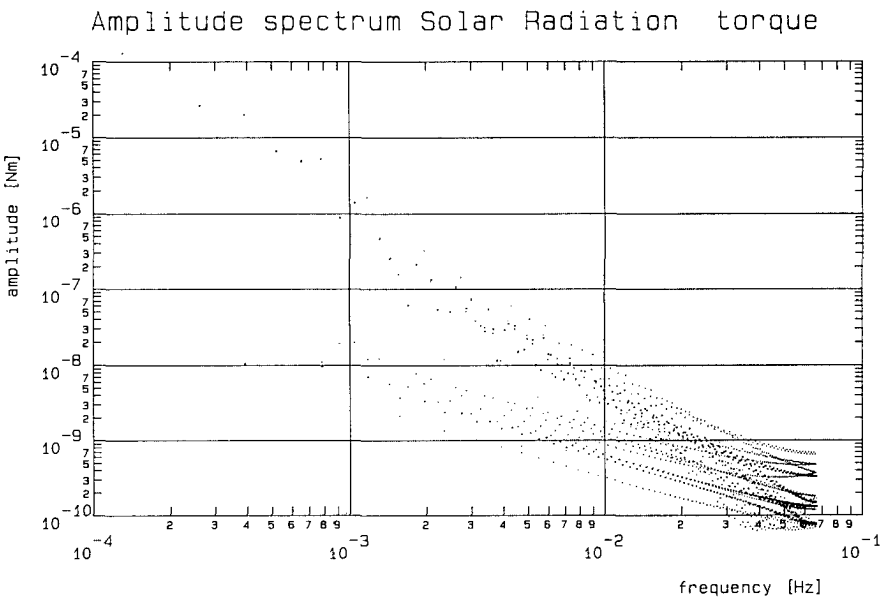
with $h(\delta,\varepsilon)$ the radiation profile of the Sun. The surface brightness of the Sun is nearly uniform over the surface of the disk, so $h(\delta,\varepsilon) = 4/\pi D^2$ for $4(\delta^2+\varepsilon^2) \leq D^2$ and 0 elsewhere. The major variations occur in the Sun azimuth

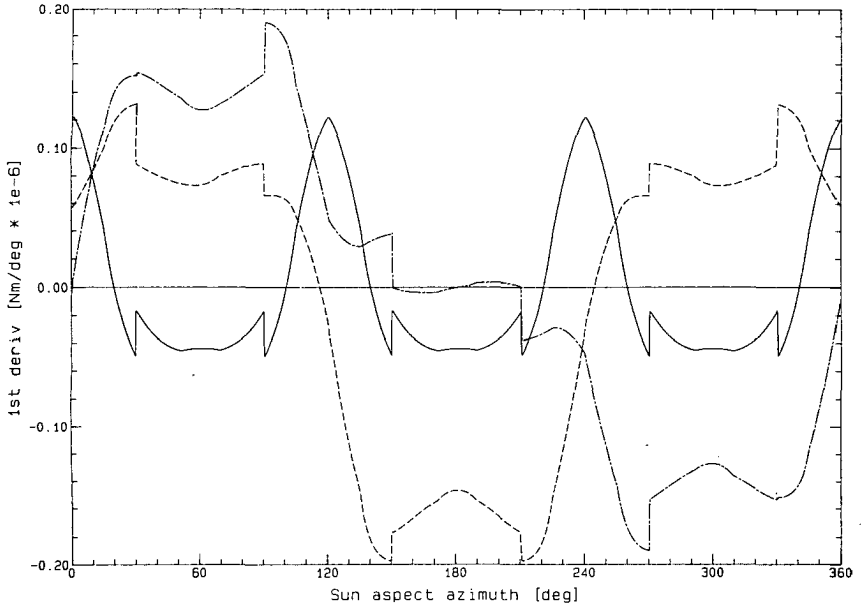
---

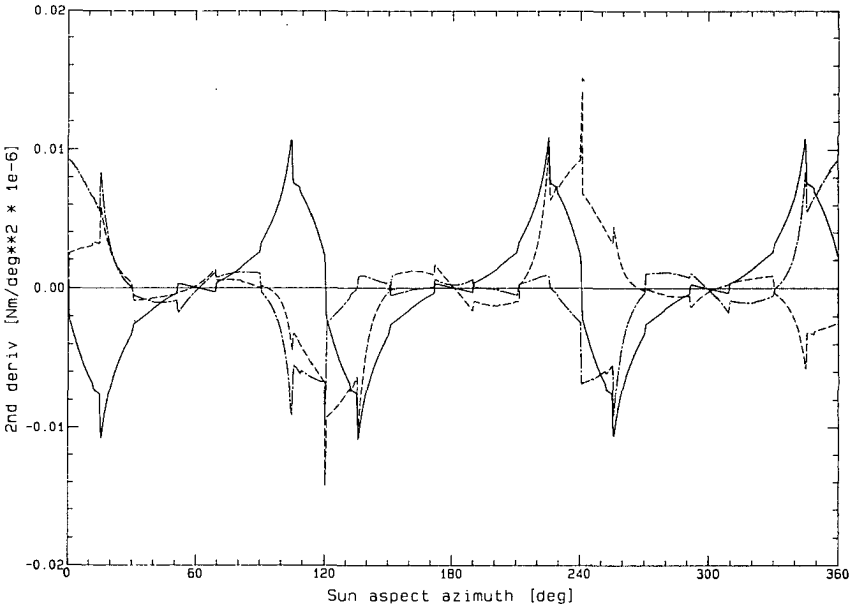
Figure 6.2 (on the right) - Solar radiation pressure torque (a), first time derivative (b) and second time derivative (c) over one rotation of the satellite, and the amplitude spectrum of the solar radiation pressure torque (d), computed from a model of the satellite. Legenda: ····· n_x , —·—·— n_y , ——— n_z

a	d
b	c


torque [Nm * 1e-6]
8.00
4.00
0.00
-4.00
-8.00
0
60
120
180
240
300
360
Sun aspect azimuth [deg]


Amplitude spectrum Solar Radiation torque
amplitude [Nm]
frequency [Hz]


1st deriv [Nm/deg * 1e-6]
0.20
0.10
0.00
-0.10
-0.20
0
60
120
180
240
300
360
Sun aspect azimuth [deg]


2nd deriv [Nm/deg**2 * 1e-6]
0.02
0.01
0.00
-0.01
-0.02
0
60
120
180
240
300
360
Sun aspect azimuth [deg]

angle $\alpha=\omega_0 t$, whereas the Sun aspect angle σ remains more or less constant (±10 arcmin). Therefore the 2-D convolution may be replaced by

$$n_x(t) = \int n'_x(t-\tau)\, h(\tau)\, d\tau \qquad |\tau| \le T=\frac{D}{2\omega_0} \tag{6.4.b}$$

with $h(\tau) = \frac{1}{\omega_0} \int h(\delta,\varepsilon)d\varepsilon = \frac{2}{\pi T} \sqrt{1-\left(\frac{\tau}{T}\right)^2}$. The Fourier transform of $h(\tau)$, with $\omega=2\pi f$, is

$$H(\omega) = \frac{2}{\omega T} J_1(\omega T) \tag{6.5}$$

where $J_1(\omega)$ is the Bessel function of the first kind of order 1 [*Gradstheyn, 1980*, p. 419]. $H(\omega)$ is real since $h(\tau)$ is an even function. According to the convolution theorem convolution in the time domain is equivalent to multiplication in the frequency domain, i.e. the spectrum $N(\omega)$ of the more realistic torque $n(t)$ is

$$N(\omega) = N'(\omega)\, H(\omega). \tag{6.6}$$

with $N'(\omega)$ the Fourier transform of $n'(t)$. The transfer function $H(\omega)$ can be represented by

$$H(\omega) = \sum_{k=0}^{\infty} \frac{(-1)^k\, (\omega T)^{2k}}{2^{2k}\, k!\,(k+1)!} \simeq 1 - \frac{(\omega T)^2}{8} + \frac{(\omega T)^4}{192} - \ldots\ldots \tag{6.7}$$

which is somewhat similar to the Bessel function $J_0(\omega T)$. In other words this transfer function acts like a low-pass filter, with $H(\omega) \simeq 1 - \frac{\pi}{4}(fT)^2$.

The solar radiation torque is not present when the satellite moves through the Earth's shadow cone, the so-called solar eclipse. Solar eclipses happen during 40 days around the spring and autumn equinoxes and last less than 75 minutes. During solar eclipse the solar radiation torque first decreases while the satellite moves through the penumbra (half-shadow) until it vanishes when the satellite enters the umbra (shadow). During the passage through the penumbra, which takes about 2.5 minutes, the solar radiation torque is modulated by the change in solar radiation. Since the surface brightness of the Sun is nearly uniform, the flux is directly proportional to the area of the solar disk which can be seen from the spacecraft. The Earth's atmosphere will absorb and scatter some light, and hence a slight increase in the size of the shadow and a general lightning of the entire umbra due to scattering will occur.

The effects of an eclipse on the attitude are twofold. First the attitude will be more smooth in the umbra because of the absence of the solar radiation torque. Secondly, at the start and end of the passage through the penumbra there will be a discontinuity in the third derivative of the attitude similar to the discontinuities caused by the panels, except that this discontinuity will be somewhat smoother due to some scattering effects of the earth atmosphere.

6.2.4 Attitude Jitter

By *attitude jitter* we designate all short periodic motions of the viewing direction which cannot be reconstituted a-posteriori by the data reduction. Structural vibrations are the primary source of attitude jitter.

They are mainly due to elasticity of the solar panels, which are excited by gas jet actuations. Other sources of attitude jitter are the high frequency components of the disturbing and control torques which are not modelled in the reduction, rate gyro noise and ABM particles shocks, caused by small particles inside the apogee booster motor (ABM) which hit at random times one of the walls of the booster. In [*Froeschlé et al., 1983*] three types of attitude jitter are distinguished: a) high frequency (>20 Hz) white noise, b) medium frequency (between 0.5 and 20 Hz) with some peaks around 3 Hz corresponding to proper frequencies of the satellite with a very low damping [*Kovalevsky, 1984*] and c) a low frequency contribution that blends into the deterministic attitude variations.

The effects of attitude jitter on the grid coordinates will be partly eliminated by the observing strategy. During each elementary observing period *T3* (see section 3.4, p. 23) stars are observed quasi simultaneously, but between successive periods a smooth attitude motion is assumed. Let $P(f)$ be the power spectral density of the attitude jitter, then the jitter induced error on the computed angles is

$$\sigma_\eta^2 = 2 \int_0^\infty P(f) W(f) \mathrm{d}f \qquad (6.8)$$

where $W(f)$ is the spectral window corresponding to the observation strategy and attitude modelling, *i.e.* for the geometric attitude model [*Matra, 1984*]

$$W(f) = \begin{cases} 15\ |4\pi f T_2 \mathrm{sinc}(T_4)|^2 & \text{if } f \le 1 \text{ Hz} \\ 2\ |\mathrm{sinc}(\pi f T_4)/\mathrm{sinc}(\pi f T_3)|^2 & \text{if } 1 \text{ Hz} < f < 15 \text{ Hz} \\ 2 & \text{if } f \ge 15 \text{ Hz} \end{cases} \qquad (6.9)$$

Computations have shown that σ_η is of the order of 2 mas [*Kovalevsky, 1984*] and, due to medium frequencies with a low damping, σ_η is practically constant over the mission except for the first few frames after a gas jet actuation. In case of smoothing the term "$\mathrm{sinc}(\pi f T_4)$" in the window function should be replaced by the corresponding sampling function, *i.e.* for smoothing with B-splines "$\mathrm{sinc}(\pi f T_5)^k$", where T_5 is the interval of the B-spline series and k the order of the splines (see section 6.5). In general this results in a slight increase of σ_η, mainly due to the low frequency jitter.

6.3 Hipparcos Attitude Modelling

In this section the different methods and models for attitude smoothing are described. We distinguish:

1. Numerical smoothing with:
   a) the semi-dynamical model,
   b) the B-spline model,
2. Dynamical smoothing.

In case of *numerical* smoothing the attitude of the telescope is represented by a finite series of base functions for each of the three attitude angles. In the *semi-dynamical* model the base functions are chosen in such a way that the series is actually the solution of the dynamic and kinematic state equations of a rigid body driven by two types of torques: a Fourier series which models the disturbance torques and an impulse train which models the control torques [*Donati et al, 1986a, Donati, 1983b*]. The *B-spline* model on the other hand uses analytical functions, namely B-splines, which are not

directly related to the physics of the spacecraft [*Van der Marel, 1983c, 1985a*]. Other analytical functions have been considered in [*Van der Marel, 1983c*], but B-splines gave the best performance in relation to the cpu time needed for the calculation. Also B-splines proved to be the most flexible and, with a suitable selection of multiple knots, they allow to extend the analytical representation over the gas jet actuations.

In case of *dynamical* smoothing a dynamical model for the spacecraft is used, with as unknowns the initial state plus some physical constants. This model is much more complicated than the models considered so far, and the computations are heavy. However, from a conceptual point of view dynamical smoothing is superior, provided that we have a very accurate dynamical model of the spacecraft.

6.3.1 Definition of the Attitude Angles

The Hipparcos attitude at time t in a reference great circle (RGC) set is defined by an ordered sequence of Euler rotations which align the spacecraft's body axes (instrumental frame) with the reference great circle frame. The body axes [x,y,z] are related to the two *central viewing directions*; the vectors from the focal point of the telescope in the direction of the projection of the zero-point of the grid, through the preceding and following field of view, on the celestial sphere. The origin of the body frame is chosen as the center of mass of the spacecraft. The z-axis is orthogonal to the two central viewing directions, and the x-axis is parallel to the bisector of the two viewing directions. The y-axis completes the triad. The external reference frame is defined by three axes [u,v,w]. The w-axis is defined by the pole of the chosen reference great circle, the u-axis defines the origin on the reference great circle and the v-axis completes the triad.

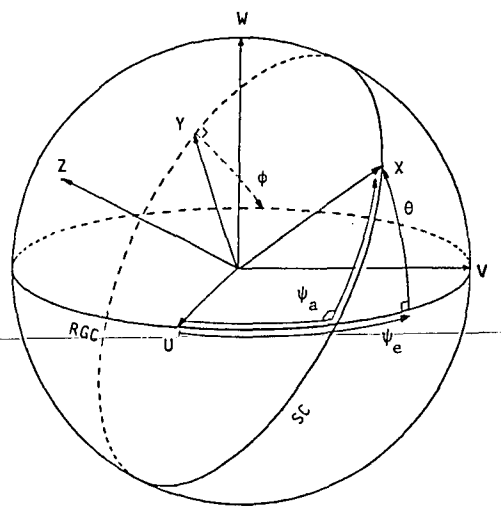


Figure 6.3 - The attitude angles

Two different sequences of attitude angles are considered. The 3-2-1 sequence of Euler rotations aligns the body axes [x,y,z] with [u,v,w] by first a rotation $-\phi$ around x which brings the y-axis in the RGC plane, followed by a rotation $-\vartheta$ around the new y-axis which brings the x-axis in the RGC plane finally followed by a rotation $-\psi_e$ around the w-axis which aligns the x and y with u and v. In the semi-dynamical model used by the

attitude reconstruction, the *Euler* angle ψ_e is substituted, for modelling reasons, by the so-called *astronomical* angle ψ_a. ψ_a is the sum of the first and last angle from the 3-1-3 sequence of Euler rotations, *i.e.* the sum of the rotations around the z and w-axis, which aligns the x-axis first with RGC node and then with the u-axis (figure 6.3).

The main advantage of ψ_a over ψ_e is that this angle is free from second order effects due ϕ and ϑ, and thus depends only on the spin motion and its perturbations. Let $\psi_e = \psi_a + \Delta\psi$, with $\Delta\psi$

$$\sin \Delta\psi = \frac{-\sin\phi \sin\vartheta}{1 + \cos\phi \cos\vartheta} \tag{6.10}$$

then $\psi_e \simeq \psi_a + \frac{1}{2}\vartheta\phi - \frac{1}{24}\vartheta^3\phi + \frac{1}{24}\vartheta\phi^3$.

6.3.2 B-spline Model

A cubic spline consists of a number of cubic polynomial segments joined end to end with continuity up to the second derivative at the joints. The cubic spline going through the joints has among all interpolating functions the property of least curvature. More precisely, the integral of the second derivative is minimal. In general splines consist of polynomial segments, of fixed degree, joined end to end with continuity in a limited number of derivatives at the joints, the so-called knots. One way to represent them is by a linear combination of shifted base functions [*De Boor, 1978*]

$$S(t;\zeta,k,N) = \sum_{i=1}^{N} \lambda_i B_i(t;\zeta_i,k) \tag{6.11}$$

with k the order of the B-spline, *i.e.* the maximum degree of the polynomial segments plus one, ζ_i, $i=1,\ldots,N+k$, the abscissae of the knots and λ_i, $i=1,\ldots,N$ the coefficients of the series. Each base function $B_i(t)$, the so-called B-spline of order k, is completely defined by the abscissae, $\zeta_i,\ldots,\zeta_{i+k}$, of $k+1$ knots. Outside the interval $[\zeta_i,\zeta_{i+k}]$ the B-spline is zero. In a B-spline series of order k two consecutive B-splines always have k knots in common. In figure 6.4 the B-spline functions of order 1, 2 and 4 are given. An example of a B-splines series is given in figure 6.5.

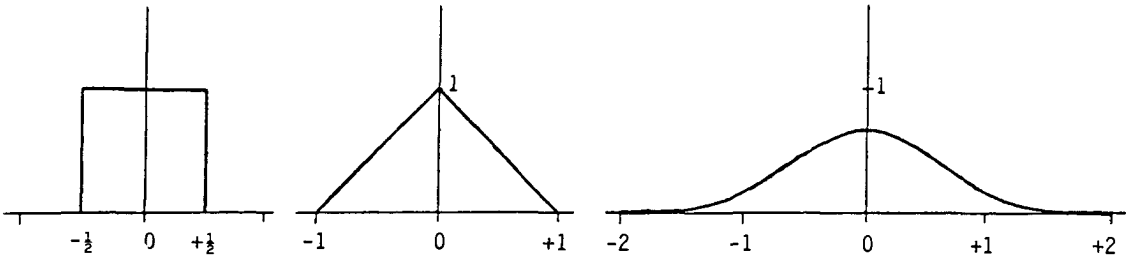


Figure 6.4 - B-splines of order 1, 2 and 4.

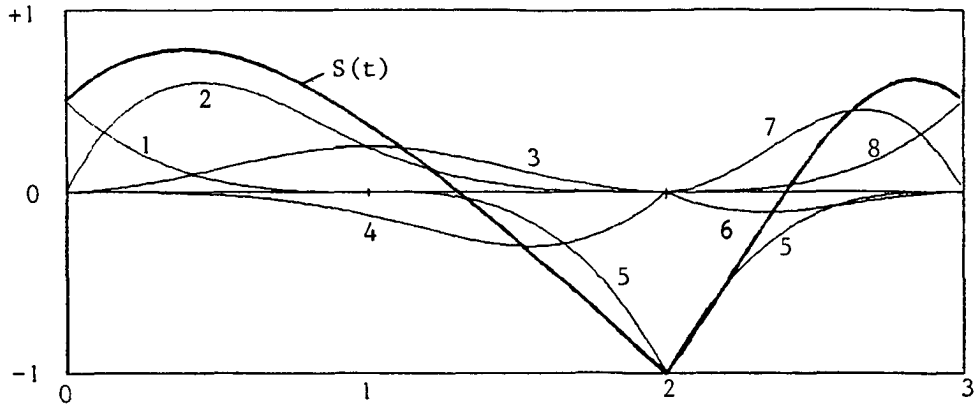


Figure 6.5 - Example of a B-spline series consisting of 8 cubic (k=4) B-splines with coefficients λ_1=.5, λ_2=1, λ_3=.5, λ_4=-.5, λ_5=-1, λ_6=-.25, λ_7=1 and λ_8=.5, and knots $\zeta_1=\zeta_2=\zeta_3=\zeta_4=0$, $\zeta_5=1$, $\zeta_6=\zeta_7=\zeta_8=2$ and $\zeta_9=\zeta_{10}=\zeta_{11}=\zeta_{12}=3$.

B-splines can be computed recursively from B-splines of lower order, starting from the B-spline of order 1, the unit step function:

$$\begin{cases} B_i(t;\zeta_i,1) = 1 \text{ for } \zeta_i \le t < \zeta_{i+1} \text{ and } B_i(t;\zeta_i,1)=0 \text{ otherwise} \\ B_i(t;\zeta_i,j) = \dfrac{t-\zeta_i}{\zeta_{i+j-1}-\zeta_i} B_i(t;\zeta_i,j-1) + \dfrac{\zeta_{i+j}-t}{\zeta_{i+j}-\zeta_{i+1}} B_{i+1}(t;\zeta_{i+1},j-1) \end{cases} \tag{6.12}$$

The (k-1)'th derivative of a B-spline series of order k is "piecewise" constant; there may be jumps at the knots, but between two consecutive knots the derivative is constant. Multiple knots, or coinciding knots, give discontinuities in the lower derivatives. *E.g.* l coinciding knots result in discontinuities in the (k-l)'th and higher derivatives of the series.

An alternative representation for a spline is the so-called piecewise polynomial representation [*De Boor, 1978*]:

$$S(t;\zeta,k,N) = \sum_{j=1}^{k} \frac{c_{ij}(t-\zeta_i)^{j-1}}{(j-1)!} \quad , \ \zeta_i \le t < \zeta_{i+1}, \ i=1,\ldots,l \tag{6.13}$$

where $\zeta_1,\zeta_2,\ldots,\zeta_{l+1}$, with $\zeta_{i+1}>\zeta_i$, are the so-called breakpoints, $l \le N$ the number of polynomial pieces and c_{ij} the j'th derivative at breakpoint ζ_i. Multiple knots in the B-spline series are counted as a single breakpoint in the piecewise polynomial representation. If the piecewise polynomial representation is derived from the B-spline representation the coefficients c_{ij} are not independent, *viz.* there are k times l coefficients c_{ij} necessary

and only l plus k B-splines (in case of only single knots, more are needed in case of multiple knots). For continuity up to the m'th derivative at the i'th breakpoint ($0<m<k-1$), which consists of $k-m-1$ coinciding knots, the right and left derivatives at ζ_i must be equal, *i.e.*

$$c_{ij} = \lim_{t \to \zeta_i} S^{(j)}(t;\zeta,k,N) \qquad \text{for } j=0,\ldots,m \tag{6.14}$$

The derivatives of the spline can be computed in a very simple way from the piecewise polynomial representation, *i.e.*

$$S^{(m)}(t;\zeta,k,N) = \sum_{j=m+1}^{k} \frac{c_{ij}(t-\zeta_i)^{j-m-1}}{(j-m-1)!} \quad , \ \zeta_i \le t < \zeta_{i+1} \ , \ i=1,\ldots,l \tag{6.15}$$

The function values, and/or derivatives, of a spline can be computed more efficiently from the piecewise polynomial representation (*e.g.* by Horner's rule) than from the B-spline representation. On the other hand the B-spline representation is preferred when a spline has to be fitted to some experimental data, because with B-splines no additional conditions between the coefficients have to be added and the degree of continuity at the breakpoints can be regulated very easily.

The Hipparcos attitude is not continuous in all its derivatives and it goes without saying that the B-spline series must have the same discontinuities. Assuming that a gas-jet actuation can be modelled as an instantaneous impulse, which seems justified in view of the relative short duration of the pulse, it will produce a discontinuity in the first derivative of the attitude. Solar eclipses, and shadows cast by the solar panels on the sides of the spacecraft, give at some instants rapid changes in the third derivative of the attitude. These changes can be modelled as discontinuities in the third derivative. By placing the right number of coinciding knots at the discontinuities the B-spline series will have the same discontinuities. The start and end of the series, which can be considered as discontinuities in the function itself, are dealt with in the same way.

The number of knots needed to deal with discontinuities, due to the start and end of the data, gas jet actuations and eclipses, are usually not sufficient to obtain the desired precision. Other knots are needed in order to model the continuous, but unknown, Hipparcos attitude *between* discontinuities with sufficient precision. These knots can be regularly distributed over the intervals between two discontinuities, as close as possible to a nominal density ($1/T_{av}$). But other strategies can be considered too (see section 6.6). Two different values for the nominal density are foreseen, one for the nominal, sunlit, case and another one in case of a solar eclipse. A-priori values for the nominal density can be obtained by experiments on simulated data, but they should in any case be verified during the mission.

The number of B-splines in the series is computed as follows:

$$N = N_{int} + l\,(N_{gas}+1) + 1 + N_{solar}, \qquad l = k-2 \tag{6.16}$$

with N_{gas} the number of gas-jets *within* the attitude sequence under consideration, k the order of the B-spline and N_{solar} the number of additional knots needed to model discontinuities other than gas jet

actuations or the start and end of the data. N_{int}, which can be compared with the role of the number of harmonics in the following section, is

$$N_{int} = \sum_{j=1}^{N_{gas}+1} Round\{ \ (t_j - t_{j-1}) \ / \ T_{av} \ \} \qquad (6.17)$$

where t_j are the times of a gas-jet actuation and T_{av} the reciprocal of the nominal knot density. Unless there is a relatively large gap in the attitude data (*e.g.* during an Earth or Moon occultation), a single B-spline series can be used over 5 or 6 scanning circles (one Reference Great Circle).

6.3.3 Semi-dynamical Model

The sum of the disturbing torques is an unknown, deterministic, quasi-periodic and band limited signal. Such a torque can be approximated by a truncated Fourier series with fundamental frequency $\omega_0=2\pi f_0$. By integrating the state equations twice it follows that the contribution of the disturbance torques to the attitude can also be modelled by a Fourier series plus a few extra, basically parabolic, functions, which - at least for the transversal angles ϑ and ϕ - are modulated by the spin motion. The disturbance torque response for ψ_a, with (unknown) parameters p, s and c, is modelled by

$$\psi_a = p_0 + p_1(t-t_0) + p_2(t-t_0)^2 + \sum_{n=1}^{N} (s_i \sin(n\omega_0 t) + c_i \cos(n\omega_0 t) \ ) \qquad (6.18.a)$$

The disturbance torque response for ϑ and ϕ is modelled by

$$\vartheta = p_0 + p_1(t-t_0) + p_s t \ \sin(\omega_0(t-t_0)) + p_c t \ \cos(\omega_0(t-t_0)) + \sum_{n=1}^{N} (s_i \sin(n\omega_0 t) + c_i \cos(n\omega_0 t) \ ) \qquad (6.18.b)$$

and similarly for ϕ, with, of course, different parameters p, s and c for each angle.

The gas jet actuations are modelled as ideal impulses of unknown intensity centered on the known actuation times. This results in a rate change in the attitude, which is modulated, as regards the transversal angles ϑ and ϕ, by the spin motion. The corresponding base functions are for ψ_a

$$(t-t_l) \ , \qquad t \geq t_l \qquad (6.19.a)$$

and for ϑ and ϕ

$$\sin \omega_0(t-t_l) \qquad 1-\cos \omega_0(t-t_l) \qquad t \geq t_l \qquad (6.19.b)$$

where t_l, $l=1,\ldots,N_{gas}$, are the times of the gas jet actuations.

The total number of base functions N_{tot} (the dimension of the vector of unknowns) depends on the number of harmonics and number of internal gas jets

$$N_{tot} = 2 \ N_{harm} + l \ (N_{gas}+1) + 2 \qquad (6.20)$$

with $l=1$ for the spin angle ψ and $l=2$ for the transversal angles ϑ and ϕ.

The unknown coefficients of the series are computed by a least squares adjustment on batches of data corresponding to one revolution. The number of

gas jet actuations is fixed by the adopted control strategy, so only the number of harmonics can be varied freely in order to obtain the desired accuracy.

6.3.4 Dynamical Smoothing

In case of dynamical smoothing an actual dynamical model of the satellite is used, with as unknowns some parameters for the initial conditions, rate changes due to gas jets, and physical constants. In fact dynamical smoothing can be separated in two separate problems: 1) attitude computation and parameter estimation for the hypothetical case of a perfect dynamical model, and 2) updating and improving the dynamical model.

Any dynamical model is completely defined by the dynamical equations, the moment of inertia tensor, the torques and some initial conditions. The equations, moment of inertia tensor and some of the torques remain reasonably constant, or can easily be computed, over the mission without having to estimate them for every great circle. The initial conditions, control torques and some other physical parameters, like the instantaneous solar flux, are not known a-priori with sufficient precision and have to be estimated more frequently because they vary. A provisional method for dynamical smoothing, based on the above observations, is presented below.

Differencing of the geometric attitude computed during reduction on circles, taking proper care of the gas jet actuation, reveals the underlying torques quite well. Assuming that the gyro induced and gravity gradient torques can be modelled with sufficient accuracy from a-priori data, which seems reasonable, we can separate the solar radiation pressure torque. This estimate of the solar radiation torque can be used to update the solar radiation torque model computed from previous great circles, while estimating the current solar flux and, possibly, some other parameters. Integration of the state equations with unknown initial conditions and unknown rate changes at the gas jet actuations returns a more accurate attitude which can be used to improve the star abscissae.

The solar radiation torque model gives the torque as function of the Sun aspect angle and azimuth. It also accounts for the influence of ageing on the absorption and reflection coefficients. The model will probably be based on the (elementary) geometry of the satellite, plus some correction terms which are computed over several RGC periods from the numerical differentiation. It is well known that numerical differentiation increases noise considerably. Therefore a preliminary smoothing with B-splines, followed by a double differentiation of the B-spline series, might be useful. This preliminary smoothing may be of a very high accuracy, involving much more parameters than used in numerical smoothing, because the results are only used to update the torque from previous RGC's, so that actually the torque is computed from data over several RGC periods.

The accuracy of the model, between gas jet actuations, has not yet been verified. Maybe some additional parameters are needed. We think that the total model is quite accurate for the medium frequencies ($10^{-2}..10^{-3}$ Hz), whereas the lower frequencies could be modelled by just a few B-splines. Another problem might be that the three attitude components are coupled, while we only have accurate information on one component. Up till now we have only experimented with numerical smoothing and we have no definite conclusions about the feasibility, accuracy and economy of dynamical smoothing. No definite plans exist for incorporating dynamical smoothing during any stage of the data reduction.

6.4 Modelling Error

6.4.1 Preliminaries and Notation

Let the function $\mathbf{x}(t)$ be a realization of one of the attitude angles ψ, ϑ or ϕ during some interval T and let $\mathbf{x}(t) \in \mathcal{C}(T)$, the linear function space of all functions continuous on T. Let us adopt the L_2 norm as norm for $\mathcal{C}(T)$, defined over the interval $T=[a,b]$, by

$$\|\mathbf{c}\|^2 = \frac{1}{b-a} \int_{t=a}^{b} \mathbf{c}(t)^2 dt \qquad (6.21.a)$$

and let $\langle \mathbf{c}, \mathbf{d} \rangle$ be the inproduct of two functions $\mathbf{c}$ and $\mathbf{d}$, defined over the interval T, by

$$\langle \mathbf{c}, \mathbf{d} \rangle = \frac{1}{b-a} \int_{t=a}^{b} \mathbf{c}(t) \cdot \mathbf{d}(t)\, dt \qquad (6.21.b)$$

Now, let $\mathcal{S}(T)$ be a linear N-dimensional subspace of $\mathcal{C}(T)$, spanned by the chosen base functions $\mathbf{b}_i(t)$, $i=1,\ldots,N$, for our attitude model, *e.g.* $\mathcal{S}(T)$ is the linear subspace of B-splines of order k and knot sequence t_i, $i=1,..,N+k$. Any function $\mathbf{s}(t) \in \mathcal{S}(T)$ can be written as

$$\mathbf{s}(t;\lambda,N) = \sum_{i=1}^{N} \lambda_i \mathbf{b}_i(t) \qquad (6.22)$$

Define a orthogonal projector $P_{\mathcal{S},\mathcal{S}^\perp}$ which projects functions $\mathbf{c}$ on $\mathcal{S}$ along $\mathcal{S}^\perp$, the orthogonal complement of $\mathcal{S}$ in $\mathcal{C}$, with $\mathcal{C} = \mathcal{S} \oplus \mathcal{S}^\perp$. Here, and in the next sections, we write for brevity $\mathcal{S}$ and $\mathcal{C}$ instead of $\mathcal{S}(T)$ and $\mathcal{C}(T)$. Now, let $\check{\mathbf{x}} = P_{\mathcal{S},\mathcal{S}^\perp}(\mathbf{x})$ be the orthogonal projection $\check{\mathbf{x}}$ of $\mathbf{x}$ on $\mathcal{S}$, then $\check{\mathbf{x}} = \mathbf{s}(t;\lambda,N)$ can be described by a few parameters λ. The *model error* $\check{\mathbf{x}}^\perp$ is defined by the orthogonal complement of $\check{\mathbf{x}}$

$$\check{\mathbf{x}}^\perp = P_{\mathcal{S}^\perp,\mathcal{S}}(\mathbf{x}) = (I - P_{\mathcal{S},\mathcal{S}^\perp})(\mathbf{x}) \qquad (6.23)$$

such that $\mathbf{x} = \check{\mathbf{x}} + \check{\mathbf{x}}^\perp$. Then $\check{\mathbf{x}}$ is the best approximation to $\mathbf{x}$ in L_2 sense, *i.e.* for all $\mathbf{s} \in \mathcal{S}$: $\|\check{\mathbf{x}}-\mathbf{x}\| \le \|\mathbf{s}-\mathbf{x}\|$. The *modelling error* $\|\check{\mathbf{x}}^\perp\|$ is $\|P_{\mathcal{S}^\perp,\mathcal{S}}(\mathbf{x})\|$.

Let $\mathbf{y} = (y_1, y_2, \ldots, y_M)^T$ be some measurements, with respect to $\mathbf{x}(t)$, at discrete times $t_1, t_2, \ldots, t_M$, with $E\{y_i\}=H(\mathbf{x},t_i)$, and let $\hat{\mathbf{x}}(t)=\mathbf{s}(t;\hat{\lambda},N)$ be the weighted least squares estimate of $\mathbf{x}(t)$ given the measurements y_i and weights w_i. Evidently, $\hat{\mathbf{x}} \in \mathcal{S}$, $\check{\mathbf{x}} \in \mathcal{S}$ and $\hat{\mathbf{x}}-\check{\mathbf{x}} \in \mathcal{S}$. Hence we can write for the *estimation error*

$$\|\hat{\mathbf{x}}-\mathbf{x}\|^2 = \|\hat{\mathbf{x}}-\check{\mathbf{x}}\|^2 + \|\check{\mathbf{x}}-\mathbf{x}\|^2 \qquad (6.24)$$

with $\|\check{\mathbf{x}}-\mathbf{x}\|$ the previously introduced *modelling error* and $\|\hat{\mathbf{x}}-\check{\mathbf{x}}\|$ the so-called *measurement induced error* which gives the effect of measurement noise on our estimate. None of the above mentioned error norms is available in practice;

only the discrete *least squares error*

$$\|\mathbf{y}-H(\hat{\mathbf{x}})\| = \sum_{i=1}^{M} w_i(y_i-H(\hat{\mathbf{x}},t_i)))^2 \qquad (6.25)$$

which is minimized by the least squares adjustment, can actually be computed.

The estimation, modelling and measurement induced error, which are in a sense more meaningful than the above mentioned discrete least squares error, can only be computed in simulation experiments, where we have simulated a "true" attitude. On the other hand, for the measurement induced error and modelling error some simple theoretical bounds can be given. But then, using the important equation 6.24, also a theoretical bound for the estimation error can be given.

The *measurement induced error* $\|\hat{\mathbf{x}}-\check{\mathbf{x}}\|$ depends mainly on the measurement noise and degrees of freedom (number of observations minus number of unknowns) in the least squares estimation. Assume $\mathbf{y}-E\{\mathbf{y}\}$ is a stationary discrete white noise with Gaussian statistics $N(0,\sigma_m^2)$ and assume for simplicity $E\{\mathbf{y}_i\}=\mathbf{x}(t_i)$, then

$$\|\hat{\mathbf{x}}-\check{\mathbf{x}}\|^2 \simeq \frac{M}{N}\sigma_m^2 \qquad (6.26)$$

with M the number of observations, N the number of parameters and σ_m the standard deviation of the uncorrelated Gaussian noise. On theoretical grounds one may assume, as will be shown in the next sections, that the *modelling error* $\|\check{\mathbf{x}}-\mathbf{x}\|$ decreases exponentially with

$$\|\check{\mathbf{x}}-\mathbf{x}\| \simeq c\ (N-N_0)^{-r} \qquad (6.27)$$

where r is the so-called *modelling error exponent*, $(N-N_0)$ the number of parameters which can be varied freely and c a constant which must follow from experiments.

A theoretical bound for the estimation error (6.24) follows directly from (6.26) and (6.27). In figure 6.6 the logarithm of the theoretical bound for the estimation, modelling and measurement induced error are plotted as function of the logarithm of the number of parameters N, with modelling error exponent r=4, N_0=25 and with the modelling error $\|\check{\mathbf{x}}-\mathbf{x}\|$ = 1 mas for N=100 as follows from simulation experiments. The measurement induced error can also be given in the form of (6.27), from (6.26) follows that the *measurement induced error exponent* is precisely -0.5. An important conclusion, which is clearly visible from the figure, is:

> *there exists an optimal choice for the number of parameters N which minimizes the estimation error* $\|\hat{\mathbf{x}}-\mathbf{x}\|$, *given the attitude model* $\mathbf{x}$ *and characteristics of the measurement noise*

The optimum is only accessible in simulation experiments, but the corresponding number of harmonics and nominal knot density can be used as starting point in case of real data. These starting values should be validated, when the satellite is operational, by statistical tests on the least squares residuals of the actual data.

6.4.2 Modelling error for B-splines

According to [*De Boor*, 1978] a theoretical upper bound for the maximum modelling error $\|\check{\mathbf{x}}-\mathbf{x}\|_\infty$, in the i'th interval, for splines is

$$\|\check{\mathbf{x}}(\lambda,N)-\mathbf{x}\|_{\infty\,[t_i,t_{i+1}]} < c_k\,|I_i|^k\,\|\mathbf{x}^{(k)}\|_{\infty\,I_i} \qquad (6.28)$$

with $\|\mathbf{x}\|_{\infty\,[a,b]}$ indicating the maximum norm over the interval $[a,b]$, and with interval and mesh size $I_i := [t_{i+2-k}, t_{i+k-1}]$ and $|I_i| := t_{i+k-1} - t_{i+2-k}$, c_k a constant and k the order of the B-spline. The function $\mathbf{x}$ must have a continuous k'th derivative. If the k'th derivative is not continuous in a discrete number of points the upper bound remains valid if the points with a discontinuity are left out of the interval and the B-spline series has the same type of discontinuity, which can be brought about by inserting the appropriate number of knots at the discontinuities.

The modelling error can be influenced in a number of ways. The interval I can be decreased uniformly, or the interval I can be decreased in those areas where the norm of the k'th derivative of f is large, which requires some a-priori knowledge about the norm of the k'th derivative. But also the order k of the splines influences the modelling error.

Assume that the knots, except those for the gas jets, are distributed more or less uniformly. Then the interval I is inversely proportional to the number of these knots. So for the maximum error, but also the L_2 error, $\|\check{\mathbf{x}}-\mathbf{x}\|_2$, assuming that the coefficients c_k and the k'th derivative are limited, we can write

$$\|\check{\mathbf{x}}-\mathbf{x}\|_2 \simeq c\,(N-N_0)^{-k} \qquad (6.29)$$

with c a constant, $(N-N_0)$ the number of - unknown - parameters of the B-spline model which can be varied freely (the reciprocal of I), k the order of the B-spline and N_0 given by

$$N_0 = (k-2)\cdot(N_{gas}+1) + 1 \qquad (6.30)$$

So, the modelling error exponent is equal to the order of the B-splines.

6.4.3 Modelling Error for the Semi-dynamical Model

The modelling error for the semi-dynamical model mainly depends on the high frequencies in the attitude spectrum, which are not modelled by the truncated Fourier series. We assume that this is the only source of errors, and the contribution of the mismodelling of gas jet response functions is neglected.

The attitude spectrum depends mainly on the spectrum of the solar radiation torque. As we have seen in section 6.2 the exponential decay of the solar torque spectrum, r_t, defined by

$$a_n = c\,n^{-r_t} \qquad (6.31)$$

is approximately 2 (40 dB Nm/decade), where n is the degree of the harmonic component, a the coefficient and c a constant. The corresponding exponential decay of the attitude spectrum is $r_t+2 = 4$ (80 dB arcsec/decade) because the torque is integrated twice.

The spectrum of the L_2 modelling error $\|\check{\mathbf{x}}-\mathbf{x}\|$ has an error exponent which is smaller, by 0.5, than the decay of the attitude spectrum, which follows from integrating the attitude spectrum over the harmonics which are not contained in the model. Therefore the modelling error for the semi-dynamical model can be written in the exponential form of (6.27). The modelling error exponent, r, is ~3.5. and the number of parameters, N_0, which cannot be varied freely, is from (6.20)

$$N_0 = l\,(N_{gas}+1) + 2 \qquad (6.32)$$

where l=1 for ψ and l=2 for ϑ and ϕ.

6.4.4 Results from Simulation Experiments

The theoretical error bounds have been verified by a number of experiments on simulated attitude angles, in collaboration with Centro di Studi sui Sistemi (CSS), Torino [*Belforte et al., 1986a*]. Attitude angles have been simulated with and without additional random noise of 10 mas, using the simulation software available at CSS (see table 6.2 and intermezzo I).

intermezzo I

The attitude has been simulated by Centro di Studi sui Sistemi (Torino) by numerical integration of the dynamic Euler equations and kinematic equations of the satellite attitude. Simulated angles were computed with respect to the nominal motion of the satellite. The simulated attitude is based on the following assumptions:
- the satellite is a rigid body,
- the symmetric tensor of inertia is:

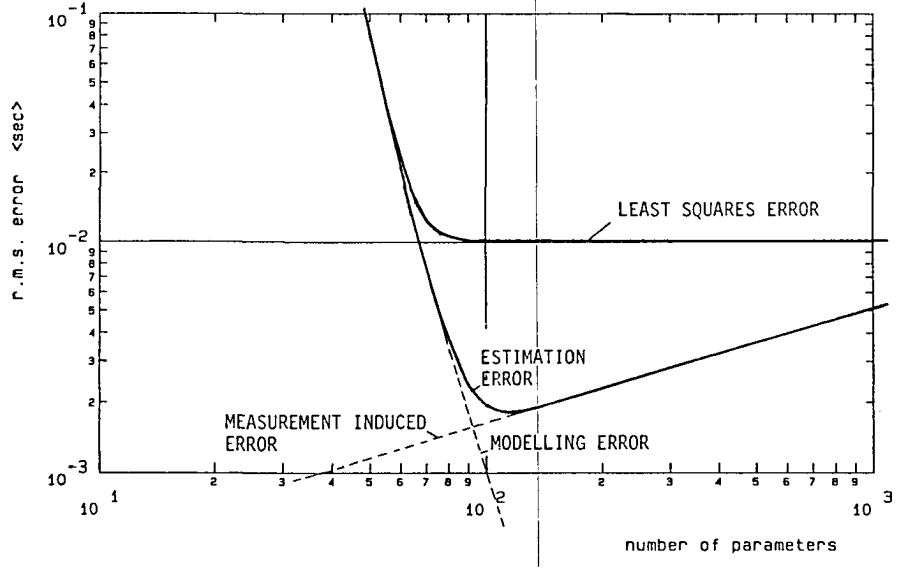
$$I = \begin{bmatrix} 385.7 & & -sym- \\ -2.6 & 430.8 & \\ .2 & -1.8 & 354.2 \end{bmatrix} \quad kg\ m^2$$

- simulated disturbance torques are: solar radiation torques, gyro moment torques and gravity gradient torques,
- the solar radiation torque is periodic and given as a Fourier series with harmonics up to degree 100. This Fourier series was derived from simulated data by CERGA [*Pinard et al.*, 1983].

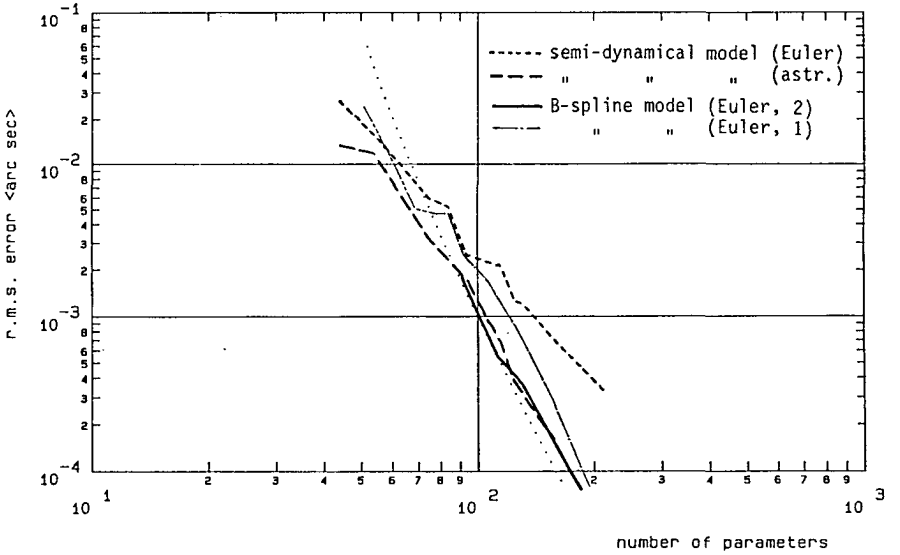
Torques which are very small and/or have spectra similar to either the solar radiation or to the gravity gradient torques may be neglected. We believe that the simulated attitude contains all the features of the real attitude, although the actual realization can be different. In other words the amplitude spectrum of the simulated attitude is fairly realistic.

The attitude data, without additional noise, have been used to study the modelling errors as function of the number of unknowns. The same data, but now perturbed by a discrete white noise with standard deviations of 100 or 10 mas, was used to study the estimation errors. The standard deviations are representative for the uncertainties in the star mapper and IDT data respectively. The experimental results have been obtained for several datasets. The majority of the simulated datasets contained one revolution of data, with a simulated attitude every 2 seconds of time.

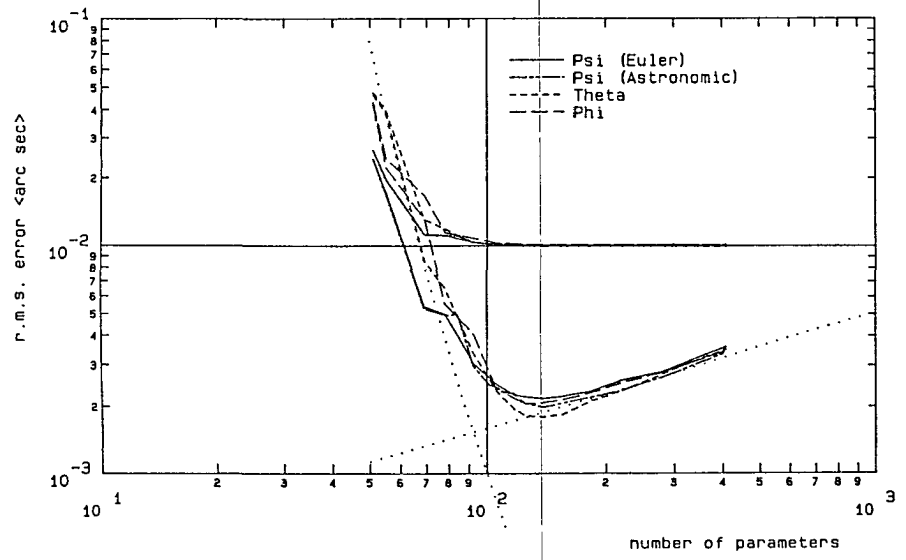
Expected r.m.s. errors



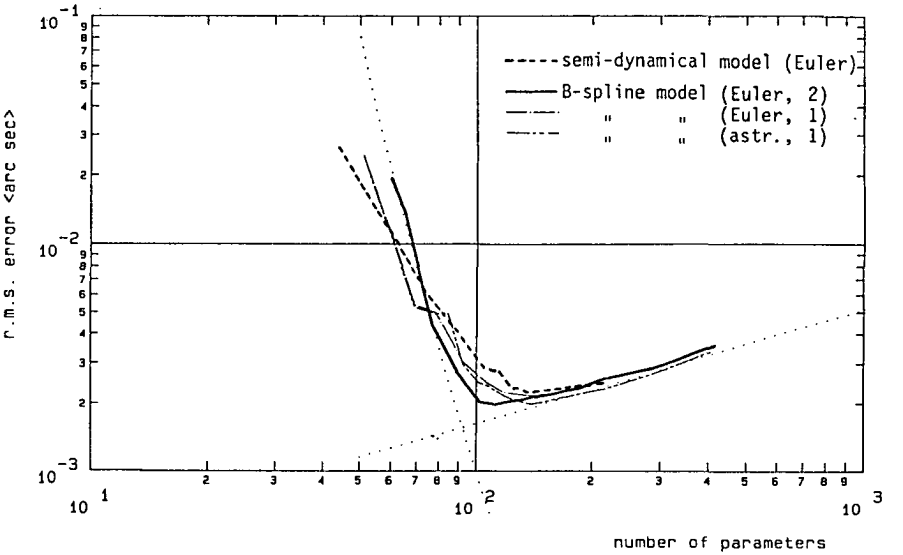
Modelling error (r.m.s.) Psi-angle



B-spline model, strategy 1 k=4



Estimation error (r.m.s.) Psi-angle



In figure 6.6 the computed modelling and estimation errors are plotted and compared with the theoretical results of sec. 6.4.1 - 6.4.3. Figure 6.6 (b) also gives the least squares error, which goes asymptotically to 10 mas, the noise level of the data. The asymptotic behaviour of the experimental errors agrees quite well with the predicted errors. For a small number of parameters (<80) the modelling error is dominant, but for a larger number of parameters (> 150) the measurement induced error becomes dominant instead. In between there is a transition zone where the estimation error reaches a minimum. Near this minimum the function is quite flat, so the choice of the number of parameters for which the minimum will be obtained is not very critical. The optimal number of parameters and corresponding estimation error in the simulated case are tabulated in table 6.3.

Table 6.3: Best experimentally obtained estimation errors in mas

	Estimation err. σ_m=10 mas				Estimation err. σ_m=100 mas			
	B-splines		semi-dynamic		B-splines		semi-dynamic	
	n	rms	n	rms	n	rms	n	rms
ψ_e	139	2.1^7	125	2.3^4	92	$17.^0$	94	$17.^7$
ψ_a	139	1.9^7	-	-	69	$16.^0$	-	-
ϑ	139	1.7^9	125	1.7^1	84	$13.^8$	66	$13.^6$
ϕ	125	2.0^6	117	1.9^3	84	$15.^9$	66	$15.^4$

Approximately 140 attitude parameters per circle are needed for the B-spline model and 125 for the semi-dynamical model. The estimation error is in both cases ~2 mas. When additional knots (12) are inserted in the B-spline series to model the discontinuities in the solar radiation torque, the optimum is shifted to 110 B-splines and the error is reduced to 1.97 mas. The ϑ angle is modelled by both approaches better than the other transversal angle ϕ or the along scan angle ψ. Somewhat disappointing is the estimation error of the Eulerian ψ angle for the semi-dynamical model. The astronomical ψ angle is modelled by the semi-dynamical model significantly better than the Eulerian ψ angle (see figure 6.6). Better results for the Euler angle - which is slightly modulated by the spin motion - are obtained when the gas-jet response functions of (6.19.b) are used instead of (6.19.a). The astronomical angle is not modulated by the spin motion, so here the gas-jet response function of (6.19.a), which has only one degree of freedom, is good enough.

Figure 6.6 (on the left) - Expected and experimentally obtained modelling and estimation errors versus the number of attitude parameters. (a) expected rms errors, (b) estimation and least squares errors for the B-spline model (strategy 1, order k=4), (c) modelling error in ψ for the B-spline model (different strategies) and semi-dynamical model (Euler and astronomical angle), and (d) estimation error in ψ for the B-spline model (different strategies and ψ angle) and semi-dynamical model. The theoretical error bounds for the modelling and measurement induced error are plotted as dotted curves.

a	c
b	d

The experimentally obtained modelling error exponent for the B-spline model is slightly better than was predicted in section 6.4.2: The modelling error exponent is close to 4, as was predicted, for small number of parameters. But for more parameters the exponent increases to 5, which is partly caused by the truncation of the Fourier series in the simulation. The experimentally obtained modelling error exponent for the semi-dynamical model appears to be 3.5, as was predicted, with about 70 parameters. But again for more harmonics the modelling error exponent goes up to more than 4, also because of the fact that the Fourier series of the solar pressure torque is truncated after 100 harmonics.

6.4.5 Conclusions

Although the two approaches for modelling the Hipparcos attitude are quite different, the results in terms of modelling accuracy and optimal number of parameters are quite similar. In combination with statistical tests, which will lead to rejections in case the modelling errors are still significant (because too few parameters were used), both methods are quite robust. The semi-dynamical model gives a slightly better modelling accuracy, especially for a small number of parameters and in particular when compared to the B-spline model with a very simple knot-placement strategy. However for the semi-dynamical model the choice of the along scan angle representation is important: the astronomical angle gives a better performance than the Euler angle. The B-spline model is quite insensitive to the choice of the along scan angle.

With respect to the computing times B-splines are superior. The normal matrix in the great circle reduction will have a limited, but variable, bandwidth. Compared to the semi-dynamical model, which gives a full normal matrix, computing times are obviously superior. This supremacy is however of little effect for the attitude reconstruction when the semi-dynamical model is used with only a few parameters. So the semi-dynamical model seems to be a good choice for the attitude reconstruction. But for the great circle reduction, which involves many more unknown parameters, the B-spline approach is preferred for its superior computing time. Although, during the great circle reduction, Fast Fourier Transform (FFT) methods could be used to improve the performance of the semi-dynamical method (during the great circle reduction the attitude data is available at regular intervals).

6.5 Harmonic Analysis of Cardinal B-splines

The spectrum of a cardinal B-spline series, i.e. polynomial splines defined over a regular infinite sequence of equally spaced knots, is investigated in [*Sünkel*, 1981]. The spectrum of the 1'st order cardinal B-spline, the unit step function, is the so-called "sinc" function. Cardinal B-splines of higher order (k) can be constructed as k'th order convolutions of the unit step function. According to the convolution theorem the corresponding Fourier transform is the k-fold product of the sinc function:

$$Q_k(f) = Q_1^{\,k}(f) = \left[\tau \frac{\sin \pi f \tau}{\pi f \tau} \right]^k \qquad (6.33)$$

The cardinal B-spline series can actually be written as a convolution, therefore, according to the convolution theorem, the spectrum of the smoothed function is

$$\mathcal{S}(f) = DFT(\lambda)\ Q_k(f) \qquad (6.34.a)$$

where $DFT(\lambda)$ is the discrete Fourier transform of the coefficients of the cardinal B-spline series. The coefficients λ may be solved analytically. This allows us to write the series as a linear combination of shifted sampling functions, with the function values at the knots as coefficients. Therefore the Fourier transform of the B-spline series can be obtained by multiplying the discrete Fourier transform of the given data by the Fourier transform of the sampling function, which acts as a transfer function, *i.e.*

$$\mathcal{S}(f) = DFT(f(t))\ F_k(f) \qquad (6.34.b)$$

The Fourier transform of the sampling function for cubic B-splines, *i.e.* k=4, is:

$$F_4(f) = \frac{3}{2 + \cos 2\pi f}\ Q_4(f) \qquad (6.35)$$

which is the transfer function in the frequency domain for the discrete Fourier transform of the data. In figure 6.7 a number of Fourier transforms of sampling functions, including the one for the semi-dynamical model, is given. If the B-spline order goes to infinity then the Fourier transform of the sampling function matches the Fourier transform of a trigonometric series (semi-dynamical model). But for cubic splines the transforms are already quite similar.

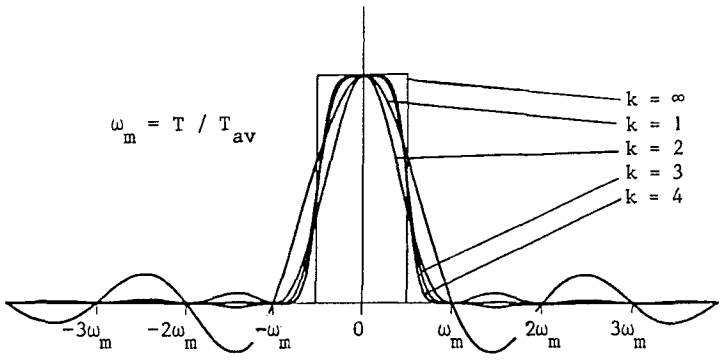


Figure 6.7 - Fourier transforms of the sampling functions (from [*Sünkel*, 1981]).

From the above, and from our experimental results in the preceding section, follows that a B-spline series forms a good and efficient approximation of a Fourier series.

6.6 Effects of B-spline Order and Knot Placement

The shape of the B-splines, and hence the function space $\mathcal{S}$ on which the attitude function is projected, depends on the chosen order of the spline and the location of the knots. Therefore, by means of the B-spline order and location of the knots, as well as the number of B-splines, the modelling error $\check{\mathbf{x}}^1$ can be influenced. In the previous sections we considered the effect of the number of B-splines on the modelling, and estimation, error. In this section the effects of B-spline order and location of the knots is investigated. Consider the theoretical upper bound (6.28) for the maximum

error $\|\check{\mathbf{x}}-\mathbf{x}\|_\infty$ for splines of section 6.4:

$$\|\check{\mathbf{x}}-\mathbf{x}\|_{\infty\ [t_i, t_{i+1}]} < c_k\ |I_i|^k\ \|\mathbf{x}^{(k)}\|_{\infty\ I_i} \qquad (6.28)$$

So, more precisely, the modelling error of the B-spline approximation depends on a) the order k of the splines, b) the choice of the intervals I, c) the norm of the k'th derivative of $\mathbf{x}$ (without discontinuities), and d) the type of discontinuities in the derivatives of the data.

6.6.1 Order of the B-splines

The effect of the B-spline order on the modelling error is captured by equation (6.29):

$$\|\check{\mathbf{x}}-\mathbf{x}\|_2 \simeq c\ (N-N_0)^{-k} \qquad (6.29)$$

where c is proportional to $c_k\|\mathbf{x}^{(k)}\|_\infty$ and $N_0=(k-2)\cdot(N_{gas}+1)+1$. When the order of the B-splines is increased, the exponent goes down, but also N_0, the number of B-splines which are needed to model the discontinuities in the data, increases and $c_k\|\mathbf{x}^{(k)}\|_\infty$ changes. Evidently the optimal order of the B-splines depends on the data and the required accuracy.

This effect of the order of the B-splines on the modelling error has been verified by simulation experiments. In figure 6.8 the modelling error is plotted for B-splines of order k=6. The exponential decrease reaches the order of the B-spline only for very small modelling errors. In the regions which are of practical interest the performance - expressed in number of parameters and estimation error - is even a little worse than that for cubic (k=4) splines, since for higher order B-splines relatively more parameters N_0 are needed to cope with gas-jets and other discontinuities. From these experiments the use of cubic B-splines seem to be justified.

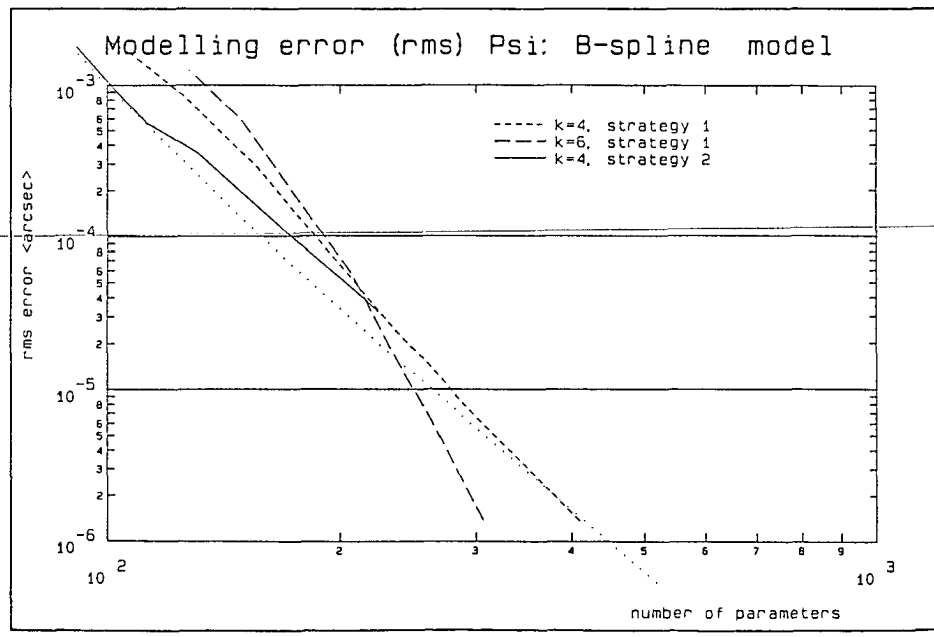


Figure 6.8 - Effect of higher order (k=6) B-splines on the rms modelling error

6.6.2 Knot Placement Strategies

Major discontinuities in the derivatives of the Hipparcos attitude occur at gas jet actuations and, of course, the B-spline series must have the same discontinuities. The torque generated by each gas jet firing can be modelled as a constant pulse of variable duration (between 50 and 500 ms). The gas-jet actuations produce two discontinuities in the second derivative of the attitude at the flanks of the pulse. However, in view of the relative short duration of the pulse compared to the frame period (2.13 s.) gas jet actuations can be modelled as an instantaneous impulse, which produces a single discontinuity in the first derivative of the attitude at the mid-time of the pulse.

This model is rather sensitive to errors in the specified mid-frames times: an error of 10 ms will give an error of 60 mas on the attitude for a gas jet actuation of 500 ms. Therefore, depending on the accuracy of the specified mid-times and the duration of the pulse, it might be better to model the gas-jet by either two discontinuities in the second derivative of the attitude at the flanks of the pulse, or a single discontinuity in the attitude itself close to the time of a gas jet firing. We prefer the last option in view of the short duration of the pulse and because it requires fewer knots for B-splines of order $k>4$.

Certain phenomena in the solar radiation torque, like solar eclipses, the hexagonal structure of the spacecraft and shadowing by the solar panels, give some rapid, but continuous, changes in the third derivative of the attitude. In other words, the fourth derivative of the attitude is significantly larger over the 5 or 6 frame periods during which these effects occur. Compared to the frame period it is hard to speak of discontinuities, on the other hand, compared to the relatively long smoothing intervals, they act like discontinuities and it certainly makes sense to insert at these places more knots in order to reduce the modelling error in the B-spline series.

This strategy is not limited to effects of solar radiation. In general, more knots are needed when the k'th derivative of the attitude $\|\mathbf{x}^{(k)}(t)\|_\infty$ is larger. This may for instance also happen right after gas jet actuations which excite some jitter and increase the probability of ABM particle shocks (section 6.2.5). On the other hand, during solar eclipse (except for the passages through the penumbra) the nominal knot density will be lower than foreseen for the Sunlit case. A-priori values for the nominal densities can be obtained by experiments on simulated data.

This strategy has been tested on simulated data (intermezzo I), which resulted in a small improvement in the modelling error over the quasi-cardinal strategy. The effect on the estimation error is illustrated in figure 6.6 (d), where the various estimation errors for ψ are plotted. Strategy 1 is a quasi-cardinal strategy, which takes only gas jet actuations into account and with regularly spaced knots in the intervals between two gas jet actuations. Strategy 2, which has been optimized for the data at hand, models also the discontinuities in the solar radiation torque. The corresponding modelling errors have already been plotted in figure 6.6 (c) and 6.8. The second strategy appears also to be quite insensitive for wrongly placed knots, *i.e.* knots which are not positioned exactly on the expected discontinuity, with the exception of knots due to the gas jet impulses.

6.7 Star Abscissae Improvement by Attitude Smoothing

Smoothing of the attitude effectively increases the longitudinal field of view, since more stars are connected directly. Especially more bright stars are now linked directly to each other, and not only by chains of measurements between fainter stars [*Lacroute, 1983*]. Smoothing will, therefore, have two favourable effects: 1) it leads to an overall increase in precision for the astrometric parameters and 2) it permits a more liberal observing strategy.

The last point, the more liberal observing strategy, can be used in two ways. One could argue that it is not necessary anymore to spend too much observing time on faint stars to get a good final accuracy, and spend this time on bright stars. On the other hand, one could argue that the bright stars will be improved anyhow, and that one should spend more time on, and improve the accuracy of, faint stars, which are often astronomically more interesting. Both arguments are valid, each with its own supporters. Another aspect of the observing strategy is that more stars may be put on the observation list. But first, let us consider what can be gained by smoothing exactly.

The precision of the star abscissae ψ_i depends on the precision of the grid coordinates and the precision by which the attitude parameters ψ_k are reconstituted (see also section 5.5), *i.e.*

$$\sigma^2_{\psi_i} \simeq \frac{1}{n_i} (\sigma^2_{x_{ik}} + \sigma^2_{\psi_k}) \qquad (6.36)$$

where we have neglected the influence of the instrument, with $\sigma_{x_{ik}}$ the average precision of the grid coordinates, approximated by

$$\sigma = \sigma_{B=9,T=1} \sqrt{T \; 10^{-0.4 \cdot (B-9)}} \qquad (6.37)$$

(B is the star colour index) and σ_{ψ_k} the precision of the reconstituted attitude parameters. In short we may write

$$\bar{\sigma}^2_{\psi_i}(B) = \bar{\sigma}^2_{x_i}(B) + \bar{\sigma}^2_{\psi_k} \qquad (6.38)$$

where $\bar{\sigma}^2_{x_i}(B)$ may be computed from (6.37) by inserting the total observing time, and $\bar{\sigma}^2_{\psi_k}$ denotes the influence of the attitude, which is reduced considerably by smoothing. An analytical expression for the attitude influence, and some examples with an enlarged field of view, were already given in section 5.5.

The improvement due to smoothing has been calculated by simulation experiments with the great circle reduction software. Grid coordinates for a reference great circle, consisting of only 1 scanning circle, with an almost regular starfield of 800 stars, were simulated using the nominal observing strategy in [*Kovalevsky & Dumoulin, 1983*]. The simulated attitude of the spacecraft is given in figure 6.9. The maximum deviation from the scanning law is 10', 11 gas jet actuations have been simulated. A more detailed description of these experiments and the results for other observing strategies are given in [*Van der Marel, 1985a*].

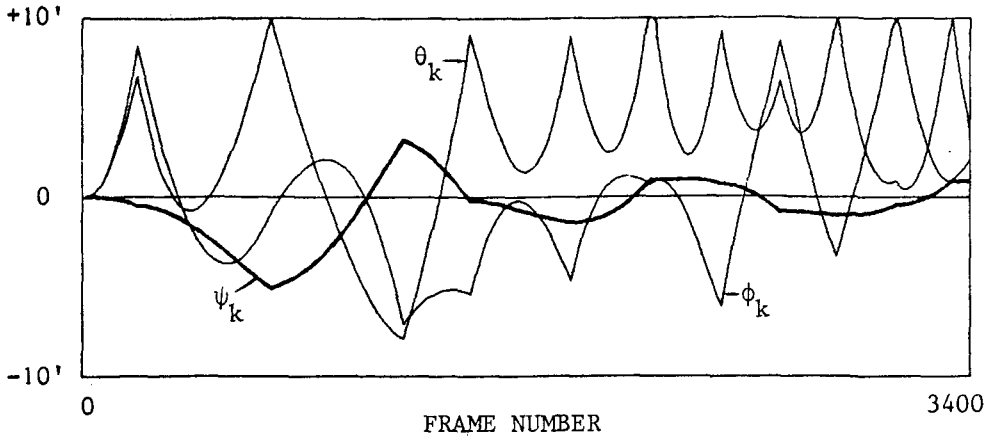


Figure 6.9 - Deviation of the simulated attitude from the scanning law

On the basis of these simulated data several GCR solutions have been computed, in the geometric mode with 3400 attitude parameters and in the smoothing mode with 138, 119, 109, 99, 89, 78, 69, 59 49, 40 and 37 B-splines. 37 B-splines is the minimum number necessary: 33 are needed for dealing with the gas jet actuations and another 4 are needed for the start and end of the data.

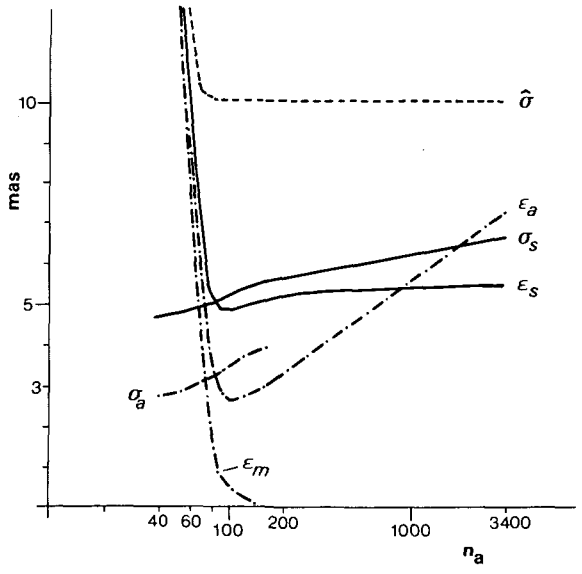


Figure 6.10 - Mean standard deviation and rms of the a-posteriori noise of the attitude and star parameters, the estimated variance factor and the modelling error of the attitude parameters versus the number of parameters in the attitude model.

In figure 6.10 the mean standard deviation and the rms value of the a-posteriori noise (estimated minus true simulated values) for the attitude and star parameters, the estimated variance factor and the modelling error in the attitude are plotted versus the number of attitude parameters used in each run. The a-posteriori noise reaches a minimum near 99 B-splines per

circle. With a smaller number of B-splines the modelling error becomes significant, for a larger number the inherent smoothness is not sufficiently exploited.

In figure 6.11 the mean standard deviation of the star abscissae is plotted versus the star magnitude for 4 different attitude models: the geometric attitude model (a), the B-spline model with 99 parameters (b), the hypothetical case of a perfect dynamical model, for which only initial conditions have to be estimated using 37 B-splines (c) and the unrealistic, but informative case of a perfectly known along-scan attitude (d). Stars brighter than magnitude 7 are improved by an factor 1.45 (with respect to the geometric attitude model) in case of smoothing with 99 B-splines. For the hypothetical case of a perfect dynamical model, simulated by the B-spline model with 37 parameters, the improvement goes up to a factor 2 for stars brighter than magnitude 7. As expected the improvement is very small for faint stars, since for faint stars $\bar{\sigma}^2_{x_i}(B) \gg \bar{\sigma}^2_{\psi_k}$, we have from (6.38) that $\bar{\sigma}^2_{\psi_i}(B) \simeq \bar{\sigma}^2_{x_i}(B)$, which is, therefore, almost not affected by an improvement in $\bar{\sigma}^2_{\psi_k}$ due to smoothing.

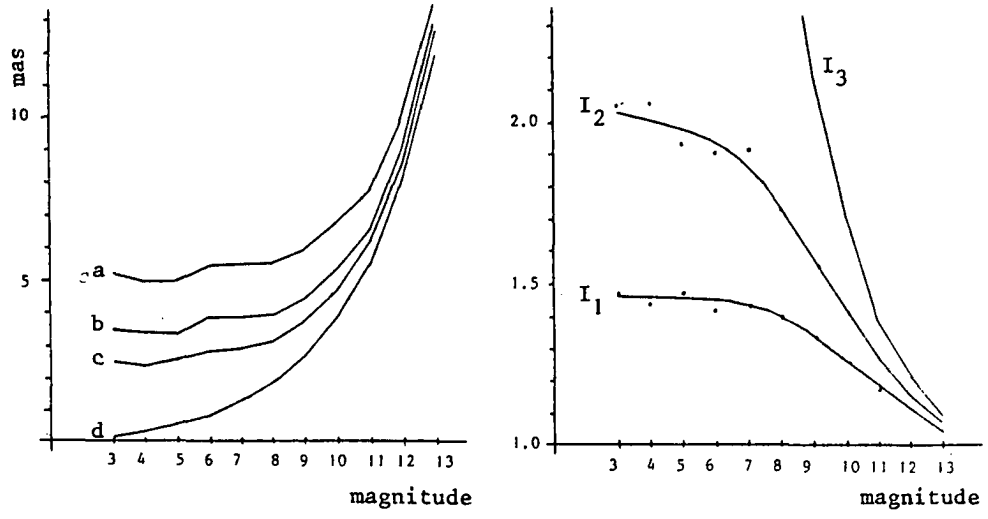


Figure 6.11 - Mean standard deviation of the star abscissae per magnitude class. The cases a, b, c and d are described in the text. $I_1=\sigma_a/\sigma_b$, $I_2=\sigma_a/\sigma_c$ and $I_3=\sigma_a/\sigma_d$.

Similar experiments have been carried out with CERGA dataset II. The modelling error in table 6.4 and figure 6.12 is computed by a great circle reduction with errorless approximate data and grid coordinates (except for a truncation error of 0.02 mas). In figure 6.12 the modelling error is given for the first 6000 attitude frames for the run with 700 B-spline parameters. The corresponding estimation errors are given in table 6.5 for two iteration runs with smoothing (good a-priori data), using 600 and 704 B-spline parameters respectively. These runs should be compared with the runs with 597 and 703 B-splines in table 6.4. Due to a tape error 40 frames (149 grid coordinates) of data were lost in iteration mode. This resulted in 2 attitude sequences instead of one, with as a consequence a number of extra B-spline parameters.

In figure 6.13 the estimation errors in the geometric and smoothed attitude are plotted for the first 6000 frames for the iteration run with 703 B-spline parameters. The corresponding modelling error has been plotted in figure 6.12.

Table 6.4: Modelling error (smoothing) for CERGA dataset II in mas. The dataset consisted of 17432 frames with 75223 grid-coordinates for 1843 active stars (out of 2411). In addition instrumental parameters were solved. 62 gas jet firings occured. Notes: 1) geometric solution, 2) additional splines for eclipses, 3) expected number of sign changes (smooth data) in brackets.

# B-splines	$\hat{\sigma}$	attitude error			star abs. error	
		rms	[min,max]	sign changes 3)	rms	max.
17432 1)	0.06	0.31	[-0.7,0.7]	-	0.06[7]	.26
703 (110) 2)	0.35	0.48	[-4.8,2.6]	1076 (704)	0.34	2.3
606 (90) 2)	0.74	0.95	[-5.3,5.4]	454 (607)	0.66	3.6
597 (90)	1.22	1.04	[-11.8,5.2]	392 (598)	0.64	3.0
509 (70) 2)	1.46	2.4	[-16.3,10.3]	287 (510)		
498 (70)	1.51	2.4	15.5	337 (499)		
447 (60)	2.40	3.8	31.7	148 (448)		

Table 6.5 - Estimation errors of CERGA dataset II in mas, with no special modelling for eclipses. In brackets the improvement wrt. the geometric solution.

	smoothing 600	smoothing 704	geometric
sample st. dev.	10.20	10.08	10.04
true attitude error			
rms	1.66	1.3	4.06
max	-10.9	7.0	60.2
smoothed-geometric att.			
rms	4.16	3.9	
max	55.9	58.1	
sign changes (8695)	6819	7497	
error star abs.			
formal st.dev.	2.70 (.73x)	2.72 (.73x)	3.71
influence attitude	2.02 (.60x)	2.05 (.63x)	3.25
rms true error	2.31 (.89x)	2.19 (.85x)	2.60
max true error	17.2	18.1	16.2

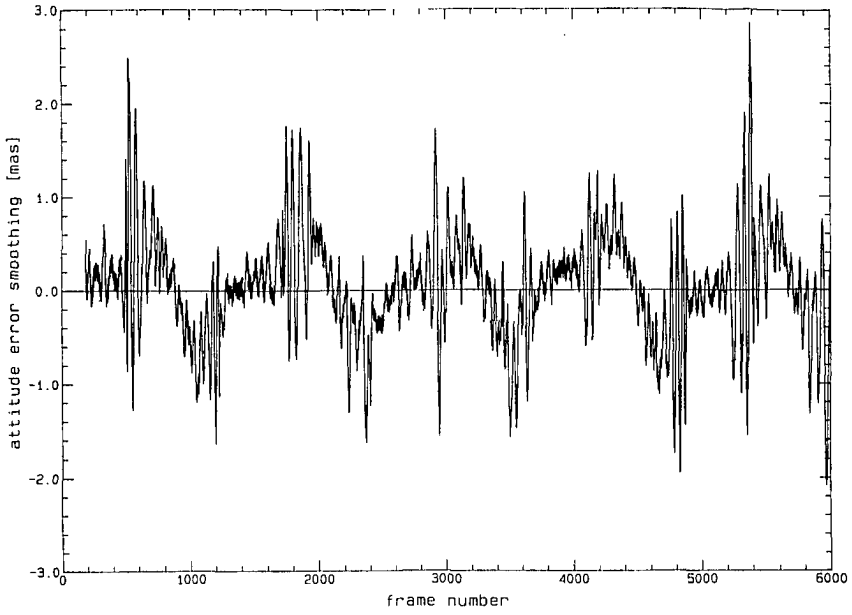


Figure 6.12 - Modelling error (smoothing) for CERGA dataset II with 703 B-splines.

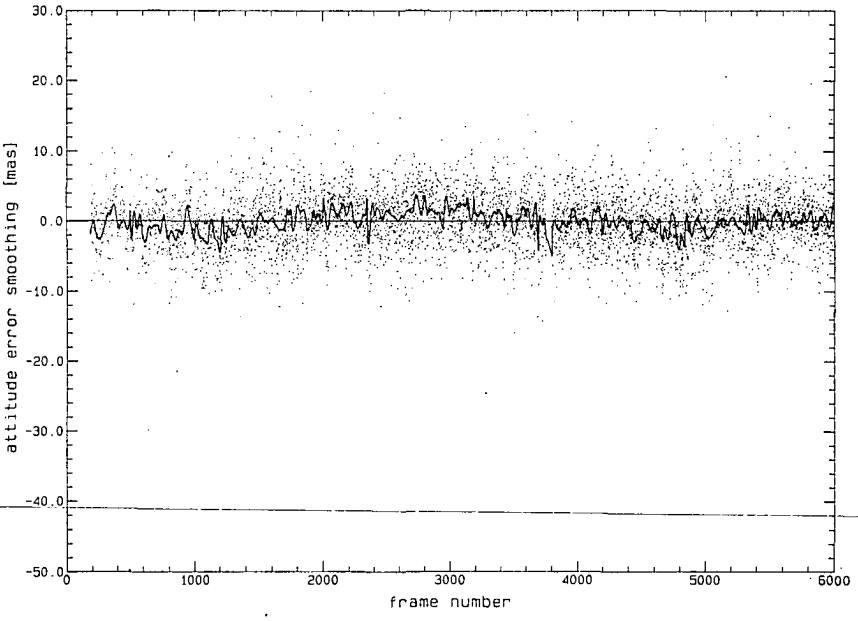


Figure 6.13 - True error in the frame by frame attitude for CERGA dataset II: geometric attititude (dots) and smoothed attitude (curve) using 704 B-spline parameters

Several more simulation experiments have been carried out for study purposes. In particular we have experimented with different observing strategies, magnitude distributions and star density variations over the RGC [*Van der Marel, 1985a, Verwaal, 1986, Van der Marel & van Daalen, 1986d*].

CHAPTER 7

NUMERICAL TECHNIQUES FOR THE GREAT CIRCLE REDUCTION

In this chapter the numerical methods actually used in FAST for solving the great circle reduction equations are considered. In the first part of this chapter the solution with one attitude parameter per frame is discussed and in the second part the smoothed solution is treated.

7.1 Introduction

The Great Circle Reduction forms a geometric adjustment problem on the celestial sphere with three types of unknowns: attitude and star abscissae along a chosen Reference Great Circle (RGC) and some instrumental parameters. The Great Circle Reduction handles about 10 hours of observations (grid phases), which cover a strip on the celestial sky of 2° wide. The reference great circle (RGC) is chosen somewhere in the middle of this strip. The adjustment problem is solved in a weighted least squares sense. Two types of attitudes can be produced, *viz.* a *geometric* and a *smoothed* attitude. The star ordinates and transversal attitude parameters are not solved; instead the great circle reduction is iterated several times with improved a-priori values obtained through sequences of preliminary versions of the Hipparcos catalogue (*i.e.* the three step procedure, see chapter 4 and section 5.3).

The linearized observation equations - in matrix notation - for the geometric solution read:

$$E\{\Delta y\} = \mathbf{A}\,\Delta x = \mathbf{A}_a \Delta x_a + \mathbf{A}_s \Delta x_s + \mathbf{A}_i \Delta x_i \qquad (7.1)$$

with Δx the vector of the small -unknown- corrections to the approximate values, Δy the vector with linearized observations, *i.e.* the observed value of y minus a value computed from approximate data on the unknown parameters, and design matrix $\mathbf{A}$ with partial derivatives $\partial y/\partial x$. The unknowns x and design matrix $\mathbf{A}$ are partitioned into an attitude, a star and an instrument part, respectively denoted by indices a, s and i. The geometric attitude is represented by one parameter per observation frame of 2.13 s. For a typical RGC-set of 5 revolutions there are about 70,000 equations (or observations) with about 17,000 geometric attitude unknowns, 2,000 star unknowns and some 50 instrumental unknowns. The submatrices $\mathbf{A}_a$ and $\mathbf{A}_s$ are very sparse, each of them contains only one non-zero element per row. $\mathbf{A}_i$ on the other hand is almost completely filled.

For attitude smoothing an additional equation $\Delta x_a = \mathbf{B}\,\Delta x_b$ has to be added. The observation equations are then

$$\Delta y = \mathbf{A}_a \mathbf{B}\,\Delta x_b + \mathbf{A}_s \Delta x_s + \mathbf{A}_i \Delta x_i \qquad (7.2)$$

The matrix elements $(\mathbf{B})_{kl} = \partial B(t_k)/\partial(x_b)_l$ and unknowns $(\Delta x_b)_l$ follow from our model for the attitude $(x_a)_t = B(t)$. In smoothing mode ~600 attitude parameters per RGC are needed, which is a considerable reduction compared to the 17,000

-geometric- attitude parameters. But the matrix $\mathbf{A}_a\mathbf{B}$, which has smaller column dimension than $\mathbf{A}_a$, is not as sparse as $\mathbf{A}_a$. The number of non-zero elements in $\mathbf{A}_a\mathbf{B}$ depends on the method of smoothing. If B-splines of order k are used the matrices $\mathbf{B}$ and $\mathbf{A}_a\mathbf{B}$ have k non-zero elements per row. The precise form of the equations are discussed in section 7.4.

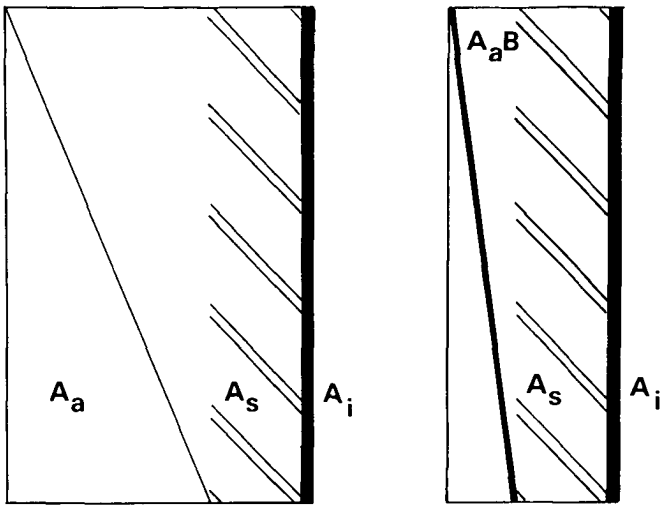


Figure 7.1: Non-zero structure of the design matrices with the unknowns in ascending order (left: geometric, right:smoothing)

The over-determined system of equations $\Delta\mathbf{y}=\mathbf{A}\cdot\Delta\mathbf{x}$ is solved in a weighted least squares sense. The solution $\Delta\hat{\mathbf{x}}$ is computed from the least squares criterion

$$\min_{\Delta\hat{\mathbf{x}}}\{ (\Delta\mathbf{y}-\mathbf{A}\cdot\Delta\hat{\mathbf{x}})\ \mathbf{W}\ (\Delta\mathbf{y}-\mathbf{A}\cdot\Delta\hat{\mathbf{x}})^T \} \qquad (7.3)$$

with $\mathbf{W} = \mathbf{C}_{yy}^{-1}$, the inverse of the covariance matrix of the observations $\mathbf{y}$. The error in the phase estimates, the observations $\mathbf{y}$, is dominated by photon noise, and therefore the observations are assumed to be uncorrelated. So a simple diagonal weight matrix $\mathbf{W}$ can be used.

The solution $\Delta\hat{\mathbf{x}}$ of the least squares problem can be computed from the so-called *normal equations*

$$\mathbf{A}^T\mathbf{W}\mathbf{A}\ \Delta\hat{\mathbf{x}} = \mathbf{A}^T\mathbf{W}\ \Delta\mathbf{y} \qquad (7.4)$$

with (semi-) positive definite normal matrix $\mathbf{A}^T\mathbf{W}\mathbf{A}$ and right hand sides $\mathbf{A}^T\mathbf{W}\ \Delta\mathbf{y}$. The normal equations for the great circle reduction are:
- regular, except for a rank defect of -usually- 1,
- well conditioned,
- very sparse (except for the instrument part).

The rank defect may be overcome through additional constraints. Two possible constraint equations are considered: 1) the correction to the so-called base star abscissae is zero, or 2) the sum of the corrections is zero (minimum norm solution). Under such constraint the normal matrix is regular and can be inverted.

Optimization of the computations is worthwhile, because there are approximately 1800 RGC sets during the mission, each of which has to be solved two or three times. The performance of the numerical techniques, in this chapter, will be evaluated in terms of *operation counts*. One operation is defined here as an multiplication, plus an addition, plus the necessary array accesses. In case the operands are floating point numbers the counts will be expressed in *flops* (floating point operations), or in more practical units as *Kflops* (10^3 flops) and *Mflops* (10^6 flops).

7.2 Choice of a Solution Method

7.2.1 Iterative versus Direct Methods

Numerical methods for the solution of linear systems of equations fall into two classes:
- iterative methods
- direct methods

The choice of an iterative method or direct method depends strongly on the problem at hand. In any case the solution method should, for reasons of computing efficiency, utilize the sparsity in the system of equations. The problem with direct methods is that they create *fill-in*, additional non-zeroes, in the matrix. But otherwise direct methods reach the exact solution (apart from round off errors) in a fixed number of arithmetic steps. Iterative methods, on the other hand, fully exploit the sparsity in the system of equations. Therefore, iterative methods require less storage and can use simpler data structures than direct methods.

The solution computed by an iterative method converges in principle to the exact solution, but it would require an infinite number iterations to attain this solution. However, by truncating the iterations, after some kind of stopping criterion has been fulfilled, a solution of the same accuracy as that of the direct solution can be computed. But also, if the desired accuracy is less than that of the direct solution method, the iteration process can be stopped earlier. In many problems the iterative solution can be computed faster than the direct solution, but there are also exceptions. On the other hand, some disadvantages of iterative methods are:
- generally the number of arithmetic operations cannot be predicted,
- every new right hand side of the equations, *e.g.* after an update for measurement errors, requires a completely new solution (although fewer iterations are required, because in most situations better start values are available),
- computation of variances is cumbersome and/or not very precise.

Direct methods can solve new right-hand sides without starting all over again. After the initial factorization, which is most of the work, new solutions are computed simply by forward and backward substitution with the different right-hand sides. Once the equations are factorized, covariance information can be obtained much faster, *viz.* in twice the time of the factorization itself (see section 7.3.5), than with iterative methods.

Both iterative and direct methods have been considered for the great circle reduction. A direct method, Choleski factorization, has been chosen because there were no *decisive* reasons to choose an iterative method. Choleski factorization is not a real bottleneck in the computations, though it requires quite a lot of computing time. Besides, iterative methods were not much faster. But a more important reason to use a direct method is that several solutions have to be computed. For instance, for the grid-step inconsistency correction several internal iterations are necessary.

7.2.2 Iterative Methods

Several iterative methods have been considered for solving the great circle reduction equations:
- conjugate gradient method [*Tommasini, 1983, Tommasini et al., 1985a*] and incomplete Choleski conjugate gradient method [*Benciolini et al., 1981a*],
- Burrows' iterative method [*Burrows, 1982*],
- Gauss-Seidel and successive overrelaxation (SOR) [*Joosten, 1986*].

The conjugate gradient method, which will be used in the sphere reconstitution, using the LSQR algorithm proposed by Paige and Saunders, has also been applied to the great circle reduction problem. The LSQR algorithm works directly on the design matrix, using column scaling as a preconditioning method. From simulation experiments it was found that, without instrumental parameters, several hundreds of iterations are necessary, but this number goes up rapidly with the number of instrumental parameters [*Tommasini et al., 1985a*]. The LSQR algorithm is used in combination with variance estimation: the quality of the estimated variances is good, although they are systematically underestimated by 10-20 % of the variances computed by inversion of the normal matrix. The LSQR algorithm was not significantly faster than Choleski factorization, especially with a realistic number of instrumental parameters.

A special iterative method is proposed by Burrows [*Burrows, 1982*]. His iteration formula also followed from our analytical treatment of the variances in section 5.5. This is typically a Jacobi or Gauss-Seidel kind of iteration process. Joosten showed, on a small regular example with 100 stars, that the results of the Burrows method correspond to SOR with a relaxation factor between the 1.8 and 1.9, resulting in already 80 iterations [*Joosten, 1986*]. Joosten also showed that, for his small example, the best results are obtained with a relaxation factor of 1.5, which resulted in 23 iterations. In general, for larger sizes of the network, the number of iterations goes up. It is expected that several hundreds of iterations are needed. Especially the results of the Burrows method, on the small example, are disappointing. Still Burrows stated that the iterative process may be stopped earlier, because the rigidity factor (see section 5.5) soon reaches its optimum value. However, as was explained by Van Daalen [*Van Daalen, 1983*], the rigidity factor is not necessarily the right function for controlling an iteration process.

A different kind of iterative procedure was briefly investigated by the author: namely, a block successive overrelaxation method on the two by two block partitioned normal matrix of equation (7.8), after elimination of the attitude parameters. This iteration process also converged rather slowly, so the experiments were stopped in an early stage. So, none of the above mentioned experiments pointed decisively in the direction of an iterative method.

7.2.3 Choice of a Direct Method

Here we consider two classes of direct methods for solving linear systems of equations as arise in least squares problems:
- factorization of the normal equations,
- orthogonalisation of the observation equations.

Both methods have in common that a triangular factor is formed. Orthogonalisation methods, *e.g.* Givens rotations and Householder transformations, work directly on the observation equations $\mathbf{Ax=y}$. The factorization methods, like the Choleski and Gauss method, operate on the normal equations $\mathbf{A}^T\mathbf{Ax}=\mathbf{A}^T\mathbf{y}$. Orthogonalisation methods are preferred for numerical stability: if programmed correctly they take advantage of the fact that the condition number of $\mathbf{A}$ is the square root of the condition number of $\mathbf{A}^T\mathbf{A}$. The

orthogonalisation methods may also work on the normal equations, but then they loose their advantages over the factorization methods.

The equations during the great circle reduction are well conditioned. Therefore we prefer the above mentioned factorization methods over the orthogonalisation methods, which are more work to compute. An estimate for the condition number (see appendix C) of the normal equations is

$$K_\infty = \| \mathbf{N} \|_\infty \| \mathbf{N}^{-1} \|_\infty \simeq 10^6 \qquad (7.5.a)$$

assuming that $\mathbf{N}$ is diagonal, then $\| \mathbf{N} \|_\infty \simeq 10^3$, the accumulated observation weight of a very bright star, and $\| \mathbf{N}^{-1} \|_\infty \simeq 10^3$, the variance of a very faint star. A different, but more common, estimate of the condition number may be obtained from the eigenvalues computed from the Fourier analysis in chapter 5 (figure 5.5):

$$K_2 = \frac{\lambda_{max}}{\lambda_{min}} \simeq 1.5\ 10^3 \qquad (7.5.b)$$

This estimate is smaller than the estimate of (7.5.a), because in the Fourier analysis of chapter 5 the star magnitudes are assumed to be the same for each star. Therefore, in the estimate of (7.5.b), only the rigidity (strength) of the network is taken into account. The first estimate for the condition number is confirmed by the error matrix $\mathbf{E}$ computed after the Choleski factorization of $\mathbf{N}$. Let

$$\mathbf{E} = \mathbf{N} - \mathbf{L}\mathbf{L}^T \qquad (7.6)$$

then, for CERGA dataset II, $\|\mathbf{E}\|_\infty \simeq 10^{-9}$ and $\|\mathbf{E}\|_2 \simeq 10^{-10}$ on a VAX 750 with a round off error of $O(10^{-15})$. So, no specific numerical condition problems are expected, hence Choleski factorization of the normal equations is chosen as solution method.

7.3 Geometric Solution

7.3.1 Introduction

The system of normal equations - partitioned in an attitude, star and instrument part - is written in matrix notation as:

$$\begin{bmatrix} \mathbf{N}_{aa} & \mathbf{N}_{as} & \mathbf{N}_{ai} \\ \mathbf{N}_{sa} & \mathbf{N}_{ss} & \mathbf{N}_{si} \\ \mathbf{N}_{ia} & \mathbf{N}_{is} & \mathbf{N}_{ii} \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x}_a \\ \Delta\mathbf{x}_s \\ \Delta\mathbf{x}_i \end{bmatrix} = \begin{bmatrix} \mathbf{b}_a \\ \mathbf{b}_s \\ \mathbf{b}_i \end{bmatrix} \qquad (7.7)$$

with $\mathbf{N}_{pq} = \mathbf{A}_p^T \mathbf{W}_{yy} \mathbf{A}_q$ and $\mathbf{b}_p = \mathbf{A}_p^T \mathbf{W}_{yy} \Delta\mathbf{y}$ for $p,q = a,s,i$, and with weight matrix $\mathbf{W}_{yy}$, where $\sigma_0^2 \cdot \mathbf{W}_{yy}^{-1} = \mathbf{C}_{yy} = E\{ (\mathbf{y}-E\{\mathbf{y}\})^T \cdot (\mathbf{y}-E\{\mathbf{y}\}) \}$ and σ_0^2 the variance of unit weight. $\mathbf{N}_{aa}$ and $\mathbf{N}_{ss}$ are diagonal, $\mathbf{N}_{as} = \mathbf{N}_{sa}^T$ is sparse and the blocks pertaining to the instrument are full or almost full.

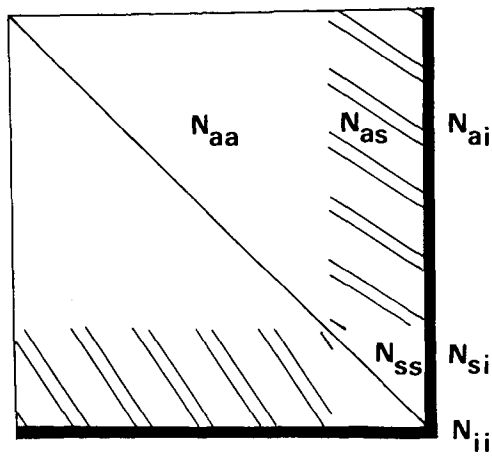


Figure 7.2: Non-zero structure of the normal matrix, with the unknowns in ascending order

Noting that the coefficients in **A** for the geometric solution are +1 or -1, with deviations less than 1%, or zero, and assuming uncorrelated observations, then the elements $(\mathbf{N}_{aa})_{kk}$ and $(\mathbf{N}_{ss})_{ii}$ are the accumulated observation weight spent in an observation frame k and on a star i respectively. An element $(-\mathbf{N}_{sa})_{ik}$ is the observation weight spent on star i in observation frame k. The number of non-zeroes in a column $(\mathbf{N}_{sa})_{*k}$ is equal to the number of observations in a frame k, *viz.* on the average 4 and maximally 10, the number of non-zeroes in a row $(\mathbf{N}_{sa})_{i*}$ is the number of frames in which a star is observed.

The full system of normal equations is never computed. Instead a reduced system, from which the attitude parameters are eliminated, is computed

$$\begin{pmatrix} \bar{\mathbf{N}}_{ss} & \bar{\mathbf{N}}_{si} \\ \bar{\mathbf{N}}_{is} & \bar{\mathbf{N}}_{ii} \end{pmatrix} \begin{pmatrix} \Delta\mathbf{x}_s \\ \Delta\mathbf{x}_i \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{b}}_s \\ \bar{\mathbf{b}}_i \end{pmatrix} \qquad (7.8)$$

with $\bar{\mathbf{N}}_{pq} = \mathbf{N}_{pq} - \mathbf{N}_{pa}\mathbf{N}_{aa}^{-1}\mathbf{N}_{aq}$ and $\bar{\mathbf{b}}_p = \mathbf{b}_p - \mathbf{N}_{pa}\mathbf{N}_{aa}^{-1}\mathbf{b}_a$ for $p,q \leftarrow s,i$. $\mathbf{N}_{aa}$ is diagonal, so the inverse matrix $\mathbf{N}_{aa}^{-1}$ can be computed on a frame by frame basis while updating the reduced normal equations. $\mathbf{N}_{aa}^{-1}\mathbf{b}_a$ is a provisional solution for the attitude parameters $\mathbf{x}_a$. $\bar{\mathbf{N}}_{ss}$ is not diagonal anymore, but still sparse: $(\bar{\mathbf{N}}_{ss})_{ij} \neq 0$ if star i and j have been simultaneously present in at least one frame. The instrumental blocks $\bar{\mathbf{N}}_{ii}$ and $\bar{\mathbf{N}}_{is}$ are almost completely full. The non-zero structure of the reduced normal equations is given in figure 7.3.

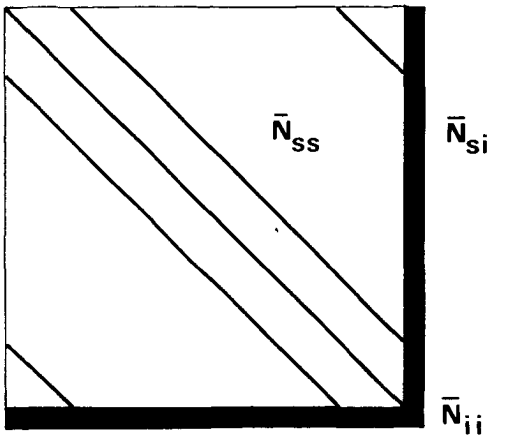


Figure 7.3: Non-zero structure of the reduced normal equations with the unknowns in ascending order

The block-partitioning sketched above not only enhances our insight in the non-zero structure of the problem, but it also offers some advantages during the computations,
1) an appropriate data structure can be used for each block,
2) at most one block at a time has to be stored *fully* in fast access computer memory.
The normal matrix blocks pertaining to the instrument are stored as full matrices, for the other blocks only the non-zeroes are stored, using the sifted format data structure (*Nz*) of appendix C.

The computation of the geometric solution is organized in the following steps:
1) computation of the reduced normal equations,
2) Choleski factorization of the block partitioned normal equations, solution of the equations and computation of the variances,
3) solution of the attitude parameters, computation of the residuals to the observations and testing of the solution.
The fill-in during Choleski factorization of the star part is reduced by changing the order of the star parameters before the factorization.

The order of the unknowns influences the amount of fill-in during factorization. This might also affect our block partitioning of the normal equations. In chapter 8 we prove that 1) the current order among the attitude, star and instrument blocks is optimal, and 2) each block of unknowns can be left intact. With a full normal matrix block for the instrument part and the current block partitioning the (internal) order of the attitude and instrument unknowns is irrelevant. Only reordering of the star unknowns reduces the fill-in during Choleski factorization. This is done by special ad-hoc ordering procedures, tailored to the structure of the problem at hand. The ordering procedures are discussed in chapter 8.

7.3.2 Computation of the Reduced Normal Equations

After linearization around approximate values for the attitude and star abscissae, ψ_k^0 and ψ_i^0, and for a -nominal- instrument $\mathbf{d}^0$, the correction equation for x_{ki}, the observed field coordinate (computed from the phase

estimate) minus the value computed from approximate data on the unknowns, of star i in frame k, becomes

$$\Delta x_{ki} = a_k \Delta\psi_k + b_{ki} \Delta\psi_i + \mathbf{c}_{ki}^T \cdot \Delta\mathbf{d} \qquad (7.9)$$

with coefficients $a_k \simeq -1$, $b_{ki} \simeq 1$ and $|(\mathbf{c}_{ki})_m| \lesssim 1$. The actual coefficients follow from table 5.1 and equation (5.7.b). It is typical for our approach that we compute the coefficients of the design matrix in a rigorous manner, instead of using some simple approximation (it is not much better, but it is also not much more work to compute). Let w_{ki} be the observation weight of Δx_{ki}. The contributions of the observations in frame k to the block partitioned normal equations are given in table 7.1, with

$$\begin{aligned} w_k &= \sum_i w_{ki} \\ \mathbf{c}_k &= \sum_i w_{ki}\mathbf{c}_{ki} / w_k \end{aligned} \qquad (7.10)$$

the total weight and mean instrumental coefficient in frame k, and

$$\Delta x_k = \sum_i w_{ki} \Delta x_{ki} \; / \; w_k \qquad (7.11)$$

the mean observation in frame k.

The blocks pertaining to the attitude parameters, $\mathbf{N}_{aa}$, $\mathbf{N}_{as}$ and $\mathbf{N}_{ai}$, are very large. Fortunately they have not to be computed explicitly since the attitude parameters can be eliminated directly from the normal equations, giving the reduced normal equations (7.8). The attitude unknowns are eliminated frame by frame, directly from the observation equations, by subtracting the mean observation Δx_k from each of the observations. The updating formulae for the normal equations are given in table 7.2. Note that a_k disappears from the equations.

Table 7.1: Normal matrix computation without attitude elimination (see also the text)

attitude part

$$(\mathbf{N}_{aa})_{kk} = w_k a_k^2 \qquad (\mathbf{N}_{as})_{ki} = w_{ki} a_k b_{ki} \quad \forall i$$

$$(\mathbf{N}_{ai})_{k*} = w_k a_k \mathbf{c}_k^T \qquad (\mathbf{b}_a)_k = w_k a_k \Delta x_k$$

updates star part

$$(\mathbf{N}_{ss})_{ii}^k = (\mathbf{N}_{ss})_{ii}^{k-1} + w_{ki} b_{ki}^2 \qquad (\mathbf{b}_s)_i^k = (\mathbf{b}_s)_i^{k-1} + w_{ki} b_{ki} \Delta x_{ki}$$

$$(\mathbf{N}_{si})_{i*}^k = (\mathbf{N}_{si})_{i*}^{k-1} + w_{ki} b_{ki} \mathbf{c}_{ki}^T \quad \forall i$$

updates instrument part

$$\mathbf{N}_{ii}^k = \mathbf{N}_{ii}^{k-1} + \sum_i (w_{ki} \mathbf{c}_{ki} \mathbf{c}_{ki}^T) \qquad \mathbf{b}_i^k = \mathbf{b}_i^{k-1} + \sum_i (w_{ki} \Delta x_{ki} \mathbf{c}_{ki})$$

Updating of the normal matrix parts $\bar{\mathbf{N}}_{si}$ and $\bar{\mathbf{N}}_{ii}$, which are almost completely full, and right hand sides $\bar{\mathbf{b}}_s$ and $\bar{\mathbf{b}}_i$ is relatively simple because these components are not stored in a compressed form. $\bar{\mathbf{N}}_{ss}$ is a sparse matrix and only the non-zero elements will be stored, using the *sifted format data structure* of appendix C.

In the great circle reduction software two different storage structures for sparse matrices are used: namely the *sifted format* for general sparse matrices and *envelope format* for sparse symmetric matrices, defined in appendix C. In the sifted format only the non-zero elements $Nz(\mathbf{A})$ are stored, where $Nz(\mathbf{A})$ is our notation for the set of non-zero elements of a sparse matrix $\mathbf{A}$. In the envelope format also some zeroes, which fall within the envelope $Env(\mathbf{A})$, are stored. The *envelope* $Env(\mathbf{A})$ of a sparse symmetric matrix $\mathbf{A}$ is the set of elements (zeroes and non-zeroes) in the lower triangle of the matrix, except the *leading* zeroes. *I.e.* $Env(\mathbf{A})$ is the set of matrix elements $(i,j) \in Env(\mathbf{A}) \iff \exists\, k \le j \quad (a_{ik} \neq 0)$.

Table 7.2: Normal matrix computation with attitude elimination (see also the text)

updates star part

$$(\bar{\mathbf{N}}_{ss})^k_{ii} = (\bar{\mathbf{N}}_{ss})^{k-1}_{ii} + w_{ki}(1-w_{ki}/w_k)b^2_{ki} \qquad \forall i$$

$$(\bar{\mathbf{N}}_{ss})^k_{ij} = (\bar{\mathbf{N}}_{ss})^{k-1}_{ij} - w_{ki}w_{ki}b_{ki}b_{ki}/w_k \qquad \forall i,\ \forall j \neq i$$

$$(\bar{\mathbf{b}}_s)^k_i = (\bar{\mathbf{b}}_s)^{k-1}_i + w_{ki}b_{ki}(\Delta x_{ki} - \Delta x_k) \qquad \forall i$$

$$(\bar{\mathbf{N}}_{si})^k_{i*} = (\bar{\mathbf{N}}_{si})^{k-1}_{i*} + w_{ki}b_{ki}(\mathbf{c}^T_{ki} - \mathbf{c}^T_k) \qquad \forall i$$

updates instrument part

$$\bar{\mathbf{N}}^k_{ii} = \bar{\mathbf{N}}^{k-1}_{ii} + \Sigma_i(w_{ki}\mathbf{c}_{ki}\mathbf{c}^T_{ki}) - w_k\,\mathbf{c}_k\,\mathbf{c}^T_k$$

$$\mathbf{b}^k_i = \mathbf{b}^{k-1}_i + \Sigma_i(w_{ki}\mathbf{c}_{ki}\Delta x_{ki}) - w_k\,\mathbf{c}_k\,\Delta x_k$$

Before we can start with the normal equation computation the precise non-zero structure $Nz(\bar{\mathbf{N}}_{ss})$ must be computed first. This is carried out in a first pass through the observations, the coefficients of the normal equations are computed in a second pass. An element $(\bar{\mathbf{N}}_{ss})_{ij}$, of the normal matrix block for the star part, is a non-zero only when the stars, i and j, are observed in the same observation frame. Therefore, during the first pass all combinations of stars i and j which are seen simultaneously in one frame are

stored in a circular list. At the same time the observation equations are linearized and written to a file on a frame by frame basis, and the observations are checked for grid step inconsistencies. At the end of the first pass active and passive stars are distinguished and the circular list is processed: Firstly it is checked that all active stars are connected to at least one other active star (not connected stars become passive), and a base star is assigned to each group of connected stars. Secondly the stars are reordered for a small profile ($|Env(\bar{\mathbf{N}}_{ss})|$), and thirdly the normal matrix administration of the -reordered- active stars ($Nz(\bar{\mathbf{N}}_{ss})$), is computed from the circular list. In the second pass the observation equations are read, and the reduced normal equations are computed on a frame by frame basis, according to the formulae in table 7.2.

The computation of the normal matrix parts $\bar{\mathbf{N}}_{si}$ and $\bar{\mathbf{N}}_{ii}$, and the elimination of the star parameters - as we will see later -, is a heavy computational burden (Table 7.3). This is a bit paradoxical, because only a relatively small number of instrumental parameters is involved. However $\mathbf{A}_i$, $\mathbf{N}_{si}$ and $\mathbf{N}_{ii}$ are practically full. In the example of table 7.3 approximately 44 Mflops are needed in the computation of $\bar{\mathbf{N}}_{si}$, $\bar{\mathbf{N}}_{ii}$ and $\bar{\mathbf{b}}_i$ (of which 10 Mflop for the attitude elimination) using the straightforward computation method, whereas for $\bar{\mathbf{N}}_{ii}$ and $\bar{\mathbf{b}}_i$ only 0.5 Mflops are needed. In the example of table 7.3 the number of instrumental parameters was quite moderate. For an increase in the number of instrumental parameters cpu times will increase quadratically.

7.3.3 Optimization of the Normal Matrix Computation

The operation counts in table 7.3 (first column of the example) are based upon straightforward computation of $\mathbf{N}_{si}$ and $\mathbf{N}_{ii}$, involving all possible products, including zero and duplicate products. More specific:

- $\mathbf{c}$ and $\mathbf{cc}^T$ contain zeroes (in case the P/F representation of section 5.4.2 is used),
- $\mathbf{cc}^T$ contains many duplicate products.

If the first property is used in the computations a speed-up by at most a factor of 2 in the computation of $\mathbf{b}_i$ and $\mathbf{N}_{is}$, and by a factor of 4 for $\mathbf{N}_{ii}$, is possible. The second property is only of some use in the computation of $\mathbf{N}_{ii}$, it gives an additional speed-up depending on the degree and form of the instrumental deformation polynomials.

The instrumental unknowns solved during the great circle reduction aim at describing the large scale distortion of the instrument (see section 5.4). The large scale distortion consists basically of three parts: 1) the distortion in the preceding field of view, 2) the distortion in the following field of view and 3) the basic angle deformation. The distortion d, using the P/F representation of section 5.4 (equation 5.15.b), is computed from

$$d = \mathbf{c}^T\mathbf{d} \qquad (7.12)$$

with $\mathbf{d}^T = (\mathbf{a}_p^T, \mathbf{a}_f^T, \mathbf{b}^T)$, and with $\mathbf{c}^T = (\mathbf{p}^T, \mathbf{0}, \mathbf{t}^T)$ or $\mathbf{c}^T = (\mathbf{0}, \mathbf{p}^T, -\mathbf{t}^T)$ for the

preceding and following field of view respectively. The vector **p** contains the powers $(B\text{-}V\ -0.5)^k\ \bar{x}^i \bar{y}^j$, $i=0,\ldots,n_k$, $j=0,\ldots,n_k-i$, $i+j\neq 0$ and $k=0,\ldots,n_a$ and the vector **t** contains the powers $(B\text{-}V\ -0.5)^l\ \bar{t}^m$, $m=0,\ldots,n_l$ and $l=0,\ldots,n_b$, where $\bar{x}$ and $\bar{y}$ are the normalized field coordinates, $B\text{-}V$ the star colour index and $\bar{t}$ the normalized RGC time. **a** and **b** are vectors with the unknown parameters a_{kij} and b_{lm} corresponding to the powers in **p** and **t**.

Table 7.3: Operation-counts (per frame) normal matrix computation, where $n^2/2 := n(n+1)/2$ and $\bar{n}_i := |Nz(\mathbf{c})|$. The Mflops are for an example with 18,000 frames, m=5 observations per frame and n_i=25 instrumental parameters: (1) straightforward method, *i.e.* $\bar{n}_i = n_i = 25$, (2) method of equation (7.13), *i.e.* $\bar{n}_i = 14$ and (3) final algorithm, *i.e.* $\bar{n}_i = 14$ and $\bar{n}_i^2$ follows from table 7.4. The " in the 2nd and 3rd column indicate that the results are identical to the 1st column.

	without att. elimin.				extra att. elimin.		
	OC	Mflops exampl. (1)	(2)	(3)	OC	Mflops exampl. (1)	(2)
w_k					$m+1$	.1	"
ϕ_k					$m+1$	.1	"
$\mathbf{c}_k$					$m\bar{n}_i+n_i$	2.6	1.6
$\mathbf{b}_s$	m	.1	"	"	1	.02	"
$\mathbf{b}_i$	$m\bar{n}_i$	2.3	1.3	"	n_i	.5	"
$\mathbf{N}_{ss}$	m	.1	"	"	$m^2/2$	.3	"
$\mathbf{N}_{is}$	$m\bar{n}_i$	2.3	1.3	"	n_i	.5	"
$\mathbf{N}_{ii}$	$m\bar{n}_i^2/2$	29.3	9.5	7.1	$n_i^2/2$	5.9	"
Total		34.1	12.3	9.9		10.0	9.0

Updating for the j'th observation, in the preceding field of view, gives for the normal matrices $\mathbf{N}_{ii}$ and $\mathbf{N}_{si}$

$$\mathbf{N}_{ii}^{(j)} = \mathbf{N}_{ii}^{(j-1)} + w_{ki}\ \mathbf{c}\ \mathbf{c}^T = \mathbf{N}_{ii}^{(j-1)} + w_{ki} \begin{bmatrix} \mathbf{p}\mathbf{p}^T & & \\ \mathbf{0} & 0 & \\ \bar{f}\ \mathbf{t}\mathbf{p}^T & 0 & \mathbf{t}\mathbf{t}^T \end{bmatrix}$$

$$\mathbf{N}_{si}^{(j)} = \mathbf{N}_{si}^{(j-1)} + w_{ki}\ b_{ki}\ \mathbf{c}^T = \mathbf{N}_{si}^{(j-1)} + w_{ki}\ b_{ki}\ (\mathbf{p}^T,\ 0,\ \bar{f}\ \mathbf{t}^T) \qquad (7.13)$$

where w_{ki} is the observation weight, b_{ki} is the coefficient of the star i and $\bar{f}$ the field index ($\bar{f}$=+1 for the preceding field of view, $\bar{f}$=-1 for the

following field of view). In order to obtain the updating formulae for the following field of view the first and second row and column have to be interchanged. If the zero multiplications are avoided considerable savings are obtained (second method of the example in table 7.3, with $\bar{n}_i$=11+3=14 instead of $\bar{n}_i$=25).

$\mathbf{pp}^T$ also contains duplicate powers. Assuming $\mathbf{p}$ to be linear in the terms *B-V*, we now introduce the notation $\mathbf{p}_{n_0,n_1}$, where n_0 is the maximum degree of the *xy* terms without colour terms *B-V*, and n_1 is the maximum degree of the *xy* terms with colour terms *B-V* (generally $n_0 > n_1$). Then $\mathbf{p}_{2n_0,2n_1}$ contains all the terms from $\mathbf{p}_{n_0,n_1}\mathbf{p}^T_{n_0,n_1}$, *e.g.* $\mathbf{p}_{0,3}\mathbf{p}^T_{0,3}$ is

$$\begin{bmatrix} y^2 & & & & \\ y^3 & y^4 & & & \\ xy & xy^2 & x^2 & & \\ xy^2 & xy^3 & x^2y & x^2y^2 & \\ x^2y & x^3y^2 & x^3 & x^3y & x^4 \end{bmatrix}$$

the powers xy^2, x^2y and x^2y^2 are redundant. The reductions for other degrees are summarized in table 7.4.

Table 7.4 - Reduction in the size of the polynomials

n_0	n_1	$\lvert\mathbf{p}_{n_0,n_1}\rvert$	$\lvert\mathbf{p}_{n_0,n_1}\mathbf{p}^T_{n_0,n_1}\rvert$	$\lvert\mathbf{p}_{2n_0,2n_1}\rvert$	perc.
2	0	5	15	12	80%
2	1	7	28	22	79%
3	0	9	45	25	56%
3	1	11	66	40	61%
3	2	16	136	55	40%
4	3	23	276	100	36%

The dimensions of $\mathbf{p}_{n_0 n_1}$, $\mathbf{p}_{n_0 n_1}\mathbf{p}^T_{n_0 n_1}$ and $\mathbf{p}_{2n_0 2n_1}$ are computed by the following formulae

$$
\begin{aligned}
|\mathbf{p}_{n_0,n_1}| &= \sum_{i=0}^{1}\left\{\frac{(n_i+1)(n_i+2)}{2} - 1\right\} \\
|\mathbf{p}_{n_0,n_1}\mathbf{p}^T_{n_0,n_1}| &= \tfrac{1}{2}\ |\mathbf{p}_{n_0,n_1}|\ (\ |\mathbf{p}_{n_0,n_1}| - 1\ ) \qquad (7.14)\\
|\mathbf{p}_{2n_0,2n_1}| &= \sum_{j=0}^{1}\sum_{i=0}^{j}\left\{\frac{(n_i+n_j+1)(n_i+n_j+2)}{2} - 3\right\}
\end{aligned}
$$

Similar optimizations for $\mathbf{t}\mathbf{t}^T$, which is almost always very small, as for $\mathbf{p}\mathbf{p}^T$ are not worthwhile. Similar optimizations for the contribution of the eliminated attitude are not possible. Firstly, $\mathbf{c}_k = \Sigma w_{ki}\mathbf{c}_{ki}$ is generally completely full, and secondly $\mathbf{c}_k\mathbf{c}_k^T$ does not contain duplicate products.

The algorithm for computing $\mathbf{N}_{ii}$ is given below. Two special vector/matrix operations are used: *Vec*(**M**) maps a matrix into a one dimensional array, *Mat*(**V**) undoes the operation.

algorithm *normal matrix computation instrument-instrument part*

1) initialize $\mathbf{N}_{ii} \leftarrow 0$, $\mathbf{C}_p \leftarrow 0$ and $\mathbf{C}_f \leftarrow 0$
2) for frame $k=1,2,\ldots,n_a$ compute
   2.1) $\mathbf{c}_k \leftarrow 0$, $w_k \leftarrow 0$
   2.2) for every observation $i=1,2,\ldots,m$ in this frame
   $w_k \leftarrow w_k + w_{ki}$
   if f =1 (preceding field of view) do
   $\mathbf{C}_p \leftarrow \mathbf{C}_p + w_{ki}\ (\mathbf{p}_{2n_0 2n_1},\ Vec(\mathbf{p}_{n_0 n_1}\mathbf{t}^T),\ Vec(\mathbf{t}\mathbf{t}^T)\ )$
   $\mathbf{c}_k^T \leftarrow \mathbf{c}_k^T + w_{ki}\ (\mathbf{p}^T,\ 0,\ \mathbf{t}^T\ )$
   if f =-1 (following field of view) do
   $\mathbf{C}_f \leftarrow \mathbf{C}_f + w_{ki}\ (\mathbf{p}_{2n_0 2n_1},\ -Vec(\mathbf{p}_{n_0 n_1}\mathbf{t}^T),\ Vec(\mathbf{t}\mathbf{t}^T)\ )$
   $\mathbf{c}_k^T \leftarrow \mathbf{c}_k^T + w_{ki}\ (0,\ \mathbf{p}^T,\ -\mathbf{t}^T\ )$
   2.3) $\mathbf{N}_{ii} \leftarrow \mathbf{N}_{ii} - \mathbf{c}_k\,\mathbf{c}_k^T / w_k$
3) $\mathbf{N}_{ii} \leftarrow \mathbf{N}_{ii} + Mat(\mathbf{C}_p) + Mat(\mathbf{C}_f)$

The savings are considerable: for instance in the example of table 7.3 not 34 Mflops but only 10 Mflops are needed for the normal matrix computation without attitude elimination, while the gain in the attitude elimination is only 1 Mflop (from 10 to 9 Mflops).

The large scale distortion of the field of view, $\mathbf{a}^T\mathbf{p}$, can be developed into a part common to both fields of view, $\mathbf{g}^T\mathbf{p}$, and a part containing half the difference between the two fields of view, $\mathbf{h}^T\mathbf{p}$. The distortion is then

$$d = (\mathbf{p}^T , f\,\mathbf{p}^T , f\,\mathbf{t}^T) \begin{bmatrix} \mathbf{g} \\ \mathbf{h} \\ \mathbf{b} \end{bmatrix} \qquad (7.15)$$

This is the so-called G/H representation of chapter 5 (equation 5.17). The parameters g_{kij} *and* h_{kij} from $\mathbf{g}$ and $\mathbf{h}$ can be computed from (5.18.a).

If the distortion is modelled in terms of $\mathbf{g}$ and $\mathbf{h}$ the updates to the normal matrices $\mathbf{N}_{ii}$ and $\mathbf{N}_{si}$ for the j'th observation become

$$\mathbf{N}_{ii}^{(j)} = \mathbf{N}_{ii}^{(j-1)} + w_{ik} \begin{bmatrix} \mathbf{p}\,\mathbf{p}^T & & \\ f\,\mathbf{p}\,\mathbf{p}^T & \mathbf{p}\,\mathbf{p}^T & \\ f\,\mathbf{t}\,\mathbf{p}^T & \mathbf{t}\,\mathbf{p}^T & \mathbf{t}\,\mathbf{t}^T \end{bmatrix} \qquad (7.16)$$

$$\mathbf{N}_{si}^{(j)} = \mathbf{N}_{si}^{(j-1)} + w_{ik}\, b_{ik}\, (\mathbf{p}^T, f\,\mathbf{p}^T, f\,\mathbf{t}^T)$$

It will be clear that straightforward computation with this definition of $\mathbf{c}$ leads to enormous duplication of work; of the 6 blocks in $\mathbf{N}_{ii}$ only 3 blocks - $\mathbf{p}\mathbf{p}^T$, $\mathbf{p}\mathbf{t}^T$ and $\mathbf{t}\mathbf{t}^T$ - are unique, and also $\mathbf{N}_{si}$ contains duplicate blocks, although the block pertaining to $\mathbf{h}$ may have a smaller degree. However, these normal matrices can also be computed indirectly from the normal matrices (7.13) for the P/F representation, even after the attitude is eliminated. Thereby all the computational advantages of the P/F representation can be used.

Due to the physics of the distortions the maximum degree of the $\mathbf{h}^T\mathbf{p}$ polynomial may be less than the maximum degree of the $\mathbf{g}^T\mathbf{p}$ polynomial. Therefore, the total amount of parameters needed for the G/H representation can be less than for the P/F representation. In this case the G/H representation should be used in the least squares estimation, *i.e.* during Choleski factorization and the forward and backward substitutions. However, the P/F representation is preferred during the normal equation computation and polynomial evaluations. Therefore, if desired, the normal matrices and right hand side pertaining to the P/F representation can be transformed to the G/H representation before the factorization of the instrument part. The solution, pertaining to the G/H representation, is transformed afterwards to the P/F representation.

7.3.4 Solving the Block Partitioned System

The normal equations which remain after attitude elimination are solved in block partitioned form. Let the Choleski factor $\mathbf{L}$ be partitioned in a star and an instrument part

$$\mathbf{L} = \begin{bmatrix} \mathbf{L}_{ss} & \\ \mathbf{L}_{si} & \mathbf{L}_{ii} \end{bmatrix} \qquad (7.17)$$

similarly to the reduced normal equations (7.8). The instrumental blocks are

stored as full matrices. The star part $\mathbf{L}_{ss}$ of the Choleski factor is sparse, but not as sparse as $\bar{\mathbf{N}}_{ss}$ in the reduced normal equations (7.8) for which only the non-zero elements $Nz(\bar{\mathbf{N}}_{ss})$ have to be stored. Depending on the ordering of the star parameters only the non-zero elements $Nz(\mathbf{L}_{ss})$ are stored ($|Nz(\mathbf{L}_{ss})| \geq |Nz(\bar{\mathbf{N}}_{ss})|$) or the envelope $Env(\mathbf{L}_{ss})$.

The Choleski factor can be computed with the *symmetric block factorization* algorithm of appendix C:

Algorithm *Symmetric block factorization (reduced normal matrix)*

1) factor $\bar{\mathbf{N}}_{ss}$ into $\mathbf{L}_{ss}\mathbf{L}_{ss}^T$
2) solve $\mathbf{L}_{si}$ from the triangular systems $\mathbf{L}_{ss}\mathbf{L}_{is}^T = \bar{\mathbf{N}}_{is}^T$
3) modify $\bar{\mathbf{N}}_{ii}$ into $\bar{\bar{\mathbf{N}}}_{ii} = \bar{\mathbf{N}}_{ii} - \mathbf{L}_{is}\mathbf{L}_{is}^T$
4) factor $\bar{\bar{\mathbf{N}}}_{ii}$ into $\mathbf{L}_{ii}\mathbf{L}_{ii}^T$

Let n_s and n_i be the dimensions of the star and instrument part respectively. Assume that before the factorization the star unknowns are reordered, *e.g.* by the modulo 60^0 ordering (see chapter 8), to reduce the bandwidth or profile of the Choleski factor. Then the operation counts for the different steps, with $n_s \simeq 1800$ and $n_i \simeq 25$, are of the order

$$\begin{aligned}
&1)\ O(\tfrac{1}{2}\omega^2 n_s) && \simeq 27 \text{ Mflop} \\
&2)\ O(\omega n_i n_s) && \simeq 7.8 \text{ Mflop} \\
&3)\ O(\tfrac{1}{2} n_i^2 n_s) && \simeq 0.6 \text{ Mflop} \\
&4)\ O(\tfrac{1}{6} n_i^3) && \simeq 2.6 \text{ Kflop}
\end{aligned} \qquad (7.18)$$

with ω the average frontwidth, the number of active rows during factorization (appendix C). In this example we assumed $\omega = 2\beta^c$, where $\beta^c \simeq 90$ is the cyclic bandwidth after the modulo 60^0 ordering (chapter 8).

The Choleski factorization of the star part is computationally intensive. Somewhat surprisingly the operation count for the second step, the elimination of star parameters from the instrument part, is of the same order of magnitude. So, taking into account also the efforts for the instrumental normal matrix computation, the computation of even a modest number of instrumental parameters involves already quite a heavy computation: *e.g.* another 27 Mflops just to compute 25 instrumental parameters (10 Mflops are needed for the normal matrix of the instrument part, 9 Mflops for the elimination of attitude parameters (see table 7.3) and another 7.8 Mflops for the elimination of star parameters (7.18)).

The storage requirements for the symmetric block factorization are at most $\omega n_s + n_s$ words ($\simeq$325 Kwords) which are needed to store $\mathbf{L}_{ss}$ completely and one vector of n_s long which is used to store columns of $\mathbf{L}_{is}$. Before the third step $\mathbf{L}_{ss}$ may be partly overwritten by $\mathbf{L}_{is}$, which is needed completely in memory ($n_s n_i \simeq$ 50-100 Kwords). A different algorithm for the block factorization, also suggested in appendix C, is the so-called a-symmetric

block factorization, which does not need $\mathbf{L}_{is}$ explicitly. A-symmetric block factorization is superior if $|Nz(\bar{\mathbf{N}}_{is})| \leq |Nz(\mathbf{L}_{is})|$. However, since this is not the case, and because $\mathbf{L}_{is}$ is needed explicitly later on for the covariance computation, the symmetric block factorization method has been chosen.

The solution of a two by two block partitioned system can be computed also with and without $\mathbf{L}_{is}$. The *standard solution scheme* uses $\mathbf{L}_{is}$, the *implicit solution scheme*, which is implemented in our software, uses $\bar{\mathbf{N}}_{is}$.

Algorithm *Implicit solution scheme*

1) solve $\Delta\mathbf{x}'_s$ $\quad \mathbf{L}_{ss}\mathbf{t}_s = \bar{\mathbf{b}}_s$ and $\mathbf{L}_{ss}^T \Delta\mathbf{x}'_s = \mathbf{t}_s$

2) compute $\quad \bar{\bar{\mathbf{b}}}_i = \bar{\mathbf{b}}_i - \bar{\mathbf{N}}_{is}\Delta\mathbf{x}'_s$

3) solve $\Delta\mathbf{x}_i$ $\quad \mathbf{L}_{ii}\mathbf{t}_i = \bar{\bar{\mathbf{b}}}_i$ and $\mathbf{L}_{ii}^T \mathbf{x}_i = \mathbf{t}_i$

4) compute $\quad \bar{\bar{\mathbf{b}}}_s = \bar{\mathbf{b}}_s - \bar{\mathbf{N}}_{is}^T \Delta\mathbf{x}_i$

5) solve $\Delta\mathbf{x}_s$ $\quad \mathbf{L}_{ss}\mathbf{t}_s = \bar{\bar{\mathbf{b}}}_s$ and $\mathbf{L}_{ss}^T \Delta\mathbf{x}_s = \mathbf{t}_s$

The star solution is computed twice, *viz.* with and without taking the instrument into account. The fourth and fifth step may be written also as an updating of the first solution, *i.e.*

4) solve $\Delta(\Delta\mathbf{x}_s)$ $\quad \mathbf{L}_{ss}\mathbf{t}_s = \bar{\mathbf{N}}_{is}^T \Delta\mathbf{x}_i$ and $\mathbf{L}_{ss}^T \Delta(\Delta\mathbf{x}_s) = \mathbf{t}_s$

5) update $\Delta\mathbf{x}_s$ $\quad \Delta\mathbf{x}_s = \Delta\mathbf{x}'_s + \Delta(\Delta\mathbf{x}_s)$

The operation counts for the star solution are of the order $O(2\omega n_s) \simeq 0.6$ Mflop (step 1 and 5), for the reduced right hand sides $O(n_i n_s) \simeq 0.04$ Mflop (step 2 and 4) and for the instrument solution $O(n_i^2) \simeq .6$ Kflop. In the standard solution scheme the star solution is computed only once.

Algorithm *Standard solution scheme*

1) solve $\mathbf{t}_s$ $\quad \mathbf{L}_{ss}\mathbf{t}_s = \bar{\mathbf{b}}_s$

2) compute $\quad \bar{\bar{\mathbf{b}}}_i = \bar{\mathbf{b}}_i - \mathbf{L}_{is}\mathbf{t}_s$

3) solve $\Delta\mathbf{x}_i$ $\quad \mathbf{L}_{ii}\mathbf{t}_i = \bar{\bar{\mathbf{b}}}_i$ and $\mathbf{L}_{ii}^T \Delta\mathbf{x}_i = \mathbf{t}_i$

4) update $\mathbf{t}_s$ $\quad \bar{\bar{\mathbf{t}}}_s = \mathbf{t}_s - \mathbf{L}_{is}^T \Delta\mathbf{x}_i$

5) solve $\Delta\mathbf{x}_s$ $\quad \mathbf{L}_{ss}^T \Delta\mathbf{x}_s = \bar{\bar{\mathbf{t}}}_s$

In the standard solution scheme $\mathbf{L}_{ss}$, $\mathbf{L}_{is}$ and $\mathbf{L}_{ii}$ are needed. In the implicit solution scheme $\mathbf{N}_{is}$ is used instead of $\mathbf{L}_{is}$. The implicit solution scheme is faster than the standard solution scheme if $|Nz(\mathbf{N}_{is})| + |Nz(\mathbf{L}_{ss})| \leq |Nz(\mathbf{L}_{is})|$. This is not the case for the great circle reduction; the standard solution scheme is a factor 2 faster. In the actual software still the implicit solution scheme is used, mainly for historical reasons, but also because the

implicit solution scheme gives some additional control over the adjustment and because it may be, if updating is used, numerically better.

An alternative method for the solution of a two by two block partitioned system is given in [*Van Daalen, 1983*]. This method is more or less similar to the implicit solution method, except that in step 2 and 4 the design matrix **A** and least squares residuals are used, and that the star parameters are eliminated from the design matrix first, before forming the instrumental normal equations and factorization of the instrument part.

7.3.5 Variance Computation

The covariance matrices of the instrument and star part are respectively $\sigma^2\bar{\bar{N}}_{ii}^{-1}$ and $\sigma^2\bar{\bar{N}}_{ss}^{-1}$. $\bar{\bar{N}}_{ii}$ is the normal matrix of the instrument part after elimination of the attitude and star part, which is computed as in the preceding section. $\bar{\bar{N}}_{ss}$ is the normal matrix of the star part, after elimination of the attitude and instrument part, *i.e.*

$$\bar{\bar{N}}_{ss} = \bar{N}_{ss} - \bar{N}_{si}\bar{\bar{N}}_{ii}^{-1}\bar{N}_{is}$$

$\bar{N}_{ii}$, $\bar{N}_{is}$ and $\bar{N}_{ss}$ are normal matrices after elimination of only the attitude part. $\bar{\bar{N}}_{ii}^{-1}$, which is very small, can be computed straightforwardly by solving the system of equations $L_{ii}(L_{ii}^T\bar{\bar{N}}_{ii}^{-1})=I$. On the other hand straightforward computation of $\bar{\bar{N}}_{ss}^{-1}$, which is completely full, is out of proportion: about $O(\omega n_s^2) \simeq 500$ Mflops are needed. Fortunately only a few terms of $\bar{\bar{N}}_{ss}^{-1}$ are interesting or will be used in subsequent stages of the reduction. In the sphere reconstitution and astrometric parameter extraction only the diagonal of the covariance matrix, which gives the variances of the star abscissae, is used. This may be done only if off-diagonal elements are not too large, which is not the case: from chapter 5 we know that "close by" stars, *e.g.* observed in the same frame, can have large covariances. However, these covariances, and some more, will be computed (as a byproduct) when the *sparse inverse* technique, described below, is used.

Partial inverses of the star part $\bar{\bar{N}}_{ss}^{-1}$ can be computed more efficiently using the *block inversion* algorithm, organized along $\bar{\bar{N}}_{ss}^{-1} = \bar{N}_{ss}^{-1} + \bar{N}_{ss}^{-1}\bar{N}_{si}\bar{\bar{N}}_{ii}^{-1}\bar{N}_{is}\bar{N}_{ss}^{-1}$ and the *sparse inverse* algorithm, which computes only those elements from the inverse $\bar{N}_{ss}^{-1}$ corresponding to $Nz(L_{ss})$. The symmetric block inversion algorithm (appendix C) is

algorithm *symmetric block inversion*

1) solve G_{is} from $L_{ii}^T G_{is} = L_{is}$ and solve F_{si} from $L_{ss}F_{si} = G_{is}^T$
2) compute $\bar{N}_{ss}^{-1}$ from L_{ss}
3) compute $\bar{\bar{N}}_{ss}^{-1} = \bar{N}_{ss}^{-1} + F_{si}F_{si}^T$

In step 2 the partial inverse is computed by the so-called *sparse inverse* technique. Typical for the sparse inverse is that only those elements in the inverse are computed which are non-zeroes in the Choleski factor. The

sparse inverse is computed column by column in a recursive manner, with formulae very much identical to the ones mentioned above (appendix C). In step 3 the contribution of the instrumental parameters to the covariances of the star part are added. The operations counts for each of the steps are

$$\begin{aligned} &1)\ O(\tfrac{1}{2}n_s(n_i^2+\omega n_i)) \simeq 4.5 \text{ Mflop} \\ &2)\ O(\omega^2 n_s) \simeq 54 \text{ Mflop (sparse inverse)} \\ &3)\ O(\eta n_i) \simeq 0.045 \text{ Mflop (only the diagonal, } i.e.\ \eta=n_s) \end{aligned} \qquad (7.19)$$

with ω the average frontwidth (the number of active rows during Choleski factorization) of the normal matrix block for the star part. The operation count for the third step depends strongly on the number of elements of the inverse which have to be computed. If only the diagonal is computed $\eta=n_s$, but if the elements corresponding to $Nz(\bar{\mathbf{N}}_{ss})$ are computed $\eta\approx\omega n_s$ and then the operation count is 7.8 Mflop. If the full inverse is computed $\eta=\frac{1}{2}n_s^2$ and the operation count is 41 Mflop.

The inverse of the instrument part $\bar{\bar{\mathbf{N}}}_{ii}^{-1}$ can be computed in a straightforward way, *viz.* by solving the systems $\mathbf{L}_{ii}(\mathbf{L}_{ii}^T(\bar{\bar{\mathbf{N}}}_{ii}^{-1}))=\mathbf{I}$. The rectangular matrix with the covariances of the instrument/star part, $\sigma^2\bar{\mathbf{N}}_{is}^{-1}$, can be computed by solving the triangular system $\mathbf{L}_{ii}^T\mathbf{N}_{is}^{-1}=\mathbf{F}_{is}$.

7.3.6 Computation of the Attitude and L.S. Residuals

The geometric attitude can be computed directly from the block partitioned normal equations: first new right hand sides

$$\bar{\bar{\mathbf{b}}}_a = \mathbf{b}_a - \mathbf{N}_{as}\Delta\mathbf{x}_s - \mathbf{N}_{ai}\Delta\mathbf{x}_i$$

are formed, then $\Delta\mathbf{x}_a = \mathbf{N}_{aa}^{-1}\bar{\bar{\mathbf{b}}}_a$ is solved. Since $\mathbf{N}_{aa}$ is diagonal, and because $\mathbf{N}_{aa}$, $\mathbf{N}_{as}$ and $\mathbf{N}_{ai}$ are were not computed explicitly, the geometric attitude is computed frame by frame from the observation equations, which are stored on disk. The correction to the geometric attitude estimate $(\hat{\mathbf{x}}_a)_k$ in the k'th frame is computed from

$$(\Delta\hat{\mathbf{x}}_a)_k = \frac{1}{a_k w_k}\sum_i \left(w_{ki}\left(\Delta x_{ki} - b_{ki}(\Delta\hat{\mathbf{x}}_s)_i - \Delta d_{ki}\right)\right) \qquad (7.20)$$

where $w_k=\Sigma w_{ki}$ and Δd_{ki} the correction to the instrumental distortion. If $a_k=-1$ then the attitude estimate is equal to the weighted mean of the least squares residuals after back substitution of the star and instrument parameters. The distortion may be computed from $\mathbf{c}^T\Delta\mathbf{d}$. However, since the coefficients $\mathbf{c}$ of the linearized equations are not stored, the instrumental deformation is computed directly by Horner's rule.

The residuals to the observations are computed at the same time as the attitude parameters, from

$$r_{ki} = \Delta x_{ki} - a_k(\Delta\hat{\mathbf{x}}_a)_k - b_{ki}(\Delta\hat{\mathbf{x}}_s)_i - \Delta d_{ki} \qquad (7.21)$$

The residuals are mainly used for testing purposes, notably the grid-step inconsistency correction.

The variances of the attitude parameters are computed in a similar way as the variances of the star part, only the computations are a bit more complicated:

$$\bar{\bar{N}}_{aa}^{-1} = N_{aa}^{-1} + N_{aa}^{-1} \begin{pmatrix} N_{as} & N_{ai} \end{pmatrix} \begin{bmatrix} \bar{\bar{N}}_{ss}^{-1} & \bar{\bar{N}}_{si}^{-1} \\ \bar{\bar{N}}_{is}^{-1} & \bar{\bar{N}}_{ii}^{-1} \end{bmatrix} \begin{pmatrix} N_{sa} \\ N_{ia} \end{pmatrix} N_{aa}^{-1} \qquad (7.22)$$

Simulation experiments indicate that the influence of the instrument on the variances of the attitude can be neglected and so the variances of the attitude are computed from

$$\bar{\bar{N}}_{aa}^{-1} = N_{aa}^{-1} + N_{aa}^{-1} N_{as} \bar{\bar{N}}_{ss}^{-1} N_{sa} N_{aa}^{-1} \qquad (7.23)$$

on a frame by frame basis. Only those elements of $\bar{\bar{N}}_{ss}^{-1}$ which are non-zero in the normal matrix, $Nz(\bar{N}_{ss})$, are needed. *I.e.*

$$(\bar{\bar{N}}_{aa}^{-1})_{kk} = \frac{1}{w_k a_k^2} \left(1 + \sum_i \sum_j \left(w_{ki} w_{kj} b_{ki} b_{kj} (\bar{\bar{N}}_{ss}^{-1})_{ij} \right) \right) \qquad (7.24)$$

7.4 Smoothed Solution

The smoothed solution is computed in a more or less similar way as the geometric solution. The equations are again partitioned in an attitude, star and instrument part, but now the star and attitude unknowns change roles: the stars are eliminated first, and the attitude unknowns - now much fewer than in the geometric mode - are reordered. Optionally the smoothed solution is computed as an update to the geometric solution.

7.4.1 Observation- and Normal Equations

Two different types of along scan attitude parameters are computed during the Great Circle Reduction: a *geometric* attitude $\mathbf{x}_a$, represented by one parameter per observation frame of 2.13 sec, and a *smoothed* attitude $\mathbf{x}_b$ which follows from the non-linear attitude model $(\mathbf{x}_a)_k = B(t_k, \mathbf{x}_b)$. After linearization around approximate values $B(t_k, \mathbf{x}_b^0)$, assuming time t_k non-stochastic, the linear relations

$$\Delta\bar{\mathbf{x}}_a = \mathbf{B}\, \Delta\mathbf{x}_b \qquad (7.25)$$

are obtained, with $(\mathbf{B})_{kl} = \partial B(t_k, \mathbf{x}_b)/\partial(\mathbf{x}_b)_l$ and $(\Delta\bar{\mathbf{x}}_a)_k = (\mathbf{x}_a)_k - B(t_k, \mathbf{x}_b^0)$. Thus the along scan attitude, computed during the geometric reduction step, can be further improved. In fact an additional adjustment of the geometric attitude is carried out. The improvement of the attitude entails also corrections to the star abscissae and instrumental parameters.

In matrix notation the complete linearized observation equations, with a geometric and smoothing reduction step, read

$$\begin{aligned} \Delta y &= A_a \Delta \bar{x}_a + A_s \Delta x_s + A_i \Delta x_i \\ \Delta \bar{x}_a &= B \, \Delta x_b \end{aligned} \qquad (7.26)$$

with Δx the vector of the small -unknown- corrections to approximate values for the unknowns, Δy the vector with linearized observations, *i.e.* the observed grid coordinate minus a value computed from approximate data, and design matrices A and B with partial derivatives as usual. The first part of the equations, which are partitioned into an attitude, star and instrument part, are identical to the observation equations for the geometric solution, except that different approximate values are now used for the attitude. The relation between the corrections to the attitude parameters in equation (7.26) and (7.1) is

$$(\Delta \bar{x}_a)_k = (x_a^0)_k - B(t_k, x_b^0) + (\Delta x_a)_k$$

The second part of the equations, the smoothing step, follows from the attitude model (7.25).

The observations equations (7.26) are to be solved in least squares sense. Two approaches are open: 1) compute the least squares solution of the combined equations

$$\Delta y = A_b \Delta x_b^s + A_s \Delta x_s^s + A_i \Delta x_i^s \qquad (7.27)$$

with $A_b = A_a B$, or 2) solve (7.26) in steps. It appears that the first approach will take less computing time, but in this way some control over the adjustment is lost. In the second approach the smoothed solution can be compared to the geometric solution, which is produced as an intermediate result containing fewer model assumptions than the smoothed solution, and thus the acceptability of the smoothed solution can be verified. Therefore the first approach is only used when the degree of smoothing has been well established and when there is already substantial experience with smoothing.

Let us assume that the geometric reduction step has been completed. Then the least squares solution $\hat{x}_b$ is computed from the normal equations

$$(B^T C_{aa}^{-1} B) \, \Delta \underline{\hat{x}}_b = B^T C_{aa}^{-1} \Delta \underline{\bar{x}}_a^g \qquad (7.28)$$

where the observations $\Delta \underline{\bar{x}}_a^g$ and weight matrix C_{aa}^{-1} follow from the geometric solution. Upper indices g for the geometric solution and upper indices s for the smoothed solution have been introduced. Underscores for stochastic variables, and a "^" for the least squares estimate, are used as usual. The "observations"

$$(\Delta \underline{\bar{x}}_a^g)_k = (\Delta \underline{\hat{x}}_a^g)_k - \left(B(t_k, x_b^0) - (x_a^0)_k \right) \qquad (7.29)$$

are identical to the attitude estimate computed during the geometric reduction step, except for a small non-stochastic term $B(t_k, x_b^0) - (x_a^0)_k$, which is necessary because different approximate values for the attitude have been used in the smoothing and geometric reduction steps. The corrections to the geometric attitude, $\Delta(\Delta \underline{x}_a^s) = \Delta \underline{x}_a^s - \Delta \underline{x}_a^g$, are equal to

$$\Delta(\Delta \underline{x}_a^s) = - \left(\Delta \underline{\bar{x}}_a^g - B \Delta \underline{\hat{x}}_b \right) \qquad (7.30)$$

The star abscissae and instrumental parameters are correlated with the geometric attitude, and therefore the improvement of the geometric attitude entails also corrections to the star- and instrumental parameters. The corrections to the star abscissae and instrumental parameters, $\Delta(\Delta\underline{x}^s_s)=\Delta\underline{x}^s_s-\Delta\underline{x}^g_s$ and $\Delta(\Delta\underline{x}^s_i)=\Delta\underline{x}^s_i-\Delta\underline{x}^g_i$, can be computed from the well known formulae

$$\begin{aligned}\Delta(\Delta\underline{x}^s_s) &= C_{sa}C_{aa}^{-1}\Delta(\Delta\underline{x}^s_a)\\ \Delta(\Delta\underline{x}^s_i) &= C_{ia}C_{aa}^{-1}\Delta(\Delta\underline{x}^s_a)\end{aligned} \qquad (7.31)$$

where C_{aa}, C_{as} and C_{ai} follow from the block partitioned inverse of the normal matrix in the geometric solution. But these formulae are not very efficient for sparse systems of equations.

A more efficient method is obtained if the observation equations for the smoothing step are written as

$$\begin{bmatrix}\Delta\bar{\underline{x}}^g_a\\ \Delta\bar{\underline{x}}^g_s\\ \Delta\bar{\underline{x}}^g_i\end{bmatrix} = \begin{bmatrix}0 & B & 0\\ I & 0 & 0\\ 0 & 0 & I\end{bmatrix}\begin{bmatrix}\Delta(\Delta\underline{x}^s_s)\\ \Delta\underline{x}^s_b\\ \Delta(\Delta\underline{x}^s_i)\end{bmatrix} \qquad (7.32)$$

where $\Delta(\Delta\underline{x}^s_s)=\Delta\underline{x}^s_s-\Delta\hat{\underline{x}}^g_s$, $\Delta(\Delta\underline{x}^s_i)=\Delta\underline{x}^s_i-\Delta\hat{\underline{x}}^g_i$, $\Delta\bar{\underline{x}}^g_s=\Delta\hat{\underline{x}}^g_s-\Delta\hat{\underline{x}}^g_s=\underline{0}$ and $\Delta\bar{\underline{x}}^g_i=\Delta\hat{\underline{x}}^g_i-\Delta\hat{\underline{x}}^g_i=\underline{0}$ are corrections to the geometric solution. The order of the blocks is different than in the geometric solution mode. It turns out that it is more efficient to solve first the star parameters and then the attitude parameters (see chapter 8). The "observations" for the star and instrument part are zero, which simplifies the equations somewhat, except when propagating variances. The normal equations, using as weight matrix the normal matrix from the geometric solution, are

$$\begin{bmatrix}N_{ss} & N_{sb} & N_{si}\\ N_{bs} & N_{bb} & N_{bi}\\ N_{is} & N_{ib} & N_{ii}\end{bmatrix}\begin{bmatrix}\Delta(\Delta\hat{\underline{x}}^s_s)\\ \Delta\hat{\underline{x}}^s_b\\ \Delta(\Delta\hat{\underline{x}}^s_i)\end{bmatrix} = \begin{bmatrix}\underline{b}^u_s\\ \underline{b}^u_b\\ \underline{b}^u_i\end{bmatrix} \qquad (7.33)$$

with $N_{pq}=A_p^T W_{yy} A_q$ for p,q=s,b,i, where $A_b=A_a B$, $N_{bb}=B^T N_{aa} B$ and $N_{*b}=N_{*a}B$ for *=a,s,i, and right hand sides $\underline{b}^u_*=N_{*a}\Delta\bar{\underline{x}}^g_a+N_{*s}\Delta\bar{\underline{x}}^g_s+N_{*i}\Delta\bar{\underline{x}}^g_i$ which can be simplified to $\underline{b}^u_*=N_{*a}\Delta\bar{\underline{x}}^g_a=A_*^T W_{yy}\Delta\bar{\underline{x}}^g_a$ for *=a,s,i, and with $\sigma_0^2\cdot W_{yy}^{-1}=C_{yy}=E\{(\Delta y-E\{\Delta y\})^T\cdot(\Delta y-E\{\Delta y\})\}$. Note that the right hand sides $\underline{b}^u_s$ and $\underline{b}^u_i$ are different from the right hand sides in the geometric reduction step.

The system of normal equations - partitioned in a star, attitude and instrument part - computed from the combined observation equations are

$$\begin{bmatrix}N_{ss} & N_{sb} & N_{si}\\ N_{bs} & N_{bb} & N_{bi}\\ N_{is} & N_{ib} & N_{ii}\end{bmatrix}\begin{bmatrix}\Delta\hat{\underline{x}}^s_s\\ \Delta\hat{\underline{x}}^s_b\\ \Delta\hat{\underline{x}}^s_i\end{bmatrix} = \begin{bmatrix}\underline{b}^s_s\\ \underline{b}^s_b\\ \underline{b}^s_i\end{bmatrix} \qquad (7.34)$$

with $N_{pq}=A_p^T W_{yy} A_q$ and $\underline{b}^s_p=A_p^T W_{yy}\Delta y$ for p,q=s,b,i. The normal matrices in (7.33) and (7.34), neglecting second order differences due to different approximate values for the attitude, are similar. However the right hand sides in (7.33), (7.34) and in the geometric reduction step are different. In

the geometric reduction step approximate values $\mathbf{x}_a^0$ for the attitude are used, in (7.34) approximate values $\mathbf{x}_b^0$, with $\mathbf{x}_a^0 \neq \mathbf{B}\mathbf{x}_b^0$, are used, while in (7.33) both of them are used. In (7.33) the smoothed solution is computed as an update to the geometric solution are computed.

So far no assumptions about the attitude model were made, except that it must be linearizable and able to represent the attitude at milliarcsecond level. In chapter 6 two different ways of smoothing are distinguished: dynamical and numerical smoothing. In this chapter especially numerical smoothing with B-splines is considered. We prefer spline functions since each spline is non-zero over a small domain and with spline functions it is possible to extend the analytical representation over the gas-jets [*Van der Marel, 1983c, 1985a*].

Let us call the domain over which a B-spline is non-zero a *superframe*. Like ordinary observation frames superframes overlap, the number of superframes which overlap is equal to the *order* k of the spline. Superframes are longer than ordinary frames and have a variable length. Contrary to ordinary frames superframes cannot be described by single attitude parameter, but k parameters are sufficient, and two consecutive superframes have usually $k-1$ parameters in common. Therefore the total amount of attitude parameters does not exceed the number of B-splines. Simulation experiments indicate that about 600 parameters $\mathbf{x}_b$ are needed to model the attitude with sufficient precision. This is a considerable reduction compared to the 17,000 geometric attitude parameters $\mathbf{x}_a$.

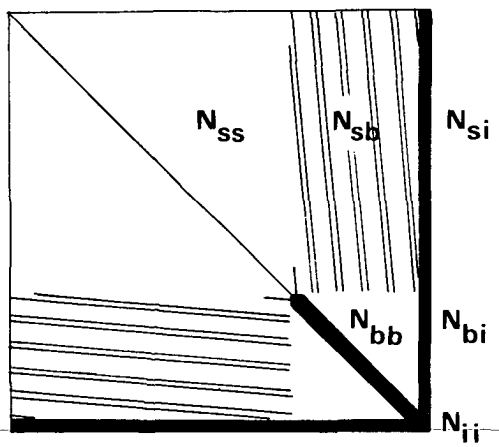


Figure 7.4: Non-zero structure of the normal matrix for smoothing.

In figure 7.4 the non-zero structure of the normal equations (7.33) is given. The non-zero structure of the star and instrument part are identical to the geometric solution; $\mathbf{A}_s$ and $\mathbf{N}_{ss}$ contain only one non-zero element per row, $\mathbf{A}_i$, $\mathbf{N}_{ii}$, $\mathbf{N}_{is}$, but also $\mathbf{N}_{bi}$, are almost completely full. Each spline is non-zero over a small domain, and therefore the matrix $\mathbf{B}$ has only few non-zero elements. Let k be the order of the B-splines, then $\mathbf{B}$ and $\mathbf{A}_a\mathbf{B}$

contain precisely k non-zeroes per row. $\mathbf{N}_{bb}=\mathbf{B}^T\mathbf{N}_{aa}\mathbf{B}$ is banded, with bandwidth k (*i.e.* $2k+1$ non-zeroes per row). The non-zero structure of $\mathbf{N}_{bs}=\mathbf{N}_{sb}^T$ are identical to $\mathbf{N}_{as}$, except for the compressed row dimension and thicker bands.

The normal matrix blocks pertaining to the smoothed attitude parameters are smaller but relatively denser than for the geometric parameters. $(\mathbf{N}_{bs})_{li}\neq 0$ if and only if star i is observed in the l'th superframe. Per scan a star is observed once in the preceding and once in the following field, during each passage it is seen on not more than $2k$ superframes. So there are approximately $2k\bar{c}$ non-zeroes in each column of $\mathbf{N}_{bs}$, with $\bar{c}$ the average number of scan circles a star is observed on. Let n_a, n_b, n_s and n_i be respectively the number of geometric attitude, spline, star and instrument parameters, and c the number of scan circles for this RGC. Then the average number of non-zeroes in a row of $\mathbf{N}_{bs}$ is approximately $2k\bar{c}(n_s/n_b)$.

7.4.2 Solving the Block Partitioned System

The block partitioned normal equations (7.33) and (7.34) of the smoothing (updating) and combined reduction step can be solved in similar ways, using the symmetric block factorization algorithm and standard solution scheme of appendix C.

Algorithm *Smoothed solution scheme*

1) *Eliminate stars*

1.1) compute $\mathbf{L}_{ss}^{-1}=\mathbf{N}_{ss}^{-1/2}$ ($\mathbf{N}_{ss}$ and $\mathbf{L}_{ss}$ are diagonal)

1.2) compute $\begin{pmatrix}\mathbf{L}_{bs}\\ \mathbf{L}_{is}\end{pmatrix}=\begin{pmatrix}\mathbf{N}_{bs}\\ \mathbf{N}_{is}\end{pmatrix}\mathbf{L}_{ss}^{-1}$ and $\mathbf{t}'_s=\mathbf{b}_s\mathbf{L}_{ss}^{-1}$.

1.3) compute $\begin{pmatrix}\bar{\mathbf{N}}_{bb} & \bar{\mathbf{N}}_{bi}\\ \bar{\mathbf{N}}_{ib} & \bar{\mathbf{N}}_{ii}\end{pmatrix}=\begin{pmatrix}\mathbf{N}_{bb} & \mathbf{N}_{bi}\\ \mathbf{N}_{ib} & \mathbf{N}_{ii}\end{pmatrix}-\begin{pmatrix}\mathbf{L}_{bs}\\ \mathbf{L}_{is}\end{pmatrix}\begin{pmatrix}\mathbf{L}_{bs}^T & \mathbf{L}_{is}^T\end{pmatrix}$

$$\begin{pmatrix}\bar{\mathbf{b}}_b\\ \bar{\mathbf{b}}_i\end{pmatrix}=\begin{pmatrix}\mathbf{b}_b\\ \mathbf{b}_i\end{pmatrix}-\begin{pmatrix}\mathbf{L}_{bs}\\ \mathbf{L}_{is}\end{pmatrix}\mathbf{t}'_s$$

2) *Factor and solve attitude/instrument part*

2.1) Factor $\begin{pmatrix}\bar{\mathbf{N}}_{bb} & \bar{\mathbf{N}}_{bi}\\ \bar{\mathbf{N}}_{ib} & \bar{\mathbf{N}}_{ii}\end{pmatrix}$ into $\begin{pmatrix}\mathbf{L}_{bb} & \\ \mathbf{L}_{ib} & \mathbf{L}_{ii}\end{pmatrix}$

2.2) Solve $\begin{pmatrix}\mathbf{L}_{bb} & \\ \mathbf{L}_{ib} & \mathbf{L}_{ii}\end{pmatrix}\begin{pmatrix}\mathbf{L}_{bb}^T & \mathbf{L}_{ib}^T\\ & \mathbf{L}_{ii}^T\end{pmatrix}\begin{pmatrix}\Delta\hat{\mathbf{x}}_b^s\\ \Delta(\Delta\hat{\mathbf{x}}_i^s)\end{pmatrix}=\begin{pmatrix}\bar{\mathbf{b}}_b\\ \bar{\mathbf{b}}_i\end{pmatrix}$

3) *Update star part*

3.1) compute $t''_s = t'_s - \begin{pmatrix} L_{bs}^T & L_{is}^T \end{pmatrix} \begin{bmatrix} \Delta\hat{x}_b^s \\ \Delta(\Delta\hat{x}_i^s) \end{bmatrix}$

3.2) solve (compute) $\Delta(\Delta\hat{x}_s^s) = L_{ss}^{-1} t''_s$

Symmetric block factorization is chosen because $Nz(\mathbf{N}_{bs})=Nz(\mathbf{L}_{bs})$, and therefore a-symmetric block factorization is not more efficient. Symmetric block factorization may be applied also in the second step, but this step can also be solved integrally. $\mathbf{N}_{bb}$, $\bar{\mathbf{N}}_{bb}$ and $\mathbf{L}_{bb}$ are stored in the envelope data structure, and the instrument part, which is full, can be appended to these matrices without loss of efficiency. $\mathbf{N}_{bs}$ and $\mathbf{L}_{bs}$ are stored in sifted format, and $\mathbf{L}_{bs}$ may overwrite $\mathbf{N}_{bs}$.

The instrumental parameters are already very well determined in the geometric reduction step. Therefore if the updated solution is computed the updating of the instrumental parameters may be omitted. However if the smoothed solution is computed directly, without computing the geometric solution first, the instrumental parameters must certainly be solved, unless they are already correct.

7.4.3 Covariance Computation

The variances of the star parameters can be computed from the the following formula, which follows from *symmetric block inversion* algorithm of appendix C,

$$\mathbf{C}_{ss} = \mathbf{L}_{ss}^{-1}\mathbf{L}_{ss}^{-T} + \begin{pmatrix} \mathbf{L}_{sb} & \mathbf{L}_{si} \end{pmatrix} \begin{bmatrix} \bar{\mathbf{N}}_{bb} & \bar{\mathbf{N}}_{bi} \\ \bar{\mathbf{N}}_{ib} & \bar{\mathbf{N}}_{ii} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{L}_{bs} \\ \mathbf{L}_{is} \end{bmatrix} \qquad (7.35.a)$$

or, written out in full,

$$\mathbf{C}_{ss} = \mathbf{N}_{ss}^{-1} + \mathbf{N}_{ss}^{-1} \begin{pmatrix} \mathbf{N}_{sb} & \mathbf{N}_{si} \end{pmatrix} \begin{bmatrix} \bar{\mathbf{N}}_{bb} & \bar{\mathbf{N}}_{bi} \\ \bar{\mathbf{N}}_{ib} & \bar{\mathbf{N}}_{ii} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{N}_{bs} \\ \mathbf{N}_{is} \end{bmatrix} \mathbf{N}_{ss}^{-1} \qquad (7.35.b)$$

The partial inverse of the attitude/instrument part may be obtained by the sparse inverse technique (appendix C), which gives only the elements in $Nz(\bar{\mathbf{N}}_{bb})$, $Nz(\bar{\mathbf{N}}_{bi})$ and $Nz(\bar{\mathbf{N}}_{ii})$. But if we only have to compute the diagonal of $\mathbf{C}_{ss}$ this is sufficient. Simulation experiments indicate that the influence of the instrument on the variances of the stars and attitude can be neglected - if the instrumental parameters are computed at all -, so the variances of the stars may be computed from

$$\mathbf{C}_{ss} = \mathbf{N}_{ss}^{-1} + \mathbf{N}_{ss}^{-1}\mathbf{N}_{sb}\bar{\mathbf{N}}_{bb}^{-1}\mathbf{N}_{bs}\mathbf{N}_{ss}^{-1}. \qquad (7.36)$$

A different approach for computing the variances of the stars can be derived from the updating formula in the smoothed solution scheme

$$\Delta(\Delta\hat{\underline{x}}_s^s) = \mathbf{N}_{ss}^{-1}\mathbf{b}_s - \mathbf{N}_{ss}^{-1}\begin{pmatrix}\mathbf{N}_{sb} & \mathbf{N}_{si}\end{pmatrix}\begin{bmatrix}\Delta\hat{\underline{x}}_b^s \\ \Delta(\Delta\hat{\underline{x}}_i^s)\end{bmatrix}$$

with $\underline{b}_s = \mathbf{N}_{sa}\Delta\bar{\underline{x}}_a^g + \mathbf{N}_{ss}\Delta\bar{\underline{x}}_s^g + \mathbf{N}_{si}\Delta\bar{\underline{x}}_i^g$, substitution gives

$$\Delta(\Delta\hat{\underline{x}}_s^s) = \Delta\bar{\underline{x}}_s^g - \mathbf{N}_{ss}^{-1}\begin{pmatrix}\mathbf{N}_{sa} & \mathbf{N}_{si}\end{pmatrix}\begin{bmatrix}\mathbf{B}\Delta\hat{\underline{x}}_b^s - \Delta\bar{\underline{x}}_a^g \\ \Delta(\Delta\hat{\underline{x}}_i^s) - \Delta\bar{\underline{x}}_i^g\end{bmatrix} \qquad (7.37)$$

So the corresponding improvement in variances, neglecting the instrumental parameters, read

$$\mathbf{C}_{ss}^s = \mathbf{C}_{ss}^g - \mathbf{N}_{ss}^{-1}\mathbf{N}_{sa}(\mathbf{C}_{aa} - \mathbf{B}\mathbf{C}_{bb}\mathbf{B}^T)\mathbf{N}_{as}\mathbf{N}_{ss}^{-1} \qquad (7.38)$$

7.5 The Rank Defect during the Great Circle Reduction

7.5.1 Base Star Solution

The design and normal matrices in the great circle reduction do not have full rank, *i.e.* their columns are linearly dependent. In case all active stars are connected, directly or indirectly, to each other, the rank defect is one. This corresponds to a unknown zero point for the abscissae. Hence, the observations are invariant under an arbitrary shift of both the star and attitude abscissae. In the exceptional case that there are several groups of active stars, which are all connected within the group, but which do not have connections between stars belonging to different groups (in graph theory: there is more than one connected component), the rank defect is equal to the number of groups. In this case the adjustment can be carried out for each group separately, with for each group a rank defect of one. Therefore, we will only consider the case for which the rank defect is one. The generalization to a multiple rank defect is straightforward.

During the great circle reduction the rank defect is overcome in a simple way: some unknown does not get a correction and is fixed on its approximate value. Generally, the abscissa of some bright star, close to one of the scan circle nodes, is fixed: this is the so-called *base star*. This corresponds to skipping the corresponding column in **A** (and the corresponding row and column in **N**). This remedy for the rank defect is very attractive for its simplicity: it results in a smaller system and less fill-in during Choleski factorization.

The above mentioned remedy for the rank defect has the same effect as the following constraint:

$$\mathbf{B}\,\Delta\mathbf{x} = \mathbf{c} \qquad (7.39)$$

with $\mathbf{c}=\mathbf{0}$, $\mathbf{C}_{cc}=\mathbf{0}$ and **B** a $1\times n$ matrix with all zeroes, except for the element corresponding to the base star (appendix C). The solution and the covariance matrix of the unknowns, for a different choice of constraints (*i.e.* a different base star), can be computed from the so-called *S-transform* of appendix C [*Baarda, 1973*]. For instance, if s is the label of the new base

star, and the old base star is labelled r, then

$$(\Delta x)_i^{(s)} = (\Delta x)_i^{(r)} - (\Delta x)_s^{(r)}$$
$$(C)_{ij}^{(s)} = (C)_{ij}^{(r)} - (C)_{sj}^{(r)} - (C)_{is}^{(r)} + (C)_{ss}^{(r)} \qquad (7.40)$$

The complete solution is shifted over a value equal to minus the correction to star s before the S-transform, so that the correction to star s, the new base star, becomes zero after the S-transform.

7.5.2 Minimum Norm Solution

In case of the minimum norm solution a different choice of constraints for equation (7.39) has been used. Namely, $\mathbf{c}=\mathbf{0}$, as before, but $\mathbf{B}$ is chosen is such a way that

$$\mathbf{A}\,\mathbf{B}^T = \mathbf{0} \qquad (7.41)$$

i.e. $\mathbf{B}^T$ is a basis for the null space of $\mathbf{A}$ (rank 1). A possible choice for $\mathbf{B}$ is a vector consisting of all ones for the attitude and star part, and zeroes for the instrument part. Actually, we will only consider here the system after elimination of the attitude parameters: a possible choice for $\mathbf{B}$ is then a vector with all ones for the star part and zeroes for the instrument part. The minimum norm solution has some nice properties:

- the solution $\Delta\mathbf{x}$ has minimum norm,
- the covariance matrix $\mathbf{C}_{xx}$ has minimum trace.

In the case of the great circle reduction also the sum of the elements in $\Delta\mathbf{x}$, the correction to the approximate values, is zero.

The minimum norm solution can be computed from the base star solution by the following S-transform (here we consider only the star part, after elimination of the attitude part):

$$(\bar{\Delta x})_i = (\Delta x)_i^{(r)} - \frac{1}{N}\sum_k^N (\Delta x)_k^{(r)}$$
$$(\bar{C})_{ij} = (C)_{ij}^{(r)} - \frac{1}{N}\sum_k^N (C)_{kj}^{(r)} - \frac{1}{N}\sum_l^N (C)_{il}^{(r)} + \frac{1}{N^2}\sum_k^N\sum_l^N (C)_{kl}^{(r)} \qquad (7.42)$$

Again, the abscissae are only shifted, but from the original covariances the column-and-row-average must be subtracted, and the overall-average must be added. From eq. 7.42 it seems as of the complete covariance matrix is needed for the computation of the transformed covariance matrix.

The minimum norm solution can be computed more efficiently from the normal matrix system of equation (C.10) in appendix C

$$\begin{bmatrix} \mathbf{A}^T\mathbf{W}\mathbf{A} & \mathbf{B}^T \\ \mathbf{B}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{x}} \\ m \end{bmatrix} = \begin{bmatrix} \mathbf{A}^T\mathbf{W}\mathbf{y} \\ 0 \end{bmatrix} \qquad (7.43)$$

with m a Lagrange multiplier. This linear system is symmetric and regular, but **not** positive definite. Therefore this system cannot be solved with Choleski factorization. However, any N-1 dimensional submatrix of $\mathbf{A}^T\mathbf{W}\mathbf{A}$ is positive definite, and can be factored by the Choleski method. The remaining 2-dimensional part, after reduction by the Choleski factor, can be inverted by Gauss' method.

Let the normal matrix be partitioned in an N-d dimensional block and two d dimensional blocks, then N-d steps of Choleski factorization, with reduction of the remaining part, give

$$\begin{bmatrix} N_{11} & N_{21}^T & B_1^T \\ N_{21} & N_{22} & B_2^T \\ B_1 & B_2 & 0 \end{bmatrix} \xrightarrow[\text{Choleski}]{N\text{-}d \text{ steps}} \begin{bmatrix} L_{11} & 0 & 0 \\ L_{21} & 0 & \bar{B}_2^T \\ \bar{B}_1 & \bar{B}_2 & -D \end{bmatrix} \tag{7.44}$$

with $\bar{B}_1 = B_1 L_{11}^{-1}$, $\bar{B}_2 = B_2 - B_1 N_{11}^{-1} N_{12} = B_2 - \bar{B}_1 L_{21}^T$, $D = B_1 N_{11}^{-1} B_1^T = \bar{B}_1 \bar{B}_1^T$ and $\bar{N}_{22} = N_{22} - N_{21} N_{11}^{-1} N_{12} = 0$ because of the rank defect. The inverse of the remaining four d-dimensional blocks can be computed by Gauss' method with pivoting, *i.e.*

$$\begin{bmatrix} 0 & \bar{B}_2^T \\ \bar{B}_2 & -D \end{bmatrix}^{-1} = \begin{bmatrix} \bar{B}_2^{-1} D \bar{B}_2^{-T} & \bar{B}_2^{-1} \\ \bar{B}_2^{-T} & 0 \end{bmatrix} \tag{7.45}$$

The remainder of the inverse follows by recursive partitioning using the sparse inverse technique.

In the operational software the solution will be computed with the base star approach. This solution can be transformed to the minimum norm solution by equation (7.42), but this is not really necessary, since in the next step of the reduction (sphere reconstitution) the zero point will be solved anyhow. It also does not matter which covariance matrix is used. However, in practice the full covariance matrix is not used, but only the diagonal. So, since the off-diagonal elements in the minimum norm covariance matrix are smaller, we believe that it is better to use the minimum norm variances than the base star variance in the next data reduction step. But also for analysis purposes the minimum norm variances are better. These are the reasons why, after the base star solution is computed, the sparse inverse for the minimum norm solution is computed from equation (7.44), with the 2-dimensional Gauss' inverse of equation (7.45) as starting point.

CHAPTER 8

ORDERING OF THE UNKNOWNS DURING THE GREAT CIRCLE REDUCTION

During the great circle reduction two large sparse systems of equations have to be solved. The order, in which the attitude, star and instrumental unknowns are computed, influences the efficiency of the adjustment process. In this chapter several possible, near optimal, orderings of the unknowns are discussed.

8.1 Introduction

The number of non-zeroes in the Choleski factor, and therefore storage requirements and computing resources, depend on the ordering (numbering) of the unknowns. The ordering does not influence the numerical stability of Choleski factorization as much as it does with Gaussian elimination. Therefore, 1) we may order without regard for numerical stability, and 2) we can order before the actual factorization takes place. Our objective is to find an ordering which minimizes computing resources, especially execution time, but also storage requirements.

During the Choleski factorization process "new" non-zero elements will be created in places where there was a zero in the normal matrix. The ideal ordering would be the one which minimizes the number of newly created non-zero elements in the Choleski factor, the so-called fill-in, but generally there is no algorithm which computes such an ordering in an tractable way. Therefore, an acceptable ordering is: 1) one which gives small, but not necessarily the minimum, fill-in, and 2) one which is produced by an efficient ordering algorithm. Two different types of ordering schemes are distinguished [*George & Liu, 1981, Duff, 1981*]:

1. schemes which cluster the non-zeroes around the diagonal, and so try to reduce the *envelope* of the non-zero elements in the normal matrix,
2. schemes which reduce the *fill-in* of the Choleski factor.

The choice for one of these schemes actually involves a choice of a data structure for the Choleski factor, *viz.* the *profile* or *variable band format* for 1) or the more general *sifted format* for 2).

Representatives of the ordering schemes that cluster the non-zeroes around the diagonal are the so-called *Reverse Cuthill-McKee* algorithm [*Cuthill-McKee, 1969, Gibbs et al, 1976, George & Liu, 1981*] and the *"banker's"* algorithm [*King, 1970, Snay, 1976*]. An example of an ordering scheme which aims at reducing the fill-in is the well known *minimum degree* algorithm [*George & Liu, 1981*]. The performance (in terms of fill-in) of the minimum degree algorithm is generally better than that of the schemes which just cluster the non-zeroes around the diagonal. However, with the minimum degree algorithm the non-zeroes will be scattered over the whole matrix and therefore a more complex data structure, *viz.* sifted format, has to be used to profit from this low number of non-zeroes, with more overhead in storage and access times. The envelope orderings use the simpler variable band or profile data structure with almost no overhead. Therefore the overall performance of the envelope orderings schemes could be better than the minimum degree algorithm, the more so since the minimum degree ordering is also more expensive to compute.

Most of the above mentioned schemes can be formulated most clearly in terms of the *graph* of the matrix, representing the connection structure of the unknowns (see the next section and appendix C). In the case of Hipparcos, the major characteristics of the graph, can be well captured by an *approximate graph*, derived solely from geometry: Two stars can only be connected if their abscissae differ by not more than the size of the field of view, possibly plus or minus the basic angle. The actual connection structure depends not only the difference of the star abscissae. Therefore, the approximate graph represents a more densely tied star network. A specific class of ordering schemes, the so-called *synthetic* orderings, are based on the approximate graph. The synthetic orderings are computed directly from the list of star abscissae and the reordering takes place before the actual normal matrix computation. An example of a synthetic ordering scheme is the *modulo* ordering (sec. 8.4.2), but also synthetic versions of the earlier mentioned graph oriented ordering schemes can be given.

In the next section first some terminology is introduced which will be used later on. Then a very special synthetic ordering is discussed, *viz.* the optimal order of the attitude, star and instrument blocks is established and proof is given that the blocks can be kept intact under the optimal overall ordering. It appears that only the star parameters in the geometric solution mode, and attitude parameters in the smoothing mode, have to be ordered further. Most of the principles of Choleski factorization, and the occurrence of fill-in are described in appendix C. Readers which are not familiar with sparse systems of equations might prefer to read this appendix first.

8.2 Terminology

Sparse matrices contain only -relatively- few non-zeroes. The zero elements have not to be stored and multiplication by a zero results again in a zero. The set of non-zero elements in a matrix **A** will be denoted by $Nz(\mathbf{A})$, *i.e.*

$$Nz(\mathbf{A}) = \{ (i,j) \mid a_{ij} \neq 0 \ , \ i=j \}$$

For symmetric matrices we prefer a slightly different definition: we will also require that $i \leq j$, *i.e.* $Nz(\mathbf{A})$ contains only the non-zero elements (i,j) from the lower triangle of the matrix. The *fill* of a sparse matrix is equal to $|Nz(\mathbf{A})|$, the number of non-zero elements in the matrix. The *bandwidth* $\beta_i(\mathbf{A})$ in the i'th row of a $N \times N$ sparse *symmetric* matrix **A** is

$$\beta_i(\mathbf{A}) = i - \min\{ j \mid a_{ij} \neq 0, \ 1 \leq j \leq i \}$$

The *envelope* $Env(\mathbf{A})$ of this matrix is defined by

$$Env(\mathbf{A}) = \{ (i,j) \mid i-\beta_i(\mathbf{A}) \leq j \leq i \}$$

i.e. the envelope of the matrix consists of all elements up to the diagonal, zeroes and non-zeroes, except the leading zeroes in a row. The *profile* $|Env(\mathbf{A})|$ of a matrix is equal to

$$|Env(\mathbf{A})| = \sum_{i=1}^{N} \beta_i(\mathbf{A})$$

the number of elements within the envelope. Obviously for any symmetric matrix **A** $Nz(\mathbf{A}) \subseteq Env(\mathbf{A})$ and $|Nz(\mathbf{A})| \leq |Env(\mathbf{A})|$.

Let L be the Choleski factor from the factorization $\mathbf{A}=\mathbf{L}\mathbf{L}^T$. Then we know from lemma (C.10) of appendix C that $Env(\mathbf{L})=Env(\mathbf{A})$. Now, let us define the frontwidth $\omega(\mathbf{A})$ of a symmetric N by N matrix $\mathbf{A}$ as

$$\omega_i(\mathbf{A}) = \left| \{ k \mid (k,i) \in Env(\mathbf{A}) \text{ for } k > i \} \right|$$

the number of rows of the envelope of $\mathbf{A}$ which "intersect" column i. But, then $\omega_i(\mathbf{L})=\omega_i(\mathbf{A})$ is the number of active rows at the i'th step in the Choleski factorization. Hence the operation count for computing the Choleski factor is

$$1/2 \sum_{i=1}^{N} \omega_i(\mathbf{A}) \, (\omega_i(\mathbf{A}) + 3) \;\leq\; 1/2 \sum_{i=1}^{N} \beta_i(\mathbf{A}) \, (\beta_i(\mathbf{A}) + 3) \quad .$$

and the operation count for solving the system of equations:

$$2 \sum_{i=1}^{N} (\omega_i(\mathbf{A}) + 1) \;\leq\; 2 \sum_{i=1}^{N} (\beta_i(\mathbf{A}) + 1)$$

The envelope $Env(\mathbf{A})$ and profile $|Env(\mathbf{A})|$ can also be computed from the frontwidth ω, using identical definitions as for the bandwidth.

The non-zero structure of a sparse symmetric matrix $\mathbf{A}$ can be associated with a graph G. A graph $G=(X,E)$ consists of a finite set X of *nodes* x_i together with a set E of unordered pairs of nodes $\{x_i,x_j\}$, called *edges*. Nodes are associated with unknowns and edges with a pair of unknowns connected through observations. A graph is *ordered* if the nodes are numbered. An ordered graph can be related to a symmetric matrix $\mathbf{A}$: the nodes can be associated with the diagonal entries in $\mathbf{A}$, edges can be associated with the off-diagonal non-zero elements in $\mathbf{A}$.

Two nodes x and y are *adjacent* if the pair $\{x,y\}$ is an element of E, the set of edges. Let Y be a subset of nodes (simple case: Y is a node), then the *adjacent set* of Y, denoted by $Adj(Y)$, is the set of nodes which are not in Y but are adjacent to at least one node in Y, and the *degree* $Deg(Y)$ of Y is simply the number of edges with nodes outside Y, *i.e.* $Deg(Y)=|Adj(Y)|$. A *section graph* consists of a subset of the nodes plus all edges between them, *i.e.* the section graph $G(Y)$ is the subgraph $G(Y,E(Y))$ of $G(X,E)$, with $Y \subset X$ and $E(Y) = \{ \{x,y\} \in E \mid x \in Y, \; y \in Y \}$.

The complete process of Choleski factorization (using the outer product formulation of appendix C) can be interpreted as a sequence of graph transformations

$$G_0 \rightarrow G_1 \rightarrow \; \ldots\ldots \; \rightarrow G_n$$

with G_0 the original graph G of the normal matrix $\mathbf{N}$, G_i the graph of the reduced normal matrix after i steps of Choleski factorization process and G_n the empty graph. The so-called *elimination* graphs G_i are computed recursively: G_{i+1} is computed from G_i by the following recipe:

1) delete the node x_i and all its incident edges,
2) add edges so that all nodes in $Adj(x_i)$ are pairwise adjacent in the new graph.

The *fill-in* of the Choleski factor L corresponds to the set of *new* edges added during the elimination process.

The order in which the nodes are eliminated influences the fill-in. The *minimum degree* reordering algorithm specifies that the node to be eliminated next in an elimination graph G_i must be of minimum degree in G_i. It will be intuitively clear that the number of edges to be added to G_{i+1} is then small. A more elaborate criterion than the minimum degree criterion is obtained if the edges which are already present in the elimination graph are not taken into account. Both methods aim at local minimization of the fill-in. In this sense minimum degree is a heuristic algorithm; it is not guaranteed to give minimum fill-in globally, although there is sufficient empirical evidence that it produces an ordering which gives a low fill-in. On the other hand global minimization of the fill-in is an NP-complete problem, and therefore generally out of the question.

8.3 Optimum Block Ordering in the Geometric Mode

Previously we have put intuitively the -geometric- attitude unknowns first, followed by the stars and finally the instrumental parameters. The questions, which are considered in this section, are 1) is this the optimal sequence, and 2) can the groups be left intact? Observe for the graph associated with the normal matrix of the great circle reduction (figure 7.2), that generally

- $Deg(\mathbf{x}_a) < Deg(\mathbf{x}_s) << Deg(\mathbf{x}_i)$, where $Deg(\mathbf{x}_a) \simeq 4$ (max. 10), and $Deg(\mathbf{x}_s) \simeq 40$, not counting the instrumental parameters, and $Deg(\mathbf{x}_i) \simeq 20,000$
- the edge set of the section graphs $G(\mathbf{x}_a)$ and $G(\mathbf{x}_s)$ are empty, *i.e.* $E(\mathbf{x}_a) = \emptyset$ and $E(\mathbf{x}_s) = \emptyset$.

This can be observed directly from the normal matrix for the geometric solution of figure 7.2: 1) the degree is equal to the number of non-zeroes in the corresponding row or column, and 2) the empty edge sets of the two sections graphs correspond with the diagonal normal matrix blocks for the attitude and star part. From the second observation follows that successive elimination of nodes of one of the section graphs first does not influence the degree (in the total graph) of the remaining nodes in the section graph.

These two observations suggest that, using the minimum degree criterion, in order to obtain minimum fill-in 1) attitude parameters must be eliminated from the graph first, because they have the lowest degree, and 2) they may be eliminated in any order, without consequences for the fill-in, because the edge set of the section graph for the attitude parameters is empty. There will be fill-in the section graph of the star parameters, but there is no fill-in in the section graph of the attitude parameters themselves. The elimination process is so simple that it is done during the normal matrix computation (sec. 7.3.2).

The *elimination* graph $G'=(X',E')$ after elimination of the attitude parameters, which can be associated with the normal matrix of figure 7.3, has the following properties:

- the edge set of the section graph of the star parameters is not empty anymore
- $Deg((\mathbf{x}_s)'_i) \geq \max_k \{ Deg((\mathbf{x}_a)_k) \mid \{(\mathbf{x}_s)_i, (\mathbf{x}_a)_k\} \in E \}$ because all stars observed simultaneously in an observation frame become connected.
- $Deg((\mathbf{x}_s)'_i)$ is usually smaller than $Deg((\mathbf{x}_s)_i)$ (*e.g.* $Deg((\mathbf{x}_s)'_i) \simeq 24$)

The degree of the star parameters, after elimination of the attitude parameters, is equal to the number of other stars to which they are connected. This is usually smaller than the degree in the original graph, *i.e.* the number of frames in which the star observed. The degree of the star parameters is on the average $Deg(\mathbf{x}'_s) \simeq 24$.

If we would have eliminated the star parameters instead of the attitude parameters then we would have:

$$Deg((\mathbf{x}_a)'_k) \geq \max_i \{ Deg((\mathbf{x}_s)_i) \mid \{(\mathbf{x}_s)_i, (\mathbf{x}_a)_k\} \in E \}$$

because all attitude parameters for the observation frames in which the same star is observed become connected. The degree of the attitude parameters -after elimination of the stars- is therefore almost certainly larger than that of the star parameters would we have eliminated the attitude first. Also considering that the system of attitude parameters is much larger, 17,000 compared to 2,000, then it will be clear that the attitude parameters must be placed first. In case of attitude smoothing the degree of the attitude parameters before elimination will be already larger than the degree of the star parameters. Then, the above reasoning does not hold.

So far we did not discuss the instrumental parameters. It will be sufficiently clear from their high degree, almost the highest attainable degree, that they must be put last. Furthermore, for the same reason, it makes no sense to order the instrumental parameters. Since the attitude unknowns may be eliminated in any order, only the star unknowns have to be ordered further. The appropriate ordering is now:

1. attitude unknowns in arbitrary order,
2. star unknowns, have to be ordered further,
3. instrumental unknowns in arbitrary order.

The ordering of the star parameters is discussed in the next section(s).

8.4 Ordering of the Star Unknowns

8.4.1 Introduction

Consider the graph $G'=(X',E')$ after elimination of the attitude parameters; two stars i and j are *connected* if and only if $\{i,j\} \in E'$. Because of the design of the Hipparcos instrument stars can only be connected to their direct neighbours and to stars at a basic angle distance. More precisely: two stars can be connected only if the difference of their (scan circle) abscissae, modulo 360°, is in the range

$$[-C-f,-C+f] \quad , \quad [-f,+f] \quad \text{or} \quad [+C-f,+C+f]$$

with C the basic angle and f the size of the field of view, although f is slightly increased in order to account for the projection on the RGC. The reverse is not true, but the probability is high that two stars are connected if the difference, modulo 360°, of their abscissae is in one of the above mentioned ranges. The actual connections between stars depend also on their ordinates and the scan pattern on the RGC. For example, pairs of stars for which the difference in abscissae is in the range $[-f,+f]$, will only be connected if the difference of the ordinates is in the same range. In general pairs of stars near the node of the scan circles on the RGC are always connected, but 90° away from the nodes those pairs will only be connected in about 50% of the cases (assuming 5 scan circles).

The normal matrix of the stars after elimination of the attitude parameters corresponds to the graph G'. When the stars are numbered in order of ascending abscissae (the natural ordering) the non-zeroes of the normal matrix are located in three small circular bands at "basic angle distance" (figure 7.3). The small bands at the upper right and lower left corner are due to the cyclic nature of our problem; *i.e.* the abscissae are modulo 360° and star indices are modulo N. The normal matrix of the star part is very sparse: the rate of fill is not more than 1.5%. Since the normal matrix has almost the *monotone profile* property (see appendix C), the spaces between the small bands are almost completely filled in the Choleski factor (figure 8.4), which gives the upper triangular factor the appearance of a roof with a chimney. Therefore, whenever such a structure occurs, we call it a *chimney* matrix, even when we work with the lower triangle. The rate of fill of the Choleski factor is approximately 50%.

Obviously the natural ordering of the stars by ascending abscissae is not a very good ordering. Therefore, in the next sections, three other types are considered:
- the so-called *modulo* ordering, which leaves the cyclic nature of the normal matrix and chimney structure of the factor intact,
- the *envelope* orderings, like the ones produced by the *reverse Cuthill-McKee* (RCM) algorithm and the "*banker's*" algorithm, which cluster the non-zeroes around the diagonal,
- the orderings produced by the *minimum degree* (MD) and *nested dissection* (ND) algorithms.

The modulo ordering is a synthetic ordering, which can be computed from the list of abscissae without computing the graph first. The other two groups operate on the graph, although synthetic versions are possible too. In particular, a synthetic version of the minimum degree algorithm (SBMD), which is somewhat similar to nested dissection, is considered. In table 8.1 the various ordering procedures which we considered are summarized. In addition to the nature of the routines (graph based or synthetic) also the -local- fill reducing criterion is listed: *m.f.* stands for minimum fill, *m.b.* for minimum bandwidth and *m.p.* for minimum profile. In the fourth and fifth column some of the properties of the Choleski factor are given: respectively the data structure to be used (envelope *Env*(L) or sifted format *Nz*(L)) and whether the factor has the *cyclic matrix* property (defined later).

Table 8.1: Ordering procedures for the star unknowns

	graph/ synth	min ?	data- struc.	cyclic	sect
NATURAL	synth	-	*Env*	Y	4.2
MOD 60^{0}	synth	m.p.	*Env*	Y	4.2
RCM	graph	m.b.	*Env*	N	4.3
BANKER'S	graph	m.p.	*Env*	N	4.4
MD, ND	graph	m.f.	*Nz*	N	4.5
SBMD	synth	m.f	*Nz*	N	4.5

The efficiency of the ordering procedures depends on three factors: 1) the amount of fill-in, 2) the data structure for the factor, and 3) the fact that synthetic orderings usually can be computed faster than graph based orderings. Computer storage comprises both storage for the element values and overhead for information about the row and column indices of the elements. The overhead is negligible when the factor is stored in envelope form, but when it is stored in the sifted format there will be a considerable overhead in both computation time and storage. The sifted format gives especially a non-negligible amount of overhead in the execution times due to 1) storage allocation for L (symbolic factorization) and 2) overheads in access times.

The "banker's" algorithm gave the best overall result in cpu-time, with the modulo 60^0 ordering and reverse Cuthill-McKee ex aequo on a good second place. The fill of the Choleski factors produced by the synthetic block minimum degree algorithm is already larger than the profile of the Choleski factor with the banker's ordering. The overall result of the synthetic block minimum degree and nested dissection, measured in cpu time, was not very good, also because the synthetic block minimum degree and nested dissection scatter the non-zeroes over the matrix, which results in a more complicated data structure and more overhead during the factorization.

8.4.2 Modulo ordering

The base-line of our a-priori ordering is that stars which are connected will get nearby indices. So this ordering belongs to the class of band (profile) minimizing algorithms, and has also the advantage that the overhead, during Choleski factorization is kept to a minimum. Our procedure works in two steps: First stars are ordered according to their abscissae modulo an angle C_m close to the basic angle, which reduces the overall bandwidth. Secondly the bandwidth of the chimney is further reduced by selecting a suitable starting point for the ordering.

8.4.2.1 Terminology

The normal matrix $\bar{\mathbf{N}}_{ss}$ of the star part -under the natural ordering by ascending star abscissae- is a *cyclic* variable band matrix. Let **A** be such a symmetric N by N cyclic variable band matrix, see figure 8.1. The non-zero structure of cyclic variable band matrices can be conceived as periodic in both the row and column directions, with period N. Therefore, some parts of the band are repeated in the upper right and lower left corner of the matrix. The *cyclic bandwidth* β_i^c in row i (or similar in column i) does not take these repetitions into account. It is defined as

$$\beta_i^c(\mathbf{A}) = i - \min\{\, j \mid a_{il} \neq 0,\ l = j \bmod(N)\ ,\ i-N/2 \leq j \leq i \,\}$$

The *cyclic frontwidth* $\nu_j(\mathbf{A})$ in column j (or similar in row j) is defined by

$$\nu_j^c(\mathbf{A}) = \max\{\, i \mid a_{kj} \neq 0,\ k = i \bmod(N),\quad j \leq i \leq j+N/2 \,\} - j$$

In these cyclic matrices the row and column indices i,j may be elements from $\mathbb{Z}$, the natural numbers, with $|i-j| \leq N/2$. The row and column indices in **A**, a_{kl}, are $k = i \bmod(N)$ and $l = j \bmod(N)$.

The following identities hold for the non-cyclic bandwidth and frontwidth of a cyclic matrix A, assuming the band is completely full,

$$\beta_i(\mathbf{A}) = \begin{cases} \min\{i-1,\ \beta_i^c(\mathbf{A})\} & \text{for} \quad 1 \le i < N-\beta_1^c(\mathbf{A}) \quad \text{(I \& II)} \\ i-1 & \text{for} \quad N-\beta_1^c(\mathbf{A}) \le i \le N \quad \text{(III)} \end{cases}$$

$$\omega_i(\mathbf{A}) = \min\{\ \nu_i^c(\mathbf{A}) + \beta_1^c(\mathbf{A}),\ N-i\ \}$$

For a not completely full band the previous formulae give an upper bound for $\beta_i(\mathbf{A})$ and $\omega_i(\mathbf{A})$. Assuming two bands with the bandwidth of the diagonal band b and the width of the off-diagonal bands $2b+1$, then the bandwidth and frontwidth of A are

$$\beta_i(\mathbf{A}) = \begin{cases} \min\{i-1, b\} & \text{for} \quad 1 \le i \le \nu_1^c(\mathbf{A}) \quad \text{(I)} \\ \beta_i^c(\mathbf{A}) & \text{for} \quad \nu_1^c(\mathbf{A}) < i < N-\beta_1^c(\mathbf{A}) \quad \text{(II)} \\ N-\nu_i^c(\mathbf{A})+(2b+1) & \text{for} \quad N-\beta_1^c(\mathbf{A}) \le i \le N \quad \text{(III)} \end{cases}$$

$$\omega_i(\mathbf{A}) = \min\{\ \nu_i^c(\mathbf{A}) + \beta_1^c(\mathbf{A}),\ N-i\ \} - \max\{\ \nu_i^c(\mathbf{A})-\nu_1^c(\mathbf{A})-(i+b)\ ,\ 0\ \} - \\ - \max\{\ \beta_1^c(\mathbf{A})-\beta_i^c(\mathbf{A})+(i+2b),\ 0\ \}$$

Assuming $b=\text{mean}(\beta_i(\mathbf{A}))$, the mean width of the diagonal band, and $c = \{\ \max\{\beta_i(\mathbf{A})-i+1\} \mid i=1,\nu_N(\mathbf{A})\ \}$, the width of the "chimney", then the percentage of fill in the Choleski factor is:

$$\sim \frac{2(b+c)N - (b+c)^2}{N^2} \cdot 100\ \%$$

The operation count for the factorization is of the order $O((b+c)^2N)$.

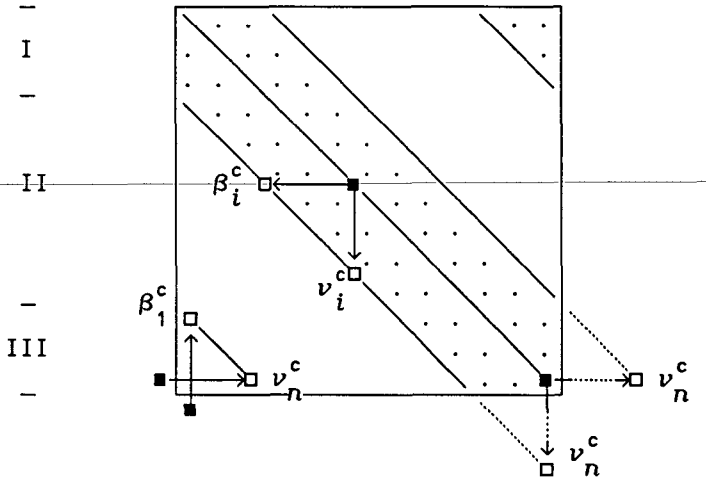


Figure 8.1: Illustration of the cyclic bandwidth

8.4.2.2 Continuous Star Graph

Consider the *continuous* Graph $G=(\psi,E)$ of the abscissae ψ $[0,2\pi)$, with edges

$$E(\psi)=\{(\psi_i,\psi_j) \mid \varepsilon C-f \le (\psi_j-\psi_i)Mrnd(2\pi) \le \varepsilon C+f,\ \varepsilon=-1,0,+1 \}\ .$$

The operator *Mrnd* is defined as $(a)Mrnd(b)=a-b\cdot Rnd(a/b)$ with range $[b/2,b/2)$ and is very similar to the *Mod* operator $(a)Mod(b)=a-b\cdot Int(a/b)$, with range $[0,b)$. The discrete version of the continuous graph, for the sequence of star abscissae ψ_i, $i=1,..N$, has more edges than the actual graph, which corresponds to the normal matrix. Especially for stars not at one of the scan circle nodes the continuous graph may give too many edges. But at least all edges present the actual graph are also present in the continuous graph. Assign a continuous numbering $Num(\psi)$ to the abscissae, $Num(\psi)$ is in the range $[0,1)$. The *cyclic* bandwidth and frontwidth are

$$\beta_i^c(G)=Max\{\ (Num(\psi_i)-Num(\psi_j))Mrnd(1)\ \mid \psi_j \in Adj(\psi_i)\ \}$$

$$\nu_i^c(G)=Max\{\ (Num(\psi_j)-Num(\psi_i))Mrnd(1)\ \mid \psi_j \in Adj(\psi_i)\ \}$$

with $Adj(\psi_i)=\{\ \psi_j \mid \{\psi_j,\psi_i\} \in E(\psi)\ \}$. The ordinary bandwidth is defined by the same formulae but without the *Mrnd* operator.

Now consider the class of orderings $Num(\psi)=(\psi\ Mod\ C_m)/C_m$, *i.e.* the abscissae are ordered modulo an angle C_m. The circular bandwidths for these orderings are:

$$\beta_i^c(G)\ =\ Max\{\ (\varepsilon C+\eta 2\pi+\delta f)Mrnd(C_m)/C_m\ \mid\ \varepsilon,\eta=-1,0,+1,\ -1\le\delta\le+1\ \}$$

and the same equation for $\nu_i^c(G)$.

proof: The formula for the cyclic bandwidth

$$\beta_i^c(G)\ =Max\{\ (Num(\psi_i)-Num(\psi_j))Mrnd(1)\ \mid \psi_j \in Adj(\psi_i)\ \}$$

can be rewritten as

$$\beta_i^c(G)=Max\{(\psi_i-\psi_j)Mrnd(C_m)/C_m \mid \varepsilon C-f \le (\psi_j-\psi_i)Mrnd(2\pi) \le \varepsilon C+f, \varepsilon=-1,0,+1\}$$

The inequality $\varepsilon C-f \le (\psi_j-\psi_i)Mrnd(2\pi) \le \varepsilon C+f$, with $\varepsilon=-1,0,+1$, is written as $\varepsilon C-f+2\pi\cdot Rnd((\psi_j-\psi_i)/2\pi) \le \psi_j-\psi_i \le \varepsilon C+f+2\pi\cdot Rnd((\psi_j-\psi_i)/2\pi)$. The difference $\psi_j-\psi_i$ can take the values $\varepsilon C+\delta f+\eta 2\pi$, $\varepsilon,\eta=-1,0,+1$ and $-1\le\delta\le+1$. Then the bandwidth $\beta^c=Max\{\ (\varepsilon C+\eta 2\pi+\delta f)Mrnd(C_m)/C_m \mid -1\le\delta\le+1,\ \varepsilon,\eta=-1,0,+1\ \}$ □

8.4.2.3 The Optimal Angle for the Modulus

The natural ordering by ascending star abscissae is defined by $Num(\psi)=\psi/2\pi$, *i.e.* $C_m=2\pi$. The circular bandwidth for the natural ordering is then

$$\beta^c(G^n)\ =\ (C+f)/2\pi\ \simeq\ .18$$

with $C=58^\circ$ and $f=.9^\circ$. Of course $C_m=2\pi$ is not a very good choice. The value

for C_m should be chosen in such a way that the bandwidth β^c is minimized. The formula for the bandwidth suggests that:

- C_m should be as large as possible, because C_m is the denominator in the formula for the bandwidth,
- C_m should be close to a divisor of 2π, $2\pi-C$ and $2\pi+C$ since $f \ll C$, *i.e.* C_m should be a divisor of 2π and C.

From these considerations it follows that $C_m=60^\circ$ is a good choice, the bandwidth is then

$$\beta^c(G^{60}) = (|C-C_m|+f)/C_m \simeq .048$$

i.e. a reduction by almost a factor of 4 compared to the natural ordering. For *e.g.* $C_m=58^\circ$ the maximum cyclic bandwidth

$$\beta^c(G^{58}) = (360+f)Mrnd(C_m)/C_m \simeq .22$$

is even larger than in case of the natural ordering.

The ordering is illustrated by a small example with 360 stars, regularly distributed over the RGC, numbered in ascending order from 1 to 360. The width of the field is in this example 1°. The modulo 60° ordering for the example is

1,61,121,181,241,301,2,62,122,182,242,302,...
.....,59,119,179,239,299,359,60,120,180,240,300,360

Star i is connected to star i-1 and i+1, observed in the same field of view, to star i+57, i+58, i+59, observed in the following field of view, and to star i-57, i-58, i-59, observed in the preceding field of view. It is easy to see - by counting - that star i is connected to stars not further than 17 positions, or unknowns, away. The bandwidth as computed by the earlier derived formula is 18. This ordering is illustrated in figure 8.2.

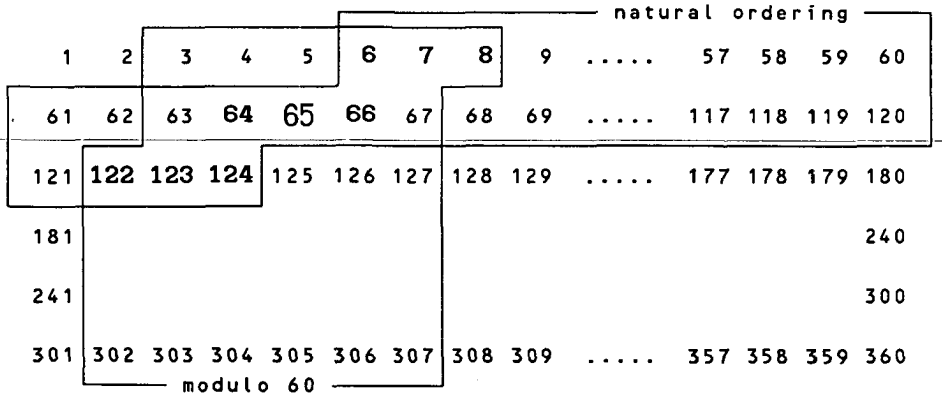


Figure 8.2: Illustration of the modulo 60° ordering; stars connected to star 65 are in bold, stars within the band are outlined for the modulo 60° and the natural ordering.

In figure 8.3 the effect of the ordering is shown for an angle C_m=58°, which is not a divisor of 360°. Now the term $|C_m - C|$ is zero, and for most unknowns the bandwidth is only $N \cdot f/C_m \simeq 0.017 \cdot N \simeq 6$. In the figure this is demonstrated for star 65, for which the bandwidth is 7. However, at least one third of the unknowns has a larger bandwidth. This is demonstrated for star 3 which has a bandwidth of 66. Here part of the connected unknowns have abscissae > 360^0 or < 0°, and so we have to take in the formula for the bandwidth the modulus. The bandwidth is then $N \cdot (f+2\pi) \bmod (C_m)/C_m \simeq .22 \cdot N \simeq 80$. Therefore, the overall profile of the modulo 58^0 ordering is larger than the modulo 60^0 ordering. Another problem is of course the small group of unknowns left: they should be placed somewhere, and will increase the bandwidth further.

star 3 — star 65

1	2	3	4	5	6	7	8	9	10	11	12	.....	57	58
59	60	61	62	63	64	65	66	67	68	69	70	.....	115	116
117	118	119	120	121	122	123	124	125	126	127	128	.....	173	174
175														232
233														290
291	292	293	294	295	296	297	298	299	300	301	302	.....	347	348
												349	....	360

Figure 8.3: Illustration of the modulo 58^0 ordering; stars connected to star 65 and 3 are in bold, stars within the band for for star 65 and 3 are outlined.

The ordering of the star unknowns according to their abscissae modulo 60^0 leads to a reduction of the bandwidth of the normal matrix from roughly (the equivalent of) $N(C+f)/360$ to $N(2+f)/60$, where C is the basic angle, f the along-scan width of the field of view, and N the number of stars on the RGC. With this ordering the Cholesky factor preserves its characteristic "roof with chimney" form, but a 70 % reduction in memory (assuming f~0.9° and C~58°) and a 90 % reduction in computing time - for the sparse inverse as well - is attained when compared with the natural ordering of the stars on the circle.

8.4.2.4 Bandwidth Optimization of the Chimney

A considerable amount of fill-in is still created in the chimney of the matrix. First note that the density of stars on an RGC is not constant, but varies from approximately 0.5 to 1.5 times the average density. After "collapsing" of the circle, i.e. superposing the separate 60^0 sectors, the density variations become smaller but remain present. By choosing a suitable starting point of the ordering (see figure 8.4), the width of the chimney can further be reduced. Since in realistic cases the density variations are larger than nominal, this small refinement can lead to an additional improvement in computing requirements from a few percent up to some 30 %, for a small additional reordering expense.

The bandwidth of the normal matrix - when the stars are ordered modulo 60^0- is a function of the cumulated star density on the RGC, *i.e.* the bandwidth $\beta_i^c(\mathbf{A})$ is proportional to the sum of the densities at $\psi_i+(k-1)*60^\circ$, $k=1,..,6$, where ψ_i is the abscissae of the i'th unknown. The density on the RGC depends on several factors: first of all near the node of the scanning circles the density is a factor 3 lower than 90^0 further along the circle, but also occultations and the variable star density of the star field itself are important factors. Due to these three factors and the ordering procedure itself it is difficult to choose the most favourable first unknown from the density function only. Therefore in the ordering routine the bandwidth itself is estimated.

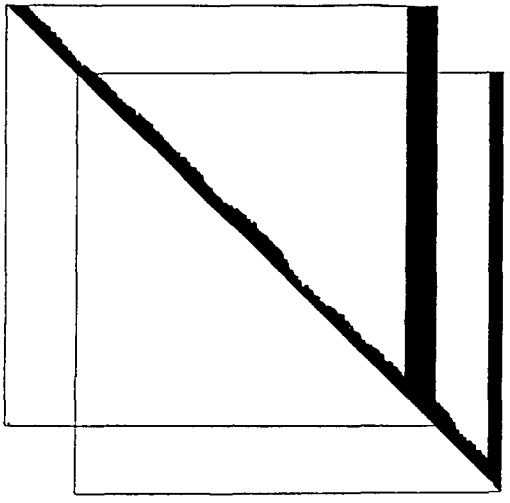
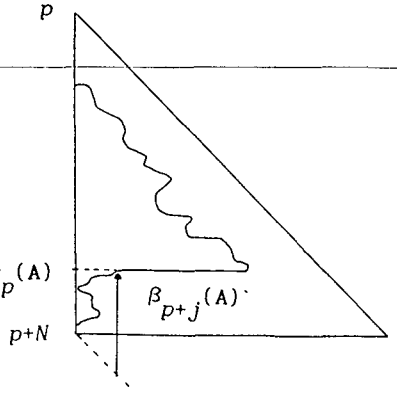


Figure 8.4: Non-zero elements of the Choleski factor of the star part after reordering (CERGA dataset II). The expected fill-in is given for two different starting points of the ordering.

The profile $|Env(\mathbf{L})|_p$ of the Choleski factor with node p as starting node is

$$|Env(\mathbf{L})|_p = \sum_{i=p}^{p+N-\chi_p(\mathbf{A})} \beta_i^c(\mathbf{A}) + \sum_{i=p+N-\chi_p(\mathbf{A})+1}^{p+N} (i-p) =$$

$$= \sum_{i=1}^{N} \beta_i^c(\mathbf{A}) + \sum_{i=N-\chi_p(\mathbf{A})+1}^{N} (i-\beta_{p+i}^c(\mathbf{A}))$$



where $\chi_p(\mathbf{A})$ is defined as

$$\chi_p(\mathbf{A}) = \max\{\ \beta_{p+j}^c(\mathbf{A})-j\ \} \quad \text{for } j=0,..,\nu_p(\mathbf{A})$$

The objective of the bandwidth optimization of the chimney is to find a starting node p such that the second part of the formula is minimized, *i.e.*

$$\min_{p} \{ \sum_{i=N-\chi_p(\mathbf{A})+1}^{N} (i-\beta^{c}_{p+i}(\mathbf{A})) \}$$

The bandwidth $\beta^{c}_{i}(\mathbf{A})$ is computed from the "continuous" synthetic graph and not from the actual graph.

8.4.3 Reverse Cuthill-McKee algorithm

The *rooted level structure* rooted *at* x is the partitioning $\mathcal{L}(x) = \{L_0(x), L_1(x), \ldots, L_{l(x)}(x)\}$ where $L_0(x)=\{x\}$ and $L_i(x)$, $i=1,\ldots,l(x)$ consists of the nodes adjacent to $L_{i-1}(x)$ and not already present in one of the previous partitions, *i.e.* $L_0(x)=\{x\}$, $L_1(x)=Adj(x)$ and $L_i(x)=Adj(L_{i-1}(x))-L_{i-2}(x)$, $i=2,3,\ldots,l(x)$.

The *Cuthill-McKee* ordering is based on this partitioning of the rooted level structure of the graph of the normal matrix. The nodes are numbered level by level. In each level internally the nodes adjacent to the lowest numbered node of the preceding level are numbered first in order of increasing degree and so on until the highest numbered node in the preceding level. The *Reverse* Cuthill-McKee ordering (RCM), the ordering obtained by reversing the Cuthill-McKee ordering, often turns out to be superior. It has been proved by Liu that the reverse ordering is never inferior [*George & Liu, 1981*]. This recipe has been implemented in the following algorithm:

Algorithm (*Reverse Cuthill-McKee*)

1. Determine a starting node (*e.g.* a pseudo peripheral node), and number it x_1.
2. For $i=1,\ldots,N$ find all unnumbered nodes in $Adj(x_i)$ and number them in increasing order of degree.
3. Reverse the ordering

The general idea behind the Cuthill-McKee ordering is to reduce the bandwidth by local minimization of the $\beta_i(\mathbf{A})$'s, the bandwidth. One point deserves further attention, *viz.* which node to select as a starting node.

Let us define a path from a node x to a node y of length $l \geq 1$ as an ordered set of distinct nodes $(v_1, v_2, \ldots, v_{l+1})$ such that $v_{i+1} \in Adj(v_i)$, $i=1,2,\ldots,l$ with $v_1=x$ and $v_{l+1}=y$. In other words y is reachable from x through the set of nodes $\{v_2,\ldots,v_l\}$. The reach of a node y through S is defined by

$$Reach(y,S) = \{ x \notin S \mid x \text{ is reachable from y through } S \}$$

The *distance* $d(x,y)$ between two nodes x and y in the connected graph is simply the shortest path joining the two nodes. The *eccentricity* $l(x)$ of a node x and the *diameter* $\delta(G)$ of the graph G are defined as

$$l(x) = \max\{ d(x,y) \mid y \in X \}$$

$$\delta(G) = \max\{ l(x) \mid x \in X \}$$

and x is called a *peripheral* node if $l(x)=\delta(G)$.

In the literature [*George & Liu, 1981*] it is suggested to choose a near peripheral node as starting node. The peripheral nodes themselves are too expensive to compute, therefore it is suggested to use an approximate peripheral node. However, for the great circle reduction problem there is no clear peripheral node. It turns out that, due to the specific structure of the measuring device, the eccentricity of the nodes is between 14 and 16, and the optimal starting node does not necessarily have the highest eccentricity. Experiments showed that in the case of Hipparcos the influence of the starting nodes on the fill-in is small; the difference in fill-in between the worst and best starting nodes is not more than 10%.

The Reverse Cuthill-McKee ordering is illustrated in figure 8.5, where we have plotted the rooted level structure for CERGA dataset II. The dots in figure 8.5 indicate stars, which are plotted on imaginary concentric circles which represent the level the star is in. Level 0, the starting node, is the dot in the middle. The results of the ordering on CERGA dataset II are given in figure 8.6 and table 8.2. The performance of the Reverse Cuthill-McKee ordering is comparable to the modulo 60^0 ordering. Also the time needed to compute the ordering is of the same order of magnitude.

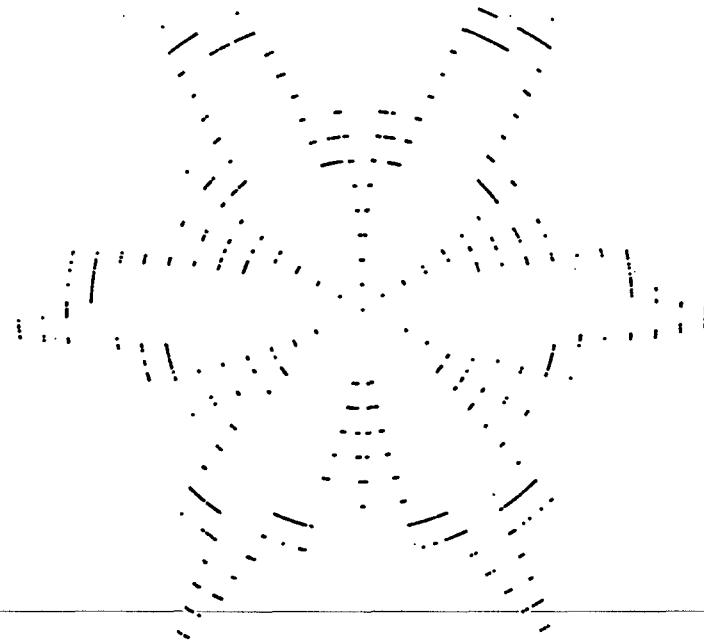


Figure 8.5: The rooted level structure: the imaginary concentric circles depict the level, the dots indicate stars; plotted as function of the level (radius) and abscissa (argument). (from [*De Jonge, 1987*])

8.4.4 Banker's algorithm

The Cuthill-McKee ordering is based on local minimization of the bandwidth. The *banker's* algorithm, suggested by [*King, 1971*] and [*Snay, 1976*], is based on local minimization of the profile or frontwidth.

The name for the banker's algorithm comes from the following analogy [*Snay, 1976*]. A banker is asked to divide the estate of a rich landlord among his tenants, but before any tenant can get a piece of land he must first

settle his affairs with other tenants. From the first time a tenant has to settle one of his affairs until he gets a plot of land he must stay in the village hotel, during this time the tenant is hopeful. The banker aims, in order to save the good name of his bank and in order to save hotel costs, at minimizing the total waiting time of all the tenants. The tenants can be associated with nodes, the total waiting time of the tenants corresponds with the profile of the matrix and giving a tenant a plot of land corresponds to numbering a node. The analogy goes a little further: if the banker would have preferred to minimize the maximum waiting time of the tenants, which corresponds to the bandwidth in the matrix, he could have used the Cuthill-McKee algorithm.

algorithm (*banker's*)

1. Pick a starting node and number it x_1.
2. For $i=1,\ldots,N-1$ first determine the set of *hopeful* nodes H and *candidates* C

$$H = Adj(\{x_1,\ldots,x_i\})$$
$$C = H \cup \{\ c \in Adj(H) \mid c \notin \{x_1,\ldots,x_i\}\ \}$$

if $C=\emptyset$ (and $H=\emptyset$) then stop; all the nodes in this connected component are numbered.

Otherwise find a node $y \in C$ with minimum

$$m_{i+1} - n_{i+1}$$

where $m_{i+1}=|M_{i+1}|$, $M_{i+1}=\{\ m \in Adj(y) \mid m \notin \{x_1,..,x_i\},\ m \notin H\ \}$ and $n_i=1$ if $y \in H$, else $n_i=0$, and number y as x_{i+1}.

$|H|+m_{i+1}-n_{i+1}$ represents the change in profile: the number of off-diagonal elements (the number of active rows) in the $(i+1)$'th column of the lower triangular factor. Note that, for efficiency reasons, not every unnumbered node is a candidate. The set of hopeful and candidate nodes are in practice computed from the following updating formulae

$$H \leftarrow H \cup \{\ y \in Adj(x_i) \mid y \notin \{x_1,\ldots,x_i\}\ \} \setminus \{x_i\}$$
$$C \leftarrow C \cup \{\ y \in \{Adj(x_i), Adj\{Adj(x_i)\}\} \mid y \notin \{x_1,..,x_i\}\ \} \setminus \{x_i\}$$

where H and C are initialized by $H=\emptyset$ and $C=\emptyset$.

A starting node for the banker's algorithm can be computed in the same way as for the Cuthill-McKee ordering. However, in the algorithm proposed by Snay an ordering is computed for ten different pseudo peripheral nodes, and then the pseudo peripheral node which gives the lowest profile is selected as starting node. This approach is not followed for Hipparcos, because almost all nodes are near peripheral and because we believe that it is not necessary to compute ten different orderings. A starting nodes is good when it gives in the first steps of the banker's algorithm only a few hopeful stars, *i.e.* the degree $|Adj(x_0)|$ of the starting node x_0, and the degree of the hopeful nodes in the first step of banker's $|Adj(x_i)\backslash Adj(x_0)|$, with $x_i \in Adj(x_0)$ hopeful nodes, must be small. These nodes can be found 1) for stars at 90^0 distance from the intersections of the scanning circles with the RGC, and 2) for stars which are observed only on the outermost scanning circle [*De Jonge, 1987*].

Results of the banker's algorithm on CERGA dataset II are given in table 8.2 and figure 8.6. The ordering produced by the banker's algorithm gives very good results, but it is rather expensive to compute. Still, the cpu times in table 8.2 for the banker's ordering are a little too pessimistic because ten orderings with different starting nodes were actually tried, as suggested by Snay. However, the best results were obtained with a starting node which was not selected by the algorithm. Once we have a good starting node the banker's ordering is not very difficult or expensive to compute. In [*De Jonge, 1987*] also an iterative scheme for improving the banker's ordering is suggested. At the cost of a few extra seconds ordering time 20-30 seconds of factorization time can be saved on CERGA dataset II.

The banker's algorithm and the ordering schemes suggested by King [*King, 1970*] and Levy [*Levy, 1971*] are very much alike. Actually, King proposed two ordering schemes: In his first scheme only the hopeful nodes are considered as candidates. His second scheme is identical to the banker's ordering. In the Levy ordering all unplaced nodes are candidates. Experiments with King's first scheme on CERGA dataset II were not very successful. But, to our surprise, the Levy ordering, which considers all nodes as candidates, did not give better results than the banker's algorithm.

8.4.5 Minimum Degree, Nested Dissection and Synthetic Block Minimum Degree

So far we discussed only ordering schemes which aim at reducing the bandwidth or profile of a matrix. We will now briefly discuss three different schemes which directly aim at (local) minimization of the fill-in, these are: the minimum degree algorithm, nested dissection and synthetic block minimum degree.

In section 8.2 we have shown that the factorization process can be modelled by a series of elimination graphs. The *minimum degree* reordering algorithm specifies that the pivot to be eliminated next in an elimination graph G_i must be of minimum degree in G_i. It will be intuitively clear that the number of edges to be added to G_{i+1} then generally tends to be small. In this sense minimum degree is a heuristic algorithm; in general it does not result in minimum fill-in globally, although there is sufficient empirical evidence that it produces an ordering which gives a low fill-in. Despite its simplicity an efficient implementation for the minimum degree algorithm turns out to be not straightforward. An efficient implementation based on so-called quotient graphs is given in [*George & Liu, 1981*]. This method has been tried on small simulations only; results for a more realistic great circle are not yet available. A synthetic version of the minimum degree algorithm, and nested dissection, have been tried instead.

In dissection schemes the graph is subdivided into smaller graphs. The idea behind one-way dissection is to find n small sets of separators S_i, which divide the original graph into $n+1$ unconnected components G_j. If the unconnected components are numbered first and the separator sets last, fill-in will only occur *in* G_j and S_i, and *between* the separator sets S_i and S_{i+1}. The idea behind nested dissection is to partition the graph into two subgraphs, using a single separator ($n=1$), which is then repeated for each of the subgraphs until no more separators can be found. Each time a separator is found their nodes are numbered as the last in the subgraph. The nested dissection ordering for the great circle reduction has been investigated in [*De Jonge, 1987*].

A similar idea underlies the synthetic block minimum degree reordering. This ordering scheme is based on the approximate star graph which we already used for the modulo ordering. The idea behind this ordering is simple: First partition the star nodes in 7 sectors. The first 6 sectors are, in terms of the star abscissae, 58^0 long, the seventh sector contains stars from the remaining 12^0. Each of the sectors is divided in small segments, which consist of a long and short part. The short parts of the segments, and 3 of the sectors, act as separators. The large parts of the segments are numbered first, starting with sector 2 and followed by sector 4, 6, 1, 3, 5 and 7. Then the small parts are numbered, in the same order of sectors. The order of the sectors, and the sizes of the large and small parts have been optimized. The small parts, which act as separator, are usually 1^0 long, a little more than the length of the field of view. The optimum length for the large parts turns out to be 4.8^0; so there go 10 segments in a sector of 58^0 long.

8.4.6 Results for CERGA dataset II

In figure 8.6 and table 8.2 the results of various orderings for CERGA dataset II have been summarized. Results for the natural, modulo 60^0, reverse Cuthill-McKee, banker's and Synthetic Block Minimum Degree ordering are given. Figure 8.6 contains the lower triangle of $Nz(\bar{\mathbf{N}}_{ss})$ and in the upper triangle $Env(\bar{\mathbf{N}}_{ss})$ is given. In table 8.2 the fill $|Nz(\mathbf{L}_{ss})|$ and profile $|Env(\bar{\mathbf{N}}_{ss})|$ of the Choleski factor, plus some computing times for a VAX 750, are given. In table 8.2 also the number of page faults is given. The VAX 750 is a virtual memory machine: Only a small part of the data can be stored in the fast internal computer memory. The remainder of the data is stored on disk. When the required data is not found in the fast memory one "page" of data (512 Kb) has to be swapped between disk and internal memory. This is a so-called page fault. Page faults cost cpu-time, therefore they are also given in the table.

The banker's ordering gave the best overall results in cpu-time, with the modulo 60^0 and reverse Cuthill-McKee ordering ex aequo on a good second place. The fill-in produced by the synthetic block minimum degree ordering is already larger than the profile from the banker's ordering. The fill-in produced by the nested dissection algorithm is smaller, but the overall results of nested dissection (and synthetic block minimum degree) are not very good because the non-zeroes will be scattered over the matrix, which results in a more complicated data structure during the factorization. Surprisingly the sifted format factorization of the normal matrix after reordering with the banker's algorithm is faster than that of the normal matrix after nested dissection, although there will be less fill-in with nested dissection. This effect is due to the sub-optimal Choleski factorization algorithm (the so-called bordering algorithm of appendix C) which was used. Another surprise is that, although the modulo 60^0 and RCM ordering produce more or less the same envelope, the results measured in factorization time are better for the RCM ordering. This effect is caused by the small number of page faults for the RCM factorization. Here it is an advantage that the RCM ordering does not generate an chimney matrix, as the modulo 60^0 ordering does. However, a different implementation of the Choleski factorization could reduce the number of page faults for the modulo 60^0 ordering (see appendix C).

Table 8.2: Results of ordering strategies for the star part. Data: CERGA dataset II. n_s=1843, $|Nz(A)|$=17949. NAT=natural ordering, M60=Modulo 60, RCM=Reversed Cuthill-Mckee, BN=bankers, ND=nested dissection and SBMD=synthetic block minimum degree.

		NAT	M60	RCM [2]	BN [1]	ND [1]	SBMD
$\|Env(L)\|$	#	819934	242178	242336	165363	-	-
	%	48.2^{8}%	14.2^{5}%	14.2^{6}%	9.7^{3}%	-	-
$\|Nz(L)\|$	#	3)	235177	226779	161066	138406	180464
	%		13.8^{4}%	13.3^{4}%	9.4^{8}%	8.1^{5}%	10.6^{2}%
cpu ordering		-	1.7	2.9	70^{4}	3.5	0.1
Envelope fact.		-					
cpu (s.)		3)	372	286	180	-	-
10^3 page faults			185	4.0	63.8	-	-
Sifted fm. fact.							
cpu - symbolic		3)	5.3	8.2	4.0	4.5	4.5
- actual [5]			556	523	338	404	763
10^3 page faults			270	6.0	139	3)	416

Notes:

1) the reordering started with the natural ordering
2) the reordering started with the modulo 60 ordering
3) data not available
4) computing time for 10 orderings (with a special starting node for the GCR the ordering time becomes 13 s.)
5) the factorization routine for sifted format matrices can still be improved

Figure 8.6 (on the next three pages) - The non-zero structure (in the lower triangle) and envelope (in the upper triangle) of the normal matrix for the star part of CERGA dataset II, as results from six different ordering schemes (the envelope is not given in figure 8.6.e and 8.6.f).

Figure 8.6.a

Ascending order

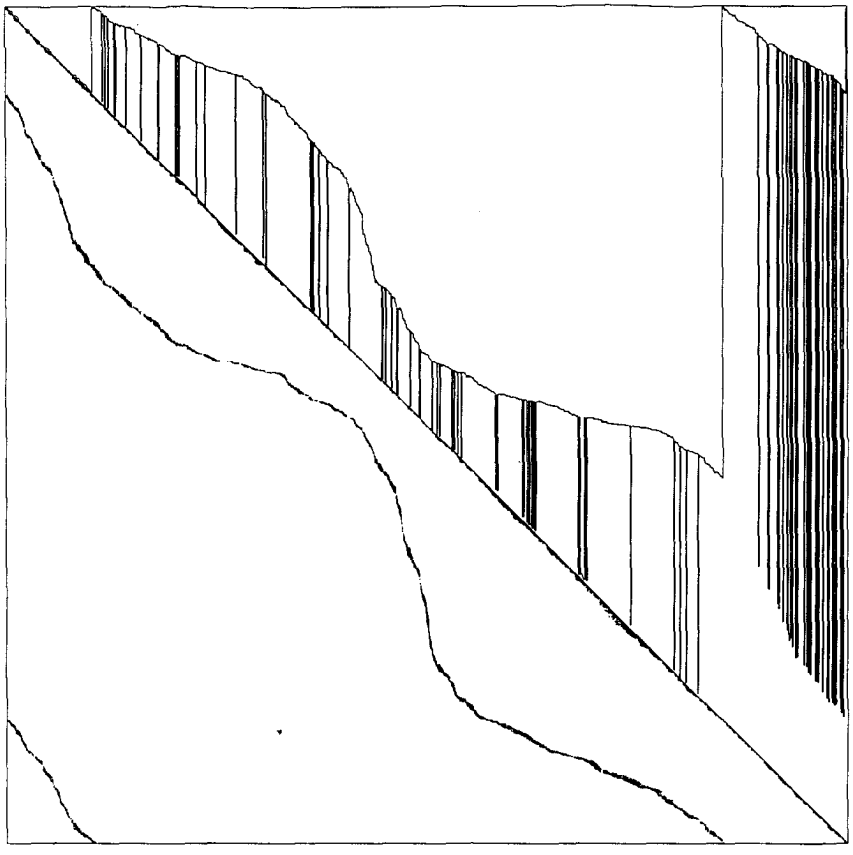
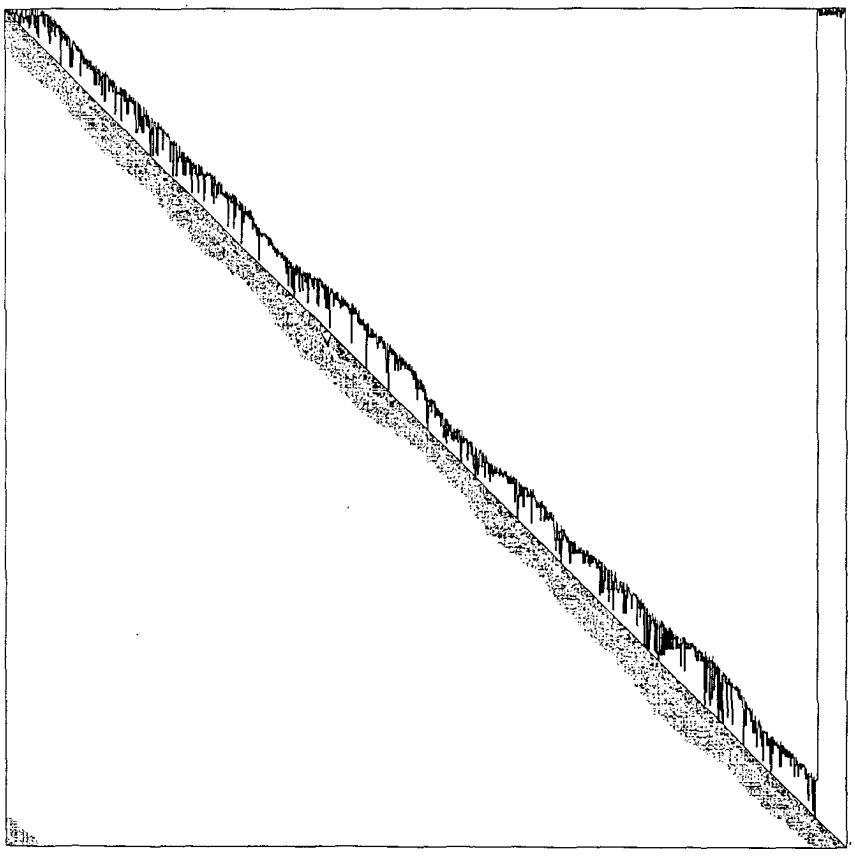


Figure 8.6.b

Modulo 60^0



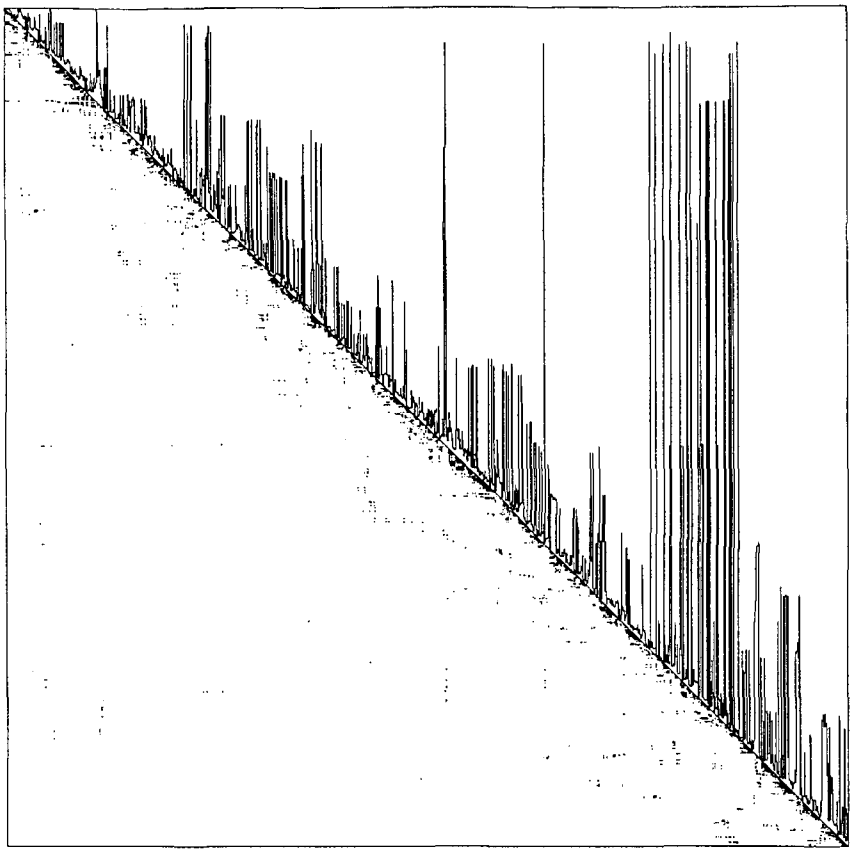


Figure 8.6.c

Banker's

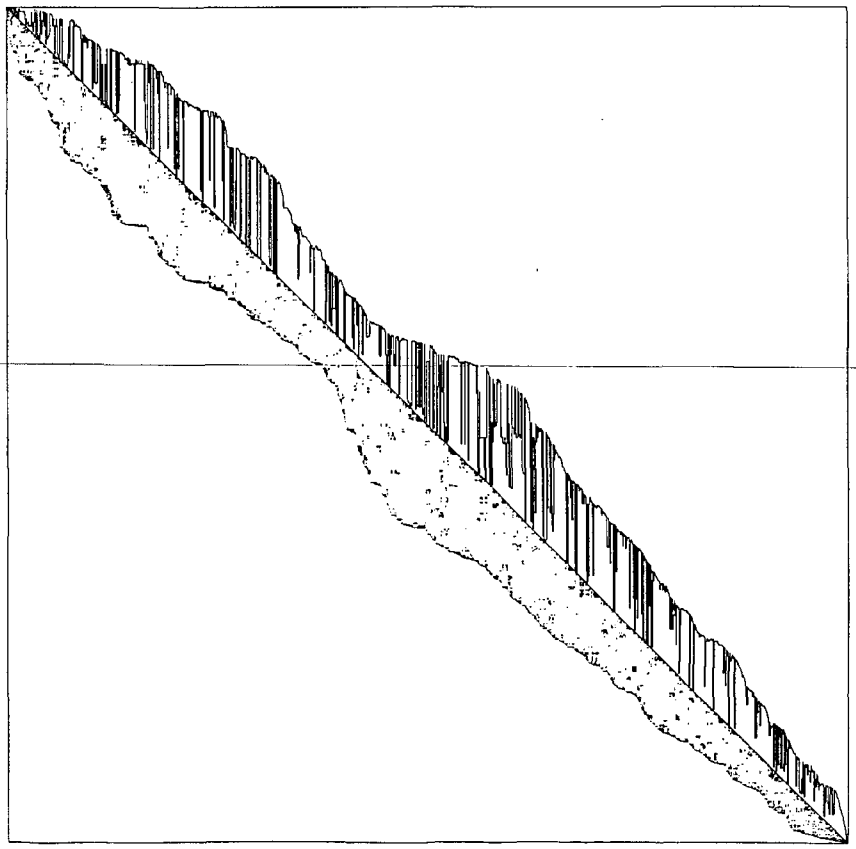


Figure 8.6.d

Reverse Cuthill
-McKee

Figure 8.6.e

Nested
Dissection

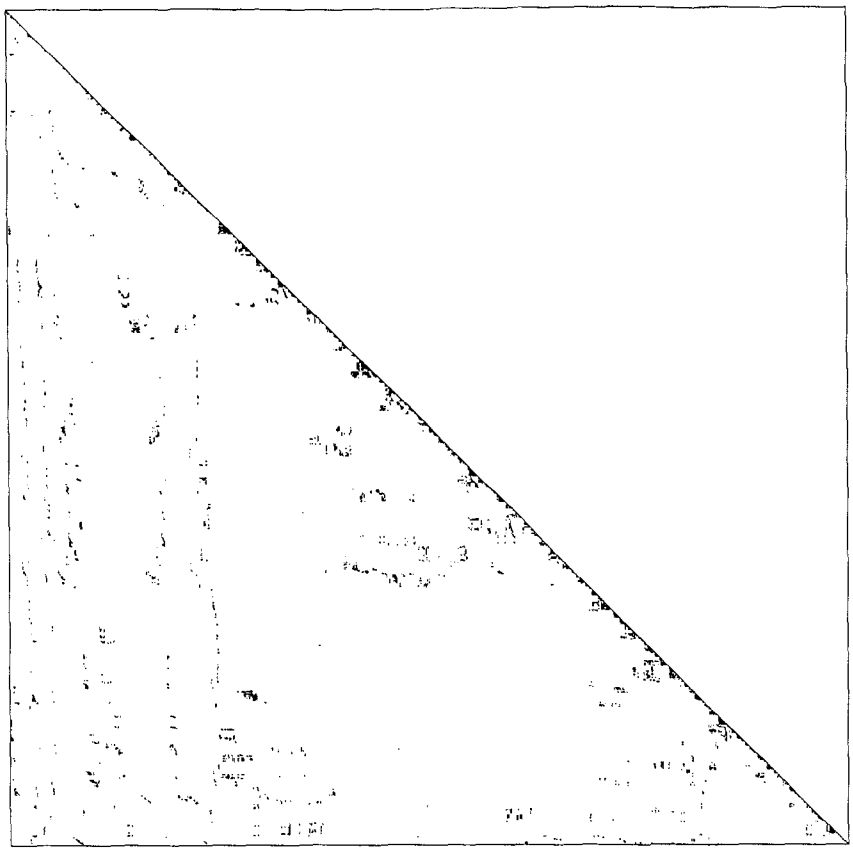
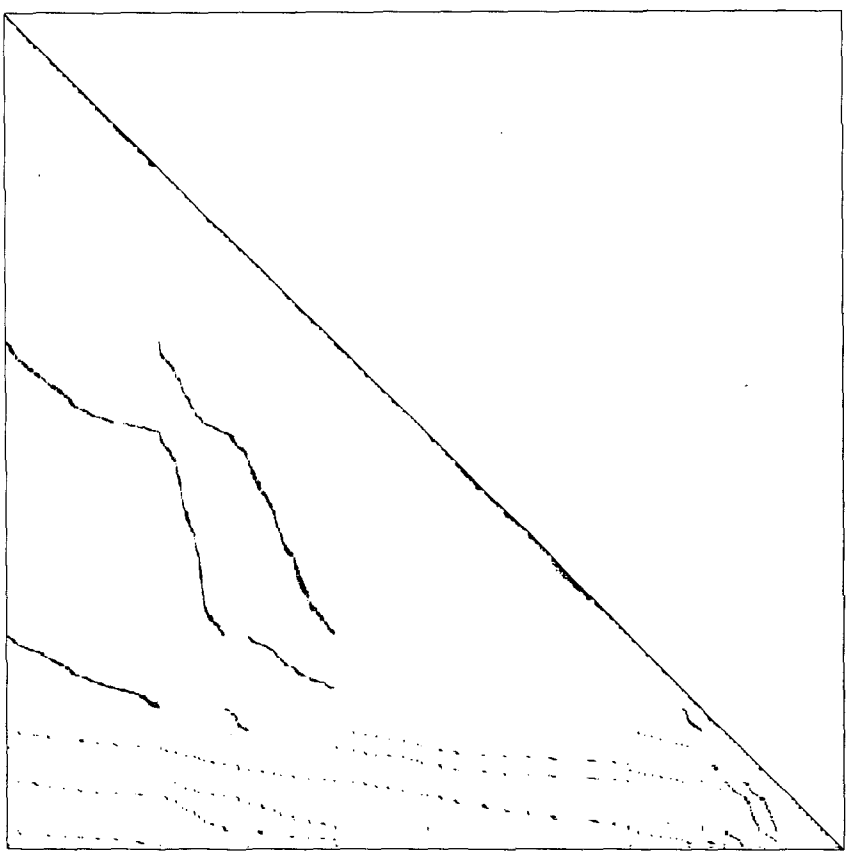


Figure 8.6.f

Synthetic Block
Minimum Degree



8.5 Optimum Block Ordering in Smoothing Mode

The optimal ordering of the unknowns in the geometric solution mode was: attitude, stars, instrument. In the smoothed solution mode the optimal ordering is: stars, attitude (B-splines), instrument.

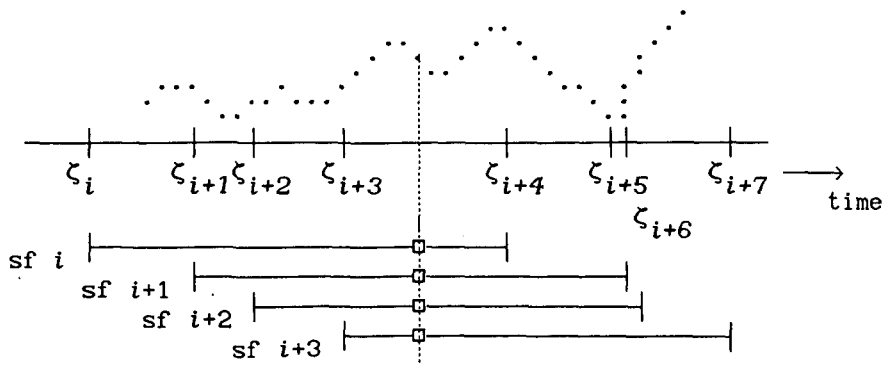


Figure 8.7 - Support of the B-splines (superframes)

In this chapter only the B-spline model is considered. Let us define the domain over which a B-spline is non-zero a *superframe*. Superframes overlap like ordinary observation frames. Generally, a superframe overlaps with the next k consecutive superframes, where k is the *order* of the spline (figure 8.7). Superframes are longer than ordinary frames and have a variable length. Contrary to ordinary frames the attitude within a superframe cannot be described by a single attitude parameter, but $2k$ parameters are needed. Two consecutive superframes have $2k$-1 parameters in common. Therefore the total amount of attitude parameters does not exceed the number of B-splines. Simulation experiments indicate that about 600 B-spline parameters are needed to model the attitude with sufficient precision. This forms a considerable reduction compared to the 17,000 geometric attitude parameters.

The non-zero structure of the normal matrix for smoothing (eq. 7.33) has already been given in figure 7.4. The non-zero structure of the star and instrument part is identical to the non-zero structure for the geometric solution; $\mathbf{A}_s$ and $\mathbf{N}_{ss}$ contain only one non-zero element per row, $\mathbf{A}_i$, $\mathbf{N}_{ii}$, $\mathbf{N}_{is}$ and $\mathbf{N}_{bi}$, are almost completely full. Each spline is non-zero over a small domain, and therefore the matrix $\mathbf{B}$ has only a few non-zero elements per row. Let k be the order of the B-splines, then $\mathbf{B}$ and $\mathbf{A}_a\mathbf{B}$ contain precisely k non-zeroes per row. $\mathbf{N}_{bb}=\mathbf{B}^T\mathbf{N}_{aa}\mathbf{B}$ is banded, with bandwidth k (*i.e.* $2k$+1 non-zeroes per row). The non-zero structure of $\mathbf{N}_{bs}=\mathbf{N}_{sb}^T$ is identical to that of $\mathbf{N}_{as}$, except for the compressed row dimension and thicker bands.

The normal matrix blocks pertaining to the smoothed attitude parameters are smaller, but denser, than those for the geometric parameters. $(\mathbf{N}_{bs})_{li}\neq 0$ if and only if star i is observed in the l'th superframe. Per scan a star is observed once in the preceding and once in the following field; during each passage it is seen in at most $\underline{k}$ superframes. Now let c be the number of scan circles for this RGC, and let $\bar{c}$ be the average number of scan circles a star is observed on. So, there are approximately $2k\bar{c}$ non-zeroes in each column of $\mathbf{N}_{bs}$. Further, let n_a, n_b, n_s and n_i be respectively the number of geometric

attitude, spline, star and instrument parameters. Then the average number of non-zeroes in a row of $\mathbf{N}_{bs}$ is approximately $2k\bar{c}(n_s/n_b)$. Hence, $Deg(\mathbf{x}_s) \simeq 2k\bar{c}$ and $Deg(\mathbf{x}_b) = 2k + 2k\bar{c}(n_s/n_b)$, not counting the instrumental parameters. So, for smoothing with cubic B-splines (order $k=4$), ± 2000 stars, with each star observed on the average on 3 scan circles ($\bar{c} \simeq 3$), and 600 B-splines, we have $Deg(\mathbf{x}_s) \simeq 2k\bar{c} \simeq 25$ and $Deg(\mathbf{x}_b) \simeq 2k + 2k\bar{c}n_s/n_b \simeq 90$, not counting the instrumental parameters, and $Deg(\mathbf{x}_i) \simeq n_s+n_b+n_i-1 \simeq 2600$. Hence, $Deg(\mathbf{x}_s) < Deg(\mathbf{x}_b) << Deg(\mathbf{x}_i)$.

According to the minimum degree criterion the unknowns with a small degree must be eliminated first in order to obtain a small fill-in in the Choleski factor. Hence, it follows that the star parameters must be eliminated first. Also observe that the edge set of the section graph $G(\mathbf{x}_s)$ is empty, *viz.* $\mathbf{N}_{ss}$ is diagonal, so successive elimination of star parameters does not influence the degree (in the total graph) of the remaining nodes in the section graph. From these two observations we may conclude that to obtain a small fill-in 1) star parameters must be eliminated from the graph first, and 2) they may be eliminated in any order, without consequences for the fill-in. There will be fill in the section graph of the attitude (spline) parameters, but there is no fill in the section graph of the star parameters themselves.

The graph $G'=(X',E')$ has, after elimination of the star parameters, the following properties:

- $Deg((\mathbf{x}_b)'_k) \geq Deg((\mathbf{x}_b)_k)$
- $Deg((\mathbf{x}_b)'_k) \geq \max_i \{ Deg((\mathbf{x}_s)_i) \mid \{(\mathbf{x}_s)_i, (\mathbf{x}_b)_k\} \in E \}$

The degree of the attitude parameters is greater than that in the original graph and larger than the degree of the stars to which the attitude parameter in question is connected. The non-zero structure of the reduced normal matrix $\bar{\mathbf{N}}_{bb}$ after elimination of the star parameters is given in figure 8.8.a. $(\bar{\mathbf{N}}_{bb})_{kl} \neq 0$ if there exists a star i observed in the l'th and k'th superframe, which typically occurs for superframes separated in time by the equivalent of one basic angle and one or more revolutions of the spacecraft. This gives the off-diagonal bands in figure 8.8.a Let c be the number of scan circles and k the order of the spline, then there will be not more than $3(2c-1)$ bands. Each band will be at least $2k$ thick, except the diagonal which is slightly thicker, $2k+1$, but the width can be larger because of irregularities in the spacing of the B-splines, especially near gas jets. Therefore the degree of the attitude parameters - after elimination of the stars - is $Deg(\mathbf{x}'_b) \simeq 3(2k)(2\bar{c}-1) \simeq 145$ (the number of B-splines is typically 600), where $\bar{c}$ is the average number of scan circles a star is observed on.

Elimination of the attitude parameters first gives a larger system with only three bands. Each of these bands is $2\bar{c}kn_s/n_b$ thick, so $Deg((\mathbf{x}_s)'_i) \simeq 6\bar{c}kn_s/n_b \simeq 250$. The degree of the star parameters - after elimination of the attitude - is almost certainly larger than for the attitude parameters, would we have eliminated the stars. Also considering that the system of star parameters is larger, 2,000 compared to 600, then it will be clear that the star parameters should be placed in front.

The instrumental parameters are of course placed last and it makes no sense to order them. Since the star unknowns may be eliminated in any order, only the attitude unknowns have to be ordered further. The ordering of the attitude parameters is discussed in the next section.

8.6 Ordering of the Attitude Unknowns during Smoothing

The ordering procedures which we have been using to order the star parameters in the geometric reduction step can also be applied to the attitude parameters in the smoothing step. Even the modulo ordering with bandwidth optimization can be used, except for a -possibly- different modulus.

Consider the B-splines series of order k with knots at t_l, $l=1,2,..,n_b+k$. The domain (superframe) of the l'th B-spline is the interval $[t_l, t_{l+k})$. Two B-splines p and q can be "connected", *i.e.* $\{(\mathbf{x}_b)_p, (\mathbf{x}_b)_q\} \in E'$, if

$$-(t_{q+k}-t_q) \leq t_q-t_p+\nu T_1+\varepsilon T_2 \leq (t_{p+k}-t_p)$$

with T_1 the revolving period, $T_2=CT_1/360^0$, the period corresponding to one basic angle, $\varepsilon=-1,0,+1$ and $\nu=-c,..,0,..c$ with c the number of scan circles. This is a necessary, but not a sufficient, condition. Define $\psi_p=t_p*360^0/T_1$ and $f_p=\psi_{p+k}-\psi_p$, *i.e.* the length of the p^{th} superframe, then two B-splines can be connected only if

$$-f_q \leq \psi_q-\psi_p+\nu*360^0+\varepsilon*C \leq f_p$$

Of course the analogy with the continuous graph of the star parameters in the geometric reduction step is clear. A difference is however that the modulo operator is missing, whereas a new term $\nu*360^0$ is introduced. Another difference is that the dimensions of the field pertaining to a superframe are variable.

Assign a continuous "numbering" $Num(\psi)$ to the attitude parameters, with $Num(\psi)$ in the range $[0,1)$. The *circular* bandwidth is defined as

$$\beta_i^c(G)=\text{Max}\{\ (Num(\psi_i)-Num(\psi_j))Mrnd(1)\ \mid\ \psi_j \in Adj(\psi_i)\ \}$$

with $Adj(\psi_i)=\{\ \psi_j \mid \{\psi_j,\psi_i\} \in E(\psi)\ \}$. Now consider the class of orderings $Num(\psi)=(\psi\ Mod\ C_m)/C_m$, *i.e.* the attitude parameters are ordered modulo an angle C_m. The circular bandwidth is then

$$\beta_i^c(G)=\text{Max}\{\ (\varepsilon C+\nu 2\pi+\delta)Mrnd(C_m)/C_m\ \mid\ \varepsilon=-1,0,+1,\ \nu=-c,..,0,..c,\ -f \leq \delta \leq f\ \}$$

The value for C_m should be chosen in such a way that the bandwidth β^c is minimized. From the formula for the bandwidth follows that:

- C_m should be as large as possible, because C_m is the denominator in the formula for the bandwidth,
- C_m should be an approximate divisor of $\nu 2\pi+\varepsilon C$ since $f<C$, *i.e.* C_m should be a divisor of 2π and C.

Obviously C_m=60° or C_m=360° are good choices, depending on which condition is the more important. If C_m=60° is chosen the circular bandwidth is

$$\beta^c(G^{60}) = (|C-C_m|+f)/C_m$$

On the other hand if C_m=360° is chosen the circular bandwidth becomes

$$\beta^c(G^{360}) = (C+f)/C_m$$

The turn-over point occurs for f=9.6°, the equivalent of 150 B-splines per circle. The circular bandwidth is then 20%. If more than 150 B-splines per circle are used the modulo 60° ordering is better, otherwise the modulo 360° ordering. These bounds give only a rough idea because fluctuations in f and gas-jet actuations were not taken into account.

In figure 8.8 and table 8.3 the results of various orderings for CERGA dataset II have been summarized. Figure 8.8 gives in the lower triangle $Nz(\mathbf{N})$ and in the upper triangle $Env(\mathbf{N})$ for the normal matrix of the B-spline part. In table 8.3 $|Nz(\mathbf{N})|$, $|Nz(\mathbf{L})|$ and $|Env(\mathbf{N})|$, plus some computing times, are given. The banker's ordering gives again the best overall result in profile and cpu time, with the modulo 360° ordering on a good second place. The modulo 60° and the reverse Cuthill-McKee ordering were not much worse. The normal matrix as ordered by the banker's algorithm can be factored three times faster than the original matrix, and the same holds for the computation of the sparse inverse. The overall gain in cpu time is considerable. The nested dissection ordering resulted in a fill-in ($Nz(\mathbf{L})$=151408) already larger than the profile of the Choleski factor for banker's. Therefore the cpu time for the factorization is certainly larger than any of the other methods, except maybe for the natural ordering.

Table 8.4: Results of reordering strategies of the B-spline parameters for CERGA dataset II. n=719, $|Nz(\mathbf{A})|$ = 38850. RCM=Reversed Cuthill-Mckee, BN=banker's, M60=Modulo 60°, M360=Modulo 360° and NAT=natural ordering

	NAT	M60	M360	RCM	BN
$\|Env(\mathbf{A})\|$ #	221563	175627	151415	159994	145955
%	85.6 %	67.8 %	58.4 %	61.9 %	56.5 %
$\|Env(\mathbf{A})\|-\|Nz(\mathbf{L})\|$	?	?	?	662	721
cpu times (sec):					
ordering	-	< 1	< 1	6.4	6.6
factorization	1008	620	403	?	374
sparse inverse	2540	1495	828	?	?

Figure 8.8 (on the next three pages) - The non-zero structure (in the lower triangle) and envelope (in the upper triangle) of the normal matrix for the B-spline part of CERGA dataset II, as results from six different ordering schemes (the envelope is not given in figure 8.8.f).

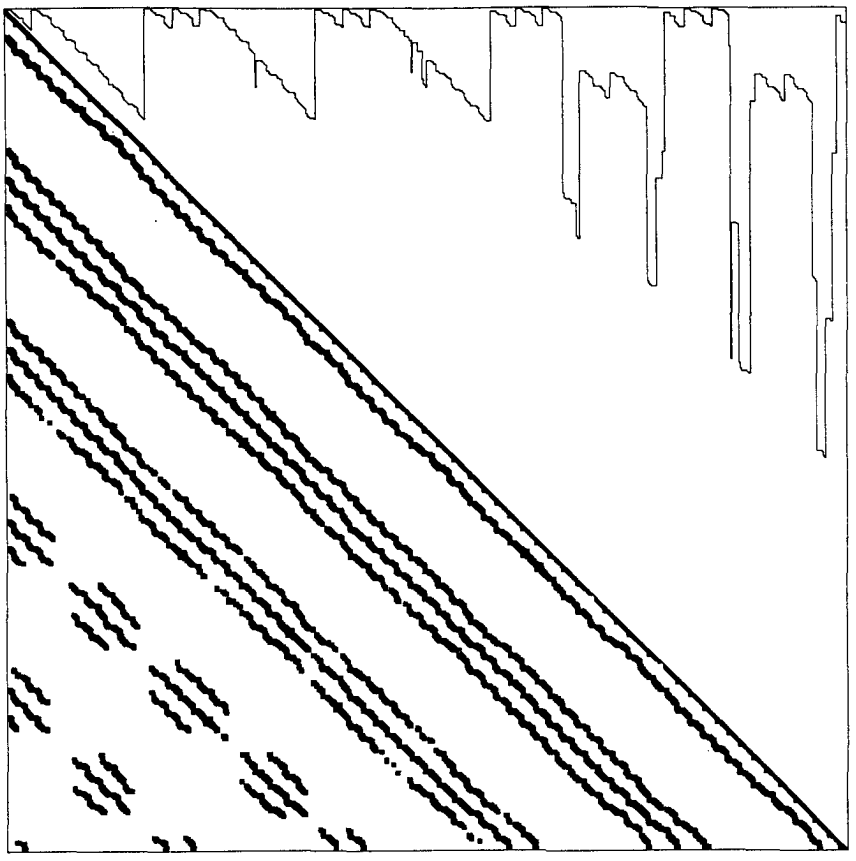


Figure 8.8.a

Natural

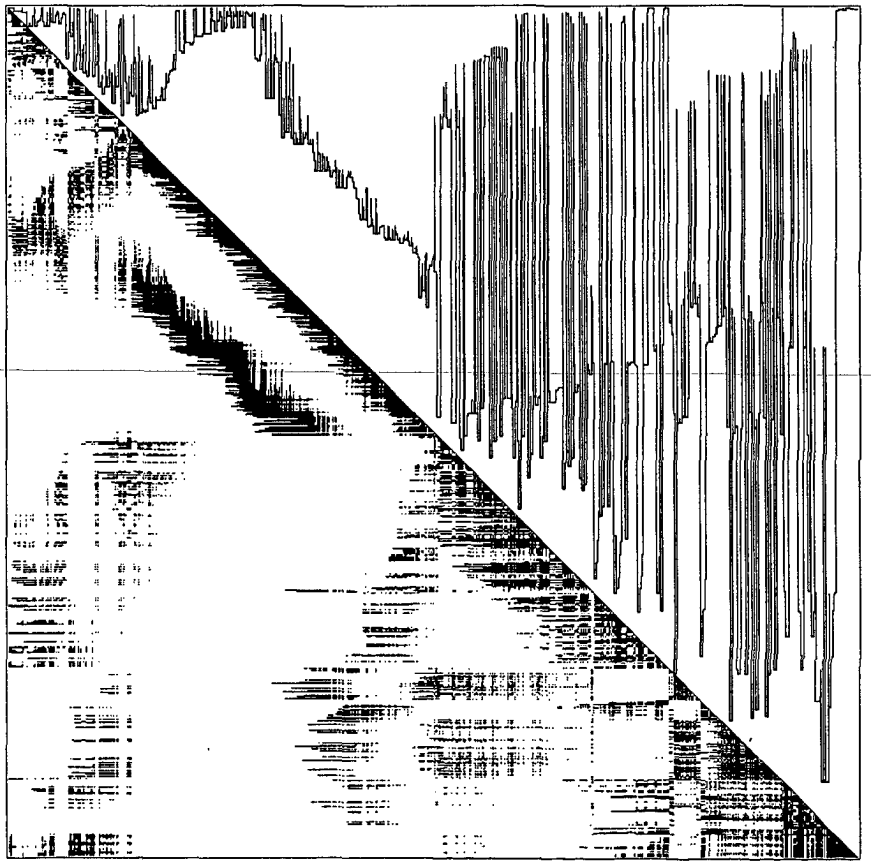


Figure 8.8.b

Banker's

Figure 8.8.c

Modulo 360^0

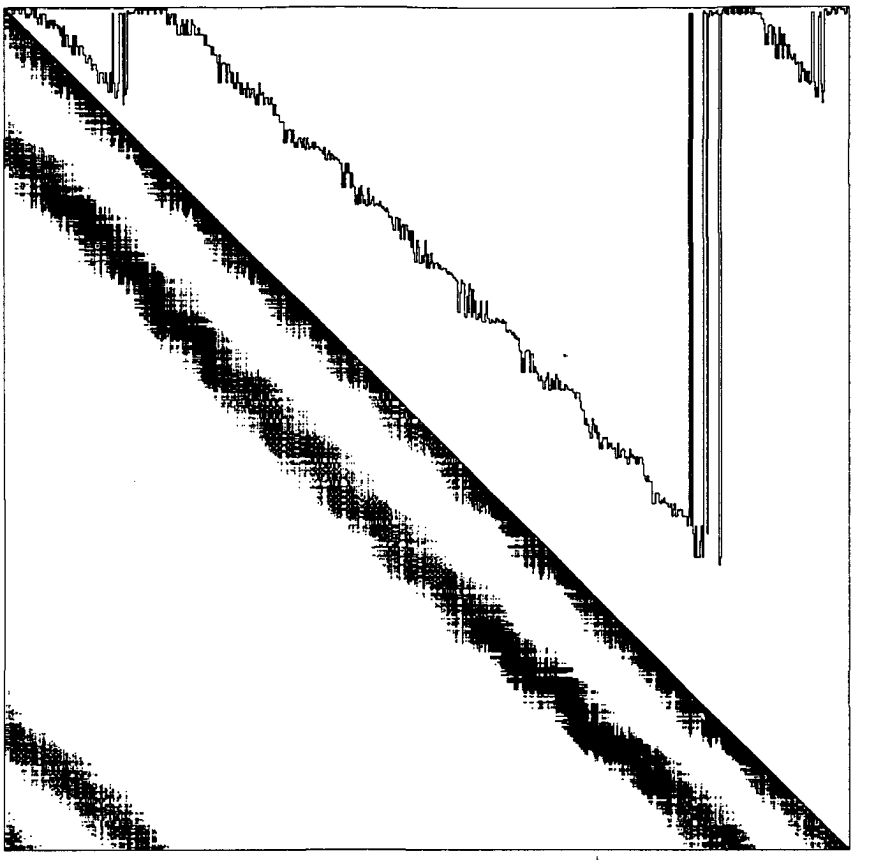
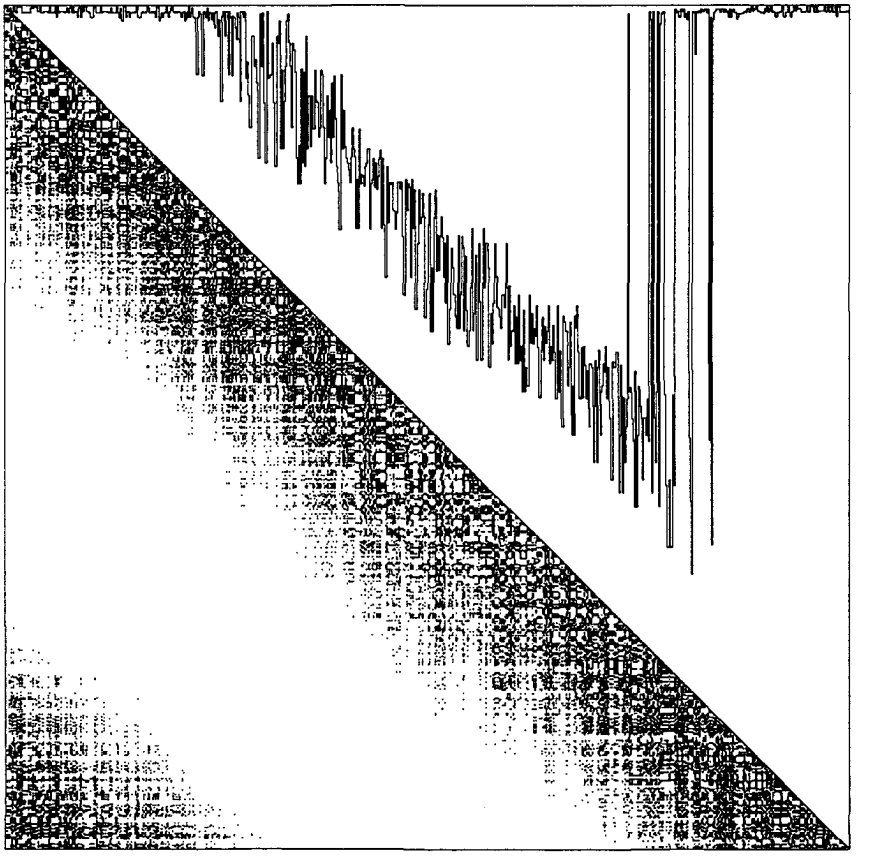


Figure 8.8.d

Modulo 60^0



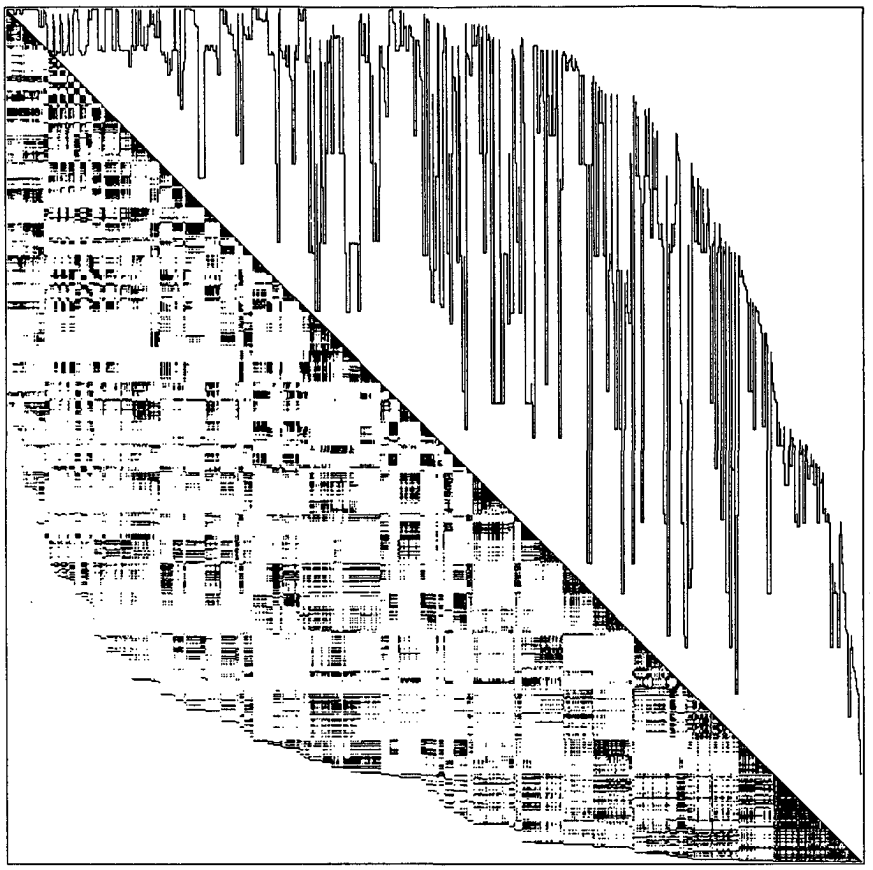


Figure 8.8.e

Reverse Cuthill-McKee

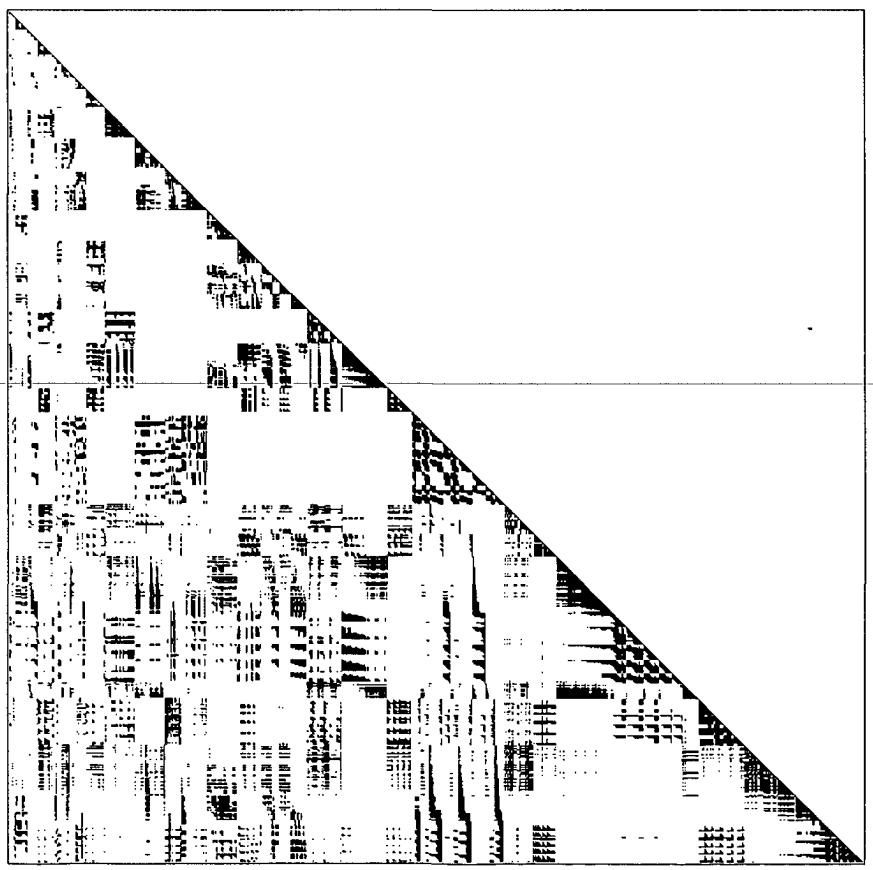


Figure 8.8.f

Nested dissection

CHAPTER 9

GRID STEP AMBIGUITY HANDLING

The grid step errors, resulting from a wrongly computed integer slit number, form a special problem in the Hipparcos data reduction. The great circle reduction is not able to find, from all possible combinations of consistent slit numbers, the combination without grid step errors. Therefore, the RGC abscissae may be wrong by one or more grid steps. One of the objectives of the great circle reduction is to make the slit number information within a RGC-set *consistent*. It is left to the sphere reconstitution and astrometric parameter extraction to repair the grid step errors in the RGC abscissae.

9.1 Introduction

The Hipparcos main instrument cannot measure the along scan grid coordinate, which is needed by the geometric adjustment, directly. The prime observables are the modulation phases of stars visible in the field of view. The along scan grid coordinate of star i, observed in a frame labelled k, can be computed from

$$g_{ki} = (n_{ki} + \varphi_{ki}) \cdot s \qquad (9.1)$$

with φ the observed main grid phase ($0 \leq \varphi < 1$), n the integer slit number and s the grid period (equation 4.13). The slit number is not observed by the main instrument; it will be computed from approximate data. An error in one of the computed slit numbers results in a large error ($\simeq$1"208) in the grid coordinate. The error in the grid coordinate due to a wrongly computed slit number is called a *grid step error*.

The observed grid phase is already corrected for the small and medium scale distortions of the grid. Therefore, we can consider the grid coordinate system defined by equation (9.1) as being attached to a perfectly regular grid in a flat plane, with a constant grid period s. The G-axis is perpendicular to the slits, the origin g=0 is chosen on a reference slit - which is also the second axis - and g is increasing in the direction of the moving star images. The second axis (H-axis) is not relevant and is not further used. The grid coordinate system is obtained through a mapping of the celestial sphere, through the Hipparcos optical system, into a perfectly regular grid in a flat plane. The mapping consists actually of two parts. In the first part the field coordinates x and y are computed from the apparent star positions and the attitude of the satellite (see chapter 4). This mapping is written symbolically as

$$\mathcal{A}: \mathbb{R}^2 \rightarrow \mathbb{R}^2 \quad (x,y) = A(\lambda,\beta,\mathbf{a}(t)) \qquad (9.2)$$

whith λ and β the two coordinates which give the apparent star position in a celestial reference frame and $\mathbf{a}(t)$ a vector with the three attitude parameters of the satellite. In the second part the grid coordinate g is computed from the field coordinates x and y (spherical angles on the celestial sphere), this is the so-called field to grid transform of chapter 4. This mapping is basically a projection of the sphere onto a flat plane,

but it also takes the large scale instrumental distortion of the instrument into account:

$$\mathcal{G}: \mathbb{R}^2 \rightarrow \mathbb{R}^1 \quad g = G(x, y; B\text{-}V, t, f) \tag{9.3}$$

where f is the field of view index, $B\text{-}V$ the star colour index, t the time. Most of the parameters in these mappings are not known a-priori with sufficient precision and must be determined during the reduction, *viz.* the attitude and instrumental parameters.

The fractional and integral part of g/s are respectively equal to the modulation phase and integer slit number, thus,

$$\begin{aligned} \varphi_{ki} &= g_{ki}/s - Entier(g_{ki}/s) \\ n_{ki} &= Entier(g_{ki}/s) \end{aligned} \tag{9.4}$$

which is the inverse of equation (9.1). Therefore, a very straightforward way for computing slit numbers is to use equation (9.4) with approximate values g^0 for the grid coordinate, which can be computed from the mappings $\mathcal{A}$ and $\mathcal{G}$, with approximate values for the star position, the attitude and the instrumental deformation. This gives the following estimate for n:

$$n^0 = Entier(g^0/s) \ , \quad g^0 = G(x^0, y^0; B\text{-}V, t, f_i) \tag{9.5}$$

A first observation on the formula for n^0 is that it does not have the same behaviour for every value of g: *viz.* for g close to a multiple of s even a small error in g^0 can result in a error in the slit number. A better formula is

$$n^0 = Round(\, g^0/s - \varphi \,), \qquad Round(x) = Entier(x+\tfrac{1}{2}) \tag{9.6}$$

which gives correct results if the error in $|g^0/s - \varphi| < 0.5$. The method of equation (9.6) is referred to as the *straightforward slit number computation.*

In the next sections we will investigate the percentage of grid step errors, resulting from the straightforward slit number computation, which can be expected during the great circle reduction [*Van den Heuvel & Van Daalen, 1985*]. Here we make some simple statistical assumptions about the error in the approximate attitude and star parameters. In the later sections we will discuss better methods for computing the slit number, and we will discuss ways for the detection and the correction of grid step errors. The great circle reduction aims at only making the slit number information consistent. It does not matter during the great circle reduction if all the grid coordinates of a star have the same grid step error, since this cannot be detected. Therefore, the computed star abscissae can still be wrong by one or more grid steps. These grid step errors in the abscissae must be corrected during the sphere reconstitution and astrometric parameter extraction [*Walter, 1983a, Bastian, 1985*].

9.2 Probability of Grid Step Errors

The occurrence of an error in the slit number, computed by the straightforward slit number computation method, depends strongly on the error in the approximate values of the grid coordinate g, which are computed from approximate values for the star, attitude and instrumental parameters. In this section a stochastic approach is chosen to compute the percentage of grid step errors among the observations of a relatively large sample of stars. Therefore, the error in the a-priori data can be treated as a stochastic quantity, which we assume to have a normal distribution.

The number of grid step errors in n^o depends on the distribution of $g^o/s-\varphi$. Let us call this intermediate quantity ν^o, *i.e* $\nu^o := g^o/s-\varphi$, then the computed slit number is simply $n^0=Round(\nu^0)$. The error in ν^0 is defined as $\varepsilon_\nu = \nu^o-E(\nu^0)$, with the mathematical expectation $E(\nu^o)=n \in \mathbb{Z}$ equal to the expected slit number n. Hence, we have for ε_ν

$$\varepsilon_\nu = \varepsilon_g/s - \varepsilon_\varphi \tag{9.7}$$

with ε_g the error in the approximate grid coordinate and and ε_φ the error in the observed grid phase. The error in the slit number, *i.e.* the grid step error, is therefore

$$\varepsilon_n = Round(\varepsilon_\nu) \in \mathbb{Z} \tag{9.8}$$

The probability of a grid step error is

$$P_{gridstep\text{-}error} = P(\ |\varepsilon_\nu| > 0.5 \ ) \tag{9.9}$$

The probability of one, two and three grid step errors is plotted in figure 9.1 against the standard deviation of ε_ν, assuming that ε_ν is normally distributed.

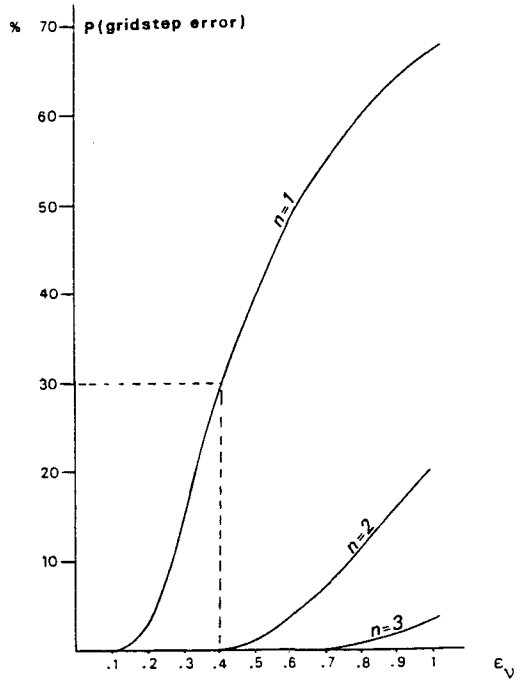


Figure 9.1: The probability of a grid step error
(from [*Van den Heuvel, 1986*])

Consider the observation equations linearized in the errors ε, then ε_g is

$$\varepsilon_g = a_a\varepsilon_a + a_s\varepsilon_s + \Sigma_j a_{ij}\varepsilon_{ij} \tag{9.10}$$

Now putting a_a=1 and a_s=-1, which is perfectly safe, and writing for $\Sigma a_{ij}\varepsilon_{ij}$

simply ε_i, we get

$$\varepsilon_g = \varepsilon_a - \varepsilon_s + \varepsilon_i$$

The standard deviation of φ (and ε_φ) is typically 0.01, very small compared to the grid step errors and the standard deviations which can be expected for g° (and ε_g). Therefore, the contribution of φ in equation (9.7) shall be neglected, i.e.

$$\varepsilon_\nu \simeq 1/s \cdot \varepsilon_g$$

The standard deviation of ε_ν, neglecting the correlations of the instrumental parameters with the star and attitude parameters, is

$$s^2\sigma_\nu^2 = \sigma_s^2 - 2\,\sigma_{sa}^2 + \sigma_a^2 + \sigma_i^2$$

The correlation between the attitude and stars is defined as $\rho := \sigma_{sa}^2/(\sigma_s\sigma_a)$. Then σ_ν is

$$s^2\sigma_\nu^2 = \sigma_s^2 + (\,1 - 2\,\rho\,\sigma_s/\sigma_a)\,\sigma_a^2 + \sigma_i^2 \qquad (9.11)$$

We will now make some simple assumption about the correlation between the star parameters and the star mapper attitude. Two cases can be distinguished: 1) the star in question has been used by the attitude reconstruction, and 2) the star has not been used. In the second case the correlation is zero (ρ=0). In the first case the star and the attitude will be correlated. Here it matters if the star gets a correction during the attitude reconstruction; then it is reasonable to assume that the star and attitude parameters are fully correlated, *i.e.* ρ=1. If, on the other hand, the star is not corrected, the correlation will be much smaller, because of the large smoothing interval used during the attitude reconstruction. Hence we shall assume in this case that ρ=0 (but it is almost certainly larger than this).

9.3 Grid Step Inconsistencies

Closely related to grid step errors are the so-called *grid step inconsistencies*. The slit numbers for a specific star are *consistent* when the observations contain no grid step errors at all or all have the same grid step error. The last situation, *viz.* consistent grid step errors, will not be noticed during the great circle reduction, but is also not harmful. Only the abscissa of the star in question will be shifted over one slit period. So data can be consistent, but still having grid step errors. On the other hand, the slit number data of one star is *inconsistent* if there are two or more smaller -but consistent- groups of slit numbers. Slit numbers which do not belong to the largest group are said to have a *grid step inconsistency*. The grid step inconsistencies result in contradictions between the observations of one star during the great circle reduction. Inconsistencies are harmful, but fortunately they can be noticed during the great circle reduction, and therefore corrective actions can be taken.

Now consider the probability of grid step *inconsistencies* among the observations of a star with an error ε_s in its approximate abscissa. The conditional probability , given ε_s, of an error of k grid steps in one observation is

$$P(\ |\varepsilon_\nu - k| < 0.5\ ;\ \varepsilon_s\ ) = \int_{k-.5}^{k+.5} f_{\varepsilon_\nu;\varepsilon_s}(x)\ dx \qquad (9.12)$$

where $f_{\varepsilon_\nu;\varepsilon_s}$ is the conditional probability density function of ε_ν given ε_s, with expectation and variance

$$\begin{aligned} E\{\varepsilon_\nu;\varepsilon_s\} &= -1/s\cdot(1 - \rho\ \sigma_a/\sigma_s)\cdot\varepsilon_s \\ s^2\sigma^2_{\nu;s} &= (\ 1 - \rho^2)\ \sigma_a^2\ + \sigma_i^2 \end{aligned} \qquad (9.13)$$

Now the conditional probability that all l observations (on one RGC) to a star are consistent is

$$P_{consistency;\varepsilon_s} = \sum_{k=-\infty}^{k=+\infty} \left(\int_{k-.5}^{k+.5} \cdots \int f_{\varepsilon_\nu;\varepsilon_s}(x_1 \ldots x_l)\ dx_1 \ldots dx_l \right) \qquad (9.14)$$

where $f_{\ldots}$ is the multivariate probability density function of ε_ν, given ε_s, over the frames $1 \ldots l$. If the error in the attitude and instrument are not correlated in subsequent frames, then

$$P_{consistency;\varepsilon_s} = \sum_{k=-\infty}^{k=+\infty} \left(\int_{k-.5}^{k+.5} f_{\varepsilon_\nu;\varepsilon_s}(x)\ dx \right)^l \qquad (9.15)$$

For practical computations only values of k around zero have to be taken into account. Figure 9.2 shows $P_{consistency;\varepsilon_s}$ for some cases.

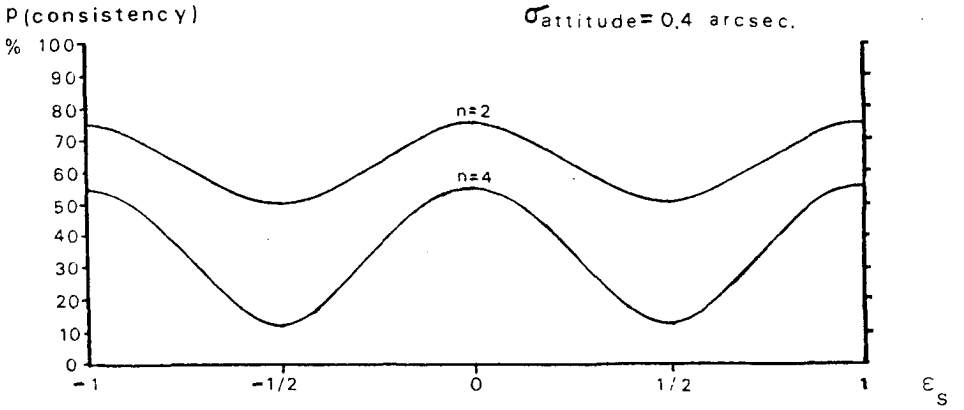


Figure 9.2: Conditional probability of consistency for a star given the error ε_s in the star position (from [*Van den Heuvel, 1986*])

The probability of consistency is at a minimum when $E\{\varepsilon_\nu;\varepsilon_s\}$ is a multiple of 0.5. In this situation, neglecting the correlation between the stars and attitude, and between the attitudes mutually, and neglecting the influence of the instrumental parameters, ε_s is a multiple of s/2. Rounding up or down is then equally probable and the probability of consistency is

then 0.5^{l}. The correlation among successive attitude parameters during the passage of a star through the field of view is actually close to 1, so in the formulae the number of frames l should be replaced by the number of passages through the field of view.

The correlation between stars and attitude can be taken into account by lowering the standard deviation of the attitude. Especially, stars which get corrections in the attitude reconstruction are correlated strongly with the attitude and have variances close to the internal accuracy of the star mapper. So, for these stars inconsistencies will be improbable.

The overall probability of the consistency of stars is

$$P_{consistency} = \int_{x=-\infty}^{x=+\infty} f_{\varepsilon_s}(x)\, P_{consistency;\varepsilon_s}\, dx \qquad (9.16)$$

with f_{ε_s} the probability density function of ε_s. In figure 9.3 the probability of consistency is plotted for several values of σ_a, σ_s and l, the number of passages of a star through the field of view, assuming no correlation between the stars and attitude, and among the attitudes themselves. The number of passages of a star through the field of view, assuming 5 scan circles per RGC, is typically 10 near the nodes of the scanning circles with the RGC, and l=2 for stars at the border of the RGC band 90^0 away from the scan circle nodes. The average number of passages of a star is ~5.

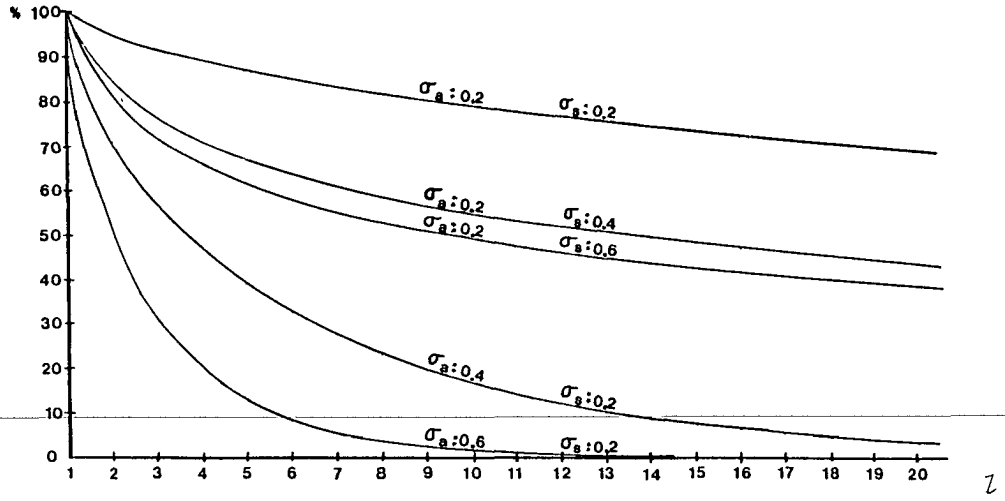


Figure 9.3: The probability of consistency
(from [*Van den Heuvel, 1986*])

9.4 Grid Step Inconsistency Handling

One of the objectives of the great circle reduction is to make the slit number information consistent. The abscissae should of course be computed from consistent, but not necessarily correct, slit number data. Three stages can be distinguished for detection and correction of grid step inconsistencies: *pre-adjustment*, *post-adjustment* and *passive stars* handling.

9.4.1 Pre-Adjustment Slit Number Handling

The pre-adjustment stage is slit number estimation rather than correction, because at this stage the grid coordinates are actually computed for the first time during the great circle reduction. We will discuss three slit number computation methods:
- *straightforward* method,
- *refined* method,
- *approximate sequential adjustment* method.

All these methods use the a-priori given approximate data on a frame by frame basis. The *straightforward* slit number computation method has been introduced already in equation (9.5). Essential is that the straightforward method does not try to improve the approximate data. The two other methods use basically the same kind of data as the first one, but are different insofar they estimate, and use, the error in the along scan attitude and/or the error in the star abscissae.

In the *refined slit number computation* method the error in the star abscissae is estimated from the first observation to each star. In the *approximate sequential adjustment* the error in the star abscissae and along scan attitude are estimated from all available data. Actually, the refined slit number computation is a special case of the sequential adjustment; only the error in the approximate star abscissa is estimated from one observation. The sequential adjustment turns out to be very efficient. This method will be discussed in section 9.5.

9.4.2 Post-Adjustment Slit Number Correction

The *post-adjustment slit number correction* is based upon the analysis of the least squares residuals of the observations after the adjustment. Typical for these residuals is that they are correlated, which results in a certain blurring of the errors. Therefore, the post-adjustment grid step correction procedures must work in an iterative fashion. *I.e.* the most evident cases are first tackled, then a new solution is computed and the residuals are inspected again. The process converges if in subsequent iterations more and more subtle cases are recognized as grid step inconsistencies. Two procedures, *frame by frame* analysis and *sequential analysis*, have been verified experimentally. A third procedure, *star by star* analysis is only used as an evaluation tool. These procedures are discussed in section 9.6.

9.4.3 Passive Stars Grid Step Inconsistency Handling

The along scan attitude and instrumental parameters are computed from the active stars only. The passive stars are computed afterwards, using the instrumental deformation and attitude, already computed. The slit number estimation for observations to passive stars, once a satisfactory solution for the active stars has been obtained, is generally not difficult any more.

The active stars are selected - by the software - by static criteria, *e.g.* the quality of the approximate star positions, double star characteristics, etc., as well as dynamically from the results of the a priori approximate sequential adjustment. The first objective of the active star selection is to have as few as possible grid step inconsistencies in the active stars during the first treatments, in order to keep the number of iterations in the post-adjustment grid step correction low. In particular all the problem stars should become passive. In this way, during internal iterations, re-weighting or rejection of observations, with the associated re-computation of the Choleski factor is not necessary. Only the right hand sides of the normal equations need updating after the grid step inconsistency

correction. On the other hand, we aim at having as few as possible passive stars during the last iteration, in order to get the best possible precision for the active star abscissae and attitude parameters. Generally, the precision of the active star abscissae and attitude parameters decreases slightly by not considering the passive star observations in the active star solution.

9.5 Approximate Sequential Adjustment

In a sequential adjustment only small batches of data are adjusted at a time. In the present case the grid coordinates are adjusted frame by frame, *i.e.* the star positions and satellite attitude are updated using the observations of one frame at a time. The updating formulae for $k=1,2,\ldots,n_a$ are

$$\begin{aligned} \mathbf{P}^{(k)} &= \mathbf{P}^{(k-1)} + \mathbf{A}_k^T \mathbf{W}_k \mathbf{A}_k \\ \mathbf{P}^{(k)}\mathbf{x}^{(k)} &= \mathbf{P}^{(k-1)}\mathbf{x}^{(k-1)} + \mathbf{A}_k^T \mathbf{W}_k \mathbf{y}_k \end{aligned} \qquad (9.17)$$

with $\mathbf{x}$ the vector with corrections to the approximate values of the star and attitude parameters and $\mathbf{P}$ the weight matrix of the unknowns $\mathbf{x}$, where $\mathbf{x}^{(0)}=\mathbf{0}$ and $\mathbf{P}^{(0)}$ the weight matrix of the approximate values, and with $\mathbf{y}_k$ the vector with the observed minus computed value of the observations in the k'th observation frame, $\mathbf{W}_k$ the diagonal weight matrix of these observations and $\mathbf{A}_k$ the design matrix. $\mathbf{A}_k$, $\mathbf{W}_k$ and $\mathbf{y}_k$ are small parts, corresponding to one observation frame, of the design matrix $\mathbf{A}$, weight matrix $\mathbf{W}$ and vector of observations $\mathbf{y}$ which arose in chapter 7. $\mathbf{P}^{(k)}-\mathbf{P}^{(0)}$ is equal to the normal matrix update after k frames. Therefore, except for the small addition $\mathbf{P}^{(0)}$, the normal equations, and hence the solution, at the end of the sequential adjustment are practically the same as the least squares solution computed in chapter 7 (in fact by adding the small weights $\mathbf{P}^{(0)}$ the normal equations are regularized, *i.e.* the rank defect disappears).

At each step of the sequential adjustment, just before the updating of the corrections to the star and attitude parameters, the observations are checked, and corrected, for grid step inconsistencies. If the observed value of the grid coordinate minus the value computed from the updates is larger than half a grid step the observed value must be corrected by one or more grid periods. The condition is

$$|\Delta g_{ki} - a_i \Delta\psi_i^{(k-1)} - a_k \Delta\psi_k^{(k-1)}| > s/2 \qquad (9.18)$$

In order to evaluate this condition it is necessary to compute for each frame the up-to-date attitude and star parameters. Let us introduce the vector $\mathbf{p}$, then the update formulae are

$$\begin{aligned} \mathbf{P}^{(k)} &= \mathbf{P}^{(k-1)} + \mathbf{A}_k^T \mathbf{W}_k \mathbf{A}_k \\ \mathbf{p}^{(k)} &= \mathbf{p}^{(k-1)} + \mathbf{A}_k^T \mathbf{W}_k \mathbf{y}_k \end{aligned} \qquad (9.19.a)$$

and $\mathbf{x}$ is computed by solving

$$\mathbf{P}^{(k)}\,\mathbf{x}^{(k)} = \mathbf{p}^{(k)} \qquad (9.19.b)$$

It is too much work to solve equation (9.19.b) in a rigorous manner for every observation frame. Therefore, an approximate sequential adjustment has been developed.

The approximate sequential adjustment is based on a recursive algorithm. There are three computation steps, which have to be carried out for each observations frame:

1. compute a prediction of the along scan attitude,
2. check the slit numbers,
3. update the attitude and star abscissae from the main grid data.

Here we consider the linearized observation equations for the set of stars P_k observed in the k^{th} frame

$$\Delta x_{ki} = a_k \Delta\psi_k^{(k-1)} + a_i \Delta\psi_i^{(k-1)} \qquad i \in P_k \tag{9.20}$$

let w_{ki} be the observation weight and let us define

$$w_i^{(k-1)} = w_i^0 + \sum_{l=1}^{k-1} w_{li} \tag{9.21}$$

the sum of all observation weights for star i over the preceding frames, with w_i^0 the weight of the approximate star abscissa.

In the first step a prediction for the along scan attitude is computed -as a weighted mean- from the following information:

1) the star mapper attitude $\Delta\psi_k (\equiv 0)$ with weight w_k^0,

2) a prediction based on the attitude of the preceding frame

$$\Delta\psi_k^{(k-1)} = \Delta\psi_{k-1}^{(k-1)} + \omega_{k-1,k} \cdot \Delta t + \psi_{k-1}^0 - \psi_k^0 \tag{9.22}$$

with ω the average scanning velocity, including the effect of possible gas jet actuations, over the interval Δt (a multiple of the frame period $T4$) determined from the star mapper data. The weight $\bar{w}_k^{(k-1)}$ of $\Delta\psi_k^{(k-1)}$ is computed by propagation of the variances, *i.e.* $\bar{w}_k^{(k-1)} = \dfrac{w_{k-1}^{(k-1)} \cdot w_\omega}{w_{k-1}^{(k-1)} + w_\omega}$.

The attitude estimate $\Delta\psi_k^{(k-1)}$ is not changed by the correction to the star mapper attitude, because the star mapper attitude has been used for the linearization of the equations. However, the weight must be increased accordingly, *i.e.*

$$w_k^{(k-1)} = \bar{w}_k^{(k-1)} + w_k^0$$

In the second step the observed grid coordinate is checked against a grid coordinate calculated from the attitude computed by the first step and the most up-to-date star abscissae, *i.e.*

$$\Delta = \Delta x_{ki} - a_k \Delta\psi_k^{(k-1)} - a_i \Delta\psi_i^{(k-1)} \qquad \forall\ i \in P_k \tag{9.23}$$

and if $|\Delta| > s/2$ then the observed grid coordinate is corrected by a number of slit periods.

In the third step the current attitude and star abscissae are improved using the main grid observations. The attitude is computed as a weighted mean

from the attitude estimate computed in the first step and the attitude estimate computed from the observed grid coordinates and corrections to the star abscissae, *i.e.*

$$\Delta\psi_k^{(k)} = \frac{1}{w_k^{(k)}} \left[\sum_{i \in P_k} \bar{w}_{ki} \left(\Delta x_{ki} - a_i \, \Delta\psi_i^{(k-1)} \right) / a_k + w_k^{(k-1)} \Delta\psi_k^{(k-1)} \right] \tag{9.24}$$

with $w_k^{(k)} = w_k^{(k-1)} + \sum_{i \in P_k} w_{ki}$ and $\bar{w}_{ki} = \frac{w_i^{(k-1)} \cdot w_{ki}}{w_i^{(k-1)} + w_{ki}}$. The star updates are not the same ones as used by the second step. This is done to eliminate certain systematic errors. We will come back to this later. At the end of the third step the star parameters are updated, using the observed grid coordinates and the attitude computed earlier in this step, *i.e.*

$$\Delta\psi_i^{(k)} = \frac{1}{w_i^{(k)}} \left[\bar{w}_{ki} \left(\Delta x_{ki} - a_k \, \Delta\psi_k^{(k)} \right) / a_k + w_i^{(k-1)} \Delta\psi_i^{(k-1)} \right] \quad \forall \; i \in P_k \tag{9.25}$$

with $w_i^{(k)} = w_i^{(k-1)} + \bar{w}_{ki}$ and $\bar{w}_{ki} = \frac{w_k^{(k)} \cdot w_{ki}}{w_k^{(k)} + w_{ki}}$. In the third step of the approximate sequential adjustment also the sum of the residuals (updates) squared for the star parameters is updated. Therefore, at the end of the sequential adjustment, the performance can be evaluated for each stars. This can be used to select active and passive stars.

The sequential adjustment was first very sensitive to errors in the approximate instrumental parameters, notably for the x-term. One would think that the instrumental errors average out after a while. However, this is not the case for the uneven powers of x, unless some precautions are taken. This can be seen as follows; during a star passage through the field of view the weight on the position and attitude updates increases. So the attitude updates computed from observations in *one* half of the field get systematically more weight; thus, in this way errors in the instrumental parameters with uneven x-terms are accumulated in the attitude and star parameters at the end of the third step. In the next frame, again, the star positions are first used to compute the attitude, and finally they are updated. In this way the error is accumulated rapidly, and even small errors (20 mas at the border of the field) result after one basic angle in large discrepancies, resulting in systematically correcting all the observations by the same number of grid steps. The remedy is very simple: update the attitude in the third step by using old star positions, *i.e.* star positions computed during the *previous* passage of the star through the field of view.

9.6 Post-Adjustment Grid Step Inconsistency Correction

9.6.1 Introduction

The *post-adjustment slit number correction* is based upon the analysis of the least squares residuals of the observations after the adjustment. It is typical for these residuals, and functions thereof, that they are correlated. Therefore, an error in one of the observations results not only

in large residuals for the erroneous observation, but also gives large residuals for some errorless observations. This effect is called *smearing*. Smearing may also result in *masking*; an erroneous observation may have a small residual due to the smearing effect of other errors. These two effects result in a certain blurring of the errors. Another effect is the inseparability: two or more errors cannot be separated. This occurs in particular for grid step inconsistency correction since there are many solutions possible for one and the same star (which differ by a grid step). Also it is not always clear in an observation frame (especially in the geometric mode) which of the observations is wrong, possibly resulting in a shift of the attitude by a grid period.

A-posteriori analysis of the least squares residuals - in the form of hypothesis testing - is a common technique in geodesy, see for instance [*Baarda, 1973, Kok, 1985*]. The observations are generally inspected one by one, using the so-called *conventional* alternative hypothesis; *i.e* that there is an error in one of the observations. This procedure is known as *data snooping*. The major problem with these techniques is a lack of *robustness*, caused by smearing and masking effects, which makes that - whatever the procedure for detection and correction is - it should be iterated. *I.e.* the most evident cases have to be tackled first, then a new solution is computed and the residuals, or testing variates, are inspected again. The process converges if in subsequent iterations more and more subtle cases are recognized as errors. This procedure can be automatized. This is for instance done by Kok in his iterated data snooping procedure [*Kok, 1985*], or by Poder [*Eeg, 1986*] in his iteratively re-weighted least squares. More robust alternatives for weighted least squares are un-weighted least squares, least squares with the weights of the a-priori values or L-1 (least absolute values) adjustment.

Our procedure for repairing grid step inconsistencies shows little resemblance with the well known data-snooping in other geodetic software. The rigorous one by one analysis of the observations by *conventional* alternative hypotheses (*i.e* assuming an error in one of the observations) is not feasible because of the large amount of observations, and the high percentage of errors among them. For the same reason it is not possible to correct observations one by one, as is done in geodetic networks which have only few errors, and which is also done in automatic procedures such as iterated data snooping [*Kok, 1985*]. Another difference is that unlike in most geodetic problems the magnitude of the error is known; and it is this fact which has led to some very successful procedures:
- *sequential analysis*,
- *frame by frame* analysis,
- *star by star* analysis.

These post-adjustment grid step correction procedures also work in an iterative fashion. Re-computation of the Choleski factor during these iterations is not necessary. Only the right hand sides of the normal equations need updating after the grid step inconsistency correction.

9.6.2 Star by Star Analysis

In the star by star analysis first all residuals per star are collected into a condensed histogram per star (figure 9.4). The histogram data is tested for normal distribution, using four central moments, minimum, maximum and the 25%, 50% and 75% quantiles as statistics. Two tests are needed, for skewness and for bi-modality. The major problem is to separate between stars which have really inconsistent grid coordinates and stars whose distribution is not normal because of smearing effects. Therefore, the most evident cases should be tackled first and then the procedure should be iterated. For each

iteration two passes through the observation equations are needed: a detection and a correction pass. In the software this technique is not used for the automatic correction of grid step inconsistencies, but it can be used for the off-line inspection. The statistics are used to select only the most interesting cases, and not all of the 2000 histograms.

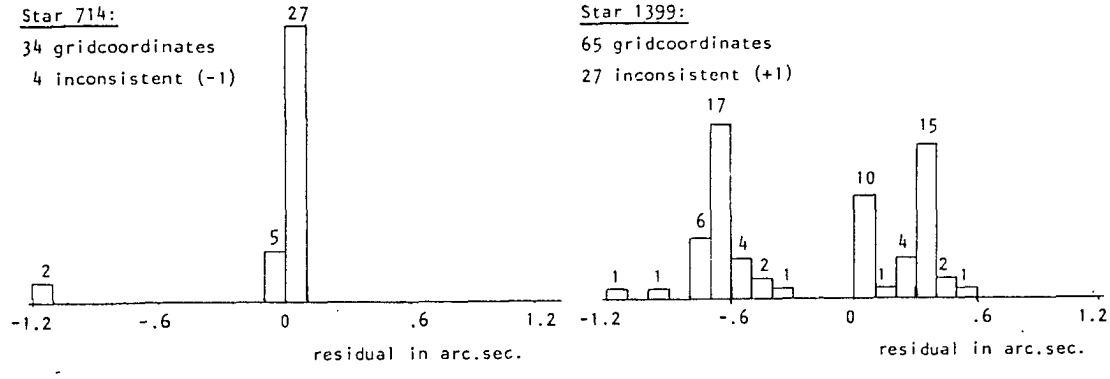


Figure 9.4 - Skew and bi-modal residual histograms of stars with inconsistent grid step numbers

9.6.3 Analysis per Frame

In the analysis per frame the, on the average 4 à 5 residuals per frame are inspected. Correction of suspected observations takes place when the residuals are larger than a certain criterion. In order to avoid *e.g.* cycling of the correction process during iteration, several restrictions have been imposed such as:
- only one grid coordinate per frame is corrected,
- all corrections to observations of one star should have the same sign.

The value of the criterion depends on the global statistical test, which is computed *before* each iteration, and the iteration number. The performance of this method turns out to be very sensitive for the value of the criterion. A rule for calculating this value from the global statistical test and iteration number is therefore not easy to give. Per iteration only one pass through the observation equation file is sufficient; the detection and correction (*i.e.* computation of new right hand sides of the normal equations) can be done on a frame by frame basis at the same time. So this method is approximately half as expensive (in computing time) as the star by star analysis method.

A slightly different procedure has been proposed and tested by [*Van den Heuvel, 1986*]. In his procedure correction of suspected observations takes place on the basis of the size and the *pattern* of the residuals. This procedure has not been implemented in the final software, because the generalization of his procedure resulted in the so-called a-posteriori sequential adjustment [*Van der Marel, 1987*].

9.6.4 A-posteriori Sequential Analysis

This is basically the same process as in the pre-adjustment stage, but now on the basis of the estimated abscissae and instrumental parameters. The attitude prediction (first step) is now computed differently. Namely, as the attitude prediction the median of the residuals themselves is used (instead of the weighted mean, which is the least squares estimator).

The sequential analysis and adjustment give the best performance. Several simulation experiments showed that the majority, say 95%, of the inconsistencies are corrected in the first iteration. The few remaining inconsistencies were, in our experiments, always solved in the second iteration, when we have again improved our star abscissae and instrumental parameters. For the residual analysis per frame in general much more iterations are needed, *i.e.* typically 5-10 iterations. Furthermore, the performance is sensitive for the value of the criterion and a general rule for computing this value has not yet been established. For these reasons the sequential adjustment will be the base line in the operational software.

9.7 Results

The straightforward grid step computation algorithm gives, in case of the Lund dataset, 4731 slit number errors, which resulted in 523 inconsistent stars (27%). This is not a very large percentage, since the attitude was already quite good (0.1"). The error in the star positions was about 1.5", but this is not very relevant for grid step inconsistencies.

Both the refined slit number estimation method and the approximate sequential analysis succeeded in repairing these inconsistencies. The rms error in the approximate instrumental parameters was 20 mas. After the adjustment, abscissae of 1400 stars (70%) were wrong by one or more grid steps. This large percentage is certainly caused by the large uncertainty in the a-priori positions.

The a-posteriori sequential adjustment and the analysis by frame method have been tested on the Lund dataset. The slit numbers were computed by the straightforward method. Two iterations with the a-posteriori sequential adjustment were needed to correct the inconsistencies: in the first iteration 4984 slit numbers were corrected (only 4728 were effective because at the same time 50 stars were shifted over one grid step). After the first iteration only 7 stars with 10 inconsistent slit numbers remained, which were all corrected in the second iteration. The sample standard deviation was reduced from 258 mas, before correction, to 21.7 mas and finally 14.27 mas after the last iteration (if modelling errors are absent the expected sample standard deviation is 10 mas). At the end of the great circle reduction 1360 stars were wrong by one or more grid steps.

APPENDIX A

FAST GREAT CIRCLE REDUCTION SOFTWARE

In this appendix a short description of the structure of the FAST great circle reduction software is given.

A.1 Software set-up

The FAST great circle reduction software has been developed at the faculty of Geodesy of the Technical University of Delft. The software will be installed at Centre National d'Etudes Spatiales (CNES) in Toulouse and at the Space Research Laboratory in Utrecht. The software consists essentially of two parts [*Van der Marel, 1986*]:

1) The *kernel software*, which performs the actual least squares adjustment,
2) The *monitoring software*, which consists of several interactive programs for inspection, analysis and control purposes.

The actual great circle reduction is performed by the kernel software, which consists of 7 Fortran 77 programs. The monitoring software is used to update the parameter files needed by the kernel software, to inspect files and to analyze the results. The monitor software is used in an interactive fashion, whereas the kernel software is almost completely automatic.

The kernel and monitoring software are supported by an extensive software library with *file handling* and *error handling* subprograms dedicated to the great circle reduction software, plus *general purpose* mathematical and statistical subprograms, as well as several utilities (printer plots, histograms, etc.). All the software is written in a portable subset of Fortran 77 [*FAST, 1986b*].

The input files for the great circle reduction are generally created by other tasks in the FAST data reduction, *viz.* the grid coordinate and attitude reconstitution task, and by the so-called reception and preparation task. The reception and preparation task converts the ESOC files into the FAST interface files [*FAST, 1986a*], but also the corrections for apparent places are computed here. The output files are either used directly by other tasks, or are processed by the data management and control system (DMCS). The data management control system (DMCS) is also responsible for the automatic running of the data reduction tasks [*Huc, 1985, Pieplu, 1986*].

A.2 Kernel software

The output of the kernel software consists of star abscissae on a reference great circle (RGC), a one-axis attitude and a set of coefficients describing the large scale disturbance of the instrument. These quantities are estimated by a large scale least-squares adjustment from the grid coordinates, which are the main input to the kernel software. The corrections for apparent places, needed to arrive at star abscissae corresponding with mean geometric positions are computed beforehand and form an additional input to the great circle reduction. Another input consists of approximate values for the geometric star positions and for the three-axis attitude used in the linearization of the equations.

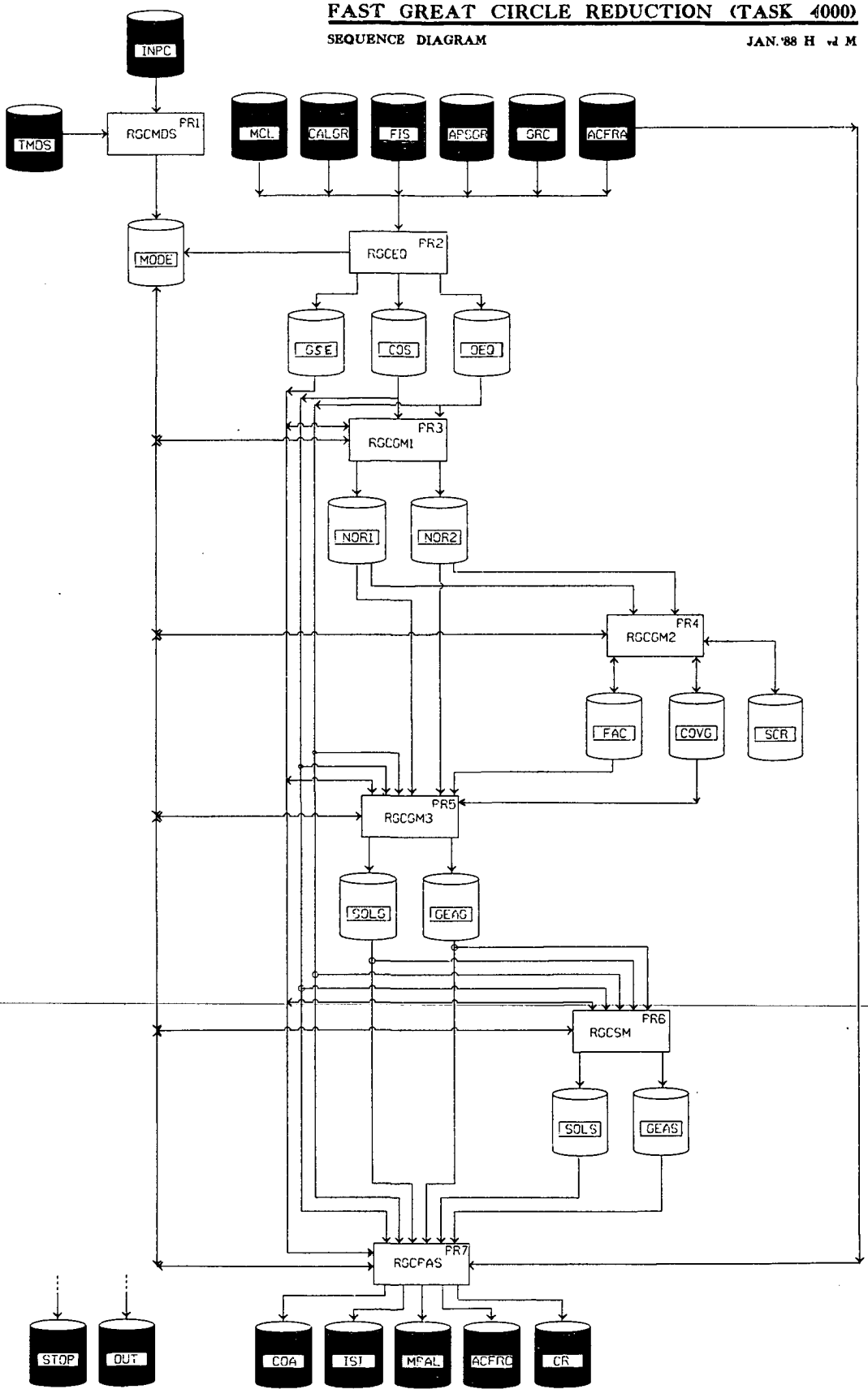


Figure A.1 - Sequence diagram GCR

The software is able to work in three different modes: *geometric*, *smoothing* and a *combined* mode. The operation modes are different with respect to the solution of the one-axis attitude. In geometric mode the attitude is solved in the form of one abscissa per frame: the geometric attitude. In the smoothing and combined mode the geometric attitude is smoothed and represented by a much smaller number of coefficients in a B-spline representation. The smoothing and combined modes differ insofar that the combined mode operates in two steps: first the geometric solution is formed explicitly, then it is smoothed. Whereas in the smoothing mode, the equations are solved in a single step without forming the geometric solution first.

In figure A.1 the configuration of the kernel programs and files in the combined operation mode are given. Programs are represented by boxes, files as "barrels" and the arrows indicate which file is read by which program [*Van der Marel, 1986a*].

A.2.1 Kernel Software Modules

RGCMDS: *Mode Selection*. In this program the operation modes and control parameters for the great circle reduction software are selected from a pre-defined table. The table, which is stored in the file TMDS, gives the values of the control parameters and operation modes as a function of the RGC-epoch (in 6 months intervals) and the iteration number. The selected parameters are written to the file MODE, which is read and updated by each of the other kernel programs.

RGCEQ: *Observation Equations*. The linearized observation equations (chapter 5) and integer slit numbers (coupled with a first grid step inconsistency correction, see chapter 9) are computed, and active stars are selected. The active stars are selected on a-priori information contained in the input files and results of the grid step inconsistency correction.

RGCGM1, **RGCGM2** and **RGCGM3**: *Geometric solution*. The least squares solution with a frame-by-frame attitude is computed. Firstly, in RGCGM1, the normal equations of the active-star and instrument part, after elimination (solution) of the frame-by-frame attitude parameters, are formed (chapter 7). The star unknowns are labelled according to the modulo 60° or banker's ordering in order to optimize the factorization time (chapter 8). In RGCGM2 the Choleski factor and the sparse inverse of the block partitioned normal matrix are computed. The instrument part and the diagonal of the star part, respectively corresponding to the covariance matrix of the instrumental parameters and the variances of the active stars, are written on the file COVG. Finally, in RGCGM3 the solution is computed, tested, and whenever necessary, grid step inconsistencies are corrected. This chain of programs is only run in the geometric and combined solution modes, but not in the smoothing mode.

RGCSM: *Smoothed Solution*. The frame-by-frame along scan attitude is smoothed in a combined adjustment with the active stars (chapter 7). The smoothed attitude is represented as a B-spline series (chapter 6). First the B-spline knots are selected using a-priori information about the gas jet actuation times and other events. Then the normal equations are formed, star parameters are eliminated, and the reduced system is solved using Choleski factorization. The order of the B-spline parameters in the reduced normal equations corresponds to the modulo 360° ordering (chapter 8). This program is run in the smoothing and combined solution modes. In the combined solution mode it is not necessary to re-compute the instrumental parameters. They are already very well determined in the geometric solution.

RGCPAS, *Passive stars*. The stars not marked as active stars in RGCEQ are treated in this module. Passive stars do not participate in the least squares solution computed in RGCGM and RGCSM. The abscissae of the passive stars are computed by this program, using the attitude, instrument and active star parameters computed in RGCGM or RGCSM. The abscissae of the active and passive stars are tested statistically and the results are written on the file COA.

The division of the great circle reduction into several programs is somewhat arbitrary. The present choice of processors aims at minimizing the use of (virtual) memory. Of course, small programs use little memory space themselves, but more important are the savings in storage locations for data. Each program keeps only in-core those data which is really needed, which is different from program to program. This cannot be brought about in a single program with the same clarity. This plays in particular a role in the geometric solution process, which is split into three programs.

A.2.2 Files

Three types of files can be distinguished: input, output and intermediate files. The input files are created by other tasks in the FAST data reduction, *e.g.* the "reception and preparation", the grid coordinates and attitude reconstitution tasks. The output files are either used directly by other tasks, or are processed by the data management and control system (DMCS) to form even larger files needed for the sphere reconstitution and astrometric parameter extraction. Intermediate files are used to exchange data between individual programs of the great circle reduction task.

Input files:

management data:

INPC: interface with the Data Management and Control System (DMCS).

TMDS: file which contains the tables with mode selection and control parameters for the great circle reduction.

MCL: file with mission control data, *e.g.* times of gas jet actuation, RGC definition.

approximate values:

FIS: approximate values for the geometric star position in the RGC reference frame, star magnitude and colour index, several quality indicators plus various flags.

CALGR: calibration data and approximate values for the large scale instrumental distortion.

ACFRA: approximate values for the three-axis attitude. The transversal attitude components are determined by the attitude reconstruction task. The along scan component comes from a previous iteration, except during first treatment of an RGC when the along scan component is also determined by the attitude reconstruction.

observations:

GRC: file with the grid coordinates (modulation phase), which are the main input for the great circle reduction, the standard deviation of the grid coordinates and the results of statistical tests on the modulation.

APSGR: file with the corrections for apparent places at the time of observation.

Output files:

adjusted values:

COA: file with the adjusted abscissae, their standard deviations and a status indicator, which gives the results of statistical tests, grid step inconsistency correction and contains several flags. This file is further processed by the data management and control system to form a single file which contains also the abscissae of other RGC's. The new file will be input to the sphere reconstruction and astrometric parameter extraction.

ISI: file with the adjusted large scale instrumental distortion and its covariance matrix. This file is read by the calibration software and is eventually used to compute improved versions of the calibration file CALGR.

ACFRC: contains the frame-by-frame three axis attitude, of which the along scan component is computed by the great circle reduction software. The other two components are copied from the ACFRA file. The along scan attitude on this file is used in the next iteration by the attitude reconstitution software. This file will be forwarded to the Tycho data reduction consortium after the final iteration of the Hipparcos data reduction.

MPAL: gives the abscissae of minor planets for each frame in which they are observed. This file is processed by an off-line task, which is going to compute the ephemeridis of minor planets.

reports:

CR: circle report file. This file contains a summary of the great circle reduction results, complementary to the other output files. This file can be processed by one of the monitor programs in order to produce an astrometric report of the great circle reduction.

OUT: operation report which contains simple messages indicating success or otherwise abnormalities, and optionally debugging results.

STOPFIL: interface with the DMCS.

Intermediate files

MODE: the "brain" of the great circle reduction software. This file contains the value of all mode selection and control parameters, plus other results collected during execution.

COS: contains condensed information from FIS, wether a star is an active star and the connection structure. The connection structure indicates which stars belong to the same connected component, and which star is selected as a base-star.

OEQ: the "vein" of the great circle reduction software. This large file contains the linearized observation equations, *i.e.* the observed grid coordinate minus the computed value from approximate data, the partial derivatives for the attitude and star part and the field coordinates. The coefficients of the instrumental parameters in the linearized equations are not stored since they can be re-computed cheaply when needed. Besides, the coefficients are not needed very often, because the instrumental dislocation can be computed as efficiently from the other estimated parameters and field coordinates.

GSE: results of the grid step inconsistency correction. This file is complementary to the observation equations file OEQ.

SOLG, SOLS: files which contain the least squares solution of the active stars (geometric and smoothed solution respectively).

COVG, COVS: files with the computed covariance information (geometric and smoothed solution respectively).

Other intermediate files, of minor importance, are: **GEAG** and **GEAS** which contain the computed frame-by-frame attitude and least squares residuals, and **NOR1**, **NOR2** and **FAC** which contain respectively the normal matrices and Choleski factor for the geometric solution.

A.2.3 Error Handling

The error handling software is responsible for the (error) messages and debugging results on the file OUT. There are three levels of (error) messages, which results in different actions:
- Fatal Error; execution is terminated immediately,
- Warning; execution is terminated only after the number of warnings exceeds a certain bound,
- Message; only a message is written in the OUT file.

In case of an fatal error or warning a message and a trace-back of calling routines is written on the OUT file.

With the error handling software also debugging results can be obtained. The amount of debugging can be selected by two types of parameters:
- the debugging level; if the debugging level is zero no intermediate output, except (error) messages, is produced on the OUT file. For levels larger than zero the amount of output grows rapidly.
- subroutine names with alternative levels; useful when only for a small part of the software debugging output is needed.

The debugging output itself should be distinguished from the output produced by the off-line monitoring software. Firstly, the debugging output is not always interesting for astrometric purposes, and secondly if more debugging output is desired the software should be run another time. With the monitoring software a variable degree of output can be obtained without re-doing the great circle reduction.

A.3 Monitoring Software

The monitoring software is in a sense complementary to the kernel software. The kernel software, which is almost completely automatic, produces only little information in readable form. The only formatted file, which can be printed, produced by the kernel software is the -normally short- OUT file. This file is only intended for the operator, and it only gives the most important test results. Of course, more output on the OUT file can be obtained with higher debugging levels, but this must be specified before start of the reduction. Another way to obtain more output is through the monitoring software. The following programs are available for these purposes at the moment:

CMSCRS: *circle report stars*. Evaluates and analyses the star abscissae and instrumental parameters computed during the great circle reduction. If it concerns simulated data a comparison with the true errorless simulated data is made.

CMSFIL: *file monitor*. Most of the input, output and intermediate files are not ASCII files, therefore, they cannot be inspected directly from the terminal. The file monitor is a program which can open each of these files, and show, or print, the complete file or just some data concerning a specific star or stars, and/or frame times. The last option is very useful to inspect large files such as APSGR, GRC and OEQ.

CMSATT: evaluation of the (smoothed) attitude.

CMSGSE: interactive inspection of L.S. residuals for grid step inconsistency correction purposes.

Besides these off-line inspection programs a few other programs are needed concerning the selection of operation modes. These programs are:

CMSTAB: interactive program for updating of the mode selection table file TMDS.
CMSMDS: interactive version of the mode selection program. It allows for basic selections from the TMDS file and modifications thereof.

Finally, for testing purposes a simulation program is available, and two programs which set-up input files from externally simulated data:

RGCSIM: Delft GCR simulation program.
CONCD2: conversion of formatted CERGA datasets into binary input files.
CONLOS: conversion of formatted Lund datasets into binary input files.

A.4 Cpu times

The CPU times for the so-called CERGA dataset II, on the VAX 750 of the faculty of geodesy of the Technical University of Delft, are given in table A.1. The CDC computer, on which the software will be run in Toulouse, is approximately a factor 20 faster.

Table A.1 - CPU times for the great circle reduction for CERGA dataset II on a VAX 750 (in s.)

RGCEQ	observation equations	190
RGCGM1	synthetic ordering of star unknowns	1.3
	normal equations formation	331
RGCGM2	Choleski factorization star part	227
	Choleski factorization instrument part (including elimination star part)	115
	sparse covariance matrix star part	497
	covariance matrix instrument part and influence on the star variances	4
RGCGM3	solution star and instrument part	18
	attitude, residuals and testing	96
RGCSM	knot selection	0.1
	normal equations B-spline part	86
	elimination of stars from B-spline part	174
	reordering of the B-spline parameters (modulo 360)	1
	Choleski factorization & solution B-splines	403
	updating of the star abscissae	1.5
	sparse covariance matrix B-spline part	828
	variances star part	16

APPENDIX B

SIMULATED DATA FOR THE GREAT CIRCLE REDUCTION

The Great Circle Reduction software has been tested extensively on simulated data. Moreover, runs with simulated data often prove to be a valuable analysis tool. This thesis is in that matter no exception. Therefore, in this appendix the characteristics of the simulated data, used throughout this thesis, are discussed.

B.1 Simulation Possibilities

There are two approaches to simulate input data for the great circle reduction:
1) simulate data especially for the great circle reduction,
2) simulate raw measurements and process this data first by the phase extraction, star mapper and attitude reconstitution software, before treating it by the great circle reduction software.

Actually, in the second approach the complete chain of software up to the great circle reduction (GCR software) is tested, rather than the GCR software alone. At present several datasets have been simulated following the first approach, but no datasets have yet been simulated according to the second approach. In the near future a comparison of the GCR results with the other consortium is foreseen, both on simulated and real data [*Van der Marel, 1987*].

There are at least three programs which can simulate data especially for the Great Circle Reduction, called, according to their origin: Lund, CERGA and Delft GCR simulation software. The programs are similar in the sense that they produce grid coordinates directly from:
- a simulated star catalogue,
- a simulated attitude,
- a simulated instrumental perturbation.

The grid coordinates are computed from the theoretical values, *viz.* the expected or "true" values, and then perturbed with noise representing photon noise plus the statistics of the grid coordinate estimation procedure. Afterward the simulated "true" star catalogue, attitude and instrument are perturbed with noise representing in an ad hoc way the a priori knowledge of the various quantities respectively the statistical properties of the estimation processes involved. In particular, no photon counts need to be simulated and no grid coordinates are estimated, so the computation costs of the simulation remain within reasonable limits.

At present three large scale data sets (each covering one RGC) have already been produced by CERGA [*Falin et al., 1985, 1986, Van der Marel et al., 1986d*] and one by Lund observatory [*Lindegren, 1986, Van der Marel, 1987*]. Furthermore, several datasets have been simulated by the Delft software. Each of the datasets simulated by CERGA is based on three different input catalogues (see sec. B.4): one of the first treatment type, one typical for the final iteration and one without errors. In addition errorless grid coordinates have been simulated. The Lund dataset which was send to Delft is typical for first treatment, but also errorless data has been given. Results obtained from the second CERGA dataset (CDII) and Lund dataset (Lund) will be

frequently used in this thesis. Therefore we will outline some of their characteristics below.

B.2 Lund data

The input data has been simulated by L. Lindegren of the Lund observatory. Below we summarize some of the major characteristics, more details are given in [*Lindegren, 1986*]. The simulation covered five revolutions of the satellite, forming an RGC set of about average size. The dimensions are summarized in table B.1. The data contained several major gaps: 1504 frames are occulted by the Earth and an additional 105 are just empty. There are no eclipses. 59 jet firings occur during this RGC, the minimum and maximum jet interval are 410 s and 1406 s. Double stars or minor planets are not observed, so every observation to each star is well defined and there are no a-priori reasons to make stars passive. However, in some of the runs only bright stars (magnitude < 10) were used.

Table B.1 - Size of the CERGA and Lund datasets

	Lund	CERGA II
total number of stars	1964	2411
... brighter than magnitude 10	1697	1843
... fainter than magnitude 10	267	568
number of revolutions	5	5
number of frames (with observations)	16392	17653
... with bright stars	16308	17392
... with only faint stars	84	261
number of grid coordinates	78300	94664
... to bright stars	67811	75074
... to faint stars	10489	19590

The "true" attitude is obtained by integrating the equations of motion, using a diagonal tensor of inertia, and including a gyro induced torque and a -simple- model of the solar radiation torques (presumably 6 harmonics). To the torques a random perturbation is added in the form of a first order Markov process, with a standard deviation of $2\ 10^{-6}$ Nm on each axis and a time constant of 100 s. However, during the great circle reduction it was found out that these random perturbations are not realistic for attitude smoothing. The gas jets are actuated when the simulated attitude exceeds preset limits on the nominal motion. The on ground attitude reconstruction (AR) estimate, which is needed by the great circle reduction, is computed from the true attitude by first adding a first-order Markov process to each angle independently (this simulates star mapper estimation errors) and then smoothing the results by fitting a cubic spline to each angle. The Markov process was characterized by a standard deviation of 0.1 arcsec on each angle and a correlation time of 200 s. The maximum cubic spline interval was approximately 40 s, the cubic spline fit was interrupted at each gas jet actuation.

The input catalogue -containing the a priori, approximate, star positions- is created by perturbing true positions with uncorrelated Gaussian noise with a standard deviation of 1"5. True grid coordinates were computed from the true attitude and star positions, accounting for corrections due to aberration, etc. at the time of observation and -true- perturbations of the instrument (which may depend on the star colour).

Observational errors are simulated by adding Gausian errors with a variance depending on the mean intensity of the star, the observing time and some constants [*Lindegren, 1986*]. This variance must account for photon noise, imperfect knowledge of the medium scale instrumental effects and attitude jitter. The simulation (of the true observations) includes various instrumental effects, among which chromaticity and basic angle variations. The present data set did not include chromaticity effects, though the Lund program has the ability to do this as well.

B.3 CERGA dataset II

CERGA dataset II has been simulated by M. Froeschlé, J.L. Falin and F. Mignard of CERGA. The simulation covers about 10.7 hours of data (five revolutions of the satellite). The dimensions of the RGC-set are given in table B.1. The data contained no major gaps due to occultations (except an involuntary gap in the data for one of the tests because of a damaged data block on magnetic tape). There is one solar eclipse, which lasted about 135 frames for each of the two passage through the penumbra and 1576 frames for the passage through the Earth shadow. During the RGC-set 62 gas jet firings occurred. Some double stars and three minor planets were simulated, but those were marked as passive stars and are not used in the least squares solution.

The true attitude is obtained by integrating the equations of motion, using a diagonal tensor of inertia, and including a gyro induced, solar radiation and gravity gradient torque [*Pinard et al., 1983*], plus some intermittent torques due to particle shocks in the apogee boost motor. The attitude reconstitution (AR) estimate is computed by adding a uncorrelated Gaussian noise to the true attitude at mid frame time, with a standard deviation of 0"1 for iterations and 0"3 for first treatments.

The input catalogue in the RGC system is created by perturbing the true positions with uncorrelated Gaussian noise with a standard deviation of 0"1 for iterations, and 0"2 (mag. $\leq$ 8) and 0"8 (mag. $>$ 8) for first treatments. True grid coordinates are computed from the true attitude and true star positions, taking into account the correction for apparent places, large scale instrumental distortion, periodic basic angle variations, attitude jitter and chromaticity effects [*Falin et al., 1986*]. Observation errors are simulated by adding Gaussian errors with a variance depending on the mean intensity of the star and observing time.

B.4 Description of the testruns

Six different tests, following the terminology introduced in the FAST software test plan for the great circle reduction [*Van Daalen & Van der Marel, 1986b*], can be distinguished: Perfect Geometric (PG), Perfect A-priori values (PA), Perfect Observations (PO), Iteration (I), First treatment (F) and Perfect Smoothing (PS). The characteristics of these tests are summarized in table B.2.

Perfect Geometric (PG) refers to a test which uses errorless data. This test is first used to check the mutual consistency of the formulae in the reduction software and simulation software, and to see if conversions of the simulated data are done properly. But eventually this test can be used to check the ultimate (pure) linearization effects and rounding errors. *Perfect Smoothing* (PS) uses the same input data as PG, but PS aims at computing the additional modelling error due to smoothing, *i.e.* shortcomings in our model for the along scan attitude.

Perfect A-priori (PA), *Iteration* (I) and *First Treatment* (F) are tests which use noisy observations. The results of the first treatment test F are seriously affected by modelling errors, caused by the bad approximate values for the star ordinates and transverse attitude components, which are not estimated in the reduction. This modelling error is still present in the iteration runs, but can be neglected there, which should be confirmed by the test with perfect a-priori data (PA). The results of the test with perfect a-priori data should not be affected by these modelling errors and indeed show only the effect of noisy measurements. The modelling error can be assessed separately in a test using errorless (perfect) observations, but perturbed a-priori values (PO). The first treatment (F) is not only affected by modelling errors, which are typically in the order of mas, but also by grid step inconsistencies. Grid step inconsistencies give even larger errors, and they must be corrected in all cases.

Table B.2 - Description of the -ideal- testruns

ID	a-priori attitude	a-priori star pos.	observations	grid-steps	smoothing
PG	perfect	perfect	perfect	none	no
PS	perfect	perfect	perfect	none	yes
PA	perfect	perfect	realistic	none	no/yes
PO	~0"1	~0"01	perfect	none	no
I	~0"1	~0"01	realistic	few	yes
F	~0"3	>0"8	realistic	many	no

Usually the tests are done with all stars active. However in some cases the faint stars (mag. > 10) are made passive. The rank deficiency of the problem is, usually, solved by fixing one star, the base star, to its approximate value.

B.5 Analysis of the results

All simulations are similar in the sense that they compute grid coordinates from simulated "true" star, attitude and instrumental parameters. The observational errors, and errors in the approximate values, are added later. The adjusted star abscissae, along scan attitude and large scale instrumental deformation parameters, computed during the great circle reduction, can be compared with the simulated true -unperturbed- parameters. More specific, the *true error*, the difference between the adjusted and true value, should be compared with the formal (co-)variances computed during the great circle reduction. Two hypotheses should be checked:

a) the mathematical expectation of the true error is zero, *i.e.* there are no systematic errors,
b) the spread in the true error conforms with the variances.

In some cases, especially in first treatments, the first hypothesis is wrong. One important source of systematic errors are those caused by bad a-priori values for the star ordinates and transverse attitude components, which are not estimated. Fortunately, the effect of this error can be calculated by a few formulae (see chapter 5) or it can be assessed by special test-runs, so that its influence on our first hypothesis can be eliminated. The true errors

can of course not be computed from the real data. In this case only the least squares residuals of the grid coordinates can be analyzed, just in order to monitor the estimation process.

Let $\hat{\mathbf{x}}$ denote the vector with adjusted values computed during the great circle reduction and let $\mathbf{Q}$ be its covariance matrix. Then, $\Delta\mathbf{x}=\hat{\mathbf{x}}-E\{\mathbf{x}\}$ is the true error, which can be computed from the true simulated values $E\{\mathbf{x}\}$. The true error $\Delta\mathbf{x}$ should be unbiassed, *i.e.* there are no systematic errors, so $E\{\hat{\mathbf{x}}\}=E\{\mathbf{x}\}$, and the "spread" of the true error should fit the formal covariance matrix $\mathbf{Q}$. This hypothesis can be tested by the well known Chi-square test

$$H_0: \quad \Delta\mathbf{x}^T\mathbf{Q}_{\Delta\Delta}^{-1}\Delta\mathbf{x} \quad \leq \quad \chi_{1-\alpha;\,n-d}$$

with χ the critical value, $1-\alpha$ the level of significance of the test, $n-d$ the number of condition equations (degrees of freedom), n is the number of stars and d is the rank deficiency (d=1).

The covariance matrix $\mathbf{Q}$ is only computed fully for the instrumental parameters. For the abscissae only a small part of the covariance matrix, including the diagonal with variances, is computed. Therefore, the usual Chi-square test, which needs $\mathbf{Q}^{-1}$, cannot be computed. However, when the influence of the instrument on the abscissae is neglected, which is very small anyhow, the inverse of the covariance matrix $\mathbf{Q}^{-1}$ is simply the reduced normal matrix of the star-part after elimination of the attitude. Although this matrix is very sparse it is not (yet) used at the comparison stage. At the comparison stage just the diagonal elements are considered, this gives an approximate testing variate for the Chi-square test

$$\sum_{i=1}^{n} \frac{1}{(\mathbf{Q}_{\Delta\Delta})_{ii}}\,\Delta x_i^2$$

for which only the variances are needed. The power of the test is reduced considerably, and it makes no sense to hold on to the critical value χ of the original test. Therefore, instead of the above mentioned test more meaningful quantities are computed:

$$\eta := \sqrt{\frac{1}{n-d}\,\Delta\mathbf{x}^t\Delta\mathbf{x}} \quad \text{and} \quad \nu := \sqrt{\frac{1}{n-d}\sum_{i=1}^{n}(\mathbf{Q}_{\Delta\Delta})_{ii}}$$

It is sufficient to compare η with ν, which is usually done separately for each magnitude class. *I.e.* the mean true error per magnitude (colour) class is compared with the square root of the mean variance per magnitude (colour) class. An other useful tool are plots of the true error and square root of the variance as function of the abscissae.

The quantities $\Delta\mathbf{x}$ are computed in the minimum norm sense, *i.e.* the sum of the $\Delta\mathbf{x}_i$ is zero and the norm is minimal. Therefore the variances should be computed in the same way; *i.e.* the sum of the (minimum norm) variances is minimal. Usually variances pertaining to the *base-star* solution are computed, which are systematically larger than the minimum norm variances. But also the covariances, which are neglected in the test, are much larger than in the minimum norm solution (see chapter 7).

There are two other ambiguities in the abscissae, which should be solved before the comparison with true values: the abscissae are only defined modulo 2π and modulo the slit width.

APPENDIX C

COMPUTER SOLUTION OF LEAST SQUARES PROBLEMS

In this appendix the computer solution of large, sparse, symmetric, positive definite systems of equations, as arising in least squares problems, is discussed. First Choleski factorization, the only method which is considered here, is derived as a generalization from the solution of a 2 by 2 block partitioned system of equations. Secondly Choleski factorization of sparse systems of equations, and the occurrence of fill-in, are treated, and finally an efficient method for computing a partial inverse of a sparse matrix is presented.

C.1 Least Squares Estimation

Many estimation problems involve the estimation of unknown parameters $\mathbf{x}$ which bear a linear(ized) relationship to measurement data $\mathbf{y}$

$$\underset{m\times 1}{\mathbf{y}} \equiv \underset{m\times n}{\mathbf{A}} \cdot \underset{n\times 1}{\mathbf{x}} \qquad m > n \tag{C.1}$$

The data $\mathbf{y}$, obtained from a measurement process, is not perfect: it may contain random errors ($\mathbf{y}\neq E\{\mathbf{y}\}$), systematic errors ($E\{\mathbf{y}\}\neq\mathbf{Ax}$) and even large blunders. In general, for $m>n$, a solution to equation (C.1) does not exist at all. The space spanned by the columns of $\mathbf{A}$, the range $R(\mathbf{A})$, is a subspace of $\mathbb{R}^m$, and it would be a mere coincidence if the observation vector $\mathbf{y} \in \mathbb{R}^m$ lies in $R(\mathbf{A}) \subset \mathbb{R}^m$. The dimension of $R(\mathbf{A})$ is $n-d$, the rank of $\mathbf{A}$, with n the column dimension of $\mathbf{A}$ ($n<m$) and d the so-called *rank deficiency*, the dimension of the null space of $\mathbf{A}$.

Although a solution $\mathbf{x}$ to (C.1) may not exist at all it makes sense to choose an estimate $\hat{\mathbf{x}}$ for $\mathbf{x}$ for which $\mathbf{A}\hat{\mathbf{x}}$ is as close as possible to the measurement data $\mathbf{y}$, *i.e.*

$$\min_{\hat{\mathbf{x}}} \| \underline{\mathbf{y}} - \mathbf{A}\,\underline{\hat{\mathbf{x}}} \| \tag{C.2}$$

with $\| \ \|$ any sensible norm. The underscores in the formula are used to emphasize the stochastic nature of the variables $\underline{\hat{\mathbf{x}}}$ and $\underline{\mathbf{y}}$. Let $\mathbf{C}_{yy}=E\{(\underline{\mathbf{y}}-E\{\underline{\mathbf{y}}\})(\underline{\mathbf{y}}-E\{\underline{\mathbf{y}}\})^T\}$ be the covariance matrix of the observations $\mathbf{y}$. Then the well known weighted least squares solution is calculated from

$$\min_{\hat{\mathbf{x}}}\{\ (\underline{\mathbf{y}}-\mathbf{A}\underline{\hat{\mathbf{x}}})^T\mathbf{C}_{yy}^{-1}(\underline{\mathbf{y}}-\mathbf{A}\underline{\hat{\mathbf{x}}})\ \} \tag{C.3}$$

i.e. from minimizing the *residual sum of squares* $\underline{E} = \underline{\mathbf{v}}^T\mathbf{C}_{yy}^{-1}\underline{\mathbf{v}}$ in the metric induced by the observations, with $\underline{\mathbf{v}}$ the *least squares residuals* $\underline{\mathbf{v}} = \underline{\mathbf{y}} - \mathbf{A}\underline{\hat{\mathbf{x}}}$. The least squares solution $\underline{\hat{\mathbf{x}}}$ has the following -nice- properties:

- the least squares solution is *unbiased* when the model is unbiassed, *i.e.* $E\{\underline{\hat{\mathbf{x}}}\}=E\{\underline{\mathbf{x}}\}$,

- the least squares solution has *minimum variance* among all linear unbiased estimators,
- the least squares residuals $\underline{v} = \underline{y} - A\hat{\underline{x}}$ are *orthogonal* to $\hat{\underline{x}}$, so $Cov(\underline{v}, A\hat{\underline{x}})=0$.
- the least squares solution is the *maximum likelihood* solution in case the measurements **y** have a multi-normal distribution with covariance matrix C_{yy}.

Clearly, if there are no systematic errors, *i.e.* $E\{\mathbf{y}\}=\mathbf{Ax}$, then $E\{\underline{v}\}=\mathbf{0}$.

Differentiating E with respect to $\hat{\mathbf{x}}$ and putting $\partial E/\partial\hat{\mathbf{x}} = \mathbf{0}$, gives the so-called *normal equations*

$$\underset{n\times n}{(\mathbf{A}^T\mathbf{W}\mathbf{A})}\ \underset{n\times 1}{\hat{\underline{x}}} = \underset{n\times 1}{\mathbf{A}^T\mathbf{W}\underline{y}} \quad , \quad \mathbf{W} = C_{yy}^{-1} \tag{C.4}$$

with *normal matrix* $\mathbf{N}=\mathbf{A}^T\mathbf{W}\mathbf{A}$ and right hand sides $\underline{b}=\mathbf{A}^T\mathbf{W}\underline{y}$. **N** is a (semi) positive definite $n\times n$ matrix, with $Rank(\mathbf{N}) = Rank(\mathbf{A}) = n-d$ (if **W** is of full rank). If $Rank(\mathbf{N})<n$, *i.e.* **N** has a rank deficiency $d>0$ so **N** is singular, there is no unique solution $\hat{\mathbf{x}}$ of the least squares problem. In other words the columns of **A** are dependent, *i.e.* for $N(\mathbf{A})$, the null space of **A**, we have $N(\mathbf{A})\neq\{\mathbf{0}\}$. Let **G** be a basis for the null space $N(\mathbf{A})$, with $Dim(N(\mathbf{A}))=d$, $\mathbf{AG}=\mathbf{0}$ and so $\mathbf{NG}=\mathbf{0}$, then every vector

$$\underset{n\times 1}{\hat{\underline{x}}} = \underset{n\times 1}{\hat{\underline{x}}^I} + \underset{n\times d}{\mathbf{G}} \cdot \underset{d\times 1}{\mathbf{m}} \tag{C.5}$$

is a solution of (C.4), where $\hat{\mathbf{x}}^I$ is a particular solution and **m** a $d\times 1$ vector of -arbitrary- parameters.

Apparently more parameters than can be estimated from the observations have been introduced in the model. One remedy, therefore, is very simple: skip as many unknowns of **x** and columns of **A** as necessary (d). These unknowns do not get a correction, *i.e.* they are fixed on their a-priori values. This remedy is very simple and also numerically attractive, because the dimension of the normal equation system is reduced to $n-d$ (columns and rows corresponding to columns of **A** are skipped) and possible zeroes in the normal matrix are preserved. The reduced normal matrix is positive definite and can be factorized by Choleski's method and inverted. Let us assume that the weight matrix $\mathbf{W}=C_{yy}^{-1}$ is calculated from the covariance matrix C_{yy} of the observations, then the covariance matrix C_{xx} of the least squares solution is equal to the inverse $\mathbf{N}^{-1}$ of the normal matrix.

Instead of skipping columns of **A**, the linear(ized) system of equations $\mathbf{Ax}=\mathbf{y}$ can be extended by constraints

$$\underset{c\times n}{\mathbf{B}} \cdot \underset{n\times 1}{\hat{\underline{x}}} = \underset{c\times 1}{\underline{c}} \quad , \quad \underset{c\times c}{C_{cc}} = E\{(\underline{c}-E\{\underline{c}\})(\underline{c}-E\{\underline{c}\})^T\} \tag{C.6}$$

The resulting system $\begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} \mathbf{x} = \begin{bmatrix} \mathbf{y} \\ \mathbf{c} \end{bmatrix}$ is regular if and only if the $c\times d$ matrix **BG** has $Rank(\mathbf{BG})=d$. In case of the minimum number of constrains $c=d$ and $Rank(\mathbf{BG})=d$ the unique solution to the parameter vector **m** in (C.5) is $\mathbf{m}=(\mathbf{BG})^{-1}(\mathbf{c}-\mathbf{B}\hat{\mathbf{x}}^I)$, showing that the solution $\hat{\mathbf{x}}$ is uniquely characterized by **B** and **c**. Substitute the solution for **m** in (C.5), then the specific solution

$\hat{\underline{x}}_{B,c}$ is obtained from *any* solution $\hat{\underline{x}}$ by the so-called *S-transform* [*Baarda, 1973*]:

$$\hat{\underline{x}}_{B,c} = S_B \hat{\underline{x}} + G(BG)^{-1}\underline{c}$$

$$C_{B,c} = S_B C_{xx} S_B^T + G(BG)^{-1} C_{cc} (BG)^{-1} G^T \qquad (C.7)$$

with the so-called *S-matrix*

$$S_B \equiv I - G(BG)^{-1}B \qquad (C.8)$$

where the "S" stands for similarity or, more general, singularity [*Teunissen, 1985*]. The solution $\hat{\underline{x}}_{B,c}$ is uniquely determined by **B** and **c**, it does not depend on the prior solution $\hat{\underline{x}}$ or on the covariance matrix C_{cc} of **c**.

Two cases are of particular interest: 1) **c=0** and **B** = **(0,I,0)** with **I** the *dxd* identity matrix, which gives exactly the same solution as obtained by skipping the columns of **A** corresponding to **I**, and 2) **c=0** and **B=G**^{T}, the so-called *minimum norm* solution. Both cases are examples of *hard* constraints, *i.e.* with C_{cc}**=0**. The second case, which gives the minimum norm solution, has also some nice properties:
- **x** has minimum norm,
- C_{xx} has minimum trace.

However, since **G** is in general full, also S_B will be almost full. The minimum norm solution and its covariance matrix can be computed from the first solution by the S-transform (C.7), although for large and sparse systems of equations it is not advisable to do so for the covariance information.

Instead of computing the constrained solution by an S-transform also a more direct approach - which preserves the sparsity - can be used. The constrained least squares problem is formulated as

$$\min_{\hat{\underline{x}}} \{ \|\underline{y} - A\hat{\underline{x}}\|_2 \mid B\hat{\underline{x}} = \underline{c} \} \qquad (C.9)$$

These equations are solved by the method of Lagrange multipliers. The Lagrange function

$$h(\underline{m}, \underline{x}) = \| \underline{y} - A\underline{x} \|_2^2 + (B\underline{x} - \underline{c})^T \underline{m}$$

has a minimum if $\partial h/\partial(\underline{x})_i = 0$ and $\partial h/\partial(\underline{m})_j = 0$ for all possible i and j. Differentiation of the Lagrange function and putting the partial derivatives to zero, gives the linear system

$$\begin{bmatrix} A^T W A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \hat{\underline{x}} \\ \underline{m} \end{bmatrix} = \begin{bmatrix} A^T W \underline{y} \\ \underline{c} \end{bmatrix} \qquad (C.10)$$

The linear system (C.10) is *symmetric* and *regular*, assuming as before **BG** regular, but the system (C.10) is *not* positive definite. Therefore this system cannot be solved completely with Choleski factorization. However any *n-d* dimensional submatrix of $A^T WA$ is positive definite and can be factored by the Choleski method, the remaining *d+d* dimensional part can be inverted by Gauss' method.

C.2 Matrix Decompositions

C.2.1 LU Decomposition (Gauss)

Matrix decompositions are the basis of an effective class of solvers of linear systems of equations. The following theorem is fundamental

Theorem (C.1)
If **A** is an $m \times n$ matrix, with non-singular leading $k \times k$ submatrices for $k = 1,..,min(m-1,n)$, then there exist a unit lower triangular $m \times m$ matrix **L** (the diagonal consists of ones) and upper triangular $m \times n$ matrix **U** such that **A = LU**.

proof see *e.g.* [*Golub & Van Loan*, 1983], page 56. □

According to this theorem a linear system of equations **Ax=y** can be decomposed into the *lower* and *upper* triangular systems

$$\begin{aligned} \mathbf{L\,g} &= \mathbf{y} \\ \mathbf{U\,x} &= \mathbf{g} \end{aligned} \qquad \text{(C.11)}$$

which can be solved by simple forward and back substitution. The upper triangular matrix **U** is computed by successive Gauss transformations, the multipliers are stored in **L**. Multiple right hand sides may be solved by repeated forward and back substitution, without repeating the matrix decomposition.

The **LU** decomposition may not always exist or may be very unstable. More specific, (1) if a leading $k \times k$ sub-matrix of **A** is singular the **LU** decomposition fails, and (2) the Gaussian elimination can be unstable because of the possibility of arbitrary small *pivots*. Interchanging the rows and columns during elimination, such that as the next pivot the matrix element with the largest absolute value is chosen, will alleviate this problem. This is called *pivoting for stability*. In fact the **LU** decomposition of the permuted system **PAQ** is computed, where **P** and **Q** denote the row and column permutations.

C.2.2 LL^T Decomposition (Choleski)

In least squares problems linear system of equations of the type $(\mathbf{A}^T\mathbf{WA})\mathbf{x}=(\mathbf{A}^T\mathbf{W})\mathbf{y}$ have to be solved, the so-called *normal equations* **Nx=b** with *normal matrix* **N** $(\mathbf{A}^T\mathbf{WA})$ and *right hand sides* **b** $(\mathbf{A}^T\mathbf{Wy})$. The normal matrix $\mathbf{A}^T\mathbf{WA}$ is (1) *square*, (2) *symmetric* and (3) *semi positive definite*. For these matrices a special variant of the **LU** decomposition is considered.

If **A** is square and symmetric a symmetric decomposition $\mathbf{A} = \mathbf{LDL}^T$, with **D** $diag(d_1,..,d_n)$, is possible. If **A** is also positive definite all elements d_i are positive, which follows directly from the definition of positive definiteness. Let **A** be *positive definite*, *i.e.* $\mathbf{x}^T\mathbf{Ax} > 0$ for all non-zero vectors **x**; let $\mathbf{e}_i$ be the i'th unit vector and let $\mathbf{x}=\mathbf{L}^{-1}\mathbf{e}_i$. Then, $\mathbf{x}^T\mathbf{Ax}=d_i > 0$. Now let $\mathbf{L}'=\mathbf{L}\mathbf{D}^{1/2}$ and $\mathbf{U}'=\mathbf{D}^{-1/2}\mathbf{U}$, then $\mathbf{A} = \mathbf{L}'\mathbf{L}'^T$. This important result is formalized in the following theorem.

Theorem (C.2)
If **A** is an $n \times n$ symmetric positive definite matrix, it has a unique triangular factorization $\mathbf{LL}^T$, where **L** is a lower triangular matrix with positive diagonal entries.

proof: Any textbook, *e.g.* [*George & Liu, 1981, p. 15*] □

The triangular factor can be computed by *Choleski factorization.*

The linear system of equations $\mathbf{Ax=b}$ can be decomposed into the *lower* and *upper* triangular systems $\mathbf{L\,g = b}$ and $\mathbf{L^T x = g}$, which are solved by simple forward and back substitution. Multiple right hand sides are solved by repeated forward and back substitution, the factorization has to be computed only once.

The importance of Choleski factorization for our purposes lies in the fact that pivoting for stability is not needed.

C.2.3 Stability Considerations

A system of equations is *ill-conditioned* if one of its equations is "almost" linearly dependent on the others. In this case small changes in the data can give large variations in the solution. A measure for the condition is the so-called *condition number K*, defined by

$$K(\mathbf{A}) = \| \mathbf{A} \| \cdot \| \mathbf{A}^{-1} \| \qquad \text{(C.12)}$$

where $\| \ \|$ is any matrix norm. Let us define the *relative errors*

$$\varepsilon_{\mathbf{v}} = \frac{\|\mathbf{v}' - \mathbf{v}\|}{\|\mathbf{v}\|} \quad \text{and} \quad \varepsilon_{\mathbf{A}} = \frac{\|\mathbf{A}' - \mathbf{A}\|}{\|\mathbf{A}\|}$$

where the vector and matrix norms are compatible. Then the relative error in the solution of the system $\mathbf{Ax=b}$ is

$$\varepsilon_{\mathbf{x}} \leq K(\mathbf{A}) \, [\, \varepsilon_{\mathbf{A}} + \varepsilon_{\mathbf{b}} \,] \; + \; O(\varepsilon^2) \qquad \text{(C.13)}$$

This bound is the best possible , *i.e.* no algorithm can do better *[Golub & Van Loan, 1983]*. For instance, in case of ill-conditioned problems, truncation errors during the factorization can also result in large variations in the solution, or can even make the equations singular.

The *stability* of a method or an algorithm says something about its sensitivity for truncation errors. Consider the round off errors during Gaussian elimination, then the computed triangular factors $\mathbf{L}$ and $\mathbf{U}$ satisfy

$$\mathbf{LU} = \mathbf{A} + \mathbf{E} \qquad \text{(C.14)}$$

and $|\mathbf{E}|$ is of the order $(dim(\mathbf{A})\ (|\mathbf{A}|+|\mathbf{L}|\cdot|\mathbf{U}|)\ \varepsilon)$, where $|\mathbf{E}|$ stands for the matrix with absolute values $|e_{ij}|$ and ε denotes the truncation error of the number representation in the computer. With Gaussian elimination there is the possibility that the term with $|\mathbf{L}||\mathbf{U}|$ becomes large, because there is nothing to prevent small pivots and small pivots give large elements in $\mathbf{L}$ and $\mathbf{U}$. Obviously the stability can be improved by selecting in each step of the Gaussian elimination as pivot the element with largest absolute value. This amounts to solving the permuted system $\mathbf{PAQ\ z = Pb}$ with solution $\mathbf{x=Qz}$. The decomposition is now $\mathbf{PAQ=LU}$.

Fortunately it can been shown that it is safe not to pivot symmetric positive definite systems. The lower triangular factor $\mathbf{L}$ of the Choleski factorization $\mathbf{A=L_c L_c^T}$ is given by $\mathbf{L_c=L_G D^{1/2}}$, where $\mathbf{U_G=DL_G^T}$ with $\mathbf{D}$ a diagonal matrix with positive entries. The equality $\|\mathbf{L_c}\|_2^2=\|\mathbf{A}\|_2$ shows that the Choleski

factor must be nicely bounded, therefore the terms $|L_c||L_c^T|$ and $|E|$ are also nicely bounded. It will now be evident that 1) pivoting for stability is not necessary and 2) the stability depends mainly on the condition number K. If the condition number $K \sim 10^{+p}$ and rounding errors $\varepsilon \sim 10^{-q}$ then the relative error in the solution $\mathbf{x}$ is of the order 10^{p-q}.

C.3 Choleski Factorization

We take the following theorem - concerning a two by two partitioned system - as the basis for Choleski factorization. Algorithms for computing the factorization follow from recursive application of this theorem.

Theorem (C.3)
If $\mathbf{A}$ is an $n \times n$ symmetric positive definite matrix, with $\mathbf{A} = \begin{bmatrix} A_{11} & A_{21}^T \\ A_{21} & A_{22} \end{bmatrix}$, and if $\mathbf{L} = \begin{bmatrix} L_{11} & \\ L_{21} & L_{22} \end{bmatrix}$ is the lower triangular matrix from the factorization $\mathbf{A}=\mathbf{L}\mathbf{L}^T$, then $L_{21}^T = L_{11}^{-1}A_{21}^T$ and $L_{22}L_{22}^T = A_{22} - A_{21}A_{11}^{-1}A_{21}^T = A_{22} - L_{21}L_{21}^T$.

proof Straightforward multiplication $\mathbf{L}\mathbf{L}^T$ gives $A_{11} = L_{11}L_{11}^T$, $A_{21}^T = L_{11}L_{21}^T$ and $A_{22} = L_{22}L_{22}^T + L_{21}L_{21}^T$. Solving L_{21}^T by simple forward substitution from $A_{21}^T = L_{11}L_{21}^T$ gives $L_{21}^T = L_{11}^{-1}A_{21}^T$. Rewriting $A_{22} = L_{22}L_{22}^T + L_{21}L_{21}^T$ and substituting the result for L_{21} gives $L_{22}L_{22}^T = A_{22} - L_{21}L_{21}^T = A_{22} - A_{21}L_{11}^{-1}L_{11}^{-T}A_{21}^T = A_{22} - A_{21}A_{11}^{-1}A_{21}^T$. □

Corollary (C.4)
L_{21} can be computed from A_{21} by forward substitution with L_{11}.

Corollary (C.5)
The Choleski sub-factor L_{22} of the factor $\mathbf{L}$ is equal to the Choleski factor of the reduced matrix $\bar{A}_{22} = A_{22} - A_{21}A_{11}^{-1}A_{21}^T$.

The last corollary shows that any theorem, lemma, corollary or algorithm applicable to $\mathbf{A}$ is also applicable to $\bar{A}_{22} = A_{22} - A_{21}A_{11}^{-1}A_{21}^T$. This opens the way for recursive application of theorem (C.3). First consider a special case of theorem (C.3).

Lemma (C.6)
If $\mathbf{A}$ is an $n \times n$ symmetric positive definite matrix, with $\mathbf{A} = \begin{bmatrix} a & \mathbf{a}^T \\ \mathbf{a} & A_s \end{bmatrix}$, and if $\mathbf{L} = \begin{bmatrix} l & \\ \mathbf{l} & L_s \end{bmatrix}$ is the lower triangular matrix from the factorization $\mathbf{A}=\mathbf{L}\mathbf{L}^T$, then $l = \sqrt{a}$, $\mathbf{l} = \frac{1}{l}\mathbf{a}$ and $L_s L_s^T = A_s - \frac{\mathbf{a}\mathbf{a}^T}{a} = A_s - \mathbf{l}\mathbf{l}^T \equiv \bar{A}_s$.

proof Straightforward multiplication $\mathbf{L}\mathbf{L}^T$ gives $a = l^2$, $\mathbf{a} = l\,\mathbf{l}$ and $A_s = L_s L_s^T + \mathbf{l}\mathbf{l}^T$. So $l = \sqrt{a}$, $\mathbf{l} = \frac{1}{l}\mathbf{a}$ and $L_s L_s^T = A_s - \mathbf{l}\mathbf{l}^T = A_s - \frac{\mathbf{a}\mathbf{a}^T}{a} \equiv \bar{A}_s$ □

Recursive application of the previous lemma on the submatrices $\bar{A}_s$, for $s = 1,..n$, with $\bar{A}_1=A$, suggests an algorithm for Choleski factorization. This algorithm is called the *outer product method* after the outer products 11^T. The algorithm is given in diagram C.1.

The outer product method can be written as a sequence of transformations **T**. Let **A** be an $n \times n$ symmetric positive definite matrix with factorization $A=LL^T$. Define the block partitioned $n \times n$ matrices

$$A_i = \begin{bmatrix} I_{11} & 0 \\ 0 & \bar{A}_{22} \end{bmatrix} \text{ and } L_i = \begin{bmatrix} L_{11} & 0 \\ L_{21} & I_{22} \end{bmatrix},$$

where the dimension of the first block is ixi, and $\bar{A}_{22} = A_{22} - A_{21}L_{11}^{-1}L_{11}^{-T}A_{21}^T = A_{22} - L_{21}L_{21}^T$, so $A_0 \equiv A$, $A_n \equiv I$, $L_0 \equiv I$ and $L_n \equiv L$. Then

(1) $A = L_i A_i L_i^T$, $i=0,..,n$

Furthermore let $T_{i+1} = L_{i+1} - L_i + I$, then

(2) $L_{i+1} = L_i T_{i+1} = L_i + T_{i+1} - I$ (C.15)

(3) $A_i = T_{i+1} A_{i+1} T_{i+1}^T$

So the Choleski factor $L = \prod_{i=1}^{n} T_i$ or $L = \sum_{i=1}^{n} T_i - (n-1)\, I$.

A different algorithm for Choleski factorization is suggested by the following lemma:

Lemma (C.7)

If **A** is an $n \times n$ symmetric positive definite matrix, with

$$A = \begin{bmatrix} A_{11} & a_i^T & .. \\ a_i & a_{ii} & .. \\ : & : & . \end{bmatrix}, \text{ and if } L = \begin{bmatrix} L_{11} & 0 & .. \\ l_i & l_{ii} & .. \\ : & : & . \end{bmatrix} \text{ is the lower triangular}$$

matrix from the factorization $A=LL^T$, then the row l_i of L is $l_i^T = L_{11}^{-1}a_i^T$ and the diagonal element $l_{ii} = \sqrt{a_{ii} - l_i l_i^T}$.

proof The lemma follows from theorem (C.3) by considering (1) that the solution L_{21} to the triangular system $L_{21}L_{11}^T = A_{21}$ can be computed row by row and (2) that $l_{ii} = 1/\sqrt{\bar{a}_{ii}}$, where $\bar{a}_{ii}$ is the first entry in $\bar{A}_{22} = A_{22} - L_{21}L_{21}^T$ and $\bar{a}_{ii} = a_{ii} - l_i^T l_i$. □

The algorithm suggested by the previous lemma is called the *bordering method*, because the provisional factor is successively bordered with a row of **A**, from which the corresponding row of L is computed by forward substitution. The Choleski factor is computed row by row. By changing the order of the computations, so that the elements of the Choleski factor are computed column by column, a third algorithm is derived, the so-called *inner product method*.

In diagram C.1 all three possible algorithms for Choleski factorization are given. Note that the indices in the algorithms presented in diagram C.1 run faster the higher they are in the alphabet. With the outer product method the Choleski factor is computed column by column, but now using outer products rather than inner products and during the computation the other elements are modified. The access to the matrix elements during factorization is shown in diagram C.2 for full matrices and diagram C.3 for envelope matrices.

With the bordering and inner product method the elements of the Choleski factor are computed directly from inner products. The precision of the computations can be increased simply by increasing the precision of the inproduct computations without using extra storage. In case of the outer product method the Choleski factor elements are not computed directly: after a new column has been computed the outer product of this column with itself is added to the remaining elements. Therefore if the stability has to be increased not only the precision of the outer product computation must be improved, but also the wordsize of the modified part must be increased.

The cost of the factorization, back and forward substitution are evaluated in terms of *operation counts* or *flops* when *flo*ating *p*oint operations are considered. An operation is defined here as an addition, plus a multiplication, plus the necessary array accesses. The operation count for the Choleski factorization of a full $n \times n$ matrix is actually $\sim\frac{1}{6}n^3$ and the operation count for back and forward substitution is $\sim\frac{1}{2}n^2$. Therefore multiple right hand sides can be solved easily and cheaply, $O(n^2)$, by repeated forward and back substitution once the initial factorization has been obtained. The operation counts decrease dramatically if the matrices are sparse, *i.e.* if the matrices only contain relatively few non-zeroes.

C.4 Sparsity Considerations

C.4.1 Introduction

In most practical applications the matrix to be factored is sparse. Sparse matrices only contain few non-zeroes; the zero elements 1) have not to be stored and 2) multiplication by a zero results again in a zero. If the sparse symmetric positive definite matrix **A** has a Choleski factorization $\mathbf{LL}^T$, then the matrix **L** is usually sparse too. $\mathbf{L}+\mathbf{L}^T$ contains at least the same non-zero elements as **A**, but also newly created non-zeroes, called *fill-in*.

If **A** is an $n \times n$ sparse symmetric positive definite matrix and **L** the lower triangular matrix from the factorization $\mathbf{A}=\mathbf{LL}^T$, then it will be clear from the previous lemma's that the element l_{ij} of **L**, $i \geq j$, is non-zero if and only if

(1) a_{ij} is a non-zero, or

(2) the inproduct $1^T_{j(j-1)} \cdot 1_{i(j-1)}$ is non-zero

with $1_{j(i)}$ the sparse vector $(l_{j1}, \ldots, l_{ji})^T$. Let $Nz(\mathbf{A})$ be the set of non-zero elements of a matrix **A**

$$Nz(\mathbf{A}) = \{ (i,j) \mid a_{ij} \neq 0 \ , \ i \neq j \}$$

and define $Nz(\mathbf{L})$ similarly. Evidently $Nz(\mathbf{A}) \subset Nz(\mathbf{L}+\mathbf{L}^T)$, *i.e.* the *fill-in* is $Nz(\mathbf{L}+\mathbf{L}^T)-Nz(\mathbf{A})$. Now, using the previously established conditions, $Nz(\mathbf{L})$ is defined by the following lemma. The proof is trivial and will be omitted.

Lemma (C.8)

If **A** is an $n \times n$ sparse symmetric positive definite matrix and **L** the lower triangular matrix from the factorization $\mathbf{A}=\mathbf{L}\mathbf{L}^T$, then an element $(i,j) \in Nz(\mathbf{L})$ if and only if, for $j<i$,

(1) $(i,j) \in Nz(\mathbf{A})$, or

(2) $l_{ik} \in Nz(\mathbf{L}) \wedge l_{jk} \in Nz(\mathbf{L})$ for some $k < j$

The second condition in the previous lemma, which determines the fill-in, is not very practical, but it shows that the order of elimination is important. Consider the fill-in in the following two example matrices, which are identical except for a row and column permutation

$$\mathbf{A}_1 = \begin{bmatrix} * & & & & & \\ * & * & & & & \\ * & & * & & & \\ * & & & * & & \\ * & & & & & * \end{bmatrix} \longrightarrow \mathbf{L}_1 = \begin{bmatrix} * & & & & & \\ * & * & & & & \\ * & * & * & & & \\ * & * & * & * & & \\ * & * & * & * & * & \end{bmatrix} \quad \text{and} \quad \mathbf{A}_2 = \begin{bmatrix} * & & & & & * \\ & * & & & & * \\ & & * & & & * \\ & & & * & & * \\ & & & & * & * \\ * & * & * & * & * & * \end{bmatrix} \longrightarrow \mathbf{L}_2 = \begin{bmatrix} * & & & & & \\ & * & & & & \\ & & * & & & \\ & & & * & & \\ & & & & * & \\ * & * & * & * & * & * \end{bmatrix}$$

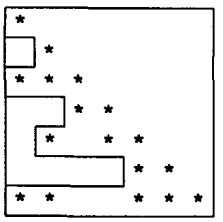
The non-zero elements are denoted by an *. The fill-in is respectively 6 and 0.

Corollary (C.9)

In the example we see clearly that $|Nz(\mathbf{L})|$ depends on the order in which the unknowns are numbered.

The order in which the unknowns are numbered does not affect the numerical stability as much as by Gaussian elimination. Therefore instead of $\mathbf{A}=\mathbf{L}\mathbf{L}^T$ we may compute $\mathbf{P}\mathbf{A}\mathbf{P}^T=\mathbf{L}\mathbf{L}^T$, where the permutation matrix **P** is chosen in such a way that the fill-in during factorization is small.

The *envelope* $Env(\mathbf{A})$ of a symmetric matrix **A** is formed by the matrix elements a_{ij}, $j \le i$, where $(i,j) \in Env(\mathbf{A})$ if and only if $\exists\, k \le j$ for which $a_{ik} \neq 0$. *I.e.* if $(i,k) \in Nz(\mathbf{A})$ for $k \le j \le i \Rightarrow (i,j) \in Env(\mathbf{A})$.



Lemma (C.10)

If **A** is an $n \times n$ sparse symmetric positive definite matrix and **L** the lower triangular matrix from the factorization $\mathbf{A}=\mathbf{L}\mathbf{L}^T$, then, neglecting numerical cancellations, $Env(\mathbf{L}) = Env(\mathbf{A})$.

proof Suppose $(i,l) \notin Nz(\mathbf{L})$ for $l=1,..,k-1$, then according to lemma (C.8) $(i,k) \in Nz(\mathbf{L})$ if and only if $(i,k) \in Nz(\mathbf{A})$. Induction on k, $k=1,\ldots j$, completes the proof □

Corollary (C.11)

If the non-zero elements of a sparse symmetric positive definite matrix **A**, with the Choleski factorization $\mathbf{A}=\mathbf{L}\mathbf{L}^T$, tend to cluster around the diagonal, this property will be maintained in the factor **L**.

Corollary (C.12)

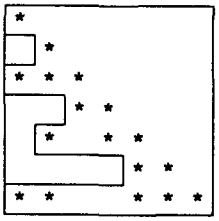
If $Nz(\mathbf{L})$ is the set of non-zeroes of the lower triangular matrix **L**, then $Nz(\mathbf{L}) \subseteq Env(\mathbf{L})$ and $|Nz(\mathbf{L})| \le |Env(\mathbf{L})|$.

$|Nz(\mathbf{L})|$ and $|Env(\mathbf{L})|$, and therefore storage requirements and computing resources, depend on the numbering of the unknowns. The ordering does not influence the numerical stability of the process as much as it does with the Gaussian elimination. Therefore we may order without regarding numerical stability, and we can order before the actual factorization takes place. The ordering procedures either
- reduce the fill $|Nz(\mathbf{L})|$, or,
- reduce the profile $|Env(\mathbf{L})|$.
The ordering procedures which aim at reducing $|Nz(\mathbf{L})|$ give usually less fill-in, but are also more complex than those which aim at reducing $|Env(\mathbf{L})|$. Besides, the advantage of a smaller $|Nz(\mathbf{L})|$ may be outweighted by the additional overhead associated to the more complex data structure for $Nz(\mathbf{L})$ than $Env(\mathbf{L})$.

Computer storage comprises both primary storage, for the element values, and overhead, needed to store some information about the row and column indices of the elements in the primary storage. Similar considerations hold for the execution time requirements; the overhead is formed by 1) the execution time needed to find an ordering, 2) storage allocation for $\mathbf{L}$ (symbolic factorization) and 3) overhead in access times. The amount of overhead depends on the storage structure chosen, and also on the number of elements to be stored.

C.4.2 Envelope Methods

Matrices for which the non-zero elements tend to cluster around the diagonal are called *envelope*, *profile* or *variable band* matrices. The fill-in during Choleski factorization is limited to non-zeroes within the envelope of the matrix, so it makes sense only to store the elements within the envelope of the matrix, without considering the non-zero structure within the envelope.



i	$\beta_i(\mathbf{A})$	$\omega_i(\mathbf{A})$
1	0	2
2	0	3
3	2	3
4	1	2
5	3	2
6	1	1
7	6	0
$\|Env(\mathbf{A})\|$	13	13

* non-zero

⌊ envelope

The *bandwidth* $\beta_i(\mathbf{A})$ in the i'th row of a symmetric matrix $\mathbf{A}$ is defined by

$$\beta_i(\mathbf{A}) = i - \min\{\, j \mid a_{ij} \neq 0,\ 1 \leq j \leq i \,\}$$

Then the *envelope* $Env(\mathbf{A})$ of a matrix $\mathbf{A}$ is the set of elements

$$Env(\mathbf{A}) = \{\, (i,j) \mid 0 < i-j \leq \beta_i(\mathbf{A}) \,\}$$

and the *profile* $|Env(\mathbf{A})|$, the number of elements in $Env(\mathbf{A})$, is

$$|Env(\mathbf{A})| = \sum_{i=1}^{n} \beta_i(\mathbf{A})$$

i.e. all the zeroes in the envelope are included.

The frontwidth[1] $\omega(\mathbf{A})$ of a matrix **A** is defined as

$$\omega_i(\mathbf{A}) = |\{ k \mid k > i \text{ and } a_{kl} \neq 0 \text{ for some } l \leq i \}|$$

the number of rows of the envelope of **A** which intersect column i. But, then

$$|Env(\mathbf{A})| = \Sigma \beta_i(\mathbf{A}) = \Sigma \omega_i(\mathbf{A}).$$

From lemma C.10 we know that the envelope of the normal matrix and Choleski factor, neglecting cancellations, are the same, but then also the frontwidth must be the same, *i.e* $Env(\mathbf{A})=Env(\mathbf{L})$ and $\omega_i(\mathbf{A})=\omega_i(\mathbf{L})$. Hence, $\omega_i(\mathbf{A})$ is the number of active rows at the i'th step in the factorization. Therefore the operation count for computing the factorization, which follows from the outer product formulation, is

$$\frac{1}{2}\sum_{i=1}^{n}(\omega_i(\mathbf{A})\ (\omega_i(\mathbf{A})+3)) \leq \frac{1}{2}\sum_{i=1}^{n}(\beta_i(\mathbf{A})\ (\beta_i(\mathbf{A})+3)) \qquad (C.16)$$

If **A** has the monotone profile property (see below) the operation count is equal to the upper bound computed from the bandwidth β. The operation count for solving the system of equations is

$$2\sum_{i=1}^{n}(\omega_i(\mathbf{A})+1) = 2\sum_{i=1}^{n}(\beta_i(\mathbf{A})+1) \qquad (C.17)$$

Let b be the mean width of the band, then the operation count for the factorization is $O(b^2n)$ and the operation count for solving is $O(bn)$.

Let $f_i(\mathbf{A})$ be $i-\beta_i(\mathbf{A})$, the column index of the first non-zero in the i'th row, then the *monotone profile property* is defined by $f_i(\mathbf{A})\leq f_j(\mathbf{A})$ for $i \leq j$.

Lemma (C.13)
Let **A** be a symmetric positive definite matrix with the monotone profile property $f_i(\mathbf{A})\leq f_j(\mathbf{A})$ for $i\leq j$, and factorization $\mathbf{A}=\mathbf{L}\mathbf{L}^T$. Then **L** has a full envelope $Env(\mathbf{L}+\mathbf{L}^T)$, *i.e.* $Env(\mathbf{L}+\mathbf{L}^T)=Nz(\mathbf{L}+\mathbf{L}^T)$.

proof Assume this assertion is true for the rows $1\ldots i-1$. From lemma (C.8) follows that $l_{ij}\neq 0$ if and only if $a_{ij}\neq 0$ or $l_{ik}\neq 0 \wedge l_{jk}\neq 0$ for some $k=f_i(\mathbf{A}),..,j-1$. Suppose for $j>f_i(\mathbf{A})$ $a_{ij}=0$ and $l_{ik}\neq 0$ for $k=f_i(\mathbf{A}),..,j-1$, then $l_{ij}\neq 0$ if and only if $l_{jk}\neq 0$, so from the previous assumption follows $k\geq f_j(\mathbf{A})$ and therefore $f_j(\mathbf{A})\leq f_i(\mathbf{A})$ (the monotone profile property). For $j=f_i(\mathbf{A})$ $a_{ij}\neq 0$ and so $l_{ij}\neq 0$, then induction on $j=f_i(\mathbf{A})+1,..,i-1$ and $i=1,..,n$ completes the proof. □

---

[1] The frontwidth of a matrix must not be confused with the *forward* bandwidth $\nu_i(\mathbf{A})$ of a matrix, which is defined as

$$\nu_i(\mathbf{A}) = \max\{ i \mid a_{ij} \neq 0,\ j \leq i \leq N \} - i$$

The *forward* envelope $Env(\mathbf{A})^*$ and profile $|Env(\mathbf{A})^*|$ are defined similarly as for the bandwidth $\beta(\mathbf{A})$, but in gereral $Env(\mathbf{A})^*\neq Env(\mathbf{A})$ and $Env(\mathbf{A})\neq Env(\mathbf{L})$.

A data structure for the envelope *Env*(**A**) is given in figure C.1. Because *Env*(**A**)=*Env*(**L**) Choleski factorization can be implemented inplace, *i.e.* **L** may overwrite **A**.

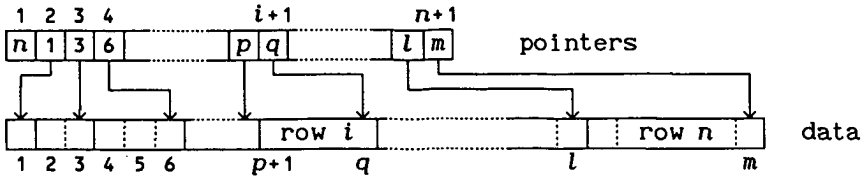


Figure C.1 - Variable band (profile) data structure for *Env*(**A**)

Note that for symmetric matrices the profile storage structure in figure C.1 applies both to the "row" envelope of the lower triangle and "column" envelope of the upper triangle. However, the row and column envelope of the lower triangle are in general not the same.

The data structure from figure C.1 is well suited for the implementation of the bordering method, which computes the Choleski factor row by row (see diagram C.1). However other methods can be implemented equally well. The choice for a specific method therefore depends on the preferred order of computations and some machine considerations. In some applications the reduced sub-matrix $\bar{\mathbf{A}}_{22}$ is needed, which is produced directly by the outer product method. For very large systems of equations the complete factor cannot always be kept in-core for most paging machines, and depending on the shape of the envelope one method may be preferred to others. For instance the factor in diagram C.3, which arises in many problems, can best be handled by the outer product method: throughout the computations only $\frac{3}{2}b^2$ storage locations are sufficient and all elements have to be read once from disc. On the other hand for vector machines with a large memory, the bordering or innerproduct method may be preferred.

C.4.3 Sifted Format Methods

A sparse symmetric matrix can be associated with a graph. A graph $G=(X,E)$ consists of a finite set X of *nodes* x_i together with a set E of unordered pairs of nodes $\{x_i,x_j\}$, called *edges*. Nodes are associated with unknowns and edges with a pair of unknowns connected through observations. A graph is *ordered* if the nodes are numbered. An ordered graph can be related to the normal matrix **N**: the nodes can be associated with the diagonal entries in **N**, edges can be associated with the off-diagonal non-zero elements in **N**.

Two nodes x and y are *adjacent* if the pair $\{x,y\}$ is an element of E, the set of edges. The *adjacent set* of y, denoted by *Adj*(y), is the set of nodes adjacent to y, *i.e.*

$$Adj(y) = \{\ x \mid \{x,y\} \in E\ \}$$

Let Y be a subset of nodes, then the *adjacent set* of Y is the set of nodes which are not in Y but are adjacent to at least one node in Y, *i.e.*

$$Adj(Y) = \bigcup_{y \in Y} Adj(y) \ / \ Y$$

We define the *degree* *Deg*(Y) of Y as the number of edges with nodes outside Y, *i.e.* $Deg(Y)=|Adj(Y)|$.

A *section graph* contains a subset of the nodes plus all edges which are pairs from the subset of nodes, *i.e.* the section graph $G(Y)$ is the subgraph $G(Y,E(Y))$ of $G(X,E)$, with $Y \subset X$ and $E(Y) \subset E$, $E(Y)=\{ \{x,y\} \in E \mid x \in Y, y \in Y \}$.

The following lemma is identical to our lemma (C.8), but now formulated in graph theory:

lemma (C.8')

The unordered pair $\{x_i,x_j\} \in E^{L+L^T}$ if and only if $\{x_i,x_j\} \in E^A$ or $\{x_i,x_k\} \in E^{L+L^T}$ and $\{x_k,x_j\} \in E^{L+L^T}$ for some $k < \min(i,j)$.

The complete process of Choleski factorization can be interpreted as a sequence of graph transformations, which is very helpful when considering the occurrence of fill-in. The sequence of graph transformations is depicted by

$$G_0 \rightarrow G_1 \rightarrow G_2 \ldots\ldots\ldots \rightarrow G_n$$

with $G_0=G(X,E^A)$ and G_n the empty graph. The so-called *elimination graph* $G_i=(X_i,E_i)$ is the graph of the reduced matrix $\bar{A}_{22}$ after i elimination step in the outer product formulation of the Choleski factorization. Elimination of a node x involves
- deleting the node x and all its incident edges, and
- adding edges so that nodes in $Adj(x)$ are pairwise adjacent in the new graph.

Fill-in occurs if *new* edges have to be added to the new graph. Choleski factorization can now be interpreted as a sequence of such elimination steps, and the fill-in **L** corresponds to the set of new edges added during the elimination process.

The previous lemma (C.8') is not very satisfactory because it gives the fill-in in terms of the matrix **L** and not the original matrix **A**. A different approach is provided by the concept of reachable sets [*George & Liu, 1981*]. A path from a node x to a node y of length $l \geq 1$ is an ordered set of distinct nodes $(v_1,v_2,\ldots,v_{l+1})$ such that $v_{i+1} \in Adj(v_i)$, $i=1,2,..,l$ with $v_1=x$ and $v_{l+1}=y$. In other words y is reachable from x through the set of nodes $\{v_2,\ldots,v_l\}$. The reach of a node y through S is defined by

$$Reach(y,S) = \{ x \notin S \mid x \text{ is reachable from y through } S \}$$

A graph $G=(V,E)$ is connected if every pair of distinct nodes is joined by at least one path, *i.e.* if

$$Reach(y,V) = V \quad \text{for} \quad \forall y \in V$$

Otherwise G consists of one or more connected components.

The next two theorems give the fill-in of the Choleski factor **L** in terms of the original matrix **A**, using reachable sets. For the proof we refer to [*George & Liu, 1981*].

Theorem (C.14)

$$E^{L+L^T} = \{ \{x_i,x_j\} \mid x_j \in Reach(x_i,\{x_1,\ldots,x_{i-1}\}) \}$$

Theorem (C.15)

Let $y \in G_i=(X_i,E_i)$, with G_i the i'th elimination graph, then $Adj(y)=Reach(y,\{x_1,\ldots,x_{i-1}\})$.

Let $\mu_i(L)$ be the number of non-zeroes in the i'th column of L, and $\eta_i(L)$ the number of non-zeroes in the i'th row of L. Then the operation count for the factorization is

$$\frac{1}{2}\sum_{i=1}^{N}\left(\mu_i(L)\ (\mu_i(L)+3)\right) \le \frac{1}{2}\sum_{i=1}^{N}\left(\eta_i(L)\ (\eta_i(L)+3)\right) \qquad (C.18)$$

In terms of the original graph $\mu_i(L) = |Reach(x_i,\{x_1,\ldots,x_{i-1}\})|$.

A data structure for $Nz(A)$ is given in figure C.2 [*Kok, 1984*]. The column indices for each row are usually given in increasing order. In order to access an arbitrary element, or columns, some testing is necessary. When a complete row has to be accessed testing is not needed which is good for the performance of any algorithm. This makes that the choice for a rowwise or columnwise (the transpose) storage structure is important for the performance of *e.g.* Choleski factorization.

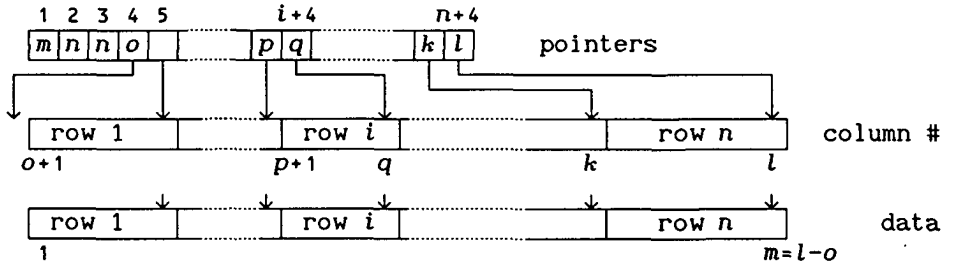


Figure C.2 - Sifted format data structure for $Nz(A)$

It is useful to distinguish the *symbolic* factorization, during which the non-zero structure $Nz(L)$ is computed, and the actual factorization. First consider the actual factorization; at first sight neither the rowwise or the columnwise storage structure seems to have profound advantages and some testing cannot be prevented. However with a simple trick testing may be prevented for the inner product method if the columnwise administration is used [*George & Liu, 1981*]. The trick is to compute the inner products as outer products, and to add these to the current column i. Only columns which have a non-zero in row i, which may follow from a temporary list, play a role. Symbolic factorization can be done with the same principles: the non-zeroes in column i are found by merging the non-zeroes in the columns which have a non-zero in row i. A different implementation of the symbolic factorization algorithm, implemented in Delft, uses the rowwise storage structure.

It is good to introduce now a clear concept of the *transpose* of a matrix: we must distinguish between the mathematical interpretation and the consequences for the data storage. *I.e.* elements of the transpose matrix are stored in the same storage locations as the original if the storage structure is transposed too. Only storage structure transposal, or only mathematical transposal have a profound influence on the way the elements are stored, and may even necessitate a different storage structure, *e.g.* $|Env(A^T)| \neq |Env(A)|$.

C.4.4 Partitioned Systems

Theorem (C.3) of section C.3 formed the basis for Choleski factorization. However, it may be worthwhile -especially for sparse matrices- to consider partitioned systems on their own. Corollary (C.4) and (C.5) suggest an algorithm for the *block factorization* of the partitioned matrix

$$A = \begin{pmatrix} A_{11} & A_{21}^T \\ A_{21} & A_{22} \end{pmatrix} \text{ into the factor } L = \begin{pmatrix} L_{11} & \\ L_{21} & L_{22} \end{pmatrix}.$$

Algorithm *Symmetric block factorization*

1) factor A_{11} into $L_{11}L_{11}^T$
2) solve L_{21} from the triangular systems $L_{11}L_{21}^T = A_{21}^T$
3) modify A_{22} into $\bar{A}_{22} = A_{22} - L_{21}L_{21}^T$
4) factor $\bar{A}_{22}$ into $L_{22}L_{22}^T$

The third step may turn out to be a nuisance, especially if L_{21} is full. Expanding the product $L_{21}L_{21}^T$ gives $(A_{21}L_{11}^{-T})(L_{11}^{-1}A_{21}^T)$, which may be computed also as $A_{21}(L_{11}^{-T}(L_{11}^{-1}A_{21}^T)) = A_{21}(L_{11}^{-T}L_{21}^T) = A_{21}\bar{L}_{21}^T$. This product can be evaluated column by column, which gives the following algorithm.

Algorithm *A-symmetric block factorization*

1) factor A_{11} into $L_{11}L_{11}^T$
2) for each column a_i of A_{21}^T
   2.1) solve the lower triangular system $L_{11}\bar{l}_i = a_i$
   2.2) solve the upper triangular system $L_{11}^T\bar{l}_i = \bar{l}_i$
   2.3) modify the i'th column of A_{22}: $(\bar{A}_{22})_{*i} = (A_{22})_{*i} - A_{21}\bar{l}_i$
3) factor $\bar{A}_{22}$ into $L_{22}L_{22}^T$

The a-symmetric block factorization evidently does not need L_{21} explicitly, instead A_{21} is used. Since $|Nz(A_{21})| \leq |Nz(L_{21})|$ the second algorithm is more efficient, only when A_{21} is full the performance is the same.

The *solution* of a two by two block partitioned system can be computed also with and without L_{21}. The so-called *standard solution scheme* uses L_{21}, the *implicit solution scheme* uses A_{21}.

Algorithm *Standard solution scheme*

1) *Forward solve*

1.1) solve $L_{11}y_1=b_1$

1.2) compute $\bar{b}_2=b_2-L_{21}y_1$

1.3) solve $L_{22}y_2=\bar{b}_2$

2) *Backward solve*

2.1) solve $L_{22}^T x_2=y_2$

2.2) compute $\bar{y}_1=y_1-L_{21}^T x_2$

2.3) solve $L_{11}x_1=\bar{y}_1$

In the standard solution scheme L_{11}, L_{21} and L_{22} are needed. In the implicit solution scheme A_{21} is used instead of L_{21}.

Algorithm *Implicit solution scheme*

1) *Forward solve*

1.1) solve $L_{11}y_1=b$ and $L_{11}^T x'_1=y_1$ (x'_1=temporary solution)

1.2) compute $\bar{b}_2=b_2-A_{21}x'_1$

1.3) solve $L_{22}y_2=\bar{b}_2$

2) *Backward solve*

2.1) solve $L_{22}^T x_2=y_2$

2.2) compute $\bar{b}_1=b_1-A_{21}^T x_2$

2.3) solve $L_{11}y_1=\bar{b}_1$ and $L_{11}^T x_1=y_1$

Alternative formulations for step 2.2) and 2.3) of the implicit solution scheme are:

2.2) compute $\bar{y}_1=y_1-t_1$, with t_1 from $A_{11}t_1=A_{21}^T x_2$

2.3) solve $L_{11}^T x_1=\bar{y}_1$

The cost of the implicit solution scheme is no greater than the cost of the standard solution scheme if $|Nz(A_{21})| + |Nz(L_{11})| \leq |Nz(L_{21})|$.

C.5 Computing the -Partial- Inverse

An example of a system with multiple right hand sides is $NC=I$, where $C=N^{-1}$ is to be computed by repeated forward and back substitution with columns e_i of I. This is a very straightforward way of computing the inverse of a matrix. The operation count for the inversion of a full matrix, not counting the initial factorization, is $\sim \frac{1}{3}n^3$. This is twice as much as for the factorization itself.

For sparse matrices repeated forward and back substitution is not a very elegant way for computing the inverse of a matrix: Take for instance an envelope matrix, and let b be the mean width of the band. Then the operation count for the Choleski factorization is of the order b^2n, and the operation

count for one solution of the order bn. For the inversion n systems have to be solved, therefore, the operation count for the inverse is $O(bn^2)$. It would be ideal if the operation count for the inversion is of the same order as the operation count for the factorization. With the present method this is only the case if only a *few* elements of the inverse have to be computed. However, as we will show below, it is possible with a different technique to compute a large number of elements of the inverse, corresponding to the non-zeroes in the Choleski factor, in $\sim \frac{1}{3}b^2n$ time. The inverse which is computed by this technique is called the *sparse inverse*. For most applications the sparse inverse is sufficient, and other elements are not needed. This technique, the so-called *recursive partitioning* or *sparse inversion*, can be based on the following lemma.

Lemma (C.16)
Let $\mathbf{A}$ be an $n \times n$ symmetric positive definite matrix, with inverse $A^{-1} = B = \begin{pmatrix} \cdot & & \\ \cdot & \cdot & \\ \cdot & B_{22} & B_{32}^T \\ \cdot & B_{32} & B_{33} \end{pmatrix}$, and let $L = \begin{pmatrix} \cdot & & \\ \cdot & \cdot & \\ \cdot & L_{22} & \\ \cdot & L_{32} & L_{33} \end{pmatrix}$ be the lower triangular matrix from the factorization $A=LL^T$, then $B_{32} = -B_{33}L_{32}L_{22}^{-1}$ and $B_{22} = L_{22}^{-T}(I_{22} + L_{32}^T B_{33} L_{32})L_{22}^{-1} = L_{22}^{-T}L_{22}^{-1} - L_{22}^{-T}L_{32}^T B_{32}$.

proof If $B=A^{-1}$, then $AB=I$. Let $A=LL^T$, multiply $AB=I$ left with L^{-1} and right with L, gives $L^TBL=I$. Straightforward multiplication L^TBL gives $L_{22}^T B_{22} L_{22} + L_{32}^T B_{32} L_{22} + L_{22}^T B_{32}^T L_{32} + L_{32}^T B_{33} L_{32} = I_{22}$ (1), $L_{33}^T B_{32} L_{22} + L_{33}^T B_{33} L_{32} = 0$ (2) and $L_{33}^T B_{33} L_{33} = I_{33}$ (3). Rewriting (2) gives $B_{32}L_{22} = -B_{33}L_{32}$, using this in (1) gives $L_{22}^T B_{22} L_{22} = I_{22} + L_{32}^T B_{33} L_{32}$ or $L_{22}^T B_{22} L_{22} = I_{22} - L_{32}^T B_{32} L_{22}$. Solving for B_{32} and B_{22} completes the proof. □

Recursive application of this lemma on partial columns of the inverse gives an operational scheme for computing the inverse of a matrix. The algorithm is presented in the form of a lemma.

Lemma (C.17)
Let $\mathbf{A}$ be an $n \times n$ symmetric positive definite matrix, with inverse $A^{-1} = B = \begin{pmatrix} \cdot & & \\ \cdot & \cdot & \\ \cdot & b_{ii} & b_i^T \\ \cdot & b_i & \bar{B}_{i+1} \end{pmatrix}$, and let $L = \begin{pmatrix} \cdot & & \\ \cdot & \cdot & \\ \cdot & l_{ii} & \\ \cdot & l_i & \bar{L}_{i+1} \end{pmatrix}$ be the lower triangular matrix from the factorization $A=LL^T$, then the inverse matrix $B=\bar{B}_1$, with $\bar{B}_i = \begin{pmatrix} b_{ii} & b_i^T \\ b_i & \bar{B}_{i+1} \end{pmatrix}$, $i = n-1, \ldots, 1$ and $\bar{B}_n = 1/l_{nn}$, where $b_i = -\bar{B}_{i+1} l_i / l_{ii}$ and $b_{ii} = (1/l_{ii} - l_i^T b_i)/l_{ii}$.

proof This lemma follows directly from recursive application of the previous lemma for partial columns of B. □

Let us consider the computation of a single element $(\mathbf{b}_i)_j$ in the i'th column of the inverse, *i.e.* $(\mathbf{b}_i)_j = \mathbf{b}_j^T \mathbf{l}_i / l_{ii}$: Assume that $\mathbf{l}_i$ is a sparse vector, then only those elements of $\mathbf{b}_j$ (the j'th column of the inverse) are needed which correspond to a non-zero in $\mathbf{l}_i$. It turns out that, when $(\mathbf{l}_i)_j$ itself is a non-zero, the required elements in $\mathbf{b}_j$ correspond also to non-zeroes in the Choleski factor. This becomes plausible when we consider that the fill-in, created in the i'th elemination step of Choleski factorization, is given by $Nz(\mathbf{l}_i \mathbf{l}_i^T)$. Hence, it is possible to compute only the elements $(\mathbf{b}_i)_j$ of the inverse which correspond to non-zeroes in the Choleski factor, *i.e.* $(i,j) \in Nz(\mathbf{L})$. The resulting -partial- inverse is called the *sparse inverse*.

DIAGRAM C.1

ALGORITHMS FOR CHOLESKI FACTORIZATION

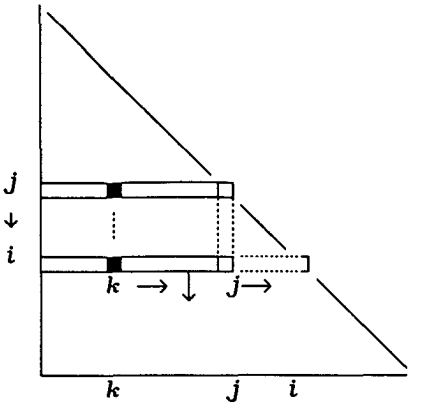
Bordering method:

for $i=1,2,\ldots.N$

for $j=1,2,\ldots\ldots,i-1$

$$l_{ij} = \left(a_{ij} - \sum_{k=1}^{j-1} l_{jk} \cdot l_{ik} \right) / l_{jj}$$

$$l_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2}$$



Outer product method:

$\mathbf{M} \leftarrow \mathbf{A}$

for $i=1,2,\ldots.N$

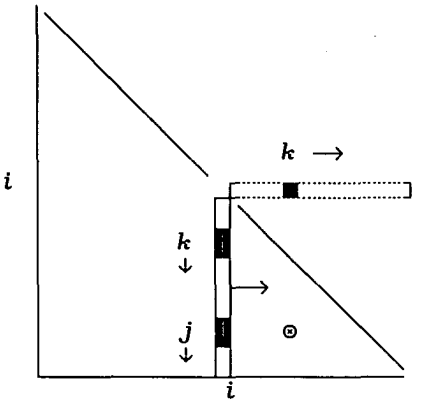
$$l_{ii} = \sqrt{m_{ii}}$$

for $j=i+1, i+2,\ldots.,N$

$$l_{ji} = m_{ji}/l_{ii}$$

for $k=i+1, i+2,\ldots.,j$

$$m_{jk} = m_{jk} - l_{ji} \cdot l_{ki}$$



Inner product method:

for $i=1,2,\ldots.N$

$$l_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2}$$

for $j=i+1, i+2,\ldots.,N$

$$l_{ji} = \left(a_{ji} - \sum_{k=1}^{i-1} l_{jk} \cdot l_{ik} \right) / l_{ii}$$

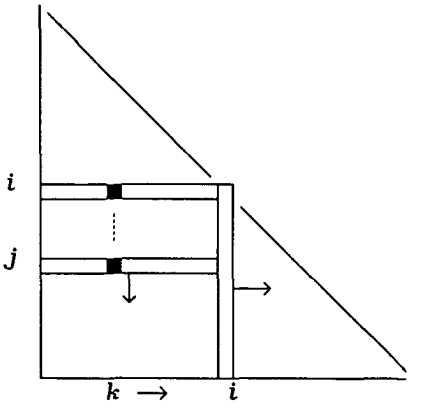
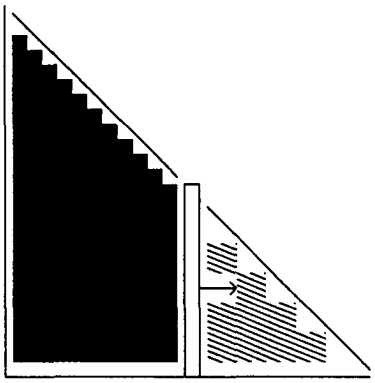


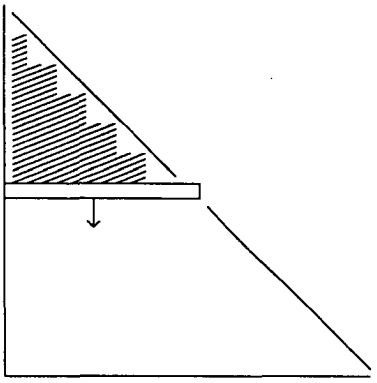
DIAGRAM C.2

ACCESS TO MATRIX ELEMENTS DURING CHOLESKI FACTORIZATION FOR A FULL MATRIX

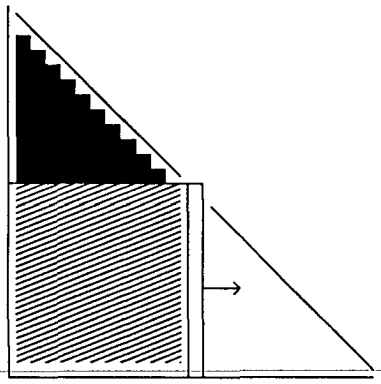
Outer product method:



Bordering method:



Inner product method:



Legenda:

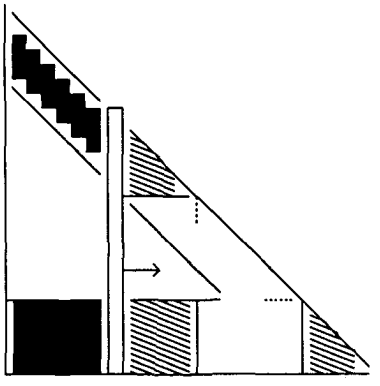
not yet accessed	}	$A = L\,L^t$
modified, currently accessed		
completed, currently accessed	}	L
completed, not currently accessed		
completed, no longer accessed		

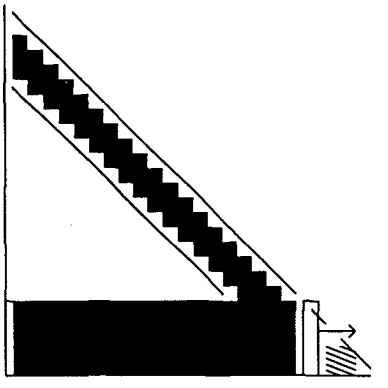
DIAGRAM C.3

ACCESS TO MATRIX ELEMENTS DURING CHOLESKI FACTORIZATION

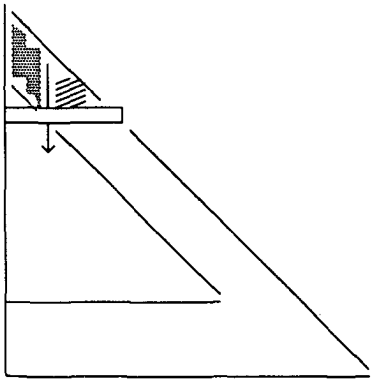
FOR A PROFILE MATRIX (CHIMNEY SHAPED)

Outer product method:





Bordering method:



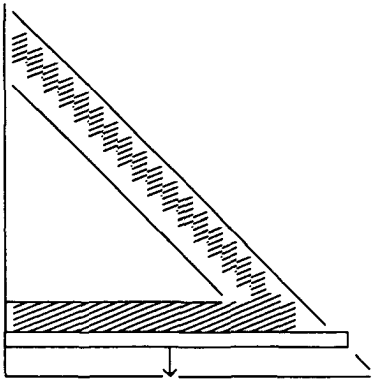
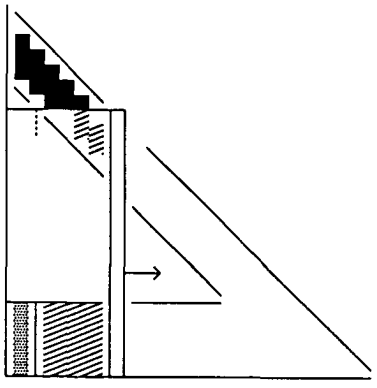
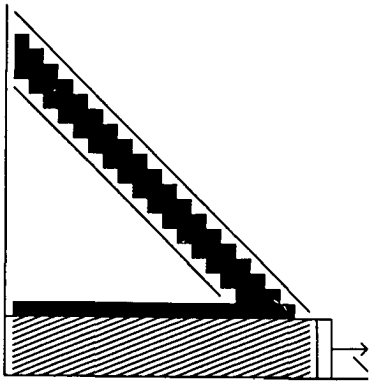


DIAGRAM C.3 - CONTINUED

Inner product method:





Legenda:

	not yet accessed	
	modified, currently accessed	$A = L\ L^t$
	completed, currently accessed	
	completed, not currently accessed	L
	completed, no longer accessed	

REFERENCES

Amoureus, L. [1984], Hipparcos: Een quaternionen formulering van reductie op cirkels (in dutch), Case study, Fac. Geodesie, TU Delft, 1984.

Baarda, W. [1968], A testing Procedure for use in Geodetic Networks, Neth. Geod. Comm., Publ. on Geodesy, Vol. 2, No. 5, 1968.

Baarda, W. [1973], S-transformations and Criterion matrices, Neth. Geod. Comm., Publ. on Geodesy, Vol. 5, No. 1, 1973

Badiali, M., M. Amoretti, D. Cardini and A. Emanuele [1986], Tests on Hipparcos Real System. In: P.L. Bernacca (ed.), Proc. third FAST thinkshop, Bari, November 1986, pp. 304-314.

Barbieri, C. and P.L. Bernacca [1979], Colloquium on European Satellite Astrometry, held in Padova, Italy, June 5-7, 1978.

Bastian, U. [1985], Treatment of the grid step ambiguity problem within task 6000. In: J. Kovalevsky (ed.), Proc. second FAST Thinkshop, Marseille, February 1985, pp. 295-298

Belforte, P., E. Canuto, F. Donati and A. Villa [1983a], An approach to on-ground attitude reconstitution. In: P.L. Bernacca (ed.), Proc. first FAST Thinkshop, Asiago, May 1983, pp. 163-174

Belforte, P., E. Canuto, D. Carlucci and B. Fassino [1983b], The jitter error in grid coordinate estimation. In: P.L. Bernacca (ed.), Proc. first FAST Thinkshop, Asiago, May 1983, pp. 197-208

Belforte, P., H. van der Marel and E. Canuto [1986a], On Two Different Ways of Modelling Hipparcos Attitude. In: P.L. Bernacca (ed.), Proc. third FAST thinkshop, Bari, November 1986, pp. 81-96.

Belforte, P. [1986b], Report on the task Attitude Reconstitution. In: P.l. Bernacca (ed.), Proc. third FAST thinkshop, Bari, November 1986, pp. 41-49.

Benciolini, B., L. Mussio and F. Sansò [1981a], The ICCG method for network adjustment. In: Proc. Smp. Geod. networks and comp., Munich, 1981, DGK, Reihe B, nr. 258/viii.

Benciolini, B., L. Mussio [1981b], Test on a reordering algorithm for geodetic networks and photogrammetric block adjustment. In: C.C. Tscherning (ed.), Proc. of the International Symposium on Management of Geodetic Data, Copenhagen, August, 1981.

Bernacca, P.L. (ed.) [1983], Processing of scientific data from the E.S.A. Astrometry satellite HIPPARCOS, proc. of the first FAST Thinkshop, held in Asiago, Italy, May 24-27, 1983.

Bernacca, P.L. (ed.) [1987], Processing of scientific data from the E.S.A. Astrometry satellite HIPPARCOS, proc. of the third FAST Thinkshop, held in Bari, Italy, November 3-6, 1986.

Bertani, D., M. Cetica and D. Iorio-Fili [1986], Field to Grid Transformations for the Real System. In: P.L. Bernacca (ed.), Proc. third FAST thinkshop, Bari, November 1986, pp. 304-314.

Bertotti, B., P. Farinella, A. Milani, A.M. Nobili and F. Sacerdote [1983], Linking reference systems from space. In: P.L. Bernacca (ed.), Proc. first FAST Thinkshop, Asiago, May 1983, pp. 403-404.

Betti, B., F. Mussio and F. Sanso [1983a], A new proposal of sphere reconstitution in HIPPARCOS project. In: P.L. Bernacca (ed.), Proc. first FAST Thinkshop, Asiago, May 1983, pp. 281-298.

Betti, B. and F. Sanso [1983b], A detailed analysis of rank deficiency in HIPPARCOS project. In: P.L. Bernacca (ed.), Proc. first FAST Thinkshop, Asiago, May 1983, pp. 317-324.

Betti, B. and F. Sanso [1985a], A continuous model for the arcwise sphere reconstitution. In: J. Kovalevsky (ed.), Proc. second FAST Thinkshop, Marseille, February 1985, pp. 255-262.

Betti, B., L. Mussio and F. Sanso [1985b], Experiments with the arcwise sphere reconstitution. In: J. Kovalevsky (ed.), Proc. second FAST Thinkshop, Marseille, February 1985, pp. 263-280.

Betti, B. and F. Sanso [1986a], The continuous analog of the sphere reconstruction in HIPPARCOS project. In: Manuscripta Geodaetica (1986) 11: pp. 133-145.

Betti, B., F. Migliaccio and F. Sanso [1986b], A rigorous approach to attitude and sphere reconstitution in HIPPARCOS project. In: P.L. Bernacca (ed.), Proc. third FAST thinkshop, Bari, November 1986, pp. 391-412.

BIH [1981-1986], Annual Reports of the Bureau International de l'Heure, 1981-1986.

Birardi, G.R. [1982], A possible contribution to African geodesy: vertical deflections by portable zenith cameras. In: Proc. IAG Tokyo, 1982, pp. 632-639.

Boor, C. de [1978], A practical guide to Splines, Applied Mathematical Sciences, Vol. 27, Springer-Verlag, Berlin.

British Aerospace Dynamics Group [1983], Control law analysis and simulation report, HIP.3076.Bae, Issue 1 (Hipparcos Technical Documentation).

Bucciarelli, B., M. Lattanzi and T. Tommasini-Montanari [1986], Sphere reconstitution: parallel methods and numerical experiments. In: P.L. Bernacca (ed.), Proc. third FAST thinkshop, Bari, November 1986, pp. 97-110.

Bürki, B., H.G.. Kahle and H.H. Schmid [1983], Das neue Zenithkamera Messsystem am Institut für Geodäsie und Photogrammetrie der ETH Zürich. In: Vermessung, Photogrammetrie, Kulturtechnik, 10/83, pp. 349-354.

Burrows, C. [1982], A new iterative approach to the great circle reduction step for the HIPPARCOS satellite, LOG-HIP-3301, sept. 1982 (Hipparcos technical documentation).

Canuto, E., B. Fassino and A. Villa [1983a], The Star Mapper data processing. In: P.L. Bernacca (ed.), Proc. first FAST Thinkshop, Asiago, May 1983, pp. 155-162.

Canuto, E., A. de Luca and B. Fassino [1983b], The photon noise error in grid coordinate estimation. In: P.L. Bernacca (ed.), Proc. first FAST Thinkshop, Asiago, May 1983, pp. 187-196

Caprioli, G. [1983], Precise apparent places of stars by vector transformations. In: P.L. Bernacca (ed.), Proc. first FAST Thinkshop, Asiago, May 1983, pp. 255-262.

Catullo, V. [1985], Apparent places of stars as seen by the HIPPARCOS satellite. In: J. Kovalevsky (ed.), Proc. second FAST Thinkshop, Marseille, February 1985, pp. 329-334.

Cuthill, E. and J. McKee [1969], Reducing the Bandwidth of Sparse Symmetric Matrices. In: Proc. 24th National Conference ACM, ACM No. P-69, Brandon Systems Press, N.J.

Daalen, D.T. van [1983], General considerations on Reduction on Circles. In: P.L. Bernacca (ed.), Proc. first FAST Thinkshop, Asiago, May 1983, pp. 235-250

Daalen, D.T. van [1985a], The current state of the Reduction on Circles. In: J. Kovalevsky (ed.), Proc. second FAST Thinkshop, Marseille, February 1985, pp. 193-196

Daalen, D.T. van [1985b], The astrometry satellite Hipparcos. In: F. Sanso (ed.), Proc. first Marussi-Hotine Symp., Roma 1985, pp. 259-284

Daalen, D.T. van, and H. van der Marel [1986a], Geodetic, Geometric and Computational Aspects of Hipparcos. In: Manuscripta Geodaetica (1986) 11: pp. 146-166

Daalen, D.T. van, and H. van der Marel [1986b], Software Testplan for Reduction on Circles, Delft, august 1986, 22 p. (FAST technical documentation)

Daalen, D.T. van, B. Bucciarelli and M. Lattanzi [1986c], Rank Deficiency during Sphere Reconstruction. In: P.L. Bernacca (ed.), Proc. third FAST thinkshop, Bari, November 1986, pp. 369-376.

Daalen, D.T. van, and H. van der Marel [1987], The Modelling Error in Reduction on Circles, Delft (internal communication), March 1987.

Donati, F., E. Canuto, D. Carlucci and A. Villa [1983a], The C.S.S. approach to attitude reconstruction and raw data treatment. In: P.L. Bernacca (ed.), Proc. first FAST Thinkshop, Asiago, May 1983, pp. 147-154

Donati, F. [1983b], On dynamical smoothing. In: P.L. Bernacca (ed.), Proc. first FAST Thinkshop, Asiago, May 1983, pp. 251-254

Donati, F., E. Canuto and P. Belforte [1985], Modelling the Hipparcos attitude motion in solar eclipse conditions, abstract in: J. Kovalevsky (ed.), Proc. second FAST Thinkshop, Marseille, February 1985

Donati, F., E. Canuto, B. Fassino and P. Belforte [1986a], High accuracy on-ground attitude reconstruction for the ESA astrometry HIPPARCOS mission. In: Manuscripta Geodaetica (1986) 11: pp. 115-123

Donati, F. [1986b], Some Considerations about HIPPARCOS Reference System. In: P.L. Bernacca (ed.), Proc. third FAST thinkshop, Bari, November 1986, pp. 377-382.

Duff, I.S. [1981], A Sparse Future. In: I.S. Duff (ed.), Sparse Matrices and their Uses, Proc. of the IMA Conference, held at the University of Reading, July 9-11, 1981, Academic Press, London, pp. 1-29

Eeg, J. [1986], On the Adjustment of Observations in the Presense of Blunders, Geodætisk Institut, Tech.report No.1, København, Danmark, 1986.

ESA [1979], Hipparcos space astrometry, Report on the phase A study. ESTEC, SCI(79)10, 1979.

FAST [1981], HIPPARCOS: proposal for scientific data processing, ESA HIP 81/02, 1981.

FAST [1986a], Interface Document, Issue 2, Fast Consortium, 1986.

FAST [1986b], Hipparcos Reference Manual Guide, Issue 1, CNES, 1986.

Falin, J.L., M. Froeschlé and F. Mignard [1985], Simulation des Donnees "Grand Cercles", CERGA/FAST Note Technique 19, Oktober 1985.

Falin, J.L., M. Froeschlé and F. Mignard [1986], Simulation des Fichiers interface pour la tache "Grand Cercles", CERGA/FAST Note Technique 23, September 1986.

Fassino, B. [1986], Report on Grid Coordinate Task. In: P.L. Bernacca (ed.), Proc. third FAST thinkshop, Bari, November 1986, pp. 51-64.

Fresneau, A. [1985], Proposals of observations with the space telescope in the domain of astrometry. In: J. Kovalevsky (ed.), Proc. second FAST Thinkshop, Marseille, February 1985, pp. 97-102

Fricke, W. [1980], The FK5, an Improved Fundamental Reference System Extended to Fainter Stars. In: Mitt. Astron. Ges., 48.

Froeschlé, M. and J. Kovalevsky [1982], The connection of a catalogue of stars with an extragalactic reference frame. In: Astron. Astrophys. 116(1982), pp. 89-94.

Froeschlé, M., J. Kovalevsky and F. Mignard [1983], Effects of various types of noise on phase determination. In: P.L. Bernacca (ed.), Proc. first FAST Thinkshop, Asiago, May 1983, pp. 209-214.

Froeschlé, M., J.L. Falin, J. Kovalevsky et F. Mignard [1985a], Le logiciel de simulation, Les donnees de la grille principale. In: J. Kovalevsky (ed.), Proc. second FAST Thinkshop, Marseille, February 1985, pp. 335-350.

Froeschlé, M., J.L. Falin and F. Mignard [1985b], Observations with the star-mapper. In: J. Kovalevsky (ed.), Proc. second FAST Thinkshop, Marseille, February 1985, pp. 351-358.

Galligani, I. and L. Montefusco [1983], Computational complexity of the three-step procedure for the reconstitution of the celestial sphere. In: P.L. Bernacca (ed.), Proc. first FAST Thinkshop, Asiago, May 1983, pp. 299-204.

Galligani, I., B. Betti and P.L. Bernacca [1986], The sphere reconstitution problem in the HIPPARCOS project. In: Manuscripta Geodaetica (1986) 11: pp. 124-132.

George, A. and J.W. Liu [1981], Computer Solution of Large Sparse Positive Definite Systems, Prentice-Hall, Englewood Cliffs, N.J.

Gibbs, N.E., W.G. Poole and P.K. Stockmeyer [1976], An Algorithm for Reducing the Bandwidth and Profile of a Sparse Matrix. In: SIAM Journal of Numerical Analysis, Vol. 13, No. 2, April 1976, pp. 236-250.

Golub, G.H. and C.F. van Loan [1983], Matrix Computations, The John Hopkins University Press, Baltimore, Maryland, 1983.

Gradstheyn, I.S. and I.M. Ryzhik [1980], Table of Integrals, Series, and Products, Academic Press, New York, 1980.

Groten, E. [1982], Geodetic applications of Hipparcos results. In: M.A.C. Perryman and T.D. Guyenne (eds.) , Proc. of a colloquium on the scientific aspects of the Hipparcos space astrometry mission, Strasbourg, February 1982, pp. 71-77.

Guyenne, T.D. and J. Hunt [1985], HIPPARCOS: Scientific Aspects of the Input Catalogue Preparation, Proc. of a Colloquium held at Aussois, Savoie (France), 3-7 June, 1985, ESA SP-234.

Hageman, L.A. and D.M. Young [1981], Applied Iterative Methods. In: Computer Science and Applied Mathematics, a series of monographs and textbooks, Academic Press, New York.

Hering, R. and S. Röser [1983], On the spatial, brightness and colour distribution of IRS stars. In: P.L. Bernacca (ed.), Proc. first FAST Thinkshop, Asiago, May 1983, pp. 131-134.

Heuvel, F.A. van den, and D.T. van Daalen [1985], Grid step inconsistency correction during Reduction on Circles. In: J. Kovalevsky (ed.), Proc. second FAST Thinkshop, Marseille, February 1985, pp. 221-228.

Heuvel, F.A. van den [1986], Astrometric Satellite Hipparcos: Gridstep Inconsistency Correction during Reduction on Circles, ingenieurs scriptie T.U. Delft, April 1986.

Hoyer, P., K. Poder, L. Lindegren and E. Hog [1981], Derivation of positions and parallaxes from simulated observations with a scanning astrometry satellite. In: Astronomy and Astrophysics, 101 (1981), pp. 228-237.

Huc, C. [1985], The general structure of the data reduction and its consequences on the exploitation phase. In: J. Kovalevsky (ed.), Proc. second FAST Thinkshop, Marseille, February 1985, pp. 393-396.

IAU [1977], Recommendations to the IAU General Assembly. In: Müller, E.A. and A. Jappel (eds), Proceedings of the sixteenth General Assembly, Grenoble, 1976, Transactions of the IAU, Vol. XVIB.

Jonge, P.J. de [1987], On the ordering of unknowns during the Hipparcos reduction on circles, ingenieurs scriptie TU Delft, November 1987.

Joosten, P. [1986], Een casestudie met betrekking tot iteratieve oplossingsmethoden in verband met Hipparcos (in dutch), case study, Fac. Geodesy, TU Delft, 1986.

King, I.P. [1970], An automatic reordering scheme for simultaneous equations derived from network systems. In: International Journal Numerical Methods Engineering, Vol. 2, pp. 495-509.

Kok, J.J. [1984a], On data snooping and multiple outlier testing, National Geodetic Survey, Rockville Md. 1984.

Kok, J.J. [1984b], Some notes on numerical methods and techniques for least squares adjustments. In: Surveying Science in Finland, 2(1984), pp. 1-35.

Kovalevsky, J. and M.T. Dumoulin [1983], Observing and dwell-time strategies. In: P.L. Bernacca (ed.), Proc. first FAST Thinkshop, Asiago, May 1983, pp. 271-274

Kovalevsky, J. [1984], Prospects for Space Stellar Astrometry. In: Space Science Reviews 39 (1984), pp. 1-63

Kovalevsky, J. (ed.) [1985a], Processing of scientific data from the E.S.A. Astrometry satellite HIPPARCOS, proc. of the second FAST Thinkshop, held in Marseille, France, January 21-25, 1985, 404 p.

Kovalevsky, J., M. Froeschlé, J.L. Falin and F. Mignard [1985b], Weighting of phases and grid coordinates. In: J. Kovalevsky (ed.), Proc. second FAST Thinkshop, Marseille, February 1985, pp. 119-124

Kovalevsky, J. [1986a], The project HIPPARCOS. In: Manuscripta Geodaetica (1986) 11: pp. 89-96

Kovalevsky, J. and F. Mignard [1986b], Phases and Grid Coordinates for Single and Multiple Stars, CERGA/FAST Technical Note 22, June 1986.

Kovalevsky, J., J. Schrijver and J.L. Falin [1986c], FAST Calibration document, version 2, December 1986 (FAST Technical documentation)

Lacroute, P. [1983], Le lissage d'attitude dans le project HIPPARCOS. In: P.L. Bernacca (ed.), Proc. first FAST Thinkshop, Asiago, May 1983, pp. 141-146

Lawson, C.L. and R.J. Hanson [1974], Solving Least Squares Problems, Prentice-Hall, Englewood Cliffs, N.J.

Leeuwen, F. van, and M.A.J. Snijders [1986], Hipparcos Data Simulations, proc. of a workshop held on 22 and 23 September 1986, Royal Greenwich Observatory, Herstmonceux.

Lestrade, J.F., R.A. Preston, R.L. Mutel, A.E. Niell and R.B. Phillips [1985], Linking the HIPPARCOS catalog to the VLBI inertial reference system, high angular resolution structures and VLBI positions of 10 radio stars. In: J. Kovalevsky (ed.), Proc. second FAST Thinkshop, Marseille, February 1985, pp. 87-98.

Levy, R. [1971], Restructuring of the structural stiffness matrix to improve computational efficiency. In: Jet Propulsion Laboratory Technical Review 1, 1971, pp. 61-70.

Lindegren, L. [1979], Scientific Data Processing, Sec. 3.6, Rep. PF-616. In: ESA, Hipparcos space astrometry, Report on the phase A study, 1979.

Lindegren, L. and S. Soderhjelm [1980], How much weight can be gained on high-priority stars?, Lund observatory, Sweden, 1980.

Lindegren, L. and F. van Leeuwen [1985a], Attitude Reconstruction by the Northern Data Analysis Consortium (NDAC). In: J. Kovalevsky (ed.), Proc. second FAST Thinkshop, Marseille, February 1985, pp. 103-110.

Lindegren, L. and C. Petersen [1985b], Great Circle Reduction by the Northern Data Analysis Consortium (NDAC). In: J. Kovalevsky (ed.), Proc. second FAST Thinkshop, Marseille, February 1985, pp. 183-192

Lindegren, L. and S. Soderhjelm [1985c], Sphere Reconstitution by the Northern Data Analysis Consortium (NDAC). In: J. Kovalevsky (ed.), Proc. second FAST Thinkshop, Marseille, February 1985, pp. 237-242.

Lindegren, L. [1986], "SIMSET": Simulation of input data for the set solution, NDAC/LO/77, august 1986.

Lund, N. [1984], Re-ordering and Dynamical Smoothing of Simulated Hipparcos data, DSRI/Hipparcos Note, No. 3, Copenhagen, 1984.

Lynn, P.A. [1973], An introduction to the Analysis and Processing of Signals, MacMillan Publ. Ltd, London, 1982.

Marel, H. van der [1983a], Solar Radiation Pressure Torque Model for HIPPARCOS, CERGA/TU Delft (stage verslag), February 1983.

Marel, H. van der [1983b], Attitude Dynamics of HIPPARCOS: description & theory of an attitude simulation program package for HIPPARCOS, CERGA/TU Delft (stage verslag), March 1983.

Marel, H. van der [1983c], Numerical Smoothing of Attitude Data. In: P.L. Bernacca (ed.), Proc. first FAST Thinkshop, Asiago, May 1983, pp. 264-270.

Marel, H. van der [1983d], Astrometric Satellite Hipparcos: Test computations for Reduction on Circles and Smoothing of Attitude Data, ingenieurs scriptie T.U. Delft, September 1983.

Marel, H. van der [1985a], Star Abscissae Improvement by Smoothing of Attitude Data. In: J. Kovalevsky (ed.), Proc. second FAST Thinkshop, Marseille, February 1985, pp. 197-208.

Marel, H. van der [1985b], Large Scale Calibration during Reduction on Circles. In: J. Kovalevsky (ed.), Proc. second FAST Thinkshop, Marseille, February 1985, pp. 209-220

Marel, H. van der [1986a], Chain Architecture document Reduction on Circles, Delft, January 1986, Issue 2 (FAST technical documentation).

Marel, H. van der, and D.T. van Daalen [1986b], On the Geodetic Aspects of the Astrometry satellite Hipparcos. In: proceedings C.S.T.G. symposium Toulouse (COSPAR), June 1986 (4 p.)

Marel, H. van der, F. van den Heuvel and D. van Daalen [1986c], Testruns Reduction on Circles on CERGA dataset I. In: F. van Leeuwen and M.A.J. Snijders (ed.), proc. Hipparcos data simulation workshop, Herstmonceux , September 1986, pp. 119-140.

Marel, H. van der, and D.T. van Daalen [1986d], Recent Results in Reduction on Circles. In: P.L. Bernacca (ed.), Proc. third FAST thinkshop, Bari, November 1986, pp. 65-80.

Marel, H. van der [1987a], Testruns Delft Reduction on Circles software on Lund dataset LOSIM3, TU Delft, May 1987, 26 p.

Marel, H. van der [1987b], Comparison of Great Circle Reduction Results between FAST and NDAC, TU Delft, Issue 2, November 1987, 29 p.

Oppenheim, A.V., A.S. Willsky and I.T. Young [1983], Signals and Systems, Prentice-Hall, Englewood Cliffs, N.J.

Perryman, M.A.C. and T.D. Guyenne (eds.) [1982], Scientific aspects of the Hipparcos space astrometry mission, Proc. of an International Colloquium organized by the European Space Agency, Strasbourg, 22-23 February 1982, ESA SP-177.

Perryman, M.A.C. [1985], Ad Astra Hipparcos - The European Space Agency's Astrometry Mission. ESA BR-24, ESA, Noordwijk, 1985.

Pieplu, J.L. [1986], Progress Status of the Data Processing Task. In: P.L. Bernacca (ed.), Proc. third FAST thinkshop, Bari, November 1986, pp.135-141.

Pinard, M., I. Stellmacher, H. van der Marel, J. Kovalevsky and L. Saint-Crit [1983], Hipparcos Attitude Simulation. In: P.L. Bernacca (ed.), Proc. first FAST Thinkshop, Asiago, May 1983, pp. 91-98

Preston, R.A., J.F. Lestrade and R.L. Mutel [1983], Linking HIPPARCOS observations to an extragalactic VLBI frame by use of optically bright radio stars. In: P.L. Bernacca (ed.), Proc. first FAST Thinkshop, Asiago, May 1983, pp. 395-402

Röser, S. [1983], Link of the HIPPARCOS system with the FK5. In: P.L. Bernacca (ed.), Proc. first FAST Thinkshop, Asiago, May 1983, pp. 391-394.

Schrijver, J. [1986], HIPPARCOS: The instrument. In: Manuscripta Geodaetica (1986) 11: pp. 97-102.

Smith, D.E. and J.G. Marsh [1986], Modelling of the Earth's Gravity Field. In: Adv. Space Res., Vol.6, No.9, 1986.

Snay, R.A. [1976a], Reducing the Profile of Sparse Symmetric Matrices, NOS NGS-4, NOAA Technical Memorandum, Rockville, Md. 1976.

Snay, R.A. [1976b], Reducing the Profile of Sparse Symmetric Matrices. In: Bulletin Geodesique, Vol. 50, 1976, pp. 341-352.

Sugawa, C. and I. Naito [1982], Final refraction problems in time and latitude observations through classical techniques. In: Proc. Gen. Meeting Int. Ass. Geodesy, IAG, Tokyo, 1982, pp. 573-577.

Sünkel, H. [1981], Cardinal Interpolation, Reports of the department of Geodetic science, Report No. 312, The Ohio State University, Columbus.

Tengstrom, E., and G. Teleki (eds.) [1978], Refractional Influences in Astrometry and Geodesy, IAU symposium No. 89, D. Reidel Publ. Comp., Holland, 1978.

Teunissen, P.J.G. [1985], The Geometry of Geodetic Inverse Linear Mapping and Non-Linear Adjustment, Netherlands Geodetic Commission Publications on Geodesy 8(1985) 1.

Tommasini-Montanari, T. [1983], Numerical methods for the solution of large scale systems. In: P.L. Bernacca (ed.), Proc. first FAST Thinkshop, Asiago, May 1983, pp. 305-312.

Tommasini-Montanari, T., B. Bucciarelli and M. Lattanzi [1985a], Continuous experiments on the great circle reduction. In: J. Kovalevsky (ed.), Proc. second FAST Thinkshop, Marseille, February 1985, pp. 229-236.

Tommasini-Montanari, T., B. Bucciarelli and C. Tramontin [1985b], Two different methods for the solution of the large least squares problem in the sphere reconstitution. In: J. Kovalevsky (ed.), Proc. second FAST Thinkshop, Marseille, February 1985, pp. 243-254.

Varga, R.S. [1962], Matrix Iterative Analysis, Prentice-Hall, Inc., Englewood Cliffs, N.J.

Vegt, C. de [1982], On the Importance of the Hipparcos Stellar Net for Photographic Catalogue Work. In: M.A.C. Perryman and T.D. Guyenne (eds.) , Proc. of a colloquium on the scientific aspects of the Hipparcos space astrometry mission, Strasbourg, February 1982, pp. 49-52.

Verwaal, R.G. [1986], Proefberekeningen met de Delftse Hipparcos Software, deel 1 t/m 3 (in dutch), Geodesy, TU Delft, 1986 (internal communication).

Walter, H.G. [1983a], Recognition of grid step ambiguities. In: P.L. Bernacca (ed.), Proc. first FAST Thinkshop, Asiago, May 1983, pp. 275-278

Walter, H.G. [1983b], Astrometric parameters determination. In: P.L. Bernacca (ed.), Proc. first FAST Thinkshop, Asiago, May 1983, pp. 325-330

Walter, H.G., R. Hering, U. Bastian and H.H. Bernstein [1985a], Astrometric parameter determination, Methods, Algorithms and program implementation. In: J. Kovalevsky (ed.), Proc. second FAST Thinkshop, Marseille, February 1985, pp. 281-290

Walter, H.G. [1985b], Generation of geometric and apparent star positions. In: J. Kovalevsky (ed.), Proc. second FAST Thinkshop, Marseille, February 1985, pp. 323-328

Walter, H.G., F. Mignard, R. Hering, M. Froeschlé and J.L. Falin [1986], Apparent and geometric star positions for the HIPPARCOS mission. In: Manuscripta Geodaetica (1986) 11: pp. 103-114

Wertz, J.R. (ed.) [1978], Spacecraft Attitude Determination and Control, Astrophysics and Space Science Library, Vol. 73, D. Reidel publishing company, Dordrecht.