Contribution à la validation statistique des données d'Hipparcos : catalogue d'entrée et données préliminaires

Frédéric Arenou

THÈSE DE DOCTORAT

ASTRONOMIE FONDAMENTALE, MÉCANIQUE CÉLESTE ET GÉODÉSIE

Présentée le 29 Mars 1993 à l'Observatoire de Paris

⋈⋈⋈⋈⋈⋈⋈⋈⋈⋈⋈

CONTRIBUTION À LA VALIDATION STATISTIQUE DES DONNÉES D'HIPPARCOS : CATALOGUE D'ENTRÉE ET DONNÉES PRÉLIMINAIRES

⋈⋈⋈⋈⋈⋈⋈⋈⋈⋈⋈

Frédéric Arenou

M^{r} D. Egret	Observatoire de Strasbourg	Rapporteur
M^{me} A. Gómez	Observatoire de Paris-Meudon	Directrice de thèse
M^{r} L. Lindegren	Observatoire de Lund, Suède	Rapporteur
M^{r} J-C. Mermilliod	Observatoire de Lausanne, Suisse	Examinateur
M^{r} F. Mignard	Observatoire de la Côte d'Azur	Président
M^{r} M.A.C. Perryman	ESTEC, Noordwijk, Pays-Bas	Examinateur
M^{r} C. Robert	Université de Rouen	Examinateur
M^{me} C. Turon	Observatoire de Paris-Meudon	Examinatrice

« *À celui à qui je dois tout*[1] »

À la lecture des pages de remerciements de diverses thèses (qui sont sans doute les pages les plus lues), on peut difficilement ne pas remarquer les témoignages de reconnaissance de pure forme, voire les flagorneries, la citation ci-dessus n'en étant que la caricature.

Comment alors exprimer une gratitude sincère ? Simplement en étant bref :

Toute ma reconnaissance va à ma directrice de thèse, Ana Gómez, pour toute l'attention dont elle a fait preuve mais aussi pour sa vision dynamique de la recherche, et avec qui les disputes resteront un excellent souvenir.

Je voudrais également dire combien j'ai apprécié les membres du bâtiment Hipparque, qui ont contribué à la réalisation de cette thèse.

Et puis, pour toutes les minutes de liberté qu'elle m'a laissées, et pour toutes les minutes de bonheur dont elle me comble, je remercie ma fille Leïla, Estrella, Eva – la nuit, l'étoile, la vie.

1. « *Ça coûte pas cher et ça fait plaisir à un tas de gens* » Claire Bretécher, *Les frustrés*

Table des matières

Introduction

La mission scientifique du satellite Hipparcos apportera sans doute de profonds bouleversements dans de nombreux domaines de l'Astronomie, en particulier la Structure et la Dynamique Galactiques. Ce sujet, dans lequel est spécialisée notre équipe, bénéficiera des données très précises d'Hipparcos qui permettront de mettre en lumière des effets importants jusqu'alors passés inaperçus. À condition, bien entendu, d'utiliser les méthodologies adéquates, notamment statistiques.

L'apprentissage de ces méthodologies est une des motivations de cette thèse. Quant au sujet, du fait de notre profonde implication dans la préparation du Catalogue d'Entrée d'Hipparcos, il a évolué au cours du temps.

Une fois le Catalogue obtenu, il est apparu logique de comparer son contenu avec les premiers résultats du satellite. Comme les parallaxes que fournira Hipparcos sont d'intérêt majeur, nous avons particulièrement étudié leurs propriétés statistiques.

Ceci nous a conduit à utiliser d'autres méthodes d'estimation des parallaxes, principalement photométriques et spectroscopiques, et donc à approfondir les problèmes liés aux calibrations des magnitudes absolues, aussi bien spectroscopiques que photométriques. Nous avons également étudié et pris en compte les biais causés par les erreurs aléatoires sur les magnitudes absolues ou sur les parallaxes.

La première partie de cette thèse concerne donc la provenance et la gestion des données que nous utilisons, essentiellement la base de données INCA, qui contient les meilleures données obtenues depuis le sol. Nous décrirons les calibrations des couleurs intrinsèques et de la magnitude absolue obtenues à partir de la photométrie $uvby{-}\beta$. Puis nous aborderons une modélisation réaliste de l'absorption interstellaire dans le visible que nous avons réalisée.

Ce modèle empirique a servi à obtenir une estimation des indices de couleurs de plusieurs dizaines de milliers d'étoiles devant être observées par Hipparcos. Outre cet aspect directement lié à la préparation de la mission Hipparcos, ce modèle très général permet de prédire l'absorption interstellaire jusqu'à plusieurs centaines de parsecs du Soleil, en fonction de la position galactique ; il devrait s'avérer utile aux études statistiques utilisant les distances d'étoiles, et il a d'ailleurs été mis à la disposition de la communauté astronomique internationale.

La seconde partie met l'accent sur l'indispensable méthodologie statistique qui permet de mener à bien l'étude des données obtenues. Nous nous intéresserons principalement aux erreurs de mesures et à la correction des biais qui en résultent, mais nous décrirons également des tests statistiques et la façon d'obtenir des estimateurs adéquats.

Cette méthodologie statistique sera notamment utilisée pour les comparaisons, effectuées dans la troisième partie, entre les données au sol et les données préliminaires du satellite. Après avoir comparé les données de position et les données photométriques, nous décrirons plusieurs méthodes permettant de valider les parallaxes d'Hipparcos, et nous les appliquerons aux parallaxes préliminaires.

Une grande partie des indicateurs de distance dans notre Galaxie sera ainsi utilisée afin d'étudier les propriétés statistiques des parallaxes préliminaires des étoiles lointaines. Nous comparerons les résultats obtenus par les consortiums de réduction des données et nous indiquerons, par des méthodes originales, comment obtenir les éventuelles erreurs systématiques, les erreurs externes, et le meilleur estimateur des parallaxes définitives.

Finalement, la dernière partie est consacrée à une étude concernant la cinématique des étoiles naines de type A dans le voisinage solaire. Faisant le lien avec les parties précédentes, nous utiliserons dans cette étude les données de la base INCA, les calibrations photométriques et des méthodes statistiques de classification. Nous montrerons que les distributions observées des vitesses peuvent être interprétées comme la contribution de quelques groupes aux caractéristiques cinématiques différentes, ceci impliquant que le temps de mélange des vitesses est au moins supérieur à deux années galactiques.

Nous indiquons en annexe les publications auxquelles nous avons collaboré et une partie des programmes informatiques réalisés pendant l'élaboration de cette thèse.

Première partie

DES DONNÉES OBSERVATIONNELLES AUX PARAMÈTRES PHYSIQUES

Tout au long de cette thèse, nous allons utiliser des données pour des calibrations, des comparaisons ou des études, et il paraît indispensable de commencer par décrire la composition ainsi que la provenance de ces données.

Sans nous appesantir réellement sur ce sujet, nous mentionnerons ce qui concerne le contenu de la base de données INCA, son utilisation, comment elle fut créée et sa finalité. Ce sera le sujet du chapitre 1. Hipparcos d'abord, et son Catalogue d'Entrée qui aura suscité une vaste opération de collecte et d'homogénéisation, formera l'essentiel des données que nous utiliserons ; ce Catalogue d'Entrée est en quelque sorte le résumé de ce qui se fait de mieux au sol. Hipparcos toujours, avec les premiers résultats envoyés par le satellite puis traités par les consortiums de réduction des données (DRC), sera la deuxième source des données que nous étudierons, afin de les comparer ultérieurement aux données au sol.

Dans un deuxième temps, chapitre 2, nous décrirons l'implantation de calibrations concernant la photométrie $uvby{-}\beta$, que l'on utilisera essentiellement afin d'obtenir la magnitude absolue individuelle d'étoiles, mais également pour connaître leur excès de couleur. Ces magnitudes absolues seront comparées aux magnitudes absolues moyennes spectroscopiques, afin de pouvoir disposer d'une estimation de la dispersion de ces magnitudes absolues moyennes.

Enfin, nous décrirons au chapitre 3 un modèle qui nous permet d'obtenir une estimation de l'absorption interstellaire ainsi que de l'excès de couleur des étoiles avec une précision raisonnable. Bien que ce modèle ait été élaboré pour les besoins de la mission Hipparcos, il n'en est pas moins beaucoup plus général, puisqu'il peut prédire une valeur approchée de l'absorption interstellaire en fonction des coordonnées galactiques.

Ainsi, partant des données observées, cette partie décrit les paramètres physiques qu'elles nous permettent d'obtenir, principalement la couleur intrinsèque et la magnitude absolue. C'est ultérieurement, dans les parties III et IV, que ces paramètres pourront être pleinement exploités.

Chapitre 1

Les données observationnelles

1.1 La mission Hipparcos

Le satellite Hipparcos, lancé le 8 août 1989, a pour mission d'observer environ 120 000 étoiles, contenues dans un Catalogue d'Entrée, afin d'obtenir leurs positions, parallaxes trigonométriques et mouvements propres. Par suite de la défaillance de son moteur d'apogée, il ne se trouve pas sur l'orbite géostationnaire qu'il aurait dû avoir, mais sur l'orbite de transfert légèrement modifiée, fortement elliptique (apogée à 36 000 km, périgée à environ 500 km) sur laquelle il tourne avec une période de $10^{\mathrm{h}}40^{\mathrm{m}}$.

La précision des paramètres astrométriques obtenus par Hipparcos doit être meilleure que 2 millièmes de seconde d'arc (mas) pour une étoile plus brillante que la magnitude 9. Hipparcos est un satellite astrométrique, certes, mais également photométrique, puisque la précision des magnitudes doit être meilleure que quelques millièmes de magnitude (typiquement 0.002 mag pour une étoile non variable de magnitude 8.5 [Mignard *et al.*, 1992]).

Une seconde expérience embarquée sur le satellite, Tycho, utilisant le repéreur d'étoiles d'Hipparcos, fournira des données photométriques et astrométriques environ 10 fois moins précises, mais pour près d'un million d'étoiles.

Avant d'indiquer comment est composé le Catalogue d'Entrée d'Hipparcos, il faut expliquer en quelques mots le fonctionnement du satellite. Pour plus de détails, on pourra se rapporter à Perryman & Hassan (1989) ou Perryman *et al.* (1992).

Hipparcos observe simultanément deux champs de $0.9^\circ \times 0.9^\circ$, séparés par un angle d'environ 58°, et balaye continûment le ciel. Le signal de chaque étoile des champs est modulé par une grille située dans le plan focal ; chaque étoile est observée de nombreuses fois pendant la mission avec un temps d'observation qui dépend de sa magnitude, de la priorité scientifique qui lui a été accordée et des autres étoiles présentes simultanément dans l'un des deux champs.

Par conséquent, il ne doit pas y avoir dans un champ trop d'étoiles, surtout si elles sont faibles. D'un autre côté, il doit y avoir suffisamment d'étoiles observées, réparties uniformément sur le ciel, faute de quoi le satellite déterminerait difficilement son attitude (son orientation dans l'espace).

La liste des étoiles à observer par Hipparcos devait donc non seulement être établie à l'avance, mais, de plus, les positions des étoiles devaient être connues à mieux que 1.5 seconde d'arc, pour qu'elles soient repérées sans ambiguïté sur la grille, et les magnitudes (dans la bande H_p proche de V qui est celle du détecteur d'Hipparcos) plus précises

que 0.5 mag, afin que l'étoile soit suffisamment «posée» mais sans gaspiller de temps d'observation.

1.2 Les étoiles à observer par Hipparcos

En tant que satellite astrométrique, le but d'Hipparcos était de fournir un catalogue uniforme sur le ciel de positions, parallaxes et mouvements propres, mais il était évident que les étoiles à observer devaient également être sélectionnées en fonction de leur intérêt astrophysique.

À la suite de l'appel à propositions lancé en 1982 par l'Agence Spatiale Européenne (ESA) à l'attention de la communauté scientifique, 214 propositions furent reçues, représentant environ 600 000 étoiles, dont une grande partie redondantes (la même étoile, mais appelée avec des noms différents).

Tout le travail du consortium INCA (acronyme de INput CAtalogue), dont le centre est basé à l'Observatoire de Paris-Meudon, consistait à faire de ces propositions une liste unique d'étoiles aux positions et magnitudes précises, choisies de manière à ce qu'un maximum d'entre elles ait un intérêt astrophysique et astrométrique prioritaire et puisse être trouvé, puis observé par le satellite avec le temps d'observation optimum : cette liste est «le Catalogue d'Entrée».

1.2.1 Les propositions d'observation

Les implications de l'avènement d'Hipparcos sont nombreuses :

- Concernant les positions des étoiles, le satellite contribuera à l'obtention d'un système de référence d'une précision jusqu'alors inégalée, qui servira aux études du mouvement de la Terre et des corps du système solaire.

- Les mouvements propres déterminés par Hipparcos seront, eux, du plus haut intérêt pour l'étude de la dynamique et de la cinématique de notre Galaxie, et en particulier des régions de formation d'étoiles et des amas.

- Mais, surtout, le facteur 5 d'augmentation de la précision des parallaxes acquises avec Hipparcos par rapport à celles obtenues depuis le sol permettra sans doute d'obtenir les résultats astrophysiques les plus spectaculaires. Pas seulement pour la connaissance de l'échelle des distances cosmiques ; les mesures directes de distances d'étoiles pour lesquelles il fallait jusqu'à présent se contenter d'évaluations indirectes amélioreront en effet considérablement les déterminations des luminosité, rayon, âge et masse de ces étoiles.

Rien d'étonnant alors qu'autant de types différents d'étoiles aient été proposés pour qu'elles soient observées par Hipparcos (tableau 1.1). Les domaines de recherche couverts sont très larges, allant de l'évolution stellaire à l'évolution galactique en passant par l'étude du milieu interstellaire...

En plus des 214 propositions d'observations, il s'est également avéré nécessaire d'avoir une liste d'étoiles brillantes, régulièrement réparties sur le ciel, à la fois pour le contrôle d'attitude du satellite et pour la réduction des données : le «Survey».

TAB. 1.1: *Types d'étoiles observées par Hipparcos.*

A-type stars	Eruptive variable stars	R stars
A stars (H-line emission)	F-type stars	R Cr B stars
Am stars	G-type stars	RR Lyrae stars
Ap stars	H- and K- emission stars	RS CVn stars
B-type stars	He abnormal stars	S stars
Barium stars	Herbig Ae and Be stars	Semi-regular var. stars
Be stars	Horizontal branch stars	Symbiotic stars
β CMa stars	Irregular variable stars	Sub-dwarf stars (cool)
Carbon stars	K-type stars	Sub-dwarf stars (hot)
Cataclysmic variable stars	λ Bootis stars	T Tauri stars
Cepheid stars	Li rich stars	VV Cephei stars
CH like star	Long-period variable stars	Weak G-band stars
CH strong stars	M-type stars	Weak metal stars
CJ stars	Marginal barium stars	White dwarfs
CS stars	Mira Ceti stars	Wolf-Rayet stars
δ Delphini stars	O-type stars	W UMa stars
δ Scuti stars	Planetary nebulae	WW Virginis stars
Early-type H-poor stars	Pulsating variable stars	X-ray binaries

(d'après Gómez, 1992)

1.2.2 Le Survey

Le «Survey» est donc une liste d'étoiles brillantes rajoutée aux propositions initiales faites par la communauté astronomique ; cette liste garantit l'observation de (quasiment) toutes les étoiles jusqu'à la magnitude limite

$$\left| \begin{array}{rcll} V_{\rm lim} & = & 7.3 + 1.1 \sin|b| & \text{si le type spectral est plus tardif que G5} \\ V_{\rm lim} & = & 7.9 + 1.1 \sin|b| & \text{si le type spectral est plus précoce ou égal à G5} \end{array} \right. \quad (1.1)$$

où b désigne la latitude galactique.

Le mot «quasiment» mérite quelques éclaircissements. Les magnitudes V utilisées pour la définition du Survey étant à l'origine celles contenues dans SIMBAD en 1984 (voir §1.3), non seulement elles avaient souvent une erreur aléatoire d'écart-type supérieur à 0.2 mag, mais également des erreurs systématiques conduisant à une surestimation de toutes les demi-magnitudes arrondies (7.00,7.50,8.00,8.50,...) [1].

De plus, les types spectraux n'étant pas présents pour toutes les étoiles, ou bien alors non homogènes (les types HD et MK n'étant pas identiques), ou parfois incorrects, il faut s'attendre à ce qu'il y ait des étoiles introduites par erreur ($\approx$ 2 500) ou au contraire supprimées par erreur ($\approx$ 1 000) dans le Survey, malgré le soin apporté à sa réalisation, ce qui relativise d'une certaine façon la notion de complétude. De plus, les contraintes

1. les observateurs ayant sans doute une préférence pour les chiffres qui tombent juste...

d'observation ont fait que 6% des 55 000 étoiles n'ont pu être inclues dans le Catalogue d'Entrée.

La forme elle-même de la définition du Survey peut être expliquée en quelques mots. La coupure en type spectral permet de limiter la contribution des géantes rouges, dont les magnitudes absolues et les âges sont plus difficilement déterminés, en faveur d'autres types d'étoiles, plus proches, et qui auront donc une parallaxe plus précise.

La dépendance en fonction de la latitude galactique, quant à elle, permet d'assurer approximativement une distribution uniforme sur tout le ciel, propriété intéressante à la fois pour le fonctionnement du satellite et pour la future exploitation scientifique des résultats de celui-ci, d'observer des étoiles autres que celles du Survey au voisinage du plan galactique.

Une étude exhaustive du Survey peut être trouvée dans Crifo (1988) et Turon *et al.* (1989).

1.3 La base de données INCA

Pour chacune des étoiles proposées, il était nécessaire de posséder des données, astrométriques et photométriques entre autres. Et pour pouvoir convenablement traiter toutes ces données, il fallait disposer d'une base de données et d'un système de gestion de base de données (SGBD).

SIMBAD, la base de données gérée par le Centre de Données de Strasbourg (CDS), qui disposait de la structuration adéquate et d'un SGBD très pratique d'utilisation, était un choix qui s'imposait de lui-même. De plus, la disponibilité et la compétence des membres du CDS était un atout qui ne s'est d'ailleurs jamais démenti depuis.

La base de données INCA fut donc créée comme «sous-base» de SIMBAD (en termes de nombre d'objets) ou «sur-base» de SIMBAD (en termes de nombre de données par étoile, et parce que $\approx 5\,000$ étoiles ont été ajoutées qui n'étaient pas dans SIMBAD). Hormis la bibliographie concernant les étoiles, qui nous était peu utile, la structure d'une étoile se compose donc de données fondamentales, de toutes les identifications, et de mesures, comme SIMBAD. Ce sont ces dernières qui devaient être mises à jour, au fur et à mesure qu'elles étaient envoyées à Meudon par les différents instituts participant à la préparation de la mission.

En effet, en plus des données contenues à l'origine dans SIMBAD, et qui se trouvèrent *de facto* également dans la base INCA, des données supplémentaires concernant directement ou non Hipparcos furent ajoutées pendant toute la préparation du Catalogue d'Entrée.

Les premières, nécessaires à la précision exigée du Catalogue d'Entrée, concernaient :

- l'astrométrie : compilation et nouvelles mesures de positions et de mouvements propres, hiérarchisés et mis à l'équinoxe 2000, époque 1990 [Jahreiß *et al.*, 1992] ;
- la photométrie : compilation et nouvelles observations de magnitudes et couleurs [Grenon *et al.*, 1992] ;
- la variabilité : calcul des magnitudes moyennes pendant la mission des étoiles variables [Mennessier *et al.*, 1992] ;

– la duplicité : compilation et nouvelles observations de positions, séparations, magnitudes d'étoiles doubles [Dommanget, 1989] ;

– les données résultant des simulations de la mission [Crézé *et al.*, 1989].

À titre d'exemple de la quantité du travail fourni, 100 000 positions furent mesurées sur des plaques photographiques, 10 000 observées sur des cercles méridiens ; 10 000 magnitudes et couleurs furent obtenues en photométrie photoélectrique [Turon, 1992]. Toutes ces données mirent des années à être acquises – grâce à une authentique collaboration Européenne –, puis vérifiées ; elles forment un ensemble cohérent qui fait l'originalité de la base INCA. Environ 21 instituts de 8 pays – sans compter les observateurs et participants occasionnels – y auront contribué. Quand l'Europe politique balbutie encore, l'Europe scientifique travaille et réussit, et ceci, depuis des années. De plus, les relations entre les différents consortiums auront été la source d'un enrichissement que l'on espère mutuel.

Signalons de façon marginale, parce que nous les utiliserons dans cette thèse, les autres données introduites dans la base INCA, et qui ont été utilisées pour des calibrations ; tout d'abord le 4$^{\text{ème}}$ volume du Michigan Spectral Survey [Houk, 1988] qui nous permit d'avoir assez de données spectroscopiques pour la création d'un modèle de l'extinction interstellaire (chap. 3) ; d'autre part le catalogue de photométrie $uvby{-}\beta$ de Hauck & Mermilliod (1990) grâce auquel nous allions (chap. 2) obtenir les couleurs dérougies et les magnitudes absolues individuelles pour plusieurs milliers d'étoiles. Enfin, nous utiliserons également dans la partie IV des données cinématiques pour lesquelles la vitesse radiale est indispensable ; pour ce faire, le catalogue de Barbier [Barbier, 1989] a également été introduit dans la base de données INCA.

L'article ci-joint [Arenou & Morin, 1988] décrit dans le détail les différentes données contenues dans la base INCA, ainsi que l'ensemble des logiciels spécifiques développés pour la préparation du Catalogue d'Entrée.

ASTRONOMY
FROM LARGE DATABASES
SCIENTIFIC OBJECTIVES AND METHODOLOGICAL APPROACHES

ESO Conference and Workshop Proceedings No. 28

THE INCA DATABASE FOR THE PREPARATION OF THE HIPPARCOS MISSION

Arenou F. & Morin D.
Observatoire de Meudon, F-92195 Meudon Principal Cédex

ABSTRACT : The INCA database, including all stars proposed for Hipparcos observation, has been created as a subset of SIMBAD database. It has now evolved independently and contains about 214,000 stars (with 5,000 stars not yet in SIMBAD). The data contained in INCA are daily improved with the help of specific softwares.

I. WHY THE INCA DATABASE

The Hipparcos project, included in the scientific programme of the European Space Agency (E.S.A) in 1980, aims at the very accurate measurement of the positions, parallaxes, and proper motions of about 110,000 preselected stars, with a precision of about 0.002" on positions and parallaxes, and of 0.002" per year for proper motions.

The INCA Consortium, leaded by C.Turon, is carrying out the preparation of the INput CAtalogue, which will contain all the stars to be observed by the Hipparcos satellite.

From the very beginning of this work, in early 1982, the support of the Centre de Données de Strasbourg (C.D.S.), and the intensive use of the SIMBAD database turned out to become essential in order to sort out redundancies among the 600,000 stars submitted by more than 200 proposers of the international astronomical community.

At that time, SIMBAD clearly was the only set of identifications and stellar data which could be interrogated using any identification, thus allowing to clarify most of the redundancies (see Fig 1). Cross-identifications were done using the Strasbourg-Cronenbourg Univac computer from Meudon Observatory.

The completeness of SIMBAD for all the stars up to about 9th magnitude allowed the construction of a basic list of bright stars (survey) accounting for half of the Input Catalogue (see Turon, Gómez & Crifo, this volume).

The first release of the catalogue (1983) contained 160,000 stars in a sequential file. Including the survey stars and stars newly identified, and eliminating redundancies, the INCA database now contains about 214,000 stars. The necessity of interrogating, of making statistics, and of updating very frequently the catalogue naturally implied, early 1985, the creation of a specific database, named INCA (acronym for INput CAtalogue), suited to our task, with the same structure and basic data as SIMBAD, thus allowing the use of all the softwares developed for SIMBAD, which have proved to be very efficient.

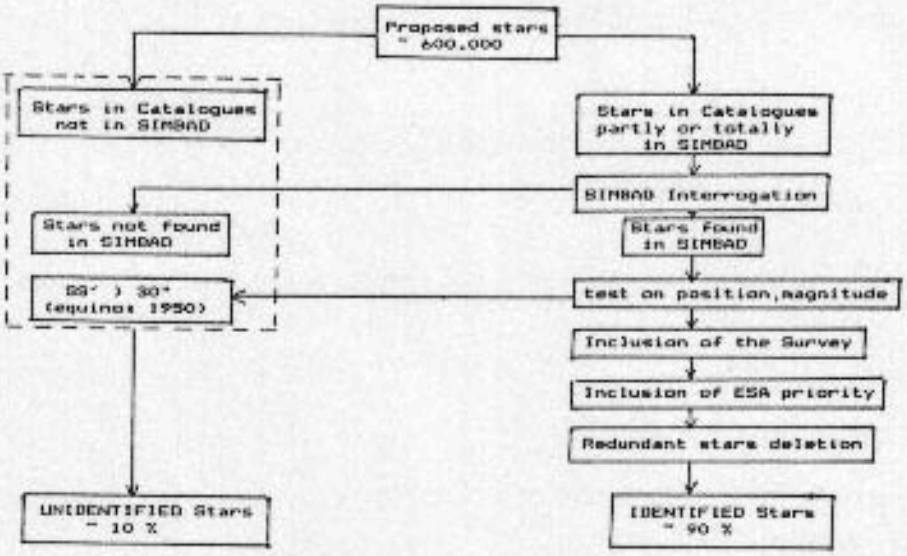


Figure 1

Situation before the INCA Database creation.

2. A SPECIFIC DATABASE

Once created, the INCA database was then developed in an autonomous way, though in close collaboration with the C.D.S. team (see Fig 2). The INCA database includes additional stars, different entries and new identifiers and data :

2.1 New stars

The INCA database now contains about 214,000 stars : 209,000 stars from SIMBAD and 5,000 additional stars which were not yet in SIMBAD.

2.2 Different entries

Due to the specificity of the Hipparcos satellite, stars in double and multiple systems received a special treatment according to the following rule :

1 entry if $\rho < 10''$ (AB) and 2 entries if $\rho \geq 10''$ (A+B), where ρ is the separation between two components A and B.

2.3 Identifiers

In addition to identifiers already in SIMBAD, some other identifiers (see Fig 3) were required to take into account :

- by which proposal each star was demanded (*"I"* identifier)
- the location of a star in 0.81 square degrees cells (size of the field of view of the satellite), to study the sky distribution (*"IBIL"* identifier) ;
- the Lausanne photometric identifier to facilitate the introduction of new photometric data provided by the Lausanne team (*"LID"* identifier) ;
- information about components of double and multiple systems given by the Working Group on Double & Multiple Stars, Observatoire Royal de Belgique (*"CCD2"* & *"CCDM"* identifiers).

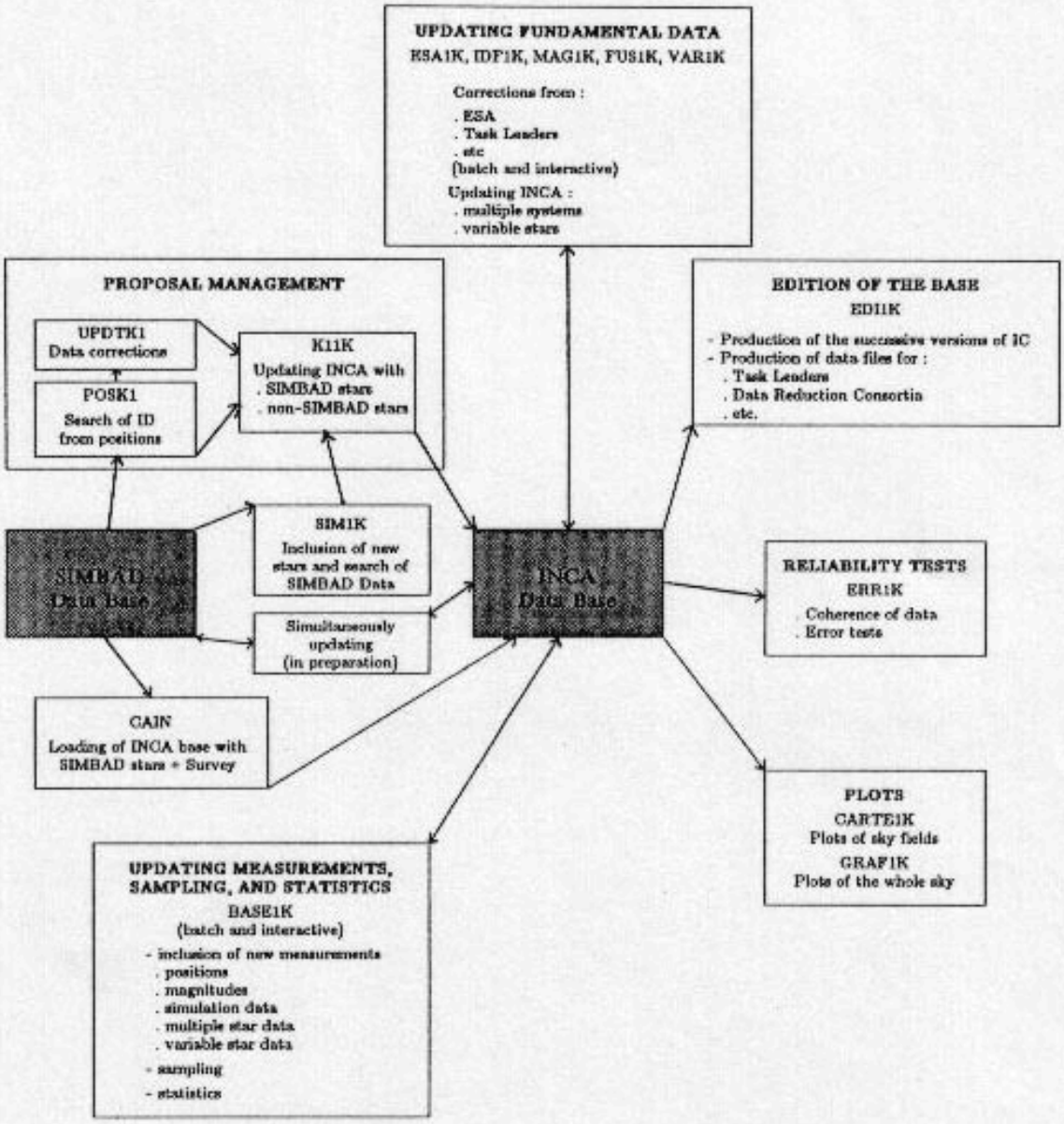


Figure 2
Input Catalogue preparation : New softwares

INCA Identifier

Identifier name	Proposal number	Running Nr in the proposal	ESA priority	Proposer's priority	error code
I	54	2182	4R	0	0

IBIL Identifier

Identifier name	Cell number in galactic latitude : IB	Cell number in galactic longitude : IL	Discriminating number of stars in the cell
IBIL	-78	134	1

Figure 3
"I" and "IBIL" identifiers.

2.4 Measurements

New measurements, continuously improved, obtained by ground-based observations and measurements or compiled in the literature by other working groups of the INCA Consortium are introduced in several catalogues. The situation in October 1987 is :

- "*pos*" and "*pmI*" containing positions and proper motions : 181,000 and 189,000 measurements respectively ;
- "*pH*" and "*vH*" for photometry about normal and variable stars : 86,000 and 12,000 measurements respectively ;
- "*sH*" for parameters concerning numerical mission simulation : 1,600,000 measurements ;
- "*CCDM*" containing data about double and multiple stars.

3. NEW SOFTWARES

Apart from the softwares written by the C.D.S. staff, many new softwares operating on the INCA database and including a total of 36,000 PL/I statements and 6,000 FORTRAN statements have been written for 3 years, to solve different problems specifically related to the elaboration of the IC, though these softwares are compatible with SIMBAD structure.

3.1 Management of proposals

This software allows the exploration of SIMBAD, within a 10 arcmin radius around each of the 50,000 stars which were proposed using identifiers not recognized by SIMBAD.

Another software introduces the data concerning the newly recognized stars from SIMBAD into INCA .

3.2 Updating the INCA database

New identifiers, fundamental data and measurements are updated following the recommendations of ESA, INCA task-leaders, working groups coordinators, and proposers. This is performed using a C.D.S. SIMBAD software as well as new softwares described in Fig 2. Another software allows corrections of the entries in double & multiple systems.

3.3 Sampling and making statistics

In order to analyse the sky distribution, the effects of the different simultations and to detect redundancies, softwares were designed to perform samples and statistics using different astrometric, photometric, and spectroscopic parameters, ESA priority, etc.

3.4 Plots

Graphic softwares were written, to quickly figure out the local and global distribution of the considered stars. Figure 4 illustrates the aspect of the sky distribution of the stars proposed by proposal #53 ; Figure 5 displays a zone of 4 square degrees around the 502th star of proposal #53.

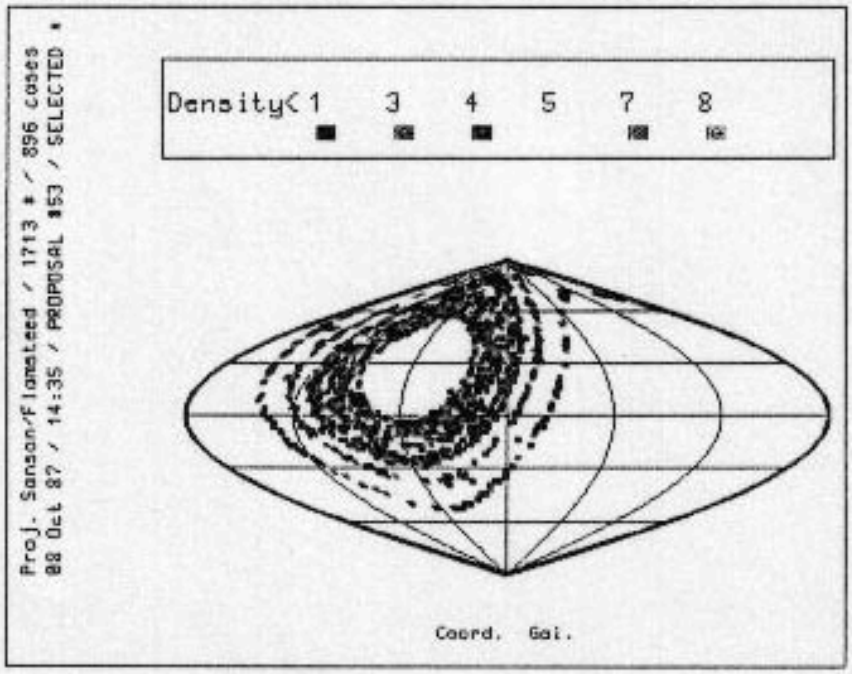


Figure 4

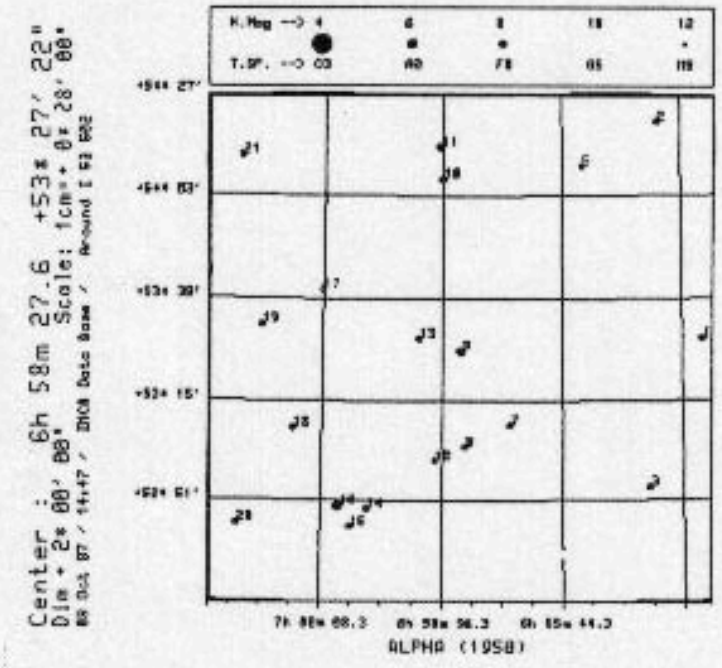


Figure 5

3.5 Edition of the INCA database

One important software allows the edition of the database for the different INCA working groups. Several editions have been necessary, each one containing improved data, for the simulation group, for ground-based observers, for ESA, for Data Reduction Consortia, etc.

3.6 Reliability tests

In order to improve the reliability of the Input Catalogue, a software has been designed to detect possible errors on astrometric, photometric and spectroscopic data, errors on cross-identifications and redundancies ; for each star, the consistency of all the data is tested : within each kind of data, between different types of data (e.g. colour index vs spectral type), and between identifiers and data.

4. TECHNICAL ASPECTS

4.1 Staff

Apart from the staff of the other working groups who provides the newly compiled or observed data for the preparation of the Input Catalogue, 3 scientists, 2 software engineers, and 4 technicians are specifically involved in the updating of the INCA database at the Observatoire de Meudon.

4.2 Internal structure of the database

The structure is the same as that of SIMBAD :

- 1 main file, containing all the data concerning one star : fundamental data, list of identifiers and measurements ; each field be variable in length ;
- 3 identifiers index files in order to allow access to an object from any of its identifiers ;
- 1 coordinate boxes data file in order to allow a quick access to stars when interrogating by their coordinates ;
- the bibliographical texts data file is the SIMBAD one ; it contains the texts of the references of the measurements.

4.3 Content of the database (October 1987)

stars : 214,000 ;

identifiers : 2,125,000 (547,000 "I" identifiers) ;

measurements : 3,166,000 ;

The INCA database contains 153 MBytes :

- 47 MBytes for fundamental data, system information and free bytes reserved for future use ;
- 63 MBytes for identifiers and measurements (90% proper to INCA) ;
- 43 MBytes for the index files ;

On average, each star is read 120 times/year and modified 12 times/year (see Fig 6)

4.4 Performances

The performances of the INCA database are similar to those of SIMBAD (Wenger, 1985) :

- interrogation by identifier : instantaneous answer in dialogue mode, 1h30m for the whole database in batch ;
- interrogation by coordinates : 2 sec in dialogue mode ;
- sampling : less than 5 min for a sample requiring the reading of the whole database.

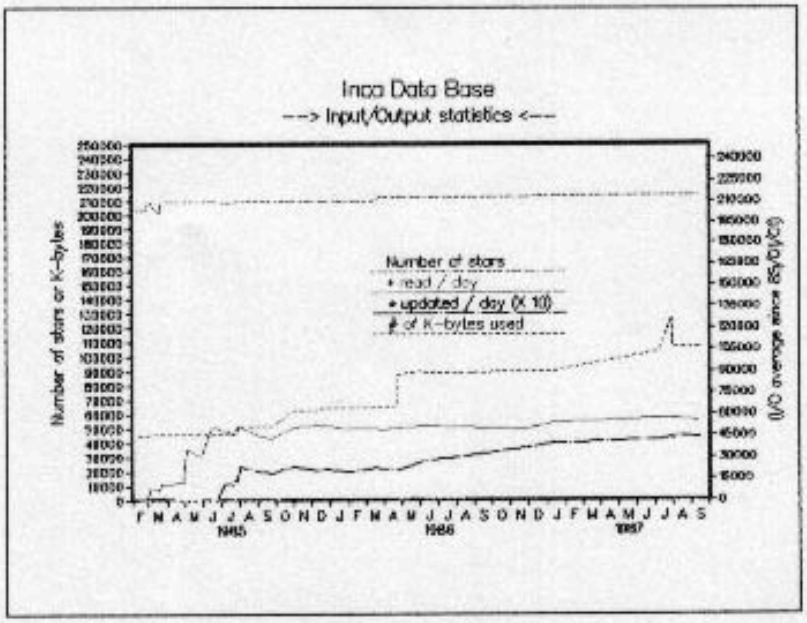


Figure 6

REFERENCES

- "Hipparcos : scientific uses of the INCA database", Turon C., Gómez A., Crifo F., this volume.
- "The INCA database, sub base of Simbad", Morin D., Arenou F., 1985, INCA Colloquium "Scientific aspects of the Input Catalogue preparation", Aussois, pub ESA SP-324, Perryman & Turon eds.
- "Construction of the Input Catalogue", Gómez A., Crifo F., Morin D., Arenou F., 1985, INCA Colloquium, Aussois, op.cit.
- "Presentation of the astronomical database SIMBAD", Wenger M., 1985, INCA Colloquium, Aussois, op.cit.

1.4 Le Catalogue d'Entrée d'Hipparcos

À partir de la base de données INCA, de la priorité scientifique donnée à chaque étoile, et des contraintes d'observation, le Catalogue d'Entrée a été créé par itérations successives. Tout ce qui concerne la formation du Catalogue d'Entrée peut être trouvé dans Perryman & Turon (1989).

Des traitements particuliers ont dû être réservés à certains types d'étoiles ; à titre d'exemple, on peut citer les étoiles composantes de systèmes doubles ou multiples. Objets à l'intérêt astrophysique indéniable, ils n'en posent pas moins nombre de problèmes lors de la mission Hipparcos, aussi bien pour les données disponibles au sol, pour l'observation par le satellite que pour la réduction des données.

Dans le premier cas, cela est dû au fait que, les étoiles dans un tel système étant voisines, les données photométriques (et autres) ne sont pas forcément connues individuellement, et, quand elles le sont, il n'est pas rare qu'il y ait confusion entre les composantes. Le catalogue CCDM [Dommanget, 1989], contenant des données individuelles pour chaque composante des étoiles doubles ou multiples, a contribué à résoudre une partie de ce problème.

Dans le deuxième cas, cela provient de la taille du champ de vision d'Hipparcos : le signal d'une étoile peut être perturbé par une étoile voisine, tout dépendant de la distance angulaire entre les deux étoiles et de leur différence de magnitude ; c'est ainsi que des étoiles non demandées ont dû être ajoutées au programme d'observation, simplement parce qu'elles perturbaient le signal de leur compagne demandée. Les composantes séparées par moins de 10 secondes d'arc forment une entrée unique dans le Catalogue d'Entrée ; nous appellerons par la suite «composantes fusionnées» de telles entrées.

Enfin, lors de la réduction, il s'agit de résoudre les couples, pour obtenir les paramètres individuels de chacune des composantes, ce qui n'est pas une mince affaire, surtout lorsqu'il s'agit de systèmes doubles qui n'étaient pas jusqu'alors connus comme tels au sol.

Le traitement des données de ces étoiles doubles ou multiples a donc nécessité une attention particulière, et c'est une des tâches que nous avons été amené à traiter à l'intérieur du consortium INCA.

On peut faire en passant une remarque qui aura son utilité lors de l'étude des parallaxes Hipparcos : mis à part le Survey, les étoiles du Catalogue d'Entrée n'ont pas grand-chose de commun entre elles, sinon de représenter le meilleur compromis entre l'intérêt scientifique et les impératifs techniques d'observabilité par Hipparcos. Ceci a une implication statistique : un échantillon d'étoiles du Catalogue d'Entrée sera tout, sauf homogène (pas de complétude en magnitude, etc). La contrepartie astrophysique de cette hétérogénéité est, bien entendu, que tous les types d'étoiles «intéressants» seront observés.

Finalement, la préparation du Catalogue d'Entrée aura été une opération assez laborieuse : on peut estimer à environ 200 année-hommes le coût humain qui lui aura été consacré [Turon, 1992]. Mais cette élaboration, pour aussi longue qu'elle fût, n'aura pas été vaine : on verra, chapitre 5, que le satellite a trouvé et observé correctement quasiment la quasi-totalité des 118 209 étoiles qui lui étaient proposées.

1.5 Les résultats des consortiums de réduction des données

Une fois les étoiles observées, commence le travail de la réduction des données fournies par le satellite. Les détails de cette réduction sont parfaitement décrits dans Perryman *et al.* (1989). Le problème était en tout cas suffisamment complexe pour que l'Agence Spatiale Européenne décide de confier cette tâche à deux consortiums indépendants et travaillant en parallèle : FAST et NDAC. Pour des raisons d'implantation géographique majoritaire, il nous arrivera parfois d'écrire consortium nord (pour NDAC) ou consortium sud (pour FAST), mais que l'on se rassure, le dialogue Nord-Sud est ici parfaitement équilibré, FAST et NDAC comparant régulièrement leurs résultats.

Il suffit sans doute de préciser que le satellite transmet au sol 24Kbits par seconde – soit environ 400 milliards d'octets sur l'ensemble de la mission, desquels il faut extraire non seulement les paramètres astrométriques des étoiles, mais également l'attitude du satellite et tous ses paramètres instrumentaux – pour que l'on devine l'étendue du travail de la réduction des données.

Mais ce n'est à vrai dire pas la quantité d'informations qui rend cette tâche ardue, mais le fait que toutes les données de la mission sont interdépendantes et doivent être traitées par étapes successives. Très schématiquement, le processus de réduction [Lindegren, 1985] se déroule en trois temps [Froeschlé, 1992c] :

1. le signal d'une étoile indique sa position sur la grille de l'instrument ;

2. cette position permet d'obtenir l'abscisse de l'étoile sur un grand cercle de la sphère céleste ;

3. au cours de la mission, cette étoile aura été observée en moyenne sur 30 des 2 500 grands cercles, soit environ 150 transits [Mignard *et al.*, 1992]. La solution sur la sphère consistera alors à résoudre 1.5 million d'équations pour 40 000 étoiles de référence, puis à traiter les autres étoiles...

Après la fin de la mission Hipparcos, lorsque l'ensemble des données aura été traité, la solution sur la sphère permettra donc d'obtenir pour chaque étoile la magnitude et les 5 paramètres astrométriques : la position (α, δ), le mouvement propre ($\mu_\alpha \cos\delta$, μ_δ), et la parallaxe π, ainsi que les variances et covariances formelles entre ces différents paramètres. Ce dernier point a de l'importance puisqu'il permet de vérifier si l'ensemble des effets instrumentaux aléatoires a été pris en compte lors de la réduction.

Néanmoins, après seulement un an de données, les deux consortiums, FAST et NDAC, décidèrent de tenter d'obtenir une solution sur la sphère. Le but évident était la vérification sur les données réelles des différents logiciels impliqués dans la réduction des données, notamment au moyen d'une comparaison entre les résultats obtenus par chaque consortium. Le mot «notamment» indique que cette comparaison entre les consortiums est nécessaire mais pas suffisante. On peut en effet très bien imaginer que les données brutes obtenues du satellite aient des problèmes inconnus (le fonctionnement nominal du satellite n'était déjà pas simple, et la mission révisée, à cause de la panne du moteur d'apogée, n'a rien arrangé) qui fassent obtenir par les consortiums des résultats identiques, mais incorrects.

Les données fournies aux consortiums par le Centre d'Opérations Spatiales Européennes (ESOC) pour cette première réduction s'étendent sur la période du 27 Novembre 1989 au 16 Décembre 1990 [Froeschlé, 1992b], [Lindegren, 1992a]. Les observations concernent la totalité des étoiles du Catalogue d'Entrée pour FAST, 115 314 étoiles pour NDAC, et totalisent approximativement 1.3 million d'abscisses sur environ 800 grands cercles.

En réalité, ce n'est pas une, mais deux solutions qui ont été obtenues par chaque consortium. La première est l'obtention des 5 paramètres astrométriques de chaque étoile (solution FAST-5P, NDAC-5P), la deuxième consiste en l'obtention des positions et de la parallaxe (solution FAST-3P, NDAC-3P) en utilisant les mouvements propres donnés par le Catalogue d'Entrée.

Pourquoi cette deuxième solution ? Simplement parce que les mouvements propres ne sont bien déterminés que si la base de temps pour les mesurer est assez longue. Un an de mission peut être insuffisant, et, par conséquent, la médiocre détermination des mouvements propres peut dégrader considérablement la qualité des parallaxes. C'est en tout cas ce qui ressort des études réalisées par M. Froeschlé et L. Lindegren en comparant les résultats des deux consortiums, et c'est ce que l'on va vérifier également dans cette thèse grâce à des comparaisons avec des données externes.

1.5.1 Les données préliminaires utilisées

Les paramètres astrométriques de ces solutions nous ont été confiés par les deux consortiums, dans le but (exclusif) d'évaluer les qualités des parallaxes préliminaires. C'est l'objet de l'étude réalisée au chapitre 6.

Divers critères ont conduit les deux consortiums à ne délivrer qu'un échantillon d'étoiles plus restreint que toutes les étoiles du Catalogue d'Entrée observées la première année.

Pour ce qui est de la solution 5 paramètres obtenue par le consortium FAST, les étoiles sélectionnées sont celles considérées comme simples par Hipparcos, observées au moins 6 fois, et dont la corrélation entre le mouvement propre et la parallaxe est inférieure à 0.6. En ce qui concerne la solution 3 paramètres, n'ont été gardées que les étoiles considérées comme simples (i.e. non doubles) dans le Catalogue d'Entrée (ceci ne préjuge pas d'un caractère éventuellement double ou multiple, Hipparcos découvrant de nouvelles étoiles non simples), observées au moins 6 fois pendant l'année d'observation, et dont l'erreur formelle sur la parallaxe est inférieure à 4 mas. Ceci représente un total de 46 716 étoiles.

En ce qui concerne la solution 5 paramètres obtenue par NDAC, les étoiles gardées sont celles dont l'erreur formelle est inférieure à 4 mas et pour lesquelles le coefficient de corrélation entre les mouvements propres et la parallaxe est inférieure à 0.6. Ce dernier point aura sans doute tendance à éliminer la contamination sur la parallaxe d'une mauvaise détermination des mouvements propres. Quant à la solution 3 paramètres, elle ne supprime que les étoiles dont l'erreur formelle sur la parallaxe est supérieure à 3 mas.

Nous ne nous intéresserons en fait qu'à la «meilleure solution» pour chacun des consortiums, de façon à limiter le nombre de comparaisons. Les consortiums ont naturellement fait la comparaison entre leurs solutions ([Froeschlé, 1992b], [Lindegren, 1992a]) et nous nous intéresserons ici essentiellement aux comparaisons avec des données *externes*. Par la suite donc, sauf mention contraire, la parallaxe FAST désignera la solution 3 paramètres et la parallaxe NDAC, la solution 5 paramètres.

Chapitre 2

Obtention des paramètres fondamentaux par la photométrie $uvby{-}\beta$

Essentiellement calibrée par Strömgren à partir des années soixante [Strömgren, 1966], et par Crawford (1975), la photométrie à bande étroite $uvby{-}\beta$ est une des méthodes permettant d'obtenir des informations sur les paramètres physiques fondamentaux qui caractérisent une étoile (magnitude absolue visuelle M_V, température effective $T_{\rm eff}$, gravité $\log g$, métallicité [Fe/H]) ainsi que sur le rougissement.

Nous nous sommes principalement intéressé dans cette thèse à l'obtention du rougissement et de la magnitude absolue, en laissant pour plus tard l'obtention des autres paramètres fondamentaux. Ce qui nous intéressait, en fait, était d'obtenir avec la meilleure précision possible la magnitude absolue, et donc la distance photométrique, dans le but de valider les parallaxes Hipparcos.

En effet, comme nous le verrons ultérieurement, la précision de la magnitude absolue obtenue par cette méthode est meilleure que celle obtenue avec les calibrations donnant la magnitude absolue en fonction du type spectral et de la classe de luminosité. Ceci provient du fait que les effets dûs à l'évolution, à la métallicité, voire à la rotation d'une étoile sont pris en compte, permettant d'obtenir une magnitude absolue individuelle, là où la spectroscopie ne donne que la magnitude absolue moyenne du groupe auquel appartient l'étoile.

Depuis trente ans, le système photométrique $uvby{-}\beta$ a été largement utilisé et de nombreux catalogues ont vu le jour. Pour mémoire, le dernier en date [Hauck & Mermilliod, 1990] ne contient pas moins de 70 000 mesures concernant 45 000 étoiles. C'est ce nombre important de données disponibles qui nous a fait choisir cette photométrie, plutôt qu'une autre (système de Genève, par exemple). De plus, dans la mesure où notre but était d'obtenir des distances d'étoiles lointaines pour la validation des parallaxes Hipparcos, la photométrie $uvby{-}\beta$ était bien adaptée aux types d'étoiles rencontrées (B, A, F naines, géantes et supergéantes).

2.1 La photométrie $uvby{-}\beta$

A partir des filtres u, v, b, y, β(étroit), β(intermédiaire), les indices utilisés dans la photométrie $uvby{-}\beta$ sont les suivants :

$(b-y)$, indicateur de température ;

$(u-b)$, indicateur de température ou de luminosité, suivant le type de l'étoile ;

$m_1 = (v-b)-(b-y)$, indicateur de métallicité ou de température ;

$c_1 = (u-v)-(v-b)$, indicateur de luminosité ou de température ;

$\beta = \beta$(étroit) - β(intermédiaire), indicateur de luminosité ou de température, quasiment insensible au rougissement interstellaire.

Strömgren (1966) a défini les paramètres insensibles au rougissement :

$$\begin{aligned}
[m_1] &= m_1 - \frac{E(m_1)}{E(b-y)}(b-y) \approx m_1 + 0.34(b-y) \\
[c_1] &= c_1 - \frac{E(c_1)}{E(b-y)}(b-y) \approx c_1 - 0.19(b-y) \\
[u-b] &= [c_1] + 2[m_1]
\end{aligned}$$

en notant $E(.)$ l'excès de couleur ; les indices et différences d'indices une fois dérougis sont notés :

$$\begin{aligned}
(b-y)_0 &= (b-y) - E(b-y) \\
m_0 &= m_1 - \frac{E(m_1)}{E(b-y)}E(b-y) \\
c_0 &= c_1 - \frac{E(c_1)}{E(b-y)}E(b-y) \\
(u-b)_0 &= (u-b) - \frac{E(u-b)}{E(b-y)}E(b-y)
\end{aligned}$$

et enfin Crawford (1975) a introduit les quantités $\delta m_0 = m_0^{\rm Hyades} - m_0$ et $\delta c_0 = c_0 - c_0^{\rm ZAMS}$.

Plusieurs calibrations empiriques des paramètres physiques fondamentaux à partir de ces indices ont été faites depuis Strömgren (1966), si bien qu'une grande partie du diagramme H-R est maintenant couverte, parfois par plusieurs calibrations : étoiles F [Crawford, 1975], étoiles B [Crawford, 1978], étoiles A [Crawford, 1979], étoiles A-F [Hilditch *et al.*, 1983], supergéantes B [Zhang, 1983], étoiles B [Balona *et al.*, 1984], étoiles G-K naines [Olsen, 1984], étoiles A*m* [Guthrie, 1987], étoiles F (population I ou population II) [Olsen, 1988], supergéantes F-G [Arellano *et al.*, 1990], supergéantes F-G [Gray, 1991], pour ne citer que les principales calibrations...

Le but, ici, n'est pas de faire une comparaison critique de ces calibrations, et encore moins d'en refaire une nouvelle. Il s'agit plutôt d'intégrer les différentes calibrations afin d'obtenir une couverture globale et cohérente du diagramme H-R, en renvoyant pour les détails le lecteur à la littérature citée ci-dessus.

L'essentiel du travail décrit ci-dessous a donc consisté :

1. à choisir pour chaque groupe les calibrations les plus adéquates, d'une part pour le dérougissement, d'autre part pour l'obtention des paramètres fondamentaux,

2. à résoudre les problèmes de frontière entre groupes,

3. à coder les calibrations choisies,

4. à obtenir une estimation des erreurs standards sur les paramètres obtenus à partir des erreurs observationnelles sur les indices.

En ce qui concerne les deux premiers points, nous nous sommes rapporté à l'algorithme donné dans Figueras *et al.* (1991), et consacré aux étoiles B tardives–A–F précoces naines ; nous y avons rajouté les calibrations des étoiles B supergéantes, F-G supergéantes, des naines et géantes F et des naines plus tardives que G2. Le problème principal a été de pouvoir choisir entre les différentes calibrations lorsque la photométrie *uvby* d'une part, l'indice β et le type spectral d'autre part, n'indiquaient pas le même groupe, et conduisaient donc à des résultats différents.

Pour ce qui est du troisième point, nous avons essayé de «coller au plus près» des calibrations données par les auteurs ; à titre d'exemple, les tableaux standards donnés par ces auteurs ont été utilisés avec des interpolations cubiques (et non par des approximations des tableaux à l'aide de régressions), et les indices dérougis ont été systématiquement obtenus par itérations successives jusqu'à convergence. En ce qui concerne les erreurs standards de mesure sur les indices dérougis et la magnitude absolue, ils ont été obtenus en faisant varier les indices initiaux dans leurs barres d'erreur.

Les programmes ont enfin été écrits de façon suffisamment modulaire pour permettre de changer ou compléter aisément les calibrations. L'ensemble (1300 lignes de code C) forme un algorithme permettant d'obtenir de manière automatique les estimateurs les plus appropriés des indices dérougis et des paramètres fondamentaux par la photométrie $uvby{-}\beta$, ainsi que leur erreur. Le détail des calibrations étant assez long, il est décrit en annexe, page 33.

2.2 Tests des calibrations

Pour vérifier la qualité du dérougissement et des magnitudes absolues obtenus avec ces calibrations, un premier test consiste à tracer un diagramme de Hertzsprung-Russel (H-R), ou plus exactement couleur/magnitude, à l'aide du $(B-V)_0$ et des magnitudes absolues M_V, pour toutes les étoiles possédant de la photométrie $uvby{-}\beta$ dans la base de données INCA (fig. 2.1).

L'indice $(B{-}V)$ intrinsèque est facilement obtenu à partir des magnitudes photoélectriques B et V contenues dans la base de données INCA, corrigées de l'excès de couleur $E(B{-}V)$ en utilisant l'approximation $E(B{-}V) \approx 1.35E(b-y)$ [Crawford & Mandewala, 1976]. L'excès $E(b-y)$ provenant justement du dérougissement, ce diagramme permet de tester à la fois les calibrations du dérougissement et celles des magnitudes absolues.

Dans ce diagramme, nous n'avons gardé que les étoiles qui avaient un $(B-V)$ photoélectrique, celles qui n'étaient pas des binaires fusionnées, et, quand la calibration donnait un excès de couleur (légèrement) négatif, celui-ci a été considéré comme nul.

On voit nettement que les calibrations implantées couvrent une bonne partie du diagramme H-R. Naturellement, ce diagramme est un peu inhabituel, puisqu'il y manque notablement les géantes rouges, les naines M et les naines blanches, toutes catégories d'étoiles pour lesquelles on ne peut pas utiliser la photométrie $uvby{-}\beta$. On peut noter

que la «ZAMS observationnelle» est bien marquée, l'essentiel des étoiles ayant une composition chimique normale ; au-dessus, on s'attend naturellement à avoir des étoiles d'âges différents, donc plus ou moins évoluées, et également des binaires.

Le raccordement des différentes calibrations se fait sans problèmes notables, ce qui était le souci principal. Une grande partie des points aberrants est due à des problèmes concernant le $(B-V)$ et non à la photométrie $uvby{-}\beta$. Pourtant, on peut naturellement craindre que des mesures effectuées sur des instruments différents soient difficiles à homogénéiser, et qu'il y ait des informations contradictoires apportées par les différents indices ; le programme de calibration réussit apparemment à détecter et traiter les étoiles dans ce cas, sauf pour les étoiles de type voisin de G0 ($(B-V)_0 \approx 0.6$, $M_V \approx 5$). À cet endroit, on passe d'une calibration utilisant essentiellement β à une calibration utilisant $(b-y)$ et le raccordement n'est pas très adéquat pour les étoiles ayant un β correspondant à 'plus tardif que G2' tandis que le $(b-y)$ correspond plutôt à 'plus précoce que F8'.

Une autre vérification des calibrations consiste à utiliser les étoiles munies de photométrie $uvby{-}\beta$ qui possèdent un identificateur d'amas, et à comparer les magnitudes absolues déduites des modules de distance d'amas (voir page 143) avec celles obtenues par la calibration décrite ici (fig. 2.2).

Pour effectuer cette comparaison, nous avons utilisé les modules de distance $V_0 - M_V$ du catalogue de Lyngå (1987), et calculé pour chaque étoile sa magnitude absolue en utilisant sa magnitude apparente corrigée de l'absorption. Cette dernière est obtenue, non pas en prenant une valeur constante pour chaque amas, mais individuellement par $A_V \approx 4.3E(b-y)$ [Crawford & Mandewala, 1976]. À nouveau, la comparaison effectuée permet donc de tester à la fois dérougissement et magnitudes absolues.

La moyenne des différences, dans le sens $M_{\rm amas} - M_{uvby-\beta}$ vaut 0.08 mag, et la dispersion (robuste) 0.58 mag, après suppression des points à plus de 5 fois l'erreur interne. La dispersion est plus importante que ce que l'on pourrait attendre de l'erreur sur la magnitude absolue $M_{uvby-\beta}$ seule (≈ 0.3 mag).

Mais dans cette dispersion intervient l'erreur sur l'absorption, l'erreur sur le module de distance, et également le problème de l'appartenance réelle de l'étoile à l'amas considéré. Pour éclaircir ce dernier point, indiquons que nous avons sélectionné les étoiles qui possèdent une identification d'amas, mais la proximité apparente dans le ciel n'implique bien évidemment pas une appartenance physique à l'amas ; nous avons éliminé certaines étoiles connues comme non-membres, mais sans doute pas toutes.

2.3 Magnitudes absolues photométriques et spectroscopiques

Pour valider les calibrations décrites ci-dessus, une méthode consisterait également à comparer les magnitudes absolues obtenues par la photométrie $uvby{-}\beta$ aux magnitudes absolues obtenues par la classification spectrale. Des comparaisons de ce type ont déjà été effectuées, voir par exemple Oblak *et al.* (1976). Compte-tenu de la taille et de la qualité du matériel disponible (le catalogue de Hauck & Mermilliod (1990) en est un exemple), nous avons tenu à refaire cette comparaison en cherchant à évaluer d'éventuelles différences systématiques entre les deux calibrations et surtout à obtenir de cette façon les dispersions des magnitudes absolues spectroscopiques dans chaque classe spectrale. Ce dernier point a

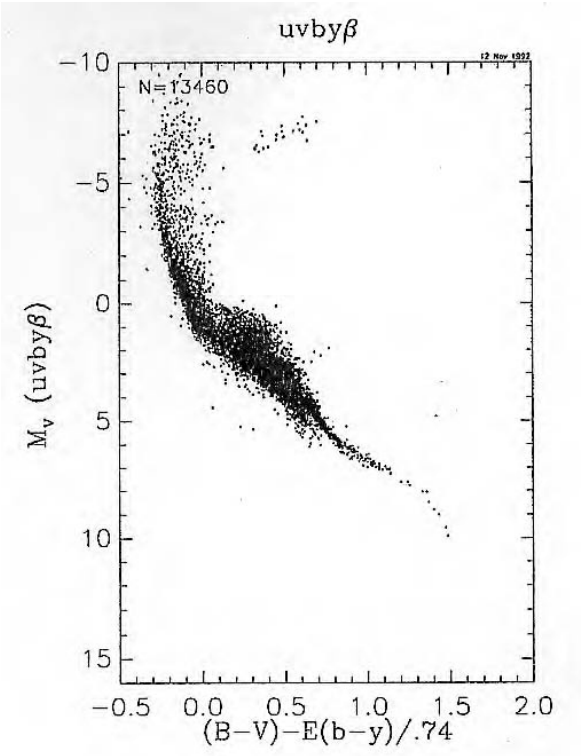


FIG. 2.1: *Diagramme $(B-V)_0/M_V$ des étoiles de la base INCA possédant de la photométrie uvby−β*

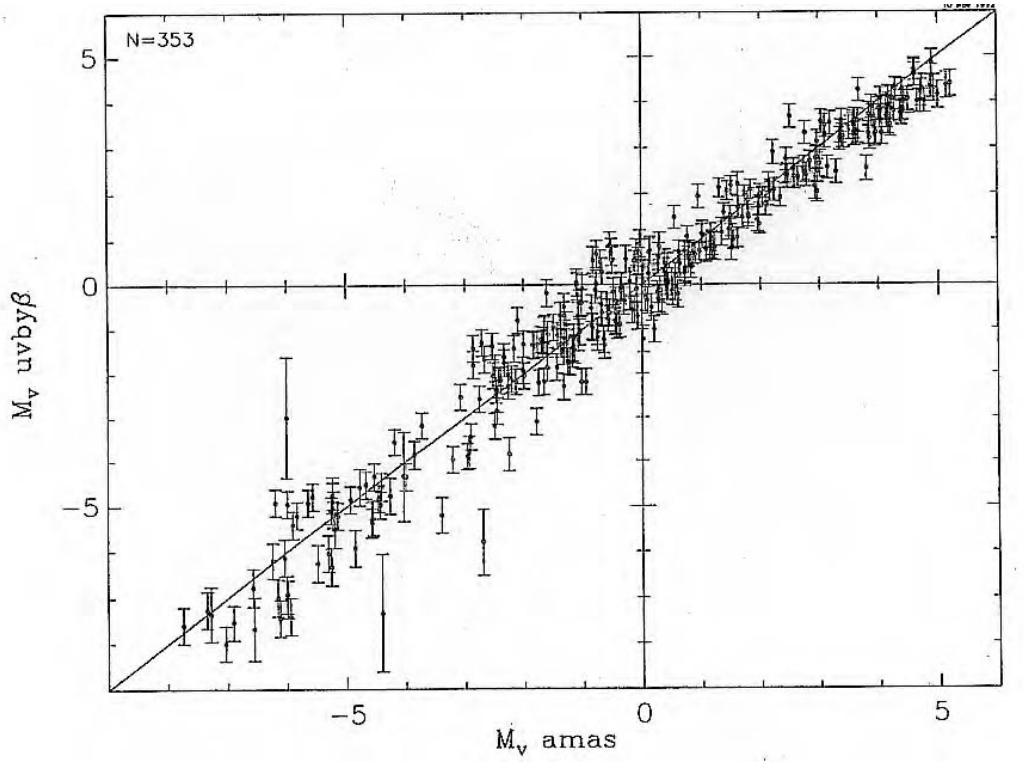


FIG. 2.2: *Comparaison des magnitudes absolues obtenues par la photométrie uvby−β et par le module de distance d'amas*

été peu étudié, malgré son importance. Il a, en effet, une implication claire sur l'estimation des erreurs des parallaxes spectroscopiques et sur le biais de Malmquist (voir page 128). Quant au premier point, la recherche de différences systématiques, nous allons voir qu'il n'y a aucune conclusion sérieuse à en attendre si l'on ne définit pas correctement l'échantillon utilisé.

2.3.1 Les magnitudes absolues spectroscopiques

Faute d'indicateur direct comme les parallaxes trigonométriques, la distance d d'une étoile peut être obtenue si l'on connaît sa magnitude apparente m_V dans le filtre V de Johnson, l'absorption interstellaire A_V, et si sa magnitude absolue M_V peut être déduite de son type spectral et de sa classe de luminosité. La loi de Pogson relie en effet ces quantités par :

$$5 \log d = m_V - A_V - M_V + 5$$

La quantité $\frac{1}{d}$ sera alors nommée parallaxe spectroscopique. Naturellement, il faut pour cela que l'on connaisse la valeur moyenne de la magnitude absolue selon le type de l'étoile.

Plusieurs calibrations de la magnitude absolue visuelle en fonction du type spectral et de la classe de luminosité existent. Citons par exemple Schmidt-Kaler (1982), Corbally & Garrison (1984), Grenier *et al.* (1985). Dans cette dernière calibration, l'influence du biais de Malmquist (voir §6.3.1) est prise en compte explicitement, et, de plus, elle présente l'avantage d'avoir été obtenue de façon homogène, alors que les deux premières sont le résultat de compilations. On notera M^S, M^C, M^G les calibrations de la magnitude absolue visuelle en fonction du type spectral et de la classe de luminosité publiées respectivement dans Schmidt-Kaler (1982), Corbally & Garrison (1984), et Grenier *et al.* (1985).

Schmidt-Kaler fournit également une calibration en fonction de l'indice de couleur intrinsèque $(B-V)_0$ et de la classe de luminosité : on notera M^{Sc} la magnitude absolue obtenue par cette méthode. Toutes ces calibrations ne sont valables que pour des étoiles de population I, ce qui ne posera pas de problème, au vu du très faible nombre d'étoiles de population II dans le Catalogue d'Entrée d'Hipparcos.

Les limites rencontrées avec ces calibrations proviennent du fait que, en général, elles ne tiennent pas compte du degré d'évolution d'une étoile, de sa composition chimique, de sa rotation, d'une binarité éventuelle, réduisant donc à deux dimensions un problème qui en mériterait beaucoup plus. De plus, l'utilisation du type spectral MK discrétise un phénomène essentiellement continu. Néanmoins, ces calibrations présentent l'avantage d'exister, et d'être applicables aux dizaines de milliers d'étoiles qui possèdent une classification spectrale.

Pour choisir la calibration la plus adéquate, nous nous sommes référé aux comparaisons effectuées dans Guarinos (1991) entre les magnitudes absolues M^S, M^C, M^G et la magnitude absolue M^{Sc}. Pour calculer cette dernière, Guarinos, en utilisant la photométrie UBV, obtient $(B-V)_0$ par une variante de la méthode Q. Cette méthode [Johnson & Morgan, 1953], utilise la quantité indépendante du rougissement

$$Q = (U-B) - \frac{E_{U-B}}{E_{B-V}}(B-V) = (U-B)_0 - \frac{E_{U-B}}{E_{B-V}}(B-V)_0$$

et permet de calculer les couleurs intrinsèques, grâce à l'approximation

$$\frac{E_{U-B}}{E_{B-V}} = 0.72 + 0.05 E_{B-V}$$

[Crawford & Mandewala, 1976] ; elle est essentiellement applicable aux étoiles B naines.

Le tableau 2.1 ci-dessous, extrait de Guarinos (1991), résume cette comparaison.

TAB. 2.1: *Différentes calibrations des magnitudes absolues.*

Différences entre les magnitudes absolues visuelles par plusieurs calibrations, sur un échantillon de naines de B0 à A0.
M^G : Grenier *et al.* (1985) (valable pour un échantillon limité en magnitude apparente),
M^C : Corbally & Garrison (1984),
M^S : Schmidt-Kaler (1982) ;
pour ces trois calibrations, la magnitude absolue est déterminée à partir du type spectral et de la classe de luminosité ;
M^{Sc} : Schmidt-Kaler (1982), à partir du $(B-V)_0$ et de la classe de luminosité ;
les erreurs standards sur les moyennes varient entre 0.01 et 0.04 et les erreurs standards sur les dispersions sont inférieures à 0.03 mag.

Type spectral	$\langle M^G - M^{Sc} \rangle$	$\langle M^C - M^{Sc} \rangle$	$\langle M^S - M^{Sc} \rangle$	$\sigma_{M^S - M^{Sc}}$
B0		-0.27	-0.26	0.58
B1		-0.22	-0.07	0.43
B2		0.07	-0.04	0.58
B3		0.15	0.15	0.51
B4		0.17	0.15	0.45
B5	0.20	0.27	0.20	0.45
B6	0.13	0.26	0.25	0.32
B7	0.14	0.36	0.39	0.39
B8	0.13	0.46	0.40	0.41
B9	0.08	0.38	0.46	0.46
A0	0.17	0.36	0.52	0.24

(d'après Guarinos, 1991)

On y a indiqué dans la dernière colonne la dispersion des $M^S - M^{Sc}$, mais pas les dispersions de $M^G - M^{Sc}$ et $M^C - M^{Sc}$. La raison en est simple : les magnitudes M^S, M^C et M^G sont *constantes* pour chaque type spectral et on doit donc obtenir

$$\sigma_{M^G - M^{Sc}} = \sigma_{M^C - M^{Sc}} = \sigma_{M^S - M^{Sc}} = \sigma_{M^{Sc}}$$

pour chaque type, aux variations d'échantillonnage près ; toutes les dispersions sur les différences ne mesurent donc en fait que la dispersion des magnitudes absolues M^{Sc} déterminées à partir du $(B-V)_0$ et de la classe de luminosité pour chaque type spectral. Et, par conséquent, on ne peut évidemment pas se servir de cette dispersion pour déterminer le meilleur choix entre les différentes calibrations, M^G, M^C ou M^S.

On pourrait alors penser choisir celle qui présente le biais le plus petit. On peut noter en effet un biais manifeste de $\langle M^C - M^{Sc} \rangle$ ainsi que de $\langle M^S - M^{Sc} \rangle$, qui croissent systématiquement avec le type spectral et donc avec l'indice de couleur. Ceci étant dit, il faut noter que les calibrations M^S, M^C et M^{Sc} sont (censées être) valables pour des échantillons limités et complets en distance. Si l'échantillon est complet en magnitude, une correction de ≈ -0.4 mag due au biais de Malmquist doit être apportée, auquel cas les différents résultats du tableau 2.1 deviennent comparables, tout en laissant des biais significatifs : en tout état de cause, l'échantillon qui a servi à effectuer cette comparaison n'est probablement pas complet en magnitude.

De plus, le résultat de Jaschek & Mermilliod (1984), montre que l'on peut s'attendre à des magnitudes absolues s'étalant sur un intervalle de 2 magnitudes pour des étoiles de même indice de couleur en haut de la séquence principale, et qu'il ne faut donc pas s'étonner de divergences entre différentes calibrations qui ne prennent pas en considération un critère lié à l'âge.

Si l'on peut donc préférer la calibration de Grenier *et al.*, ce ne sera donc pas grâce aux comparaisons ci-dessus, mais plutôt parce qu'elle a été effectuée de manière homogène. Malheureusement, elle ne s'applique qu'aux étoiles naines de B5 à F5 et aux géantes de B5 à F2.

Bien sûr, ce n'est pas le lieu, ici, de refaire une calibration des magnitudes absolues, parce que, pour cela, il faudrait donc ajouter des critères supplémentaires à la classification MK, et surtout parce que c'est une des applications très attendues des prochains résultats d'Hipparcos. Mais comme il nous faut choisir une calibration, nous utiliserons donc faute de mieux la calibration M^S(MK) de Schmidt-Kaler (1982) qui couvre la quasi-totalité du diagramme H-R.

2.3.2 Comparaison des magnitudes absolues

Malheureusement, cette calibration ne donne pas vraiment d'indication de la dispersion des magnitudes absolues, qui est pourtant un renseignement utile, par exemple pour connaître l'erreur formelle sur la parallaxe spectroscopique. Nous allons donc chercher à assigner des dispersions pour chaque type spectral à l'aide d'une comparaison avec les magnitudes absolues photométriques.

Notons M_i la vraie magnitude absolue de l'étoile i, M_{p_i} la magnitude absolue obtenue par la photométrie $uvby{-}\beta$, M_s la magnitude absolue moyenne du groupe spectral auquel l'étoile appartient. Pour simplifier, on supposera que :

- la magnitude absolue déduite de la photométrie $uvby{-}\beta$ peut s'écrire : $M_{p_i} = M_i + \epsilon_{p_i}$; avec ϵ_{p_i}, erreur gaussienne due à la calibration et aux erreurs sur les indices $uvby{-}\beta$, d'écart-type σ_{p_i}.

- la vraie magnitude absolue est répartie normalement autour de la magnitude absolue moyenne du groupe : $M_i = M_s + \epsilon_{s_i}$ avec ϵ_{s_i} de dispersion σ_{M_s}.

La différence entre les magnitudes absolues photométriques et spectroscopiques s'écrit alors : $M_{p_i} - M_s = \epsilon_{s_i} + \epsilon_{p_i}$ et, si ϵ_{s_i} et ϵ_{p_i} ne sont pas corrélées, on se trouve dans le cas du modèle simple abordé au §4.2.1, ce qui va nous permettre de calculer σ_{M_s}.

La comparaison a donc été effectuée sur les étoiles qui ne sont pas des binaires fusionnées, avec un type spectral MK et de la photométrie $uvby{-}\beta$, et pour lesquelles le

type MK n'indique pas de particularité (`e`, `m`, `n`, `p`, ...), soit environ 8 000 étoiles. Le tableau 2.2 calcule, pour chaque type spectral et classe de luminosité, les différences moyennes entre magnitudes absolues spectroscopiques et photométriques ; le nombre d'étoiles est compris entre 10 et 556 étoiles, et l'erreur standard sur la moyenne varie entre 0.02 et 0.25 mag.

TAB. 2.2: *Différences entre les magnitudes spectroscopiques et photométriques.*

Moyenne des différences entre magnitudes spectroscopiques et photométriques ; dispersion résiduelle des magnitudes spectroscopiques, en fonction du type spectral et de la classe de luminosité.

Type	$\langle M_s - M_{p_i} \rangle$				σ_{M_s}			
	I-II	III	IV	V	I-II	III	IV	V
O9				+0.05				0.63
B0	-0.30	-0.43		+0.00	0.66	0.43		0.24
B1	-0.27	-0.50	-0.10	+0.00	0.71	0.49	0.56	0.29
B2	-0.06	-1.05	-0.53	+0.41	0.69	0.47	0.50	0.53
B3	+0.23	-0.83	-0.60	+0.12	0.55	0.76	0.34	0.32
B5	+0.44	-0.81	-0.35	+0.17	0.71	0.64	0.43	0.22
B7		-0.63	-0.51	+0.20		0.41	0.35	0.34
B8	-0.12	-0.57	-0.42	+0.14	0.86	0.33	0.44	0.34
B9		-0.42	-0.34	+0.19		0.44	0.45	0.42
A0		-0.39	-0.38	+0.28		0.57	0.23	0.40
A1			-0.50	+0.01			0.73	0.49
A2		-0.97	-0.03	+0.21		0.58	0.72	0.61
A3		-0.48	-0.07	+0.23		0.56	0.63	0.62
A5		-0.72	-0.34	+0.40		0.80	0.64	0.69
A7		-0.67	+0.08	+0.36		0.68	0.61	0.62
A8		-0.65	-0.10	+0.33		0.59	0.60	0.56
F0		-0.43	+0.08	+0.35		0.57	0.58	0.60
F2		-0.23	-0.05	+0.77		0.60	0.57	0.40
F5		-0.54	-0.36	+0.15		0.54	0.41	0.34
G0				+0.46				<0.37
...				<0.38				...
M2				+0.19				<0.37

Le tableau indique également la dispersion résiduelle des magnitudes absolues spectroscopiques lorsqu'il était possible de la déduire de la dispersion totale autour de la moyenne et des erreurs standards de la photométrie. À partir de G0, en effet, la dispersion pour chaque type est du même ordre de grandeur que les erreurs standards sur les magnitudes absolues obtenues par la calibration $uvby{-}\beta$, c'est-à-dire que l'on ne peut obtenir qu'une valeur supérieure sur cette dispersion.

Les différences moyennes montrent de façon étonnante comment le fait de ne pas utiliser une échantillon bien sélectionné peut produire des différences systématiques im-

portantes. L'absence de complétude (en distance ou en magnitude) jointe au mélange d'étoiles d'âges et de métallicités différents fait que les différences pour les géantes sont toutes négatives. Si correction de Malmquist il doit y avoir, elle les rendrait encore plus négatives. Pour les naines, si l'on fait l'hypothèse que l'échantillon est limité en magnitude, on retrouve les mêmes différences (à 0.1 mag près) que Grenier *et al.* (1985) lors de la comparaison entre leur calibration et celle de Schmidt-Kaler (1982).

Crawford (1978), p 60, lors de sa calibration de la photométrie des étoiles B, avait fait une comparaison entre les magnitudes absolues qu'il obtenait et celles de Schmidt-Kaler (1965), pour les géantes et les naines. La comparaison était faite sur des étoiles B plus brillantes que $V = 6.5$ ou appartenant à un amas d'une part, et également en utilisant la moyenne des paramètres photométriques pour chaque type spectral pour calculer la magnitude absolue correspondante. Nous ne nous trouvons donc pas dans le même cas, aussi bien pour les données que pour la manière de les traiter (celle de Crawford étant plus adéquate). Néanmoins, pour les géantes, on retrouve la même tendance négative dans les écarts (Schmidt-Kaler (1982) moins Crawford), quoique de taille nettement inférieure à celle qui apparaît sur le tableau 2.2.

En ce qui concerne les dispersions, on peut noter que pour les naines, elle est maximum pour les A tardives ; elle est en moyenne de 0.5 mag pour les géantes. La dispersion pour les naines est en général inférieure à celle obtenue avec la calibration de Schmidt-Kaler (1982), à partir du $(B-V)_0$ et de la classe de luminosité (dernière colonne du tableau 2.1).

Ces dispersions nous serviront ultérieurement pour avoir une estimation de la dispersion de la parallaxe spectroscopique, tout en étant bien conscient que l'échantillon utilisé ici n'est pas forcément représentatif du Catalogue d'Entrée dans son ensemble.

2.4 Annexe : détail des calibrations

Nous décrivons ici en détail comment ont été effectuées les calibrations permettant de dérougir les indices $uvby-\beta$ et d'obtenir une estimation de la magnitude absolue à partir de ces indices.

Pour résumer la procédure utilisée, sont indiqués ci-dessous pour chacun des groupes :

- les meilleurs indicateurs de température et de luminosité,
- les étoiles concernées par les calibrations utilisées,
- les rapports d'excès $\frac{E(m_1)}{E(b-y)}$ et $\frac{E(c_1)}{E(b-y)}$ en utilisant Crawford & Mandewala (1976),
- la procédure de dérougissement,
- la procédure d'estimation de la magnitude absolue visuelle, et une estimation de l'erreur standard moyenne sur cette magnitude (provenant de la modélisation, d'un décalage du point-zéro des magnitudes absolues, d'éventuelle duplicité...), telle qu'elle a été obtenue par les différents auteurs des calibrations à l'aide de comparaisons externes.

Mais la première tâche consiste d'abord à pouvoir classer les étoiles dans le groupe adéquat.

2.4.1 Séparation en groupes d'étoiles

Schématiquement, les différents groupes d'étoiles que l'on va traiter sont les suivants :

- Groupe précoce (B0-A0)
- Groupe intermédiaire (A0-A3)
- Groupe tardif, T1 (A3-F0)
- Groupe tardif, T2 (F0-G2)
- Groupe tardif, T3 (naines G1-M2)
- Groupe des supergéantes B
- Groupe des supergéantes F-G

Compte-tenu du fait que les paramètres utilisés changent de propriété suivant le type spectral de l'étoile, il faut tout d'abord trouver le groupe auquel l'étoile appartient avant d'adopter la calibration adéquate. Pour cela, nous allons (comme dans Figueras *et al.*, 1991) utiliser deux critères que l'on nommera «critère de Strömgren» et «critère de Moon», qui définissent plus précisément les groupes, et qui permettent d'avoir le plus de sécurité possible dans la classification automatique des étoiles dans ces groupes. On résout ultérieurement les contradictions qui peuvent résulter d'un classement différent par les deux critères.

Critère de Strömgren

Le premier critère utilisé suit le schéma directeur de Strömgren, tel qu'il est défini en fonction des paramètres photométriques $[u-b]$, $[m_1]$, $[c_1]$ dans Strömgren (1966), Tables I, II, et mis sous forme d'organigramme dans Figueras *et al.*, fig. 1. Nous y avons rajouté les tests suivants :

- si $[c_1] > 12.5[m_1] - 0.45$ et $[c_1] < -4[m_1] + 1.7$: groupe B-supergéantes
- si $[m_1] \in [0.20, 0.75]$, $[c_1] > 0.25$, $[c_1] - (-6.88659[m_1]^3 + 14.844[m_1]^2 - 11.0525[m_1] + 3.11241) > -0.20$: groupe F/G-supergéantes
- si $[c_1] < 0.6$ et $[m_1] < 0.33$: groupe tardif T2
- si $[c_1] < 0.6$ et $[m_1] \geq 0.33$: groupe tardif T3

Critère de Moon

Cette procédure, utilisée dans Moon (1985) se sert de l'indice β, et surtout du type spectral de l'étoile. Il serait en effet dommage de ne pas se servir de l'information spectrale qui peut permettre de trouver le groupe auquel appartient l'étoile.

- Groupe précoce : type B0-A0 et $2.59 < \beta < 2.88$
- Groupe intermédiaire : type A0-A3 et $2.87 < \beta < 2.93$
- Groupe tardif, T1 : type A3-F0 et $2.72 < \beta < 2.88$
- Groupe tardif, T2 : type F0-G2 et $2.58 < \beta < 2.72$
- Groupe tardif, T3 : type G1-M2, classe V
- Groupe des supergéantes B : type B0-B9, Ia-Iab-Ib-II
- Groupe des supergéantes F-G : type F5-G5, Ia-Iab-Ib-II et $2.52 < \beta < 2.63$

Il est important de ne considérer que les naines pour le groupe tardif T3, et nous utilisons donc la classe de luminosité, si elle existe, et sinon nous ne traitons pas l'étoile.

2.4.2 Règlement des conflits

Parce qu'il existe un recouvrement entre les différents groupes, une étoile peut se retrouver classée dans un groupe par la première méthode, et dans un autre groupe par la deuxième. Il est également possible que ces deux groupes ne soient pas contigus, ou que l'un reste indéterminé (par exemple en raison d'erreurs observationnelles sur β ou d'un problème dans la classification spectrale).

Dans ce dernier cas (l'un des groupes est indéterminé), on choisit un ou deux groupes contigu au premier pour effectuer la discrimination suivante :

- Groupe précoce – groupe intermédiaire : l'étoile est dérougie comme si elle appartenait au groupe précoce ; si $(b-y)_0 > -0.02$, elle est dérougie dans le groupe intermédiaire et conservée dans ce groupe si $E(b-y) \geq 0$, $(b-y)_0 \geq -0.02$ et $\delta m_0 \leq 0.08$, voir [Figueras *et al.*, 1991] ;

– Groupe intermédiaire – groupe tardif, T1 : l'étoile est dérougie comme si elle appartenait au groupe tardif ; si $(b-y)_0 < 0.06$, elle est dérougie dans le groupe intermédiaire et conservée dans ce groupe si $E(b-y)^{\rm inter} \geq 0$ ou $E(b-y)^{\rm inter} > E(b-y)^{\rm tardif}$ [Figueras *et al.*, 1991] ;

– Groupe précoce – groupe tardif, T1 : si $\beta < 2.72 - \epsilon_\beta$, ou $[m_1] < 0.15$ ou $(b-y) < -0.01$, choix du groupe précoce ;

– Groupe tardif, T1 – groupe tardif, T2 : l'étoile est dérougie comme si elle appartenait au groupe T2 ; si $(b-y)_0 \leq 0.226$, elle est dérougie dans le groupe T1 et conservée dans ce groupe si $E(b-y)^{\rm T1} \geq 0$ ou $E(b-y)^{\rm T1} > E(b-y)^{\rm T2}$;

– Groupe tardif, T2 – groupe tardif, T3 : l'étoile est dérougie comme si elle appartenait au groupe T3 ; si $(b-y)_0 \leq 0.38$, elle est dérougie dans le groupe T2 et conservée dans ce groupe si $E(b-y)^{\rm T2} \geq 0$ ou $E(b-y)^{\rm T2} > E(b-y)^{\rm T3}$;

– Groupes des supergéantes – autres groupes : l'étoile est dérougie comme si elle n'était pas supergéante : si $\beta < 0.195[u-b] + 2.5 + \epsilon_\beta$ ou ($\delta m_0 \geq 0.05$ et $E(b-y) \geq 0.1$) ou $\delta m_0 \geq 0.1$, l'étoile est considérée comme supergéante ;

Comme on peut le constater, le choix a souvent été de garder le dérougissement qui ne donnait pas un excès $E(b-y)$ négatif, et de laisser «un peu de place» aux erreurs observationnelles ($\epsilon_\beta \approx 0.01$ correspond approximativement à deux fois l'erreur de mesure sur β).

L'ensemble de la procédure de discrimination permet de fournir le groupe correct avec une bonne probabilité. Des tests sont néanmoins faits sur les indices δm_0 et δc_0 pour vérifier que l'étoile n'est pas particulière, auquel cas, évidemment, on se garde bien de faire une quelconque calibration.

2.4.3 Groupe précoce

Domaine de validité : types spectraux B0-A0, classes de luminosité III-IV-V
indicateur de température : c_0 (et $(u-b)_0$)
indicateur de luminosité : β
rapports d'excès : $\frac{E(m_1)}{E(b-y)} = -0.330$; $\frac{E(c_1)}{E(b-y)} = 0.191$

Dérougissement

Par itérations successives, avec $c_0 \approx c_1$ comme valeur initiale, on utilise une interpolation de la relation standard $(b-y)_0(c_0)$ tabulée dans Crawford (1978), Table 1, et l'expression approchée $c_0 \approx c_1 - \frac{E(c_1)}{E(b-y)} E(b-y)$.

Magnitude absolue

La magnitude absolue M_V est obtenue par la calibration de Balona *et al.* (1984) :

$$[g] = \log_{10}(\beta - 2.515) - 1.60 \log_{10}(c_0 + 0.322)$$

$$M_V = 3.4994 + 7.2026 \log_{10}(\beta - 2.515) - 2.3192[g] + 2.9375[g]^3$$

Erreur sur M_V : $\langle \sigma_{M_V} \rangle \approx 0.3$ ($\sigma_{M_V} \approx 3.3 \frac{\sigma_\beta}{\beta - 2.515}$)

2.4.4 Groupe intermédiaire

Domaine de validité : types spectraux A0-A3, classes de luminosité III-IV-V
indicateur de température : $a = (b-y)_0 + 0.18((u-b)_0 - 1.36)$
indicateur de luminosité : $r = 0.35[c_1] + 2.565 - \beta$
rapports d'excès : $\frac{E(m_1)}{E(b-y)} = -0.337$; $\frac{E(c_1)}{E(b-y)} = 0.193$

Dérougissement

On utilise la calibration de Hilditch *et al.* (1983), [Hilditch *et al.*, 1983], en prenant avant la première itération :

$$(b-y)_0 \approx 4.2608[m_1]^2 - 0.5392[m_1] - 0.0235$$

puis en itérant sur

$$\left| \begin{array}{rcl} m_0 & \approx & m_1 + \frac{E(m_1)}{E(b-y)} E(b-y) \\ (b-y)_0 & = & 14.0881 m_0^2 - 3.36225 m_0 + 0.175709 \end{array} \right.$$

Cette procédure ayant tendance à sur-corriger du rougissement les étoiles avec un m_1 important, on utilise le dérougissement du groupe T1 quand $(b-y)_0 > 0.04$ [Figueras *et al.*, 1991].

Magnitude absolue

On adopte la méthode de Strömgren (1966), en introduisant les indicateurs a et r définis ci-dessus :

$$M_V = 1.5 + 6.0a - 17.0r$$

Erreur sur M_V : $\langle \sigma_{M_V} \rangle \approx 0.2$

2.4.5 Groupe tardif, T1

Domaine de validité : types spectraux A3-F0, classes de luminosité III-IV-V, étoiles Am
indicateur de température : β (et $(b-y)_0$)
indicateur de luminosité : c_0
rapports d'excès : $\frac{E(m_1)}{E(b-y)} = -0.338$; $\frac{E(c_1)}{E(b-y)} = 0.190$

Dérougissement

La calibration de $(b-y)_0$ est donnée par Crawford (1979) :

$$(b-y)_0 = 2.946 - 1.0\beta - 0.1\delta c_0 - K$$

où $K = 0.25\delta m_0$ si $\delta m_0 < 0.$, et 0 sinon.

Magnitude absolue

$M_V^{\rm ZAMS}(\beta)$ est la valeur de la magnitude absolue à la ZAMS, tabulée en fonction de β dans Crawford (1979), Table I. La magnitude absolue de l'étoile est calculée à partir de $M_V^{\rm ZAMS}(\beta)$, corrigée d'un facteur d'évolution :

$$M_V = M_V^{\rm ZAMS}(\beta) - 9\delta c_0$$

Les étoiles Am ont été étudiées par [Guthrie, 1987] qui dérive un nouveau $\delta {c_0}'$ tenant compte des effets de métallicité et de rotation :

$$\delta {c_0}' = \delta c_0 - 1.2\delta m_0 - 1.1\,\mathrm{x}\,10^{-6}(V \sin i)^2$$

$$M_V = M_V^{\rm ZAMS}(\beta) - (9.1\delta {c_0}' + 0.1)$$

Figueras *et al.* (1991) appliquent la deuxième procédure à *toutes* les étoiles de ce groupe ; ceci peut avoir pour effet d'augmenter la magnitude absolue des étoiles non Am de plusieurs dizièmes de magnitudes. Nous avons donc utilisé la procédure de Crawford pour les étoiles avec $\delta m_0 \geq 0.01$ et celle de Guthrie quand $\delta m_0 < 0.01$. Dans ce dernier cas, contrairement à Figueras *et al.*, lorsque $V \sin i$ est inconnu, nous ne le prenons pas nul (ce qui reviendrait à biaiser le résultat de la calibration, quoique l'effet de la rotation ne soit pas en général trop important) mais égal à la valeur moyenne (0.49) des $V \sin i$ de l'échantillon qui a servi à établir la calibration.
Erreur sur M_V : $\langle \sigma_{M_V} \rangle \approx 0.2$

2.4.6 Groupe tardif, T2

Domaine de validité : types spectraux F0-G2, classes III-IV-V, populations I, II intermédiaire, (II extrême ?)
indicateur de température : β (et $(b-y)_0$)
indicateur de luminosité : c_0 (et $(b-y)_0$)
rapports d'excès : $\frac{E(m_1)}{E(b-y)} = -0.331$; $\frac{E(c_1)}{E(b-y)} = 0.184$

Dérougissement

La calibration utilisée est celle d'Olsen (1988), où $\Delta\beta = 2.72 - \beta$:

$$(b-y)_0 = 0.217 + 1.34\Delta\beta + 1.6\Delta\beta^2 + C\delta c_0 - D$$

où

$$C = 4.9\Delta\beta + 32.2\delta m_0 - 262{\delta m_0}^2 - 1.31$$

C étant borné par $C_{\rm inf} \leq C \leq 1.6\Delta\beta$ avec $C_{\rm inf} = 0.13$ si $\delta m_0 > 0.08$, $C_{\rm inf} = -0.05$ sinon

$$\text{et } D = \left| \begin{array}{ll} (0.16 + 4.5\delta m_0 + 3.5\Delta\beta)\delta m_0 & \text{si} \quad \delta m_0 > 0.06 \\ 0.24\delta m_0 + 0.035 & \text{si} \quad \delta m_0 \leq 0.06 \end{array} \right.$$

Compte-tenu de la comparaison effectuée par Nissen [Schuster *et al.*, 1989], on soustrait 0.01 à $(b-y)_0$ si $\delta m_0 > 0.135$ pour mieux tenir compte des étoiles déficientes en métaux.

Magnitude absolue

$M_V^{\rm ZAMS}(\beta)$ est interpolée dans Crawford (1975), Table I, et la magnitude absolue est calculée par :

$$M_V = M_V^{\rm ZAMS}(\beta) - (9 + 20(2.72 - \beta))\delta c_0$$

Le dernier terme étant la correction d'évolution (due au fait que deux étoiles de même température n'ont pas la même magnitude absolue) trouvée par Crawford pour les étoiles de ce groupe.
Erreur sur M_V : $\langle \sigma_{M_V} \rangle \approx 0.25$

2.4.7 Groupe tardif, T3

Domaine de validité : types spectraux G0-M2, classe de luminosité V
indicateur de température : $(b-y)_0$
indicateur de luminosité : c_0
rapports d'excès : $\frac{E(m_1)}{E(b-y)} = -0.330$; $\frac{E(c_1)}{E(b-y)} = 0.180$

Dérougissement

Les procédures (dérougissement et estimation de la magnitude absolue) découlent des calibrations préliminaires de Olsen (1984). Pour dérougir, on utilise la linéarisation de Moon (1985) :

$$(b-y)_0 = \left| \begin{array}{ll} \frac{(u-b)+0.8975}{5.8651} & \text{si} \quad (b-y)_0 < 0.65 \\ \frac{0.6585-c_0}{0.7875} & \text{si} \quad (b-y)_0 \in [0.65, 0.79[\\ \frac{0.3645+c_0}{0.5126} & \text{si} \quad (b-y)_0 \geq 0.79 \end{array} \right.$$

Après avoir noté que cette linéarisation n'était pas adéquate (dans le sens où elle créait des excès de couleur négatifs) pour les plus petits $(b-y)_0$, on utilise :

$$(b-y)_0 = -1.8719 + 2.5826(u-b) - 0.7165(u-b)^2 \text{ si} (b-y)_0 < 0.45$$

Magnitude absolue

$M_V^{\rm ZAMS}(b-y)$, $m_0^{\rm Hyades}(b-y)_0$ et $c_0^{\rm ZAMS}(b-y)_0$ sont interpolés dans Olsen (1984), Table VI :

$$M_V = M_V^{\rm ZAMS}(b-y) - f\delta c_0 + 3.2\delta m_0 - 0.07$$

$$\text{avec } f = \left| \begin{array}{ll} 10 - 80((b-y)_0 - 0.380) & \text{si} \quad (b-y)_0 <= 0.505 \\ 0. & \text{si} \quad (b-y)_0 > 0.505 \end{array} \right.$$

Erreur sur M_V : $\langle \sigma_{M_V} \rangle \approx 0.29$

2.4.8 Groupe des étoiles supergéantes B

Domaine de validité : types spectraux B0-B9, classes de luminosité Ia-Iab-Ib-II
indicateur de température : c_0 (et $(u-b)_0$)
indicateur de luminosité : β
rapports d'excès : $\frac{E(m_1)}{E(b-y)} = -0.329$; $\frac{E(c_1)}{E(b-y)} = 0.190$

Dérougissement

Nous utilisons la calibration de Zhang (1983). Néanmoins, elle présente le désavantage d'être dépendante de la classe de luminosité de l'étoile. Sans prendre le parti d'utiliser des calibrations basées uniquement sur les données photométriques (on utilise bien le critère de Moon, §2.4.1), il est préférable de ne pas faire intervenir ici la classe de luminosité, ne serait-ce qu'en cas d'erreur de classification. Nous avons donc modifié la calibration de Zhang pour qu'elle soit valable quelle que soit la classe de luminosité entre Ia et II. Pour cela, on remarque que la solution

$$c_0 = 0.7(u-b)_0$$

rentre parfaitement dans les barres d'erreurs des régressions trouvées par Zhang dans les différentes classes. Une fois cette première estimation de c_0 obtenue, on détermine une classe de luminosité à partir de c_0 et β, ce qui permet d'utiliser les régressions de Zhang sans prendre en compte le type MK.

Magnitude absolue

Une fois c_0 connu, $\beta_{\rm ZAMS}$ (valeur de β pour les étoiles sur la séquence principale d'âge zéro) est interpolé dans Crawford (1978), Table I, puis $M_V(\beta_{ZAMS})$ interpolé dans Crawford (1978), Table V. On calcule ensuite $\Delta\beta = \beta_{ZAMS} - \beta$ et enfin ΔM_V, correction d'évolution, est interpolé en fonction de $\Delta\beta$ dans Zhang (1983), Table VIII. La magnitude absolue est alors donnée par :

$$M_V = M_V^{\rm ZAMS}(\beta) - \Delta M_V$$

Erreur sur M_V : $\langle\sigma_{M_V}\rangle \approx 0.4$

2.4.9 Groupe des étoiles supergéantes F et G

Domaine de validité : types spectraux F5-G5, classes de luminosité Ia-Iab-Ib-II
indicateur de température : m_0 (et $(b-y)_0$)
indicateur de luminosité : β
rapports d'excès : $\frac{E(m_1)}{E(b-y)} = -0.33$; $\frac{E(c_1)}{E(b-y)} = 0.16$

Dérougissement

Deux calibrations récentes existent : compte-tenu de l'existence de certaines erreurs systématiques qui se trouvent dans Arellano *et al.* (1990), la calibration de Gray (1991) a été préférée :

$$(b-y)_0 = 1.004[m_1] - 0.139\Delta m_1 - 0.436{\Delta m_1}^2$$

où Δm_1 est l'écart de m_1 au tableau standard donné par Gray.

Magnitude absolue

On utilise ici la procédure de Arellano *et al.* (1990) :

$$M_V = 109.2 - 87.9(b-y)_0 - 42.4\beta - 152.6[c_1] + 3.8[c_1](b-y)_0 + 56.3\beta[c_1] + 30.3\beta(b-y)_0$$

Erreur sur M_V : $\langle\sigma_{M_V}\rangle \approx 0.38$

Chapitre 3

Modélisation de l'extinction interstellaire au voisinage solaire

3.1 Objet de l'étude de l'extinction

La matière dont est rempli l'Univers n'est pas seulement composée d'étoiles : le milieu interstellaire est constitué de nuages de gaz atomique (H I, H II) ou moléculaire (H_2, CO principalement) et de poussières. Alors que le milieu interstellaire ne contenait initialement que de l'hydrogène et de l'hélium, les étoiles perdant lentement de la masse, voire explosant pour les supernovae, restituent les éléments qu'elles ont synthétisé dans leur noyau. C'est au sein des nuages interstellaires que se sont formées et se forment les étoiles, enrichies en éléments plus lourds par les générations précédentes.

Autant dire que la connaissance de la matière interstellaire est particulièrement importante pour l'étude de la structure et de l'évolution de notre Galaxie. De plus, la matière interstellaire constitue une fraction importante de la masse de notre Galaxie [Gliese *et al.*, 1986] et a donc une implication sur sa dynamique.

Paradoxalement, ce ne sont justement pas ces propriétés importantes de la matière interstellaire qui motivent ce chapitre, mais bien plutôt la gêne que son existence implique : la matière interstellaire absorbe, diffuse et polarise la lumière des étoiles, et de plus, de façon sélective : plus la longueur d'onde diminue, plus l'absorption augmente, et donc plus une étoile dont la lumière est absorbée apparaîtra rougie.

Par conséquent, non seulement les distances d'étoiles affectées par le phénomène de l'absorption seront systématiquement surestimées, mais la couleur de ces étoiles sera également plus rouge qu'en réalité, l'excès de couleur étant dû à l'absorption sélective.

Ce dernier point n'était pas sans importance en ce qui concernait le Catalogue d'Entrée d'Hipparcos, le bon fonctionnement du satellite nécessitant en effet une connaissance *a priori* de la couleur des étoiles à observer.

3.2 Un modèle empirique de l'extinction interstellaire

Bien avant le lancement du satellite, il devint vite évident qu'il n'y aurait pas de la photométrie photoélectrique pour toutes les étoiles du Catalogue : sur l'ensemble de

la base de données INCA (214 000 étoiles), 130 000 étoiles avaient un indice de couleur $(B-V)$ dont la précision était moins bonne que 0.3 magnitudes.

Pour obtenir une meilleure précision sur cette couleur «rougie» $(B-V) = (B-V)_0 + E(B-V)$, nous avons estimé la couleur intrinsèque $(B-V)_0$ des étoiles considérées en fonction de leur type spectral et de leur classe de luminosité, et l'excès de couleur $E(B-V)$ à l'aide d'un modèle d'extinction.

Bien que l'extinction interstellaire dans notre Galaxie (et dans d'autres galaxies proches) soit un sujet souvent étudié (voir par exemple Guarinos (1991) pour un panorama complet), il n'existait pas jusqu'à présent de cartographie complète de l'extinction (notamment à moyenne et haute latitude galactique), ni de modèle analytique à la fois réaliste et informatisé.

Le nombre d'étoiles à traiter dans notre cas prohibant tout traitement manuel, ce modèle devait donc être créé.

Parallèllement, Guarinos (1991) étudiait également l'extinction interstellaire de façon détaillée dans le voisinage solaire. La différence essentielle entre son étude et celle-ci provient du fait qu'il cherchait à cartographier l'extinction, alors qu'ici, on cherchait plutôt à la modéliser afin de pouvoir l'évaluer aisément pour plus de 100 000 étoiles.

Par conséquent, on a d'abord calibré l'absorption dans le visible en fonction de la distance au soleil et des coordonnées galactiques, à partir des étoiles possédant de la photométrie photoélectrique et un bon type spectral MK.

Ceci fut réalisé en découpant le ciel en un nombre de cases tel que le nombre d'étoiles dans chacune soit statistiquement suffisant pour pouvoir ajuster à l'absorption observée un modèle quadratique jusqu'à une certaine hauteur au-dessus du plan galactique, et linéaire ensuite. On en déduit un excès de couleur non biaisé en fonction du type spectral.

On montre ensuite que cette calibration est en accord avec l'absorption estimée dans le système photométrique de Walraven dans une petite zone du ciel connue pour son absorption, d'une part, et avec l'excès de couleur tel qu'on peut le déduire de la distribution d'hydrogène neutre dans notre galaxie à moyenne et haute latitude galactique, d'autre part.

Enfin, n'oublions pas que l'objectif initial était d'obtenir une couleur pour des étoiles n'en possédant pas, ou trop imprécise : on démontre *a posteriori*, en utilisant les premières mesures photométriques du satellite, l'amélioration procurée par le modèle. L'article ci-joint [Arenou *et al.*, 1992], décrit le modèle dans le détail.

On peut trouver également, page 107, une autre comparaison entre les couleurs déduites par cette méthode et les couleurs obtenues ensuite par les repéreurs d'étoiles du satellite, cette fois-ci en magnitudes Tycho (il faut diviser l'indice de couleur Tycho par 1.2 environ pour obtenir l'indice Johnson, sauf pour les étoiles les plus rouges [Grenon, 1989]).

Bien qu'il reste de petits effets en fonction des coordonnées galactiques, cette comparaison est globalement satisfaisante.

Astron. Astrophys. 258, 104–111 (1992)

ASTRONOMY AND ASTROPHYSICS

16. A tridimensional model of the galactic interstellar extinction

F. Arenou[1], **M. Grenon**[2], **A. Gómez**[1]

[1] URA 335 du CNRS, Observatoire de Meudon, DASGAL, F-92195 Meudon, France
[2] Observatoire de Genève, CH-1290 Sauverny, Switzerland

Received July 18, 1991; accepted January 24, 1992

Abstract. The distribution of interstellar extinction has been mapped over the whole sky, using all available spectral and photometric data. The colour excess distribution is modelled as a function of galactic latitude, longitude and distance within about 1 kpc from the Sun. The model was used to predict the reddened Tycho and Johnson ($B - V$) colours, with the associated accuracy, from HD or MK spectral type and one magnitude (B or V), for stars without photoelectric photometry. This model is of special interest at intermediate and high galactic latitudes where colour excesses cannot be obtained from early type stars.

Key words: interstellar extinction – colour excess – intrinsic colours – photometry – Hipparcos

1. Introduction

In order to achieve the announced accuracy on the astrometric parameters expected from Hipparcos, photometric information was required to allocate the adequate observing time to each programme star. This time is a function of the magnitude called H_p – the star magnitude measured by the wide-band Hipparcos detector – which had to be estimated with an uncertainty smaller than 0.5 mag. The consequence of the large bandwith of the H_p magnitude is its dependence not only on the visual apparent brigthness of a star, but also on its colour ($B - V$) (Grenon 1989). In spite of the enormous effort made within the INCA Consortium to obtain new ground-based photoelectric measurements (Grenon 1989), about half of Hipparcos stars has no photoelectric photometry and the available B and V magnitudes come from inhomogeneous sources.

The main mission detector is potentially able to produce very high accurate photometry of programme stars provided that ageing effects on the optics are properly corrected through an on-orbit calibration. These effects are chromatic and the reduction to a standard system is possible only through the knowledge of accurate colours. Ideally the precision on ($B - V$) should be better than 0.04 mag in order to maintain a precision on H_p better than 0.003 mag through the mission.

The colours of stars without photoelectric photometry are far too inaccurate when deduced from blue photographs and photovisual magnitudes both for the H_p prediction and the photometric calibration. A serious improvement of the precision on colours may be expected by using the spectral classification, provided that the colour excess is sufficiently well mapped as a function of the galactic coordinates and distances.

The distribution of colour excess and interstellar reddening material in the solar neighborhood has been studied by several authors, notably by Lucke (1978), FitzGerald (1968) and Neckel & Klare (1980). These previous results were not useful for our purpose: they are presented in a graphical form and give no or poor information about the mean colour excesses outside the Galactic plane. Using all available spectral and photoelectric data contained in the INCA Database (Gómez et al. 1989; Turon et al. 1991), an analytical three dimensional model of the galactic interstellar extinction has been built and compared with results found in the literature.

For each spectral type, and luminosity class in case of MK classification, the colours deduced from the model were compared with photoelectric data. Systematic differences as functions of the colour excess were used to correct the intrinsic colours and to tune the model.

The method was applied to about 130 000 stars without photoelectric photometry, among which 60 000 stars are in the Input Catalogue. The validity of the procedure was finally checked by comparing calculated colours with the preliminary colours obtained from the star mappers of the Hipparcos satellite.

2. The data

The present study used all the stars contained in the INCA Database having at the same time spectral types (MK or HD) and B, V photometry. The choice of Johnson B, V photometry, instead of Geneva, Strömgren or Walraven photometry, is simply due to the large size of the final available sample.

The INCA Database contains the 215 000 stars proposed by the astronomical community to be observed by Hipparcos and a basic list of bright stars ('survey') required for satellite operation and data reduction. Some catalogues established for astrophysical purposes are highly represented, e.g. the Catalogue of extinction data (Neckel et al. 1980) and the Michigan Spectral Survey (Houk & Cowley 1975; Houk 1978, 1982).

During the preparation of the Hipparcos Input Catalogue, all the spectroscopic data available in SIMBAD (the Database of the Strasbourg Stellar Data Center) (Egret 1985) were incorporated into the INCA Database. Spectral types were mainly obtained from SIMBAD plus those published more recently in the Fourth volume of the Michigan Spectral Survey (Houk 1988). Other classifications published in the literature were also included. Finally, about 77 000 stars had a MK spectral type and 94 000 stars a HD spectral type.

Send offprint requests to: F. Arenou

Table 1. Number of stars with MK spectral type (26 052 stars)

Class \ Sp.	O	B	A	F	G	K	M
I	15	455	67	75	65	47	12
II	9	846	81	113	146	167	14
III	23	1554	505	463	1269	3061	341
IV-V	81	4340	2873	5661	2502	1063	125
VI	7	2	1	45	19	1	4

Table 2. Number of stars with HD spectral type (16 569 stars)

Spectral type	Number	Spectral type	Number
dO	28	dG	1211
dB	1774	gG	984
dA	4961	gK	2066
dF	5407	gM	138

Photoelectric photometry published in the seven major systems was extracted from the Database maintained at Lausanne Institute of Astronomy and completed with the newly observed measurements obtained in the framework of the INCA Consortium (Grenon 1989). Through adequate transformation equations, all data were expressed in terms of Johnson B, V, Tycho B_T, V_T and Hipparcos H_p magnitudes. A total of 58 000 colours were available for the model construction.

The final sample contains neither variable stars nor doubles closer than 10 arcsec. No special selection to limiting apparent magnitude was made, but the INCA Database does not contain stars fainter than V = 13 mag. Tables 1 and 2 give, respectively, the number of stars with MK and HD spectral types having photoelectric photometry.

Stars having MK spectral types were used to build the interstellar extinction model (see Sect.3) while the whole sample was used to estimate the colour index $(B-V)$ (see Sect.4). In the case of stars with HD classification, the following luminosity classes have been adopted: stars earlier than type G5 were considered as dwarfs (V), otherwise as giants (III).

3. The galactic interstellar extinction model

Previous work on this subject (op.cit.) clearly showed that the galactic interstellar extinction is a function of the position in the Galaxy (r: heliocentric distance and l_{II} and b_{II}: galactic coordinates). Analytical three dimensional models do not exist, probably due to the difficulty of modelling with small samples the reddening material distribution up to about 2 kpc. It is well known that the Parenago formula (Parenago 1940) as well as the cosecant law do not reproduce the observed interstellar monochromatic extinction. Using a large sample of stars having MK spectral types and photoelectric photometry in the spectral range O to F8 (about 17 000 stars), which constitutes a sub-sample of the data described in Sect. 2, the sky has been divided in cells and for each cell an analytic expression of the interstellar monochromatic extinction at V-magnitude $A_V(r, l_{II}, b_{II})$ has been obtained.

For each star the distance r and the visual extinction A_V have been computed using the well-known relations:

$$E_{B-V} = (B-V)_{ph} - (B-V)_0 \quad (1)$$

$$R = 3.30 + 0.28(B-V)_0 + 0.04E_{B-V} \quad (2)$$

$$A_V = R\ E_{B-V} \quad (3)$$

$$r = 10^{(V-M_V+5-A_V)/5} \quad (4)$$

where $(B-V)_{ph}$ is the photoelectric colour index, M_V and $(B-V)_0$ are the absolute magnitude and the intrinsic colour index respectively. The factor R depends upon the amount of reddening, the energy distribution in the stellar spectrum and the position in the Milky Way. The M_V and $(B-V)_0$ calibrations versus spectral type and luminosity class and the R formula quoted before were taken from Schmidt-Kaler (1982). From Eq. (1)-(3), the uncertainty of A_V is about 0.15 mag, depending primarily of the error on $(B-V)_0$. Assuming an uncertainty on M_V of roughly 0.5 mag, the relative error on the distance is 25%.

The sky has been divided in 199 cells in galactic coordinates and their boundaries were chosen following the Fig. 3 of Lucke (1978). Our goal was to derive an analytic expression for A_V, as a function of the distance and the galactic coordinates, representative of the general trend of the interstellar extinction and not of the local irregularities. The graphics shown by Neckel & Klare (1980) suggest that A_V could be represented in a first approximation by a quadratic relationship. In each cell, the following quadratic relation has been adopted up to a distance r_0:

$$A_V(r, l_{II}, b_{II}) = \alpha(l_{II}, b_{II})r + \beta(l_{II}, b_{II})r^2, \qquad \text{if } r \leq r_0 \quad (5)$$

where r_0 represents the distance limit of the absorbing layers[1] . As a first estimate $r_0 = r_{thick} = 0.2/\sin|b_{II}|$ was taken, 0.2 kpc being the adopted half thickness of the galactic plane. As it is shown later, the definition of r_0 depends on the studied region. For the regions $|b_{II}| < 5°$, r_{thick} was fixed equal to 2 kpc. Out of the absorbed regions A_V should remain constant and identically to $A_V(r_0, l_{II}, b_{II})$. In the case of cells including a large part of the galactic plane, $|b_{II}| < 15°$ regions, a better representation of the interstellar extinction was obtained adopting beyond r_0 the following linear regression relation:

$$A_V(r, l_{II}, b_{II}) = A_V(r_0, l_{II}, b_{II}) + \gamma(l_{II}, b_{II})(r - r_0) \quad (5bis)$$

In order to adopt in each cell the best r_0 value, some constraints on A_V have been taken into account. For instance, A_V cannot diminish with r. Figs. 1a and 1b show that A_V arrives at a maximum value for $r = r_{dec}$ and then decreases. This decreasing is spurious, and is related to the presence of higher interstellar extinction and to the lack of fainter stars in the sample. On the other hand, A_V has not to go beyond the largest observed value in a given direction, corresponding to a distance $r = r_{larg}$ in Figs. 1a and 1b. Moreover, this last maximum observed value has to be compared with the results found in the literature. We adopted:

$A_{max} = 0.1$, if $60° \leq |b_{II}| < 90°$
$A_{max} = 1.2$, if $45° \leq |b_{II}| < 60°$
$A_{max} = 3.0$, if $|b_{II}| < 45°$

The corresponding distance called r_{max} is absent from Figs. 1a & 1b because it is reached at r = 9 kpc in the example shown in Fig. 1a and never reached in the other case.

Consequently, the r_0 value used in formulae (5) and (5bis) corresponds to the smallest A_V value between $A_V(r_{thick})$, $A_V(r_{dec})$, $A_V(r_{larg})$ and $A_V(r_{max})$, r_0 being $\leq r_{dec}$. The regression coefficients α, β and γ have been determined using the least square method

[1] If the parabola decreases at the beginning (from r = 0) and then increases, A_V takes negative values in a distance interval. In these few cases, we adopted $A_V \equiv 0$ in this interval.

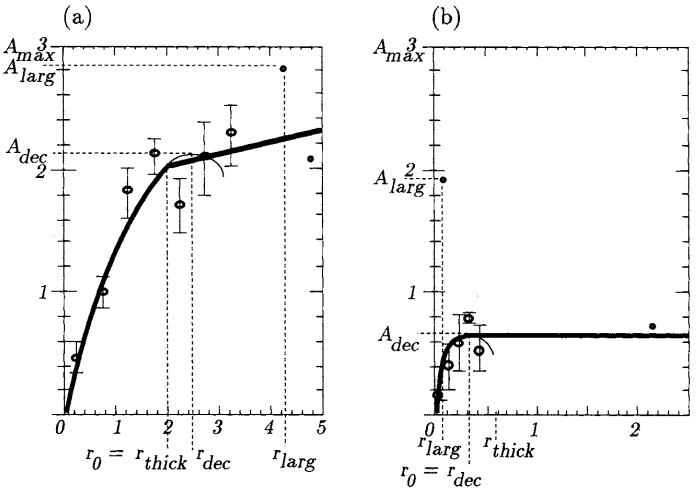


Fig. 1. Absorption (mag) as a function of distance (kpc): Example in the galactic plane ($180° \leq l_{II} < 190°$, $-5° \leq b_{II} < 5°$, bins of 500pc) (a) and at intermediate galactic latitude ($20° \leq l_{II} < 40°$, $15° \leq b_{II} < 30°$, bins of 100pc) (b)

and are given for each cell in Appendix; the r_0 values are also indicated. The last column gives the relative error $\frac{\sigma_{A_V}}{A_V}$, expressed in percentages. For each star, the error on the computed $A_V(r)$ has been estimated by $\sqrt{0.15^2 + (\frac{\sigma_{A_V}}{A_V} A_V(r))^2}$, where 0.15 mag is the assumed uncertainty of A_V calculated from Eq. (1)-(3). High relative errors (> 100%) may exist in cells containing a few number of stars or where the obtained extinction is small compared to the size of the computed error. This last case could be the result of irregularities of the distribution of the interstellar matter. The average relative error on the whole sky is about 35%.

One may ask how well our model compares with previous studies in small regions. Due to the relatively large sizes of the cells, this comparison has little meaning. The model gives the general trend of the interstellar extinction in each region and does not take into account the irregularities of the distribution of the absorbing material. On the other hand, we expect in general, for small distances ($r \leq$ 200pc), a slightly overestimation of A_V. This fact is shown in Fig. 2 which gives the comparison between the visual extinction in the Johnson UBV-System calculated using Walraven photometry for the region of Upper-Centaurus Lupus subgroup (de Geus et al. 1989) and our model's values. However, the overestimation of A_V can be considered negligible if we take into account the estimated errors on A_V. We have also compared our results with those of Neckel and Klare (1980) : for regions having $A_V < 0.5$ mag the agreement is very satisfactory, otherwise it depends on the irregularities of the distribution of the interstellar matter in the cell.

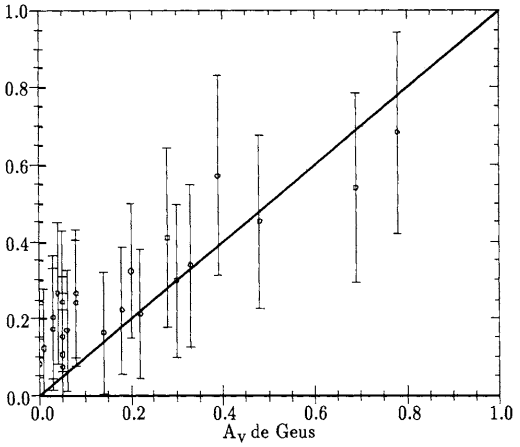


Fig. 2. Estimated $A_V(r)$ versus A_V from de Geus et al. (1989) in the Upper-Centaurus Lupus region.

4. $(B - V)$ colour index estimation

For each of the stars given in Tables 1 and 2 the (B-V) colour index has been estimated using the interstellar extinction model described in section 3. This colour index, called $(B - V)_{red}$, has been compared to $(B - V)_{ph}$. The distribution of the differences $\delta_{B-V} = (B - V)_{red} - (B - V)_{ph}$ has been evaluated as a function of the spectral type and the luminosity class. The $\overline{\delta}_{B-V}$ values in four bins of E_{B-V} are given in Tables 3 and 4, corresponding to stars having MK and HD classification respectively. The stars having $E_{B-V} > 0.35$ or $|\delta_{B-V}| > 0.75$ mag were not included. We expect to obtain $\overline{\delta}_{B-V}$ values approximatively null. In general, they increase with the colour excess and for late spectral types. These figures are the systematic corrections to be applied to the $(B - V)_{red}$ colour index in order to estimate the corresponding $(B - V)$ colour index. The rms error σ_{B-V}, given in the Tables, is the error on the colour index. It includes the errors in the distance determinations, in the adopted intrinsic colour indexes and in the spectroscopic and photometric data. These errors obtained on the estimated colour indexes are mostly smaller than 0.15 mag for MK stars and 0.2 mag for HD stars. These results allowed us to apply the method to stars without photoelectric photometry.

To summarize the colour estimation process, given a star with a V magnitude (or a B magnitude, as $V = B-(B-V)_0-E_{B-V}$) and M_V, $(B-V)_0$ derived from its spectral type, Eqs. (2)-(4) are iterated, beginning with $A_V = 0$ and then with $A_V(r)$ given in appendix. This

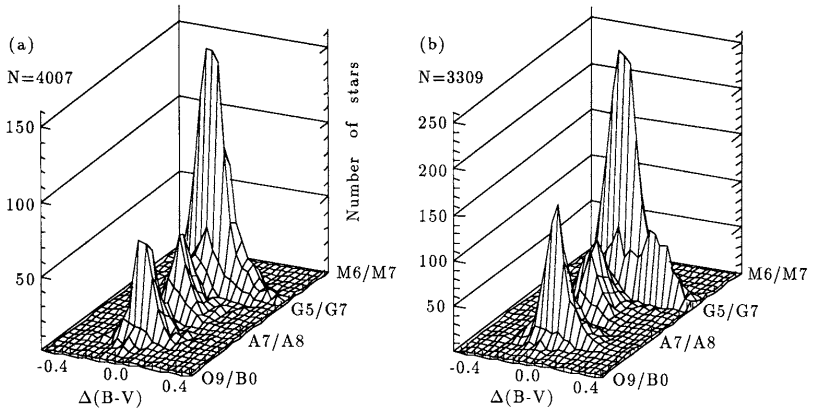


Fig. 3. Distribution of differences between $(B-V)_{NDAC}$ and $(B-V)_{INCA}$ and spectral types, for stars with MK spectral type (a) and for stars with HD spectral type (b)

Table 3. Corrections on $(B-V)$ and precision on $(B-V)$ as a function of spectral type, luminosity class and bins of colour excess, for stars with MK spectral type

	$E_{B-V} \leq .05$		].05,.15]		].15,.25]		].25,.35]	
	$\overline{\delta}_{B-V}$	σ	$\overline{\delta}_{B-V}$	σ	$\overline{\delta}_{B-V}$	σ	$\overline{\delta}_{B-V}$	σ
B I	.030	.15	.025	.110	.020	.164	.030	.170
A I	.050	.15	.030	.054	.000	.165	.000	.111
F I	-.060	.133	-.060	.162	-.060	.080	-.060	.115
G I	.040	.203	.030	.081	.040	.085	.090	.164
K I	.080	.081	.030	.168	.030	.137	.090	.237
M I	-.020	.15	-.020	.15	-.020	.15	-.020	.106
B II	.010	.057	.040	.093	.060	.104	.040	.156
A II	-.070	.102	-.070	.142	-.050	.138	-.020	.165
F II	-.010	.060	-.020	.118	-.010	.111	.050	.169
G II	.030	.145	.020	.141	.060	.122	.160	.114
K II	.020	.117	.030	.147	.090	.170	.200	.25
M II	.080	.10	.070	.088	.120	.20	.240	.25
B III	.005	.054	.015	.081	.015	.104	.015	.124
A III	-.015	.069	-.020	.099	-.020	.15	-.020	.152
F III	.010	.067	.010	.107	.020	.15	.030	.25
G III	.020	.15	.030	.142	.030	.156	.030	.377
K III	-.020	.12	-.010	.134	.000	.214	.000	.232
M III	.025	.063	.025	.092	.025	.115	.025	.198
BIV,V	.005	.055	.010	.064	.005	.095	.000	.128
AIV,V	-.010	.061	.000	.10	-.010	.106	-.020	.122
FIV,V	-.005	.10	-.005	.071	.000	.15	.000	.25
GIV,V	.000	.096	.020	.133	.040	.118	.060	.20
KIV,V	-.040	.150	-.040	.239	-.040	.25	-.040	.25
MIV,V	.040	.133	.040	.20	.040	.25	.040	.25
F VI	.010	.053	.010	.035	.000		.000	
G VI	.070	.110	.000		.000		.000	

gives E_{B-V}, and finally the colour index $(B-V) = (B-V)_0 + E_{B-V} - \overline{\delta}_{B-V}$ and its associated error.

Table 4. Corrections on $(B-V)$ and precision on $(B-V)$ as a function of spectral type and bins of colour excess, for stars with HD spectral type

	$E_{B-V} \leq .05$		].05,.15]		].15,.25]		].25,.35]	
	$\overline{\delta}_{B-V}$	σ	$\overline{\delta}_{B-V}$	σ	$\overline{\delta}_{B-V}$	σ	$\overline{\delta}_{B-V}$	σ
dB	-.025	.093	.005	.1	-.005	.139	-.035	.180
dA	-.080	.15	-.020	.14	.020	.141	.000	.155
dF	-.025	.072	.005	.107	.010	.1	.000	.15
dG	-.030	.126	.000	.155	.090	.086	.220	.20
gG	.070	.195	.090	.198	.220	.209	.330	.178
gK	-.040	.197	.020	.218	.080	.243	.260	.317
gM	.010	.082	.010	.143	.010	.20	.010	.170

5. Comparison with the preliminary data obtained with the Hipparcos star mappers

Even if the model described above is self-consistent, it is worth comparing the colours predicted within the INCA consortium, $(B-V)_{INCA}$, with those obtained from the preliminary measurements coming from the NDAC reduction consortium (van Leeuwen et al. 1992). These data are the result of the analysis of 12 weeks of observation from the Hipparcos star mappers; as they are B_T and V_T magnitudes in the Tycho bands (Grenon et al. 1992), they must beforehand be transformed to Johnson magnitudes. Moreover, for a safe comparison, we kept only the stars with $|V_{NDAC} - V_{INCA}| < .05$ in order to reduce the errors in the estimation of the distance, and thus in the reddening. Figure 3 shows the distribution of $\Delta(B-V) = (B-V)_{NDAC} - (B-V)_{INCA}$ separately for the stars with MK spectral type and HD spectral type. This figure provides the distributions of the spectral types of the stars in the samples as well as the distribution of $\Delta(B-V)$ for each spectral type; each bin is 0.04 in mag and about 2 in spectral subtypes (from O9/B0 to M6/M7). As we can see, the distributions of $\Delta(B-V)$ are centered near 0., with mean/standard deviation of $-.008/0.091$ for stars with MK spectral type and 0.02/0.17 for stars with HD spectral type. The distribution is clearly asymmetric for HD spectral type G5. This type contains a mixture of giants and dwarfs with $(B-V)_0 = 0.88$ and 0.68 respectively. The ratio dwarfs/giants is a function of distance and b_{II}; the distribution tail is due to giants. For the other stars, and this is

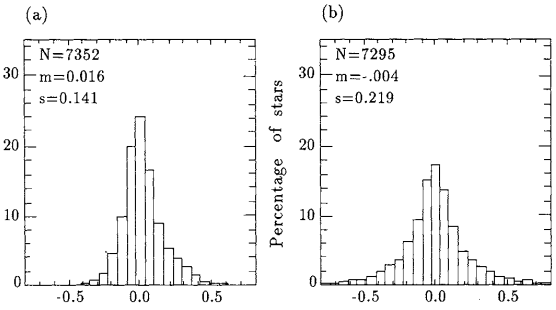


Fig. 4. Distribution of differences between $(B-V)_{NDAC}$ colour index and computed colours (a) and of differences between $(B-V)_{NDAC}$ and colours deduced from B and V magnitudes coming from heterogeneous sources (b)

also true for stars with MK spectral type, we may still notice a small tail towards $\Delta(B-V) > 0$. This is because the concerned stars are fainter, farther than about 1kpc and the estimated interstellar extinction is probably underestimated. However, all these effects had been accounted for in the $(B-V)$ error bars.

We may now ask: what would be the differences $\Delta(B-V)$ if we had not built the extinction model, that is in using $B-V$ colours with B & V magnitudes coming from heterogeneous sources. The answer is given in Fig. 4, where MK and HD stars are considered together. The number of stars is given as well as the mean and standard deviation of the differences. The slight asymmetry mentioned before is visible in Fig. 4a; however the improvement of colours with the extinction model is perfectly clear.

6. Comparison with extinction from H I distribution

At intermediate galactic latitudes, the total extinction, computed for each cell at $r = r_0$, may be compared to that deduced from surveys of neutral hydrogen (HI) column densities and deep galaxy counts. Maps of E_{B-V} contours by Burstein and Heiles (1982) were used to check both the asymptotic behaviour of the relations given in annex and the amplitude of the scatter due to irregularities of the interstellar matter distribution. In the galactic latitude ranges +15° to +45° and −45° to −15°, reddenings are read on a mesh of 9 to 12 points, depending on the cell size. Since the mapping precision and the reading resolution are about 0.01 to 0.02 mag, in moderately reddened areas the scatter in E_{B-V} is representative of the true irregularity of the interstellar matter distribution. Due to the rather large cell size, the irregularities lead to a ratio $\frac{\sigma_{A_V}}{A_V}$ of about 0.42, a value similar to that computed in well documented areas.

The comparison between the computed total mean colour excess E_{B-V} and those from Burstein and Heiles is given in Fig. 5 for the precited ranges of galactic latitudes. Considering the noise on both colour-excess determinations the agreement is quite satisfactory. Apart from a possible zero–point shift, there is some evidence for the observed E_{B-V} to be systematically larger by about 20% than those computed from HI.

In the north galactic cap, E_{B-V} is directly deduced from HI maps through the relation : $E_{B-V} = 0.485 N_H$ where N_H is given in units of 10^{15} atoms.cm^{-2}. The comparison of excesses for cells with b_{II} from +60° to +90° is given in Fig. 6. The differences between the two approaches are within 0.01 mag on E_{B-V}.

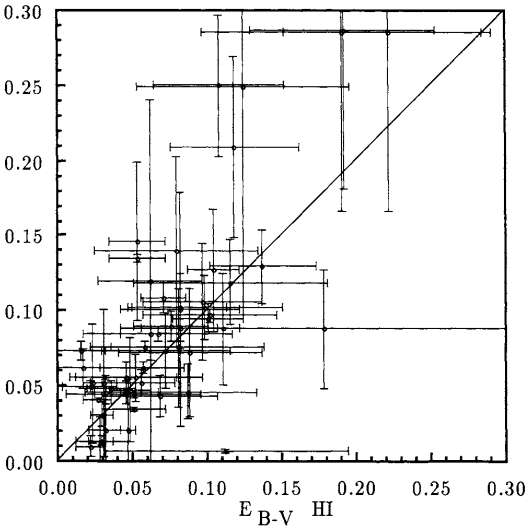


Fig. 5. Comparison between the total colour excesses $E_{B-V_{mod}}$ computed from the extinction model at $r = r_0$, and those read on maps of E_{B-V} built from HI distribution and deep galaxy counts from Burstein and Heiles. The ranges in galactic latitude are −45° to −15° and +15° to +45°.

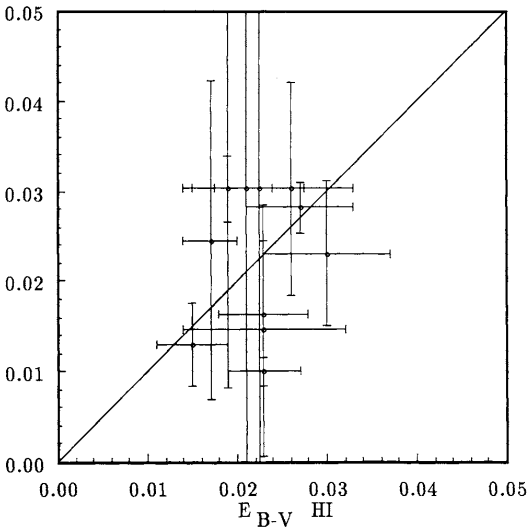


Fig. 6. Comparison between the total colour excesses E_{B-V} from the model and those from HI distribution in the north galactic cap ($b_{II} > 60°$).

7. Conclusion

The primary goal of the model was the production of colours (and ultimately instrumental magnitudes) for stars without photoelectric photometry to be observed by the Hipparcos satellite. Indeed it is of a more general application since it provides a three dimensional distribution of the reddening with a mean accuracy of about 40% for

stars within and outside the dust lane. It complements the existing description of the interstellar matter distribution for the intermediate and high galactic latitudes. This model gives the general trend of the interstellar extinction and does not take into account the local irregularities of the absorbing material; it should be used preferably for investigations of statistical character.

The C or FORTRAN source program may be obtained upon request at MESIOA::ARENOU (SPAN–DECNET) or ARENOU at FRMEU51 (EARN-BITNET).

Acknowledgements. We would like to thank the referee Dr Knude for his useful comments.

References

Burstein D., Heiles C., 1982, Astron. J., 1165
de Geus E.J., de Zeeuw P.T., Lub J., 1989, A&A 216,44
Egret D., 1985, ESA–SP 234, 105
FitzGerald M.P., 1968, AJ 73, 983
Gómez A., Morin D., Arenou F., 1989, ESA–SP 1111, 23
Grenon M., 1989, ESA–SP 1111, 129
Grenon M., Mermilliod M., Mermilliod J.C., 1992, A&A, this issue (paper 13)
Houk N., Cowley A.P., 1975, Michigan Catalog of Two Dimensional Spectral Types for the HD Stars, Vol 1, Ann Arbor
Houk N., 1978, Michigan Catalog of Two Dimensional Spectral Types for the HD Stars, Vol 2, Ann Arbor
Houk N., 1982, Michigan Catalog of Two Dimensional Spectral Types for the HD Stars, Vol 3, Ann Arbor
Houk N., 1988, Michigan Catalog of Two Dimensional Spectral Types for the HD Stars, Vol 4, Ann Arbor
van Leeuwen F. et al., 1992, A&A, this issue (paper 19)
Lucke P.B., 1978, A&A 64, 367
Neckel Th., Klare G., 1980, A&AS 42, 251
Parenago P., 1940, AZh 17,3
Schmidt-Kaler Th., 1982, in Landolt-Börnstein, Springer Verlag, vol. 2, p. 1
Turon C. et al., 1991, Databases & On-line Data in Astronomy, eds. M.A. Albrecht and D. Egret, p. 67

Appendix

Regression parameters for $A_V(r)$ and relative error on $A_V(r)$ by galactic longitude and latitude bins; α and γ in mag.kpc^{-1}, β in mag.kpc^{-2}, r_0 in kpc, $\frac{\sigma_{A_V}}{A_V}$ in %

$-90 \leq b_{II} < -60$	α	β	r_0	γ	$\frac{\sigma_A}{A}$
$0 \leq l_{II} < 29$	2.22534	-6.00212	0.052	0.0	13
$29 \leq l_{II} < 57$	3.35436	-14.74567	0.035	0.0	40
$57 \leq l_{II} < 85$	2.77637	-9.62706	0.042	0.0	15
$85 \leq l_{II} < 110$	4.44190	-19.92097	0.025	0.0	36
$110 \leq l_{II} < 150$	4.46685	-26.07305	0.026	0.0	28
$150 \leq l_{II} < 180$	7.63699	-46.10856	0.014	0.0	38
$180 \leq l_{II} < 210$	2.43412	-8.69913	0.050	0.0	36
$210 \leq l_{II} < 240$	3.34481	-13.93228	0.035	0.0	38
$240 \leq l_{II} < 270$	1.40733	-3.43418	0.091	0.0	30
$270 \leq l_{II} < 300$	1.64466	-3.97380	0.074	0.0	28
$300 \leq l_{II} < 330$	2.12696	-6.05682	0.056	0.0	14
$330 \leq l_{II} < 360$	2.34636	-8.17784	0.052	0.0	16

$-60 \leq b_{II} < -45$	α	β	r_0	γ	$\frac{\sigma_A}{A}$
$0 \leq l_{II} < 30$	2.77060	-9.52310	0.145	0.0	16
$30 \leq l_{II} < 60$	1.96533	-5.63251	0.174	0.0	06
$60 \leq l_{II} < 110$	1.93622	-13.31757	0.073	0.0	26
$110 \leq l_{II} < 180$	1.05414	-2.36540	0.223	0.0	74
$180 \leq l_{II} < 210$	1.39990	-1.35325	0.252	0.0	10
$210 \leq l_{II} < 240$	2.73481	-11.70266	0.117	0.0	08
$240 \leq l_{II} < 270$	2.99784	-11.64272	0.129	0.0	03
$270 \leq l_{II} < 300$	3.23802	-11.63810	0.139	0.0	07
$300 \leq l_{II} < 330$	1.72679	-6.05085	0.143	0.0	07
$330 \leq l_{II} < 360$	1.88890	-5.51861	0.171	0.0	14

$-45 \leq b_{II} < -30$	α	β	r_0	γ	$\frac{\sigma_A}{A}$
$0 \leq l_{II} < 30$	1.98973	-4.86206	0.205	0.0	06
$30 \leq l_{II} < 60$	1.49901	-3.75837	0.199	0.0	16
$60 \leq l_{II} < 90$	0.90091	-1.30459	0.329	0.0	73
$90 \leq l_{II} < 120$	1.94200	-6.26833	0.155	0.0	18
$120 \leq l_{II} < 160$	-0.37804	10.77372	0.210	0.0	100
$160 \leq l_{II} < 200$	-0.15710	5.03190	0.294	0.0	24
$200 \leq l_{II} < 235$	3.20162	-10.59297	0.151	0.0	09
$235 \leq l_{II} < 265$	1.95079	-4.73280	0.206	0.0	21
$265 \leq l_{II} < 300$	1.91200	-4.97640	0.192	0.0	17
$300 \leq l_{II} < 330$	2.50487	-8.63106	0.145	0.0	28
$330 \leq l_{II} < 360$	2.44394	-9.17612	0.133	0.0	07

$-30 \leq b_{II} < -15$	α	β	r_0	γ	$\frac{\sigma_A}{A}$
$0 \leq l_{II} < 20$	2.82440	-4.78217	0.295	0.0	32
$20 \leq l_{II} < 40$	3.84362	-8.04690	0.239	0.0	46
$40 \leq l_{II} < 80$	0.60365	0.07893	0.523	0.0	22
$80 \leq l_{II} < 100$	0.58307	-0.21053	0.523	0.0	53
$100 \leq l_{II} < 120$	2.03861	-4.40843	0.231	0.0	60
$120 \leq l_{II} < 140$	1.14271	-1.35635	0.421	0.0	34
$140 \leq l_{II} < 160$	0.79908	1.48074	0.523	0.0	61
$160 \leq l_{II} < 180$	0.94260	8.16346	0.441	0.0	42
$180 \leq l_{II} < 200$	1.66398	0.26775	0.523	0.0	42
$200 \leq l_{II} < 220$	1.08760	-1.02443	0.523	0.0	45
$220 \leq l_{II} < 240$	1.20087	-2.45407	0.245	0.0	06
$240 \leq l_{II} < 260$	1.13147	-1.87916	0.301	0.0	16
$260 \leq l_{II} < 280$	1.97804	-2.92838	0.338	0.0	21
$280 \leq l_{II} < 300$	1.40086	-1.12403	0.523	0.0	19
$300 \leq l_{II} < 320$	2.06355	-3.68278	0.280	0.0	42
$320 \leq l_{II} < 340$	1.59260	-2.18754	0.364	0.0	15
$340 \leq l_{II} < 360$	1.45589	-1.90598	0.382	0.0	21

-15≤ b_{II} <-5	α	β	r_0	γ	$\frac{\sigma_A}{A}$
0 ≤ l_{II} <10	2.56438	-2.31586	0.554	0.0	37
10 ≤ l_{II} <20	3.24095	-2.78217	0.582	0.0	38
20 ≤ l_{II} <30	2.95627	-2.57422	0.574	0.08336	41
30 ≤ l_{II} <40	1.85158	-0.67716	1.152	0.0	04
40 ≤ l_{II} <50	1.60770	0.35279	0.661	0.0	24
50 ≤ l_{II} <60	0.69920	-0.09146	0.952	0.12839	02
60 ≤ l_{II} <80	1.36189	-1.05290	0.647	0.16258	45
80 ≤ l_{II} <90	0.33179	0.37629	1.152	0.0	62
90 ≤ l_{II} <100	1.70303	-0.75246	1.132	0.0	31
100≤ l_{II} <110	1.97414	-1.59784	0.618	0.12847	35
110≤ l_{II} <120	1.07407	-0.40066	1.152	0.17698	14
120≤ l_{II} <130	1.69495	-1.00071	0.847	0.08567	28
130≤ l_{II} <140	1.51449	-0.08441	0.897	0.0	12
140≤ l_{II} <150	1.87894	-0.73314	1.152	0.0	23
150≤ l_{II} <160	1.43670	0.67706	0.778	0.42624	03
160≤ l_{II} <180	6.84802	-5.06864	0.676	0.0	50
180≤ l_{II} <190	4.16321	-5.80016	0.359	0.60387	51
190≤ l_{II} <200	0.78135	-0.27826	1.152	0.0	04
200≤ l_{II} <210	0.85535	0.20848	1.152	0.0	17
210≤ l_{II} <220	0.52521	0.65726	1.152	0.0	07
220≤ l_{II} <230	0.88376	-0.44519	0.993	0.06013	40
230≤ l_{II} <240	0.42228	-0.26304	0.803	0.0	26
240≤ l_{II} <250	0.71318	-0.67229	0.530	0.20994	55
250≤ l_{II} <260	0.99606	-0.70103	0.710	0.01323	48
260≤ l_{II} <270	0.91519	-0.39690	1.152	0.01961	48
270≤ l_{II} <280	0.85791	-0.29115	1.152	0.0	19
280≤ l_{II} <290	1.44226	-1.09775	0.657	0.0	39
290≤ l_{II} <300	2.55486	-1.68293	0.759	0.0	31
300≤ l_{II} <310	3.18047	-2.69796	0.589	0.0	40
310≤ l_{II} <320	2.11235	-1.77506	0.595	0.0	29
320≤ l_{II} <330	1.75884	-1.38574	0.635	0.0	25
330≤ l_{II} <340	1.97257	-1.55545	0.634	0.00043	34
340≤ l_{II} <350	1.41497	-1.05722	0.669	0.03264	46
350≤ l_{II} <360	1.17795	-0.95012	0.620	0.03339	46

-5≤ b_{II} <5	α	β	r_0	γ	$\frac{\sigma_A}{A}$
0 ≤ l_{II} <10	2.62556	-1.11097	1.182	0.00340	57
10 ≤ l_{II} <20	3.14461	-1.01140	1.555	0.0	42
20 ≤ l_{II} <30	4.26624	-1.61242	1.323	0.0	34
30 ≤ l_{II} <40	2.54447	-0.12771	1.300	0.0	30
40 ≤ l_{II} <50	2.27030	-0.68720	1.652	0.04928	52
50 ≤ l_{II} <60	1.34359	-0.05416	2.000	0.0	32
60 ≤ l_{II} <70	1.76327	-0.26407	2.000	0.0	37
70 ≤ l_{II} <80	2.20666	-0.41651	2.000	0.0	36
80 ≤ l_{II} <90	1.50130	-0.09855	1.475	0.0	45
90 ≤ l_{II} <100	2.43965	-0.82128	1.485	0.01959	36
100≤ l_{II} <110	3.35775	-1.16400	0.841	0.00298	35
110≤ l_{II} <120	2.60621	-0.68687	1.897	0.0	36
120≤ l_{II} <130	2.90112	-0.97988	1.480	0.0	32
130≤ l_{II} <140	2.55377	-0.71214	1.793	0.0	38
140≤ l_{II} <150	3.12598	-1.21437	1.287	0.15298	23
150≤ l_{II} <160	3.66930	-2.29731	0.799	0.33473	40
160≤ l_{II} <170	2.15465	-0.70690	1.524	0.14017	37
170≤ l_{II} <180	1.82465	-0.60223	1.515	0.20730	29
180≤ l_{II} <190	1.76269	-0.35945	2.000	0.08052	28
190≤ l_{II} <200	1.06085	-0.14211	2.000	0.0	48
200≤ l_{II} <210	1.21333	-0.23225	2.000	0.0	57
210≤ l_{II} <220	0.58326	-0.06097	2.000	0.36962	30
220≤ l_{II} <230	0.74200	-0.19293	1.923	0.07459	48
230≤ l_{II} <240	0.67520	-0.21041	1.604	0.16602	49
240≤ l_{II} <250	0.62609	-0.25312	1.237	0.14437	73
250≤ l_{II} <260	0.61415	-0.13788	2.000	0.26859	42
260≤ l_{II} <270	0.58108	0.01195	2.000	0.07661	40
270≤ l_{II} <280	0.68352	-0.10743	2.000	0.00849	50
280≤ l_{II} <290	0.61747	0.02675	2.000	0.0	49
290≤ l_{II} <300	1.06827	-0.26290	2.000	0.0	44
300≤ l_{II} <310	1.53631	-0.36833	2.000	0.02960	37
310≤ l_{II} <320	1.94300	-0.71445	1.360	0.15643	36
320≤ l_{II} <330	1.22185	-0.40185	1.520	0.07354	48
330≤ l_{II} <340	1.79429	-0.48657	1.844	0.0	43
340≤ l_{II} <350	2.29545	-0.84096	1.365	0.0	32
350≤ l_{II} <360	2.07408	-0.64745	1.602	0.12750	36

$5 \leq b_{II} < 15$	α	β	r_0	γ	$\frac{\sigma_A}{A}$
$0 \leq l_{II} < 10$	2.94213	-2.09258	0.703	0.05490	41
$10 \leq l_{II} < 30$	3.04627	7.71159	0.355	0.0	45
$30 \leq l_{II} < 40$	3.78033	-3.91956	0.482	0.0	42
$40 \leq l_{II} < 50$	2.18119	-2.4050	0.453	0.0	27
$50 \leq l_{II} < 60$	1.45372	-0.49522	1.152	0.0	31
$60 \leq l_{II} < 70$	1.05051	-1.01704	0.516	0.0	02
$70 \leq l_{II} < 80$	0.48416	-0.27182	0.891	0.08639	94
$80 \leq l_{II} < 90$	0.61963	0.41697	1.152	0.47171	35
$90 \leq l_{II} < 100$	4.40348	-2.95611	0.745	0.0	52
$100 \leq l_{II} < 120$	2.50938	-0.56541	1.152	0.0	27
$120 \leq l_{II} < 130$	0.44180	1.58923	0.949	0.0	04
$130 \leq l_{II} < 140$	3.96084	-3.37374	0.587	0.34109	40
$140 \leq l_{II} < 160$	2.53335	-0.40541	1.152	0.0	38
$160 \leq l_{II} < 170$	2.03760	-0.66317	1.152	0.0	23
$170 \leq l_{II} < 200$	1.06946	-0.87395	0.612	0.29230	29
$200 \leq l_{II} < 210$	0.86348	-0.65870	0.655	0.09089	79
$210 \leq l_{II} < 230$	0.30117	-0.16136	0.933	0.07495	17
$230 \leq l_{II} < 240$	0.75171	-0.57143	0.658	0.00534	12
$240 \leq l_{II} < 250$	1.97427	-2.02654	0.487	0.0	67
$250 \leq l_{II} < 260$	1.25208	-1.47763	0.424	0.31600	19
$260 \leq l_{II} < 270$	0.89448	-0.43870	1.019	0.0	05
$270 \leq l_{II} < 280$	0.81141	-0.51001	0.795	0.03505	27
$280 \leq l_{II} < 290$	0.83781	-0.44138	0.949	0.02820	50
$290 \leq l_{II} < 300$	1.10600	-0.86263	0.641	0.03402	28
$300 \leq l_{II} < 310$	1.37040	-1.02779	0.667	0.05608	28
$310 \leq l_{II} < 320$	1.77590	-1.26951	0.699	0.06972	37
$320 \leq l_{II} < 330$	1.20865	-0.70679	0.855	0.02902	35
$330 \leq l_{II} < 340$	2.28830	-1.71890	0.666	0.22887	42
$340 \leq l_{II} < 350$	3.26278	-0.94181	1.152	0.0	38
$350 \leq l_{II} < 360$	2.58100	-1.69237	0.763	0.0	53

$15 \leq b_{II} < 30$	α	β	r_0	γ	$\frac{\sigma_A}{A}$
$0 \leq l_{II} < 20$	6.23279	-10.30384	0.302	0.0	42
$20 \leq l_{II} < 40$	4.47693	-7.28366	0.307	0.0	29
$40 \leq l_{II} < 60$	1.22938	-1.19030	0.516	0.0	05
$60 \leq l_{II} < 80$	0.84291	-1.59338	0.265	0.0	04
$80 \leq l_{II} < 100$	0.23996	0.06304	0.523	0.0	32
$100 \leq l_{II} < 140$	0.40062	-1.75628	0.114	0.0	16
$140 \leq l_{II} < 180$	0.56898	-0.53331	0.523	0.0	41
$180 \leq l_{II} < 200$	-0.95721	11.69217	0.240	0.0	02
$200 \leq l_{II} < 220$	-0.19051	1.45670	0.376	0.0	01
$220 \leq l_{II} < 240$	2.31305	-7.82531	0.148	0.0	95
$240 \leq l_{II} < 260$	1.39169	-1.72984	0.402	0.0	06
$260 \leq l_{II} < 280$	1.59418	-1.28296	0.523	0.0	36
$280 \leq l_{II} < 300$	1.57082	-1.97295	0.398	0.0	10
$300 \leq l_{II} < 320$	1.95998	-3.26159	0.300	0.0	11
$320 \leq l_{II} < 340$	2.59567	-4.84133	0.268	0.0	37
$340 \leq l_{II} < 360$	5.30273	-7.43033	0.357	0.0	37

$30 \leq b_{II} < 45$	α	β	r_0	γ	$\frac{\sigma_A}{A}$
$0 \leq l_{II} < 20$	2.93960	-6.48049	0.227	0.0	77
$20 \leq l_{II} < 50$	1.65864	-9.99317	0.083	0.0	>99
$50 \leq l_{II} < 80$	1.71831	-7.25286	0.118	0.0	28
$80 \leq l_{II} < 110$	1.33617	-10.39799	0.064	0.0	>99
$110 \leq l_{II} < 160$	-0.31330	1.35622	0.329	0.0	24
$160 \leq l_{II} < 190$	1.51984	-8.69502	0.087	0.0	>99
$190 \leq l_{II} < 220$	-0.50758	4.73320	0.250	0.0	78
$220 \leq l_{II} < 250$	1.25864	-12.59627	0.050	0.0	70
$250 \leq l_{II} < 280$	1.54243	-3.76065	0.205	0.0	10
$280 \leq l_{II} < 320$	2.72258	-7.47806	0.182	0.0	05
$320 \leq l_{II} < 340$	2.81545	-5.52139	0.255	0.0	10
$340 \leq l_{II} < 360$	2.23818	0.81772	0.329	0.0	19

$45 \leq b_{II} < 60$	α	β	r_0	γ	$\frac{\sigma_A}{A}$
$0 \leq l_{II} < 60$	1.38587	-9.06536	0.076	0.0	03
$60 \leq l_{II} < 90$	2.28570	-9.88812	0.116	0.0	03
$90 \leq l_{II} < 110$	1.36385	-8.10127	0.084	0.0	04
$110 \leq l_{II} < 170$	0.05943	-1.08126	0.027	0.0	50
$170 \leq l_{II} < 200$	1.40171	-3.21783	0.218	0.0	99
$200 \leq l_{II} < 230$	0.14718	3.92670	0.252	0.0	14
$230 \leq l_{II} < 290$	0.57124	-4.30242	0.066	0.0	10
$290 \leq l_{II} < 330$	3.69891	-19.62204	0.094	0.0	05
$330 \leq l_{II} < 360$	1.19568	-0.45043	0.252	0.0	09

$60 \leq b_{II} < 90$	α	β	r_0	γ	$\frac{\sigma_A}{A}$
$0 \leq l_{II} < 30$	0.69443	-0.27600	0.153	0.0	>99
$30 \leq l_{II} < 60$	1.11811	0.71179	0.085	0.0	73
$60 \leq l_{II} < 90$	1.10427	-2.37654	0.123	0.0	>99
$90 \leq l_{II} < 120$	-0.42211	5.24037	0.184	0.0	12
$120 \leq l_{II} < 150$	0.87576	-4.38033	0.100	0.0	35
$150 \leq l_{II} < 180$	1.27477	-4.98307	0.128	0.0	72
$180 \leq l_{II} < 210$	1.19512	-6.58464	0.091	0.0	49
$210 \leq l_{II} < 240$	0.97581	-4.89869	0.100	0.0	95
$240 \leq l_{II} < 270$	0.54379	-0.84403	0.207	0.0	35
$270 \leq l_{II} < 300$	-0.85054	13.01249	0.126	0.0	39
$300 \leq l_{II} < 330$	0.74347	-1.39825	0.207	0.0	10
$330 \leq l_{II} < 360$	0.77310	-4.45005	0.087	0.0	16

This article was processed by the author using Springer-Verlag TEX A&A macro package 1992.

3.3 Perspectives

L'étude précédente est appelée à des développements dans les années à venir. D'abord, et c'est bien évident, parce qu'Hipparcos va donner les véritables couleurs des étoiles observées, couleurs qui n'avaient été qu'estimées par cette méthode. Mais plus généralement, et plus quantitativement, à l'aide de toutes les données bientôt disponibles :

- Les parallaxes ainsi que les mouvements propres que déterminera Hipparcos permettront d'améliorer les calibrations des magnitudes absolues et donc de diminuer l'incertitude sur l'absorption ;
- l'expérience Tycho va fournir 2 magnitudes (B_T et V_T) avec une précision variable suivant la magnitude, et meilleure en tout cas que 0.03 à la magnitude $B_T = 10.5$ [Høg *et al.*, 1992], [Scales *et al.*, 1992], pour les 500 000 étoiles les plus brillantes ; une seule magnitude sera disponible pour les 500 000 autres étoiles ;
- Tycho va permettre d'obtenir également des parallaxes, certes nettement moins précises que celles de la mission principale (30 mas à $B_T = 10.5$ mag), mais pour 500 000 étoiles [Høg *et al.*, 1992] ;
- les photométries de Genève et de Strömgren donnent des excès de couleur plus précis que la photométrie UBV ; de plus, le rapport de l'extinction dans le visible A_V sur l'excès de couleur en $(b-y)$, varie moins avec la couleur que R= $\frac{A_V}{E(B-V)}$ [Crawford & Mandewala, 1976], ce qui fournit une extinction plus précise également ; le nombre d'étoiles ayant de la photométrie $uvby-\beta$ est en augmentation constante ;
- avec le Michigan Spectral Survey, le nombre d'étoiles ayant un type spectral MK homogène devient conséquent (126 656 étoiles pour les volumes 1–4 déjà disponibles, avant d'atteindre les 225 000 étoiles au volume 7) ;

On peut ainsi très bien envisager, comme prolongement aux travaux sur l'extinction déjà effectués, un projet visant à obtenir simultanément :

- la calibration des couleurs intrinsèques ;
- la calibration des magnitudes absolues ;
- l'excès de couleur $E(B-V)$ pour un nombre important d'étoiles ;
- une estimation plus précise de l'absorption (meilleure que 0.2 mag) par un modèle tridimensionnel, sur des «cases» de moins d'un degré carré, jusqu'à 2 kpc ;
- le rapport R de l'absorption dans le visible sur l'excès de couleur, dans chaque région du ciel ;

et ainsi contribuer, par exemple, à la mise en évidence des régions de formation d'étoiles, à l'amélioration de la connaissance de l'extinction circumstellaire, etc.

Deuxième partie

MÉTHODES STATISTIQUES

Dans différentes parties de cette thèse, nous avons eu recours à des méthodes statistiques. Nous avions le choix soit d'alourdir l'exposé en les traitant là où nous en avions besoin, soit de jeter un voile pudique sur ces méthodes. Nous avons choisi une troisième solution, en les traitant ensemble dans le chapitre suivant.

Cette partie regroupe donc les méthodes statistiques, très générales, qui ont été nécessaires dans les autres parties de cette thèse et tente de les exposer en détail.

Nous nous sommes plus particulièrement intéressé dans cette partie à l'implication des erreurs de mesure sur les distributions étudiées, principalement dans le but d'essayer d'obtenir une estimation des variables initiales (sans erreur de mesure).

Loin d'être un simple exercice théorique, les méthodes que nous exposerons sont les réponses que nous avons trouvées à des problèmes très concrets auquels nous avons été confronté. Les parallaxes préliminaires d'Hipparcos, notamment, compte-tenu de leur qualité et donc de la difficulté à mettre en évidence un éventuel effet systématique qu'elles pourraient contenir, auront à elles seules suscité une grande partie du chapitre suivant.

Chapitre 4

Étude de distributions comportant des erreurs de mesures

Il est difficile d'analyser des données sans utiliser des méthodes d'analyse de données. Ceci sonne comme une évidence, mais les méthodes statistiques sont souvent soit mal aimées[1], soit mal comprises[2]. Et pourtant, toutes les données physiques que l'on mesure, avec les erreurs inhérentes à cette mesure, sont également des réalisations de variables aléatoires, et nécessitent par conséquent un traitement statistique.

Nous avons été confronté dans cette thèse à plusieurs cas de biais d'estimateurs, provenant d'erreurs de mesures, et nous avons donc été amené à trouver les estimateurs les plus adaptés à tenir compte de ces erreurs, d'une part, et à nous affranchir des biais, d'autre part.

4.1 Généralités concernant l'estimation

Sans trop s'attarder sur l'aspect analytique – on renvoie dans ce cas à la bibliographie correspondante – les méthodes statistiques sont décrites ci-dessous en utilisant les notations de manière informelle. On supposera qu'aucune confusion n'est à craindre entre une variable aléatoire X, ses valeurs théoriques x et leurs réalisations x_i ; de plus, si X est une variable aléatoire, par abus de notation on écrira souvent $f(x)$ au lieu de $f_X(x)$ sa fonction de densité de probabilité, $F(x)$ au lieu de $F_X(x)$ sa fonction de distribution, et $f(x|y)$ au lieu de $f_{X|Y=y}(x|y)$ la densité conditionnelle de X sachant $Y = y$. Enfin, l'espérance de la variable X sera notée $E[X]$.

Dans cette thèse on parlera à plusieurs reprises d'estimation et de propriétés d'estimateurs, et on va donc clarifier ici le vocabulaire employé.

En règle générale, on possède un échantillon et un paramètre θ (par exemple la moyenne) que l'on veut connaître. Un estimateur t_n de θ est une fonction des n variables x_i, et une estimation est la valeur de cet estimateur sur l'échantillon observé.

1. «*La statistique est la première des sciences inexactes*». E. et J. Goncourt
2. «*Je ne crois qu'aux statistiques que j'ai falsifiées moi-même*». W. Churchill

4.1.1 Propriétés des estimateurs

Le biais d'un estimateur t_n est la quantité $B_n(\theta)$ telle que $E[t_n] = \theta + B_n(\theta)$. Le caractère non biaisé, ou exactitude, de cet estimateur interviendra si B est nul. De même, un estimateur sera asymptotiquement non biaisé si le biais $B_n(\theta)$ tend vers 0 quand la taille de l'échantillon augmente.

Quant à la précision d'un estimateur, elle est reliée à sa variance. Plus celle-ci est petite, plus l'estimateur est précis. Mais il existe une limite inférieure à la variance des estimateurs non biaisés, la borne de Fréchet : introduisons la quantité

$$I(\theta) = E[\frac{\partial \log f(x|\theta)}{\partial \theta}]^2 = -E[\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2}]$$

nommée information de Fisher au point θ. Un estimateur non biaisé dont la variance atteint la borne de Fréchet $\frac{1}{nI(\theta)}$ est dit efficace. Mais rien ne dit qu'il existe un tel estimateur. Parfois, on peut même préférer un estimateur biaisé mais ayant une petite variance.

Il existe bien évidemment d'autres propriétés pour les estimateurs (convergence, etc), et on peut se référer par exemple à Tassi (1989) pour ces questions d'estimations. Nous en mentionnerons une dernière, qui est importante compte-tenu des nombreux points aberrants en Astronomie, la robustesse. Nombre de propriétés des estimateurs sont valables sous certaines hypothèses (de normalité par exemple) ; que deviennent ces propriétés lorsque les échantillons ne respectent que peu les hypothèses initiales ? Un estimateur qui conservera son efficacité dans ce cas de figure sera dit robuste. Nous aborderons ce point ultérieurement, page 63. Lecoutre et Tassi (1987), par exemple, traitent exhaustivement de la robustesse.

4.1.2 L'estimation bayésienne

Introduite par Bayes et Laplace, l'approche bayésienne consiste à supposer que le paramètre à estimer est une variable aléatoire, suivant une loi *a priori*. L'estimation bayésienne, en utilisant à la fois les observations x ainsi qu'une connaissance *a priori* sur le paramètre θ à estimer, améliore l'information sur ce paramètre au vu des observations, par l'intermédiaire de la formule de Bayes, donnée ici dans le cas continu :

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int_\Theta f(x|\theta)f(\theta)d\theta}$$

où $f(\theta)$ est la densité de la loi *a priori*, $f(x|\theta)$ la densité de la loi conditionnelle dite *vraisemblance de x sachant* θ, $f(x) = \int_\Theta f(x|\theta)f(\theta)d\theta$ la densité *prédictive* de la loi marginale, et $f(\theta|x)$ la densité de la loi *a posteriori*. D'une certaine manière, l'estimation bayésienne consiste à «*probabiliser l'inconnu*» [Robert, 1992], et, par l'intermédiaire de la formule de Bayes, à découvrir les causes à partir des effets.

Prenons l'exemple des parallaxes trigonométriques, pour lesquelles on suppose que la parallaxe observée π' a une erreur gaussienne autour de la vraie parallaxe π. Dans l'approche non-bayésienne, la vraie parallaxe d'une étoile est une valeur *fixée* inconnue. Dans le cadre bayésien, utilisant le fait que l'on connaît *a priori* – même approximativement – la distribution des vraies parallaxes (devenues variables aléatoires) dans la population

étudiée, la densité conditionnelle $f(\pi'|\pi)$ permet d'obtenir la densité *a posteriori* par l'intermédiaire de la formule de Bayes. Un estimateur bayésien de la vraie parallaxe sachant la parallaxe observée est alors obtenu en prenant le mode, la médiane ou l'espérance de la loi *a posteriori*.

L'espérance

$$\delta = \int_\Theta \theta f(\theta|x) d\theta$$

est l'estimateur le plus courant, car on peut montrer que c'est la valeur de l'estimateur d qui minimise l'espérance *a posteriori* de la perte quadratique $(d-\theta)^2$:

$$\delta = \min_d \int_\Theta (d-\theta)^2 f(\theta|x) d\theta$$

4.2 Estimations tenant compte des erreurs

Il arrive fréquemment que l'on ait à étudier une distribution (en général, trouver la moyenne et la dispersion des variables) mais que les variables observées soient entâchées d'erreurs de mesure. La plupart du temps, on se contente du raisonnement suivant : si les erreurs de mesures sont petites, on les ignore, et si elles sont importantes, on pondère les variables par leur variance. Ce raisonnement est très (trop ?) approximatif, surtout quand il existe une variation importante parmi les erreurs standards individuelles. La recherche d'une solution plus adéquate motive ce paragraphe.

On commence d'abord par trouver la solution dans le cas gaussien par maximum de vraisemblance, puis une solution plus robuste quand on s'éloigne de ce cas, et on décrit les tests statistiques utilisés ; enfin, des simulations montrent la qualité des estimateurs obtenus.

4.2.1 Modèle gaussien simple

Soient y_i ($i = 1, 2, \ldots\ n$) des variables aléatoires distribuées selon une loi gaussienne $\mathcal{N}(\mu, \sigma^2)$, et affectées par une erreur de mesure que l'on supposera également gaussienne et indépendante des variables y_i. Les variables observées sont les $x_i = y_i + \varepsilon_{x_i}$ avec $\varepsilon_{x_i} \rightsquigarrow \mathcal{N}(0, (\sigma_{x_i})^2)$. Dans la pratique, x_i a souvent été obtenue comme la moyenne de plusieurs mesures individuelles, et σ_{x_i} est l'écart-type de cette moyenne.

Les estimateurs de la moyenne μ et de l'écart-type σ sont obtenus par la méthode du maximum de vraisemblance et leurs propriétés sont démontrées en annexe, page 92 :

1 °) *L'estimateur*

$$\widehat{m} = \frac{\sum_{i=1}^{n} \frac{x_i}{\sigma^2+\sigma_{x_i}^2}}{\sum_{i=1}^{n} \frac{1}{\sigma^2+\sigma_{x_i}^2}} \tag{4.1}$$

est un estimateur non biaisé, efficace et convergent de la moyenne μ.

Sa variance est $s^2_{\widehat{m}} = \frac{1}{\sum_{i=1}^{n} \frac{1}{\sigma^2+\sigma_{x_i}^2}}$

Naturellement, en ce qui concerne l'absence de biais, toute moyenne pondérée, dont la somme des pondérations vaudrait 1, ferait l'affaire pour estimer le centre de la distribution. L'intérêt principal de cet estimateur est le fait qu'il soit efficace, c'est-à-dire que sa variance est la plus petite. En d'autres termes, cet estimateur est le plus précis possible.

2 °) *Notant* $p_i = \frac{1}{\widehat{s}^2+\sigma_{x_i}^2}$, *l'estimateur* $\widehat{s}$ *de* σ *est solution de :*

$$\widehat{s}\sum_{i=1}^{n} p_i(p_i(x_i-\mu)^2-1) \;=\; 0 \tag{4.2}$$

$$\textit{avec } I_n(\widehat{s}) = nI(\widehat{s}) = -\sum_{i=1}^{n}(p_i(1-4\widehat{s}^2p_i^2)(p_i(x_i-\mu)^2-1) - 2\widehat{s}^2p_i^2) > 0$$

Cet estimateur $\widehat{s} = \widehat{s}_n$ *est convergent et donc asymptotiquement non biaisé. Sa variance asymptotique est* $\frac{1}{I_n(\widehat{s})}$, *et il est donc asymptotiquement efficace.*

Il n'existe pas, dans le cas général, de solution analytique pour l'estimateur $\widehat{s}$ de σ. De plus, cet estimateur ne jouit d'aucune des «bonnes» propriétés de la moyenne pondérée $\widehat{m}$. En particulier, cet estimateur est biaisé.

On peut étendre les résultats précédents au cas où les données suivent une loi normale conjointe p-dimensionnelle. Notant $X_i = (x_i^{(j)})$ le p-vecteur colonne de l'observation i, Σ_{x_i} la $(p \times p)$-matrice (des covariances) des erreurs sur les $(x_i^{(j)})$, $M = (\mu^{(j)})$ le p-vecteur moyenne, et Σ la $(p \times p)$-matrice des covariances, la densité de probabilité de l'observation X_i est alors :

$$f(X_i, \Sigma_{x_i}\,;\; M, \Sigma) = \frac{1}{2\pi^{\frac{p}{2}}|\Sigma+\Sigma_{x_i}|^{\frac{1}{2}}} e^{\,-\frac{1}{2}(X_i-M)'(\Sigma+\Sigma_{x_i})^{-1}(X_i-M)}$$

où $(.)'$ désigne la transposée et $|.|$ le déterminant.

En notant $P_i = (p_i^{(j)(k)}) = (\Sigma+\Sigma_{x_i})^{-1}$, le logarithme de la fonction de vraisemblance de l'échantillon s'écrit :

$$\ln\,\mathcal{L} = -\frac{np}{2}Ln2\pi + \frac{1}{2}\sum_{i=1}^{n} Ln\,|P_i| \;-\frac{1}{2}\sum_{i=1}^{n}(X_i-M)'P_i(X_i-M)$$

Et on établit sans peine qu'un estimateur de la moyenne est

$$\widehat{M} = (\widehat{m}^{(j)}) \quad \text{où } \widehat{m}^{(j)} = \frac{\sum_{i=1}^{n} p_i^{(j)(j)} x_i^{(j)}}{\sum_{i=1}^{n} p_i^{(j)(j)}}$$

En pratique la matrice Σ_{x_i} est diagonale, car les covariances des erreurs sont rarement connues ; elles contribuent donc aux covariances de Σ que l'on cherche à déterminer.

Résolution des équations de vraisemblance

Les équations de vraisemblance (eq. 4.1) et (eq. 4.2), dépendant à la fois de m et s^2, doivent être résolues simultanément. Une valeur initiale de la résolution par itérations successives est aisée à déterminer : pour l'estimateur de la moyenne, on prend $m_0 = \frac{1}{n}\sum_{i=1}^{n} x_i$, moyenne empirique. Pour la variance, on remarque que si toutes les erreurs standards σ_{x_i} étaient égales à σ', on aurait (eq. 4.2)

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = \widehat{s}^2 + \sigma'^2$$

et on choisit donc

$$s_0{}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_0)^2 - \frac{1}{n}\sum_{i=1}^{n}\sigma_{x_i}{}^2$$

comme valeur initiale de la variance recherchée.

Dans le cas le plus favorable – correspondant au cas où la variance des erreurs $\sigma_{x_i}{}^2$ est petite devant la variance intrinsèque σ^2 de l'échantillon – la convergence est obtenue très rapidement après quelques itérations par la méthode de Newton.

4.2.2 Tests statistiques

On a commencé le §4.2.1 en faisant des hypothèses sur la nature de la distribution étudiée (la normalité). Naturellement, il faut pouvoir vérifier la véracité de ces hypothèses (ou, plus exactement, le fait qu'il n'existe pas d'indication que ces hypothèses sont erronées) : c'est le rôle des tests statistiques décrits ci-dessous.

Si l'on en croit Dudewicz & Mishra (1988), une des définitions des «statistiques» est «*la science de la prise de décision*». Les tests statistiques d'hypothèses peuvent être vus comme un problème de décision. En ce sens, on ne peut que désapprouver le bon mot de Lord Thorneycroft [3] : ce serait résumer les statistiques à une analyse purement descriptive. Mais cette étape exploratoire, quoiqu'indispensable, n'est pas suffisante : il faut pouvoir prendre des décisions si l'on ne veut pas se condamner à stagner. Et pour prendre une décision, faute de disposer d'autres renseignements, il est nécessaire de s'appuyer sur des tests.

En général, les tests statistiques sont de deux types :

- les tests d'adéquation, d'une loi empirique à une loi théorique,
- les tests paramétriques, comparant un paramètre d'une loi donnée à une valeur de référence.

On aborde ci-dessous essentiellement les tests d'adéquation.

Ces tests sont cités, sans être décrits, pas plus que ne sera décrite l'implémentation qui en a été effectuée. Pour plus de renseignement, on peut se reporter à la bibliographie, par exemple [Aïvazian *et al.*, 1986], [Dudewicz & Mishra, 1988], [Tassi, 1989], [Lecoutre & Tassi, 1987].

3. «*Il ne faut pas utiliser les statistiques commes les ivrognes utilisent les réverbères : pour s'appuyer et non pour s'éclairer*»

Dans cette thèse, lorsque, sans plus de précision, on indiquera que telle distribution est gaussienne, qu'il y a indépendance entre deux distributions, etc, cela signifiera que le test bilatéral adéquat aura ou non été significatif, au seuil de 5%.

Tests de normalité

Le Théorème Central Limite (TCL) indique que la moyenne arithmétique d'une série de variables aléatoires – quelle que soit leur distribution – converge vers la loi normale. Plus précisément, si les (x_i) sont des variables aléatoires indépendantes et identiquement distribuées, d'espérance μ et de variance σ^2, $\frac{\sqrt{n}(\langle x_i \rangle - \mu)}{\sigma}$ converge en loi vers $\mathcal{N}(0, 1^2)$ lorsque $n \to \infty$. Grâce à (ou à cause[4] du) TCL, la loi normale est la distribution la plus utilisée, pas toujours à bon escient, parce qu'il existe souvent des distributions plus adaptées à l'échantillon étudié et parce que le TCL fonctionne mal aux ailes des lois. Dans la pratique, le recours à la loi normale ne fait que mal cacher la méconnaissance du phénomène sous-jacent.

Quoi qu'il en soit, si l'on suppose que la distribution observée est gaussienne, il faut pouvoir disposer de tests de normalité. Les tests utilisés ici sont les suivants :

- le test de Kolmogorov [Aïvazian *et al.*, 1986], la moyenne et la variance de la loi normale étant connue; la statistique calculée est la distance (verticale) maximum entre la distribution empirique et la distribution normale ayant cette moyenne et cette variance ;
- le test de Lilliefors, variante du test de Kolmogorov, utilisée dans le cas (*le plus fréquent*) où moyenne et variance sont déterminés empiriquement à partir de l'échantillon ;
- le test sur l'asymétrie, en utilisant le coefficient de Fisher $\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}}$; on calcule l'asymétrie empirique g_1 qui a pour variance $v_1 = \frac{6n(n-1)}{(n-2)(n+1)(n+3)}$; asymptotiquement, la statistique $\frac{g_1}{\sqrt{v_1}} \rightsquigarrow \mathcal{N}(0, 1^2)$ [Tassi, 1989] ;
- le test sur l'aplatissement $\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3$ de Fisher ; on peut montrer [Tassi, 1989] que la statistique empirique g_2 associée à γ_2 a pour variance $v_2 = \frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}$ et qu'asymptotiquement $\frac{g_2}{\sqrt{v_2}} \rightsquigarrow \mathcal{N}(0, 1^2)$

De nombreux autres tests existent : Shapiro-Wilk, Lin-Muldokar, Vasicek... (se référer à Lecoutre & Tassi (1987)).

Autres tests d'adéquations

Il va de soi que la distribution gaussienne est un cas limite, et il faut pouvoir tester si un échantillon observé suit une autre distribution théorique (uniforme, poissonnienne, etc). Le test d'adéquation utilisé ici est le test de Kolmogorov.

Plus généralement, si l'on veut comparer deux distributions observées quelconques, nous avons implanté deux tests : le test du χ^2 de Pearson et le test de Kolmogorov.

4. «*Chacun est convaincu de la véracité de la loi normale : les expérimentateurs, parce qu'ils pensent que c'est un théorème de mathématique ; les mathématiciens, parce qu'ils pensent que c'est un fait expérimental*», Lipman, cité par H.Poincaré

Ce dernier est à préférer car il est plus puissant que le test du χ^2. Cela se comprend aisément puisque le test du χ^2 impose un regroupement des données en catégories, faisant évidemment perdre de l'information par rapport à l'information apportée par les données brutes.

Tests de corrélation et d'indépendance

On est souvent amené à se poser la question de l'association d'une variable avec une autre, par exemple au §6.5, où il nous faudra vérifier que l'erreur systématique sur la parallaxe Hipparcos n'est pas liée aux caractéristiques physiques des étoiles.

Pour clarifier les termes utilisés, rappelons que deux variables aléatoires sont non corrélées si et seulement si $E[XY] = E[X]E[Y]$; deux v.a. sont indépendantes si et seulement si $f(x, y) = f(x)f(y)$. En conséquence, deux v.a. indépendantes sont non corrélées, mais deux v.a. non corrélées ne sont pas forcément indépendantes.

Le coefficient de corrélation usuel n'est rien d'autre qu'une mesure de la relation linéaire qui peut exister entre deux variables. On s'interessera donc plus particulièrement aux tests d'indépendance. Quand les distributions étudiées sont quelconques, et les éventuelles relations entre elles le sont également, on utilisera le test du τ de Kendall [Lecoutre & Tassi, 1987] parce qu'il est non paramétrique (on ne fait pas d'hypothèse sur la distribution des variables, contrairement au test sur le coefficient de corrélation de Pearson), non linéaire (contrairement au test de Spearman), et robuste dans le sens où il est peu perturbé par des points aberrants.

Certes, ce test détecte essentiellement les associations monotones entre les variables étudiées ; de plus, comme il s'agit d'un test de rang, donc perdant de l'information par rapport aux données initiales, il peut ne pas reconnaître une dépendance entre les variables ; en revanche, s'il détecte une dépendance, c'est très probablement qu'elle existe réellement.

4.2.3 Écarts à la loi normale – robustesse

Restant dans le cadre du modèle gaussien simple décrit au §4.2.1, nous pouvons nous demander si la solution trouvée est robuste.

L'hypothèse que les erreurs sur les données suivent une loi gaussienne est en pratique peu contraignante. Parfois il est possible, au moyen d'un changement de variable, de retrouver une erreur gaussienne (par exemple, l'erreur sur une parallaxe spectroscopique est log-normale, si l'on suppose gaussienne l'erreur sur la magnitude absolue).

Plus gênante est en revanche l'hypothèse de normalité en ce qui concerne la distribution des variables sans erreurs de mesures, telle qu'on l'a faite dans le modèle simple du §4.2.1. Pour vérifier cette hypothèse de normalité, on construit les données «normalisées» $x_i' = \frac{x_i - m}{\sqrt{\sigma^2 + \sigma_{x_i}^2}}$. D'après les hypothèses qui ont été faites, $x_i' \rightsquigarrow \mathcal{N}(0, 1^2)$. Au vu des x_i' et grâce aux tests indiqués au §4.2.2, on a les moyens d'accepter ou de refuser la normalité.

Naturellement, on écarte le cas où l'on sait dès le départ que la distribution n'est pas gaussienne, auquel cas il faudrait recalculer les équations de vraisemblance en utilisant la densité de la distribution en présence. En particulier, si cette loi est multimodale, et si l'on peut faire l'hypothèse d'un mélange de populations gaussiennes, alors les résultats précédents peuvent être appliqués sur chacune des populations... pour autant que l'on

ait réussi à les séparer. Le §4.4.1 traite ce problème délicat, directement dans le cas multidimensionnel.

Mais si la non-normalité n'est dûe qu'à des observations anormales venant polluer la distribution, on peut garder l'hypothèse de normalité et procéder ainsi :

Il est connu que la moyenne – et surtout la variance – empiriques sont extrêmement sensibles aux queues des distributions. Il existe heureusement d'autres estimateurs qui y sont moins sensibles ; par exemple, pour estimer le centre d'une distribution, on peut se servir d'une moyenne symétriquement tronquée à $\alpha\%$ (on supprime les αn points les plus petits et les αn points les plus grands) ou de la médiane ; pour estimer l'étendue de la distribution, on peut se servir de la moyenne des valeurs absolues des écarts à la moyenne (écart absolu moyen) ou bien de distances inter-quantile (par exemple la distance semi-interquartile $\frac{1}{2}(x_{(0.75)} - x_{(0.25)})$), etc. En revanche, la variance de ces estimateurs est plus grande ; le tableau 4.1 le montre bien.

TAB. 4.1: *Variance asymptotique d'estimateurs.*

Variance asymptotique de différents estimateurs de centre ou d'étendue d'une distribution supposée normale .

Moyenne empirique	$\frac{\sigma^2}{n}$
Moyenne tronquée à 38%	$1.36\frac{\sigma^2}{n}$
Médiane	$\frac{\pi}{2}\frac{\sigma^2}{n}$
Écart-type empirique $\approx$	$\frac{\sigma^2}{2n}$
Écart absolu moyen ($\times\sqrt{\frac{\pi}{2}}$)	$(\pi-2)\frac{\sigma^2}{2n}$
Distance semi-interquartile ($\times 0.741$)	$2.72\frac{\sigma^2}{2n}$

(d'après Aïvazian *et al.* (1986), Lecoutre & Tassi (1987), Tassi (1989))

Dans ce tableau, le coefficient multiplicatif de l'écart absolu moyen et de la distance semi-interquartile proviennent du fait que, sans cela, ces estimateurs sont des estimateurs biaisés de l'écart-type.

L'augmentation de la variance pour ces estimateurs n'est pas vraiment importante. Comme l'écrivait P. Huber (cité dans Lecoutre & Tassi, 1987) : «*la robustesse est une sorte d'assurance : je suis prêt à payer une perte d'efficacité de 5 à 10% par rapport au modèle idéal pour me protéger de mauvais effets de petites déviations de celui-ci : je serai bien sûr heureux que ma procédure fonctionne bien sous de gros écarts, mais je n'y prête pas réellement attention car faire de l'inférence à partir d'un modèle aussi faux n'a que peu de signification concrète*».

En règle générale, sous l'hypothèse gaussienne, nous utilisons fréquemment la médiane comme estimateur du centre de la distribution, et comme estimateur de l'écart-type nous utilisons $\frac{1}{2}(x_{(0.8415)} - x_{(0.1585)})$, de variance $0.89\frac{\sigma^2}{n}$; nous les désignerons ultérieurement sous le nom d'estimateurs «à base de quantile».

On cite ici le problème des points aberrants, mais on s'intéresse à la robustesse d'un estimateur dans le sens général de la faible sensibilité à un petit écart par rapport aux hypothèses initiales ; en l'occurence ici l'hypothèse de normalité.

Dans le cadre du modèle qui nous préoccupe, les estimateurs trouvés au §4.2.1 sont peu robustes. Pour trouver des estimateurs robustes $\widetilde{m}_{\mathrm{r}}$ de μ et $\widetilde{s}_{\mathrm{r}}$ de σ, nous procédons de manière itérative, en prenant lors de la première itération $\widetilde{m}_{\mathrm{r}} = \text{mediane}(x_i)$ et $\widetilde{s}_{\mathrm{r}}^2 = \text{Écart absolu moyen}(x_i)^2 - \frac{1}{n}\sum_{i=1}^{n}{\sigma_{x_i}}^2$. Ensuite on calcule, avec les estimateurs à base de quantile, la moyenne et la dispersion de la distribution des données «normalisées» ; on s'attend à ce qu'elles soient proches respectivement de 0 et 1. L'écart à ces valeurs permet alors de corriger les estimateurs $\widetilde{m}_{\mathrm{r}}$ et $\widetilde{s}_{\mathrm{r}}$, et on itère le processus tant que cet écart n'est pas négligeable.

Naturellement, on peut comparer ces estimateurs à d'autres estimateurs. On aborde ce point page 66.

4.2.4 Simulations

Dans plusieurs cas, notamment pour tester différents estimateurs, nous avons eu à effectuer des simulations sur des échantillons «tirés au hasard», censés représenter une distribution connue. Il existe plusieurs manières de générer des variables «aléatoires» suivant une distribution donnée :

- La première est à utiliser lorsque l'on connaît analytiquement la densité de probabilité $f(t)$ de la distribution que l'on veut simuler, et que l'on peut calculer $z = F(t) = \int_{-\infty}^{t} f(u)du$ puis sa réciproque $F^{-1}(z)$. On sait [Aïvazian *et al.*, 1986] que $F(t)$ prise comme variable aléatoire a une distribution uniforme sur $[0, 1]$. On va donc générer une variable aléatoire $z = F(t)$ suivant une loi uniforme, et la variable $F^{-1}(z)$ sera donc distribuée comme t (on note de façon identique les variables aléatoires et leurs valeurs théoriques).

- Si l'on a une distribution observée que l'on veut simuler, et s'il ne s'agit pas d'une distribution «connue» (pour laquelle on a analytiquement la densité), on peut se servir de la Distribution Lambda généralisée [Ramberg *et al.*, 1979], qui utilise les 4 premiers moments empiriques de la distribution : elle est définie par

$$F^{-1}(z) = \lambda_1 + \frac{z^{\lambda_3} - (1-z)^{\lambda_4}}{\lambda_2}$$

  où $z \in [0, 1]$. Les 4 premiers moments peuvent s'exprimer en fonction des paramètres $\lambda_1, ... \lambda_4$. Il suffit donc d'ajuster ces paramètres en fonction des moments empiriques, puis de «tirer» une variable uniforme z.

- Une autre consiste à utiliser les propriétés de la loi à simuler, quand on les connaît. À titre d'exemple :

  - la variable aléatoire $\sqrt{\frac{12}{n}}\sum_{i=1}^{n}(z_i - 0.5)$ suit approximativement une loi normale réduite $\mathcal{N}(0, 1^2)$, si z_i suit une loi uniforme sur $[0, 1]$;

– la loi de Cauchy a pour densité $f(x) = \frac{1}{\pi}\frac{c}{c^2+(x-a)^2}$; le rapport de deux variables aléatoires suivant $\mathcal{N}(0,1^2)$ suit une loi de Cauchy de paramètres $a = 0$ et $c = 1$;
– la somme des carrés de n variables aléatoires suivant $\mathcal{N}(0,1^2)$ suit une loi du χ^2 à n degrés de liberté ;

Comme application de la dernière méthode, on peut citer le cas où l'on a besoin de simuler des erreurs standards de mesure s_{x_i}. Si l'on suppose que x_i a été obtenue à l'aide de p mesures et que l'erreur de mesure sur x_i est gaussienne d'écart-type σ_{x_i}, on utilise le fait que $(p-1)\frac{{s_{x_i}}^2}{{\sigma_{x_i}}^2}$ suit une loi du χ^2 à $(p-1)$ degrés de liberté. Pour simuler l'erreur standard de mesure s_{x_i}, on calcule donc

$$s_{x_i} = \sigma_{x_i}\sqrt{\frac{1}{p-1}\sum_{j=1}^{p-1}{N_i}^2} \quad \text{où } N_i \text{ est une v.a. tirée suivant} \rightsquigarrow \mathcal{N}(0,1^2)$$

Pour toutes ces méthodes, il est clair qu'il suffit seulement de disposer de variables z «aléatoires» suivant une loi uniforme, pour générer une variable suivant la distribution que l'on veut simuler.

Des générateurs de nombres quasi-aléatoires suivant une loi uniforme sont disponibles, voir par exemple dans Press (1990), et qui ont l'avantage d'être portables sur différentes machines, mais c'est la fonction `random()`, présente sur les systèmes Unix, qui a été choisie, après des tests, pour sa longue période ($\simeq 3.4\ 10^{10}$) et son absence de corrélation entre des tirages successifs. Pour générer des nombres suivant une loi gaussienne, la fonction `gasdev()` extraite de Press (1990), p. 216, a été utilisée.

Comparaison entre estimateurs

Nous avons développé des estimateurs dans le cadre du modèle gaussien du §4.2.1, ainsi que des estimateurs analogues mais plus robustes, au §4.2.3. On peut naturellement se demander s'ils ont un intérêt supérieur à des estimateurs existants, et si oui, dans quelle proportion.

Pour cela, on va introduire une mesure de cette qualité, l'efficacité relative asymptotique, qui va nous permettre de comparer deux estimateurs $t_{1,n}$ et $t_{2,n}$:

$$R(t_{1,n}/t_{2,n}) = \lim_{n\to\infty}\frac{V(t_{2,n})}{V(t_{1,n})}$$

où $V(.)$ désigne la variance, si les deux estimateurs sont non biaisés. Sinon, on remplace la variance par $\frac{V(t_n)}{(\frac{\partial E[t_n]}{\partial\theta})^2}$. Donc $R(t_{1,n}/t_{2,n})$ sera plus grand que 1 si $t_{1,n}$ est plus efficace que $t_{2,n}$.

Nous allons prendre comme variance de référence celle de l'estimateur empirique de la moyenne (respectivement de l'écart-type), et mesurer l'efficacité de plusieurs estimateurs du centre (resp. de l'étendue) de la distribution relativement à l'estimation empirique, à l'aide de simulations. Pour ces simulations, on calculera les estimateurs sur 400 échantillons de 5000 points.

Les estimateurs du centre de la distribution que nous allons tester sont :

1. la moyenne de la distribution tronquée à $[-3\sigma, +3\sigma]$;

2. la moyenne pondérée (eq. 4.1) ;

3. la médiane ;

4. la moyenne de la distribution tronquée à 38% ;

5. la moyenne pondérée robuste (§4.2.3) ;

Quant aux estimateurs de l'étendue, il s'agit de :

1. l'écart-type de la distribution tronquée à $[-3\sigma, +3\sigma]$;

2. l'écart-type pondéré (eq. 4.2) ;

3. l'étendue à base de quantiles (cf p. 64) ;

4. l'écart absolu moyen ;

5. l'écart-type pondéré robuste (§4.2.3) ;

La distribution qui est simulée est une gaussienne $\mathcal{N}(\mu, \sigma^2)$ avec une erreur de mesure sur chaque variable, la moyenne des erreurs quadratiques de mesure étant notée $k\sigma$. Si l'on fait varier k, c'est-à-dire que l'on fait augmenter la taille moyenne des erreurs de mesures, on peut observer sur une simulation l'efficacité relative des estimateurs du centre (fig. 4.1) et de l'étendue (fig. 4.2).

Clairement, les estimateurs pondérés apparaissent plus efficaces que les autres, et toujours strictement supérieurs à 1. Sur la première figure, on voit que l'efficacité des estimateurs robustes rejoint celle de la moyenne au fur et à mesure que les erreurs de mesure deviennent prépondérantes. Sur la seconde, quand ces erreurs augmentent, il ne faut pas s'étonner que les estimateurs robustes, valables sous l'hypothèse gaussienne, sont extrêmement peu efficaces quand on s'éloigne de cette hypothèse.

Maintenant, si l'on fixe la moyenne des erreurs quadratiques de mesure à une valeur peu élevée (on a pris ici 0.5σ), que se passe-t-il si l'on a des points aberrants, ou une distribution à «queue lourde» ? Pour simuler cela, on va ajouter à la gaussienne initiale $\mathcal{N}(\mu, \sigma^2)$ une certaine proportion d'une gaussienne $\mathcal{N}(\mu, (4\sigma)^2)$, puis les erreurs de mesure. L'efficacité relative des estimateurs du centre (fig. 4.3) et de l'étendue (fig. 4.4), sont représentés quand on fait varier cette proportion de 0 à 1.

La situation est ici plus contrastée. La moyenne pondérée reste plus efficace, avec une efficacité relative toujours supérieure à 1, même quand la pollution devient importante, et la moyenne pondérée robuste (ici superposé à la médiane) a une efficacité relative à la moyenne empirique souvent inférieure à 1. Il apparaît clair que les estimateurs connus comme robustes (médiane, moyenne symétriquement tronquée) perdent rapidement leur efficacité quand il y a des erreurs de mesure, même peu élevées.

Quant aux estimateurs de l'étendue, on voit que si l'estimateur pondéré robuste n'est pas le plus efficace partout, son efficacité possède l'avantage de ne varier que peu avec le taux de points aberrants, dès que l'on a 4 points sur mille qui sont aberrants. On peut noter que la situation serait très différente si la pollution choisie était dissymétrique ou si l'on travaillait sur des petits échantillons.

On voit donc grâce à ces simulations tout l'intérêt qu'apportent les estimateurs trouvés dans le cadre du modèle gaussien avec erreurs, et comment ils remplacent avantageusement

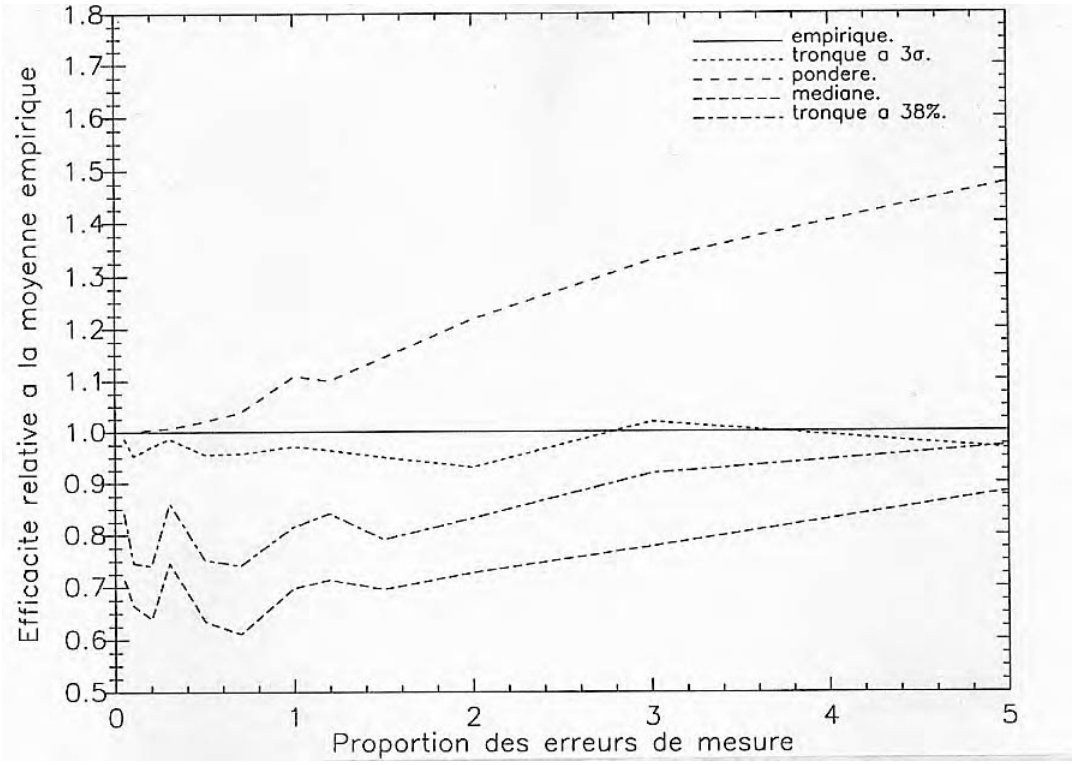


FIG. 4.1: *Efficacité relative asymptotique d'estimateurs de la moyenne dans le cadre du modèle gaussien avec erreurs, en fonction des erreurs de mesure.*

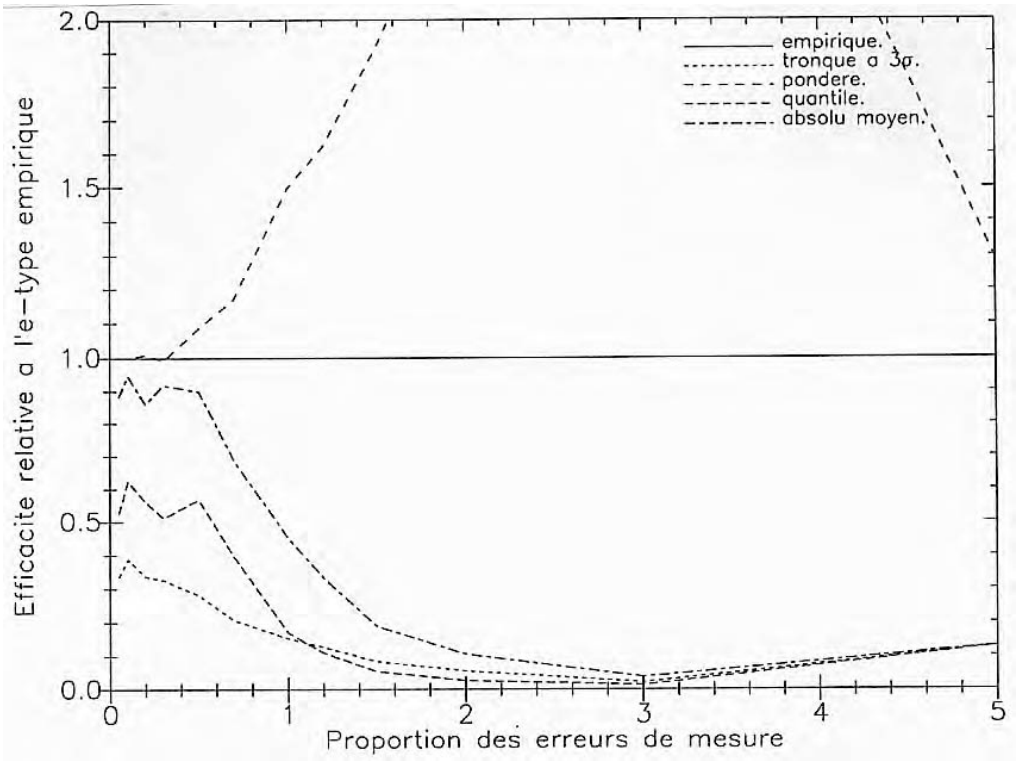


FIG. 4.2: *Efficacité relative asymptotique d'estimateurs de l'écart-type dans le cadre du modèle gaussien en fonction des erreurs de mesure.*

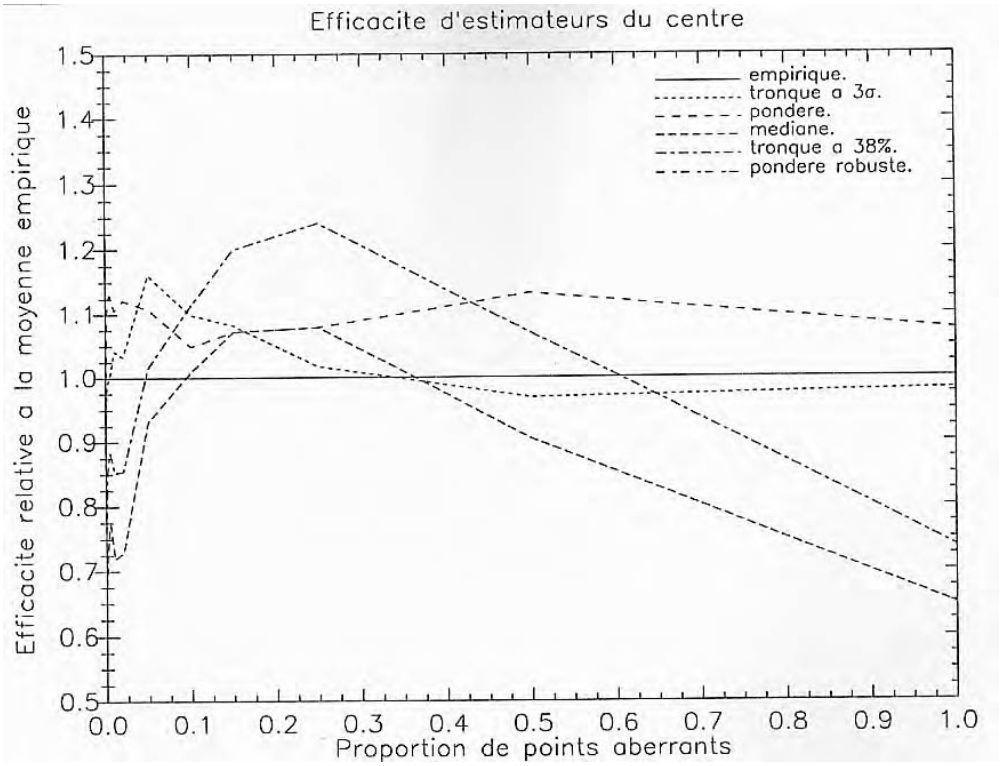


FIG. 4.3: *Efficacité relative asymptotique d'estimateurs de la moyenne dans le cadre du modèle gaussien avec erreurs en fonction du taux de pollution par* $\mathcal{N}(\mu, (4\sigma)^2)$.

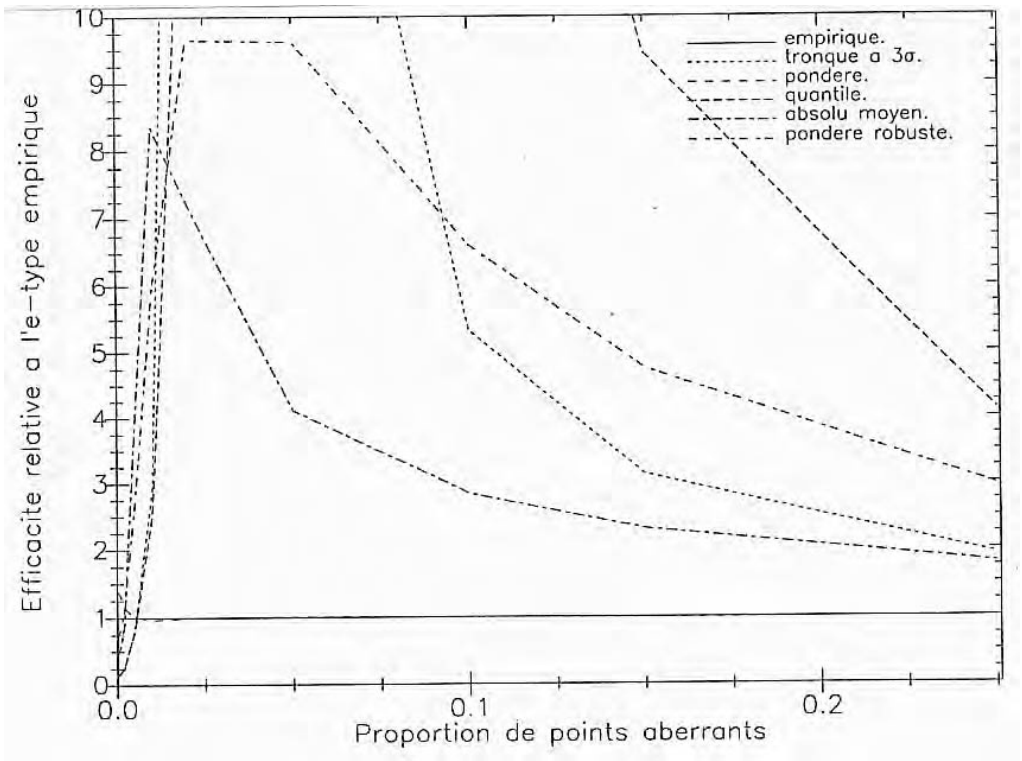


FIG. 4.4: *Efficacité relative asymptotique d'estimateurs de l'écart-type dans le cadre du modèle gaussien avec erreurs en fonction du taux de pollution par* $\mathcal{N}(\mu, (4\sigma)^2)$.

ceux que l'on aurait pu avoir tendance à utiliser si l'on néglige les erreurs de mesure. Ceci est principalement dû au fait que si la distribution des variables sans erreurs peut être supposée normale, la distribution observée avec des erreurs de mesure s'éloigne rapidement de l'hypothèse gaussienne.

4.3 Déconvolution des erreurs

À partir de maintenant, on va s'intéresser à obtenir une estimation des variables sans erreur. L'utilisation, pour ce qui nous concerne, sera l'estimation de vraies parallaxes, connaissant les parallaxes observées. Mais l'application en est beaucoup plus large, pour toute distribution observée avec des erreurs de mesure ; dans un premier temps, on va supposer que les variables initiales sont distribuées suivant une gaussienne $\mathcal{N}(\mu, \sigma^2)$.

4.3.1 Aspect bayésien dans le modèle gaussien

Si les variables sans erreur y sont distribuées normalement, leur densité de probabilité s'écrit :

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}}$$

Quant à la densité conditionnelle des variables entâchées d'erreur x, sachant les variables y, elle s'écrit naturellement :

$$f(x|y) = \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-y)^2}{\sigma_x^2}}$$

Enfin, la densité des x est indépendante des y (voir annexe) :

$$f(x) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2+\sigma_x^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2+\sigma_x^2}}$$

En utilisant la formule des densités conditionnelles,

$$f(y|x)f(x) = f(x|y)f(y)$$

on trouve l'expression de la densité de probabilité des variables sans erreur y sachant les x observés :

$$f(y|x) = \frac{\sqrt{\sigma^2+\sigma_x^2}}{\sigma_x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}[\frac{(x-y)^2}{\sigma_x^2}+\frac{(y-\mu)^2}{\sigma^2}-\frac{(x-\mu)^2}{\sigma^2+\sigma_x^2}]}$$

La valeur la plus probable des y, notée $\widehat{y}$, est obtenue par maximisation de la densité, ce qui annule sa dérivée par rapport à y :

$$-\frac{1}{2}(2\frac{-(x-\widehat{y})}{\sigma_x^2} + 2\frac{(\widehat{y}-\mu)}{\sigma^2}) = 0$$

On trouve ainsi l'estimateur bayésien de la variable sans erreur y :

$$\widehat{y} = x + (\mu - x)\frac{\sigma_x^2}{\sigma^2+\sigma_x^2} \tag{4.3}$$

Ceci correspond à l'idée intuitive que l'on peut en avoir : si la variable observée est loin de la moyenne de la distribution, alors la variable sans erreur de mesure doit être plus proche du centre, et ce d'autant plus que l'erreur observationnelle est importante. On remarque qu'ici la déconvolution des erreurs s'effectue de manière très simple.

De plus, on peut calculer la variance de cet estimateur

$$\begin{aligned}\mathrm{Var}(\widehat{y}) &= \mathrm{Var}(x\tfrac{\sigma^2}{\sigma^2+\sigma_x^2} + \mu\tfrac{\sigma_x^2}{\sigma^2+\sigma_x^2}) \\ &= \mathrm{Var}((x-y)\tfrac{\sigma^2}{\sigma^2+\sigma_x^2} + (y-\mu)\tfrac{\sigma_x^2}{\sigma^2+\sigma_x^2} + \mu)\end{aligned}$$

d'où la variance

$$\mathrm{Var}(\widehat{y}) = \frac{\sigma^2\sigma_x^2}{\sigma^2+\sigma_x^2} \tag{4.4}$$

On est ici dans un cas de figure où l'on sait calculer analytiquement $f(x)$; dans le paragraphe suivant, cela ne sera plus le cas. On voit comment cette approche consiste à estimer la variable y en calculant la densité de probabilité *a posteriori* $f(y|x)$, connaissant la densité de probabilité *a priori* $f(y)$.

4.3.2 Biais dûs aux erreurs de mesure

Quittons l'hypothèse simplificatrice d'une distribution gaussienne des variables sans erreur y. Prenons une distribution unidimensionnelle de densité quelconque pour les y ; la seule hypothèse que l'on fera est qu'elle soit deux fois dérivable et tendant vers 0 en $\pm\infty$. La figure 4.5 en est un exemple. Tout au long des paragraphes qui suivent, on conservera cet exemple tout à fait quelconque d'une distribution bimodale. On a pris ici la densité de probabilité

$$f(y) = \frac{1}{3\pi}(\frac{1}{1+(y-4)^2} + \frac{1}{1+(y/2)^2})$$

parce qu'elle permettait de présenter la méthode sans compliquer les calculs. On peut vérifier qu'il s'agit bien d'une densité de probabilité puisque $f(y) \geq 0$ et $\int_{-\infty}^{+\infty} f(y)dy = 1$.

Conservant dans cet exemple une loi d'erreur gaussienne pour les x, on va étudier ce qu'il se passe quand on veut comparer les x aux y, en fonction de la variable observée x. Ce type de comparaison est très courant ; l'exemple en est cité à deux reprises, §6.2.3 et §6.5.1, en étudiant la variation de la différence entre deux déterminations de la parallaxe en fonction de la parallaxe observée.

En prenant une dispersion des erreurs standards de mesure de 0.5, la distribution simulée des x se trouve sur la figure 4.6. Pour faire cette simulation (cf §4.2.4), on a généré une variable aléatoire $z = F(y)$ suivant une loi uniforme. On calcule $y = F^{-1}(z)$ par

$$z = F(y) = \frac{1}{3\pi}(\mathrm{Arctg}(y-4) + 2\mathrm{Arctg}(y/2) + \frac{3\pi}{2})$$

d'où, en posant $\alpha = 3\pi z - \frac{3\pi}{2}$, on est amené à résoudre

$$y^3 - (5\alpha+4)y^2 + (16\alpha-8)y + 4\alpha + 16 = 0$$

On est dans le cas où l'on a trois racines réelles, équiprobables, et l'on en tire donc une au hasard. A partir de cette «vraie» variable y, on tire ensuite une variable observée $x \rightsquigarrow \mathcal{N}(y, 0.5^2)$.

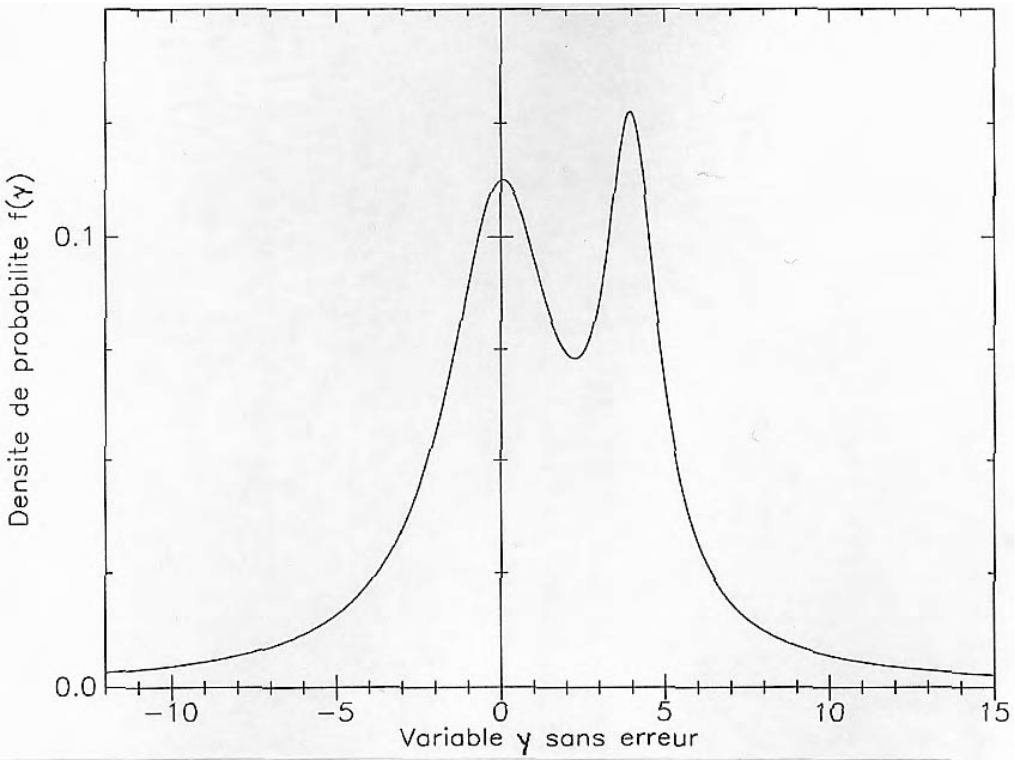


FIG. 4.5: *Exemple de distribution d'une variable sans erreur.*

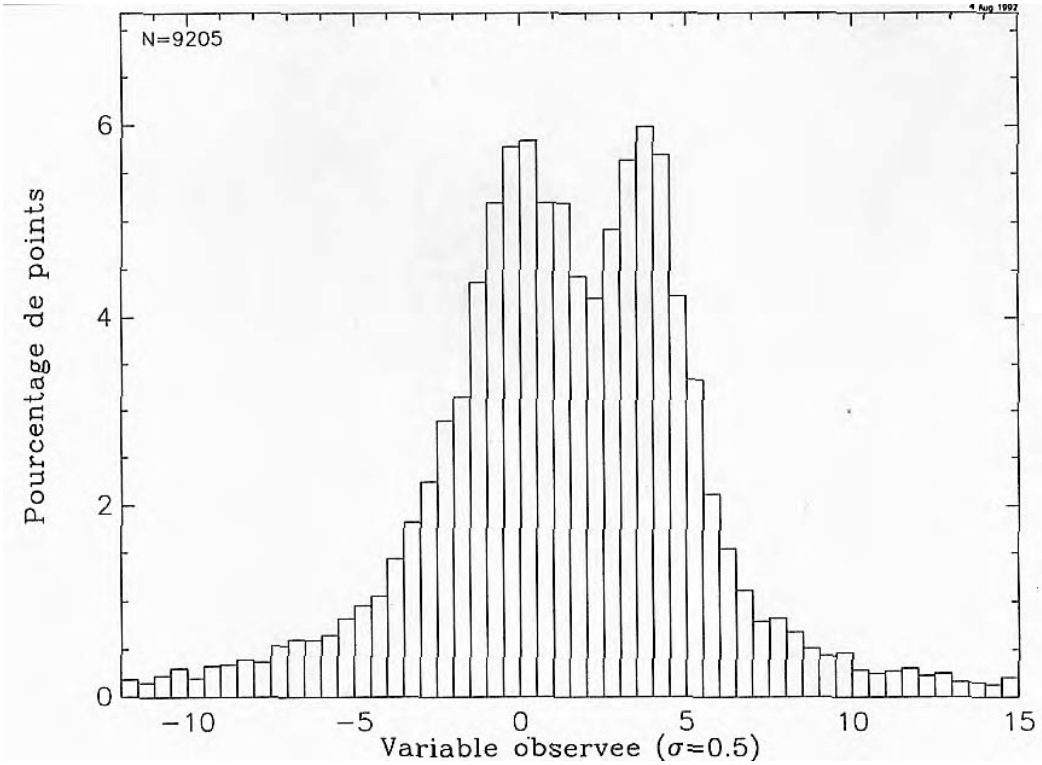


FIG. 4.6: *Simulation de la variable observée avec erreur.*

Calculons maintenant la différence $x - y$, et regardons sa variation en fonction de la variable observée x. Comme on a beaucoup de points, et pour que le dessin reste lisible, on va faire des moyennes des $x - y$ sur des «tranches» de x (ici, on a pris pour chaque point en abscisse la médiane des ordonnées des 500 points situés de part et d'autre de x). Et l'on voit apparaître une variation très importante en fonction des x, figure 4.7. Ceci pourrait sembler inattendu puisque les x sont distribués symétriquement autour des y et que l'on a $E[X] = Y$.

Cet artefact provient du fait que les variables observées sont la convolution d'une loi d'erreur et d'une distribution non uniforme des variables sans erreur. Intuitivement, on comprend bien que (par définition) la vraie variable a plus de probabilité de se trouver sur un mode de la distribution que de part et d'autre de ce mode ; à cause des erreurs, la variable observée va donc se retrouver plus fréquemment sur les ailes de la distribution que la variable sans erreur de mesure.

Le lissage ne crée pas ce biais, il ne fait que le mettre mieux en évidence ; en effet, chacun des points étant moyenné sur n observation, l'estimation de la position des ordonnées est améliorée d'un facteur $\propto \frac{1}{\sqrt{n}}$. Le biais sera visible dès que seront faites des moyennes sur une variable avec une erreur de mesure, et que l'on regardera le comportement d'une autre variable entâchée d'erreur et non indépendante de la première en fonction de ces moyennes. Si l'on avait fait le graphe des $x - y$ en fonction des y, il n'y aurait pas eu de biais, parce que les y n'ont pas d'erreur de mesure. Autrement dit, on a $E[X] = Y$ (d'où $E[X - Y|Y] = 0$) mais $E[X - Y|X] \neq 0$.

Si l'ensemble de la distribution est considérée, le biais se compensera. Mais dès que l'on contraindra d'une quelconque manière une variable affectée d'une erreur de mesure (en ne gardant dans un échantillon que celles inférieures/supérieures à telle limite sur la variable observée), alors une statistique calculée à partir des données observées sera biaisée par rapport à celle qui serait obtenue avec les mêmes variables sans erreur.

Pour prendre des exemples en Astronomie, si l'on veut calibrer la magnitude absolue d'un groupe d'étoiles en utilisant les parallaxes trigonométriques avec l'erreur relative la plus petite, cela revient à prendre les parallaxes (observées) les plus grandes, créant un biais sur les magnitudes absolues qui en résultent. C'est le biais décrit par de nombreux auteurs, notamment Trumpler & Weaver (1953) et Lutz & Kelker (1973).

Calcul du biais

Il est possible de calculer analytiquement le biais que l'on observe, et c'est heureux puisque l'on ne peut pas éviter ce biais. Pour cela, la démarche suivie est bayésienne : la loi des x est prise conditionnellement à y, et l'on suppose que l'on connaît la loi *a priori* des y.

Pour chaque variable observée x, quelle est l'espérance conditionnelle $E[Y|X]$ de la variable sans erreur y sachant x ?

La densité de probabilité conditionnelle $f(y|x)$ s'écrit par la formules de Bayes :

$$\begin{aligned} f(y|x) &= \frac{f(x|y)f(y)}{f(x)} \\ &= \frac{f(x|y)f(y)}{\int_{-\infty}^{+\infty} f(x|y)f(y)dy} \end{aligned}$$

et par définition de l'espérance mathématique

$$\begin{aligned} E[Y|X] &= \int y f(y|x) dy \\ &= \frac{\int y f(x|y) f(y) dy}{\int f(x|y) f(y) dy} \end{aligned}$$

Ici $\widehat{y} = E[Y|X]$ est l'estimateur bayésien ponctuel de la variable sans erreur y connaissant la variable x. Dans le paragraphe précédent (4.3.1), par contre, on a calculé l'estimation qui maximisait la densité conditionnelle *a posteriori*. Ce sont deux des formes d'estimation bayésienne ; on montre d'ailleurs [Aïvazian *et al.*, 1986, p. 240] que ces deux estimations convergent vers l'estimation du maximum de vraisemblance lorsque $n \to \infty$, et ceci *indépendamment du choix de* $f(y)$.

En l'appliquant à l'exemple que nous avons pris plus haut, on a

$$\begin{aligned} f(x|y) &= \frac{1}{0.5\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-y)^2}{0.5^2}} \\ f(y) &= \frac{1}{3\pi}\left(\frac{1}{1+(y-4)^2} + \frac{1}{1+(y/2)^2}\right) \end{aligned}$$

ce qui nous permet de calculer $\widehat{y}$. Quant au biais $x - \widehat{y}$, il est représenté en fonction de x sur la figure 4.8.

L'exemple que l'on a choisi montre bien le comportement du biais, combien il peut être important, mais surtout l'intérêt de l'analyse bayésienne qui permet de trouver la formulation analytique. Qu'on ne s'y trompe pas, l'estimation

$$\widehat{y} = \frac{\int y f(x|y) f(y) dy}{\int f(x|y) f(y) dy} \tag{4.5}$$

n'est pas une simple méthode pour calculer le biais, mais véritablement la mise en évidence d'un estimateur meilleur (au sens du risque) pour estimer la vraie variable, quand on connaît la variable observée, que ne l'est cette variable observée.

Ce qui pose problème dans toute approche bayésienne, c'est naturellement le choix de la distribution *a priori* et c'est d'ailleurs ce qui divise les statisticiens. Comme nous l'indiquons au paragraphe suivant, il existe néanmoins un cas particulier où la connaissance de la densité *a priori* n'est pas indispensable.

4.3.3 Estimation sans loi *a priori*

Reprenons l'estimation bayésienne des variables sans erreurs en l'écrivant

$$\widehat{y} = \frac{\int y f(x|y) f(y) dy}{f(x)}$$

Supposons maintenant que les erreurs de mesure sont distribuées normalement et indépendantes de x et y, et notons $\sigma_x^2 = s^2$ (supposant donc que les erreurs standards sont indépendantes de x). On a alors :

$$f(x|y) = \frac{1}{s\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-y)^2}{s^2}}$$

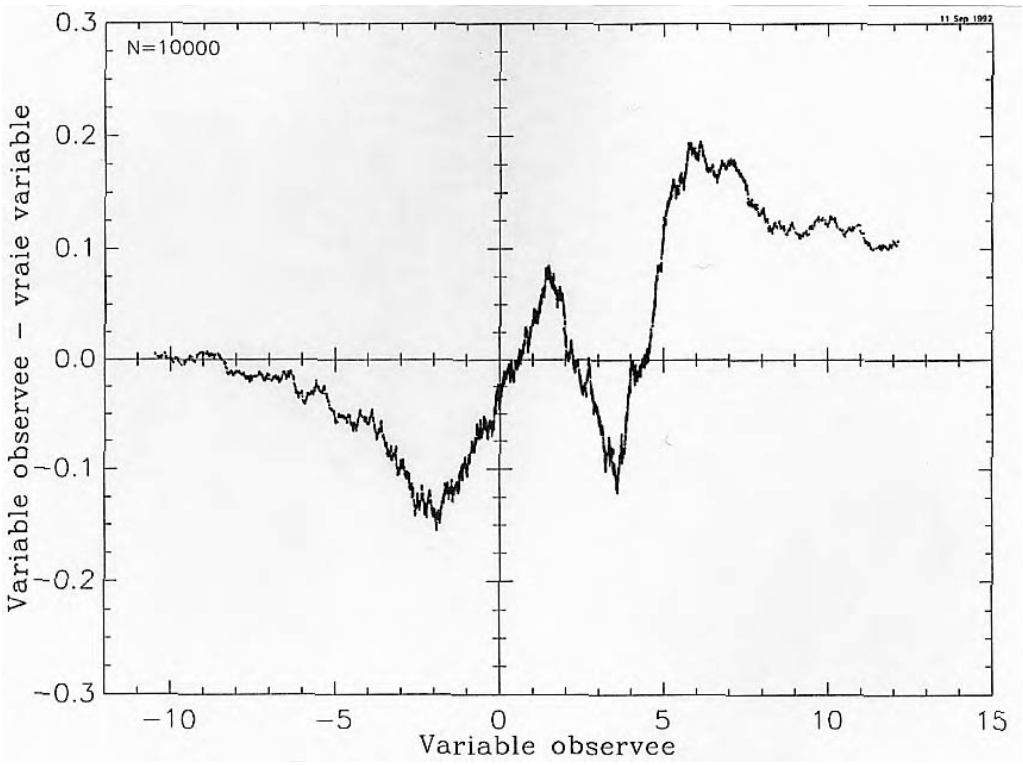


FIG. 4.7: *Lissage des différences $x - y$ en fonction des x.*

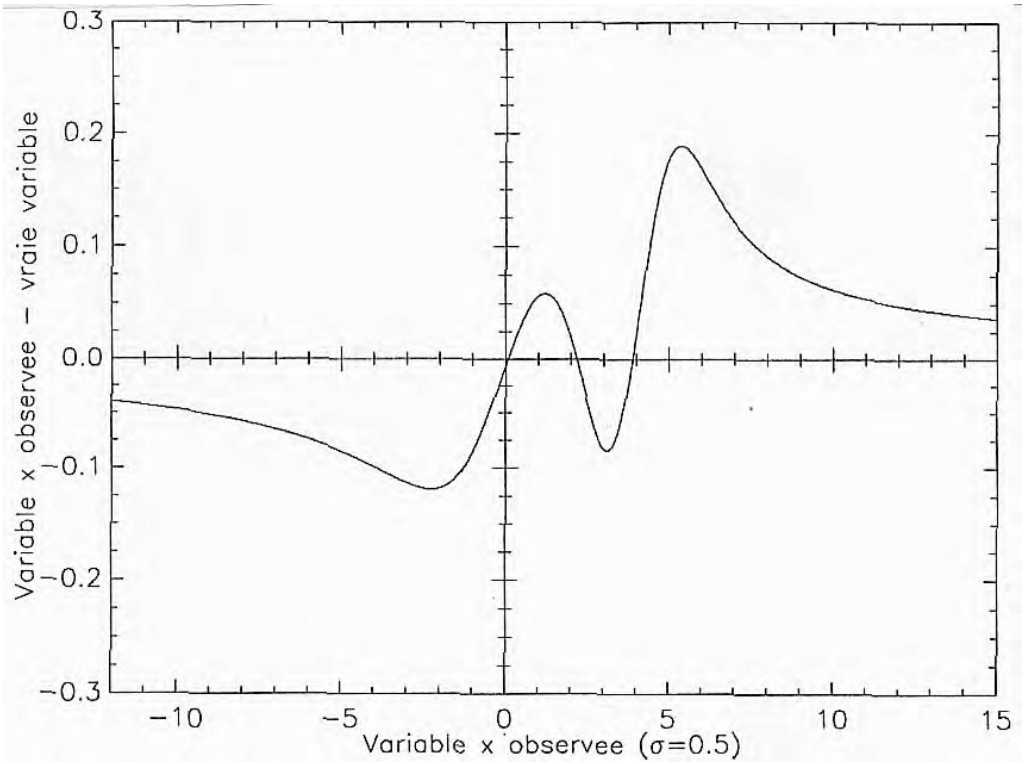


FIG. 4.8: *Biais calculé analytiquement.*

Ce qui permet de calculer sa dérivée par rapport à x, sous l'hypothèse que cette densité est dérivable en tout point :

$$\begin{aligned} f'(x) &= \int f'(x|y)f(y)dy \\ &= \int -\frac{(x-y)}{s^2}\frac{1}{s\sqrt{2\pi}}e^{-\frac{1}{2}\frac{(x-y)^2}{s^2}}f(y)dy \\ &= \int -\frac{(x-y)}{s^2}f(x|y)f(y)dy \\ &= -\frac{x}{s^2}f(x) + \frac{1}{s^2}\int yf(x|y)f(y)dy \\ &= -\frac{x}{s^2}f(x) + \frac{1}{s^2}\widehat{y}f(x) \\ &= \frac{f(x)}{s^2}(-x+\widehat{y}) \end{aligned}$$

D'où l'on déduit une expression de l'estimateur des variables sans erreurs :

$$\widehat{y} = x + s^2\frac{f'(x)}{f(x)} \tag{4.6}$$

Cette formule permet de se passer de la distribution *a priori*, puisque l'on n'utilise que la densité observée et sa dérivée. Comme le note Smart (1968), p. 36, ce résultat avait déjà été trouvé par A.S. Eddington, en utilisant une approche plus classique.

Mais, en adoptant cette optique conditionnelle, il est également possible de trouver un résultat supplémentaire qui donne la précision de l'estimateur obtenu. La variance conditionnelle s'écrivant :

$$\mathrm{Var}(\widehat{y}|x) = \frac{\int y^2 f(x|y)f(y)dy}{f(x)} - \widehat{y}^2$$

Le calcul est analogue au précédent en calculant la dérivée seconde de la densité marginale :

$$f''(x) = -s^2 f(x) - 2xs^2 - x^2 + \frac{\int y^2 f(x|y)f(y)dy}{f(x)}$$

On obtient alors après quelques calculs :

$$\mathrm{Var}(\widehat{y}|x) = s^2(1 + s^2(\frac{f'(x)}{f(x)})') \tag{4.7}$$

Cette précision sur l'estimateur de la variable sans erreur s'exprime donc elle aussi en fonction des dérivées de la densité marginale observée. On peut noter que ces estimateurs doivent être légèrement modifiés pour tenir compte du fait que les moyennes et variances sont déterminés empiriquement à partir de l'échantillon [Robert, 1992, p. 288].

Rappelons que ce résultat est valable parce que l'on a fait l'hypothèse d'une erreur gaussienne. En particulier, si l'on suppose de plus que la distribution des variables sans erreur est gaussienne, on retrouve dans 4.6 et 4.7 les expressions données respectivement par les équations 4.3 et 4.4.

La forme des eq. 4.6 et 4.7 conduit logiquement à rechercher comment obtenir un bon estimateur de la densité observée, et qui soit deux fois dérivable.

4.3.4 Estimation empirique de la densité de probabilité observée

Pour avoir une idée de la distribution des variables, la première idée est de tracer l'histogramme des données. S'il n'y a pas d'erreur grossière dans ces données, on se doute

que les vraies variables doivent avoir approximativement la forme de distribution observée. Le problème avec les histogrammes, c'est que l'on ne sait jamais quel taille d'intervalle choisir, et entre quelles bornes il faut prendre les données : c'est gênant pour estimer la densité.

Il y a plusieurs méthodes pour obtenir une estimation de la densité empiriquement. Nous en citerons trois :

– La première consiste à estimer la fonction de répartition $F(t)$ par une fonction continue $F_n(t)$ et dérivable – sauf en un nombre fini (n) de points –, et à calculer la densité comme dérivée de cette fonction $F_n(t)$. Supposons les variables observées x_i, $i = 1, \ldots, n$, triées par ordre croissant. Notons x_0 et x_{n+1} les «extrémités» de la distribution empirique, c'est-à-dire $F_n(x_0) = 0$ et $F_n(x_{n+1}) = 1$. On choisit $F_n(t)$ linéaire sur chaque segment

$$x \in [\frac{x_i + x_{i-1}}{2}, \frac{x_i + x_{i+1}}{2}]$$

telle que

$$F_n(\frac{x_i + x_{i+1}}{2}) = F_n(\frac{x_i + x_{i-1}}{2}) + 1$$

Par conséquent, on peut facilement trouver un estimateur de la densité par dérivation :

$$\widehat{f}_n(x) = \frac{2}{n(x_{i+1} - x_{i-1})}$$

si $x \in]\frac{x_i+x_{i-1}}{2}, \frac{x_i+x_{i+1}}{2}[$ et 0 sinon.

Malheureusement, cette estimation convient sur un petit échantillon mais possède beaucoup trop de variations quand on a un nombre important de points, et nécessiterait donc un lissage.

– La deuxième utilise la Distribution Lambda généralisée définie au §4.2.4. On estime les paramètres $\lambda_1, \ldots \lambda_4$ au vu des 4 premiers moments empiriques. On peut montrer facilement que la densité correspondante est

$$f(x) = f(F^{-1}(z)) = \frac{\lambda_2}{\lambda_3 z^{\lambda_3 - 1} + \lambda_4 (1 - z)^{\lambda_4 - 1}}$$

Naturellement, cette estimation est approximative et ne peut d'ailleurs convenir que si la distribution étudiée est unimodale.

– On peut obtenir une densité lissée, en adoptant la démarche [Aïvazian *et al.*, 1986, p. 273] qui consiste à utiliser la statistique

$$\widehat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} k(\frac{x - x_i}{h})$$

pour estimer la densité de probabilité, où $k(t)$ est une fonction positive, symétrique, d'intégrale 1, et tendant vers 0 quand $|t| \to \infty$. Quant à h, c'est un paramètre arbitraire, tout comme l'est d'ailleurs la largeur de l'intervalle de classe dans un histogramme.

C'est cette dernière méthode que nous avons utilisée, parce qu'elle permet d'avoir une estimation continue et dérivable de la densité observée, ce qui est important au vu des équations mises en évidence au paragraphe précédent.

Le fait que cette estimation dépende d'un paramètre h donne l'impression que l'on n'a pas beaucoup progressé par rapport à l'histogramme des données. En fait, ce paramètre est dépendant de n, et il permet d'ajuster la taille de la «fenêtre» à travers laquelle on compte le nombre d'observations.

L'estimateur que l'on a cité ci-dessus est un cas particulier d'une méthode dite du noyau de convolution [Bosq & Lecoutre, 1987]. Cette méthode a reçu de nombreux développements théoriques, et il est en particulier possible d'obtenir – de façon compliquée, certes – une première estimation de h_n qui assure la convergence la plus rapide vers la vraie densité. Si l'on note σ' la valeur la plus petite entre l'écart-type empirique et la distance semi-interquartile de l'ensemble de la distribution, cette estimation est $h_n = \sigma' n^{-\frac{1}{5}}$. Puisque l'on connaît la loi des erreurs, on va choisir une estimation voisine qui va lisser la densité en supprimant les pics non significatifs :

$$\widehat{f_n}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sigma_{i,n,x}\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-x_i)^2}{\sigma^2_{i,n,x}}}$$

où $\sigma_{i,n,x} = h_n \frac{\sigma_{x_i}}{\langle \sigma_{x_i} \rangle}$

Reprenant l'exemple du §4.3.2, on montre sur la figure 4.9 l'allure de la densité lissée obtenue à partir de la simulation effectuée, à comparer avec l'histogramme tracé figure 4.6. Le biais calculé avec la formule 4.6 et cette densité lissée peut alors être calculé, et nettement réduit, comme on peut le voir sur la figure 4.10, puisqu'il n'est plus significativement différent de 0.

Il s'agit donc de démarches empiriques, qui ont l'avantage d'être non paramétriques (la loi des variables n'étant pas présupposée), et qui peuvent s'avérer utiles si l'on veut estimer la densité d'une population à partir d'un échantillon représentatif. Mais, dans les trois cas cités, on estime la densité observée qui n'est pas la vraie densité si la variable observée a une erreur de mesure. Si celle-ci est gaussienne, l'équation 4.6 permet de s'affranchir de ce problème.

Dans le cas général, donné par l'équation 4.5, il nous faut connaître la densité *a priori*. Or, en général, on ne la connaît pas, ou on n'en a qu'une vague idée. Problématique également est le cas où la distribution générale est connue, mais où l'on sait que l'échantillon dont on dispose a été mal sélectionné, et ne représente pas la population générale.

A titre d'exemple, on peut faire allusion à un échantillon d'étoiles dont on veut calibrer la magnitude absolue à partir des parallaxes trigonométriques : il est courant de supposer que la distribution spatiale des étoiles est uniforme, mais, dans la pratique, on ne peut pas affirmer que l'échantillon est *complet* en distance (on ne connaît pas la vraie parallaxe, sinon le problème serait règlé) ou complet en magnitude (à cause de l'absorption, entre autre). Ainsi, utiliser une distribution théorique *a priori* sans tenir compte de ces problèmes peut faire plus de mal que de bien. C'est pourquoi on utilisera fréquemment, d'une façon ou d'une autre, les distributions observées qui, quand le nombre d'étoiles est très important, contiennent plus «d'information» qu'une distribution théorique, et notamment les biais de sélection. Naturellement les distributions observées sont marginales, et, sauf si l'erreur de mesure est négligeable, ce qu'il faudra en réalité c'est trouver une distribution *a priori* dont la convolution avec la loi d'erreur donne la distribution observée.

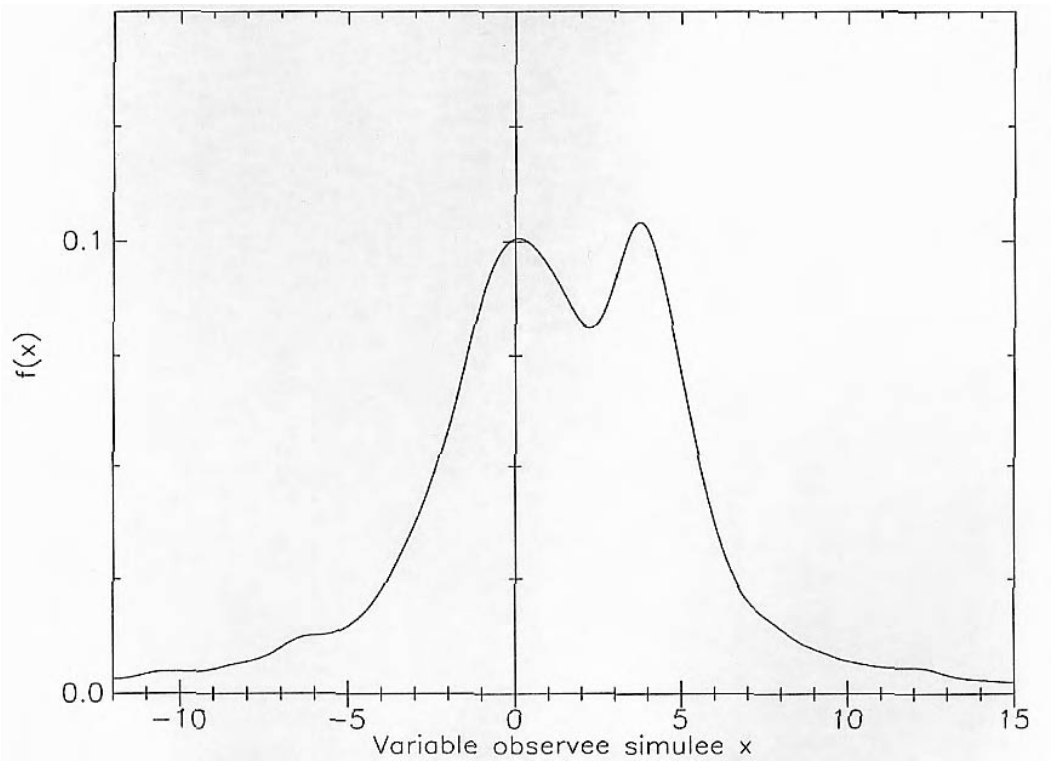


FIG. 4.9: *Densité lissée obtenue à partir des données simulées.*

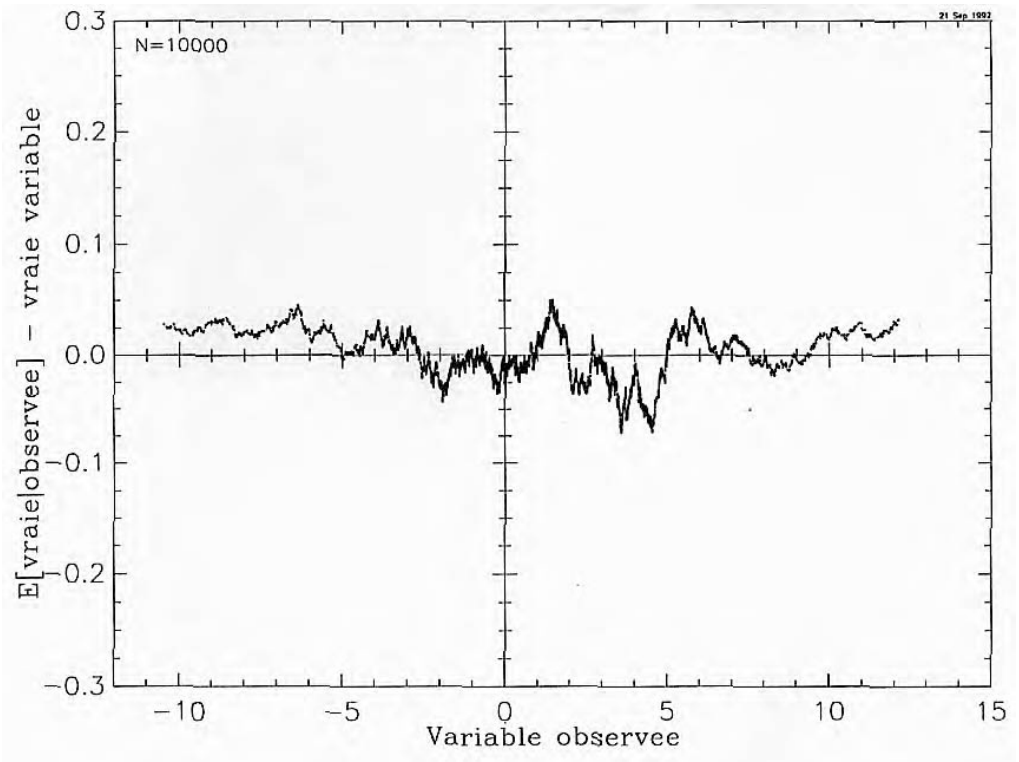


FIG. 4.10: *Lissage des différences* $(x + s^2 \frac{f'(x)}{f(x)}) - y$ *en fonction des* x*.*

4.4 Estimations multivariées

Nous abordons ici le problème de la prise en compte des erreurs de mesures, mais dans le cadre multidimensionnel, avec tout d'abord le problème de la séparation de composants gaussiens dans un échantillon.

4.4.1 Mélange de populations gaussiennes

Introduction

Il arrive fréquemment que l'on ait besoin de mettre en évidence et de séparer différentes populations dans un échantillon.

En particulier, dans le cadre d'un travail en cinématique stellaire (cf. chapitre 7), nous avons été amené à utiliser des méthodes statistiques permettant de séparer des populations d'étoiles, de manière paramétrique ou non. Nous nous intéresserons dans ce paragraphe à une méthode paramétrique, la séparation de populations gaussiennes.

Jusqu'à présent, la séparation de composants gaussiens avait été traitée [Soubiran, 1988], mais sans tenir compte des erreurs de mesure des données cinématiques, et nous nous sommes donc intéressé à l'implication des erreurs de mesures sur la séparation des populations. Bien évidemment, les algorithmes mis au point pourront s'appliquer à des domaines beaucoup plus vastes.

Pour situer le problème, précisons que l'on constate, de façon multidimensionnelle, la présence de plusieurs modes dans la distribution des variables étudiées, les vitesses d'étoiles relatives au soleil en l'occurence. Il y a donc un mélange de lois, que l'on suppose chacune gaussienne, de moyenne et de dispersion différentes, représentant des groupes d'étoiles au comportement cinématique différent. Tout le problème statistique consiste à rechercher le nombre de composants du mélange, estimer les paramètres des composantes (proportions, moyennes, variances), et savoir affecter une observation au composant auquel elle appartient. On s'intéresse donc ici à une méthode complètement paramétrique, mais dont le nombre de paramètres n'est pas déraisonnable.

Pour nuancer l'affirmation précédente, précisons que le problème d'estimation n'est pas si simple : pour m composants en dimension k, le nombre de paramètres à estimer est $m\frac{(k+1)(k+2)}{2} - 1$ puisque pour la moyenne de chaque composant, il y en a k, pour chaque matrice de variance-covariance, $\frac{k(k+1)}{2}$, et $m-1$ proportions de chaque composantes. Par exemple, pour les vitesses spatiales, avec 3 composants gaussiens, cela fait 29 paramètres à déterminer. Inutile de dire que pour un petit échantillon, il est difficile de parvenir à une bonne solution.

Bien que l'on puisse utiliser la méthode des moments (égaler les moments théoriques et empiriques) pour résoudre ce problème, la méthode la plus utilisée, et la plus efficiente, est le maximum de vraisemblance, et en particulier l'algorithme EM (Estimation-Maximisation), [Redner & Walker, 1984] dont on décrit brièvement le principe dans l'article page 86. Cette méthode nécessite malheureusement la connaissance *a priori* du nombre de composants du mélange, et d'une solution initiale. De plus, la vitesse de l'algorithme est fortement dépendante de cette solution initiale, et enfin il peut passer par un maximum pathologique, notamment si la proportion d'un des composants est petite, ou si les composants sont mal séparés.

Pour répondre à ces limitations, Celeux et Diebolt (1986) ont développé l'algorithme stochastique SEM (SEMMUL dans le cas multidimensionnel), qui présente la même structure que l'algorithme EM, mais avec une étape d'apprentissage probabiliste. À part pour de très petits échantillons – pour lesquels les perturbations aléatoires perturbent réellement la solution (!) – SEM n'a pas les problèmes de lenteur de convergence de EM, estime correctement le nombre de composants (il suffit simplement de lui donner un majorant de ce nombre), n'a pas tendance à rester près d'un «col» de la vraisemblance, et surtout ne nécessite pas de conditions initiales.

Variante de l'algorithme stochastique SEM, SAEM [Celeux & Diebolt, 1989] est un algorithme de type recuit simulé (la solution est contrainte au fur et à mesure des itérations) qui converge presque sûrement vers un maximum local de la vraisemblance, et qui est plus adapté aux petits échantillons.

L'approche bayésienne classique de ce problème d'estimation est peu utile puisque toutes les partitions possibles de l'échantillon doivent être prises en compte, conduisant à des temps de calcul extrêmement prohibitifs. Il existe néanmoins une alternative, l'échantillonnage bayésien [Robert, 1992], qui permet, si le nombre de composants de l'échantillon est connu, d'obtenir les estimations bayésiennes des paramètres [Diebolt & Robert, 1990], [Robert & Soubiran, 1991].

Notre problème était donc le suivant : nous devions tenter de séparer des populations grâce aux composantes de la vitesse spatiale, mais celles-ci souffrent d'une erreur de mesure plus ou moins importante, cette erreur dépendant de la précision sur les mouvements propres, la vitesse radiale, et surtout sur la distance. Cette erreur variant beaucoup d'une étoile à l'autre, il est prévisible que la détermination des composants gaussiens que l'on recherche va en souffrir.

G.Celeux et J.Diebolt (1989a, 1989b) ont développé, à partir de l'algorithme SEM, une méthode tenant compte de ces erreurs, que nous avons implémentée et testée, à la fois pour EM, pour SEM, et pour SEMMUL. Avant l'utilisation de cette méthode sur des données réelles, le paragraphe suivant montre, à l'aide de deux exemples, la façon dont se comportent les algorithmes.

Simulations

Nous nous limiterons, pour ces simulations, à deux dimensions, ce qui simplifie largement la visualisation des résultats. Nous générons deux populations de points (en proportion 50%/50%), et testons la reconnaissance, par les différents logiciels cités plus haut, des paramètres des deux populations. À dire vrai, pour les logiciels en question, les tests ci-dessous s'apparentent plus à une torture qu'à une mise en valeur ; en effet, l'échantillon simulé ne contient que 100 points et les populations se recouvrent partiellement ; à voir les figures 4.11 et 4.12, les deux populations semblent peut-être séparées, mais les distributions marginales montrent deux modes fort peu éloignés, et ce qui semble évident avec l'œil et la connaisance du contenu de l'échantillon ne l'est pas forcément pour une reconnaissance automatique. De plus, il faut noter que l'on teste en même temps les qualités du générateur aléatoire.

Pour les deux tests montrés ici, on a donc simulé un échantillon composé de deux populations de 50 points et dont les valeurs sont du même ordre de grandeur que des vitesses spatiales ; les erreurs standards de mesure sont en moyenne 7 et les moyennes des

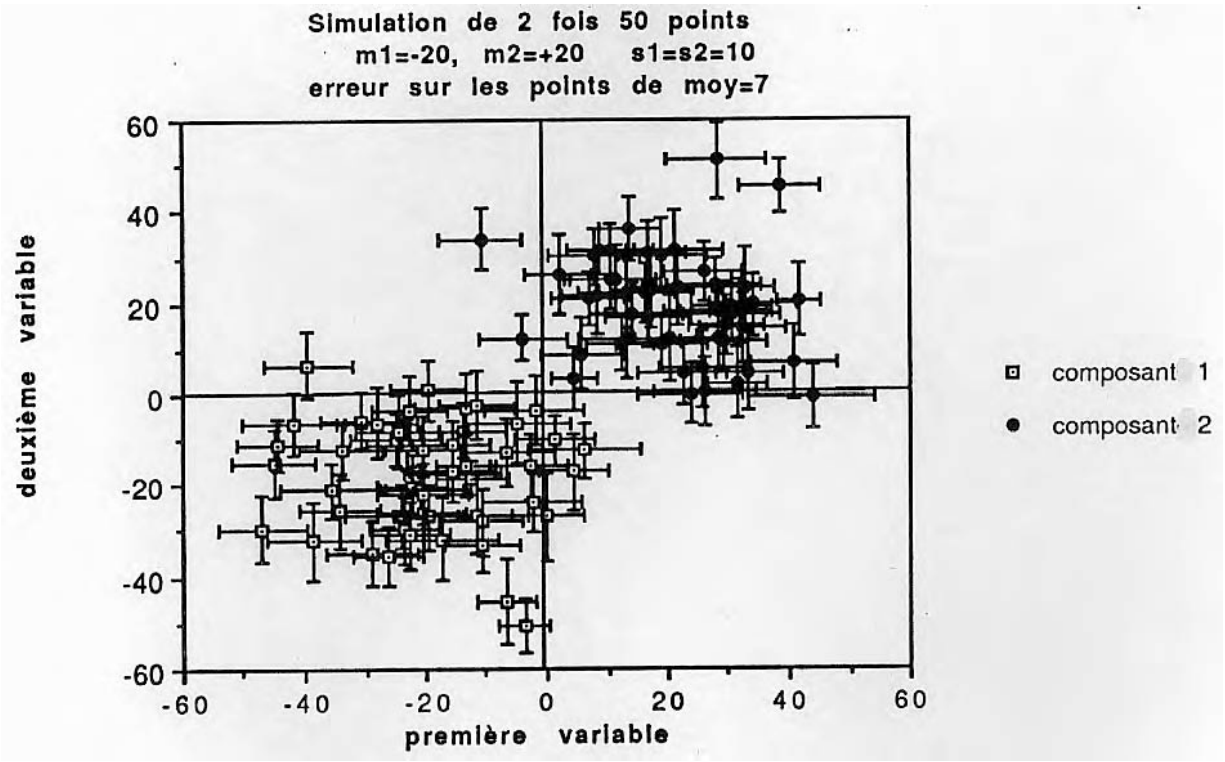


FIG. 4.11: *Simulation des 2 populations* $\mathcal{N}(-20, 10^2)$ *et* $\mathcal{N}(20, 10^2)$.

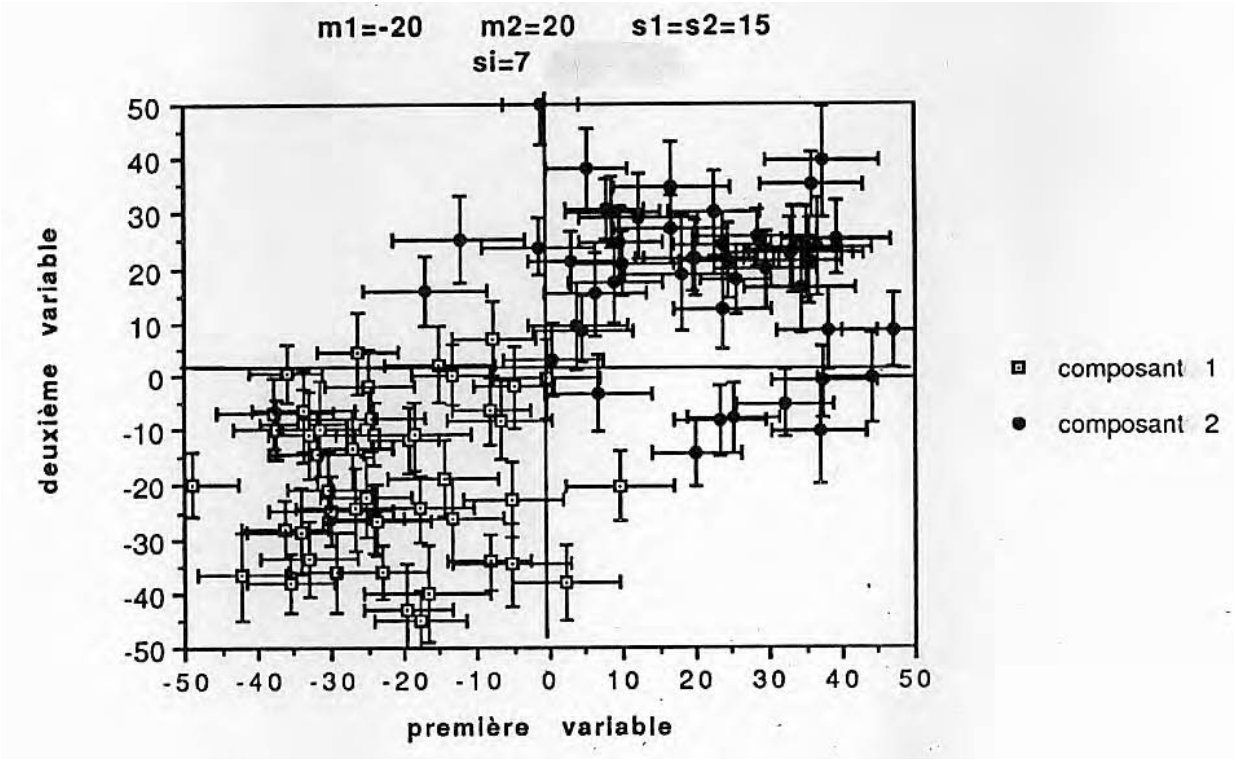


FIG. 4.12: *Simulation des 2 populations* $\mathcal{N}(-20, 15^2)$ *et* $\mathcal{N}(20, 15^2)$.

deux populations sont respectivement -20 et 20 ; l'écart-type de chaque population est 10 dans le premier test et 15 dans le second.

Rappelons que les différents logiciels testés ne sont pas sur le même pied d'égalité : par rapport à EM, SEM doit découvrir le nombre de composants (1,2 ou 3) ; dans le cas multidimensionnel, il y a deux fois plus de points, mais encore plus de paramètres à deviner ; quant aux versions avec gestion des erreurs de mesure, elles doivent en plus simuler des données manquantes (l'erreur de mesure en chaque point, dont on ne connaît que l'écart-type).

Alors que EM converge toujours vers une même solution si on l'initialise aux mêmes valeurs, les programmes SEM et SEMMUL, qui ont une étape stochastique, ne fournissent pas forcément les mêmes solutions d'une exécution à l'autre. On a donc effectué plusieurs exécutions et indiqué un résultat moyen sur les tableaux ci-dessous. Ces résultats sont donc naturellement à prendre à titre indicatif, le chiffre significatif après le point décimal étant peut-etre superflu [5].

Pour le premier test, les paramètres trouvés par les différents programmes pour la première variable sont sur le tableau 4.2. Pour les programmes multidimensionnels (SEMMUL et SEMMUL avec gestion des erreurs), on n'a indiqué que la solution trouvée pour la première variable, et non les matrices des moyennes et les matrices des variances-covariances.

On peut noter très nettement que la gestion des erreurs et les versions multidimensionnelles permettent de mieux séparer les deux composants, la solution la plus proche de la vraie solution (50,-20,10 ; 50,20,10) étant obtenue avec la version SEMMUL avec gestion des erreurs. C'est un résultat encourageant.

Les résultats du deuxième test, là où les deux populations commencent à se recouvrir, sont indiqués sur le tableau 4.3. Les résultats sont ici plus médiocres : il n'y a pas de solution pour SEM, qui ne trouve souvent qu'un composant, et les programmes avec gestion des erreurs trouvent une première population systématiquement trop petite. Ceci n'est à vrai dire pas très étonnant à la vision de la figure 4.12.

On pourrait tester de façon plus exhaustive les logiciels, en calculant les résultats moyens sur quelques centaines de simulations, en faisant varier la taille de l'échantillon, le nombre de composants, leur degré de recouvrement, la taille moyenne et la variation des erreurs de mesure, et nous n'avons montré qu'un aperçu du comportement des logiciels testés. Dans un premier temps, cela indique tout de même leur capacité à être proche des vraies solutions dans des conditions de tests un peu draconiennes.

Une autre question restée en suspens est la stabilité des résultats pour les algorithmes SEM et SEMMUL. Un élement de réponse se trouve au paragraphe suivant.

Stabilité des algorithmes

Restait donc à vérifier à la fois la stabilité des résultats obtenus par SEMMUL, et leur cohérence avec les résultats d'autres méthodes statistiques (analyse factorielle, classification). Cette stabilité a été étudiée à l'aide de données observées, et non simulées.

L'article ci-joint [Bougeard & Arenou, 1989] concerne la séparation en deux groupes d'un échantillon d'étoiles de type A2V du voisinage solaire, à l'aide des composantes

5. «*Dans toute statistique, l'inexactitude du nombre est compensée par la précision des décimales.*» G. Elgozy

TAB. 4.2: *Séparation de populations gaussiennes.*

Paramètres obtenus par différents programmes de séparation de composants gaussiens avec 2 populations $\mathcal{N}(-20, 10^2)$ et $\mathcal{N}(20, 10^2)$.

Programme	1ère population			2ème population		
	$\%_1$	m_1	s_1	$\%_2$	m_2	s_2
EM	55	-16.8	14.7	45	22.3	10.7
EM+erreur	53	-17.2	12.5	47	21.6	9.3
SEM	55	-16.9	14.9	45	22.6	10.9
SEM+erreur	50	-18.5	12.1	50	20.7	10.3
SEMMUL	50	-19.0	13.3	50	21.2	11.7
SEMMUL+erreur	50	-19.0	12.1	50	20.8	10.3

TAB. 4.3: *Séparation de populations gaussiennes.*

Paramètres obtenus par différents programmes de séparation de composants gaussiens avec 2 populations $\mathcal{N}(-20, 15^2)$ et $\mathcal{N}(20, 15^2)$.

Programme	1ère population			2ème population		
	$\%_1$	m_1	s_1	$\%_2$	m_2	s_2
EM	44	-25.3	12.5	56	18.0	18.1
EM+erreur	41	-26.4	9.8	59	16.6	17.6
SEM						
SEM+erreur	47	-24.4	11.6	52	19.8	16.1
SEMMUL	49	-23.9	13.5	51	20.7	16.8
SEMMUL+erreur	43	-25.8	5.7	57	18.4	14.4

(U, V, W) de la vitesse spatiale. La finalité de cette séparation est expliquée en détail dans le chapitre 7.

Après avoir testé la stabilité des solutions trouvées lors de plusieurs exécutions indépendantes du programme SEMMUL, on compare les résultats obtenus avec ceux d'une analyse en composante principale, d'une classification ascendante hiérarchique (algorithme des n plus proches voisins), et d'une analyse discriminante linéaire.

Concernant toujours la comparaison des résultats de SEMMUL avec les résultats de méthodes de classification non paramétriques, on montre dans une autre publication [Arenou, 1990], par classification à l'aide d'agrégation autour de centres mobiles [Lebeaux, 1986], que 97% des étoiles se retrouvent dans les classes trouvées par SEMMUL. On montre également [Arenou & Bougeard, 1992], qu'une classification ascendante hiérarchique est en accord avec les classes fournies par SEMMUL pour 92% des étoiles.

Les résultats obtenus paramétriquement apparaissent donc très fiables, sauf dans le cas des petits échantillons (ici un échantillon de 31 étoiles Ap), où la stabilité des résultats

n'est pas garantie.

Errors, bias and uncertainties in astronomy

EDITORS
C. Jaschek
Strasbourg Observatory, Strasbourg, France
and
F. Murtagh
European Southern Observatory, Garching, West Germany

Multivariate mixture distributions in stellar kinematics: *Statistical and numerical stability of the SEM algorithm*

Bougeard M.L.(1)(2), Arenou F.(3)

(1) Univ. Paris X, IUT, 1 chemin Desvallières, F-92410 Ville d'Avray
(2) URA 1125 CNRS, Observatoire de Paris, F-75014 Paris
(3) URA D0335 CNRS, Observatoire de Paris-Meudon, F-92195 Meudon

Summary :*In this paper, we are concerned with the problem of estimating the parameters of a gaussian mixture density. Here we tackle the problem of analysing the convergence stability of the SEM process by performing several independent runs. Then, the results of the most stable SEM solution are compared to classical clustering and classification techniques. The method is applied to samples of A type population I stars.*

Keywords: - *Gaussian mixture - Maximum likelihood - Stellar kinematics*

1. Introduction, notations and statistical background

Of interest in this paper is the parametric family of mixture of k normal multivariate densities, i.e the family of density functions of the form

$$f(x,\theta) = p_1 f_1(x|m_1,\Sigma_1) + \ldots\ldots + p_k f_k(x|m_k,\Sigma_k), \quad x \in \mathbb{R}^n$$

where the proportions p_i, i=1-k are constrained to be nonnegative and to amount to one. Each component f_i, i=1-k of the mixture is a n-multivariate gaussian density with n-vector mean $m_i \in \mathbb{R}^n$ and (n x n) covariance matrix Σ_i.. Much has been written on methodology for estimating the unknown parameters $\theta(k) = (p_i, m_i, \Sigma_i, i=1\text{-}k)$. For a review, we refer to (Titterington & al 1985; Redner &Walker 1984). A class of iterative procedures for numerically approximating maximum likelihood estimates is known as EM algorithm.which is an algorithm used for incomplete data problems (Dempster & al 1977).

It acts as follows for a N-sample of observations $(x_j, j=1\text{-}N)$, N>>k, $x_j \in \mathbb{R}^n$. Let $\theta^c = (p_i^c, m_i^c, \Sigma_i^c, i=1\text{-}k)$ be a current approximate maximizer of the log-likelihood function of the sample $L(\theta)$, and θ^{c+1} the next one.The **Expectation** step computes for i=1-k , j=1-N,

$$p^c(i,x_j) = p_i^c f_i(x_j|m_i^c, \Sigma_i^c) / f(x_j,\theta^c)$$

that is an estimate of the posterior probability that x_j belongs to the ith component given the approximate estimate θ^c. Then, the **Maximization** step of the EM algorithm yields θ^{c+1} maximizing L(.), given by

$$p_i^{c+1} = (1/N)\sum_{j=1}^{N} p^c(i,x_j) \qquad m_i^{c+1} = \left\{ \sum_{j=1}^{N} p^c(i,x_j).x_j \right\} / \left\{ N p_i^{c+1} \right\}$$

$$\Sigma_i^{c+1} = \left\{ \sum_{j=1}^{N} p^{c}(i,x_j).(x_j-m_i^{c+1})(x_j-m_i^{c+1})^t \right\} / \left\{ Np_i^{c+1} \right\}$$

The EM algorithm (Redner & Walker 1984) possesses several attractive properties (low computational cost, convergence, constraints on θ satisfied) compared to that of several alternative methods (Newton, scoring, quasi-Newton) for numerically approximating maximum likelihood estimates. The SEM algorithm, described in (Celeux &Diebolt, 1986), is a recent improvement of this algorithm : it incorporates a Stochastic step to accelerate the (a priori low) convergence. Nevertheless, even in the context of gaussian univariate mixture, the resulting likelihood surface is littered with singularities (Titterington & al 1985, ex. 4.3.2, p83). However, with a *good initialization* - obtained, for instance through graphical methods (Bougeard & al, 1989b)- and reasonable *sample size*, one can expect a (S.)E.M. iteration sequence to converge to a *local maximizer* of the log-likelihood function.

Of interest here is the convergence stability of the SEM algorithm in application to the study of the 3 dimensional x=(U,V,W) velocity distribution of 2 samples of A type population I stars, defined in Grenier & al (1985): an A2V sample (N=97) and an Ap sample (N=36).

2. Pertinence of a gaussian mixture model for the A2V sample

On a bidimensional graph UxV, one can foresee the presence of a potential mixture of two populations. The pertinence of the use of a parametric *gaussian* mixture model was shown in (Bougeard & al, 1989c) using the U component of the velocity. Nevertheless, it is known to be unsufficient to check univariate k' gaussian mixture for each variable U,V,W in order to be able to reject the possibility of a k mixture, k>k' (for example, see Titterington & al, 1985, fig 4.10, p68). So, a multivariate analysis has to be performed.

3. Numerical stability of the SEM algorithm

Assuming that the distribution of the (U,V,W) velocity sample is a mixture of multivariate gaussian components, we study the convergence stability of the SEM algorithm by performing several independent runs : each run represents a 200 iteration sequence.

3.1. Firstly, 31 runs of the SEM algorithm have been performed for the A2V sample using an initialization with K=3 as upper bound of the number of components. Fig 1 shows the respective estimates in U found at each run. A 3 mixture solution appears as very unstable : 19 runs lead to a two mixture solution, 6 runs find no mixture.

3.2. At this stage, the same process has been performed by initializing SEM with K=2. The results are summarized on Fig 2 for the proportions (p_1>p_2) found at each run and on Fig 3 for the distributions in U,V,W.Table 1 gives the most stable solution (21 runs over 31) .Due to the fact that it is a multidimensional analysis, the result is slightly different in the U estimations from those obtained in an univariate context by Soubiran & al (1989), Bougeard & al (1989c). For interpretation in terms of star formation bursts, see Gómez & al (1989).

4. Statistical stability of the SEM estimates

The SEM algorithm provides also the probability for each star to belong to one of the estimated components (see Section 1). We compare the resulting classification with the results of classical multivariate data analysis methods.

Firstly, a Principal Component analysis (PCA) was performed on the correlation matrix (variables:U,V,W), by which it became apparent that the first axis (53% of the variance) was highly correlated with U,V lying in the galactic plane. Axis 2 (33.5% of the variance) is correlated with W perpendicular to this plane.The centers of the two gaussian components, projected as supplementary points, are highly correlated with the first axis and 7 stars are not in the same class if we perform a SEM univariate classification only on the U component.

A hierarchical classification was also performed with the reciprocal neighbour algorithm (Lebeaux, 1986; Lebart & al, 1984). The two clusters obtained by the top-level of the hierarchy are in good agreement with the SEM clusters (Bougeard & al, 1989a,b).

Table 1 : *Sample A2V (U,V,W) - SEM stable solution*

Component #1	Component #2
Proportion : 0.64	Proportion : 0.36
$m_1(u)$ $m_1(v)$ $m_1(w)$	$m_2(u)$ $m_2(v)$ $m_2(w)$
-20.8 -14.3 -6.8	11.4 1.7 -7.4
variance-covariance matrix	variance-covariance matrix
176.9 10.8 17.5	50.2 3.8 -13.2
10.8 103.4 -21.6	3.8 33.6 -3.8
17.5 -21.6 75.3	-13.2 -3.8 51.2

Fig. 1: *Sample A2V(U,V,W)- SEM results per run, initialization with K=3*
Graph of $m_i(u)$ per run, the error bar is the square root of the respective variance $v_i(u)$ in Σi

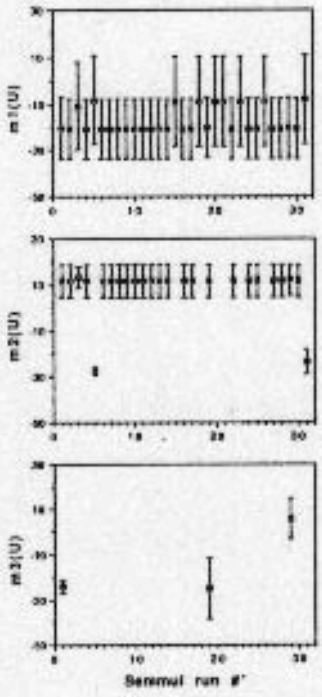


Fig. 3 : *Same as Fig 1, initialization with K=2*

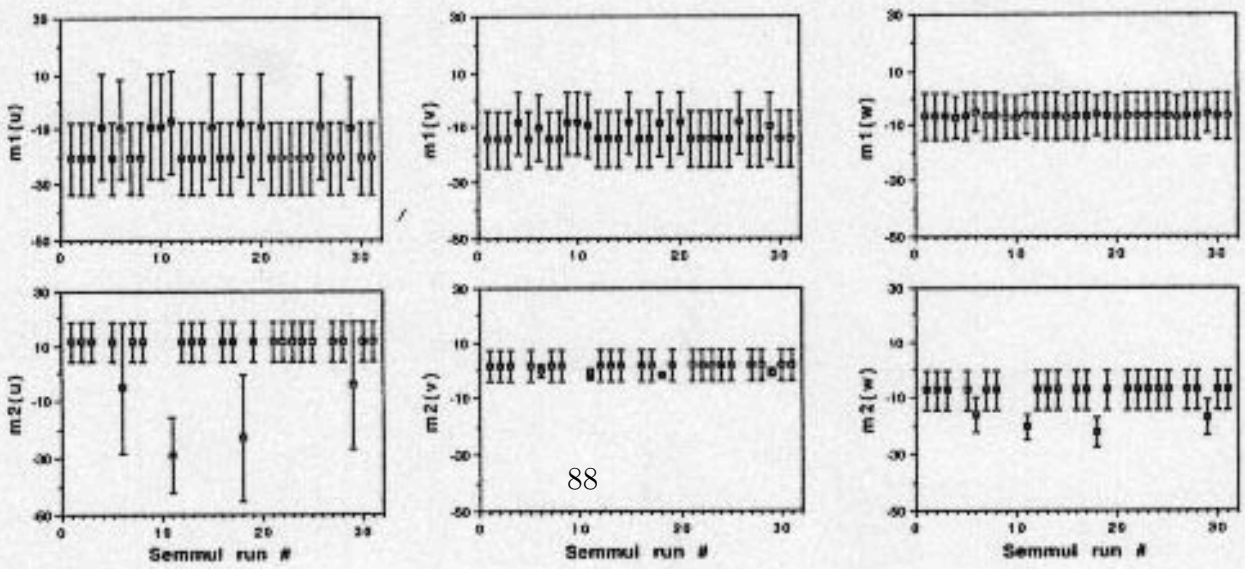


Fig.2 :Sample A2V(U,V,W), SEM (p_j) results per run, initialization with k=2

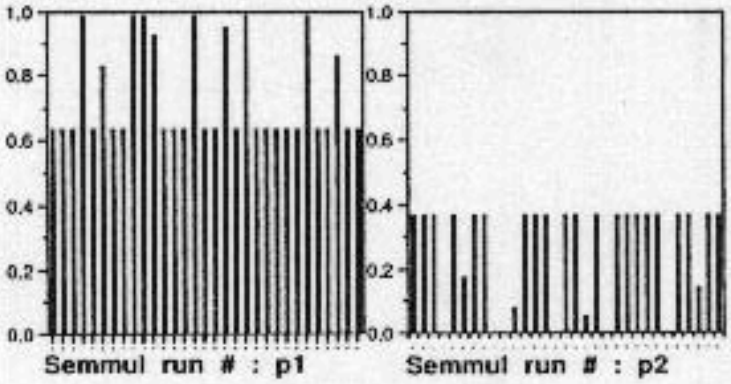
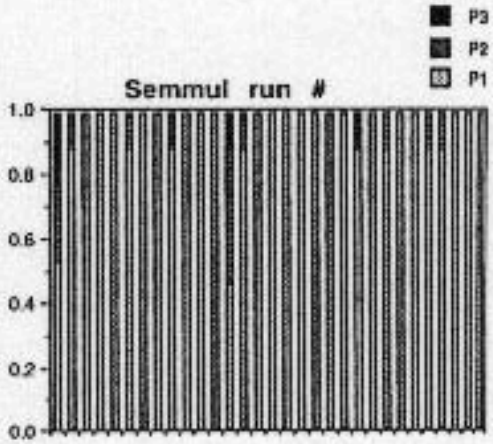


Fig.4:Sample Ap(U,V,W), SEM (p_j) results per run, initialization with K=3



Finally, a linear discriminant analysis based on Fisher linear discriminant function and Mahalanobis distance was also performed to assess the discrimination between the two groups found by SEM . Only 15 stars which were on group 1 according to SEM are affected to group 2 ; this yields to an agreement of 84.5% of well classified stars.

5. Sensitivity of the SEM algorithm to the sample size

Finally, 31 SEM runs have been performed on the Ap sample (N=36), using an initialization with K=3 as upper bound of the number of the components. Two components were expected (Gómez & al,1989), but Fig 4 shows a high instability (no mixture is found in 20 runs over 31). The main reason is that the sample size is far too small and components are overlapping too much. We note, in the studied application, that a 3 (resp. 2) mixture model yields a SEM estimation of 2+3x3+3x6=29 (resp. 19) unconstrained parameters.

6. Conclusion

If the sample size is large enough and if the components are well separated, the SEM algorithm has been seen to provide a reasonable good convergence in the estimation of the parameters of gaussian mixtures in stellar kinematics. But it cannot be used rashly in other cases. In the particular case of the A2V sample studied here, SEM results have appeared as nearly stable and in good agreement with other clustering and classification techniques.

7. Acknowledgements

Principal component analysis, hierarchical classification and discriminant analysis were performed with SAS-ADDAD software on an IBM computer at CIRCE (F-Orsay). We thank Dr Celeux and Dr Diebolt (INRIA, F-Rocquencourt) for allowing the use of SEM software.

8. References

Bougeard M.L., Arenou F., Gómez A.; 1989a Bull.47th Int. Stat. Inst.: contr. papers, vol1,p161-162 , Paris
Bougeard M.L., Arenou F., Gómez A.;1989b, "Mélanges gaussiens en cinématique stellaire. Une approche comparative de méthodes paramétriques et non paramétriques"(in preparation)
Bougeard M.L., Arenou F., Soubiran C., Gomez A., Grenier S.;1989c, (this issue)
Celeux G., Diebolt J.;1986, Revue Stat. Appl., 34, n°2,
Dempster A., Laird N., Rubin D.;1977,J. Royal Stat. Soc. B, 39, p1-38
Gómez, Delhaye, Grenier, Jaschek, Jaschek 1989: Astr. &Astroph. (soumis)
Grenier S. ,Gómez A., Jaschek C., Jaschek M., Heck A.: 1985 Astron. & Astrophys. 145, 331
Lebart L., Morineau A., Warwick K.;1984," Multivariate Descriptive Statistical Analysis", Wiley
Lebeaux M-O.;1986 Manuel de référence ADDAD&AS, CIRCE, Orsay, france
Redner R., Walker H.;1984 SIAM,26,n°2, p195-239
Soubiran C., Bougeard M.L., Gomez A., Arenou F.; 1989 (this issue)
Titterington D., Smith A., Makov H.;1985, "Statistical Analysis of finite mixtures", Wiley

4.4.2 Estimations par les moindres carrés

L'estimation par moindre carrés est une forme d'estimation classique en Astronomie, et d'ailleurs originaire de ce domaine. En effet, d'après Aïvazian *et al.* (1986), elle apparaît en 1805, lors du travail de Legendre sur les «Nouvelles méthodes de définition des orbites des comètes», puis justifiée théoriquement par Gauß en 1809 et 1821. Elle permet d'obtenir les estimateurs non biaisés qui, pour certaines distributions, ont les variances les plus petites possible. Malheureusement, l'estimation n'est pas très robuste, en règle générale.

Cherchant non seulement la robustesse, mais également la prise en compte des erreurs de mesures pour chacune des variables dans l'estimation par les moindres carrés, nous avons trouvé un excellent logiciel développé pour le Télescope Spatial (en particulier la réduction astrométrique de plaques photographiques) : Gaussfit. Ce logiciel est gratuit et accessible (par la méthode de transfert de fichiers classique sous Unix : «anonymous ftp») à l'adresse `bessel.as.utexas.edu`.

Le principal avantage du logiciel est qu'il permet de spécifier un modèle quelconque à l'aide d'un langage structuré analogue au langage C. La souplesse d'utilisation est un atoût : il n'y a pas de limite à la complexité du modèle que l'on formule. Les équations aux conditions, éventuellement non linéaires, sont données implicitement, et le logiciel calcule analytiquement les dérivées partielles. Enfin, utilisant les travaux de Huber (1981), l'algorithme est capable de donner des estimations robustes des paramètres.

Nous le citons ici car nous l'avons appliqué pour partie dans l'estimation des paramètres du modèle de l'extinction interstellaire décrit au chapitre 3.

4.5 Conclusion

Nous n'avons pas tenté de faire, dans ce chapitre, le tour d'horizon de tout ce qui concerne les erreurs de mesure, mais plus simplement d'analyser quelques outils qui peuvent s'avérer utiles. Dans la plupart des cas, nous indiquons la bibliographie qui s'y rapporte et permettrait d'approfondir les questions traitées.

Différentes méthodes ont donc été analysées afin de répondre à des problèmes concrets auxquels nous avons été confronté. Nous n'avons fait d'ailleurs qu'effleurer le thème de l'estimation statistique, mais en utilisant certaines ressources qu'offre ce vaste sujet :

- comment trouver les meilleurs estimateurs, ou les plus robustes, dans le cas gaussien, et les valider par simulation,
- comment calculer les estimateurs conditionnels des variables sans erreur,
- obtenir des estimations de la densité,
- mettre en évidence et corriger des biais dûs aux erreurs de mesure,
- séparer des populations gaussiennes affectées d'erreurs.

La quasi-totalité des procédures exposées ici a ensuite été codée en langage C, et forme une bibliothèque dont les modules sont décrits en annexe, page 209.

On peut d'ailleurs se demander pourquoi réécrire ce que l'on peut trouver dans des bibliothèques existantes. Dit sous forme de boutade, il faut bien avouer que les bibliothèques

de programmes statistiques sont souvent tellement générales que l'on n'y trouve pas la solution précise à des problèmes particuliers.

Une partie des méthodes exposées dans cette partie sont d'ailleurs, soit dispersées dans des bibliothèques multiples, soit, à notre connaissance, peu répandues. De plus, la simplicité d'utilisation, le prix, la clarté de leur documentation et des résultats qu'elles fournissent ne font en général pas partie des qualités premières des bibliothèques informatiques existantes. À tous égards, les *Numerical recipes* [Press, 1990] font exception à la règle et certaines de leurs procédures ont donc été utilisées.

Les réponses indiquées dans ce chapitre pourront ainsi nous être utiles ailleurs dans cette thèse ; en particulier, en ce qui concerne Hipparcos, même si l'on ne pourra jamais connaître les vraies parallaxes des étoiles que l'on étudie[6], les méthodes utilisées permettront d'en améliorer l'estimation.

6. «*La statistique est un bikini. Ce qu'elle révèle est suggestif, ce qu'elle cache est vital.*» A. Kœstler

4.6 Annexe

Démontrons les propriétés annoncées au §4.2.1 :

– On a $x_i = y_i + \varepsilon_{x_i}$ et $y_i = \mu + \varepsilon_i$ donc $x_i = \mu + \varepsilon_i{}'$ où $\varepsilon_i{}' \rightsquigarrow \mathcal{N}(0, \sqrt{\sigma^2 + {\sigma_{x_i}^2}^2})$ puisque ε_i et ε_{x_i} sont indépendantes et distribuées normalement de variance respective σ^2 et ${\sigma_{x_i}}^2$. La densité de probabilité de l'observation x_i est alors :

$$\begin{aligned} f(x_i, {\sigma_{x_i}}^2 ;\ \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2 + \sigma_{x_i}^2}} e^{\ -\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2+\sigma_{x_i}^2}} \\ &= \sqrt{\frac{p_i}{2\pi}} e^{\ -\frac{1}{2}p_i(x_i-\mu)^2} \quad \text{en notant } p_i = \tfrac{1}{\sigma^2+\sigma_{x_i}^2} \end{aligned}$$

La densité est donc indépendante des y_i. Compte-tenu de l'indépendance des observations, la fonction de vraisemblance de l'échantillon s'écrit :

$$\begin{aligned} \mathcal{L}(x_1,\ \ldots\ ,\ x_n,\ {\sigma_{x_1}}^2,\ \ldots\ ,\ {\sigma_{x_n}}^2 ;\ \mu,\ \sigma^2) &= \prod_{i=1}^{n} f(x_i,\ {\sigma_{x_i}}^2 ;\ \mu,\ \sigma^2) \\ &= (2\pi)^{-\frac{n}{2}} \sqrt{\prod_{i=1}^{n} p_i}\ e^{\ -\frac{1}{2}\sum_{i=1}^{n} p_i(x_i-\mu)^2} \end{aligned}$$

et son logarithme vaut $\ln\ \mathcal{L} = -\frac{n}{2} Ln\, 2\pi\ + \frac{1}{2}\sum_{i=1}^{n} Ln\, p_i\ - \frac{1}{2}\sum_{i=1}^{n} p_i (x_i - \mu)^2$

L'estimateur $\widehat{m}$ de la moyenne μ est la valeur qui maximise la vraisemblance, soit :

$$\left.\begin{array}{rcl} \frac{\partial}{\partial\mu} \ln\ \mathcal{L} & = & 0 \\ \frac{\partial^2}{\partial\mu^2} \ln\ \mathcal{L} & ¡ & 0 \end{array}\right\}$$

$\frac{\partial}{\partial\mu} \ln\ \mathcal{L}(x_1,\ \ldots\ ,\ x_n ;\ \widehat{m}, \sigma^2) = \sum_{i=1}^{n} p_i (x_i - \widehat{m}) = 0$

$\Rightarrow\ \sum_{i=1}^{n} p_i x_i = \sum_{i=1}^{n} p_i \widehat{m}\ \Rightarrow\ \widehat{m} = \frac{\sum_{i=1}^{n} p_i x_i}{\sum_{i=1}^{n} p_i}$

et l'on a bien $\frac{\partial^2 \ln\, \mathcal{L}}{\partial\mu^2} = -\sum_{i=1}^{n} p_i < 0$ car $p_i > 0$, $\forall i \in [1, n]$

– Montrons que cet estimateur est non biaisé, c'est-à-dire que son espérance est égale à la moyenne :

$$E[\widehat{m}] = \frac{\sum_{i=1}^{n} p_i E[x_i]}{\sum_{i=1}^{n} p_i}$$

et l'on a

$$E[x_i] = \mu + E[\varepsilon_i{}'] = \mu$$

ce qui entraine

$$E[\widehat{m}] = \mu \frac{\sum_{i=1}^{n} p_i}{\sum_{i=1}^{n} p_i} = \mu$$

– La variance de $\widehat{m}$ est

$$s^2_{\widehat{m}} = \frac{\sum_{i=1}^{n} p_i^2 \; Var(x_i)}{(\sum_{i=1}^{n} p_i)^2}$$

avec

$$Var(x_i) = Var(\varepsilon_i{}') = \sigma^2 + \sigma^2_{x_i} = \frac{1}{p_i}$$

donc

$$s^2_{\widehat{m}} = \frac{\sum_{i=1}^{n} p_i}{(\sum_{i=1}^{n} p_i)^2} = \frac{1}{\sum_{i=1}^{n} p_i}$$

– Montrons que $\widehat{m}$ est efficace. $I(\theta)$ étant l'information de Fisher, on a le résultat suivant [Dudewicz & Mishra, 1988]

$$I_n(\theta) = nI(\theta) = -E[\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}]$$

et si on calcule $Var(\widehat{m})$, on a alors

$$Var(\widehat{m}) = \frac{1}{-(-\sum_{i=1}^{n} p_i)} = \frac{1}{nI(\mu)}$$

donc $\widehat{m}$ est efficace ; tout autre estimateur non biaisé T de la moyenne aura une variance plus grande (ou égale) à celle de $\widehat{m}$ (inégalité de Rao-Cramer).

– Enfin, la convergence de cet estimateur $\widehat{m} = \widehat{m}_n$ – c'est-à-dire qu'il converge en probabilité vers la moyenne lorsque la taille de l'échantillon tend vers l'infini – peut se démontrer en utilisant l'inégalité de Bienaymé-Tchébychev :

$$P(|\widehat{m} - E[\widehat{m}]| \geq \epsilon) \leq \frac{Var(\widehat{m})}{\epsilon^2} \quad \forall \epsilon > 0$$

Ce qui implique, compte-tenu des résultats précédents :

$$P(|\widehat{m} - \mu| \geq \epsilon) \leq \frac{1}{\epsilon^2 \sum_{i=1}^{n} p_i} \quad \forall \epsilon > 0$$

or $p_i > 0$, donc $\lim_{n\to\infty} \sum_{i=1}^{n} p_i \to \infty$, ce qui implique

$$\lim_{n\to\infty} P(|\widehat{m} - E[\widehat{m}]| \geq \epsilon) = 0 \quad \forall \epsilon > 0$$

En ce qui concerne les propriétés concernant l'écart-type, les équations sont trouvées immédiatement en calculant les différentielles du logarithme de la fonction de vraisemblance par rapport à σ. En utilisant des propriétés générales du maximum de vraisemblance, on peut montrer les propriétés asymptotiques indiquées. On peut se rapporter pour cela par exemple à Pelat (1989), p. 56, ou Tassi (1989), p. 186.

Troisième partie

VALIDATION DU CATALOGUE D'ENTRÉE ET DES RÉSULTATS PRÉLIMINAIRES D'HIPPARCOS

Cette partie centrale traite des différentes comparaisons effectuées entre les données au sol et les résultats obtenus par le satellite Hipparcos, par l'intermédiaire des consortiums de réduction des données.

Nous avons séparé, quelque peu arbitrairement, en deux chapitres les comparaisons entre les données au sol et les données d'Hipparcos. Tout d'abord au chapitre 5, nous comparons les positions, magnitudes et couleurs des étoiles. Les données d'Hipparcos utilisées pour ces comparaisons sont celles qui ont été obtenues après six mois de mission avec les repéreurs d'étoiles du satellite.

Le chapitre 6, qui ne traite que des parallaxes préliminaires obtenues après un an de mission, est le plus long, mais le sujet le méritait : parce qu'elles ont une importance primordiale pour de nombreux problèmes astrophysiques, les parallaxes Hipparcos sont attendues par la communauté astronomique et il fallait être certain de leur qualité ; des comparaisons que nous avons effectuées, on déduira sans doute que l'on peut d'ores et déjà être rassuré.

Néanmoins, en aucun cas ces comparaisons ne doivent conduire à extrapoler les qualités ou les défauts des données d'Hipparcos utilisées ici, qui, rappelons-le, sont très préliminaires. Le chapitre 6, en particulier, a pour but de réfléchir à certains aspects de méthodologie pouvant éventuellement être appliqués plus tard aux données définitives, et les résultats obtenus sur les données préliminaires donnent essentiellement un ordre de grandeur.

Dans ce chapitre 6, nous commencerons par comparer les solutions obtenues par les deux consortiums de réduction des données, puis nous nous pencherons sur la comparaison entre les parallaxes préliminaires et les différentes estimations externes des parallaxes que l'on peut avoir depuis le sol.

Chapitre 5

Positions et magnitudes Hipparcos

Rarement un catalogue d'étoiles a eu à subir si vite l'épreuve des faits, comme cela a été le cas pour le Catalogue d'Entrée d'Hipparcos. Le satellite allait-il trouver les bonnes étoiles, à la bonne place, avec la bonne magnitude ? Les réponses à ces questions mesureraient l'adéquation au cahier des charges initial du consortium INCA, et, à travers lui, la qualité de l'ensemble des mesures effectuées au sol depuis des dizaines d'années. Autant dire que les résultats préliminaires de la mission Hipparcos étaient attendus avec impatience.

La précision des mesures au sol étaient importante à deux titres : d'abord pour optimiser les temps d'observation par le satellite de chaque étoile, ensuite pour faciliter la tâche des DRC, qui utilisent ces mesures comme première approximation. En ce qui concerne l'observabilité des étoiles du Catalogue d'Entrée, le bilan est largement positif. Sur les 118 000 étoiles du programme d'observation, moins de 0.1% n'ont pu être vues, essentiellement des étoiles de systèmes doubles ou des étoiles plus faibles que prévu, une grande partie de ces erreurs ayant d'ailleurs été corrigée par la suite [Crifo *et al.*, 1992].

D'autre part, avant de lancer la réduction «automatique» des données, il fallait vérifier que l'ensemble du processus de réduction donnait bien les résultats attendus.

L'article qui suit [Turon, Arenou, Evans, van Leeuwen, 1992] représente donc la première comparaison entre les données au sol et les positions et magnitudes obtenues à partir des repéreurs d'étoiles du satellite. Cette comparaison couvre à la fois l'astrométrie et la photométrie, globalement et suivant l'origine des données au sol.

Naturellement, il peut être utile d'insister sur le fait que les données du satellite utilisées ici proviennent des repéreurs d'étoile et non du champ principal, après seulement six mois de mesures, et que d'autre part, ils ont été obtenus avec les résultats d'un seul consortium. Par conséquent, il serait délicat d'extrapoler ce qui est mis en évidence.

Comme on pourra le constater, les seuls problèmes rencontrés concernent :

- des effets mis en évidence dans certains catalogues astrométriques du sud (voir tableau 1)
- une variation avec la position (fig. 6) de la différence entre les positions au sol et les positions Hipparcos. Cette variation n'a toujours pas reçu d'explication satisfaisante.
- une léger problème lors de la réduction photométrique (biais et grande dispersion visibles sur la figure 8), qui ont été interprétés ultérieurement [Evans, 1992] comme probablement dûs à une modélisation inadéquate du bruit de fond du ciel causé par la lumière zodiacale et le passage dans les ceintures de van Allen.

Mis à part ces problèmes mineurs, cette comparaison peut être considérée comme satisfaisante dans la mesure où elle prouve la qualité des résultats préliminaires du satellite d'une part, de la réduction des données d'autre part, et des données au sol enfin.

Astron. Astrophys. 258, 125–133 (1992)

ASTRONOMY AND ASTROPHYSICS

20. Comparison of the first results from the Hipparcos star mappers * with the Hipparcos Input Catalogue

C. Turon[1], F. Arenou[1], D.W. Evans[2], and F. van Leeuwen[2]

[1] URA 335 du CNRS et GDR Hipparcos, Observatoire de Meudon, DASGAL, F-92195 Meudon, France
[2] Royal Greenwich Observatory, Madingley Road, Cambridge CB3 0EZ, UK

Received July 18, accepted November 30, 1991

Abstract. Preliminary positions and magnitudes derived from the analysis of 12 weeks of observations from the Hipparcos star mappers are systematically compared with the various sources of ground-based data used in the Hipparcos Input Catalogue. These comparisons allow to cross-check the accuracies claimed by the various sources of ground-based data and by the analysis method of star mapper data. The parameters obtained for double stars, relative position and orientation, are also compared with ground-based data.

Key words: Hipparcos – catalogues – astrometry – reference frames – photometry – double stars

1. Introduction

The transits of Hipparcos programme stars through the satellite star mappers are recorded by the two photometers of the Tycho experiment (Høg et al. 1992). The knowledge of the attitude of the satellite at the epochs of the star transits, added to the determination of the transit times with respect to the star mapper grid, allows us to obtain corrections to the assumed positions of the observed stars on the sky. In addition, the photon counts, calibrated by the observation of photoelectric standard stars, allow the determination of B_T and V_T magnitudes. B_T and V_T stand for magnitudes in the Tycho bands (Grenon 1988). The profiles of these bands are close to the B and V of the Johnson system, with some discrepancies for the reddest stars. The process is limited to stars brighter than $V_T = 10$ approximately.

The analysis of these data have been performed for about 47 000 stars of the 118 000 of the Hipparcos Input Catalogue as part of the data reduction work performed at the Royal Greenwich Observatory (RGO) within the frame of the Hipparcos Northern Data Analysis Consortium (NDAC). Although still far from reaching the performances ultimately expected from a full analysis of the complete Hipparcos data set, these preliminary data already match the quality of most ground-based data. A total of 1.2 million transits were used. The positions and magnitudes obtained from these data (hereafter called the 'RGO catalogue') are compared with the data collected by the Hipparcos INCA Consortium which was responsible for the construction of the observing programme for Hipparcos (Turon et al. 1992). Extensive compilations and new observation programmes were undertaken by this Consortium to fulfil the ESA requirements about positions at epoch 1990 and magnitudes of programme stars (Jahreiß et al. 1992, Grenon et al. 1992). For each programme star, the best data for positions, proper motions, magnitudes and colours available within the 'INCA Database' were retained. The comparison of these data with the first results obtained from the Hipparcos star mappers allow a reciprocal check of both sets of data.

Send offprint requests to: C. Turon

* Based on observations made with the ESA Hipparcos satellite, and on work performed within the INCA and NDAC Consortia.

2. Data obtained from the Hipparcos star mappers

2.1. Positions

The star mapper data stream as received in RGO consists of stretches of 250 sampling periods around the predicted transit times of stars from the Input Catalogue. The star mapper photon count records are reduced to transit times and intensities, which in combination with the assumed positions of the stars involved provide information on the orientation of the satellite axes. Transits from the two fields of view and through the inclined and vertical slit groups describe in this way the evolution of the payload frame of reference, providing the reconstructed attitude. In NDAC the satellite attitude is determined relative to a dynamical model, strengthened by means of gyro readings (see van Leeuwen et al. 1992, Paper I). This allows the amount of information that has to be extracted from only the star mapper transits to be minimal, thus leaving information on the individual positions of the stars involved almost undisturbed in the form of transit time residuals.

The transit time residuals are collected as described in Paper I. In the reduction of the data from the satellite already distributed some 1.2 million transit time residuals from 12 weeks of data spread over 1.2 years were collected. 51 000 stars each had between 8 and 200 independent observations, which were used to improve the positions and magnitudes of these stars (47 000 stars from the Input Catalogue and 4 000 additional stars used for the 'Initial Star Pattern Recognition', i.e. for the initial attitude acquisition). The positional system defined by these updated positions is a combination of the original Input Catalogue and the smoothing effect of the attitude reconstruction process. In the attitude reconstruction two strips of sky with a length of 12–18° and separated by the basic angle of 58°, are used to determine the attitude of the satellite over one jet-firing interval. Transits through the vertical slits in both fields of view determine the 'spin-phase'. Transits through the inclined slits determine the spin-axis position. If systematic errors are present in either or both of these strips, then, in the case of the transits through the vertical slits,

the **differences** between these errors will enter the residual transit times, and will get removed from the catalogue. In the inclined slits the attitude will model the systematic errors, and only remove the individual errors.

About half the sky was covered by scans in different directions and thus the attitude reconstruction combined the data in these areas with various other areas on the sky. One third of the sky was covered by only one scan direction. The updating process was repeated several times over all 12 weeks of data, using the previous updates as starting points. The internal consistency figures clearly showed a system slowly converging. This way, some of the smaller scale systematic errors were automatically removed from the Input Catalogue. Larger scale systematic errors cannot be removed easily in this process, but were in general reduced (see also Lindegren et al. 1992).

2.2. *Double stars*

As was described in Paper I, double stars received a special treatment in the star mapper processing. The main reason to reduce the double star transits through the star mapper are to provide the processing of image dissector tube double star data with starting points on separation and orientation. In addition, it was necessary to provide better absolute positions for double stars than there were available from ground-based measurements. In the star mapper processing, the aim is to process double stars with separations above 1.5 arcsec in much the same way as single stars once they have had their positions updated. A properly resolved and recognized double star is unlikely to disturb the attitude reconstruction in the way an unresolved double can do it.

The accuracy of the relative position and orientation is, as always in differential measurements, higher than the absolute positions. The transit time differences are not affected by errors in the attitude reconstruction, and reflect directly the separation on the sky along the direction of the scan (and at an angle of 45°to the scan for transits through the inclined slits). The accuracy of the transit time differences is thus set by the accuracy of the transit time determinations. The rms accuracy for single transits under 'apogee-conditions' (low background signal), which is an indication of the best transit time accuracies available, ranges from 5 milli-arcsec at 5 mag to 40–60 milli-arcsec at 8-10 mag.

2.3. *Photometry*

The reduction of the star mapper data provides intensities in the B_T and V_T channels. These intensities have been calibrated to one system, removing effects of positional and colour dependence. They are collected (as described in Paper I) as intensities in the catalogue, with a simple relation to magnitudes. This avoids the creation of biases that would occur if magnitudes were collected in the catalogue. The calibration of the magnitudes was in an experimental phase during the processing of the provisional data, and it is therefore not surprising that some minor effects are still left in the data. The current comparison exercise is one of the tools helping us to recognize and remove these last discrepancies before the bulk processing of the data starts.

3. Data included in the Hipparcos Input Catalogue

Due to the detection system of the Hipparcos satellite and to its operational mode, the positions and magnitudes of the programme stars had to be known in advance with some accuracy. The specifications of ESA were ±1.5 arcsec on the 1990 positions and ±0.5 mag on the B or V magnitude for all programme stars, and a somewhat better accuracy on positions for a sub-set of stars used for real-time satellite attitude determination. As the stars were submitted for observation with Hipparcos on the grounds of scientific proposals, not taking into account the availability of accurate positions or magnitudes, extensive programmes of compilation and new observations or measurements were undertaken by the INCA Consortium (Turon et al. 1992).

3.1. *Astrometric data*

Astrometric data for 25 000 stars did not match the required accuracy (Jahreiß et al. 1992). New observations with Automatic Meridian Circles (10 000 stars observed at Bordeaux and La Palma) and plate measurements (100 000 stars measured on the ESO Sky Survey or CPC2 plates) were undertaken. For plate measurements, it was, indeed, decided to remeasure all candidate stars present on each plate. This yielded to a considerable overlap with earlier results, and allowed the detection of possible errors (mostly errors in star identification) not only in the plate measurements themselves but also in earlier measurements.

In parallel, the contents and precision of the available astrometric catalogues were investigated and a hierarchy established. Moreover, when possible, all available positions and proper motions were reduced to FK5. Finally, when all newly obtained data were available, the best positions and proper motions were selected to be retained in the final version of the Hipparcos Input Catalogue.

At the end of this extensive work, it was concluded that the final positional accuracy of the Hipparcos programme stars for epoch 1990 is better than 0.5 arcsec in the northern hemisphere, and better than 0.7 arcsec in the southern hemisphere, and that no systematic trend with respect to the FK5 system is present if the whole catalogue is considered. A complete description of the astrometric data included in the Hipparcos Input Catalogue can be found in Réquième (1989), Jahreiß (1989) and Jahreiß et al. (1992); references of all catalogues used can be found in Jahreiß (1989).

3.2. *Photometric data*

The specifications of ESA were only requiring 'one approximate magnitude, B or V, to within ±0.5 mag'. It rapidly appeared that, for reaching the accuracy expected on the astrometric parameters, an adequate observing time should be allocated to each programme star, as a function of its magnitude in the Hipparcos band (H_p). As a result, it was realized that the an accuracy of ±0.5 mag on the Hipparcos magnitude itself was desirable. This band has an effective wavelength close to that of the V band of the Johnson system, but much wider, and the differences $H_p - V$ are significant for very red or very blue stars (Grenon 1988). Thus, one magnitude (B or V) **and** a colour had to be obtained for all programme stars.

The photometric data available for the 214 000 proposed stars at the start of the Input Catalogue work, coming from the SIMBAD database or from the proposers, was very heterogeneous: accurate photoelectric photometry was available for about 26 000 stars, acceptable B and V magnitudes were obtained for about 145 000 stars from photographic photometry or estimates of blue and visual magnitudes, but about 17 000 stars had only incomplete or unreliable photometric information. Extensive observation programmes were performed in various photoelectric systems, and new observations were obtained for about 7 700 stars in 3 to 7 bands (Grenon 1992);

at the end of the Input Catalogue work, as a result of new observations and extensive compilations, B and V photoelectric photometry was available for about 46 000 stars and V photoelectric photometry coming from the Carlsberg Automatic Meridian Circle (CAMC) was available for about 13 000 stars. All these new data were used to obtain the data required for the mission: H_p, B_T and V_T.

In addition to this observational work, a new extinction model was derived to improve the determination of the reddened Johnson and Tycho colours obtained from the available MK or HD spectral types when only one magnitude was considered as reliable (Arenou et al. 1992). Colours were obtained in this way for about 60 000 single stars of the Input Catalogue.

3.3. Data on double and multiple stars

The situation was still worse for double and multiple stars, and a considerable effort was devoted first to make the available data easy to handle and avoid component mis-identification, and then to complement these data by new observations or measurements of positions or magnitudes where necessary (Dommanget 1989, Jahreiß et al. 1992). As for single stars, the knowledge of positions and magnitudes was required for each system, or for each observable component, but, in addition, the knowledge of the geometry of the systems and the relative magnitudes of the components was highly desirable to correct for the possible perturbing effect(s) caused by the presence of additional component(s) not taken into account for direct observation (Turon et al. 1989).

4. Comparison of the astrometric data

4.1. Global comparison

The differences between the data of the Hipparcos Input Catalogue and those obtained from star mappers are illustrated in Fig. 1, considering the 47 000 stars for which data are available from the analysis of the star mapper signals, i.e. for about 40 per cent of the complete observing programme. These two histograms show the differences in arcseconds between the $\alpha\cos\delta$ and δ from RGO and from INCA. The patterns are nearly symmetrical, with respective means of -0.01 and 0.05 arcsec and widths * of about 0.3 arcsec. This is in agreement with the values obtained in Paper I and Lindegren et al. (1992), and confortably within the initial specifications of ESA recalled in Sect. 3.

The variations of these differences with equatorial and ecliptic coordinates are shown in Fig. 2 and 3 respectively. Some features are striking:

- $(\Delta\alpha\cos\delta)_\alpha$ and $(\Delta\alpha\cos\delta)_\lambda$ stay close to zero with almost no significant deviation (one exception is a negative $\Delta\alpha\cos\delta$, about 0.060 arcsec, for α towards 3-4 hours, and about 0.080 arcsec for λ towards 320°).

- $(\Delta\delta)_\alpha$ and $(\Delta\delta)_\lambda$ are almost always positive, with little significant variations.

- $(\Delta\alpha\cos\delta)_\delta$ and $(\Delta\alpha\cos\delta)_\beta$ show significant negative deviations in the southern hemisphere (δ between $-40°$ and $-60°$ and between $-10°$ and $-20°$, β between $-20°$ and $-50°$ and south of $-60°$), a significant positive deviation in δ towards +20°, and a possible trend in β, increasing from $-90°$ to +90°.

- $(\Delta\delta)_\delta$ and $(\Delta\delta)_\beta$ show a significant positive deviation between $-70°$ and +30° in delta and between $-50°$ and +30° in ecliptic latitude.

The variations of $\Delta\alpha\cos\delta$ and $\Delta\delta$ with respect to δ described above are very similar to the curves obtained by Lindegren et al. (1992) (Figs 4 to 7) for the differences 'sphere minus Input Catalogue', but also, to a lesser extent, for the differences 'sphere minus RGO'. They are, in fact, the differences between these two figures.

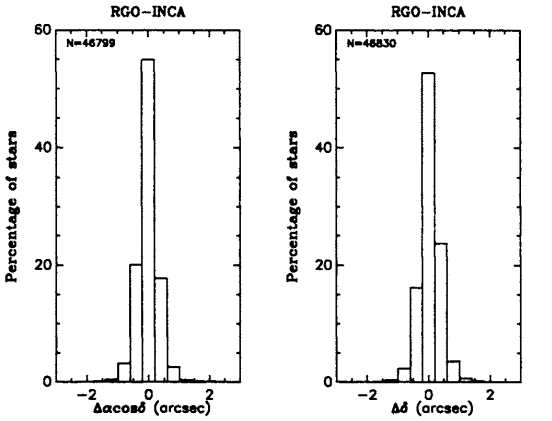


Fig. 1. Histograms of the differences between RGO and INCA in $\alpha\cos\delta$ and δ for the 47 000 considered stars

4.2. Comparison by source catalogue

The different source catalogues used in the Hipparcos Input Catalogue are considered here separately. The histograms of the differences between the RGO catalogue and each of these sources are presented in Fig. 4. The percentage of stars in each bin of $\Delta\alpha\cos\delta$ and $\Delta\delta$ with respect to the total number of stars in each source are given, in order to ease the comparison of the different figures. It shows clearly that SSSC catalogue is not centred.

Due to large scale systematic errors in the Input Catalogue, which could not be removed in the RGO Catalogue, the dispersion of the positions in the RGO Catalogue is about 0.09 arcsec as given by Lindegren et al. (1992). This prevents any direct comparison with FK5 since the order of precision of the positions given in this catalogue is about 0.04 arcsec. For the other catalogues, the comparison with the positional errors quoted in the Input Catalogue (Jahreiß et al. 1992) shows a close agreement (Table 1) and also gives an upper limit of 0.21 arcsec for the positional error of SRS catalogue at epoch 1990.

* In order to characterize the scatter of these differences, a width based on distribution percentiles is used as dispersion estimate instead of a rms scatter, which is too sensitive to heavy tail distributions and outliers. This estimate is used even if the distribution is intrinsically non-gaussian but the result of the mixing of differences of positions with accuracies ranging from 0.03 (FK5) to 3 arcsec.

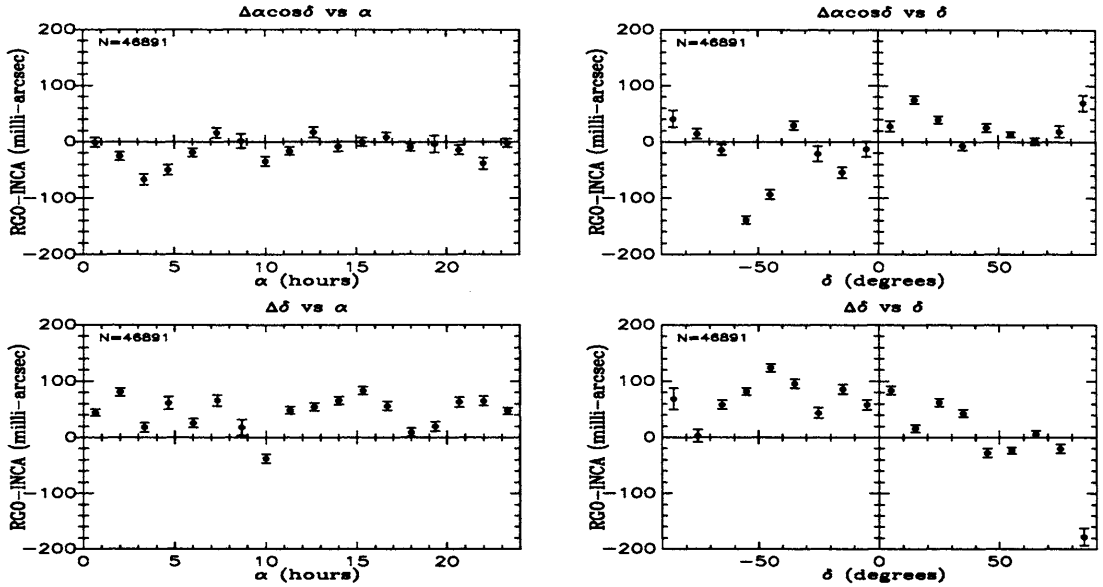


Fig. 2. Differences between RGO and INCA in $\alpha \cos\delta$ and δ for the 47 000 considered stars, as a function of equatorial coordinates; bins of 80 minutes in α, 10°in δ; the error bars are standard errors on the averages estimated from the dispersion in each bin

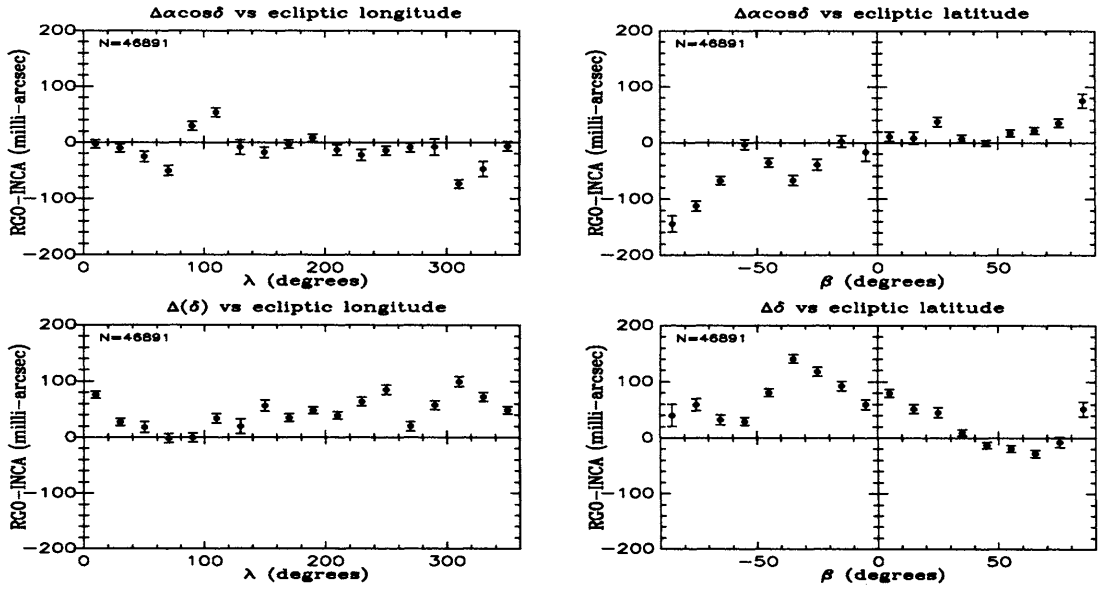


Fig. 3. Differences between RGO and INCA in $\alpha \cos\delta$ and δ as a function of ecliptic coordinates; bins of 20°in longitude, 10°in latitude

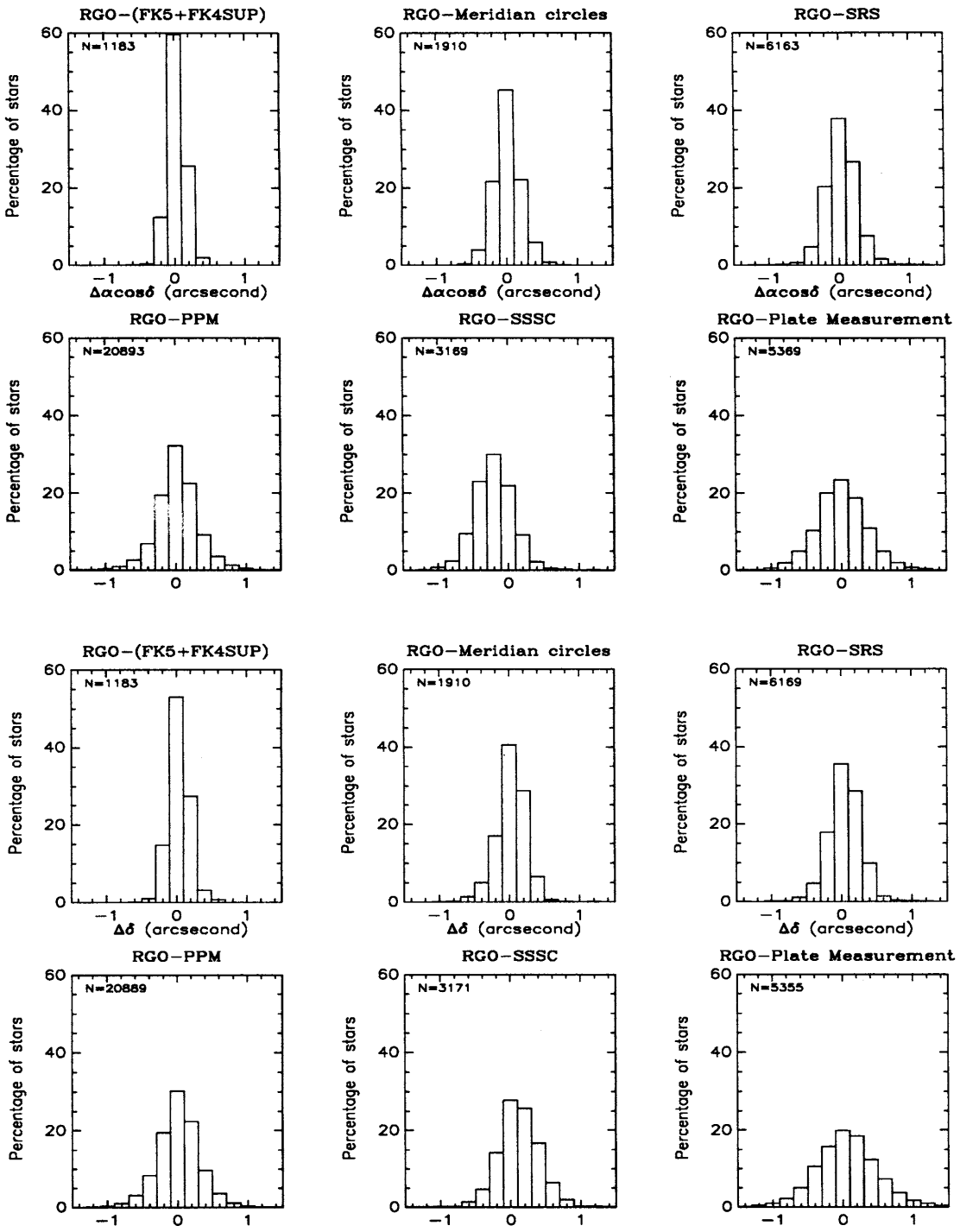


Fig. 4. Differences between RGO and INCA in $\alpha \cos \delta$ (upper 6 histograms) and δ, for each major catalogue source of astrometric data in the Input Catalogue

Table 1. Median and width (arcsec) of distribution of differences between RGO and INCA in $\alpha \cos\delta$ and δ for each catalogue

	$\alpha\cos\delta$		δ	
	median	width	median	width
Meridian circles	0.01	0.18	0.03	0.19
SRS	0.03	0.20	0.05	0.21
PPM	0.02	0.26	0.01	0.27
SSSC	-0.21	0.26	0.10	0.27
Provisional CPC2	-0.16	0.33	0.13	0.29
Plate measurements	0.00	0.35	0.05	0.42

In order to understand the variations in equatorial and ecliptic coordinates, the possible effects of some specific catalogues was investigated. For example, the stars whose position sources were the SSSC (Sydney Southern Star Catalogue, King & Lomb 1983) and the provisional CPC2 ** (Nicholson et al. 1984, 1985), were eliminated from the considered sample. The resulting variations with respect to equatorial coordinates are shown in Fig. 5.

The most striking effect, when compared with Fig. 2, is to suppress completely the two dips in $\Delta\alpha\cos\delta$ versus δ (for $\delta = -50°$ and $-15°$). As a result, there is now a positive excess in $(\Delta\alpha\cos\delta)_\delta$ in the southern as well as in the northern hemisphere, which is reflected at all right ascensions (the differences stay negative only for α between 1 and 5 hours). This can probably be explained by the fact that the RGO catalogue is 'linked' to the Input Catalogue as a whole. If catalogues for which the mean deviation in $\alpha\cos\delta$ lies between -0.15 and -0.20 arcsec are not considered, the whole solution for the remaining stars is pushed towards positive values of $\Delta\alpha\cos\delta$.

This bias towards positive $\Delta\alpha\cos\delta$ (RGO-INCA) is also clearly visible in Fig. 4 for all source catalogues other than SSSC. As the central epoch of SSSC and provisional CPC2 is about 1960, the effect of 30 years of proper motion was investigated in order to explain this bias; it appeared that the bias remains present whatever the source of proper motions is (CPC, CPC2, SAO, SSSC, ...), with only slight variations. Therefore a possible explanation may be that some southern catalogues could be poorly linked to the FK5 system (since the FK5 catalogue does not show this bias). However, it should be kept in mind that this analysis is only tentative, being based on very preliminary results from the Hipparcos mission, and on only 12 weeks of observations (only about 300 stars from the FK5 are included in this comparison).

The suppression of the stars from the SSSC and provisional CPC2 also show up very clearly on the variations of $\Delta\alpha\cos\delta$ and $\Delta\delta$ versus β. These are shown in Fig. 6. A sinusoidal trend may be seen on both graphs, more marked on the differences in δ. Such an effect may come from the uneven coverage of the sky, or/and from the uneven range of orientations of the scanned great circles. This is still under investigation.

** Final CPC2 is presented in Zacharias et al. (1992)

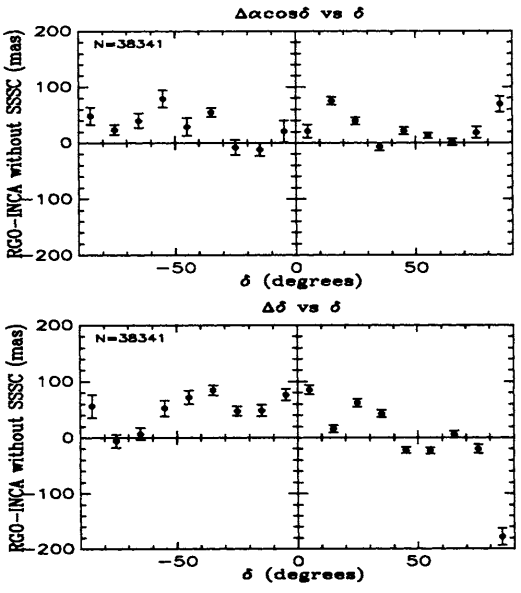


Fig. 5. Differences between RGO and INCA in $\alpha\cos\delta$ and δ for INCA stars, not considering stars whose source of position is SSSC or a provisional version of CPC2, as a function of δ

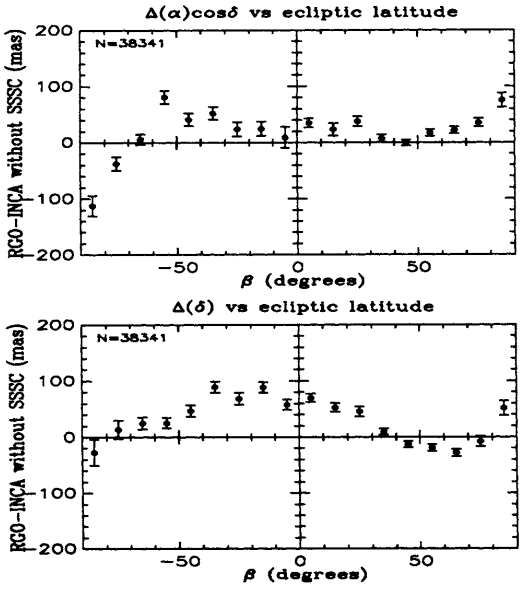


Fig. 6. Differences between RGO and INCA in $\alpha\cos\delta$ and δ for INCA stars, not considering stars whose source of position is SSSC or a provisional version of CPC2, as a function of ecliptic latitude

4.3. Single stars and double stars

The positions of double and multiple stars in the Input Catalogue are known to be less accurate than the positions of single stars. This is verified in the comparison with the RGO catalogue. Histograms of the differences RGO-INCA for double and single stars are given

separately in Fig. 7: the widths are 0.27 arcsec for single stars or stars considered as single for Hipparcos observation (perturbation due to the secondary component(s) considered negligible), and 0.53 for double stars (two entries in the Input Catalogue, or one entry which is the photo-centre or the geometric centre of the system).

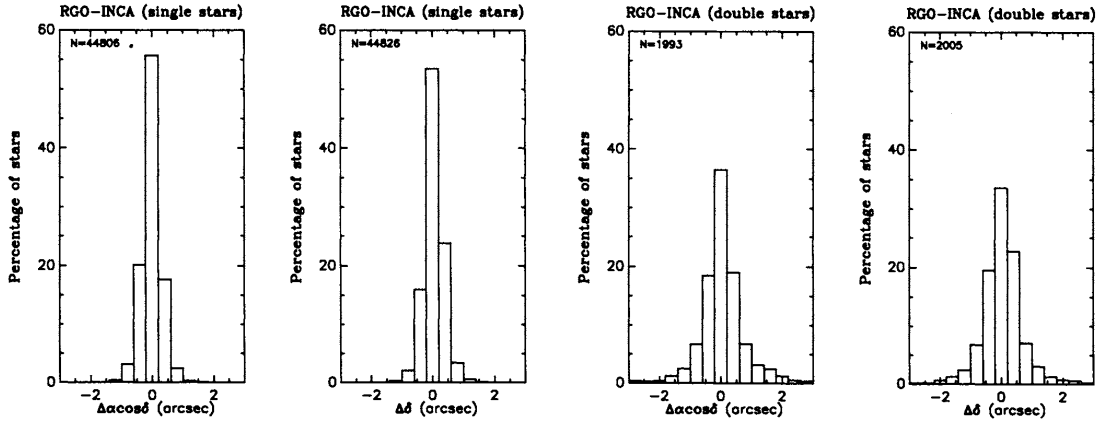


Fig. 7. Histograms of the differences between RGO and INCA in $\alpha \cos \delta$ and δ: single stars (left), double stars (right)

5. Comparison of the photometric data

A comparison has been made between the B_T and V_T as calibrated by the RGO team and as given by the INCA consortium. The three main sources of photometry in the Input Catalogue are respectively:

1) photoelectric photometry,

2) photoelectric V coming from the CAMC, and $B-V$ derived from spectral type and an extinction model,

3) V coming from very heterogeneous sources, mainly from visual observations, and $B-V$ derived from spectral type and an extinction model.

1) In this preliminary version of RGO updated Catalogue, there is a small bias in B_T and V_T magnitudes, as can be seen in Fig. 8. This bias will soon be corrected; no special trend of the differences RGO-INCA with position (e.g. ecliptic coordinates) may be noticed.

2) For stars from the CAMC, the colour was derived from spectral type and Fig. 9 shows the differences ΔB_T vs B_T and ΔV_T vs V_T. The method used to obtain colours may be tested on this sample

The difference between RGO and INCA photometry as a function of B_T and V_T magnitudes for these three main sources are presented in Fig. 8, 9, 11; medians and widths of these differences are indicated in Table 2.

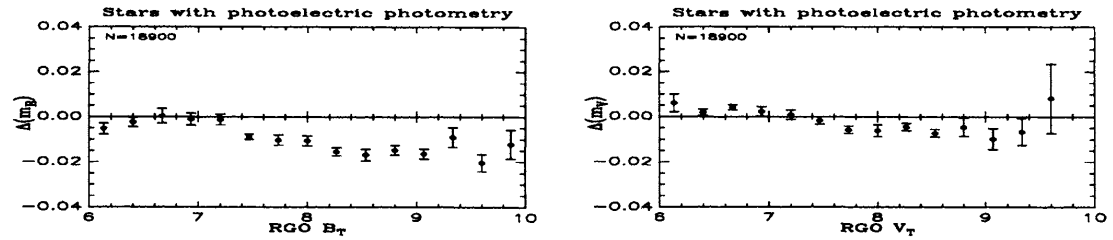


Fig. 8. Differences between RGO and INCA in B_T as a function of B_T and differences in V_T as a function of V_T for stars with photoelectric photometry and $V_T < 9.5$

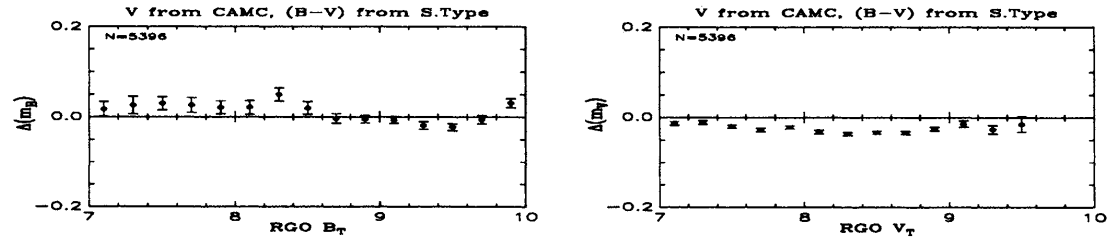


Fig. 9. Differences between RGO and INCA in B_T as a function of B_T and differences in V_T as a function of V_T for stars with photoelectric V and $B-V$ derived from spectral type

Table 2. Median and width (magnitudes) of the distribution of the differences between RGO and INCA in B_T and V_T for the three major sources of photometric data in the Input Catalogue

	B_T		V_T	
	median	width	median	width
Photoelectric B & V	-0.010	0.055	0.000	0.040
Photoelectric V	-0.010	0.195	-0.020	0.075
Heterogeneous V	0.020	0.235	0.020	0.200

6. Comparison of data on double stars

Double or multiple systems were given by the INCA Consortium as a single entry when the separation between components was below 10 arcsec. For systems with separation between 1.5 and 10 arcsec, the star mapper reduction is able to separate the components (Paper I). The comparison between RGO measurements and CCDM ground-based measurements is given in Fig. 12, both in separation and in position angle between components. Fig. 12 shows two perfect correlations, with median values/widths of 0.01/0.16 arcsec for differences in separation, and $-0.04/2.5°$for differences in position angle.

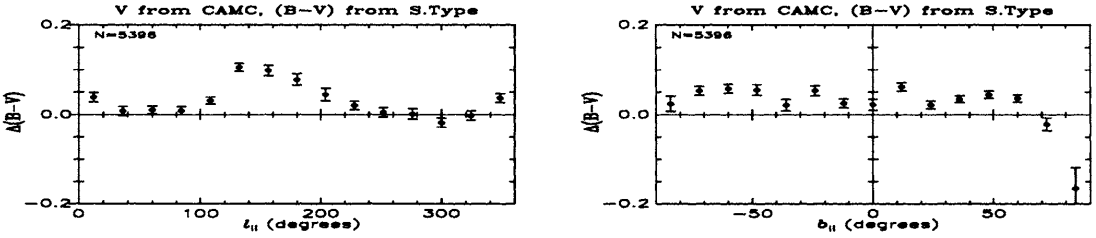


Fig. 10. Differences between RGO and INCA in $(B_T - V_T)$ as a function of galactic longitude and latitude for stars with photoelectric V and $B-V$ derived from spectral type

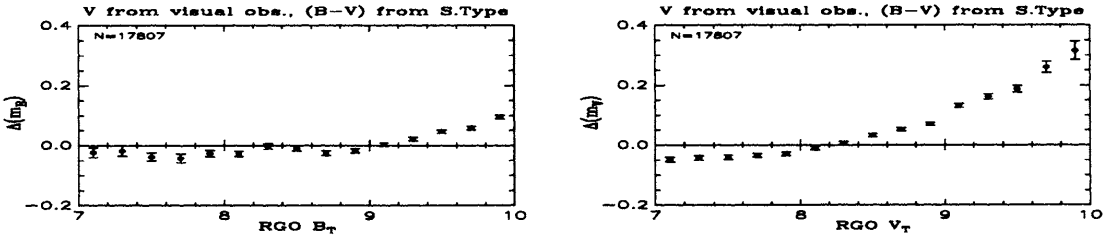


Fig. 11. Differences between RGO and INCA in B_T as a function of B_T and differences in V_T as a function of V_T for stars with heterogeneous sources in V and $B-V$ derived from spectral type

as V_T is precise and does not introduce supplementary scatters in the estimation of the colours. The overall accuracy of colours obtained by this method is of about 0.18 (Tycho) magnitudes; however small systematic effects may be noticed when plotting the differences RGO-INCA as a function of galactic coordinates (Fig. 10): at north galactic pole, the negative differences is explained by a small number of stars with bad HD spectral classification, and wrongly considered as giants. Apart from this region, the differences are slightly positive due to distant stars (the model is less accurate for distances larger than 1 kpc) – this is especially visible between 140°-180°of galactic longitude; however there is also a contribution of erroneous spectral classifications.

3) Finally, stars which had photoelectric photometry neither in B nor in V are presented in Fig. 11. On the right side, it appears clearly how heterogeneous sources of photometry – mainly visual observations – systematically underestimate the magnitude. Without deriving the colour of these stars from their spectral type, the difference RGO-INCA in B_T (left side) would have had the same systematic trend (or even worse) as in ΔV_T.

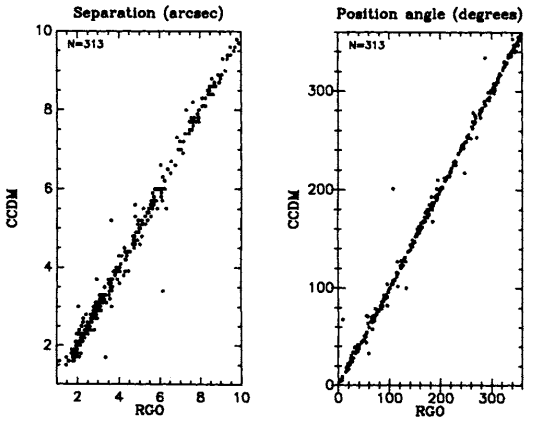


Fig. 12. Comparison between RGO and ground-based measurements of separation between components (left) and position angle (right) for double or multiple systems with $1.5 \leq \rho \leq 10$ arcsec

7. Conclusion

Although preliminary, this work shows how fruitful is the collaboration between Hipparcos Consortia. The INCA Consortium, as supplier of the input data, improves its knowledge of the astronomical content of its data and, in return, the Data Reduction Consortia will probably find in these results some answers to questions appearing during the reduction process.

Apart from the minor effects described above, the preliminary data obtained from the Hipparcos star mapper are clearly consistent with most ground-based data; part of the updated positions are already used for the real time attitude determination of the satellite. Of course, they are still far from reaching the ultimately expected Hipparcos performances, and other questions (real positional accuracy of ground-based catalogues, systematic errors) will receive a definitive answer as soon as the comparison between input data and the sphere solution is done.

Acknowledgements. We would like to thank M. Crézé, L.V. Morrison and Y. Réquième for very useful discussions about the interpretation of these results.

References

Arenou F., Grenon M., Gómez A., 1992, A&A, this issue (paper 16)
Carlsberg Meridian Catalogue, 1, 1985, Copenhagen Univ. Obs., Royal Greenwich Obs. and Instituto y Obs. de Marina.
Carlsberg Meridian Catalogue, 2, 1986, id
Carlsberg Meridian Catalogue, 3, 1987, id
Dommanget, J., 1989, ESA–SP 1111, Vol. II, 149
Fricke, W. et al., 1989, Fifth Fundamental Catalogue (FK5), Part I, The Basic Fundamental Stars, Veroff. Astron. Rechen-Instituts, Heidelberg, 32; Verlag G.Braun, Karlsruhe
Grenon, M., 1988, in 'Scientific Aspects of the Input Catalogue Preparation', Torra, J. & Turon, C. (eds), 21
Grenon M., Mermilliod M., Mermilliod J.C., 1992, A&A, this issue (paper 13)
Høg E., Bastian U., Egret D., Grewing M., Halbwachs J.L., Wiecenec A., Bässgen G., Bernacca P.L., Donati F., Kovalevsky J., van Leeuwen F., Lindegren L., Pedersen H., Perryman M.A.C., Petersen C., Scales D.R., Snijders M.A.J., Wesselius P.R., 1992, A&A, this issue (paper 27)
Hughes J.A., 1978, Southern Reference System (SRS), in IAU Coll. 48, Prochazka F.V., Tucker R.H. (eds), 497
Jahreiß H., 1989, ESA–SP 1111, Vol. II, 115
Jahreiß H., Réquième Y., Argue A.N., Dommanget J., Rousseau M., Lederle T., Le Poole R.S., Mazurier J.M., Morrison L.V., Nys O., Penston M.J., Périé J.P., Prévot L., Tucholke H.J., de Vegt C., 1992, A&A, this issue (paper 12)
King D.S., Lomb N.R., 1983, Sydney Southern Star Catalogue, Sydney Obs. Papers No 96
van Leeuwen F., Penston M.J., Perryman M.A.C., Evans D.W., Ramamani N., 1992, A&A, this issue (paper 8)
Lindegren L., van Leeuwen F., Petersen C., Perryman M.A.C., Söderhjelm S., 1992, A&A, this issue (paper 21)
Nicholson W., Penston M.J., Murray C.A., de Vegt C., 1984, MNRAS, **208**, 911
Nicholson W., 1985, private communication
Polozhentsev D.D., 1978, Southern Reference System (SRS), in IAU Coll. 48, Prochazka F.V., Tucker R.H. (eds), 489
Réquième Y., 1989, ESA–SP 1111, Vol. II, 107
Röser S., Bastian U., 1989, Positions and Proper Motions (PPM) of 181731 stars north of -2.5°declination, Astron. Rechen-Inst. Heidelberg, (F.R.G.)
Turon C., Kovalevsky J., Lindegren L.,1989, ESA–SP 1111, Vol. 2, 65
Turon C., Gómez A., Crifo F., Crézé M., Perryman M.A.C., Morin D., Arenou F., Nicolet B., Chareton M., Egret D., 1992, A&A, this issue (paper 11)
Zacharias N., de Vegt Chr., Nicholson W., Penston M.J., 1992, A&A, in press

This article was processed by the author using Springer-Verlag TEX A&A macro package 1992.

Chapitre 6

Étude des parallaxes préliminaires Hipparcos

6.1 Introduction

Les parallaxes absolues obtenues avec Hipparcos doivent avoir une précision moyenne d'environ deux millièmes de seconde d'arc (mas) et les erreurs systématiques doivent être nettement inférieures à un mas [Lindegren, 1989, page 323]. Pourquoi des erreurs systématiques ? Tout simplement parce qu'il est possible que certains effets ne soient pas modélisables ; par exemple, il peut y avoir un décalage systématique du point-zéro global des parallaxes de l'ordre de quelques dizièmes de mas, dû à des variations périodiques de l'angle de base du miroir du satellite, avec une période telle que l'effet soit impossible à décorréler de la parallaxe [Lindegren, 1989, 1992b].

Sans doute cet effet est très faible ; il n'empêche : s'il existe un moyen de le mettre en évidence, il faudra soustraire ce décalage des parallaxes Hipparcos définitives, lorsqu'elles seront obtenues. Le but des paragraphes qui suivent est donc de réfléchir aux méthodes qui pourraient permettre de le déterminer, de montrer qu'il provient bien d'un problème instrumental (plus exactement qu'il est indépendant des caractéristiques des étoiles observées) et de vérifier la qualité de ces parallaxes, par la détermination de leur erreur externe. Nous disposons pour cela de deux échantillons de parallaxes préliminaires, provenant l'un de la réduction après un an de données faite par le consortium NDAC (solution 5 paramètres), l'autre du consortium FAST (solution 3 paramètres).

Dans les pages qui suivent, un certain nombre de travaux ont été faits avec ces parallaxes préliminaires. On peut remarquer que nous n'étudierons que la qualité statistique des parallaxes ; jamais des comparaisons individuelles ou de groupes homogènes d'étoiles ne seront abordées. C'est volontaire. Les données préliminaires d'Hipparcos, outre leur aspect préliminaire (c'est-à-dire éventuellement imparfait...) appartiennent à l'Agence Spatiale Européenne et ne seront publiées que quand celle-ci le jugera approprié. D'autre part, les scientifiques qui ont fait les propositions d'observation d'étoiles par Hipparcos ainsi que les membres des consortiums ont des droits sur ces données : travailler avant eux sur les données préliminaires relèverait de ce qu'en d'autres domaines on appelle le délit d'initié.

Une remarque liminaire concernant l'estimation du point-zéro z peut être faite : l'effet à mettre en évidence doit être indépendant de la position, de la magnitude et de la

couleur des étoiles, etc. À condition que l'erreur systématique sur la parallaxe remplisse ces critères, il devrait donc suffire de regarder la moyenne de la distribution des parallaxes des étoiles les plus lointaines.

En admettant que l'on ait un échantillon de n étoiles *très* lointaines, c'est-à-dire avec une (vraie) parallaxe négligeable, et si l'erreur sur la parallaxe est en moyenne 2 mas, l'erreur standard sur la moyenne des parallaxes Hipparcos de cet échantillon serait $\sigma_z \approx \frac{2}{\sqrt{n}}$mas.

Si l'on voulait avoir une précision de l'ordre de 0.05 mas sur z, il faudrait 1 600 étoiles ; pour $\sigma_z = 0.01$ mas, il faudrait 40 000 étoiles. Comme il n'y a pas dans le Catalogue d'Entrée ce nombre d'étoiles très lointaines à notre disposition (le tableau 6.1 en donne l'ordre de grandeur), et encore moins dans les échantillons à notre disposition, cela a deux implications.

TAB. 6.1: *Nombre d'étoiles lointaines.*

Ordre de grandeur du nombre d'étoiles dans le Catalogue d'Entrée ayant une parallaxe spectroscopique inférieure à la parallaxe $\pi_{\max}$.

$\pi_{\max}$ (mas)	2.0	1.5	1.0	< 0.5
Nombre d'étoiles	10335	5562	2760	1025

D'abord qu'il ne faut pas s'attendre à une grande précision sur z, ensuite qu'il faudra utiliser des étoiles avec une parallaxe (spectroscopique ou photométrique) non négligeable, de l'ordre de 2 mas et moins. Cette dernière remarque signifie qu'il faudra trouver l'estimateur adéquat des parallaxes spectroscopiques et prendre en compte les erreurs sur ces parallaxes. Plus exactement, si l'on note $\pi_{\rm H}$ la parallaxe Hipparcos et $\pi_{\rm S}$ la parallaxe spectroscopique, en calculant z par

$$z = \frac{1}{n}\sum_{i=1}^{n}(\pi_{\rm H} - \pi_{\rm S}) = \frac{1}{n}\sum_{i=1}^{n}\pi_{\rm H} - \frac{1}{n}\sum_{i=1}^{n}\pi_{\rm S}$$

c'est l'estimateur de la moyenne empirique de parallaxes spectroscopiques qu'il faut déterminer avec soin.

On pourrait penser que si les étoiles sont lointaines, l'erreur absolue sur la parallaxe spectroscopique moyenne est négligeable ; c'est exact, mais seulement en ce qui concerne l'erreur aléatoire. S'il existe une erreur systématique sur l'estimation de la parallaxe spectroscopique – disons de 20% pour des étoiles de 1 mas –, alors l'effet sur le calcul de z sera de quelques dizièmes de mas, c'est-à-dire de l'ordre de grandeur éventuellement de z lui-même...

Nous montrerons que ces effets systématiques existent et c'est pourquoi nous nous attarderons longuement sur des problèmes de nature statistique : choix d'estimateurs, biais d'échantillonnage, etc. Mais tout d'abord, tentons d'avoir un aperçu de la qualité des parallaxes préliminaires d'Hipparcos sans faire intervenir des comparaisons avec des données externes.

6.2 Les parallaxes FAST et NDAC - Comparaisons internes

6.2.1 Aperçu des parallaxes préliminaires

Comme nous l'avons indiqué en introduction, page 21, nous disposons des parallaxes préliminaires obtenues après un an de mission, et réduites par chacun des consortiums, NDAC (consortium nord) et FAST (consortium sud). Comme le but fixé n'est pas de comparer les qualités ou défauts respectifs des réductions de chaque DRC, mais plutôt de commencer à prévoir comment la parallaxe Hipparcos définitive pourra être validée, nous nous intéresserons prioritairement à l'aspect méthodologique de cette validation.

Rappelons que chaque consortium a calculé une solution 5 paramètres (position (α, δ), mouvement propre ($\mu_\alpha \cos\delta$, μ_δ) et parallaxe π) et une solution 3 paramètres (les mouvements propres utilisés étant ceux du Catalogue d'Entrée). Nous avons choisi d'utiliser ici principalement la solution 3 paramètres (3P) du consortium FAST et la solution 5 paramètres (5P) du consortium NDAC. La place manquerait en effet pour traiter toutes les solutions acquises (quand on compte également celles qui avaient été obtenues avec six mois de données).

Les distributions de ces parallaxes sont indiquées sur les figures 6.1 et 6.2 pour les solutions FAST-3P et NDAC-5P. L'écart-type de ces distributions autour de leur moyenne est respectivement de 10.511 mas et 9.808 mas ; le mode des distributions est voisin de 3.3 mas. Signalons pour qui s'en étonnerait qu'une partie des parallaxes obtenues sont négatives ; il peut paraître de prime abord étonnant d'obtenir une distance (inverse de la parallaxe) négative. Mais l'estimation de la parallaxe qui est obtenue est clairement le résultat d'une mesure affectée d'une erreur, et pour une étoile très lointaine ($\pi \approx 0$ mas), l'estimation (réalisation d'une variable aléatoire) varie autour de 0, et peut donc naturellement être négative. C'est ainsi que nous avons 5 632 étoiles avec une parallaxe négative pour FAST-3P et 3 982 étoiles pour NDAC-5P.

Signalons au passage que la tentation d'éliminer les étoiles avec une parallaxe négative serait une erreur (si l'on ose dire). Cette troncature biaiserait gravement la distribution des parallaxes des étoiles lointaines, en ne gardant que celles qui ont une erreur positive. Certes les parallaxes négatives peuvent sembler inutilisables. En réalité, elles fournissent tout de même l'indication que l'étoile en question est lointaine, renseignement qui peut avoir son utilité, et elles indiquent également l'ordre de grandeur des erreurs sur les parallaxes.

Pour chaque étoile, l'erreur sur la parallaxe Hipparcos peut être considérée comme dépendant de la magnitude d'une part (tableau 6.2), et de la latitude écliptique d'autre part (tableau 6.3), si bien que la précision nominale de 2 millièmes de secondes d'arc qui est souvent annoncée pour Hipparcos peut varier selon ce modèle entre 0.6 et 5.6 mas. La magnitude et la latitude écliptique permettent ainsi aux consortiums de calculer une erreur formelle[1] s_H sur la parallaxe. Pour chaque étoile, on dispose donc, comme données, de la parallaxe et de son erreur formelle associée.

1. estimation, à l'aide d'un modèle, de l'écart-type de l'erreur aléatoire et que l'on appellera également par la suite «erreur interne». Nous utiliserons l'indice F pour FAST, N pour NDAC, H pour l'une des deux parallaxes d'Hipparcos ; P pour photométrique, S pour spectroscopique ; s_H désignera l'erreur interne et σ_{π_H} l'erreur externe.

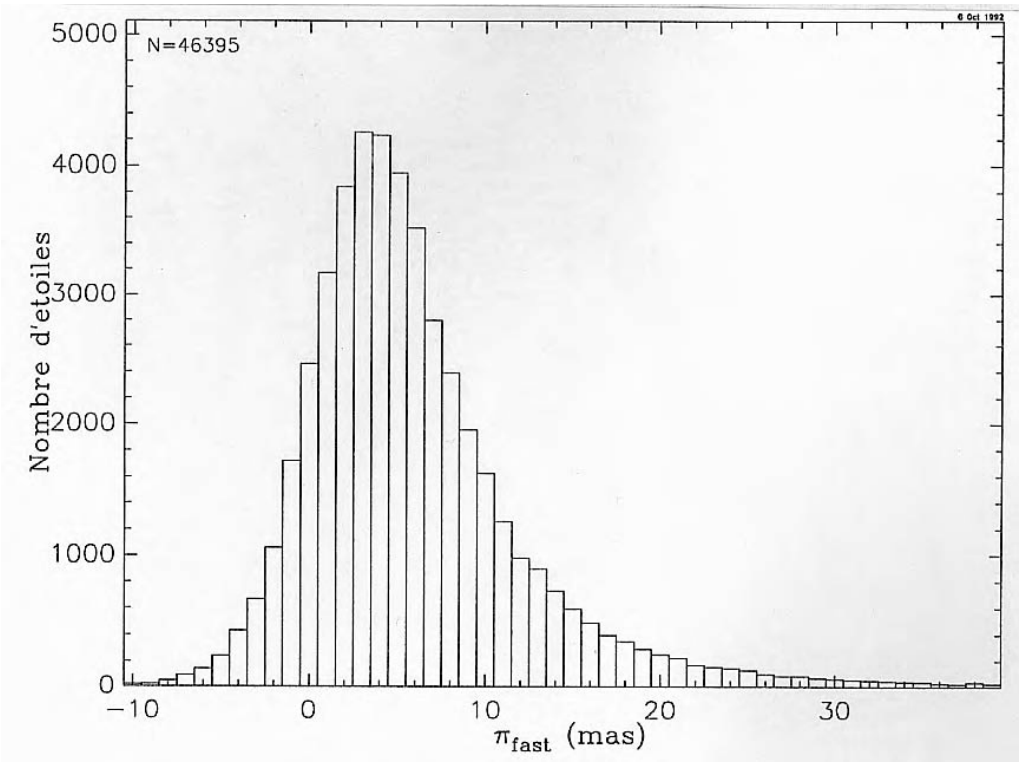


FIG. 6.1: *Distribution des parallaxes préliminaires 3 paramètres obtenues après un an de données par le consortium FAST (mas)*

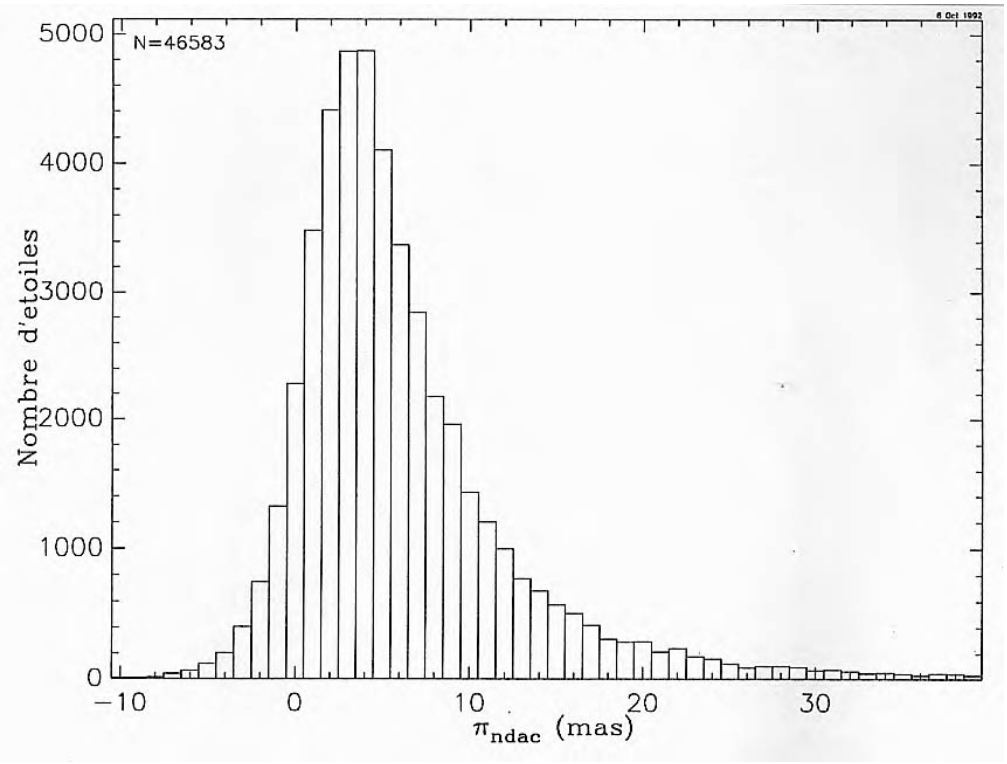


FIG. 6.2: *Distribution des parallaxes préliminaires 5 paramètres obtenues après un an de données par le consortium NDAC (mas)*

TAB. 6.2: *Erreur formelle en fonction de la magnitude.*

Variation de l'erreur formelle sur la parallaxe Hipparcos (pour une mission de 4 ans) en fonction de la magnitude H_p.

H_p	≤ 6.5	6.5-7.5	7.5-8.5	8.5-9.5	9.5-10.5	10.5-11.5	11.5-12.5	> 12.5
σ_π	0.9	1.0	1.3	1.6	2.1	2.9	3.5	4.8

(d'après Lindegren & Kovalevsky, 1992)

TAB. 6.3: *Erreur formelle en fonction de la latitude.*

Variation de l'erreur formelle sur la parallaxe Hipparcos en fonction de la valeur absolue de la latitude écliptique β. Il s'agit d'un facteur multiplicatif à apporter au tableau 6.2.

$\lvert\beta\rvert$	$0° - 24°$	$24° - 37°$	$37° - 44°$	$44° - 53°$	$53° - 64°$	$64° - 90°$
$\frac{\sigma_\pi}{\sigma_\pi(moyen)}$	1.2	1.1	1.0	0.9	0.8	0.7

(d'après Lindegren, 1989, p 321)

La variance des erreurs changeant suivant les étoiles, on se doute qu'en conséquence, lorsqu'on verra la distribution des parallaxes Hipparcos pour des étoiles très lointaines, cette distribution n'aura aucune raison d'être gaussienne (puisque provenant de variables aléatoires de variances différentes), sauf si les étoiles sont de même magnitude, à la même latitude écliptique, et ont le même nombre de mesures. Par contre, on espère que la distribution normalisée (parallaxe divisée par l'erreur interne) soit gaussienne, et, idéalement, qu'elle suive la loi normale réduite $\mathcal{N}(0, 1^2)$. Cela signifierait qu'il n'y a pas d'erreur systématique, et que, tous les effets ayant été pris en compte, l'erreur interne rejoint l'erreur externe.

Les distributions des erreurs formelles s_F et s_N calculées par les consortiums sont indiquées fig. 6.3 et 6.4 pour les solutions FAST-3P et NDAC-5P respectivement. Les parallaxes dont nous disposons sont celles qui ont une erreur interne inférieure à 4 mas. L'erreur interne est en moyenne $\langle s_\mathrm{F} \rangle = 2.217$ mas pour FAST-3P et $\langle s_\mathrm{N} \rangle = 2.158$ mas pour NDAC-5P. Les deux distributions de parallaxes en présence sont donc comparables.

Ces erreurs internes sont importantes : calculées avec soin, elles apportent une information complémentaire à la parallaxe observée. En effet, si l'erreur aléatoire est gaussienne d'espérance nulle autour de la vraie parallaxe, il suffit d'avoir l'erreur interne pour connaître la distribution complète de l'erreur (les deux premiers moments suffisant à déterminer une loi normale). Par conséquent, nous étudierons dans ce qui suit à la fois les parallaxes et leur erreur interne, et nous nous servirons des deux pour obtenir le meilleur estimateur de la parallaxe Hipparcos.

Pour l'instant, on peut déjà comparer les parallaxes FAST-3P et NDAC-5P, pour les étoiles communes aux deux solutions (22 820 étoiles), grâce à l'histogramme des différences de parallaxes, fig. 6.5, et avec l'histogramme des différences normalisées, fig. 6.6. Pour ce dernier, les différences sont divisées par l'erreur formelle sur la différence, que l'on choi-

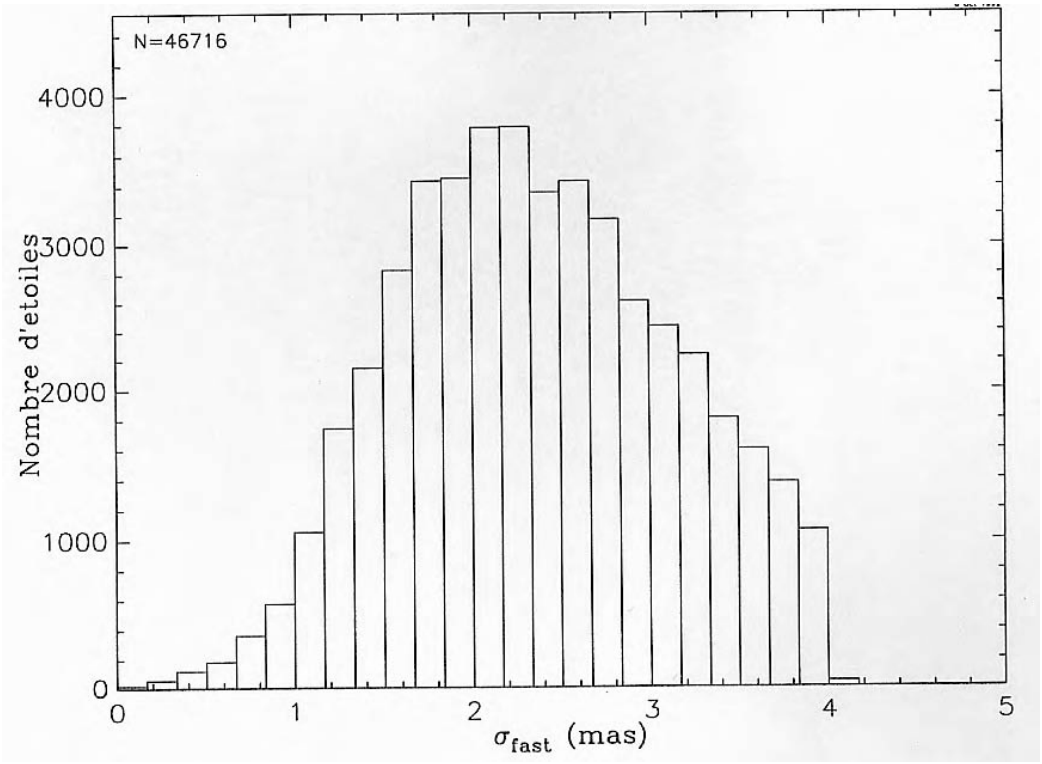


FIG. 6.3: *Distribution des erreurs internes (s_F) sur les parallaxes préliminaires obtenues avec un an de données par le consortium FAST (mas)*

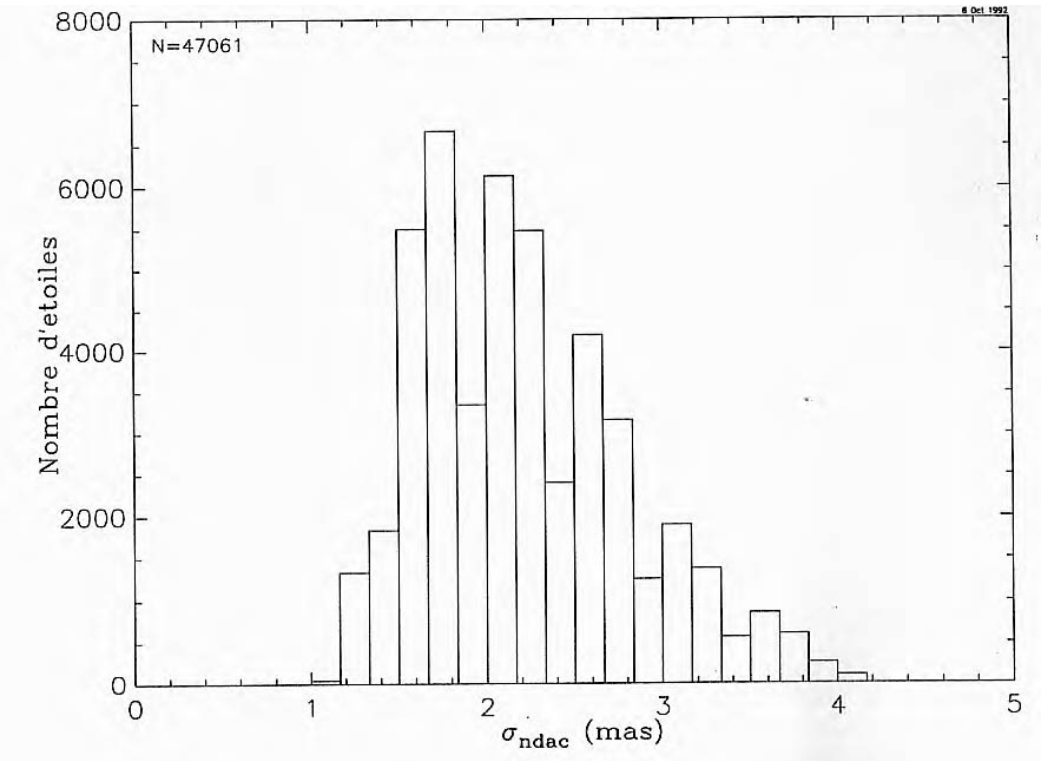


FIG. 6.4: *Distribution des erreurs internes (s_N) sur les parallaxes préliminaires obtenues avec un an de données par le consortium NDAC (mas)*

sit comme étant $\sqrt{(0.95 s_{\mathrm{F}})^2 + (0.8 s_{\mathrm{N}})^2}$. Les coefficients 0.95 et 0.8 qui peuvent sembler surprenants ont été choisis pour une raison qui sera ultérieurement explicitée. Indiquons seulement qu'ils sont dûs au fait que les erreurs sur les parallaxes sont corrélées (puisque les deux consortiums ont utilisé les mêmes données obtenues par le satellite).

Comme on peut le constater sur la première figure, la moyenne de la différence des parallaxes (-0.04 ± 0.018) n'est pas significativement différente de 0. Sur la deuxième figure, on voit que les erreurs sur la différence des parallaxes sont probablement distribuées suivant une loi gaussienne. Le nombre de points est trop important pour que tous les tests de normalité soit positifs[2], mais des sous-échantillons de taille plus réduite (5 000 étoiles) voient ces tests acceptés. Cette première comparaison est donc très encourageante, et notamment le fait que la distribution des différences de parallaxes soit si «propre» est à dire vrai inespéré.

6.2.2 Aperçu des erreurs sur les parallaxes préliminaires

En comparant les deux distributions des parallaxes (fig. 6.1 et 6.2), on a le sentiment que les erreurs sur les parallaxes NDAC-5P sont plus petites que celles sur les parallaxes FAST-3P, ceci étant corroboré par le nombre de parallaxes négatives (3 982 dans le premier cas, 5 632 dans le second cas). On pourrait d'ailleurs penser se servir de ces parallaxes négatives pour connaître l'ordre de grandeur des erreurs, comme cela a souvent été le cas pour les parallaxes acquises dans le passé.

Pour cela, on notait que, pour les plus petites parallaxes, les erreurs sur les parallaxes étaient beaucoup plus grandes que les parallaxes elles-mêmes. Par conséquent, si ces erreurs étaient supposées gaussiennes, alors la distribution des plus petites parallaxes pouvait également être supposée normale ; on pouvait alors en déduire la variance des erreurs [Upgren & Carpenter, 1977]. Clairement, on commettait là un biais puisque les vraies parallaxes, pour aussi petites qu'elles fussent, n'étaient pas nulles. À cela s'ajoute le problème que les variances des erreurs sur les parallaxes étant différentes d'une étoile à l'autre, on va probablement obtenir dans les parallaxes les plus négatives celles qui ont les variances des erreurs les plus grandes. Cette méthode suppose également que l'échantillon conservé soit représentatif de la population d'une part, et que les erreurs sur la parallaxe soient indépendantes de la parallaxe. Pour toutes ces raisons, il est, semble-t-il, inutile d'utiliser cette méthode sur les parallaxes préliminaires.

On pourrait alors faire le test suivant : prenant toutes les parallaxes négatives de FAST-3P, concernant donc des étoiles lointaines, on calcule pour ces étoiles la dispersion des parallaxes de NDAC-5P (2.59 mas) ; réciproquement avec les parallaxes négatives de NDAC-5P, il est possible de calculer les dispersions des parallaxes de FAST-3P (2.87 mas). Mais il faut alors remarquer également que ces chiffres ne sont pas réellement utilisables parce que les erreurs sur les parallaxes FAST et NDAC sont corrélées. On comprend facilement pourquoi : les deux consortiums ont utilisé les mesures provenant d'un même satellite, avec le même nombre de photons par étoile. Quantitativement, il suffit de calculer la dispersion de $\pi_{\mathrm{FAST}} - \pi_{\mathrm{NDAC}}$. Cette dispersion est 2.713 mas alors que l'on s'attendrait à $\sqrt{2.59^2 + 2.87^2} = 3.87$ mas si les erreurs n'étaient pas corrélées.

2. Paradoxe connu sous le nom de *problème de Kepler* (!) selon lequel on finira toujours, pour peu que le nombre d'observations soit assez grand, par rejeter une hypothèse nulle – alors qu'en statistique bayésienne on finira par l'accepter [Robert, 1992, p. 178]

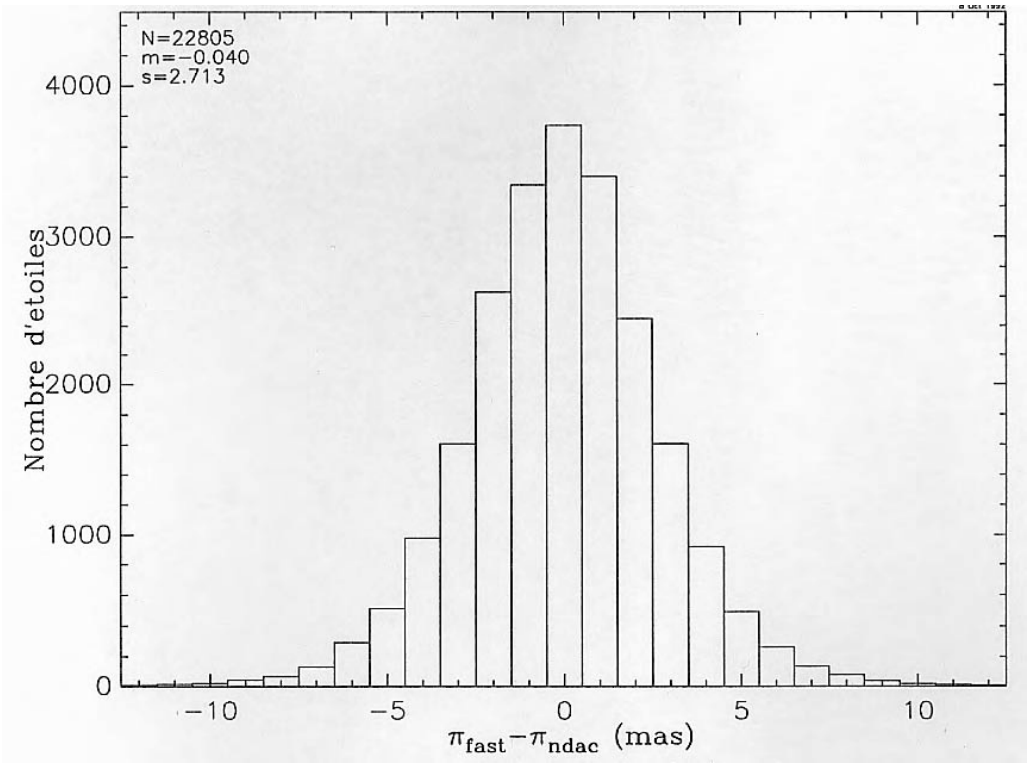


FIG. 6.5: *Distribution des différences de parallaxes FAST-3P - NDAC-5P*

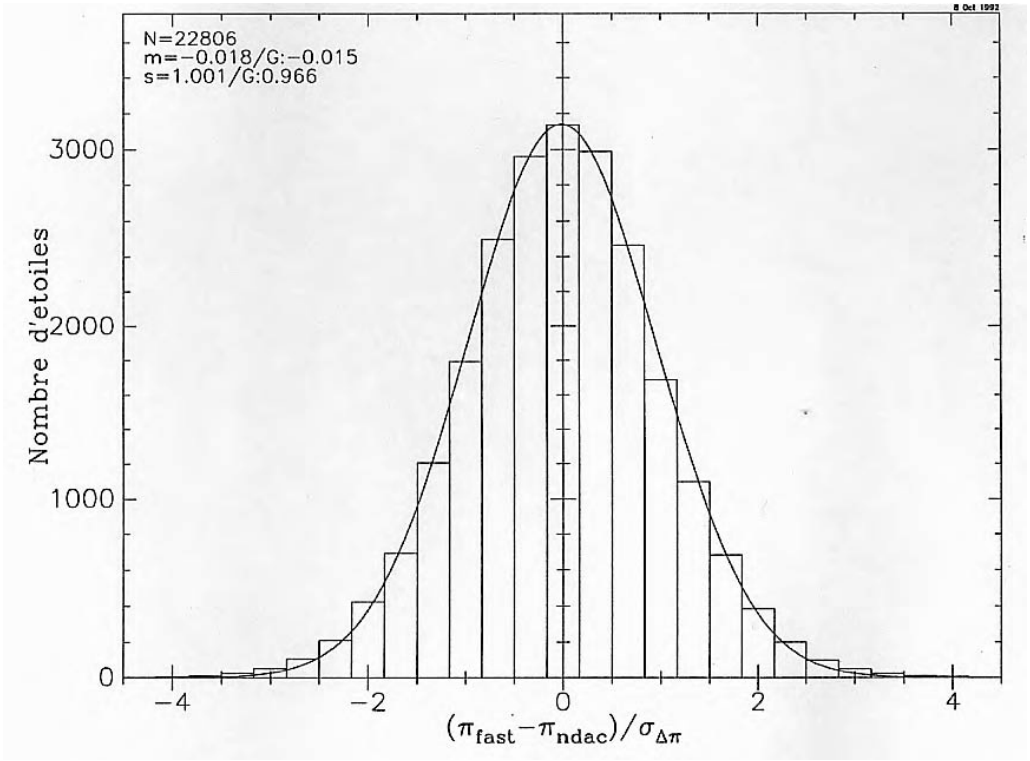


FIG. 6.6: *Distribution des différences de parallaxes FAST-3P - NDAC-5P divisées par l'erreur standard sur cette différence*

Cette corrélation ne facilite donc pas l'évaluation des erreurs sur les parallaxes, ainsi que l'évaluation du point-zéro global de ces parallaxes.

6.2.3 Biais en fonction de la parallaxe

La comparaison des parallaxes obtenues par chacun des DRC peut être effectuée de manière globale, comme ci-dessus.

Le fait qu'il n'y a pas de différence systématique est un renseignement intéressant, mais la question qui se pose également est de savoir s'il existe une variation de cette différence avec la parallaxe elle-même. Pour cela, on peut regarder ce que donnent les variations de $\pi_{\mathrm{FAST}} - \pi_{\mathrm{NDAC}}$ en fonction de π_{FAST} (fig. 6.7) et en fonction de π_{NDAC} (fig. 6.8). Une éventuelle variation serait dramatique. D'une part, elle indiquerait un problème soit de fonctionnement du satellite, soit lors de la réduction des données ; d'autre part, il serait impossible d'essayer de déterminer le point-zéro avec les parallaxes les plus lointaines puisque ce point-zéro serait dépendant de la parallaxe elle-même.

La figure 6.7 a de quoi surprendre de prime abord. On y voit un effet qui suggère une grave erreur systématique. Le même graphique fait en permutant les résultats des consortiums (fig. 6.8) montre un biais également présent, de taille légèrement plus réduite.

En réalité, le lecteur attentif aura remarqué que l'effet était tout à fait prévisible puisque les parallaxes FAST et NDAC ont des erreurs de mesure et sont corrélées, et ce biais a été étudié au §4.3.2. Cette étude statistique avait justement été suscitée par la vision de la figure 6.7 sur laquelle M. Froeschlé (1992a) avait attiré notre attention.

Le résultat que l'on peut en tirer est qu'il faut utiliser l'estimation conditionnelle de la parallaxe. Notant π_{H} la parallaxe observée par l'un des deux consortiums, $\sigma_{\pi_{\mathrm{H}}}$ la dispersion de cette parallaxe, et $f(\pi_{\mathrm{H}})$ la densité de probabilité de la distribution des parallaxes observées, l'estimateur conditionnel de la parallaxe s'écrit $\widehat{\pi} = \pi_{\mathrm{H}} + \sigma^2_{\pi_{\mathrm{H}}} \frac{f'(\pi_{\mathrm{H}})}{f(\pi_{\mathrm{H}})}$

Précisons un point de détail concernant la manière dont sont effectués ces graphiques. Sur ces figures et celles qui vont suivre, on n'a pas tracé les points individuellement car leur nombre empêcherait toute lisibilité. On a effectué un lissage en procédant de la manière suivante : pour chaque point, on a pris les $2n$ points les plus proches en abscisse, n de chaque coté, et on a calculé la médiane des ordonnées de ces $2n + 1$ points, le choix de la médiane résultant de son caractère robuste et de l'hypothèse que la distribution des erreurs est symétrique. Près des extrémités des graphiques, là où cette opération ne peut pas s'effectuer, faute d'avoir les $2n + 1$ points nécessaires, on n'a pas tracé les n premiers et n derniers points.

En général, pour cette «médiane glissante» on a choisi $n = 500$, ce qui permet de diminuer le «bruit» des données d'un facteur 40. En effet, si l'on suppose que les ordonnées des 1001 points sont distribuées normalement avec un écart-type σ, leur médiane empirique sera distribuée normalement avec un écart-type $\frac{\sigma}{\sqrt{n}}\sqrt{\frac{\Pi}{2}} \approx \frac{\sigma}{40}$. Par souci de clarté, les barres d'erreur en chaque point ne sont pas tracées, mais l'approximation ci-dessus en permet le calcul.

En particulier, si l'on prend une erreur formelle de 2.5 mas en moyenne sur les parallaxes FAST et NDAC, l'erreur standard sur chaque ordonnée des graphiques est au pire d'environ 0.09 mas, et le biais mis en évidence ci-dessus est donc largement significatif.

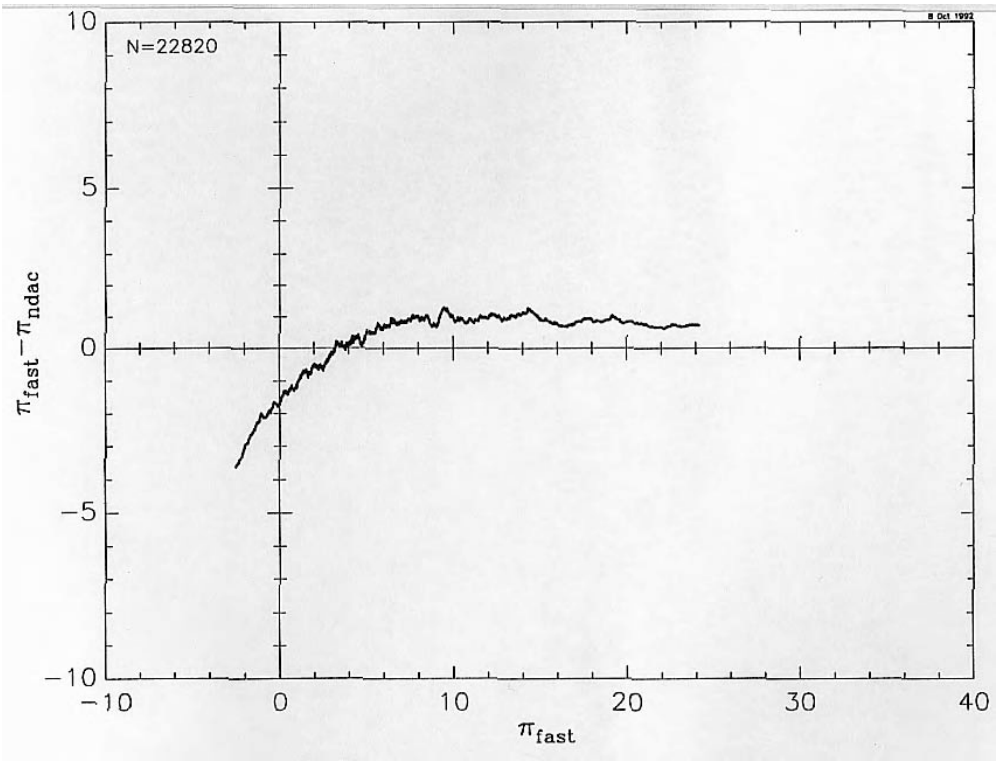


FIG. 6.7: *Lissage (1001 points) des différences* $\pi_{\mathrm{FAST}} - \pi_{\mathrm{NDAC}}$ *en fonction de* π_{FAST}

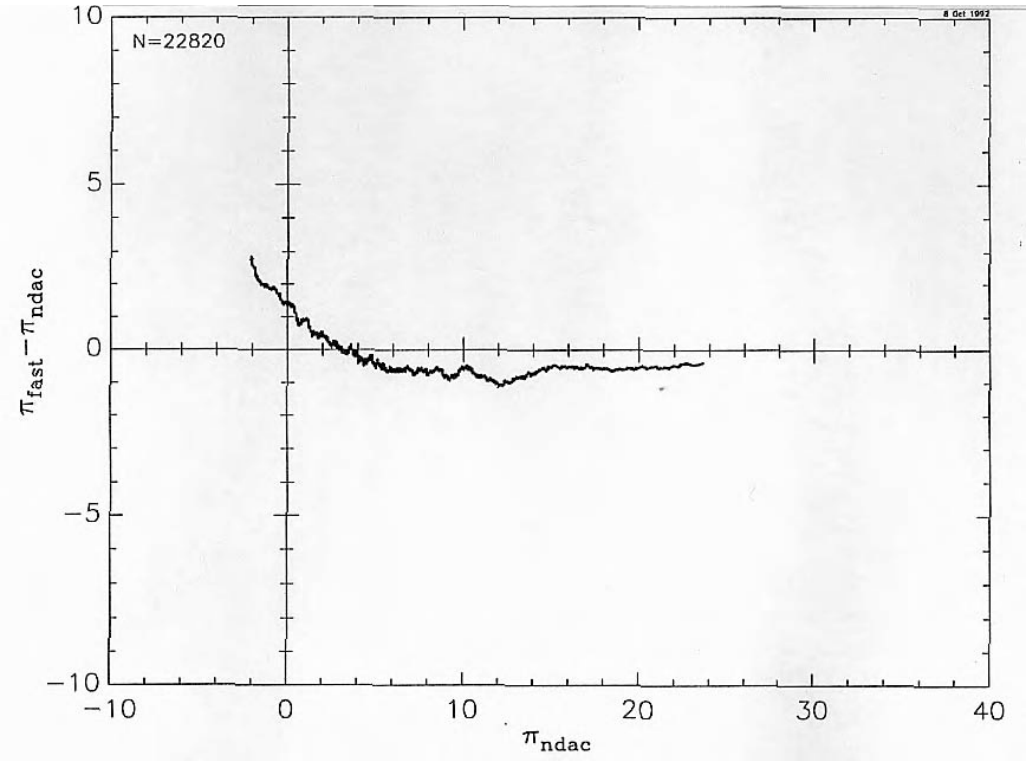


FIG. 6.8: *Lissage (1001 points) des différences* $\pi_{\mathrm{NDAC}} - \pi_{\mathrm{FAST}}$ *en fonction de* π_{NDAC}

6.2.4 Estimation des erreurs externes

Non seulement il est possible de s'affranchir de ce biais, mais en plus il est possible de l'*utiliser* pour déterminer approximativement les variances des erreurs propres à chaque consortium, de l'erreur minimum que l'on peut atteindre (en l'occurrence avec un an de données), et enfin du facteur à appliquer aux erreurs internes individuelles calculées pour chaque parallaxe pour estimer l'erreur externe.

Pour cela, regardons un peu plus attentivement à quoi correspond l'erreur sur la parallaxe dans le cadre d'un modèle simple d'erreurs additives. Pour chaque consortium, on suppose que la parallaxe obtenue est la somme de la vraie parallaxe, d'une erreur aléatoire commune aux consortiums (dûe aux effets instrumentaux) et d'une erreur de modélisation propre à la réduction de chaque consortium :

$$\begin{aligned} \pi_{\mathrm{FAST}} &= \pi + \varepsilon_C + \varepsilon_F \\ \pi_{\mathrm{NDAC}} &= \pi + \varepsilon_C + \varepsilon_N \end{aligned}$$

où ε_C, ε_F et ε_N sont par hypothèse mutuellement non corrélées, leurs variances respectives étant notées σ_C^2, σ_F^2, σ_N^2.

Les erreurs ε_F et ε_N pourraient être supposées d'espérance nulle, ce qui serait loisible puique l'on a vu que la différence $\pi_{\mathrm{FAST}} - \pi_{\mathrm{NDAC}}$ n'était pas significativement différente de 0. On ne peut pas faire de même avec $E[\varepsilon_C]$, puisque c'est justement le point-zéro de la parallaxe Hipparcos. Avec ce modèle, les erreurs «externes» sur les parallaxes s'écrivent naturellement $\sigma_{\pi_{\mathrm{F}}} = \sqrt{\sigma_C^2 + \sigma_F^2}$ et $\sigma_{\pi_{\mathrm{N}}} = \sqrt{\sigma_C^2 + \sigma_N^2}$.

Les erreurs de modélisation ε_F et ε_N ne signifient pas qu'il y a un mauvais modèle adopté pour la réduction des données, mais simplement que la solution sur la sphère a été obtenue par approximations successives, d'une part, et que d'autre part, dans le cadre de la solution 3P il y a des erreurs dans les mouvements propres du Catalogue d'Entrée, et pour la solution 5P il y a eu pollution réciproque des mouvements propres (mal déterminés avec un an de données) et de la parallaxe. On peut noter que σ_C^2 est la précision ultime que chaque consortium peut espérer atteindre après un an de mission s'il pouvait supprimer ses erreurs de modélisation.

Naturellement, la dispersion totale sur la parallaxe Hipparcos peut être réduite si l'on effectue une pondération adéquate des parallaxes FAST et NDAC, réduisant ainsi la variance des erreurs aléatoires. Nous aborderons cette question au §6.2.5.

Erreurs aléatoires de réduction

Nous allons déterminer σ_C, σ_F, σ_N, dans le cas du modèle d'erreur simple ci-dessus, sans supposer pour autant que ces variances sont les mêmes pour toutes les étoiles.

Pour cela, nous allons nous servir du biais constaté sur les figures précédentes, et qui est l'espérance de $(\pi_{\mathrm{FAST}} - \pi_{\mathrm{NDAC}})$ sachant π_{FAST} :

$$\begin{aligned} E[\pi_{\mathrm{F}} - \pi_{\mathrm{N}} | \pi_{\mathrm{F}}] &= \pi_{\mathrm{F}} - E[\pi_{\mathrm{N}} | \pi_{\mathrm{F}}] \\ &= \pi_{\mathrm{F}} - E[\pi + \varepsilon_C + \varepsilon_N | \pi + \varepsilon_C + \varepsilon_F] \\ &= \pi_{\mathrm{F}} - E[(\pi + \varepsilon_C) | (\pi + \varepsilon_C) + \varepsilon_F] \\ &= \pi_{\mathrm{F}} - (\pi_{\mathrm{F}} + \sigma_F^2 \tfrac{f'(\pi_{\mathrm{F}})}{f(\pi_{\mathrm{F}})}) \\ &= -\sigma_F^2 \tfrac{f'(\pi_{\mathrm{F}})}{f(\pi_{\mathrm{F}})} \end{aligned}$$

l'avant-dernière égalité provenant du fait que $\pi_{\rm F}$ est distribuée de façon gaussienne autour de $\pi + \varepsilon_C$ avec la variance σ_F^2. Compte-tenu de l'étude effectuée au §4.3.3, on peut appliquer alors l'équation 4.6, ce qui conduit à ce résultat. De manière analogue, on obtient également

$$E[\pi_{\rm F} - \pi_{\rm N}|\pi_{\rm N}] = \sigma_N^2 \frac{f'(\pi_{\rm N})}{f(\pi_{\rm N})}$$

Pour évaluer ces espérances, il nous faut donc déterminer les densités observées $f(\pi_{\rm F})$ et $f(\pi_{\rm N})$ et leurs dérivées, et nous utiliserons pour cela l'estimation empirique mise au point au §4.3.4.

Pour calculer σ_F et σ_N, nous allons simplement supposer que ces dispersions ont été prises en compte dans le calcul des erreurs internes $s_{\rm F}$ (resp. $s_{\rm N}$) par les consortiums, et nous écrirons $\sigma_F \approx k_F s_{\rm F}$ (resp. $\sigma_N \approx k_N s_{\rm N}$). Nous allons calculer la valeur moyenne de la constante de proportionnalité k_F (resp. k_N) simplement en découpant la distribution des $\pi_{\rm F}$ (resp. $\pi_{\rm N}$) en un certain nombre de quantiles. Cette opération a pour conséquence de mettre en évidence le biais sur un certain nombre de groupes de taille comparable. Le résultat est visible sur le tableau 6.4.

Ce rapport est assez stable, sauf aux alentours des modes des deux distributions (où $f'(\pi_{\rm H})$ s'annulle en changeant de signe), et on peut prendre comme valeur moyenne $k_F = 0.95$ et $k_N = 0.8$. On peut le vérifier immédiatement en refaisant les graphiques 6.7 et 6.8, mais cette fois-ci en appliquant respectivement les corrections $E[\pi_{\rm F} - \pi_{\rm N}|\pi_{\rm F}]$ et $E[\pi_{\rm F} - \pi_{\rm N}|\pi_{\rm N}]$ à $\pi_{\rm F} - \pi_{\rm N}$. Les graphiques correspondant 6.9 et 6.10 montrent effectivement que l'on a réussi à corriger le biais par cette méthode. Ces graphiques indiquent encore plus clairement que, une fois supprimé l'artefact, les erreurs propres à la réduction sont indépendantes de la parallaxe

Avec les valeurs adoptées $k_F = 0.95$ et $k_N = 0.8$, et sachant que l'erreur interne moyenne vaut $\langle s_{\rm F} \rangle = 2.217$ mas pour FAST-3P et $\langle s_{\rm N} \rangle = 2.158$ mas pour NDAC-5P, on trouve donc que l'ordre de grandeur des erreurs propres à la réduction est en moyenne $\langle \sigma_F \rangle \approx 2.106$ mas et $\langle \sigma_N \rangle \approx 1.726$ mas.

On peut d'ailleurs vérifier ce résultat puisque $\sigma_{\pi_{\rm F}-\pi_{\rm N}}$ a été calculé (fig. 6.5) et vaut 2.713 mas. Or $\sigma_{\pi_{\rm F}-\pi_{\rm N}} = \sqrt{\sigma_F^2 + \sigma_N^2} \approx 2.723$ mas, ce qui est très voisin.

Une deuxième manière de le vérifier consiste à calculer l'écart-type empirique de la distribution entière des parallaxes (après élimination des rares points tels que $|\pi_{\rm F} - \pi_{\rm N}| > 4\sigma_{\pi_{\rm F}-\pi_{\rm N}}$) pour les étoiles communes aux deux solutions : elle atteint 7.100 mas dans le cas de FAST-3P et 7.000 mas dans le cas de NDAC-5P. On s'attend à ce que la différence quadratique entre les deux soit voisine de la différence quadratique entre σ_F et σ_N. Dans le premier cas, cette différence vaut 1.19 mas et 1.21 mas dans le second. Un tel accord signifie marginalement qu'il n'y a que peu de points aberrants dans les deux distributions, et que la loi des erreurs n'a pas une queue «lourde». Dans le cas contraire, l'écart-type empirique aurait été un très mauvais estimateur de la largeur de la distribution.

Erreurs communes

Reste à déterminer σ_C. Ceci semble une gageure parce que l'erreur à mettre en évidence est commune aux deux distributions, et on peut à juste titre se demander comment la

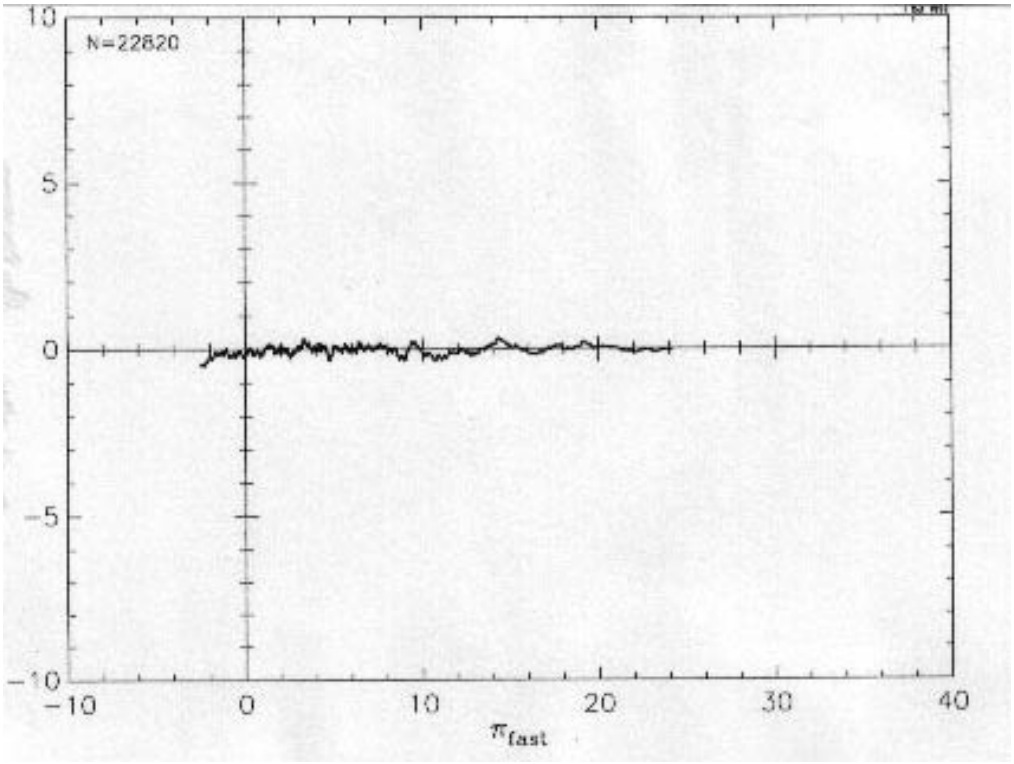


FIG. 6.9: *Variation de* $(\pi_{\mathrm{FAST}} - \pi_{\mathrm{NDAC}}) - E[\pi_{\mathrm{FAST}} - \pi_{\mathrm{NDAC}}|\pi_{\mathrm{FAST}}]$ *(mas) en fonction de la parallaxe FAST (lissage 1001 points)*

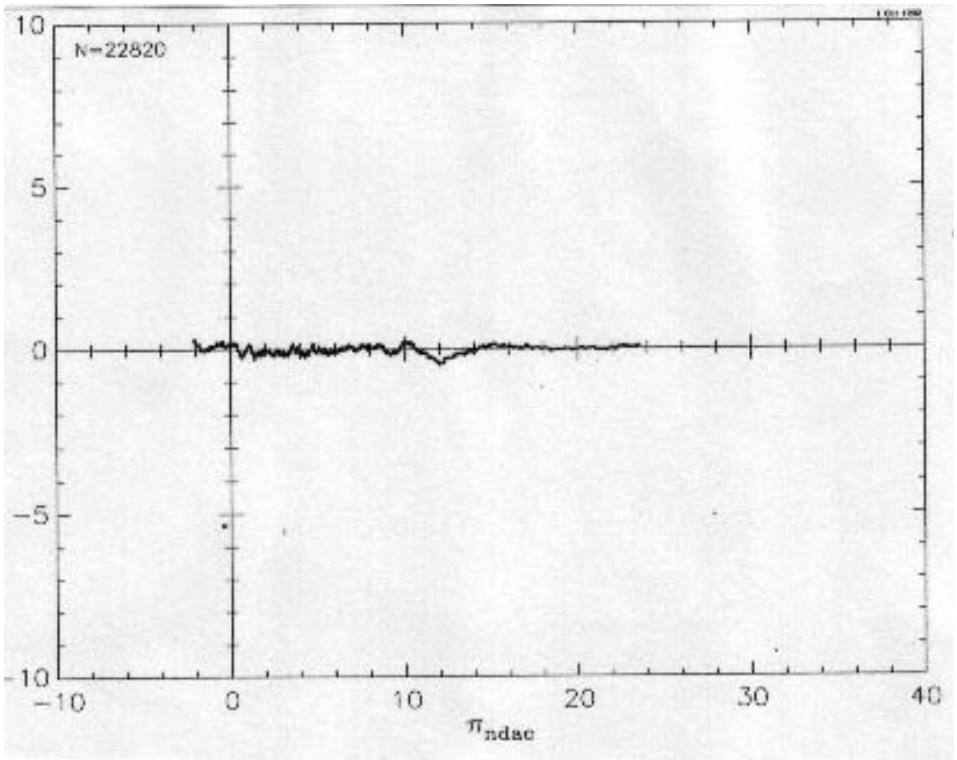


FIG. 6.10: *Variation de* $(\pi_{\mathrm{FAST}} - \pi_{\mathrm{NDAC}}) - E[\pi_{\mathrm{FAST}} - \pi_{\mathrm{NDAC}}|\pi_{\mathrm{NDAC}}]$ *(mas) en fonction de la parallaxe NDAC (lissage 1001 points)*

TAB. 6.4: *Erreur propre à la réduction.*

Calcul dans chaque quantile de la distribution des parallaxes FAST-3P (colonnes gauche) et NDAC-5P (colonnes droite) du coefficient à appliquer à l'erreur interne pour obtenir l'erreur propre à la réduction du consortium.

π_{FAST}	$\sqrt{\frac{-\langle \pi_{\mathrm{F}} - \pi_{\mathrm{N}} \rangle}{\langle {s_{\mathrm{F}}}^2 \frac{f'(\pi_{\mathrm{F}})}{f(\pi_{\mathrm{F}})} \rangle}}$	π_{NDAC}	$\sqrt{\frac{\langle \pi_{\mathrm{F}} - \pi_{\mathrm{N}} \rangle}{\langle {s_{\mathrm{N}}}^2 \frac{f'(\pi_{\mathrm{N}})}{f(\pi_{\mathrm{N}})} \rangle}}$
< -1.5	1.00	< -1.1	0.83
$[-1.5, -0.4[$	1.00	$[-1.1, 0.0[$	0.87
$[-0.4, 0.4[$	0.99	$[0.0, 0.6[$	0.85
$[0.4, 1.0[$	1.00	$[0.6, 1.2[$	0.89
$[1.0, 1.5[$	0.92	$[1.2, 1.7[$	0.82
$[1.5, 2.0[$	0.94	$[1.7, 2.1[$	0.87
$[2.0, 2.4[$	1.01	$[2.1, 2.5[$	0.84
$[2.4, 2.8[$	1.13	$[2.5, 2.9[$	0.86
$[2.8, 3.2[$	0.73	$[2.9, 3.2[$	0.74
$[3.2, 3.6[$	-	$[3.2, 3.6[$	0.85
$[3.6, 4.0[$	0.76	$[3.6, 3.9[$	0.94
$[4.0, 4.4[$	1.12	$[3.9, 4.3[$	1.43
$[4.4, 4.8[$	0.78	$[4.3, 4.7[$	1.20
$[4.8, 5.3[$	0.95	$[4.7, 5.2[$	0.63
$[5.3, 5.8[$	0.96	$[5.2, 5.7[$	0.65
$[5.8, 6.4[$	1.00	$[5.7, 6.3[$	0.76
$[6.4, 7.0[$	0.95	$[6.3, 6.9[$	0.31
$[7.0, 7.8[$	0.99	$[6.9, 7.7[$	0.71
$[7.8, 8.7[$	0.86	$[7.7, 8.7[$	0.53
$[8.7, 9.9[$	0.95	$[8.7, 9.9[$	0.63
$[9.9, 11.6[$	0.85	$[9.9, 11.5[$	0.79
$[11.6, 14.1[$	0.94	$[11.5, 14.0[$	0.75
$[14.1, 19.1[$	0.94	$[14.0, 18.8[$	0.38
> 19.1	0.90	> 18.8	-
moyenne	$0.95_{\pm.02}$		$0.80_{\pm.05}$

mettre en évidence sans faire appel à des données externes. On pourrait se donner une distribution *a priori* des parallaxes et la convoluer avec les erreurs ; pour distribution, on pourrait prendre une loi beta ou une loi log-normale (ce qui sous-entendrait que la distribution des modules de distance des étoiles observées par Hipparcos est gaussienne). Les paramètres de ces distributions seraient déterminés à partir des deux distributions observées FAST et NDAC ; mais faire varier ces paramètres consisterait aussi à faire varier la dispersion commune aux consortiums (plus on suppose grande la variance de la loi *a priori*, plus on diminue involontairement la variance de l'erreur commune). Il semble donc difficile de déterminer la variance de l'erreur commune en prenant une distribution *a priori*.

Abandonnant donc cette voie, et après avoir exploré un certain nombre d'idées qui se sont révélées infructueuses, nous allons voir que, dans notre cas particulier, l'estimation conditionnelle peut répondre à la question.

Utilisant les parallaxes NDAC, nous savons que l'espérance et la variance conditionnelles de la vraie parallaxe sachant la parallaxe observée peuvent s'écrire (équations 4.6

et 4.7) :

$$\begin{aligned} E[\pi|\pi_{\rm N}] &= \pi_{\rm N} + \sigma^2_{\pi_{\rm N}} \frac{f'(\pi_{\rm N})}{f(\pi_{\rm N})} \\ &= \pi_{\rm N} + (\sigma_C^2 + \sigma_N^2) \ln' f(\pi_{\rm N}) \\ {\rm Var}(\pi|\pi_{\rm N}) &= \sigma^2_{\pi_{\rm N}} (1 + \sigma^2_{\pi_{\rm N}} \ln'' f(\pi_{\rm N})) \end{aligned}$$

D'autre part, nous avons l'égalité suivante

$$ {\rm Var}(\pi) = E[{\rm Var}(\pi|\pi_{\rm N})] + {\rm Var}(E[\pi|\pi_{\rm N}]) \tag{6.1}$$

que l'on peut démontrer ainsi

$$\begin{aligned} {\rm Var}(\pi) &= E[\pi^2] - E^2[\pi] \\ &= E[E[\pi^2|\pi_{\rm N}]] - E^2[E[\pi|\pi_{\rm N}]] \\ &= E[E[\pi^2|\pi_{\rm N}]] - E[E^2[\pi|\pi_{\rm N}]] + E[E^2[\pi|\pi_{\rm N}]] - E^2[E[\pi|\pi_{\rm N}]] \\ &= E[E[\pi^2|\pi_{\rm N}] - E^2[\pi|\pi_{\rm N}]] + E[E^2[\pi|\pi_{\rm N}]] - E^2[E[\pi|\pi_{\rm N}]] \\ &= E[{\rm Var}(\pi|\pi_{\rm N})] + {\rm Var}(E[\pi|\pi_{\rm N}]) \end{aligned}$$

A l'aide de simulations, nous avons constaté que quand on fait varier σ_C, le second membre de l'équation 6.1 passe par un minimum. L'explication de ce minimum provient du comportement de l'estimateur conditionnel $E[\pi|\pi_{\rm N}]$, qui s'écrit comme une correction à la parallaxe observée, tendant à rapprocher une étoile du mode de la distribution des parallaxes : si σ_C est sous-estimé, la correction est trop faible, l'étoile n'est pas assez proche du mode, la distribution des $E[\pi|\pi_{\rm N}]$ est trop large et donc ${\rm Var}(E[\pi|\pi_{\rm N}])$ est trop grand. Inversement, si σ_C est surestimé, $E[\pi|\pi_{\rm N}]$ va avoir tendance à passer de l'autre coté du mode, et comme la distribution des parallaxes est asymétrique, la variance ${\rm Var}(E[\pi|\pi_{\rm N}])$ va être trop grande. Par conséquent ${\rm Var}(\pi)$ va également passer par un minimum, si le terme $E[{\rm Var}(\pi|\pi_{\rm N})]$ ne croît pas trop vite, le minimum étant atteint par une valeur voisine de la vraie dispersion des erreurs. Précisons que cette méthode est conjecturale et n'est pas applicable dans un cas général : elle semble fonctionner dans notre cas parce que nous avons une distribution asymétrique, leptokurtique, et des erreurs de mesure en moyenne peu élevées, mais ayant une certaine variation.

Nous allons donc calculer la valeur de σ_C qui minimise le second membre de l'équation 6.1. Ecrivant $\sigma_C \approx k s_{\rm N}$, la variance des erreurs externes sur la parallaxe NDAC s'écrit $\sigma^2_{\pi_{\rm N}} \approx (k^2 + 0.8^2) {s_{\rm N}}^2$. Après avoir calculé pour chacune des 22 820 étoiles, l'espérance et la variance conditionnelles, on détermine l'expression 6.1 en faisant varier k. Le tableau 6.5 montre que le minimum (deuxième colonne) est atteint pour $k \approx 0.88$.

Il faut signaler que la même méthode devrait être applicable, et donc vérifiable, en utilisant les parallaxes FAST-3P. Ceci n'a pas été possible, y compris lors de simulations, parce que nous ne sommes pas dans les conditions restrictives, mentionnées plus haut, pour lesquelles la méthode peut s'appliquer.

L'ensemble des résultats ci-dessus est résumé dans le tableau 6.6 qui indique les dispersions moyennes obtenues pour les 22 820 étoiles étudiées.

6.2.5 Meilleur estimateur de la parallaxe Hipparcos

Il est possible de vérifier l'estimation précédente des erreurs externes. Pour cela, on peut se demander quel va être le meilleur estimateur de chaque parallaxe Hipparcos définitive, sachant que l'on en aura deux...

TAB. 6.5: *Erreur commune.*

Recherche de la dispersion commune aux consortiums qui minimise la variance de la parallaxe Var(π) calculée avec les estimateurs conditionnels ; dispersion correspondante de la distribution des meilleures parallaxes Hipparcos (voir §6.2.5).

$\langle \frac{\sigma_C}{s_N} \rangle$	$\sqrt{\mathrm{Var}(\pi)}$	$\sqrt{Var(\pi_H)}$
0.00	6.805	6.9180
0.50	6.755	6.9156
0.60	6.739	6.9152
0.70	6.726	6.9150
0.80	6.716	6.9148
0.86	6.714	6.9149
0.88	6.713	6.9149
0.90	6.714	6.9149
1.00	6.722	6.9150
1.50	7.050	6.9161
2.00	8.191	6.9171

TAB. 6.6: *Bilan des erreurs.*

Erreurs standards moyennes (mas) sur les parallaxes estimées uniquement à partir des distributions FAST-3P et NDAC-5P.

Solution	interne	commune	réduction	totale	tot/int
FAST-3P	2.22	1.90	2.11	2.84	1.32
NDAC-5P	2.16	1.90	1.73	2.57	1.19

La réponse se trouve être dans l'étude que l'on a faite au §4.2.1. Par maximum de vraisemblance, on a montré que l'estimateur non biaisé de variance minimum se trouve être :

$$\widehat{\pi_H} = \frac{\frac{\pi_F}{\sigma_C^2+\sigma_F^2} + \frac{\pi_N}{\sigma_C^2+\sigma_N^2}}{\frac{1}{\sigma_C^2+\sigma_F^2} + \frac{1}{\sigma_C^2+\sigma_N^2}} \approx \frac{\frac{\pi_F}{(1.32 s_F)^2} + \frac{\pi_N}{(1.19 s_N)^2}}{\frac{1}{(1.32 s_F)^2} + \frac{1}{(1.19 s_N)^2}}$$

La variance de cet estimateur étant alors $\frac{1}{\frac{1}{(1.32 s_F)^2}+\frac{1}{(1.19 s_N)^2}}$ soit en moyenne environ 1.93 mas. Naturellement, cet estimateur n'a pas grand sens actuellement, puisque l'on utilise une solution 5 paramètres pour NDAC et une solution 3 paramètres pour FAST, mais c'est ainsi que l'on pourrait trouver la meilleure manière de combiner les parallaxes des deux consortiums à la fin de la mission.

Cet estimateur est le meilleur possible au sens de la variance. En conséquence, c'est aussi celui qui doit donner une dispersion totale de la distribution des parallaxes de notre échantillon qui soit minimale ; c'est approximativement ce que l'on constate sur la troisième colonne du tableau 6.5 (avec une barre d'erreur de 0.035 mas). Ceci permet

donc de valider le rapport des erreurs externes sur les erreurs internes que l'on a trouvé au paragraphe précédent.

Il est intéressant de noter que dans la solution 5 paramètres obtenue par le consortium Nord, les erreurs de modélisation sont semble-t-il assez faibles, peut-être grâce au traitement particulier effectué pour modéliser l'attitude du satellite, notamment l'utilisation des gyroscopes [Donati *et al.*, 1989]. Les résultats précédents indiquent également, contrairement à ce que l'on pourrait croire, que les mouvements propres obtenus en seulement une année ne dégradent pas de façon importante les parallaxes obtenues, en tout cas un peu moins que l'utilisation des mouvements propres du Catalogue d'Entrée. Certes, l'erreur moyenne de 1.90 mas commune aux consortiums n'est que la précision après un an de mission. Il serait prudent de penser que ce sera également la précision définitive, mais il est clair que l'amélioration des mouvements propres au fil du temps pourra permettre de diminuer cette erreur.

L. Lindegren (1992a) avait, le premier, fait une estimation des dispersions moyennes globales (commune et propre aux consortiums) avec ce modèle d'erreurs additives, en comparant les solutions 3 paramètres des deux consortiums, et en utilisant un modèle de la distribution des vraies parallaxes de densité $f(\pi) = c(\frac{\pi}{4.9})^{2.05}(1+\frac{\pi}{4.9})^{-2.05-4.46}$ à laquelle était convoluée une double gaussienne, d'écarts-type 0.83σ et 1.66σ, σ étant déterminé par ajustement. Il obtenait comme résultat $\langle\sigma_C\rangle = 2.07$ mas et $\langle\sigma_F\rangle = 1.99$ mas (et 1.44 mas pour l'erreur $\langle\sigma_N\rangle$ sur la parallaxe NDAC-3P). Résultats concordants, donc, avec ceux trouvés ci-dessus. Le fait que l'on trouve ici une dispersion commune plus petite (1.90 mas) est explicable, si l'on se souvient que l'erreur des mouvements propres au sol doit forcément contaminer l'erreur commune quand on compare deux solutions 3 paramètres.

Nous avons tendance à préférer la méthode développée plus haut pour plusieurs raisons. D'abord, cette méthode est non paramétrique, et n'utilise que les données observées, et aucun autre modèle ; en effet, compte-tenu de la façon dont a été fabriqué le Catalogue d'Entrée, il sera sans doute difficile de trouver un modèle vraiment adéquat pour représenter la distribution des vraies parallaxes ; pour s'en convaincre, il suffit de regarder la densité des parallaxes spectroscopiques, fig. 6.30, ou photométriques, fig. 6.31, qu'il aurait été bien difficile d'imaginer *ex nihilo*. Ensuite, la seule hypothèse qui est faite est la nature gaussienne de la loi des erreurs, dont on a vu le caractère raisonnable. De plus, on peut trouver de façon naturelle le meilleur estimateur de chaque parallaxe. Enfin, cette méthode permet d'obtenir une erreur «externe» individuelle pour chaque parallaxe, et non pas simplement une erreur moyenne.

Les résultats précédents montrent clairement que l'estimation conditionnelle était un outil puissant pour l'analyse des données. Rappelons qu'elle seule nous a permis d'avoir une estimation «externe» des erreurs instrumentales, des erreurs de réduction, et du fait que celles-ci sont indépendantes de la parallaxe. On a également obtenu le facteur multiplicatif (1.32/1.19) à apporter aux erreurs internes qui s'avèrent donc assez près de la véritable dispersion des erreurs de la parallaxe. Ceci nous permet alors de calculer le meilleur estimateur de chaque parallaxe Hipparcos et son erreur formelle associée. Le tout, avec seulement deux distributions.

Il semble à propos de faire les réserves d'usage : les résultats obtenus peuvent difficilement être qualifiés d'erreurs externes dans la mesure où ils proviennent d'une comparaison de données provenant du même satellite et il faudra les vérifier avec des données externes.

Ensuite, nous avons adopté un modèle d'erreurs additives qui suffit dans un premier temps, mais qu'il faudrait éprouver. Enfin, cette méthode nécessiterait des raffinements, en particulier des simulations pour obtenir des barres d'erreur sur les résultats et pour vérifier le domaine de validité de la conjecture que nous avons fait lors de l'estimation de l'erreur commune.

Après avoir vu ces comparaisons internes, nous allons donc dans toute la suite nous intéresser à comparer les parallaxes préliminaires avec des données externes. Entre autres, les parallaxes spectroscopiques. C'est pourquoi il faut d'abord se demander comment on peut calculer ces parallaxes de manière à en avoir un estimateur non biaisé.

6.3 Estimation des parallaxes spectroscopiques

Nous avons déjà mentionné l'existence des parallaxes spectroscopiques ou photométriques calculées à l'aide des magnitudes apparentes, des magnitudes absolues, et de l'absorption interstellaire. Compte-tenu du nombre d'étoiles possédant des données spectroscopiques ou photométriques, ces parallaxes devraient être particulièrement bien adaptées pour tester les parallaxes Hipparcos.

Mais le but étant de comparer des parallaxes trigonométriques aux parallaxes spectroscopiques, la première question que l'on doit se poser est : comment calculer une parallaxe spectroscopique, connaissant la magnitude apparente et la magnitude absolue ?

Dans un article [Smith, 1985] concernant l'estimation de la vraie parallaxe d'une étoile en utilisant à la fois sa parallaxe spectroscopique et sa parallaxe trigonométrique, Smith Jr apporte une réponse à cette question en apparence simpliste. Son propos était de montrer que la moyenne pondérée de ces deux parallaxes – et il faisait référence à ce calcul fait dans le catalogue de Gliese (1969) – n'était pas un estimateur adéquat de la vraie parallaxe. Et ceci principalement à cause du biais de Malmquist, bien que cela ne soit pas le seul problème.

6.3.1 Biais de Malmquist

Si l'on a un groupe d'étoiles de même type et que l'on veut connaître la magnitude absolue moyenne de ces étoiles, il paraît logique de calculer la moyenne des magnitudes absolues individuelles. C'est oublier l'existence des biais de sélection : si chaque étoile a été choisie aléatoirement dans la population générale des étoiles de ce type, notre groupe va être représentatif de la population ; par contre – et c'est le cas le plus fréquent – si notre groupe d'étoiles a été choisi parce qu'il s'agit de toutes les étoiles jusqu'à la magnitude apparente m, alors nous avons un biais d'échantillonnage, puisque nous n'avons gardé que les étoiles les plus brillantes. Dans ce cas, la magnitude absolue moyenne va être trop brillante : il s'agit du biais dit «de Malmquist».

De façon quantitative, si l'on suppose que la distribution[3] des magnitudes absolues est gaussienne pour chaque type d'étoile :

$$\phi(M) = \frac{\phi_0}{\sigma\sqrt{2\Pi}} e^{-\frac{1}{2}\frac{(M-M_0)^2}{\sigma_M^2}}$$

3. On note $\Pi = 3.14...$ pour réserver π à la notation de la parallaxe

où M_0 est la (vraie) valeur moyenne et σ_M la dispersion des magnitudes absolues pour ce type, et ϕ_0 la densité spatiale des étoiles de ce type, alors, pour les étoiles ayant une magnitude apparente m, Malmquist (1920, 1936) a montré que

$$\left| \begin{array}{rcl} \overline{M_m} & = & M_0 - \sigma_M^2 \frac{d \ln n(m)}{dm} \\ \sigma_m^2 & = & \sigma_M^2 \left(1 + \sigma_M^2 \frac{d^2 \ln n(m)}{dm^2}\right) \end{array} \right. \tag{6.2}$$

où $\overline{M_m}$, σ_m, $n(m)$ désignent respectivement la moyenne et la dispersion de la magnitude absolue et le nombre d'étoiles ayant la magnitude apparente m par unité de magnitude. Ce résultat reste vrai pour un échantillon de $N(m)$ étoiles complet [4] jusqu'à la magnitude m (en changeant en conséquence les définitions de $\overline{M_m}$ et σ_m), et même si l'on tient compte de l'absorption interstellaire. Autrement dit, en sélectionnant un échantillon d'étoiles jusqu'à une certaine magnitude m, la magnitude absolue moyenne $\overline{M_m}$ de l'échantillon sera plus brillante que la vraie magnitude absolue moyenne M_0, que l'on obtiendrait en sélectionnant toutes les étoiles dans un volume de l'espace.

Si l'on suppose que la distribution spatiale des étoiles est uniforme, le nombre d'étoiles jusqu'à la magnitude m s'écrit $N(m) = k.10^{0.6m}$, et on obtient alors :

$$\left| \begin{array}{rcl} \overline{M_m} & = & M_0 - 1.38\sigma_M^2 \\ \sigma_m^2 & = & \sigma_M^2 \end{array} \right. \tag{6.3}$$

On note au passage que la pente 0.6 qui est utilisée couramment est une pente théorique et qu'en toute rigueur il faudrait calculer $\frac{d \ln N(m)}{dm} = \ln 10 \frac{d \log N(m)}{dm}$ pour chaque type d'étoiles considéré, suivant sa distribution spatiale.

Récemment, Luri *et al.* (1992) ont fait le calcul du biais en utilisant une distribution spatiale réaliste des étoiles utilisées.

Si l'on continue sur le thème de la rigueur, notons également que l'on a fait l'hypothèse d'une distribution gaussienne de l'erreur sur la magnitude absolue. Cette hypothèse est vraiment du premier ordre et mériterait d'être complètement revue, dans la mesure où il n'y a *aucune raison* pour que la dispersion autour de la magnitude absolue moyenne d'un groupe soit gaussienne, pour une simple raison physique qui tient compte de l'évolution des étoiles (par exemple, une étoile classée naine ne peut pas se trouver sous la ZAMS, et, selon la théorie de l'évolution stellaire, le temps de vie n'est pas le même sur toute la largeur de la séquence principale). Une distribution plus adéquate devrait sans doute être dissymétrique et tronquée, et la correction de Malmquist est très sensible à la forme de la distribution choisie [Jaschek & Gómez, 1985].

On retiendra donc que l'expression du biais de Malmquist la plus fréquemment utilisée ($-1.38\sigma_M^2$) n'est justifiée qu'à l'aide de trois hypothèses simplificatrices : une distribution gaussienne autour de la magnitude absolue moyenne, une distribution spatiale uniforme, et un échantillon complet jusqu'à une magnitude apparente donnée.

6.3.2 Calcul des parallaxes spectroscopiques

Maintenant, plaçons-nous dans le cas inverse : supposons que nous ayons un groupe d'étoiles d'un certain type, de magnitude apparente m, de magnitude absolue M supposée distribuée normalement autour de la (vraie) magnitude absolue moyenne M_0. En tenant

4. i-e représentatif de la population complète des étoiles plus brillantes que la magnitude m

compte du biais de Malmquist, la distribution des magnitudes absolues pour des étoiles de magnitude apparente m, supposées de densité constante, devient :

$$\Phi'_m(M) = \frac{\Phi_0}{\sigma_M\sqrt{2\Pi}} e^{-\frac{1}{2}\frac{(M-\overline{M_m})^2}{\sigma_M^2}}$$

Utilisant [5] ensuite $M = m + 5\log\pi + 5$, Smith Jr trouve l'expression de la distribution des parallaxes spectroscopiques des étoiles de magnitude apparente m :

$$\phi_m(\pi) = k(m, M_0, \sigma_M) e^{-\frac{1}{2}\frac{25(\log\pi - \log\pi^*)^2}{\sigma_M^2}}$$

où $\pi^* = \pi_0 \cdot 10^{-0.368\sigma_M^2}$ est donc l'estimateur le plus probable de la parallaxe spectroscopique des étoiles de magnitude apparente m, et $\pi_0 = 10^{-\frac{(m-M_0+5)}{5}}$ l'estimateur que l'on a l'habitude d'utiliser, en utilisant l'expression de la loi de Pogson.

Smith Jr utilise ensuite la formule de Bayes pour trouver l'estimateur le plus probable de la vraie parallaxe, connaissant la parallaxe spectroscopique la plus probable et la parallaxe trigonométrique.

Comme le raisonnement apparaît correct, et que le but fixé est de comparer des parallaxes spectroscopiques aux parallaxes d'Hipparcos, faut-il donc utiliser cet estimateur π^* à la place de π_0 ?

Tout d'abord, on peut voir la différence entre ces deux estimateurs comme la combinaison de deux effets, par ordre d'importance :

- l'utilisation de la magnitude absolue moyenne biaisée $\overline{M_m}$ au lieu de M_0, c'est-à-dire la correction de Malmquist «à l'envers»,
- l'utilisation de la valeur la plus probable de la parallaxe.

En ce qui concerne le deuxième point, on peut raisonner de la façon suivante :

Notons $\langle M\rangle$ la moyenne des magnitudes absolues (suivant le cas, M_0 ou $\overline{M_m}$). Par hypothèse, la magnitude absolue M de l'étoile suit une loi gaussienne $\mathcal{N}(\langle M\rangle, {\sigma_M}^2)$, donc

$$\frac{\ln 10}{5}(M - m - 5) \rightsquigarrow \mathcal{N}(\frac{\ln 10}{5}(\langle M\rangle - m - 5), (\frac{\ln 10}{5}\sigma_M)^2)$$

Alors la vraie parallaxe de l'étoile $\pi = e^{[\frac{\ln 10}{5}(M-m-5)]}$ suit une loi log-normale dont on sait qu'en l'occurence :

- sa médiane (valeur equiprobable) est $\pi_{\text{med}} = e^{\frac{\ln 10}{5}(\langle M\rangle - m - 5)} = 10^{-\frac{(m-\langle M\rangle+5)}{5}}$,
- sa moyenne (espérance) est $\pi_{\text{med}} \times 10^{\frac{1}{2}\frac{\ln 10}{25}\sigma_M^2}$,
- son mode (valeur la plus probable) $\pi_{\text{med}} \times 10^{-\frac{\ln 10}{25}\sigma_M^2}$,
- et sa variance est ${\pi_{\text{med}}}^2 \times 10^{\frac{\ln 10}{25}\sigma_M^2}(10^{\frac{\ln 10}{25}\sigma_M^2} - 1)$ (on n'utilise en règle générale que l'approximation du premier ordre $(0.4605\pi_{\text{med}}\sigma_M)^2$)

Smith Jr a choisi le mode parce que celui-ci intervient naturellement dans l'expression de la densité, mais il n'y a pas *a priori* de raison de choisir l'une plutôt qu'une autre des 3 caractéristiques du centre de groupement de la variable aléatoire π.

5. On suppose que la magnitude apparente a été corrigée de l'absorption

Simulation

Comme nous voulons ultérieurement utiliser la moyenne de parallaxes spectroscopiques pour l'étude des parallaxes préliminaires d'Hipparcos, nous devons essayer d'obtenir son estimateur le moins biaisé, et d'abord de savoir si ce biais est vraiment important.

Pour en avoir le cœur net, nous nous sommes donc livré à une simulation d'un échantillon d'étoiles pour lesquelles la parallaxe était fixée à l'avance. Cette simulation est conduite de la manière suivante :

1. On tire suivant une loi uniforme des positions (X, Y, Z) d'étoiles (entre 20pc et $r_{\rm lim} = 500$pc), et on calcule les distances $r = \frac{1}{\pi} = \sqrt{X^2 + Y^2 + Z^2}$ en limitant r à la distance $r_{\rm lim}$

2. On fixe la vraie magnitude absolue M_0 du groupe (ici $M_0 = 0$), et on tire suivant une loi gaussienne une magnitude absolue «individuelle» M de moyenne M_0 et de dispersion intrinsèque σ_M (ici $\sigma_M = 0.5$).

3. On déduit de la parallaxe π et de la magnitude absolue M la magnitude apparente m de chaque étoile.

On a alors un échantillon d'étoiles d'un certain type, de magnitude apparente connue, de densité constante, complet en volume, et pour lequel la distribution de la magnitude absolue autour de la magnitude absolue moyenne est gaussienne. Dans le cas particulier indiqué entre parenthèses ($M_0 = 0$), cela correspond approximativement aux naines B tardives du voisinage solaire.

On calcule ensuite les parallaxes spectroscopiques,

1. avec l'estimateur usuel : $\pi_0 = 10^{-\frac{(m-M_0+5)}{5}}$ qui correspond donc à la médiane de la distribution des parallaxes, connaissant m,

2. avec l'estimateur le plus probable, connaissant la magnitude apparente, tel qu'il est indiqué par Smith Jr : $\pi^* = \pi_0 \cdot 10^{-0.368\sigma_M^2}$

3. avec l'estimateur de la moyenne d'une distribution log-normale : $\widetilde{\pi} = \pi_0 \cdot 10^{0.046\sigma_M^2}$ si l'échantillon est limité en volume et $\widetilde{\pi} = \pi_0 \cdot 10^{-0.230\sigma_M^2}$ si l'échantillon est limité en magnitude.

On compare ces trois estimateurs π_0, π^* et $\widetilde{\pi}$ à la «vraie» parallaxe π,

– dans cet échantillon complet en volume jusqu'à $r = r_{\rm lim}$,

– dans un sous-échantillon limité à la magnitude apparente $m_{\rm lim}$ que l'on estime quasiment complet (à 98%) en choisissant $m_{\rm lim} = (M_0 - 2\sigma_M) + 5 \log r_{\rm lim} - 5$

Comme ces estimateurs varient de manière importante avec la distance de l'étoile (et pour cause), plutôt que d'estimer la différence absolue entre la parallaxe estimée et la vraie, on étudie la variation relative $\delta = \frac{\pi_{\rm estimateur} - \pi_{\rm vraie}}{\pi_{\rm vraie}}$ dont les distributions se trouvent sur la figure 6.11.

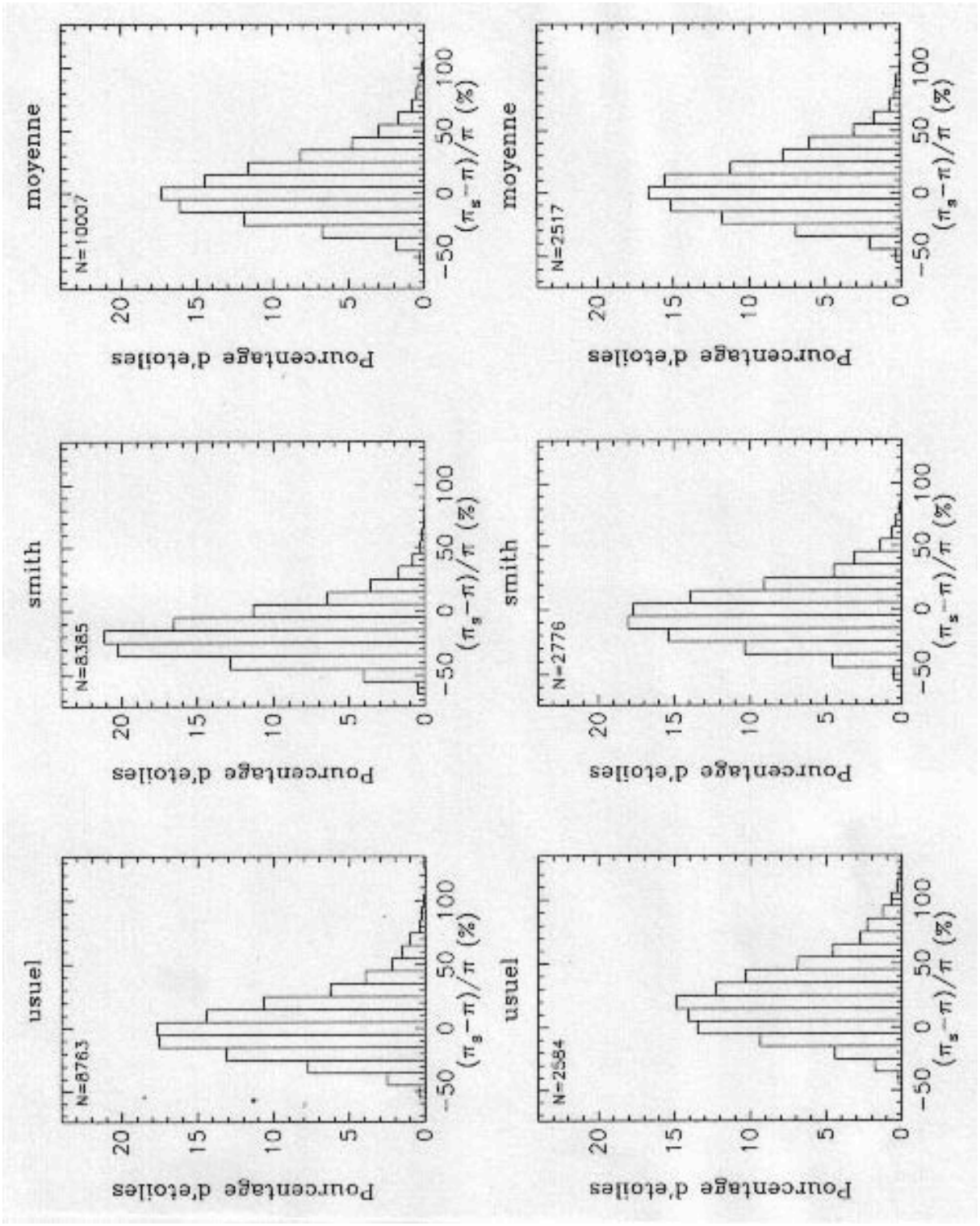


FIG. 6.11: *Comparaison des différents estimateurs de la parallaxe spectroscopique, pour l'échantillon limité à la distance* $r_{\rm lim} = 500pc$ *(en haut), pour l'échantillon limité à la magnitude apparente* $m_{\rm lim} \approx 7.5$ *(en bas) ; en abscisse, il s'agit de la différence relative des parallaxes (en pourcentage), et en ordonnée du pourcentage d'étoiles.*

De façon qualitative, on note immédiatement que dans le cas d'un échantillon limité en volume, l'estimateur de Smith Jr sous-estime d'environ 20% la vraie parallaxe, et dans le cas d'un échantillon limité en magnitude apparente, l'estimateur usuel la sur-estime d'environ 20%. Ces chiffres correspondent au cas où la dispersion des magnitudes absolues est de 0.5 magnitudes, et peuvent augmenter considérablement si σ_M est voisin de 1 magnitude.

Si l'on veut maintenant connaître plus quantitativement $\langle\delta\rangle$, on remarque que les distributions de δ sont log-normales. Ceci pose donc le problème de l'estimateur de centre à choisir... Les moyennes, médianes et modes des distributions de δ sont donc indiquées dans le tableau 6.7 dans le cas d'un échantillon limité en volume, et dans le tableau 6.8 dans le cas d'un échantillon limité en magnitude apparente. Ces résultats ont été obtenus sur 2000 simulations et les barres d'erreurs résultent de ces simulations (ces erreurs sont gaussiennes).

TAB. 6.7: *Estimateurs pour une limite en volume.*

Différence relative moyenne en % entre les différents estimateurs de la parallaxe et la vraie parallaxe, dans le cas de l'échantillon limité en volume. Cette différence moyenne est estimée à partir de la moyenne empirique, de la médiane et du mode de la distribution des différences.

Estimateur	usuel	Smith Jr	moyenne
Moyenne	2.7 ±0.1	-16.9 ±0.2	5.4 ±0.1
Médiane	-0.1 ±0.2	-19.2 ±0.2	2.6 ±0.2
Mode	-5.2 ±3.4	-23.7 ±3.2	-2.8 ±3.6

TAB. 6.8: *Estimateurs pour une limite en magnitude.*

Même légende que le tableau précédent, dans le cas de l'échantillon limité en magnitude apparente.

Estimateur	usuel	Smith Jr	moyenne
Moyenne	21.5 ±0.5	-2.7 ±0.2	5.4 ±0.3
Médiane	18.6 ±0.6	-5.0 ±0.4	2.9 ±0.4
Mode	12.9 ±5.4	-9.8 ±4.4	-2.4 ±4.7

Ces tableaux confirment le résultat qualitatif et indiquent de plus que lorsqu'on comparera les parallaxes spectroscopiques aux parallaxes Hipparcos, la manière de calculer la différence moyenne *conditionnera* le choix de l'estimateur de la parallaxe spectroscopique : mais si l'on veut utiliser la moyenne empirique, on constate qu'*aucun* des estimateurs de la parallaxe spectroscopique ne conduit à $\overline{\delta} \approx 0$.

On pouvait s'en douter analytiquement : en effet, si l'on utilise l'estimateur usuel de

la parallaxe spectroscopique, la moyenne empirique $\overline{\delta}$ s'écrit :

$$\begin{aligned}
\overline{\delta} = \frac{1}{n}\sum_{i=1}^{n}\frac{\pi_{0i}-\pi_i}{\pi_i} &= \frac{1}{n}\sum_{i=1}^{n}10^{\frac{[-(m_i-\langle M\rangle+5)+(m_i-M_i+5)]}{5}} - 1 \\
&= \frac{1}{n}\sum_{i=1}^{n}e^{\frac{\ln 10}{5}(\langle M\rangle - M_i)} - 1 \\
&\approx \frac{1}{n}\sum_{i=1}^{n}\left[1 + \frac{\ln 10}{5}(\langle M\rangle - M_i) + \frac{1}{2}(\frac{\ln 10}{5})^2(\langle M\rangle - M_i)^2 + ...\ \right] - 1 \\
&\approx \frac{1}{n}\frac{\ln 10}{5}\sum_{i=1}^{n}(\langle M\rangle - M_i) + \frac{1}{2}(\frac{\ln 10}{5})^2\frac{1}{n}\sum_{i=1}^{n}(\langle M\rangle - M_i)^2 \\
&\approx \frac{1}{2}(\frac{\ln 10}{5})^2\sigma_M^2
\end{aligned} \tag{6.4}$$

en négligeant dans (6.4) les termes d'ordre ≥ 4 (le terme d'ordre 3 s'annulant si la distribution des magnitudes absolues est gaussienne), ce qui représente une erreur relative de $\approx 3\%$ pour $\sigma_M = 0.5$.

Dans ce cas, cela suggère la correction $10^{\ -\frac{1}{2}\frac{\ln 10}{25}\sigma_M^2}$ à l'estimateur usuel, ce que l'on vérifie immédiatement en refaisant le calcul précédent avec ce nouvel estimateur. En définitive, on a donc trouvé l'estimateur de π_S qu'il faut prendre si l'on veut utiliser la moyenne empirique de parallaxes spectroscopiques :

$$\pi_S = \begin{cases} \pi_0.10^{\ -\frac{1}{2}\frac{\ln 10}{25}\sigma_M^2} & \approx \pi_0.10^{-0.046\sigma_M^2} \quad \text{(limite en volume)} \\ \pi_0.10^{-(\frac{\ln 10}{5}\frac{d\log N(m)}{dm}+\frac{1}{2}\frac{\ln 10}{25})\sigma_M^2} & \approx \pi_0.10^{-0.322\sigma_M^2} \quad \text{(limite en magnitude)} \end{cases} \tag{6.5}$$

La justesse du choix de cet estimateur de la parallaxe, quand on utilise la moyenne empirique, est montrée sur le tableau 6.9 ; malgré l'approximation utilisée, l'estimateur est valable sur la plage de variation de la dispersion des magnitudes absolues, et nettement meilleur que les trois estimateurs (π_0, π^*, $\widetilde{\pi}$) ci-dessus.

TAB. 6.9: *Variation de π_S avec la magnitude absolue.*

Différence relative en % entre l'estimateur π_S de la parallaxe spectroscopique et la vraie parallaxe, dans le cas de l'échantillon limité en volume, en fonction de la dispersion de la magnitude absolue.

σ_M	0.2	0.4	0.6	0.8	1.0	1.2	1.4
$\overline{\delta}$	-.000$_{\pm.005}$	-.001$_{\pm.023}$	.002$_{\pm.061}$	.000$_{\pm.132}$	.008$_{\pm.254}$	.017$_{\pm.448}$	.019$_{\pm.738}$

Conclusion

1. On ne peut pas calculer un estimateur de la moyenne des parallaxes spectroscopiques d'un groupe d'étoiles sans se poser la question de la forme de sélection de l'échantillon,

2. Le facteur correctif $\times 10^{-\frac{\ln 10}{5}\frac{d\log N(m)}{dm}\sigma_M^2}$ doit être appliqué à la parallaxe d'une étoile provenant d'un échantillon limité en magnitude, si l'on veut calculer la parallaxe moyenne de l'échantillon,

3. Si cette parallaxe moyenne de l'échantillon est estimée avec la moyenne empirique, il faudra appliquer le facteur correctif $\times 10^{-0.046\sigma_M^2}$.

Le dernier problème à régler concerne maintenant la notion de limitation en volume. Pour limiter un échantillon en volume, encore faut-il connaître la distance de chaque étoile, et c'est justement ce que l'on cherche à déterminer. On ne peut évidemment pas utiliser l'estimateur de la parallaxe spectroscopique ou photométrique, puisque limiter des étoiles de magnitude absolue M_0 à la distance $r_{\rm lim} = 10^{\frac{(m-M_0+5)}{5}}$ revient exactement au même que conserver celles avec $m < M_0 + 5\log r_{\rm lim} - 5$; autrement dit, en utilisant cet estimateur pour limiter en volume, on limite en fait en magnitude. Si l'on ne dispose que de la magnitude apparente pour sélectionner les étoiles, il n'est donc pas possible d'obtenir un échantillon complet en volume.

En revanche, si l'on connaît la parallaxe trigonométrique – même affectée d'une erreur de mesure – on pourrait penser s'en servir pour définir son échantillon. Ceci ne veut pas dire que l'on s'est affranchi des biais de sélection : comme la parallaxe trigonométrique a une erreur de mesure et qu'il y a plus d'étoiles en dehors du volume délimité qu'à l'intérieur, on va faire rentrer plus d'étoiles lointaines que l'on va faire sortir d'étoiles proches, biaisant de nouveau la parallaxe moyenne de l'échantillon [Trumpler & Weaver, 1953]. Mais il existe d'autres cas où l'on a une limitation en volume, et c'est par exemple le cas quand on utilise des étoiles qui sont approximativement à la même position spatiale (un amas).

6.3.3 Complétude des échantillons utilisés

Après avoir déterminé dans le cas général l'estimateur à utiliser, il faut maintenant savoir dans quel cas de figure on se trouve pour utiliser les parallaxes spectroscopiques du Catalogue d'Entrée. Quels sont les biais de sélection de l'échantillon dont on dispose ?

Il existe une limite en magnitude, qui est d'environ $H_p = 12.4$ mag, pour qu'une étoile puisse être observée par Hipparcos ; en pratique, comme le montre la figure 6.12, le nombre d'étoiles du Catalogue d'Entrée n'augmente que peu après la magnitude 9 (mais il augmente quand même). Si l'on se limite aux étoiles les plus lointaines ($\pi_{\rm S} < 2$ mas), la situation n'est d'ailleurs pas fondamentalement différente (fig. 6.13) pour ce qui est de la pente, sauf évidemment pour les étoiles faibles ou très brillantes.

La seule garantie de complétude est donnée par le «Survey», inclus dans le Catalogue d'Entrée, (presque) complet jusqu'à une magnitude $V_{\rm lim}$ qui dépend de la latitude galactique et du type spectral.

Au-delà de cette limite, non seulement il n'y a plus complétude, mais de plus on ne connaît pas les biais de sélection : comme on l'a vu au chapitre I, les étoiles ont été demandées par des proposants de la communauté astronomique pour des motifs scientifiques divers et donc avec des biais variés, puis des priorités scientifiques ont été appliquées et enfin la stratégie d'observation a conduit au Catalogue d'Entrée définitif : inutile de dire qu'il est très difficile de connaître les biais d'échantillonnage des étoiles plus faibles que $V_{\rm lim}$. À titre d'exemple des biais, on s'attend dans les étoiles faibles à trouver des naines K/M proches sélectionnées par leur grand mouvement propre, etc.

En ce qui concerne l'estimation des parallaxes spectroscopiques que nous allons utiliser par la suite, nous corrigerons donc la magnitude absolue par le biais de Malmquist, mais en

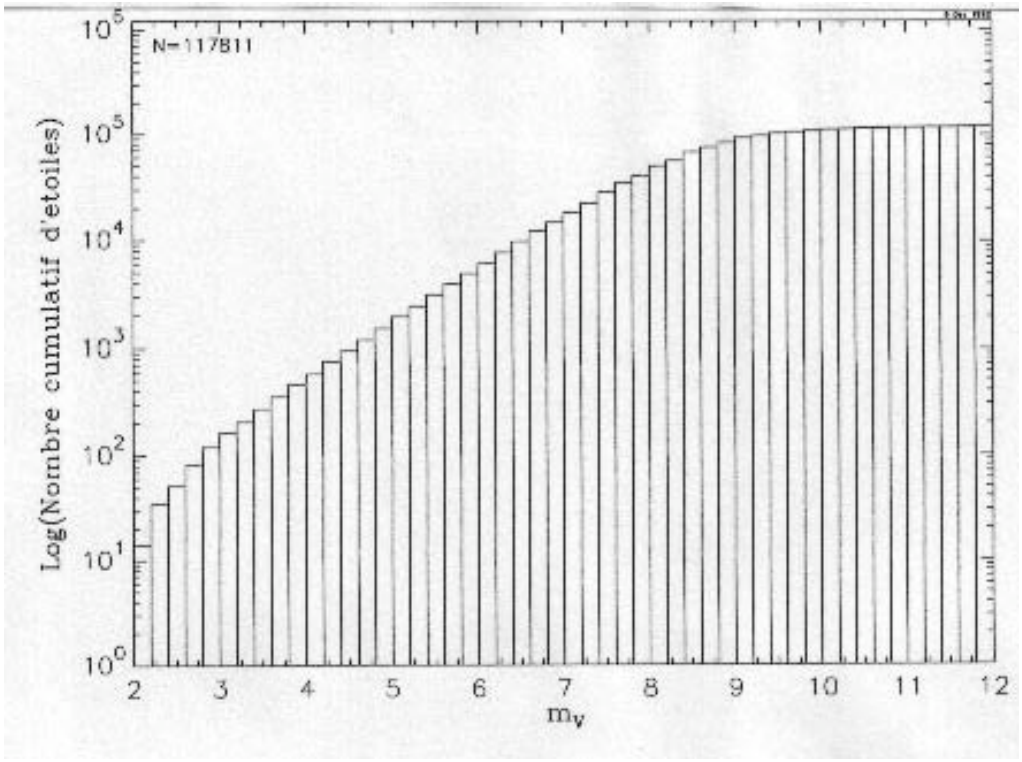


FIG. 6.12: *Distribution cumulative des magnitudes apparentes (m_V) dans le Catalogue d'Entrée d'Hipparcos.*

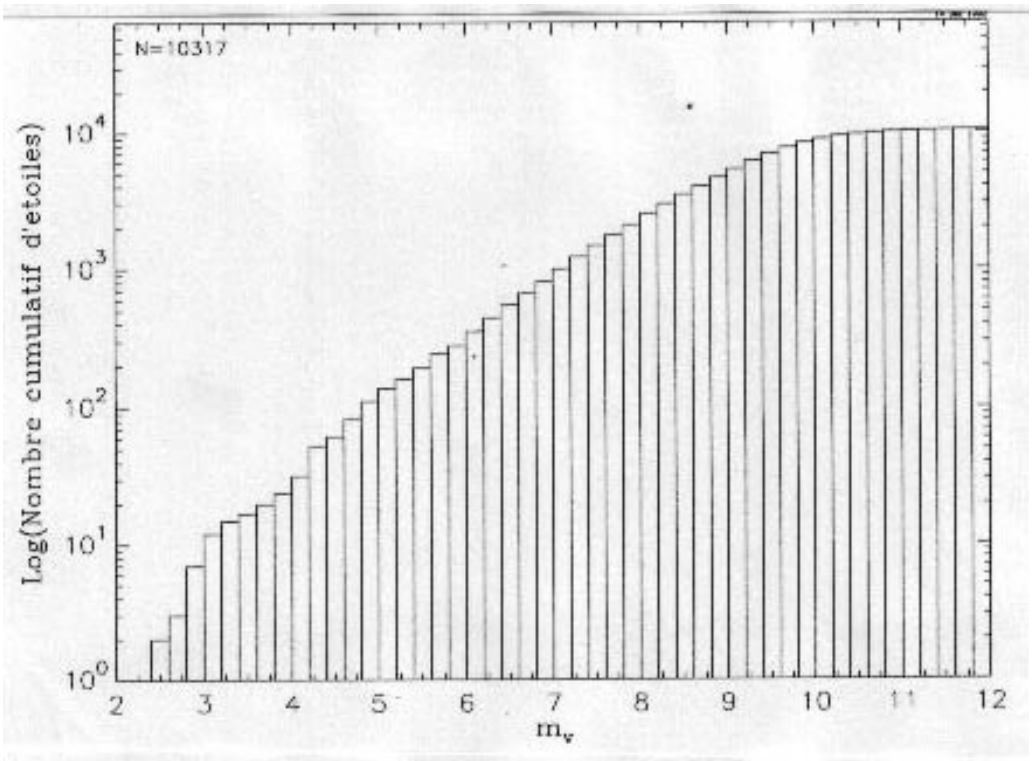


FIG. 6.13: *Distribution des magnitudes apparentes (m_V) des étoiles de parallaxe spectroscopique inférieure à 2 mas dans le Catalogue d'Entrée d'Hipparcos.*

tenant compte de la forme de $\frac{d \log N(m)}{dm}$ en fonction de m. Cette correction n'est sans doute pas optimale, mais il serait illusoire de vouloir obtenir mieux compte-tenu des données et des calibrations dont nous disposons.

Après cette longue digression sur l'estimation des parallaxes spectroscopiques, nous allons maintenant entrer dans le vif du sujet des comparaisons des parallaxes préliminaires Hipparcos avec différentes estimations des parallaxes.

6.4 Comparaison avec des estimations externes

L'étude des parallaxes préliminaires serait bien incomplète si l'on ne s'attachait pas à les comparer à des déterminations externes des parallaxes. Par déterminations externes, on entend l'essentiel des estimations de la parallaxe que l'on peut obtenir depuis le sol avec des techniques variées : utilisation des magnitudes absolues spectroscopiques ou photométriques, des modules de distance d'amas ouverts, de la distance des nuages de Magellan, des parallaxes dynamiques de systèmes binaires, et, naturellement, des parallaxes trigonométriques existantes.

Bien entendu, toutes ces estimations ont des précisions très variables, concernent un nombre d'étoiles également très variable, et aucune ne peut donc prétendre à elle-seule permettre de vérifier la qualité des parallaxes préliminaires d'Hipparcos. Ces comparaisons ne sont pas non plus exhaustives : on pourrait également s'intéresser aux distances des céphéides, etc.

Dans un premier temps, nous nous intéresserons aux différences entre les parallaxes préliminaires et ces estimations ; en analysant la forme de ces distributions et leurs deux premiers moments, ceci doit nous permettre d'obtenir une idée de l'exactitude et de la précision des parallaxes préliminaires.

Dans un deuxième temps nous étudierons la manière dont varient les éventuelles erreurs systématiques des parallaxes préliminaires, en fonction des positions, mouvements propres, magnitudes et couleurs des étoiles, et également en fonction des parallaxes elles-mêmes.

Enfin, on essaiera d'envisager la méthodologie qui pourrait permettre de déterminer les erreurs externes des parallaxes Hipparcos, ainsi que leur point-zéro. Cette méthodologie sera appliquée sur les données préliminaires.

6.4.1 Parallaxes trigonométriques

Bien évidemment, c'est tout d'abord aux parallaxes trigonométriques obtenues depuis le sol que l'on souhaite comparer les parallaxes Hipparcos. Sans vouloir opposer les deux techniques d'acquisition, qui s'avèrent finalement complémentaires, il apparaît clair que la délicate calibration des différents paramètres instrumentaux, variant avec les instruments utilisés, jointe aux problèmes de correction de la réfraction atmosphérique, etc, défavorisent le sol par rapport à l'espace.

C'est pourquoi, en comparant les parallaxes existantes à celles obtenues par Hipparcos, on est tenté de penser que l'on en apprendra plus sur les erreurs des premières que sur les erreurs des secondes.

Nous avons eu accès au CD-ROM édité par le «Astronomical Data Center» qui contient 114 catalogues concernant l'astrométrie, la photométrie et la spectroscopie [Brotzman & Gessner, 1991]. Ce CD-ROM, qui nous a été gracieusement donné par le «National Space

Science Data Center», contient la version préliminaire du Catalogue Général de parallaxes trigonométriques stellaires (GCTSP) [van Altena *et al.*, 1991]. Ce catalogue, établi avec 15 349 parallaxes pour 7 879 étoiles est une compilation où les parallaxes relatives obtenues par les différents Observatoires ont été transformées en parallaxes absolues à l'aide d'un modèle de galaxie.

Ceci nous fournit donc un échantillon important d'étoiles pour lequel on peut comparer la parallaxe «sol» avec la parallaxe «espace». Utilisant pour cette dernière la parallaxe FAST-3P, la comparaison présentée figure 6.14 montre nettement que les parallaxes sol sont beaucoup plus dispersées. On note également la présence de points aberrants, dont nous ne dévoilerons pas le nom. Les différences entre les parallaxes, dans le sens $\pi_{\rm FAST} - \pi_{\rm GCTSP}$, sont représentées sur la figure 6.15.

Sur cette figure, ainsi que sur les autres histogrammes qui suivent, on a noté en haut à gauche le nombre de points représentés, la moyenne/médiane, l'écart-type/une largeur à base de quantiles. La gaussienne qui est dessinée utilise ces deux moments robustes (médiane et largeur), et n'est représentée que dans le but d'illustrer la symétrie et l'aplatissement de la distribution des différences, et non parce qu'on supposerait normale cette distribution.

On constate que la largeur de cette distribution est d'environ 12 mas, donc de l'ordre de grandeur de la moyenne des erreurs formelles sur les parallaxes GCTSP. On note également le biais de 2 mas de la différence qui, si l'on suppose qu'il provient des parallaxes GCTSP, est surtout dû aux étoiles les plus lointaines, de parallaxe inférieure à 30 mas. Ces étoiles sont donc supposées plus proches qu'elles ne sont réellement. Dans la mesure où les parallaxes au sol sont mesurées par rapport à des étoiles supposées très lointaines, donc au mouvement négligeable, le biais peut s'expliquer si cette hypothèse n'est pas vérifiée.

6.4.2 Parallaxes spectroscopiques

La deuxième comparaison effectuée, et la plus prometteuse – si l'on s'en tient aux nombre d'étoiles en présence – consiste à calculer la différence entre la parallaxe Hipparcos préliminaire et la parallaxe spectroscopique, c'est-à-dire celle calculée au §6.3 à partir de la magnitude apparente, de l'absorption interstellaire et d'une estimation de la magnitude absolue.

En ce qui concerne la magnitude apparente, nous n'utiliserons que des magnitudes photoélectriques. L'absorption interstellaire sera calculée à partir de l'excès de couleur. Pour calculer celui-ci, on n'a gardé que les étoiles dont la couleur $B-V$ est également photoélectrique et on a utilisé la couleur intrinsèque qui provient du type spectral et de la classe de luminosité par l'intermédiaire de la calibration MK $\rightsquigarrow (B-V)_0$ de Schmidt-Kaler (1982).

Comme nous l'avons indiqué page 28, pour obtenir l'estimation des magnitudes absolues à partir des types spectraux et classes de luminosité nous prendrons la calibration MK $\rightsquigarrow M_V$ de Schmidt-Kaler (1982). Nous utilisons les types MK contenus dans la base de données INCA, et excluons les étoiles avec un type particulier (à émission, raies métalliques, etc...) et les étoiles binaires fusionnées.

Nous avons déjà évoqué ce problème de calibration M_V(MK), et notamment le fait qu'il n'est pas impossible qu'il y ait un décalage des magnitudes absolues pour certains types d'étoiles ; notamment, la calibration des géantes rouges est extrêmement incertaine.

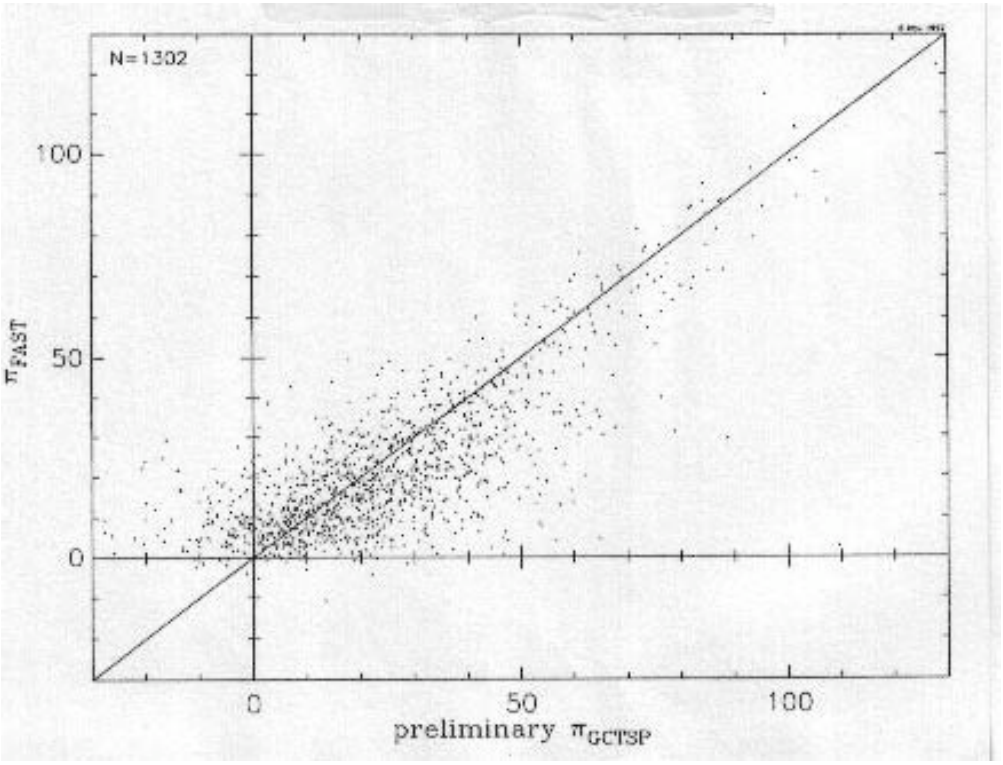


FIG. 6.14: *Comparaison des parallaxes au sol (GCTSP) et des parallaxes FAST-3P*

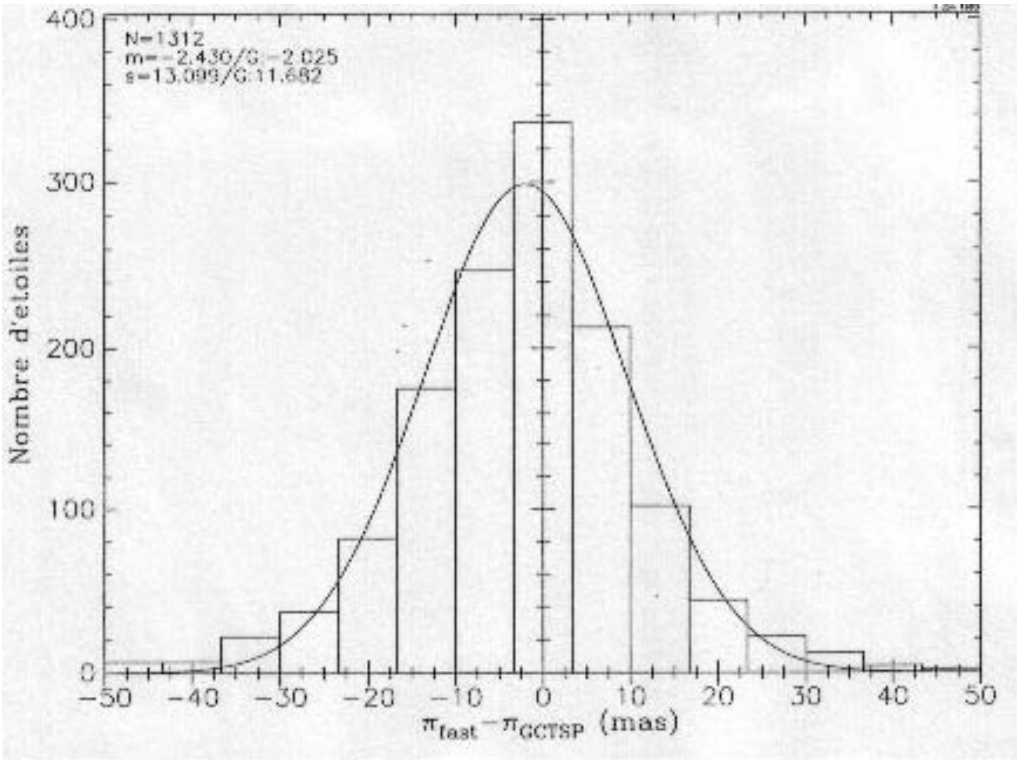


FIG. 6.15: *Différences entre les parallaxes* $\pi_{\mathrm{FAST-5P}}$ *et les parallaxes* π_{GCTSP} *en mas*

De plus, pour les étoiles assez proches, l'incertitude sur la parallaxe spectroscopique sera importante, puisque l'on a $\sigma_\pi \approx 0.46\sigma_M\pi$; pour calculer cette incertitude, on peut maintenant utiliser l'estimation que l'on a faite au §2.3.2 de la dispersion σ_M des magnitudes absolues spectroscopiques pour chaque type.

Enfin, le calcul de la parallaxe spectroscopique elle-même peut se faire grâce à l'estimation effectuée au §6.3. Nous rappelons que cette estimation doit tenir compte de la forme (log-normale) de l'erreur sur la parallaxe spectroscopique et du biais de Malmquist, et n'est donc pas la simple application de la loi de Pogson.

En comparant les parallaxes Hipparcos aux parallaxes spectroscopiques, on s'attend également à une queue de distribution importante, due à des erreurs éventuelles de classification spectrale, et la figure 6.16 ne nous détrompe pas sur ce point. Cette figure montre la distribution des différences entre la parallaxe NDAC et la parallaxe spectroscopique ; rappelons que la gaussienne qui est superposée n'implique nullement que l'on croit à l'hypothèse (fausse) de normalité.

La distribution apparaît d'ailleurs asymétrique, et ceci peut être dû à des classifications comme naines, d'étoiles qui sont en réalité géantes.

Si maintenant on se restreint aux étoiles spectroscopiquement lointaines ($\pi_S < 2$ mas), la figure 6.17 montre comment la dispersion se réduit considérablement, passant de 3.5 à 2.2 mas. Soit de l'ordre des erreurs internes sur la parallaxe NDAC, et c'est donc la première fois que par une comparaison externe on montre la précision des parallaxes Hipparcos.

Il serait extrêmement prématuré de s'intéresser au point-zéro des parallaxes préliminaires, et nous verrons pourquoi en détail au §6.5.1. Sans déflorer le sujet, on peut noter que sur l'ensemble de la distribution (fig. 6.16) le point-zéro est négatif, alors qu'il est positif quand on se limite à $\pi_S < 2$ mas, et ceci parce que l'on a tronqué la distribution à l'aide de la variable observée π_S.

6.4.3 Parallaxes photométriques

Dans toute la suite, nous appellerons parallaxes photométriques les parallaxes déduites des magnitudes absolues obtenues par les calibrations de la photométrie $uvby{-}\beta$ décrites au chapitre 2.

Nous avons vu à cette occasion que nous pouvions obtenir des magnitudes absolues individuelles dont l'incertitude est souvent plus petite que 0.3 mag, donc meilleure que la dispersion des magnitudes absolues moyennes spectroscopiques. Néanmoins, les parallaxes photométriques des étoiles proches sont tout de même incertaines. De plus, il n'est pas impossible que l'hétérogénéité des mesures des indices de la photométrie $uvby{-}\beta$ puisse créer des points aberrants. Le tracé des différences entre les parallaxes Hipparcos et les parallaxes photométriques (fig. 6.18) ne doit donc pas être pris comme la distribution des erreurs de la parallaxe Hipparcos.

Si l'on se réfère à la comparaison avec les parallaxes spectroscopiques, il est clair que la dispersion est effectivement nettement réduite, la distribution moins asymétrique, et la queue de distribution moins importante.

Limitons-nous maintenant aux parallaxes les plus petites ($\pi_P < 2$ mas), figure 6.19 ; la dispersion (2.1 mas) est identique à celle trouvée au §6.4.2, mais le mode est plus prononcé, peut-être parce que les magnitudes absolues utilisées proviennent de plusieurs calibrations dont les erreurs aléatoires sont sensiblement différentes. On verrait donc également sur

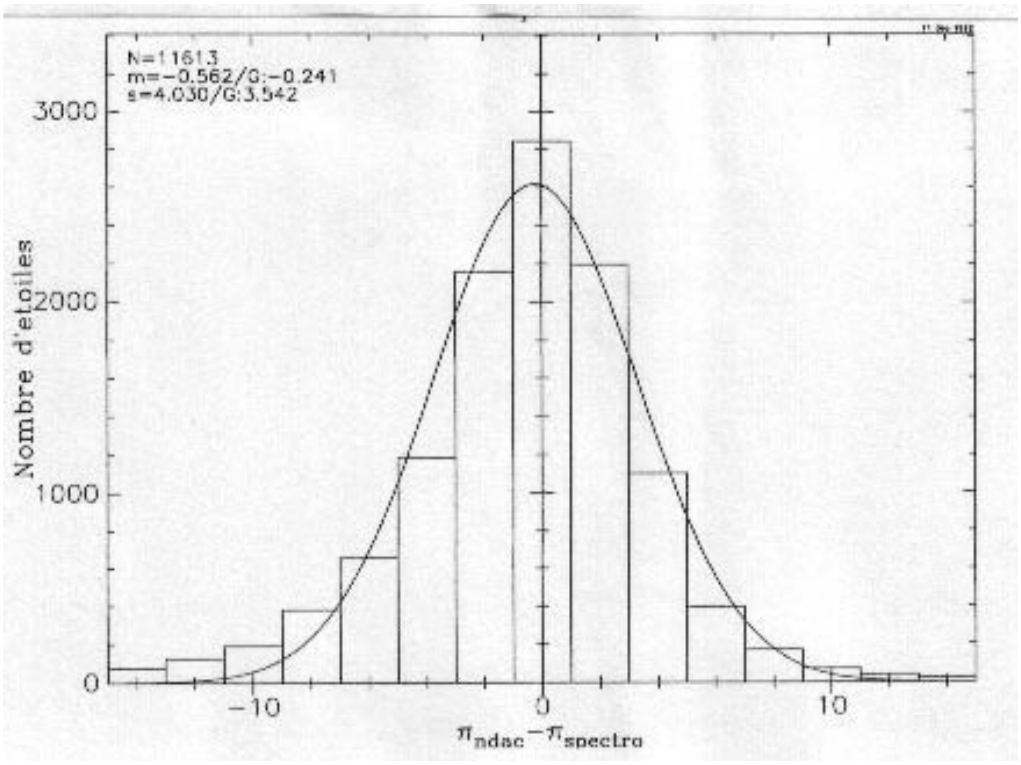


FIG. 6.16: *Différences entre les parallaxes* π_{NDAC} *et les parallaxes spectroscopiques*

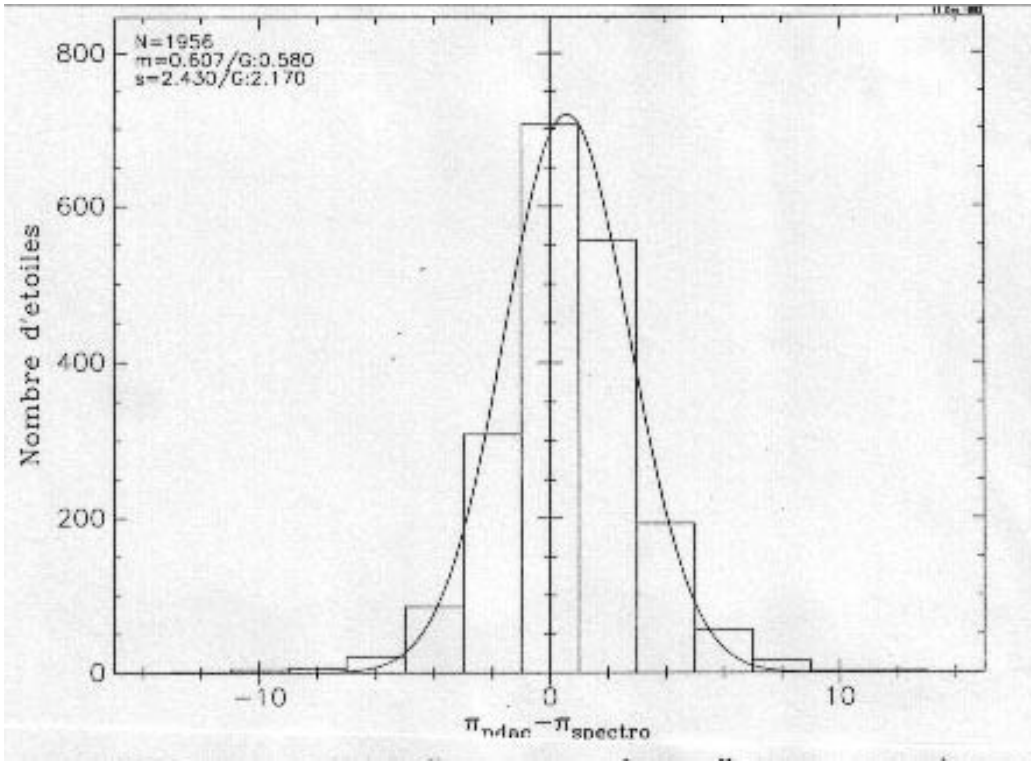


FIG. 6.17: *Différences entre les parallaxes* π_{NDAC} *et les parallaxes spectroscopiques pour* $\pi_{\mathrm{S}} < 2$ *mas*

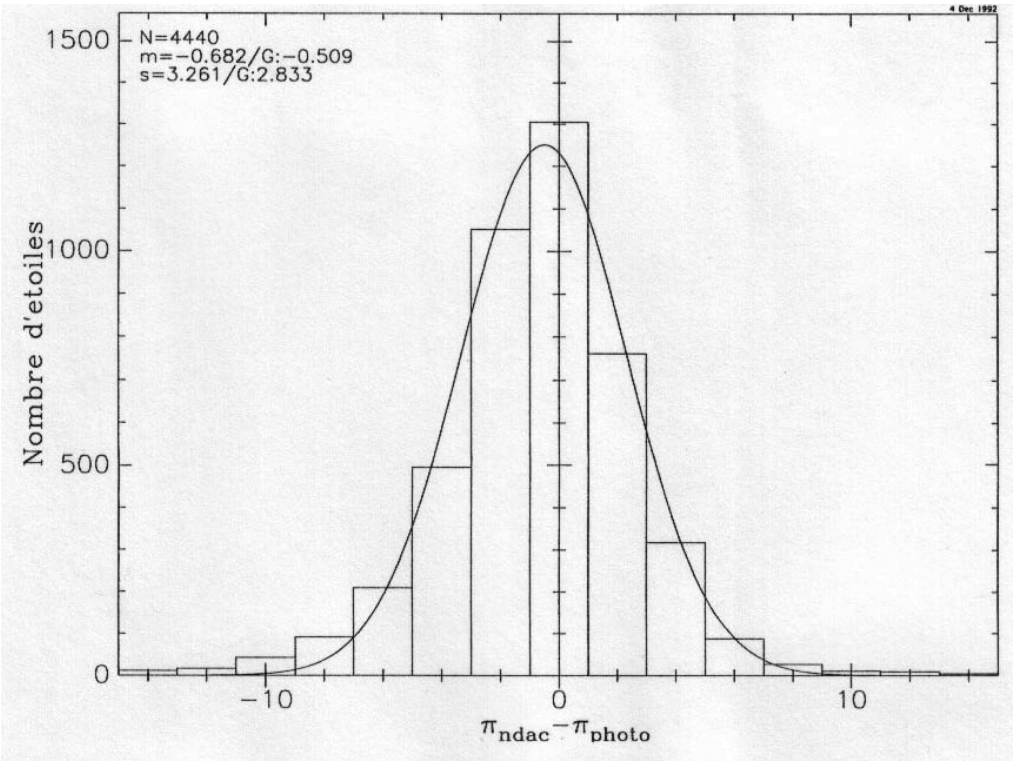


FIG. 6.18: *Différences entre les parallaxes* π_{NDAC} *et les parallaxes photométriques*

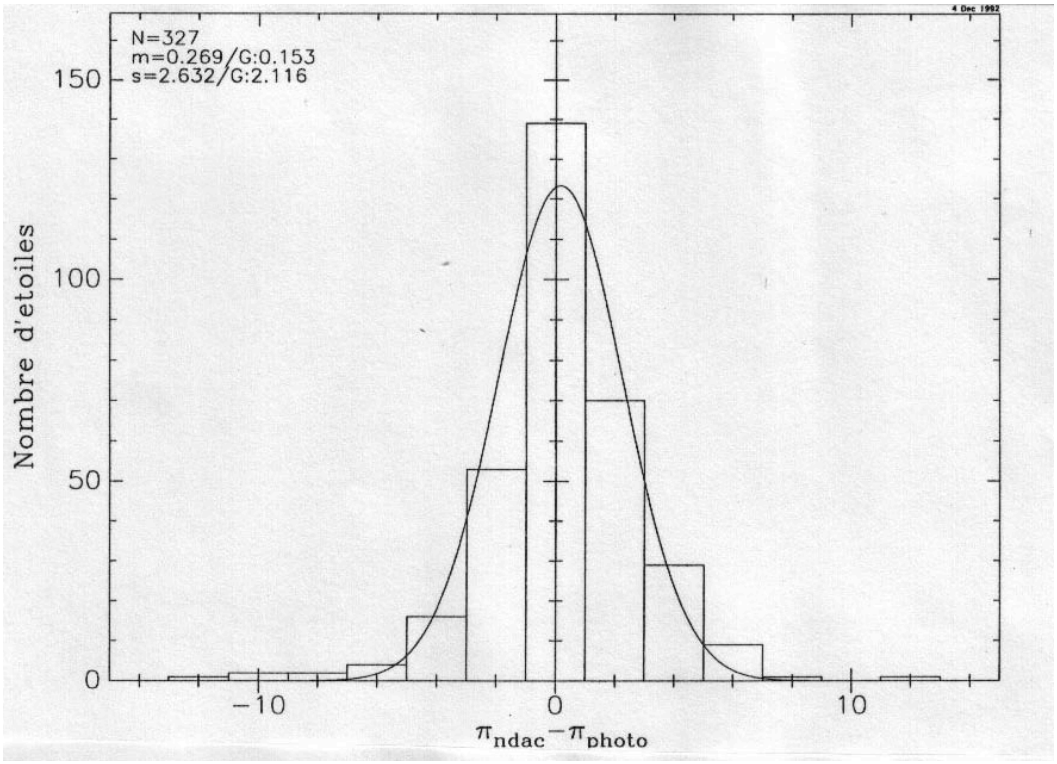


FIG. 6.19: *Différences entre les parallaxes* π_{NDAC} *et les parallaxes photométriques pour* $\pi_{\mathrm{P}} < 2$ *mas*

cette distribution une contribution des erreurs des parallaxes photométriques, bien que les étoiles soient lointaines. Si tel est le cas, la dispersion sur la parallaxe Hipparcos est donc plus petite que 2.1 mas. Ce qui n'est pas impossible si l'on se souvient que les parallaxes Hipparcos ont une erreur plus petite pour les étoiles les plus brillantes, ce qui est généralement le cas des étoiles qui possèdent de la photométrie $uvby{-}\beta$.

Quoi qu'il en soit, on constate ici encore que le point-zéro est négatif lorsque l'ensemble de la distribution est utilisé, et positif si l'on se restreint aux parallaxes observées $\pi_{\rm P} < 2$ mas.

Nous reviendrons ultérieurement beaucoup plus longuement sur les comparaisons entre parallaxes Hipparcos et parallaxes photométriques ou spectroscopiques, et allons nous intéresser maintenant à d'autres données externes.

6.4.4 Parallaxes d'amas ouverts

Pour les amas suffisamment lointains, on peut assimiler la distance de l'amas et celle des étoiles qui le composent. Par conséquent, compte-tenu de la précision des modules de distance des amas (typiquement 0.10-0.20 mag), on peut espérer une erreur interne relative meilleure que 10% sur la parallaxe de chaque étoile.

C'est dire que ces étoiles d'amas devraient fournir un excellent étalon de comparaison pour les parallaxes d'Hipparcos. Le principal problème provient du fait qu'il faut être certain que l'étoile appartient à l'amas, et qu'il ne s'agit pas d'une étoile de champ.

J.C. Mermilliod a implanté sa base de données d'amas ouverts [Mermilliod, 1992], sur notre ordinateur et la maintient à jour périodiquement. Si des données photométriques ou spectroscopiques sont disponibles pour une étoile, il est possible de lui affecter une probabilité d'appartenance à l'amas. Pour un certain nombre d'amas on dispose donc d'une liste d'étoiles suspectées non-membres.

Nous avons donc sélectionné de manière automatique les étoiles du Catalogue d'Entrée qui possèdent un identificateur d'amas, desquelles on a supprimé les étoiles suspectées non-membres, et nous avons affecté aux restantes la parallaxe de l'amas auxquelles elles appartiennent, extraite du catalogue de Lyngå(1987).

La comparaison avec les parallaxes Hipparcos se trouve figure 6.20. On peut noter la présence d'une asymétrie et d'une longue queue de distribution, dues, très probablement, aux étoiles non-membres qui n'ont pu être éliminées. Naturellement, quand la parallaxe Hipparcos définitive sera obtenue, c'est elle qui pourra servir de critère d'appartenance.

Pour l'instant, les étoiles d'amas peuvent permettre également de tester le comportement des parallaxes Hipparcos sur une petite zone du ciel. Compte-tenu de la taille du champ principal du satellite ($0.9^\circ \times 0.9^\circ$), on sait qu'il faut s'attendre en effet à des corrélations à petite échelle ($< 2^\circ$), si bien que la précision sur la parallaxe moyenne d'un groupe serré de n étoiles sera environ $\frac{\sigma_\pi}{n^{0.35}}$ au lieu du $\frac{\sigma_\pi}{\sqrt{n}}$ que l'on pourrait espérer [Lindegren, 1989, page 320]. La figure 6.21 permet de comparer la parallaxe Hipparcos moyenne de 26 amas (ceux pour lesquels on a au moins 5 étoiles) avec la parallaxe déduite du module de distance photométrique.

Si l'on ne constate aucun problème global, un amas (CL Blanco 1) illustre bien, en revanche, le caractère préliminaire des parallaxes obtenues avec un an de données. Avec un module de distance de 6.93 ± 0.06 [Westerlund *et al.*, 1988], on pourrait s'attendre à une parallaxe de 4.11 ± 0.11 mas pour cet amas, mais la parallaxe NDAC moyenne sur 12 étoiles est de 0.15 ± 0.73 mas. Après étude de l'amas en question, on note que

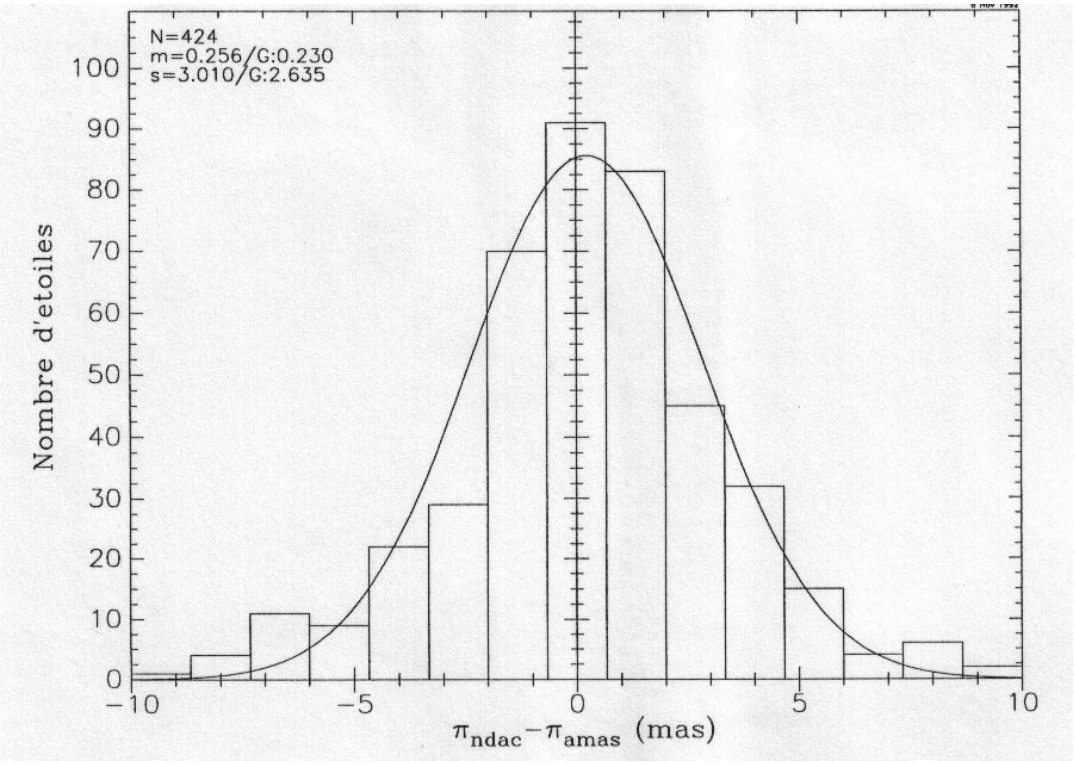


FIG. 6.20: *Différences entre les parallaxes* π_{NDAC} *et celles obtenues à partir des modules de distance d'amas*

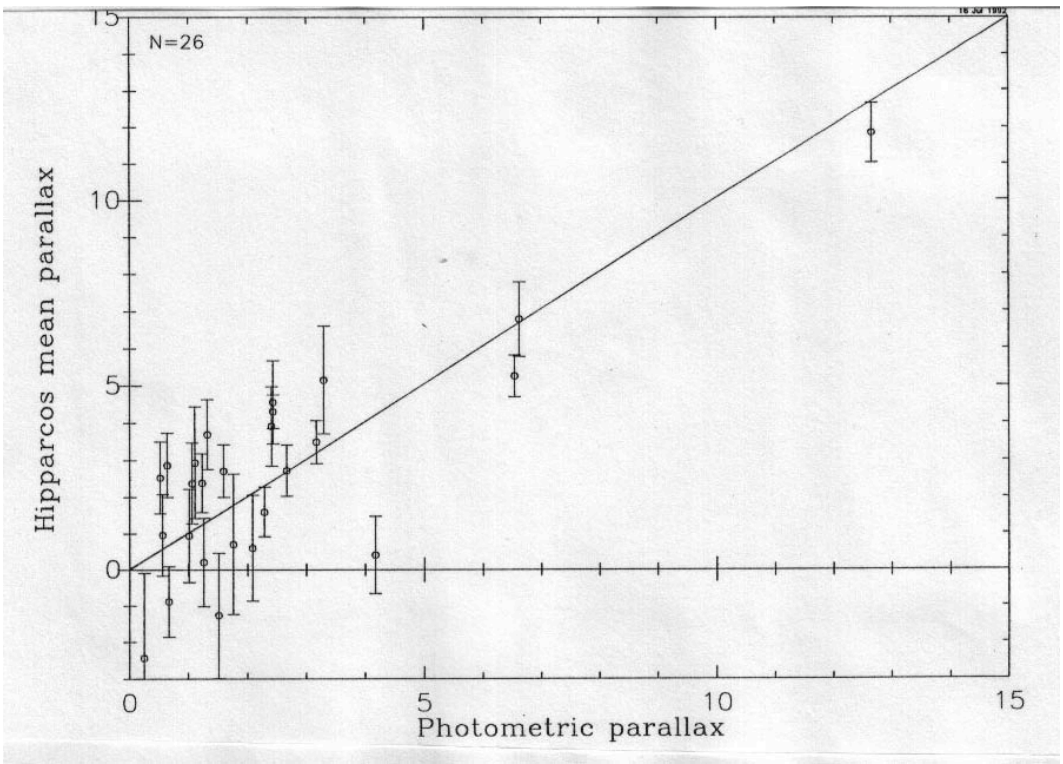


FIG. 6.21: *parallaxe de quelques amas, en abscisse déduite du module de distance, en ordonnée de la moyenne des parallaxes NDAC des étoiles de l'amas*

cette différence (significative) ne peut pas être due à la présence d'étoiles non-membres. L'hypothèse la plus plausible est que cet amas est trop proche de l'écliptique, là où les parallaxes au bout d'un an ne sont pas encore assez bien définies, compte-tenu de la loi de balayage du satellite, et la parallaxe moyenne serait donc incorrecte. Les parallaxes obtenues après un an et demi de mission devraient permettre de confirmer rapidement cette hypothèse.

6.4.5 Étoiles des nuages de Magellan

Compte-tenu de leur distance, les étoiles du grand nuage de Magellan ($\approx$ 50kpc) et celles du petit nuage de Magellan ($\approx$ 65kpc), n'auront bien évidemment pas une parallaxe Hipparcos utilisable.

Si 11 étoiles du petit nuage et 35 étoiles du grand nuage de Magellan ont été soigneusement sélectionnées pour être observées par Hipparcos [Prévot, 1992], c'est en fait dans l'espoir d'obtenir un mouvement propre, ou tout au moins une borne supérieure sur celui-ci.

Quoi qu'il en soit, ces 46 étoiles sont des candidats rêvés pour tester l'allure de la distribution des erreurs sur la parallaxe Hipparcos. Cette distribution, parfaitement gaussienne[6], est en figure 6.22, pour les 32 étoiles qui ont déjà une mesure de parallaxe. Compte-tenu du petit nombre d'étoiles et de l'absence de pollution, on peut utiliser la moyenne et l'écart-type empiriques, pour estimer les deux premiers moments de la distribution. Le résultat, dans le tableau 6.13 pour NDAC, montre que le «point-zéro» de la parallaxe Hipparcos préliminaire est voisin de 0. Il n'y a malheureusement pas assez d'étoiles pour en dire plus.

6.4.6 Parallaxes dynamiques

Si l'on connaît la parallaxe d'une étoile double, la détermination de son orbite permet de fournir la masse totale du système. Réciproquement, si l'on utilise l'orbite du système et si on a une idée de la masse (grâce à la relation masse-luminosité), on peut trouver l'expression de la parallaxe (dite «dynamique»).

Il y a dans le Catalogue d'Entrée d'Hipparcos 369 parallaxes dynamiques provenant de Dommanget (1967) et Dommanget & Nys (1982). Il peut alors être intéressant de comparer la parallaxe préliminaire d'Hipparcos à la parallaxe dynamique pour vérifier s'il n'y a pas de problèmes lors de la réduction des données sur des systèmes doubles.

La parallaxe dynamique s'écrit [Dommanget, 1992] :

$$\begin{array}{rcccccccc} \log\pi & = & 0.4096 & \log(\frac{a^3}{P^2}) & +0.0458 & m'_A & -0.4096 & \log\mu \\ & & \pm 0.002 & & \pm 0.001 & & & \end{array}$$

où a est le demi-grand axe, P la période, m'_A la magnitude bolométrique de l'étoile A et $\mu = 1 + \frac{\mathcal{M}_A}{\mathcal{M}_B}$. En règle générale, $\log(\frac{a^3}{P^2}) < 7$, $m'_A < 12$ et $\log\mu < 0.3$; l'erreur relative sur la parallaxe dynamique due aux erreurs sur les coefficients est donc inférieure à 2%.

Dans les meilleurs cas (en choisissant des orbites dont l'erreur relative sur $\frac{a^3}{P^2}$ est inférieure à 10%), l'erreur relative totale sur la parallaxe dynamique est donc inférieure à

6. les étoiles ayant approximativement la même magnitude, et étant à la même latitude écliptique, ont une erreur interne sur la parallaxe voisine

5% [Dommanget, 1992]. Il faut ajouter à l'erreur absolue une erreur uniforme d'arrondi (les parallaxes dynamiques sont données au mas près dans le Catalogue d'Entrée) de 0.3 mas. Donc lors de la comparaison avec les parallaxes Hipparcos, si l'on prend toutes les étoiles avec une parallaxe inférieure à 20 mas, on s'attend à une dispersion de moins de 1 mas sur les parallaxes dynamiques, et l'erreur sur la parallaxe Hipparcos devrait alors être prépondérante.

La distribution ($\pi_{\mathrm{NDAC}} - \pi_{\mathrm{dynamique}}$) est tracée sur la figure 6.23. Deux étoiles (HIC 95951, HIC 20765) ont été supprimées, pour lesquelles on avait une estimation externe de la parallaxe (photométrique ou spectroscopique) suggérant une erreur sur la parallaxe dynamique. Les deux premiers moments de cette distribution sont :

(NDAC-dynamique)	-0.8	± 0.45
Largeur de la distribution	3.15	± 0.39

La valeur -0.8 (médiane) est à prendre à titre indicatif, compte-tenu de la précision (au mas près) des parallaxes dynamiques. Cela signifie simplement que la distribution est (très) approximativement centrée.

Il reste néanmoins une queue de distribution non négligeable, visible sur la figure 6.23. Ces étoiles sont des couples fusionnés, dont les deux composantes sont très rapprochées (< 1 seconde d'arc) et de magnitudes voisines. Lorsque c'était possible, on a calculé une parallaxe spectroscopique et une parallaxe photométrique par la photométrie de Strömgren, pour l'étoile la plus brillante, en étant conscient du caractère dangereux de l'opération. Les valeurs obtenues sont proches de la parallaxe dynamique. On ne peut donc pas exclure un problème lors de la réduction des données de telles binaires.

La largeur de la distribution semble également plus grande que ce que l'on pourrait attendre, compte-tenu des erreurs internes des parallaxes et des erreurs externes sur la parallaxe Hipparcos que l'on a déjà évaluées. Supprimons la queue de distribution, et regardons la distribution normalisée (la différence des parallaxes divisée par l'erreur interne sur cette différence). On s'attend à une distribution gaussienne $\mathcal{N}(0, 1^2)$. Au seuil de 5%, un test de Kolmogorov ne rejette pas cette hypothèse. Mis à part les quelques couples mentionnés plus haut, ceci suggère que les parallaxes (NDAC et dynamiques) n'ont pas d'effet systématique important, et que leurs erreurs standards sont correctement estimées. Ceci-dit, compte-tenu des incertitudes sur les orbites et sur la relation masse-luminosité, on ne peut naturellement pas se servir des parallaxes dynamiques pour valider les parallaxes Hipparcos.

6.5 Variations des erreurs systématiques de la parallaxe

Nous avons déjà précisé l'importance d'une bonne connaissance des erreurs sur la parallaxe Hipparcos. Si l'on montre que l'éventuelle erreur systématique sur la parallaxe ne dépend pas de la position, des mouvements propres, de la magnitude ou la couleur, ni, enfin, de la parallaxe elle-même des étoiles, on pourra légitimement penser qu'il n'y a de problèmes ni lors de l'acquisition, ni lors de la réduction des données. On pourra alors essayer de déterminer le point-zéro des parallaxes préliminaires avec les étoiles les plus lointaines, puisque l'erreur systématique sera indépendante de la parallaxe.

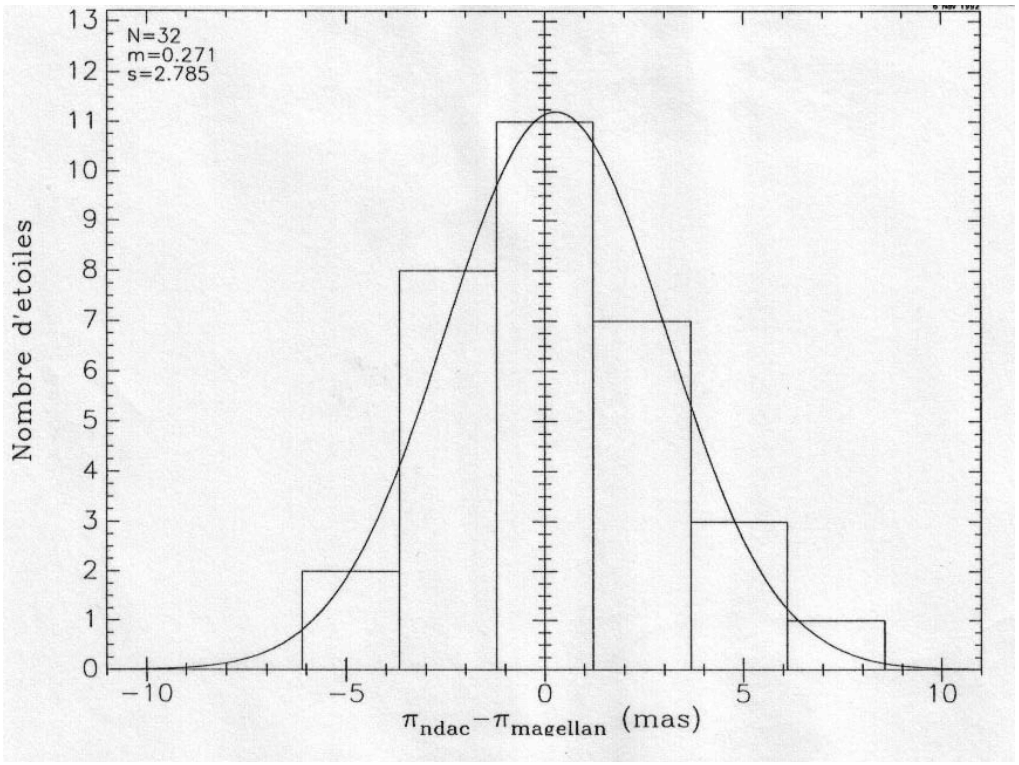


FIG. 6.22: *Différences entre les parallaxes* π_{NDAC} *et celle des étoiles des nuages de Magellan*

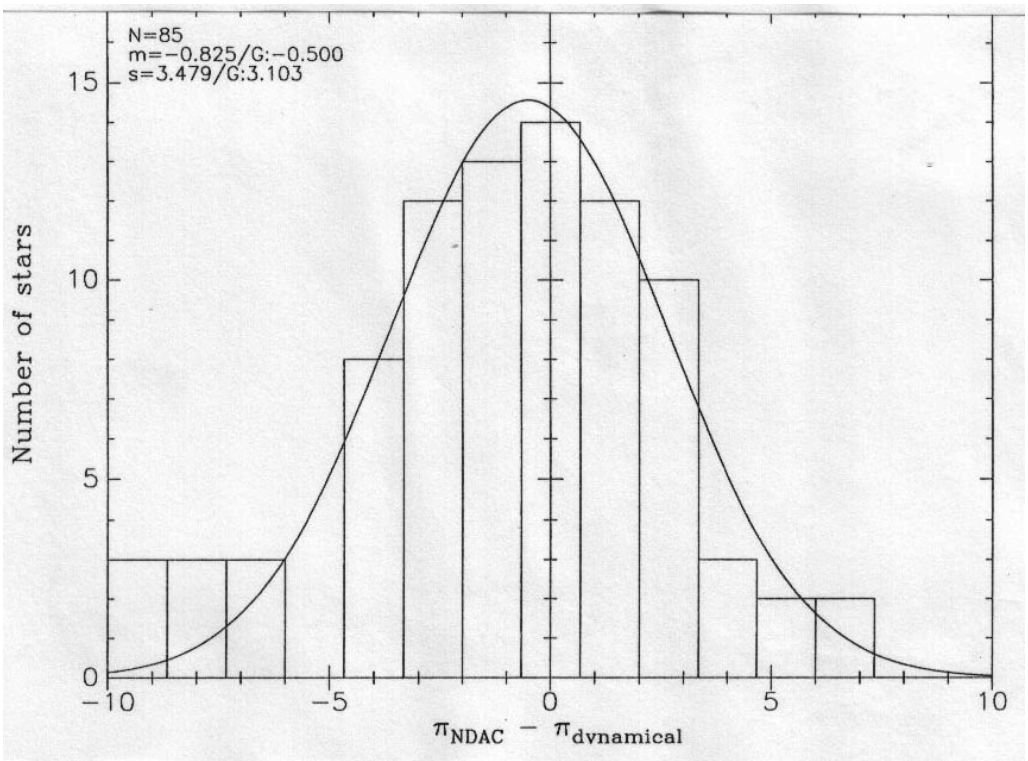


FIG. 6.23: *Différences entre les parallaxes* π_{NDAC} *et les parallaxes dynamiques*

6.5.1 Variation avec la parallaxe

Nous abordons donc ici une partie qui concerne la variation des erreurs systématiques sur la parallaxe en fonction de la parallaxe. Pour étudier cette variation, nous devons utiliser une estimation externe de la parallaxe, et les seules estimations externes pour lesquelles on ait suffisamment d'étoiles aussi bien proches que lointaines sont les parallaxes spectroscopiques et photométriques.

Dans la mesure où nous utilisons l'estimation $\langle \pi_{\rm H} - \pi_{\rm S} \rangle$ ou $\langle \pi_{\rm H} - \pi_{\rm P} \rangle$ pour étudier l'éventuelle erreur systématique sur la parallaxe Hipparcos, nous parlerons de différence systématique.

Mise en évidence du biais sur la parallaxe

L'idée consiste donc à tracer la différence $\pi_{\rm N} - \pi_{\rm S}$ en fonction de $\pi_{\rm S}$ où on note $\pi_{\rm S}$ la parallaxe spectroscopique. De même, notant $\pi_{\rm P}$ la parallaxe photométrique, on veut donc étudier les différences systématiques de la parallaxe NDAC, $\pi_{\rm N} - \pi_{\rm P}$, en fonction de $\pi_{\rm P}$. On trace donc à l'aide d'un lissage ces variations (figures 6.25 et 6.24), et on observe alors un biais positif pour les petites parallaxes, puis de plus en plus négatif au fur et à mesure qu'augmente la parallaxe. Comme ce biais atteint plusieurs mas, surtout pour les parallaxes spectroscopiques, il y a lieu de s'inquiéter.

On peut alors penser que ce biais est dû au fait que $\pi_{\rm P}$ a une erreur qui est corrélée avec $\pi_{\rm N} - \pi_{\rm P}$. Si l'on suppose que l'erreur sur la parallaxe spectroscopique et l'erreur sur la parallaxe photométrique sont indépendantes, ce qui est le cas si la première dépend d'une classification spectrale inadéquate et la deuxième des erreurs observationnelles sur les indices photométriques, alors on pourrait penser que $(\pi_{\rm N} - \pi_{\rm S})$ en fonction de $\pi_{\rm P}$ ne devrait pas exhiber de biais. La même observation devrait être valable pour $(\pi_{\rm N} - \pi_{\rm P})$ en fonction de $\pi_{\rm S}$. Malheureusement les figures 6.26 et 6.27 montrent qu'il n'en est rien.

Pour expliquer ce biais, on a alors pensé écarter l'idée d'un décalage global des magnitudes absolues et revenir sur l'hypothèse initiale d'indépendance entre l'erreur sur la parallaxe spectroscopique et celle sur la parallaxe photométrique.

En prenant les étoiles ayant à la fois une parallaxe NDAC, spectroscopique et photométrique, et plus proches que 50 pc (c'est-à-dire une parallaxe Hipparcos plus précise que 10%), on peut calculer une magnitude absolue – corrigée du biais de «Lutz-Kelker» par la formule analytique donnée par Smith Jr (1987d) – avec une incertitude inférieure à 0.22 mag. Pour cet échantillon de 294 étoiles, nous pouvons calculer des modules de distance NDAC, spectroscopiques et photométriques, puis les magnitudes absolues correspondantes, si l'on suppose que l'on n'a que peu d'erreur sur la magnitude apparente et sur l'absorption interstellaire. $M_{\rm spectro} - M_{\rm NDAC}$ apparaît alors nettement corrélée avec $M_{uvby-\beta} - M_{\rm NDAC}$ (fig. 6.28).

Le coefficient de corrélation linéaire est 0.5, et le τ de Kendall est 0.3, de probabilité quasiment nulle, et on rejette donc l'hypothèse d'indépendance. Si ceci pourrait expliquer le biais, il reste à connaître la raison de cette dépendance entre les erreurs sur les magnitudes absolues spectroscopiques et photométriques.

Si l'on étudie dans l'échantillon les étoiles avec les erreurs les plus élevées, aucune cause ne semble satisfaisante : pas de binarité nouvellement mise en évidence (par le consortium FAST), ni de problème de calibration portant sur un type particulier d'étoiles (ces étoiles sont un mélange de naines F, G ou K). Il semble d'autre part peu probable que cela

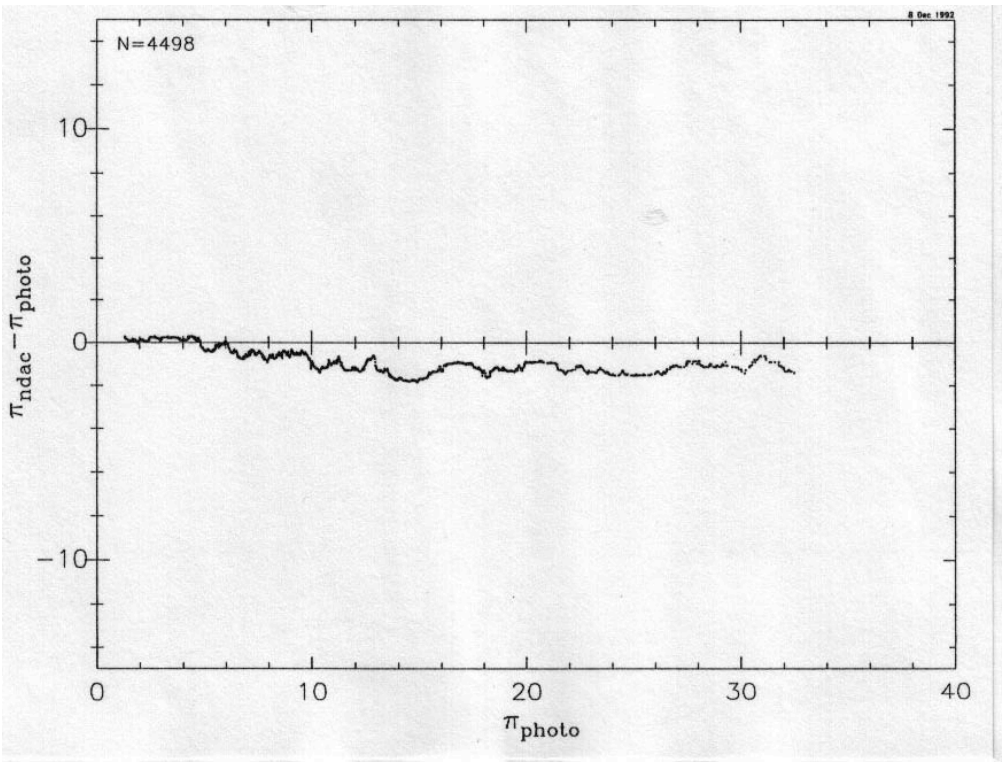


FIG. 6.24: *Lissage (401 points) des différences* $\pi_{\mathrm{NDAC}} - \pi_{\mathrm{P}}$ *en fonction de* π_{P}

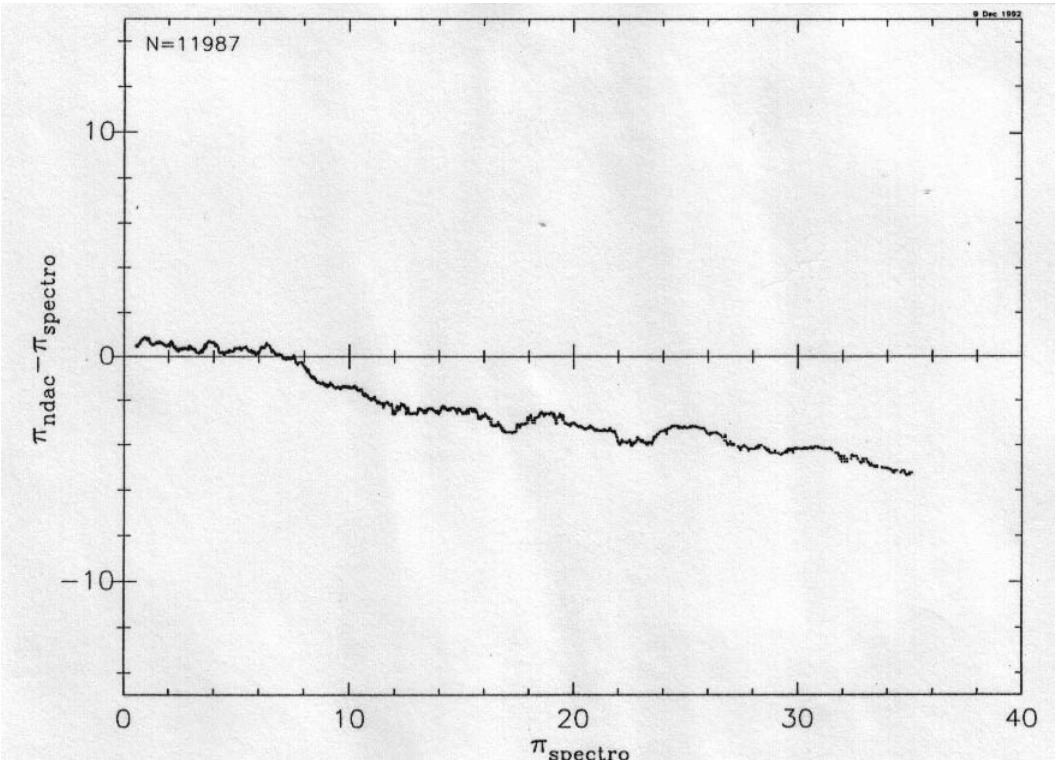


FIG. 6.25: *Lissage (801 points) des différences* $\pi_{\mathrm{NDAC}} - \pi_{\mathrm{S}}$ *en fonction de* π_{S}

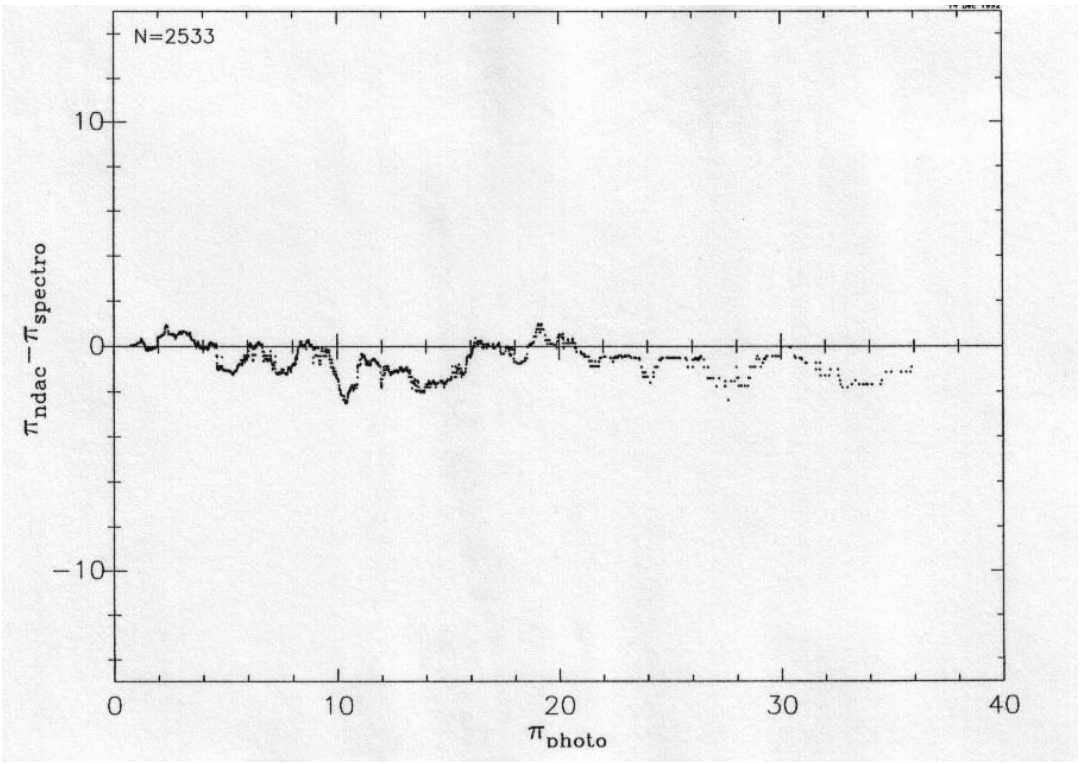


FIG. 6.26: *Lissage (401 points) des différences* $\pi_{\mathrm{NDAC}} - \pi_{\mathrm{S}}$ *en fonction de* π_{P}

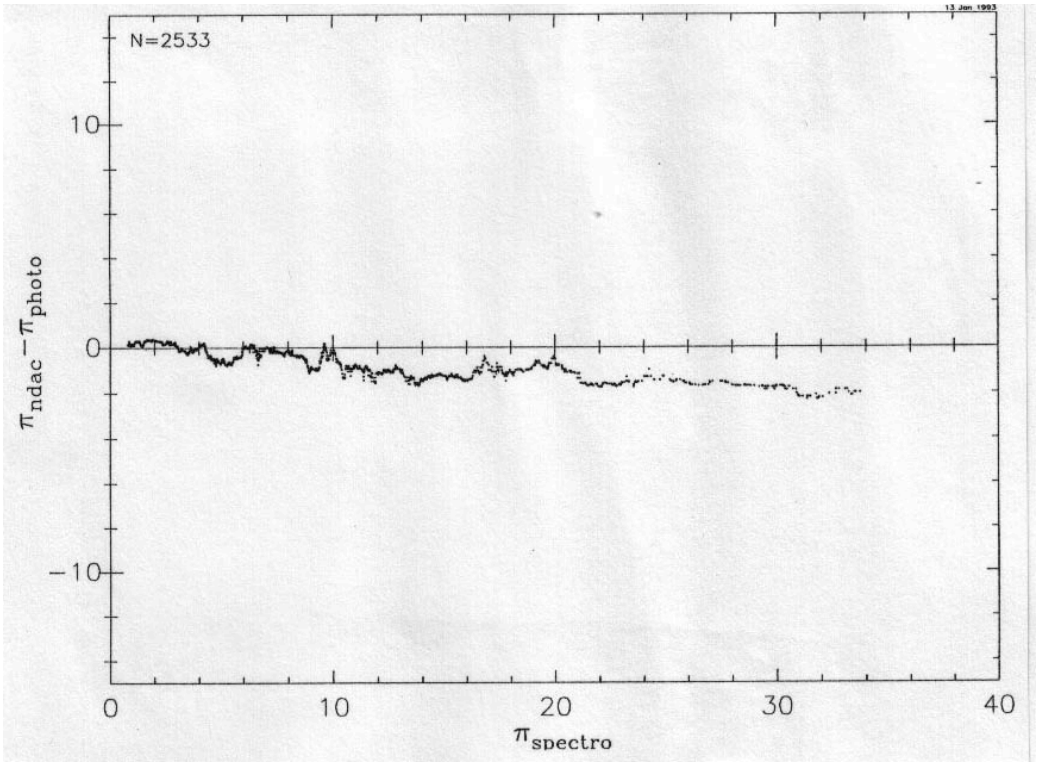


FIG. 6.27: *Lissage (401 points) des différences* $\pi_{\mathrm{NDAC}} - \pi_{\mathrm{P}}$ *en fonction de* π_{S}

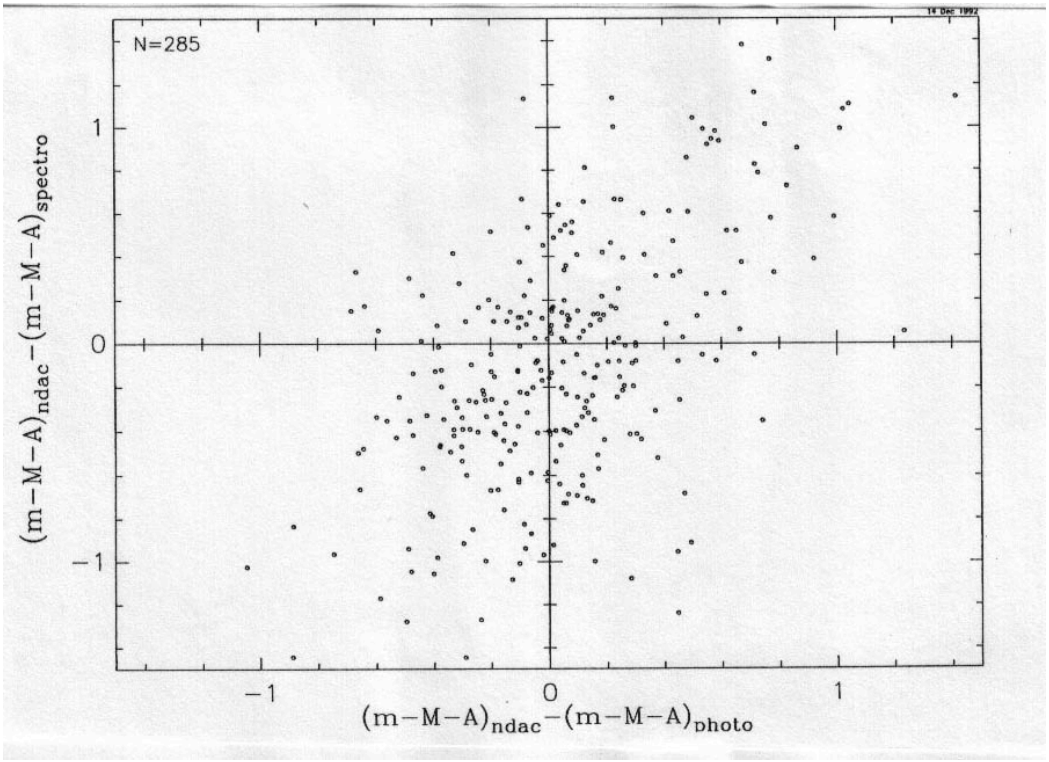


FIG. 6.28: *Corrélation entre* $M_{\rm spectro} - M_{\rm NDAC}$ *et* $M_{uvby-\beta} - M_{\rm NDAC}$ *pour les étoiles avec* $\pi_{\rm NDAC} > 20$ *mas*

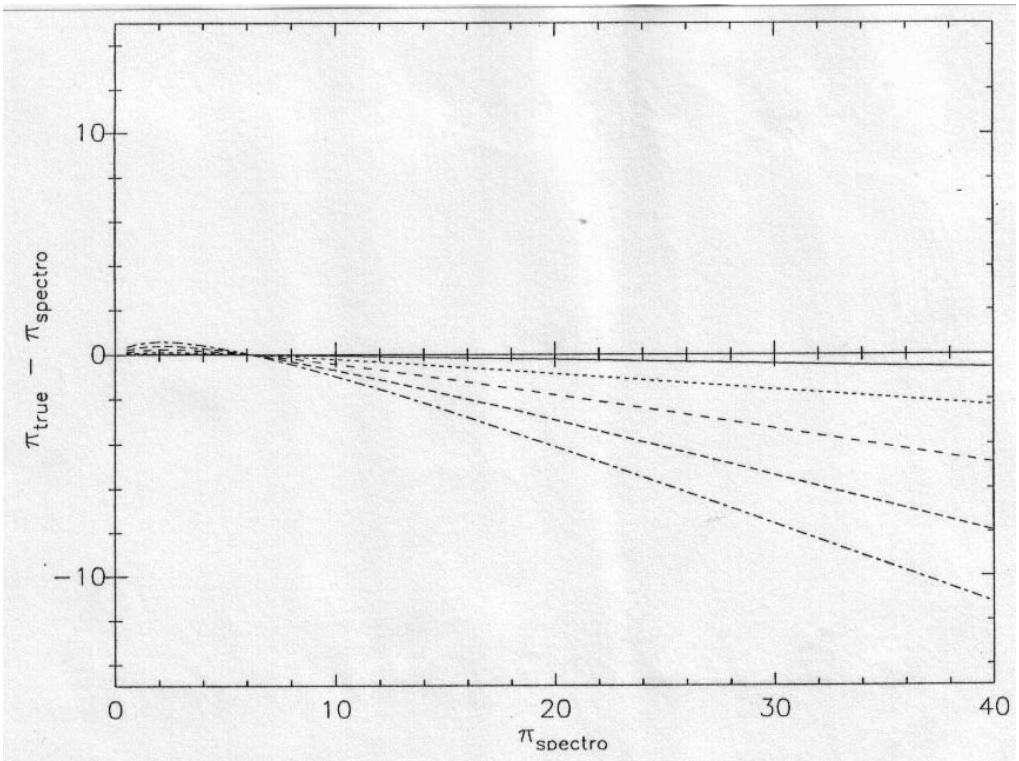


FIG. 6.29: *Biais théorique dû à la dispersion (0.2, 0.4, 0.6, 0.8, 1 mag) des magnitudes absolues spectroscopiques*

provienne de la magnitude apparente (photoélectrique) ou de l'absorption (faible jusqu'à 50 pc) que l'on utilise.

Sans rentrer dans le détail des calibrations des magnitudes absolues, nous pouvons nous demander s'il n'existe pas alors une erreur systématique sur les magnitudes absolues, qui pourrait expliquer à la fois les biais mis en évidence et la corrélation mentionnée ci-dessus.

A titre d'exemple, prenons la calibration obtenue par Crawford (1975) pour les étoiles F grâce à la photométrie de Strömgren. Crawford utilise les parallaxes trigonométriques les plus précises [Wooley *et al.*, 1970] pour fixer le point-zéro des magnitudes absolues. Sur les 43 étoiles utilisées, 18 ont une mesure préliminaire de la parallaxe par NDAC. Pour ces étoiles, la différence moyenne entre la magnitude absolue calculée à partir de la parallaxe NDAC et la magnitude absolue utilisée par Crawford n'est que de -0.03 ± 0.06 mag. Il n'y a donc pas d'évidence d'un décalage des magnitudes absolues.

Il n'est pas possible de pousser plus loin la recherche sans déborder du cadre général qui nous concerne. Notons simplement que sur ce même échantillon d'étoiles, on peut obtenir une estimation de l'erreur moyenne et de la dispersion des magnitudes absolues :

	médiane		dispersion	
$M_{uvby-\beta} - M_{\rm NDAC}$	0.02	± 0.03	0.38	± 0.03
$M_{\rm spectro} - M_{\rm NDAC}$	-0.13	± 0.03	0.55	± 0.04

On notera que le décalage des magnitudes absolues spectroscopiques est dans le sens du biais de Malmquist et sans doute dû à ce biais. Les dispersions sont, quant à elles, à peine plus grande que ce que l'on pourrait attendre.

Nous ne nous attarderons pas sur ce point, notamment parce qu'il touche à la calibration des magnitudes absolues, mais également et surtout parce que le biais des figures 6.24 et 6.25 qui nous préoccupe provient en réalité d'une cause purement statistique que l'on va maintenant expliciter.

Calcul du biais

Comment déterminer analytiquement la grandeur du biais observé ? L'approche initiale est dûe à L. Lindegren (1992c) ; nous allons l'écrire de façon bayésienne.

Si $\pi_{\rm H}$ est la parallaxe trigonométrique Hipparcos, ce que l'on doit calculer est l'espérance de $(\pi_{\rm H} - \pi_{\rm P})$ sachant la parallaxe $\pi_{\rm P}$ (désignant ici indifféremment la parallaxe photométrique ou la parallaxe spectroscopique).

$$\begin{aligned} E[\pi_{\rm H} - \pi_{\rm P}|\pi_{\rm P}] &= E[\pi_{\rm H} - \Pi|\pi_{\rm P}] + E[\pi|\pi_{\rm P}] - E[\pi_{\rm P}|\pi_{\rm P}] \\ &= 0 + E[\pi|\pi_{\rm P}] - \pi_{\rm P} \end{aligned} \tag{6.6}$$

où π est la vraie parallaxe de l'étoile. Le premier terme du membre de droite est nul sous l'hypothèse que les erreurs sur la parallaxe Hipparcos sont indépendantes de la parallaxe et d'espérance nulle. Si elle n'est pas nulle (point-zéro), alors on pourra la mettre en évidence par un décalage de $(\pi_{\rm H} - \pi_{\rm P}) - E[\pi_{\rm H} - \pi_{\rm P}|\pi_{\rm P}]$.

Or la densité de probabilité *a posteriori* $f(\pi|\pi_{\rm P})$ s'écrit par la formules de Bayes :

$$f(\pi|\pi_{\rm P}) = \frac{f(\pi_{\rm P}|\pi) f(\pi)}{\int f(\pi_{\rm P}|\pi) f(\pi) d\pi}$$

et par définition de l'espérance mathématique

$$\begin{aligned} E[\pi|\pi_{\rm P}] &= \int f(\pi|\pi_{\rm P})\pi d\pi \\ &= \frac{\int f(\pi_{\rm P}|\pi)f(\pi)\pi d\pi}{\int f(\pi_{\rm P}|\pi)f(\pi)d\pi} \end{aligned} \tag{6.7}$$

On note que $\widehat{\pi} = E[\pi|\pi_{\rm P}]$ est l'estimateur bayésien ponctuel de la vraie parallaxe sachant la parallaxe $\pi_{\rm P}$.

Si l'on fait maintenant l'hypothèse que la distribution de la magnitude absolue utilisée M_P autour de la vraie magnitude absolue $M = m - A_V + 5\log\pi + 5$ de l'étoile est gaussienne, on obtient :

$$\begin{aligned} f(M_P|M) &= \frac{1}{\sigma_M\sqrt{2\Pi}} e^{-\frac{1}{2}\frac{(M_P-M)^2}{\sigma_M^2}} \\ f(\pi_{\rm P}|\pi) &= f(M_P(\pi_{\rm P})|\pi)|\frac{\partial M_P}{\partial \pi_{\rm P}}| \\ &= ke^{-\frac{1}{2}\frac{25(\log\pi_{\rm P}-\log\pi)^2}{\sigma_M^2}} \end{aligned} \tag{6.8}$$

où k désigne un terme indépendant de π, que l'on ignore parce qu'il se simplifie dans l'équation 6.7.

Pour calculer $E[\pi_{\rm H} - \pi_{\rm P}|\pi_{\rm P}]$, il reste donc à prendre une densité *a priori* $f(\pi)$. En utilisant une loi beta, L. Lindegren (1992c) obtenait une estimation du biais (fig. 6.29). Nous utiliserons pour notre part une loi *a priori* à l'aide de données externes.

Distribution des vraies parallaxes

Même si la vraie parallaxe reste inconnue, on a une certaine idée *a priori* de la distribution des parallaxes. Dans le cas où l'on aurait une population infinie d'étoiles réparties uniformément, $f(\pi)$ serait proportionnel à $\frac{1}{\pi^4}$; la distribution serait également calculable dans le cas d'un échantillon limité en magnitude apparente.

Mais dans la pratique, compte-tenu de la composition du Catalogue d'Entrée d'Hipparcos, avec notamment la présence du Catalogue de Gliese (1969), d'un survey avec une limite variable en magnitude, et d'étoiles des nuages de Magellan, etc, une distribution théorique des parallaxes est tout à fait compliquée à imaginer. Pour illustrer cette dernière remarque, on a tracé, figures 6.30 et 6.31, la densité lissée des parallaxes spectroscopiques et photométriques, qu'on pense être assez proche des vraies densités pour les plus petites parallaxes. Naturellement, ces deux densités ne sont pas supposées être les mêmes puisque l'échantillon d'étoiles ayant un type spectral n'est pas le même que l'échantillon des étoiles ayant de la photométrie $uvby{-}\beta$. Il n'en reste pas moins que les irrégularités de ces densités montrent bien qu'il est difficile de remplacer la vraie distribution des parallaxes par une distribution théorique.

Il semble donc plus réaliste de se servir de la distribution observée des parallaxes dans l'échantillon étudié. Pour cela, on va, non pas étudier la distribution en parallaxe, mais la distribution des modules de distance $m - M - A_V$ calculés avec les magnitudes absolues dont on dispose. Comme on peut faire l'hypothèse que la magnitude absolue et la magnitude apparente, corrigée de l'absorption interstellaire, ont une distribution approximativement gaussienne autour de leur vraie valeur, on peut considérer également que les modules de distance ont une distribution gaussienne. On peut donc obtenir par la méthode décrite au §4.3.4 une estimation continue de la densité des modules de distance observés.

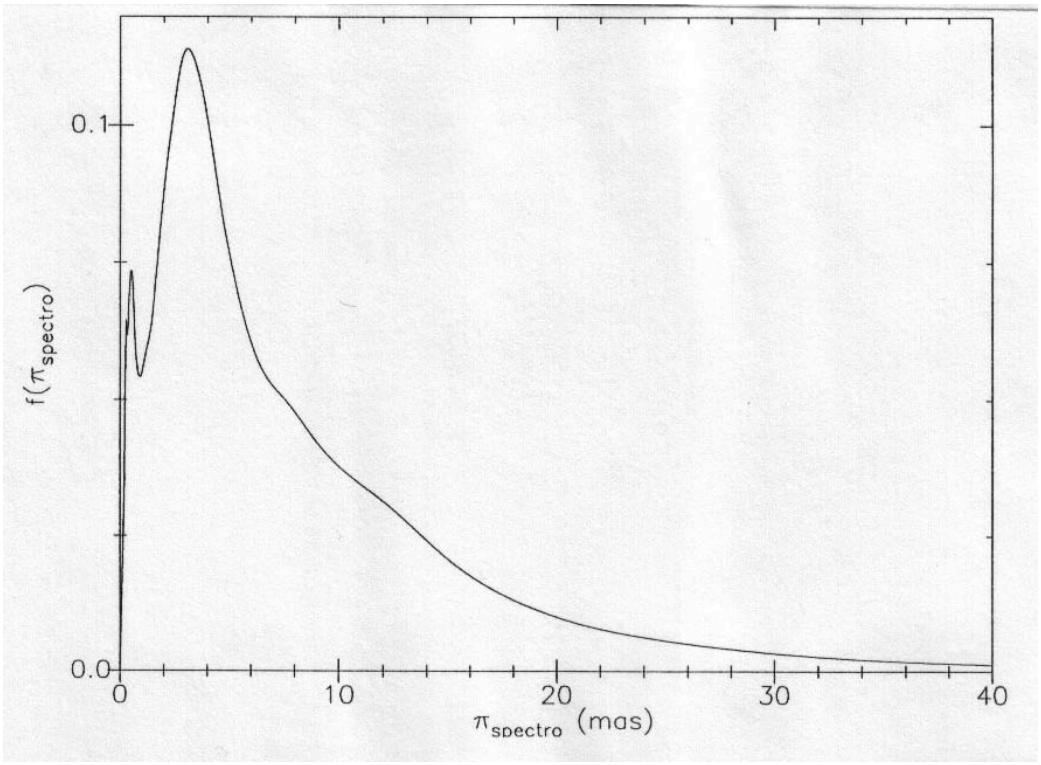


FIG. 6.30: *Densité lissée des parallaxes spectroscopiques pour les étoiles ayant une parallaxe FAST-3P*

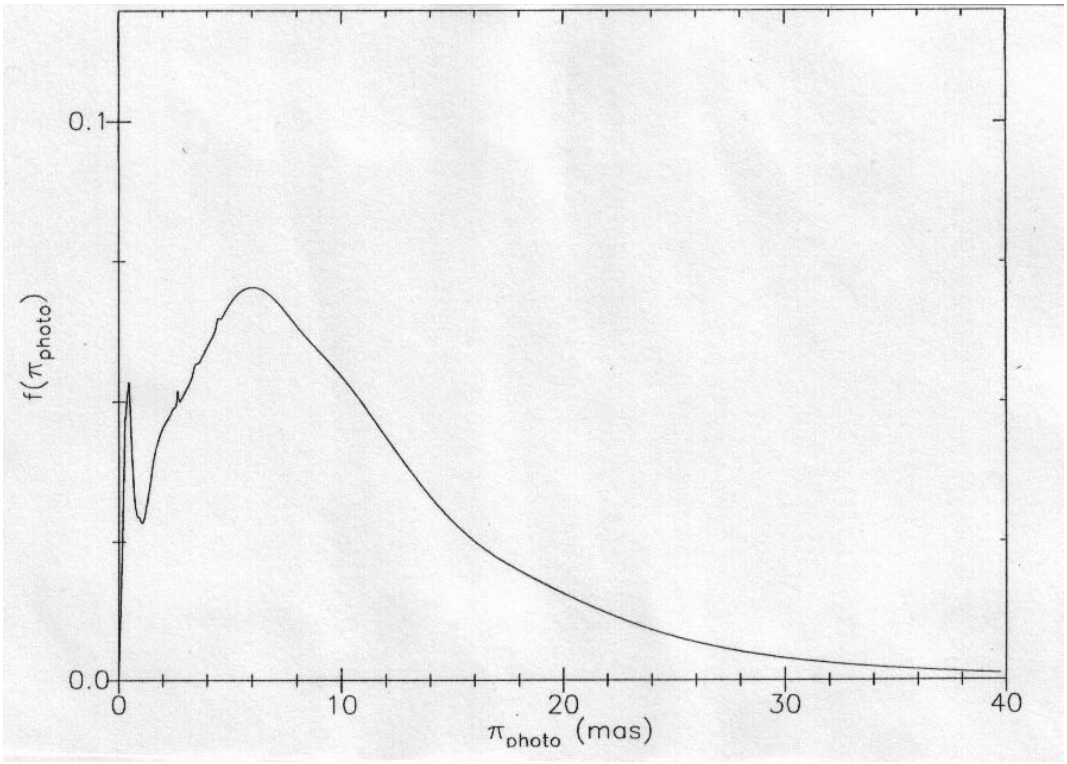


FIG. 6.31: *Densité lissée des parallaxes photométriques pour les étoiles ayant une parallaxe FAST-3P*

On prend maintenant par essais successifs une densité *a priori* voisine de cette densité observée et on la convolue avec les erreurs jusqu'à ce qu'on obtienne une densité marginale proche de la densité observée.

On peut maintenant refaire les graphiques initiaux (fig. 6.24 et 6.25), mais en corrigeant l'ordonnée de chaque étoile par le biais théorique calculé ci-dessus (fig. 6.32 et 6.33). On constate bien que les différences moyennes $(\pi_{\rm H} - \pi_{\rm P})$ en fonction de $\pi_{\rm P}$ ne sont plus vraiment différentes de 0, bien que légèrement négatives. Par contre les différences moyennes $(\pi_{\rm H} - \pi_{\rm S})$ continuent à exhiber un biais très négatif, preuve sans doute que l'on n'a pas réussi soit à prendre une densité *a priori* satisfaisante, soit, plus probablement, à estimer correctement les dispersions des magnitudes absolues spectroscopiques ; mais il faudrait multiplier par 2 ces dispersions pour réduire le biais, ce qui semble étonnant.

On peut noter également que si l'on n'avait pas corrigé du biais de Malmquist les parallaxes spectroscopiques, le biais serait encore plus grand ; cette correction n'est peut-être pas non plus adéquate. Encore une fois, si l'on voulait en savoir plus, il faudrait se pencher sur les calibrations des magnitudes absolues moyennes, ce que nous ne pouvons pas effectuer pour l'instant.

Il est néanmoins possible d'obtenir l'indication qu'il n'y a pas de variation des erreurs moyennes sur les parallaxes préliminaires avec la parallaxe en utilisant l'estimation conditionnelle de la parallaxe trigonométrique que l'on a déjà montrée (eq. 4.6). Pour cela, notant que $(\pi_{\rm N} - \pi) - E[\pi_{\rm N} - \pi|\pi_{\rm N}] = E[\pi|\pi_{\rm N}] - \pi = \pi_{\rm N} + \sigma^2_{\pi_{\rm N}} \frac{f'(\pi_{\rm N})}{f(\pi_{\rm N})} - \pi$, on trace les variations de $(E[\pi|\pi_{\rm N}] - \pi_{\rm P})$ (fig. 6.34) et $(E[\pi|\pi_{\rm N}] - \pi_{\rm S})$ (fig. 6.35) en fonction de $\pi_{\rm N}$. On constate que le biais est fortement réduit, mais qu'il reste des effets vers $\pi \approx 10$ mas. Ceci indique également que la correction du biais statistique des parallaxes est beaucoup plus facile avec les parallaxes Hipparcos qu'avec les parallaxes spectroscopiques ou photométriques, et c'est sans doute dû à leur qualité (erreurs gaussiennes, peu de points aberrants).

À la suite des études précédentes, il apparaît clair que si les parallaxes spectroscopiques et photométriques doivent être utilisées pour déterminer le point-zéro des parallaxes Hipparcos, alors il faudra se plonger attentivement sur l'étude des dispersions des magnitudes absolues pour chaque type d'étoile, quitte à utiliser les parallaxes Hipparcos pour les déterminer.

En résumé, les principaux résultats de ce paragraphe sont :

- Il n'y a pas de variation sensible des différences systématiques sur la parallaxe préliminaire Hipparcos avec la parallaxe elle-même ;
- il n'y a pas d'évidence statistique d'un décalage global sensible des magnitudes absolues photométriques ou spectroscopiques, tout au moins sur les étoiles plus près que 50 pc ;
- pour ces étoiles, provenant essentiellement du bas de la séquence principale, la dispersion des magnitudes absolues est d'environ 0.35 mag pour celles provenant de la photométrie de Strömgren et de 0.5 mag pour les magnitudes absolues spectroscopiques ;
- une corrélation peut être mise en évidence entre les erreurs sur les magnitudes absolues photométriques ou spectroscopiques dont l'origine reste à déterminer.

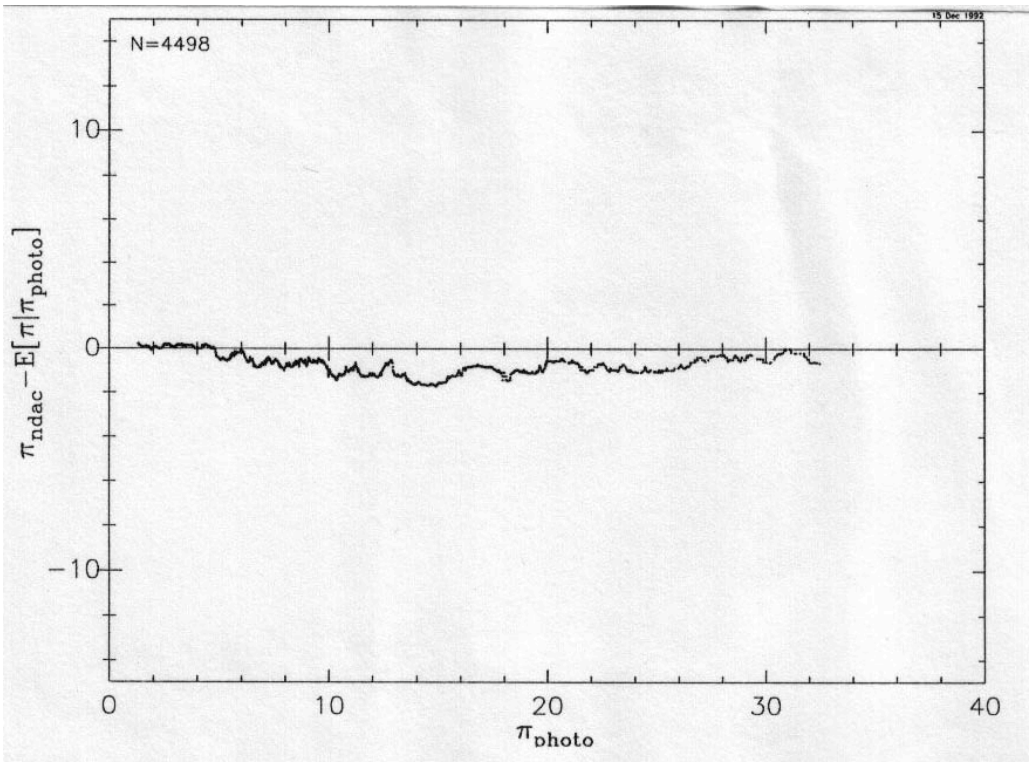


FIG. 6.32: *Lissage (401 points) des différences* $\pi_{\mathrm{NDAC}} - E[\pi|\pi_{\mathrm{P}}]$ *en fonction de* π_{P}

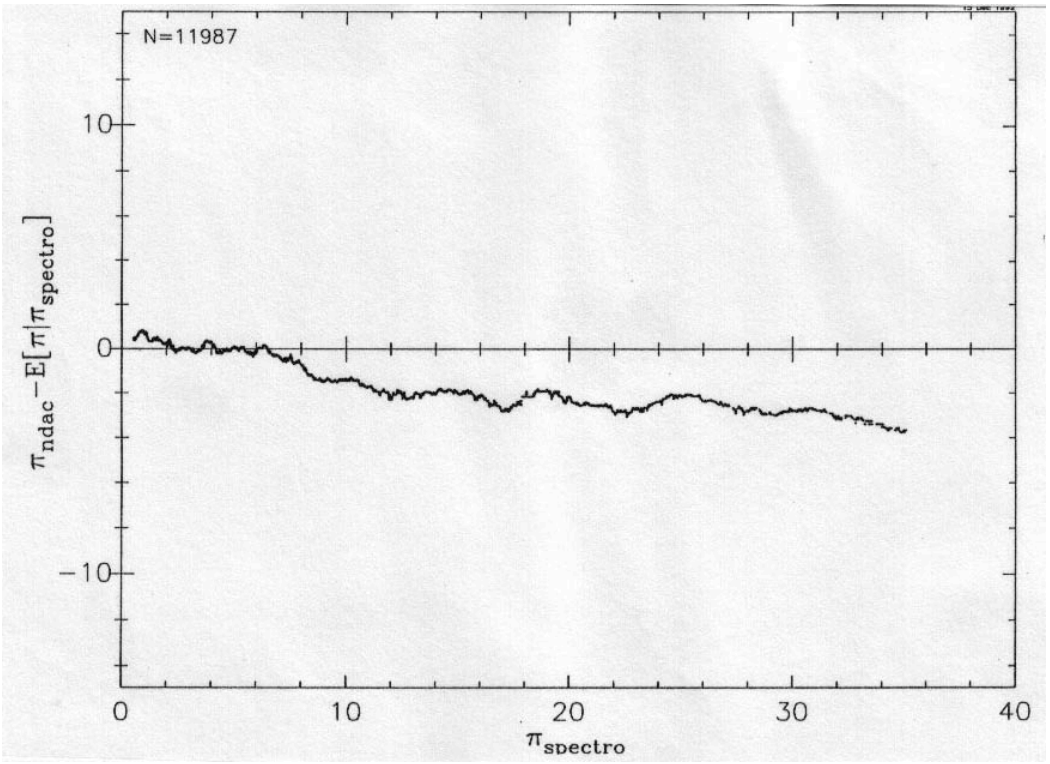


FIG. 6.33: *Lissage (401 points) des différences* $\pi_{\mathrm{NDAC}} - E[\pi|\pi_{\mathrm{S}}]$ *en fonction de* π_{S}

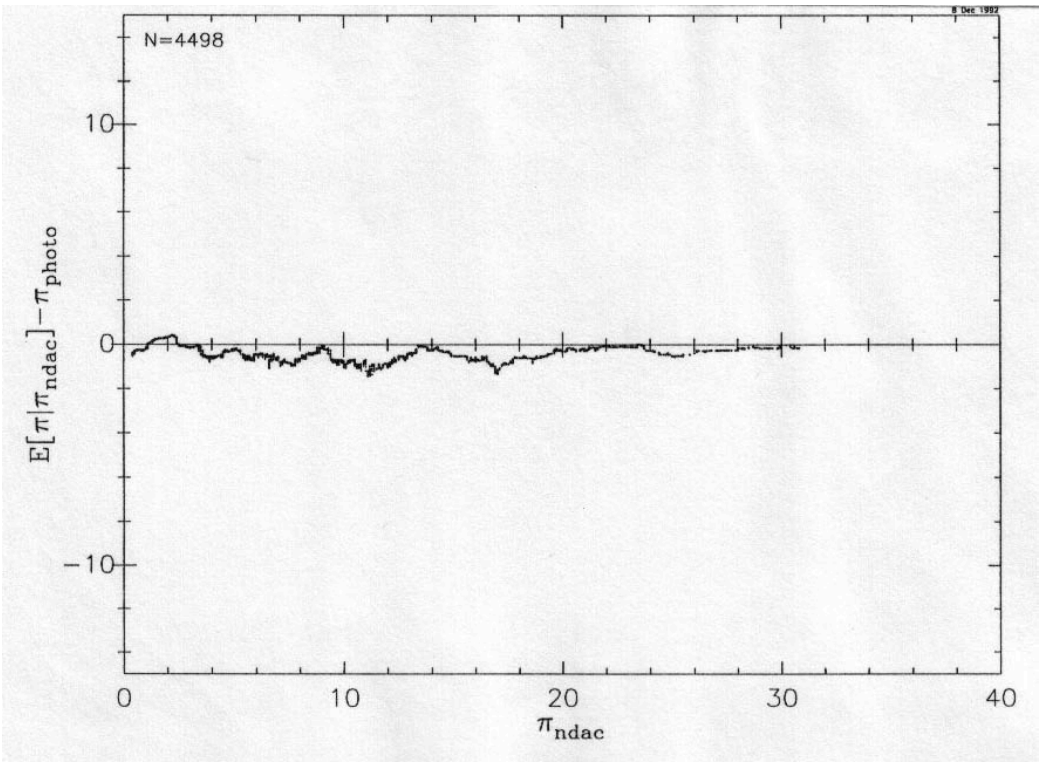


FIG. 6.34: *Lissage (801 points) des différences* $E[\pi|\pi_{\mathrm{NDAC}}] - \pi_{\mathrm{P}}$ *en fonction de* π_{NDAC}

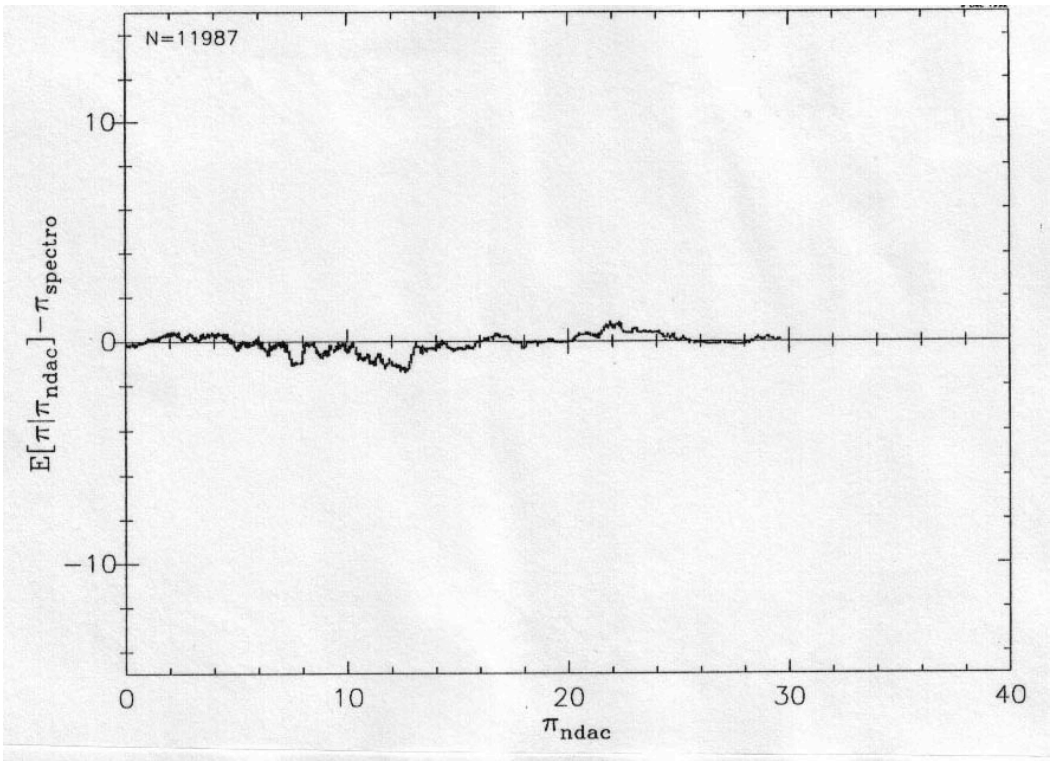


FIG. 6.35: *Lissage (801 points) des différences* $E[\pi|\pi_{\mathrm{NDAC}}] - \pi_{\mathrm{S}}$ *en fonction de* π_{NDAC}

6.5.2 Variations avec les données astrométriques et photométriques

Nous abordons ici les comparaisons les plus intéressantes pour voir de façon fine la qualité des parallaxes préliminaires Hipparcos. Nous avons déjà précisé la finalité de ces comparaisons : montrer que les erreurs systématiques sur la parallaxe Hipparcos sont indépendantes des autres paramètres déterminés à l'issue de la réduction, et, par conséquent, nous analysons les variations des différences systématiques en fonction des positions, mouvements propres, magnitude et couleur, l'ensemble de ces données provenant, bien évidemment, d'une source externe, le Catalogue d'Entrée, c'est-à-dire toutes les données acquises depuis le sol.

Pour ce faire, nous utiliserons les parallaxes spectroscopiques, parce qu'elles seules fournissent suffisamment d'étoiles ; les mêmes comparaisons ont été faites avec les parallaxes photométriques et les parallaxes d'étoiles d'amas, mais il n'est souvent pas possible de mettre en évidence une éventuelle variation (plus précisément, de montrer que les différences systématiques sont significativement différentes de 0) à cause du faible nombre d'étoiles. Nous effectuons donc les comparaisons avec les parallaxes spectroscopiques plus petites que 2 mas.

Avec cette valeur, nous sommes assuré que les erreurs aléatoires sur les parallaxes spectroscopiques sont négligeables. Pour fixer les idées, notons qu'avec une erreur sur la magnitude absolue de 0.3 mag, pour une étoile à 500 pc (2 mas) l'erreur sur la parallaxe spectroscopique (ou photométrique) n'est que 0.3 mas, soit 8 fois plus petite environ que l'erreur sur la parallaxe trigonométrique Hipparcos.

Dû à la limite sur la parallaxe observée (2 mas), on s'attend néanmoins à un biais qui doit rendre légèrement positive (≈ 0.4 mas) la différence $\langle \pi_{\rm H} - \pi_{\rm S} \rangle$. C'est pour cette raison que nous utilisons le terme 'différence systématique' au lieu de 'erreur systématique', parce que nous savons que le décalage observé ne provient pas seulement des parallaxes préliminaires, mais également d'un biais statistique. Mais ce décalage statistique est sans importance pour ce qui suit, dans la mesure où l'on va regarder les variations de cette différence, plus que sa valeur moyenne.

En fonction de la position

Commençant par observer les différences systématiques sur les parallaxes NDAC-5P en fonction des positions, ce qui frappe de prime d'abord à la vision de la figure 6.36, c'est la répartition non uniforme sur tout le ciel, surtout au niveau de l'écliptique, région qui apparaît à la fois dépeuplée et peu «stable». On pourrait penser que c'est dû à la loi de balayage du satellite, mais pour les parallaxes FAST (fig. 6.37d) c'est plutôt aux pôles que manquent les étoiles. On voit de plus nettement apparaître sur cette dernière figure une certaine oscillation autour de la moyenne. En longitude écliptique, c'est également le cas.

Il n'y a donc clairement pas indépendance entre les différences systématiques sur la parallaxe et la position. Pour s'en convaincre plus quantitativement, il n'y a pas besoin de développer un test d'indépendance compliqué : il suffit de trouver deux zones dont la différence des moyennes soit significativement différente de 0 ; c'est le cas, par exemple

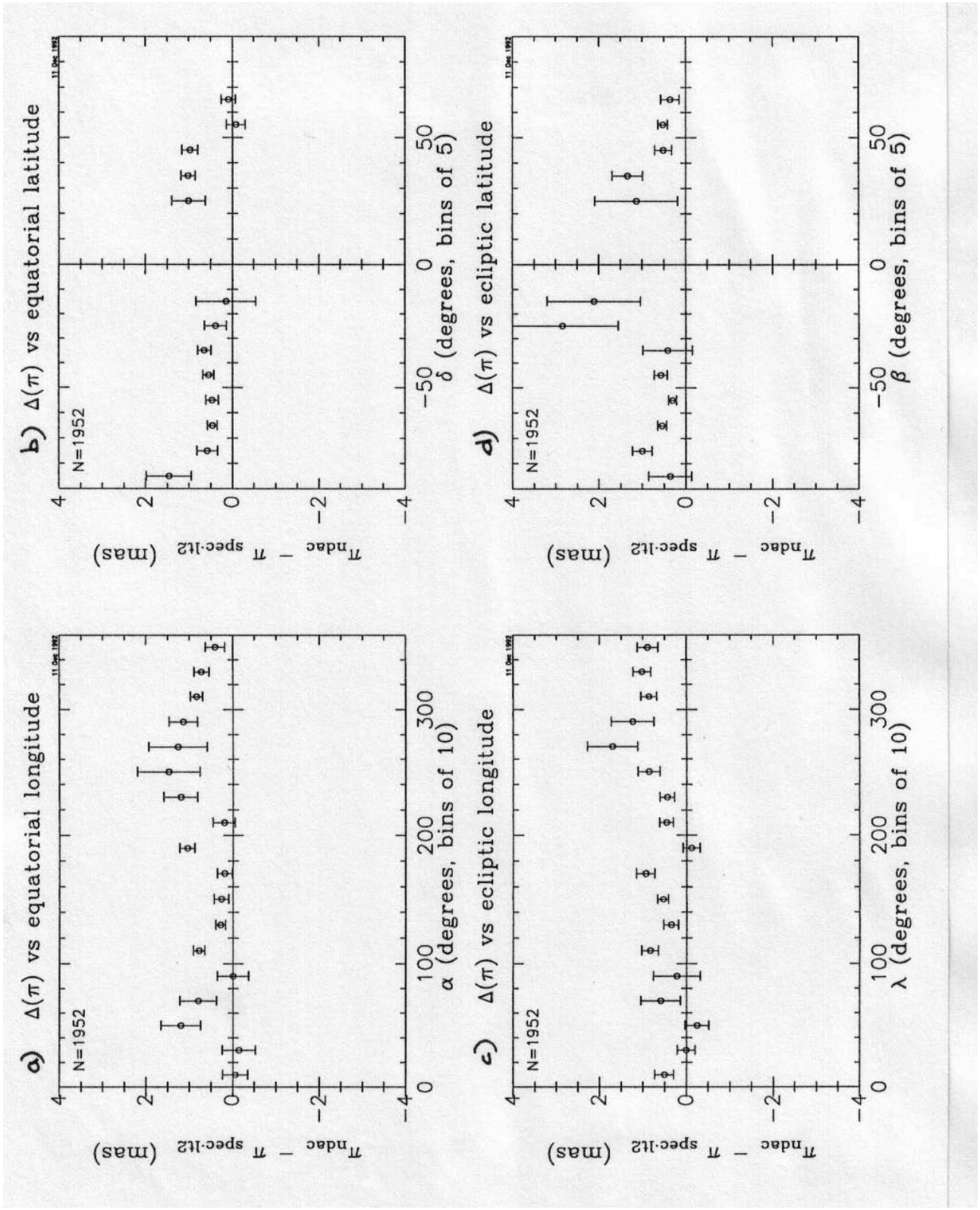


FIG. 6.36: *Erreur sur la parallaxe NDAC-5P (comparée à la parallaxe spectroscopique) en fonction des coordonnées équatoriales et écliptiques*

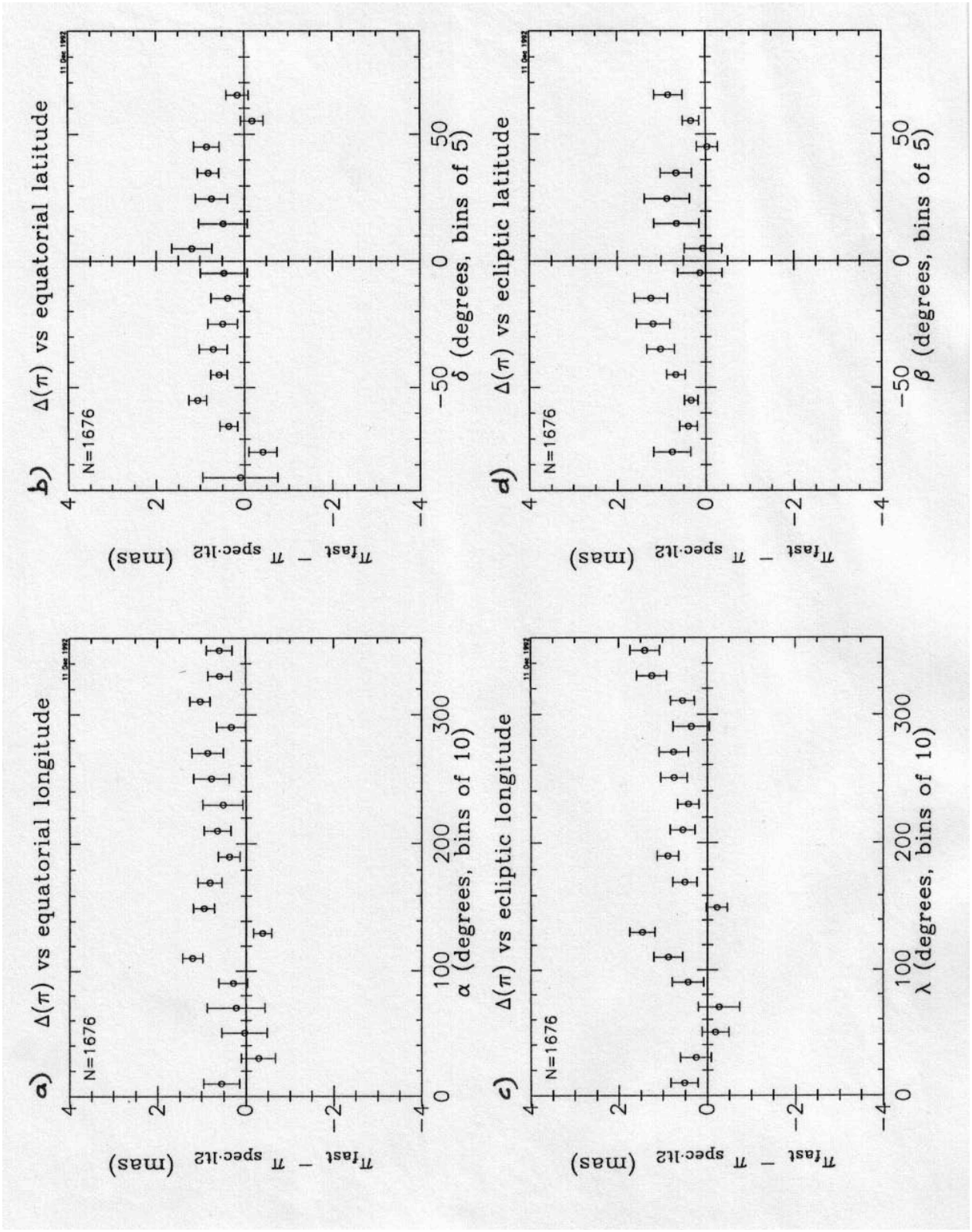


FIG. 6.37: *Erreur sur la parallaxe FAST-3P (comparée à la parallaxe spectroscopique) en fonction des coordonnées équatoriales et écliptiques*

sur la figure 6.36b, pour les zones $40° < \delta < 50°$ et $50° < \delta < 60°$ et sur la figure 6.37b, pour les zones $-60° < \delta < -50°$ et $50° < \delta < 60°$.

En fonction des mouvements propres

L'étude de la variation des différences systématiques sur la parallaxe en fonction du mouvement propre est intéressante à la fois pour la solution 5 paramètres et pour la solution 3 paramètres.

Pour la première, cela doit permettre de voir si une seule année de données suffisent à extraire convenablement les parallaxes, ou bien si la qualité probablement médiocre des mouvements propres déterminés ne va pas avoir tendance à les contaminer.

Pour la solution 3 paramètres, comme les mouvements propres proviennent du Catalogue d'Entrée, on devrait voir s'ils ont dégradé, ou non, les parallaxes. Et pour être sûr des conclusions que l'on pourrait en tirer, nous regarderons à la fois la solution 3 paramètres de FAST et la solution 3 paramètres de NDAC.

Les parties a et b des figures 6.38, 6.39 et 6.40 sont assez éloquentes. Pour les étoiles avec un mouvement propre proche de 0 (± 0.005"/an), c'est-à-dire une grande majorité des étoiles (on le voit à la grandeur des barres d'erreur), les différences systématiques sur la parallaxe sont voisines de leur moyenne (≈ 0.5 mas). Par contre, dès que le mouvement propre augmente en valeur absolue, on assiste à une variation très importante des différences systématiques. On peut remarquer que c'est essentiellement vrai pour les mouvements propres négatifs, et moins pour les mouvements propres positifs.

Tout d'abord, il serait légitime de se demander si ces variations ne proviennent pas tout simplement des parallaxes spectroscopiques que l'on utilise. En effet, nous avons pris comme étoiles de comparaison celles qui ont une parallaxe spectroscopique inférieure à 2 mas. Imaginons qu'une de ces étoiles ait été mal classée spectroscopiquement, et qu'en réalité elle soit proche, donc sans doute avec un mouvement propre non négligeable. Dans ce cas, la différence $\pi_{\rm H} - \pi_{\rm S}$ devrait être nettement positive. Dans la mesure où on a éliminé lors de ces comparaisons les étoiles avec une différence supérieure en valeur absolue à 3 fois l'erreur formelle, on a sans doute supprimé une partie des étoiles mal classées. De plus, on constate sur les graphes que l'on a également des différences négatives pour des «grands» mouvements propres. On supposera donc que les effets constatés proviennent des parallaxes Hipparcos et non des parallaxes spectroscopiques.

Regardons tout d'abord la solution 5 paramètres (fig. 6.38). On peut noter une forme parabolique des différences systématiques, qui pourrait s'expliquer ainsi : les plus grands mouvements propres n'étant pas bien déterminés au bout d'un an, ils ont eu tendance à augmenter la parallaxe Hipparcos, et donc à créer une différence $\pi_{\rm H} - \pi_{\rm S}$ positive. Pour montrer que cette variation est significative, on calcule le τ de Kendall entre la distribution des erreurs et celle de la valeur absolue des mouvements propres, et ce test suggère qu'il n'y a pas indépendance. Si l'on utilise la valeur absolue, c'est naturellement pour mettre en lumière le fait que l'erreur sur la parallaxe augmente lorsque le mouvement propre augmente en module.

Pour le mouvement propre en ascension droite, les deux solutions 3 paramètres sont semblables entre elles, et différentes de la solution 5 paramètres, dans la mesure où les différences systématiques sont essentiellement négatives pour les mouvements propres compris entre -30 et -20 mas/an. En déclinaison, on voit à un moindre degré des similitudes entre les deux solutions 3 paramètres, les différences systématiques étant ici négatives pour les

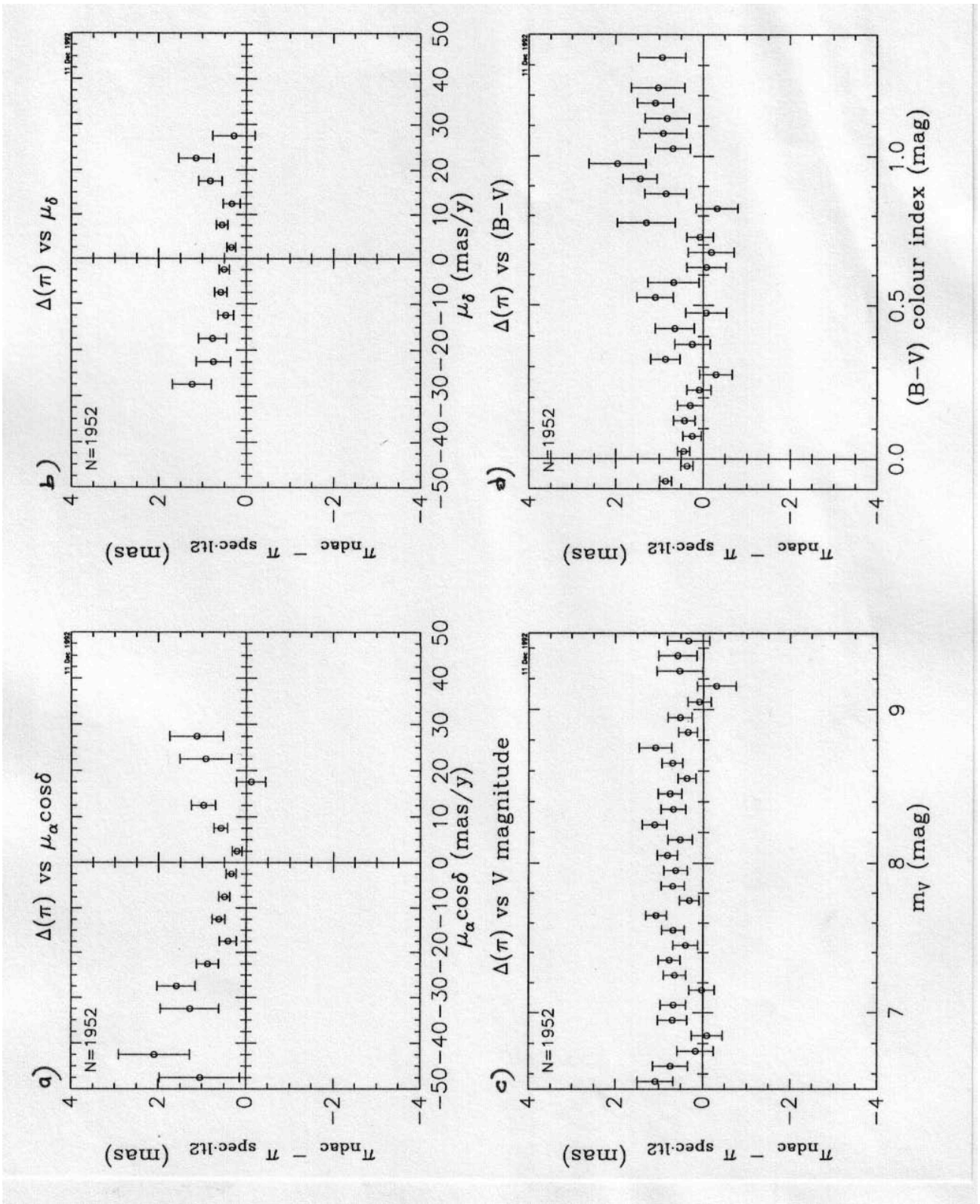


FIG. 6.38: *Erreur sur la parallaxe NDAC-5P (comparée à la parallaxe spectroscopique) en fonction des mouvements propres, de la magnitude V et de l'indice de couleur $(B-V)$*

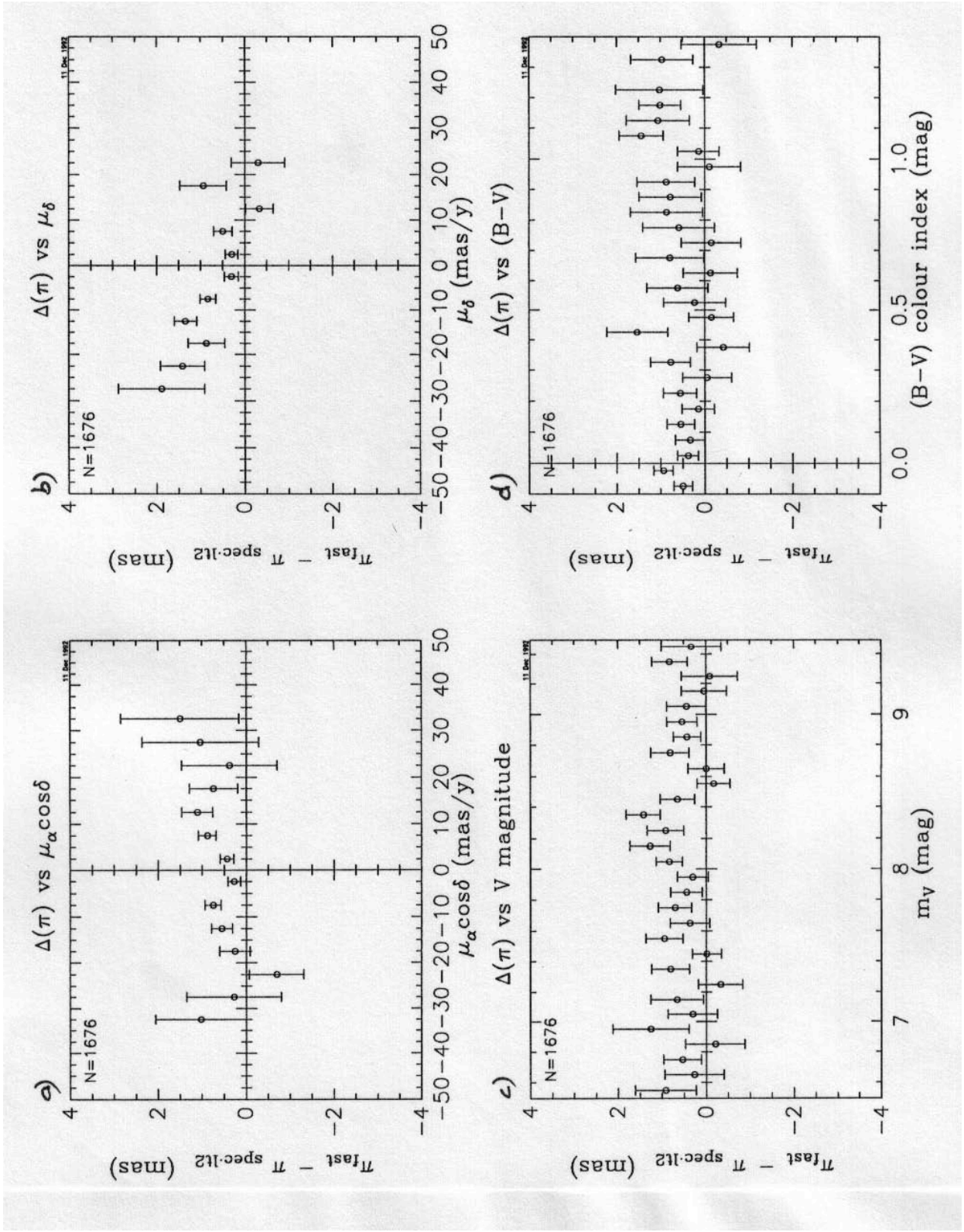


FIG. 6.39: *Erreur sur la parallaxe FAST-3P (comparée à la parallaxe spectroscopique) en fonction des mouvements propres, de la magnitude V et de l'indice de couleur $(B-V)$*

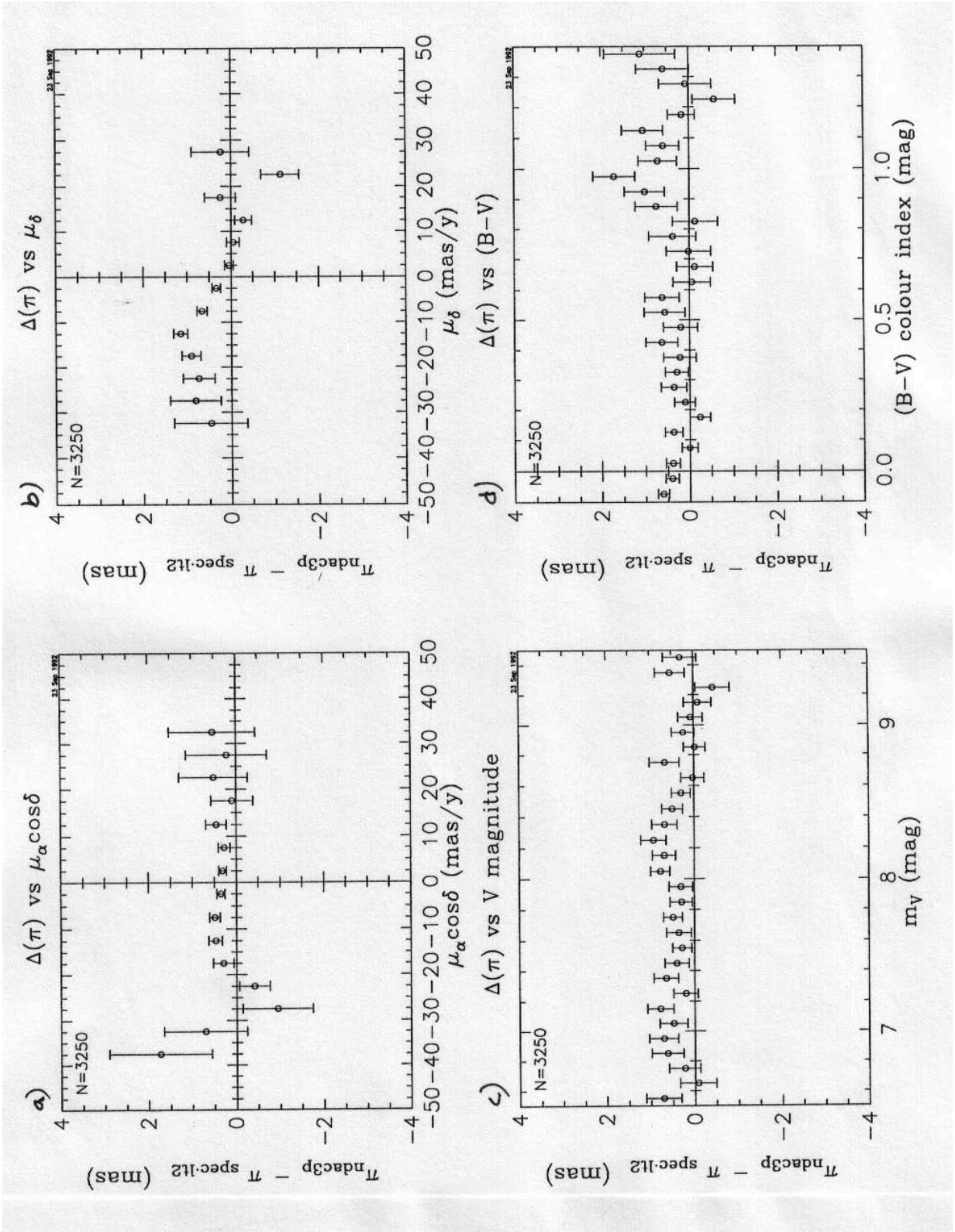


FIG. 6.40: *Erreur sur la parallaxe NDAC-3P (comparée à la parallaxe spectroscopique) en fonction des mouvements propres, de la magnitude V et de l'indice de couleur $(B-V)$*

mouvements propres positifs.

On constate donc une dépendance entre le mouvement propre et les différences systématiques sur la parallaxe ; par exemple, pour $-15 < \mu_\delta < -10$ mas/an, les différences systématiques sur les parallaxes FAST-3P et NDAC-3P sont significativement différentes de la moyenne.

En fonction de la magnitude et de la couleur

Enfin, une étude des variations des différences systématiques en fonction de la magnitude V et de la couleur $B-V$ s'imposait. En effet, la magnitude H utilisée dans le champ principal du satellite dépend de V et de $B-V$. On pourrait craindre une mauvaise affectation du temps d'observation, ou un effet chromatique qui se traduirait inévitablement par une variation des erreurs systématiques sur les paramètres astrométriques en fonction de V ou de $B-V$.

Ici encore, certains problèmes peuvent être mis en évidence, peut-être liés à la réduction de chaque consortium. Les deux solutions NDAC montrent en effet des variations semblables, avec la magnitude et surtout avec l'indice de couleur ; pour les étoiles les plus rouges, les différences systématiques deviennent extrêmement dispersées et certains points ($B-V \approx 0.9$) atteignent presque 2 mas. Pour la solution FAST-3P, il est plus difficile de conclure, dans la mesure où les barres d'erreur sont plus grandes et les points apparaissent d'ailleurs plus dispersés, y compris en magnitude. La taille de ces barres d'erreur ne nous permet pas de mettre de façon significative en évidence la dépendance entre les différences systématiques sur la parallaxe préliminaire et les magnitudes, quoique cette dépendance existe probablement.

Bilan des comparaisons

Comme on a pu le mettre en évidence ci-dessus, il existe une certaine dépendance entre les différences systématiques sur la parallaxe et les autres données, astrométriques et photométriques, bien que le niveau de cette dépendance ne puisse pas être considéré comme dramatique pour des parallaxes obtenues après seulement un an de données pour les premières solutions sur la sphère.

Beaucoup d'autres comparaisons mériteraient d'être également faites. D'une part, prendre les mouvements propres les plus précis du Catalogue d'Entrée, et comparer leur différence d'avec les mouvements propres Hipparcos en fonction des positions, magnitudes, et couleurs, comme ci-dessus ; d'autre part étudier la variation des erreurs sur la parallaxe Hipparcos en fonction des erreurs sur le mouvement propre Hipparcos. Malheureusement, nous ne disposons pas des mouvements propres obtenus par le consortium NDAC pour la solution 5 paramètres. Ce que l'on devrait mettre en évidence, c'est une éventuelle contamination entre les mouvements propres et la parallaxe, ce qui serait sans surprise pour une solution 5 paramètres obtenue avec seulement un an de données.

Nous espérons que cette comparaison pourra se faire avec un an et demi de données. Au fur et à mesure de l'accumulation des données, qui conduisent à des solutions de plus en plus précises, ce sont sans aucun doute les erreurs des données au sol qui seront mises en évidence, bien plus que les erreurs sur les solutions préliminaires Avec plus d'étoiles, on pourra également utiliser d'autres données (photométriques) qui permettront de lever

l'ambiguïté sur la responsabilité des dépendances exhibées (on ne peut actuellement pas distinguer les contributions respectives des parallaxes préliminaires et spectroscopiques).

6.6 Point-zéro des parallaxes préliminaires

Bien que ce ne soit pas encore vraiment le cas avec les parallaxes obtenues après un an de données, nous allons supposer que les erreurs systématiques sur la parallaxe Hipparcos sont indépendantes des caractéristiques des étoiles (position, vitesse, luminosité, température), et dépendantes uniquement de caractéristiques instrumentales du satellite, et que l'on peut donc tenter de déterminer le point-zéro des parallaxes préliminaires à l'aide des différences systématiques.

6.6.1 Estimation directe

Comme les erreurs systématiques sur la parallaxe Hipparcos sont indépendantes de la distance, il suffit donc d'étudier la distribution des parallaxes pour les étoiles les plus lointaines. Il faut également que ces étoiles soient assez nombreuses pour avoir la précision nécessaire sur le point-zéro (voir §6.1).

Le dilemme est là : si l'on utilise les étoiles des nuages de Magellan, la précision est insuffisante à cause du faible nombre d'étoiles. On a beaucoup plus d'étoiles d'amas, mais les étoiles non-membres risquent de perturber la détermination de la moyenne.

Restent les parallaxes spectroscopiques et photométriques qui sont assez nombreuses, mais qui sont biaisées si l'on prend les plus lointaines, parce que l'on aura pris une limite supérieure sur la parallaxe observée.

Pour supprimer ce biais, ou tout au moins le diminuer, nous remplaçons la parallaxe spectroscopique ou photométrique par l'estimateur bayésien de la vraie parallaxe sachant la parallaxe spectroscopique $E[\pi|\pi_{\rm S}]$ ou sachant la parallaxe photométrique $E[\pi|\pi_{\rm P}]$ (§6.5.1).

Après les avoir corrigées de ce biais par l'estimateur bayésien, ces parallaxes fournissent le contingent le plus important pour effectuer une dernière vérification : le point-zéro (défini ici comme la différence entre la parallaxe préliminaire et de l'estimation bayésienne de la vraie parallaxe) doit être indépendant :

- de la limite choisie en parallaxe spectroscopique (ici 2 mas), sinon cela mettrait en évidence un biais sur la parallaxe spectroscopique : soit le biais statistique dû à la distribution des parallaxes (censé être corrigé), soit un biais dans la calibration des magnitudes absolues spectroscopiques.

- des erreurs formelles sur la parallaxe. Dans le cas contraire, cela ne permettrait plus de prendre l'ensemble des parallaxes pour calculer la moyenne de cette distribution.

En ce qui concerne le premier point, on va regarder le comportement des différences $z_N = \langle \pi_{\rm N} - E[\pi|\pi_{\rm S}] \rangle$ pour les étoiles avec $\pi_{\rm S} < 2$ mas. Pour cela, on va «découper» la distribution de ces parallaxes en dix déciles (tableau 6.10, colonnes 1 à 3).

Pour chacun, on va calculer le point-zéro z_N et le rapport moyen $k_N = \langle \frac{(\pi_{\rm N} - E[\pi|\pi_{\rm S}])}{s_{\rm N}} \rangle$ des erreurs externes sur les erreurs internes ; dans chaque décile (195 étoiles), l'erreur

TAB. 6.10: *Variation de z_N et k_N avec la parallaxe spectroscopique.*

Estimation du point-zéro et du rapport moyen des erreurs externes sur les erreurs internes pour chaque décile de la distribution des parallaxes spectroscopiques inférieures à 2 mas ; 1) à partir de la différence entre la parallaxe NDAC et l'estimation bayésienne sachant la parallaxe spectroscopique ; 2) à partir de la différence entre la parallaxe NDAC et la parallaxe photométrique.

$(\pi_N - E[\pi\|\pi_S]), \pi_S < 2$			$(\pi_N - \pi_P), \pi_S < 2$		
π_S	z_N	k_N	π_S	z_N	k_N
$[0.00, 0.36[$	0.25	0.92	$[0.00, 0.33[$	0.23	1.03
$[0.36, 0.52[$	0.33	1.03	$[0.33, 0.50[$	0.11	1.19
$[0.52, 0.72[$	0.70	1.17	$[0.50, 0.72[$	0.68	0.89
$[0.72, 0.97[$	0.52	1.07	$[0.72, 0.93[$	-0.35	0.98
$[0.97, 1.19[$	0.43	1.04	$[0.93, 1.11[$	0.33	0.98
$[1.19, 1.40[$	0.42	1.04	$[1.11, 1.31[$	0.61	0.91
$[1.40, 1.58[$	0.28	1.15	$[1.31, 1.49[$	-0.11	1.25
$[1.58, 1.73[$	0.09	1.16	$[1.49, 1.64[$	0.29	0.98
$[1.73, 1.89[$	0.01	1.03	$[1.64, 1.83[$	-0.02	1.32
$[1.89, 2.00[$	0.55	1.10	$[1.83, 1.99]$	0.46	0.73

TAB. 6.11: *Variation de z_N et k_N avec les erreurs internes.*

Estimation du point-zéro et du rapport moyen des erreurs externes sur les erreurs internes pour chaque décile de la distribution des erreurs formelles sur la parallaxe NDAC-5P ; 1) à partir de la différence entre la parallaxe NDAC et l'estimation bayésienne sachant la parallaxe spectroscopique ; 2) à partir de la différence entre la parallaxe NDAC et la parallaxe photométrique pour les parallaxes spectroscopiques inférieures à 2 mas.

$(\pi_N - E[\pi\|\pi_S]), \pi_S < 2$			$(\pi_N - \pi_P), \pi_S < 2$		
s_N	z_N	k_N	s_N	z_N	k_N
$[0.0, 1.5]$	0.37	1.07	$[0.0, 1.4]$	0.39	0.83
$[1.5, 1.6]$	0.05	1.08	$[1.4, 1.5]$	-0.19	1.34
$[1.6, 1.7]$	0.43	0.96	$[1.5, 1.6]$	0.17	0.89
$[1.7, 1.8]$	0.44	0.93	$[1.6, 1.7]$	0.01	1.21
$[1.8, 1.9]$	0.61	0.92	$[1.7, 1.8]$	0.54	0.83
$[1.9, 2.1]$	0.44	1.05	$[1.8, 1.9]$	0.34	0.97
$[2.1, 2.3]$	0.50	1.06	$[1.9, 2.1]$	0.53	0.92
$[2.3, 2.5]$	-.10	0.97	$[2.1, 2.4]$	0.20	0.81
$[2.5, 2.9]$	0.82	1.03	$[2.4, 2.7]$	-0.01	0.87
$[2.9, 4.0]$	0.01	1.14	$[2.7, 3.9]$	0.20	1.23

standard sur z est d'environ 0.15 mas. On espère naturellement que z_N soit voisin de 0 et k_N soit proche de 1.

Mais on voit que l'on n'a pas réussi à supprimer complètement le biais car il y a une variation (croissance puis décroissance) de z_N qui a exactement la forme du biais théorique calculé eq. 6.7 (fig. 6.29).

Pour ce qui est du second point, on va procéder de manière analogue, mais avec la distribution des erreurs internes sur la parallaxe préliminaire (tableau 6.11, colonnes 1 à 3). Il y a ici encore une variation sensible du point-zéro (l'erreur standard sur z varie de 0.11 à 0.22 mas), mais la forme de la dépendance avec l'erreur formelle n'est pas évidente.

Comme on n'a pas réussi à éliminer complètement le biais, on pourrait donc penser que l'on ne pourra pas se servir des parallaxes spectroscopiques et photométriques pour estimer les deux paramètres z et k.

Il existe heureusement une solution à ce problème : se servir des parallaxes spectroscopiques lointaines pour calculer la différence entre la parallaxe préliminaire et la parallaxe photométrique, et réciproquement. Cette méthode n'est naturellement valable que s'il n'y a pas, pour les étoiles lointaines, de corrélation entre les erreurs sur les parallaxes spectroscopiques et les erreurs sur les parallaxes photométriques. Ce qui semble le cas à la vision du tableau 6.10 ($5^{\text{ème}}$ colonne), où z_N ne semble pas montrer de biais ($\sigma_z \approx 0.37$), bien qu'il reste quelques valeurs différentes de façon significative, probablement dues à des points aberrants.

L'inconvénient est que le nombre d'étoiles a largement diminué (≈ 300), puisque l'on doit disposer pour chaque étoile d'une parallaxe préliminaire, d'une parallaxe spectroscopique et d'une parallaxe photométrique. Du coup, les variations de z_N avec l'erreur interne ($5^{\text{ème}}$ colonne du tableau 6.11) ne sont plus significatives ($0.24 < \sigma_z < 0.56$).

Résumé des comparaisons

Concernant les estimations du point-zéro et des erreurs externes, effectuées en observant les deux premiers moments de la distribution des différences (Hipparcos-externe), nous pouvons maintenant conclure. Les tableaux 6.12 et 6.13 résument ainsi les estimations obtenues respectivement pour la parallaxe FAST-3P et la parallaxe NDAC-5P, à l'aide de toutes les comparaisons directes que nous avons effectuées jusqu'à présent avec des données externes.

Compte-tenu des différents problèmes que nous avons soulevés ci-dessus, on ne peut pas se baser sur ces résultats pour en tirer des conclusions hâtives, d'autant que les parallaxes sont préliminaires. On constate néanmoins que le point-zéro est voisin de 0, avec la plupart de ces estimations. Les erreurs externes sont également proches des erreurs internes dans la plupart des cas et celles de la solution 5 paramètres légèrement inférieures à celles de la solution 3 paramètres.

Nous allons maintenant voir une dernière méthode qui peut permettre de déterminer z et k, et qui n'utilise que les données observées et une loi *a priori*.

6.6.2 Estimation avec les fonctions de répartitions

La loi des erreurs sur les parallaxes spectroscopiques ou photométriques étant lognormale, on sait que, pour des étoiles lointaines, la distribution de ces parallaxes doit être

TAB. 6.12: *Point-zéro et erreur externe pour FAST-3P*

Estimation par différentes méthodes du point-zéro (mas), de l'erreur externe moyenne et du rapport moyen des erreurs externes sur les erreurs internes pour les parallaxes FAST-3P.

Méthode	nombre	z	largeur	k
parallaxes spectroscopiques ($\pi_{\rm P} < 2$)	266	0.22 ± 0.24	3.11 ± 0.26	1.46 ± 0.12
parallaxes photométriques ($\pi_{\rm S} < 2$)	312	0.01 ± 0.21	2.91 ± 0.22	1.38 ± 0.10
Distances d'amas	427	0.20 ± 0.24	3.94 ± 0.20	1.83 ± 0.07

TAB. 6.13: *Point-zéro et erreur externe pour NDAC-5P*

Estimation par différentes méthodes du point-zéro (mas), de l'erreur externe moyenne et du rapport moyen des erreurs externes sur les erreurs internes pour les parallaxes NDAC-5P.

Méthode	nombre	z	largeur	k
parallaxes spectroscopiques ($\pi_{\rm P} < 2$)	266	0.17 ± 0.15	1.94 ± 0.16	1.01 ± 0.08
parallaxes photométriques ($\pi_{\rm S} < 2$)	305	0.24 ± 0.14	1.97 ± 0.15	1.01 ± 0.08
Distances d'amas	437	0.27 ± 0.17	2.84 ± 0.15	1.33 ± 0.07
Nuages de Magellan	32	0.27 ± 0.48	2.74 ± 0.34	0.90 ± 0.11

proche de la distribution des vraies parallaxes.

Autrement dit, on peut se servir de la densité des parallaxes spectroscopiques ou photométriques comme estimation *a priori* de la densité des vraies parallaxes, et à partir de laquelle on peut calculer la densité marginale des parallaxes, comparable à la densité observée (celle des parallaxes de FAST ou NDAC). En tout cas pour les parallaxes des étoiles les plus lointaines, cette estimation ne sera pas très différente de la densité des vraies parallaxes.

Bien que cela puisse sembler surprenant de prendre comme densité *a priori* la densité des parallaxes spectroscopiques (ou photométriques) observées – donc une densité marginale –, une simple simulation suffit à le justifier ; cette simulation consiste simplement à générer pour chaque étoile une parallaxe à partir de sa magnitude absolue photométrique et d'une erreur gaussienne sur celle-ci : la comparaison entre la densité des parallaxes initiales et la densité des parallaxes simulées montre que ces deux densités sont quasiment identiques jusqu'à 3 mas environ et ne commencent à vraiment être différentes qu'à partir de 5 mas.

En pratique, pour éviter d'avoir à estimer la densité observée, on ne va pas travailler avec les densités de probabilité mais avec les fonctions de répartition. On va montrer dans ce paragraphe que cette comparaison entre une répartition théorique et la répartition

observée peut nous permettre d'obtenir simultanément une estimation du point-zéro et une estimation des erreurs externes sur les parallaxes.

Quel est l'avantage d'utiliser cette méthode, plutôt que de comparer directement les parallaxes préliminaires aux parallaxes spectroscopiques (ou photométriques) ? D'abord une certaine robustesse ; ceci peut se comprendre puisque l'on n'utilise que la densité et non les valeurs individuelles des parallaxes spectroscopiques (ou photométriques). De plus, on peut mettre une limite sur la parallaxe observée sans créer de biais. Enfin, la fonction de répartition empirique ne consiste qu'à compter le nombre d'étoiles observées, c'est-à-dire sans transformation des parallaxes Hipparcos.

Connaissant la densité des vraies parallaxes $f_\pi(\pi)$, et la densité des erreurs externes sur la parallaxe Hipparcos $f_{\sigma_\pi}(\sigma_\pi)$, la fonction de répartition des parallaxes Hipparcos s'écrit :

$$F(\pi) = \int_{-\infty}^{\pi} \int_0^{+\infty} \int_0^{+\infty} f(\pi_\mathrm{H}|\pi,\sigma) f_\pi(\pi) f_\sigma(\sigma) d\pi d\sigma d\pi_\mathrm{H}$$

Cette distribution cumulative peut être comparée à la fonction de répartition empirique,

$$F_n(\pi) = \sum_{i:\pi_{\mathrm{N}i}<\pi} \frac{1}{n}$$

On va donc prendre comme densité des vraies parallaxes celle des parallaxes spectroscopiques (ou photométriques) et comme densité des erreurs externes celle des erreurs internes de NDAC (ou FAST), et on calcule numériquement la fonction $F(\pi)$ avec la méthode de Gauß, et sa contrepartie empirique $F_n(\pi)$ pour chaque étoile, ce qui nécessite, soit dit en passant, un temps de calcul conséquent. Les étoiles utilisées pour une telle comparaison sont celles qui ont à la fois une parallaxe spectroscopique (ou photométrique) et une parallaxe NDAC (ou FAST).

La comparaison des fonctions de répartition (théorique et empirique) est représentée sur les figures 6.41 (12 000 étoiles) et 6.42 (5 500 étoiles). La première est obtenue en utilisant la densité des parallaxes spectroscopiques, la seconde en utilisant la densité des parallaxes photométriques. Dans les deux cas, les distributions théoriques et empiriques apparaissent superposées jusqu'à 3 mas environ.

On peut noter que les fonctions de répartition devraient être légèrement décalées si le point-zéro des parallaxes Hipparcos était différent de zéro. Un décalage est d'ailleurs légèrement visible sur la figure 6.44, comparaison entre la fonction de répartition empirique des parallaxes FAST et la fonction théorique obtenue en utilisant la densité des parallaxes photométriques. Un décalage encore plus important est visible si l'on utilise les parallaxes spectroscopiques (fig. 6.43), mais il provient peut-être de la densité utilisée pour les parallaxes. C'est pourquoi nous ne nous servirons dans un premier temps que des parallaxes photométriques parce que leurs erreurs sont plus petites que celles sur les parallaxes spectroscopiques, et donc leur densité supposée plus proche de la vraie densité.

On a cité ici deux paramètres libres, le point-zéro et le rapport moyen entre les erreurs externes sur les erreurs internes, pour introduire l'idée que l'ajustement entre les fonctions de répartition théorique et empirique peut servir à déterminer ces paramètres z et k, par l'intermédiaire de

$$f(\pi_\mathrm{H}|\pi,\sigma) = \frac{1}{k\sigma\sqrt{2\Pi}} e^{-\frac{1}{2}(\frac{\pi_\mathrm{H}-(\pi+z)}{k\sigma})^2}$$

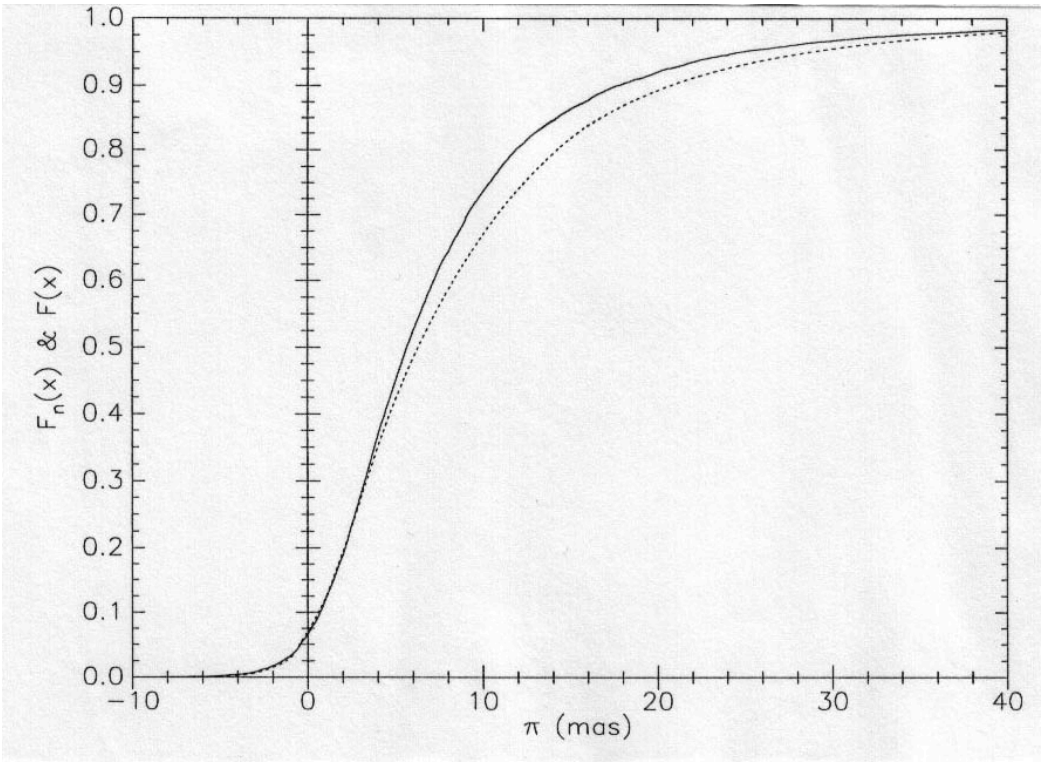


FIG. 6.41: *Comparaison entre la fonction de répartition empirique NDAC (trait continu) et celle calculée à partir de la densité des parallaxes spectroscopiques (pointillés)*

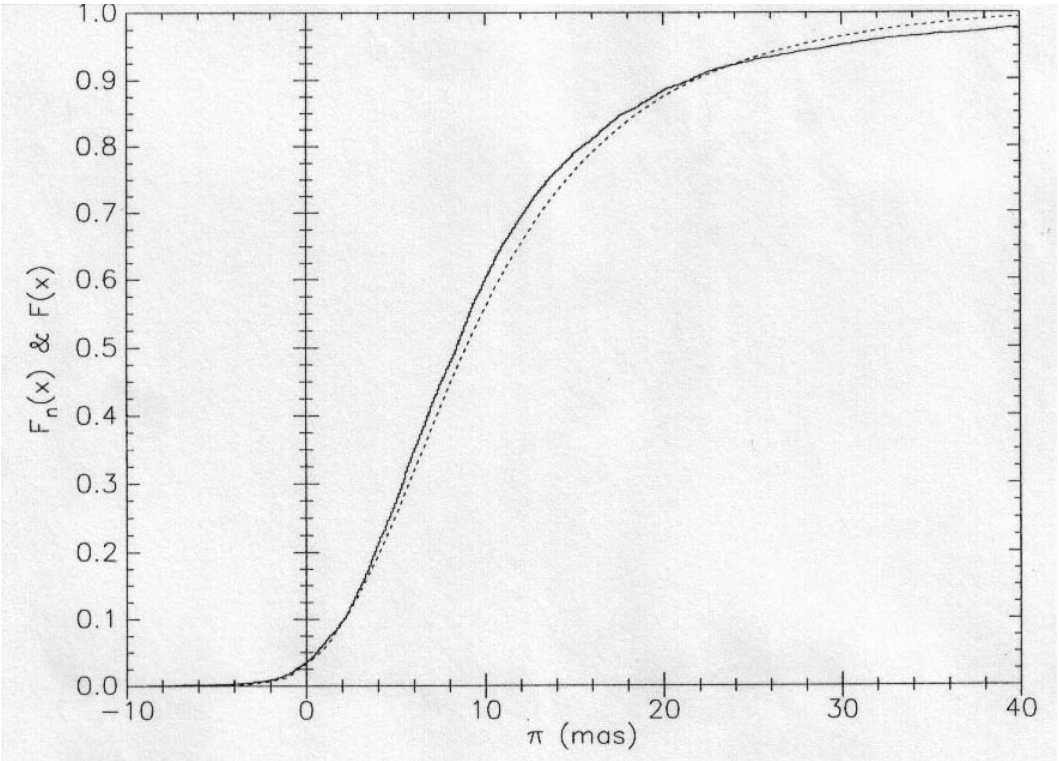


FIG. 6.42: *Comparaison entre la fonction de répartition empirique NDAC (trait continu) et celle calculée à partir de la densité des parallaxes photométriques (pointillés)*

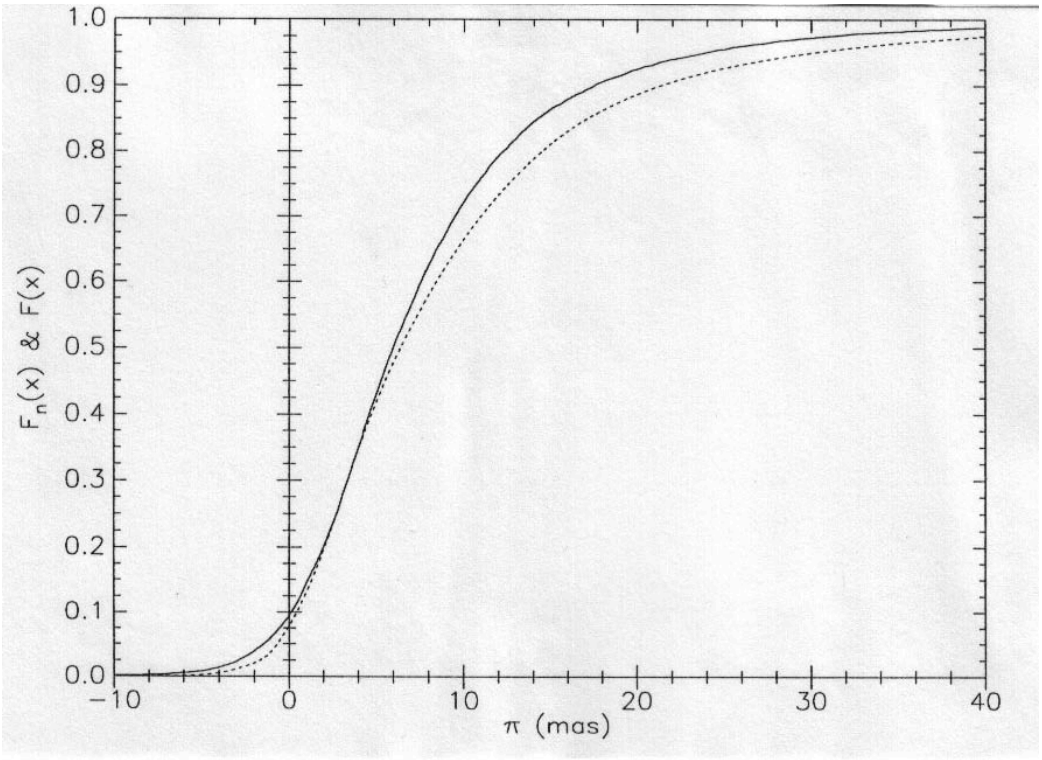


FIG. 6.43: *Comparaison entre la fonction de répartition empirique FAST (trait continu) et celle calculée à partir de la densité des parallaxes spectroscopiques (pointillés)*

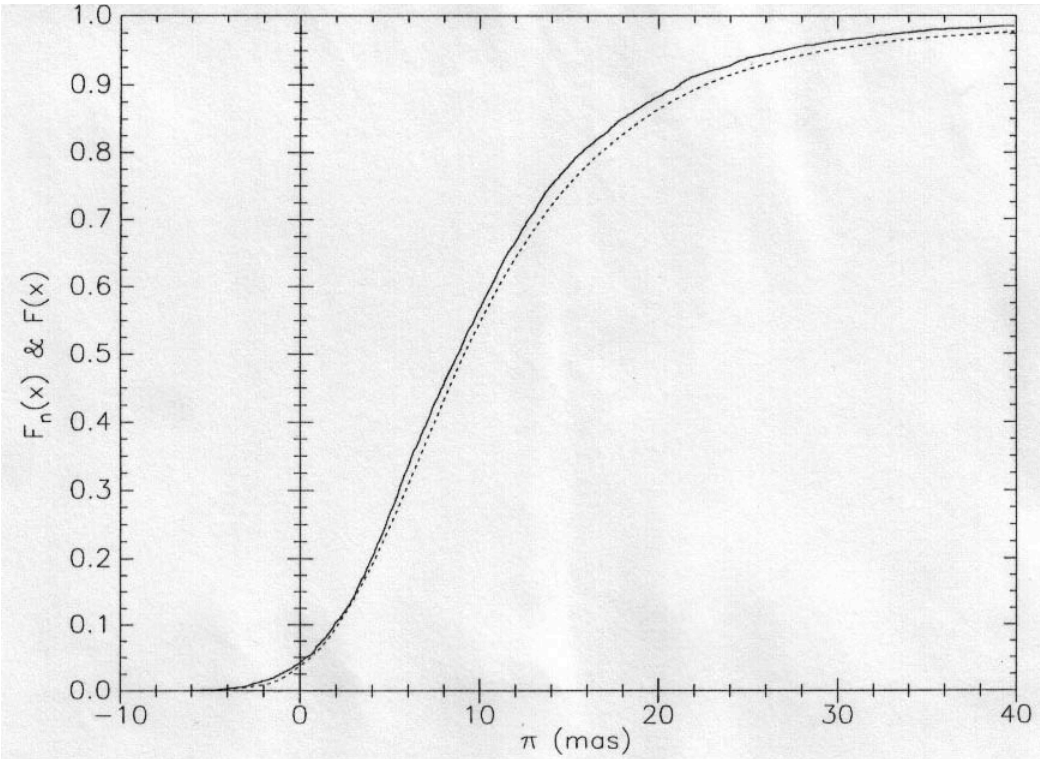


FIG. 6.44: *Comparaison entre la fonction de répartition empirique FAST (trait continu) et celle calculée à partir de la densité des parallaxes photométriques (pointillés)*

Quel est le critère d'ajustement ? Quand on compare deux distributions, pour tester statistiquement l'identité de ces distributions, on utilise le test de Kolmogorov, plus puissant que le test du χ^2. On se servira donc de la statistique $K_n = \sup_\pi |F_n(\pi) - F(\pi)|$ comme critère d'adéquation pour déterminer les deux paramètres d'intérêt.

L'ajustement a été calculé en deux temps, d'abord par une recherche aléatoire des couples (k, z) minimisant la statistique K_n, puis itérative pour parvenir à la solution définitive.

On pourrait s'attendre à ce que la solution obtenue se dégrade si l'on utilise les parallaxes les plus grandes, parce que la densité des parallaxes photométriques s'éloigne de la densité des vraies parallaxes. C'est pourquoi on présente sur le tableau 6.14 les différentes solutions obtenues pour (k, z) suivant la limite en parallaxe observée que l'on a utilisée ; il n'y a pas nettement de variation.

TAB. 6.14: *Variation de z avec la limite sur la parallaxe*

Estimation du point-zéro z (mas) et du rapport moyen k des erreurs externes sur les erreurs internes pour les parallaxes NDAC et FAST, en fonction de la limite sur la parallaxe observée.

parallaxe	z_F	k_F	z_N	k_N
$-6.0 < \pi < 1.0$	-0.034	1.316	-0.037	1.091
$-6.0 < \pi < 1.5$	-0.041	1.313	-0.051	1.102
$-6.0 < \pi < 2.0$	-0.031	1.318	-0.053	1.105
$-6.0 < \pi < 2.5$	-0.040	1.313	-0.037	1.100
$-6.0 < \pi < 3.0$	-0.043	1.312	-0.024	1.105

Nous avons effectué un certain nombre de simulations qui ont montré que cette méthode permettait d'obtenir le résultat correct. Ces simulations ont été effectuées en prenant les parallaxes photométriques, puis en générant des erreurs gaussiennes autour de ces parallaxes, puis en calculant les moyennes et les dispersions sur les paramètres obtenus par cette méthode pour une centaine d'échantillons de ce type.

Ceci a donc également fourni une estimation des erreurs standards sur les paramètres obtenus. L'erreur calculée sur k et z est de l'ordre de 0.03, donc très petite, mais cela ne mesure que l'erreur de la méthode employée ; cela ne prend par exemple pas en compte l'erreur sur la densité *a priori* que l'on a utilisée. D'un autre côté, le nombre d'étoiles en présence est assez faible (≈ 500), et les dispersions devraient donc diminuer si l'on augmente ce nombre d'étoiles, ce qui se produira lorsque toutes les parallaxes Hipparcos seront obtenues.

Avec les parallaxes étudiées, on obtient ainsi

$$\begin{array}{rclrcll} z_F & = & -0.03 & \pm 0.027 & \text{mas}, & k_F = 1.31 \pm 0.028 & \text{pour FAST-3P} \\ z_N & = & -0.04 & \pm 0.027 & \text{mas}, & k_N = 1.10 \pm 0.028 & \text{pour NDAC-5P} \end{array}$$

Avec ces valeurs, on voit sur les figures 6.45 et 6.46 que l'ajustement est parfait. L'aspect «en escalier» de la fonction de répartition NDAC est simplement dû au fait que nous

n'avons que le dizième de mas comme précision sur les parallaxes NDAC ; pour l'ajustement, on génère dans ce cas des décimales «aléatoires» supplémentaires (l'erreur d'arrondi n'est d'ailleurs pas prise en compte dans les erreurs standards calculées ci-dessus).

Comme on peut le constater, le point-zéro obtenu n'est pas différent de 0 de manière significative, aussi bien pour les parallaxes préliminaires FAST que NDAC. On notera que la différence des points-zéro entre les deux consortiums $z_F - z_N \approx 0.01 \pm 0.04$ mas n'est pas sensiblement différente de la différence moyenne empirique des parallaxes $\langle \pi_{\rm F} - \pi_{\rm N} \rangle \approx -0.04 \pm 0.02$ mas.

Nous n'avons utilisé qu'un petit sous-ensemble des données initiales, et il est possible que le point-zéro puisse varier légèrement suivant les échantillons choisis, tout au moins pour cette solution préliminaire. De même, il n'est pas impossible qu'il y ait quelques points aberrants dans la distribution de ces parallaxes préliminaires, ce qui aurait pour effet d'augmenter légèrement les rapports k_F ou k_N. Pour y remédier, une comparaison entre les distributions des parallaxes des deux consortiums devrait permettre de détecter et d'éliminer une partie de ces points. Pour plus de prudence, il serait également préférable de déterminer les rapports k_F et k_N par une autre méthode et de calculer avec celle-ci z_F et z_N uniquement.

Appliquant la même méthode, mais cette fois-ci avec comme distribution *a priori* celle des parallaxes spectroscopiques, on obtient

$$\begin{array}{rclcrcl} z_F & = & 0.09 \ \text{mas}, & & k_F & = & 1.27 \ \text{pour FAST-3P} \\ z_N & = & 0.07 \ \text{mas}, & & k_N & = & 1.02 \ \text{pour NDAC-5P} \end{array}$$

en utilisant les ≈ 2200 parallaxes plus petites que 2 mas. Mais cette fois-ci, l'ajustement est moins bon. L'estimation des rapports k est également différente de celle obtenue avec les parallaxes photométriques, mais dans tous les cas le point-zéro est très proche de 0.

6.7 Conclusions et perspectives

6.7.1 La parallaxe Hipparcos

L'ensemble de ce travail est naturellement préliminaire, puisqu'il n'a été appliqué qu'aux parallaxes préliminaires... Avant d'en tirer néanmoins quelques conclusions, nous rappelons toutefois qu'un des buts était de réfléchir à la méthodologie pour l'étude des parallaxes Hipparcos. De ce point de vue méthodologique, on a abordé sans doute une grande partie de ce qui était possible avec les données dont on dispose actuellement, en utilisant le moins possible de modèles particuliers, et le plus possible les données observées.

C'est ainsi que nous avons vu plusieurs méthodes pour déterminer le point-zéro ainsi que les erreurs externes des parallaxes. D'abord par des comparaisons internes entre consortiums, qui ne peuvent naturellement pas donner une estimation du point-zéro, mais qui, en revanche, permettent d'estimer les erreurs externes ainsi que la corrélation entre les erreurs des parallaxes des consortiums. Ensuite à l'aide de comparaisons externes : nous avons vu deux manières qui permettent d'obtenir les estimations des moyennes et erreurs externes des parallaxes trigonométriques.

La première consiste à prendre des étoiles suffisamment lointaines pour que les erreurs log-normales sur les estimations au sol soient négligeables. En utilisant donc ces étoiles

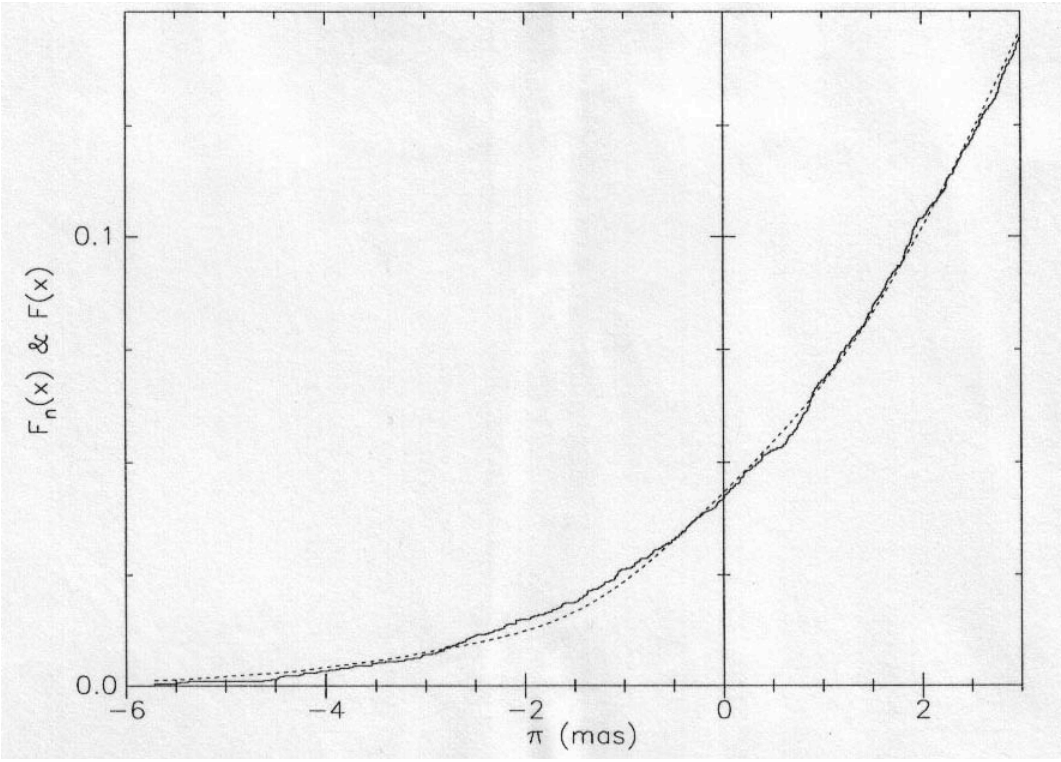


FIG. 6.45: *Comparaison entre la fonction de répartition empirique FAST (trait continu) et celle ajustée à partir de la densité des parallaxes photométriques et de la meilleure estimation de* (k_F, z_F) *(pointillés)*

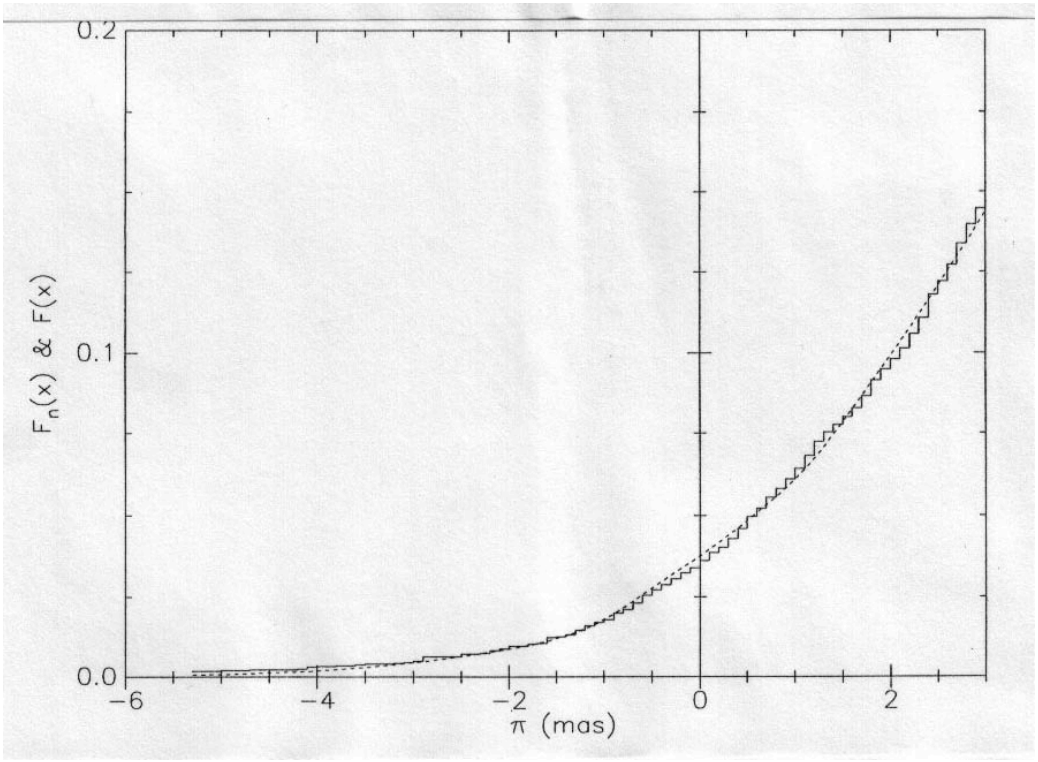


FIG. 6.46: *Comparaison entre la fonction de répartition empirique NDAC (trait continu) et celle ajustée à partir de la densité des parallaxes photométriques et de la meilleure estimation de* (k_N, z_N) *(pointillés)*

lointaines, on obtient une estimation des erreurs standards sur les parallaxes Hipparcos à l'aide de la moyenne et de la dispersion de la différence entre ces parallaxes et les estimations externes.

La seconde consiste à ajuster la fonction de répartition théorique des parallaxes, obtenue en prenant la densité des parallaxes spectroscopiques ou photométriques comme loi *a priori*, à la fonction de répartition empirique des parallaxes Hipparcos.

Pour obtenir les parallaxes Hipparcos définitives, une méthode consisterait sans doute :

1. à utiliser la première méthode ainsi qu'une comparaison interne entre les consortiums de manière à obtenir les erreurs externes et les corrélations des erreurs. La variation des erreurs externes devra être étudiée en fonction de divers paramètres (magnitude, latitude, ...). Il faudra préalablement à cette étape détecter et éliminer d'éventuels points aberrants ;

2. le rapport moyen entre les erreurs externes sur les erreurs internes étant ainsi fixé, le point-zéro global serait calculé à l'aide de la seconde méthode. ;

3. les points-zéro et les erreurs externes ayant été calculés pour chaque consortium, les parallaxes pourraient être combinées de façon à en obtenir pour chaque étoile le meilleur estimateur.

On aura sans doute constaté que les différentes méthodes appliquées aux parallaxes préliminaires ne donnent pas des résultats strictement identiques, preuve que ces méthodes nécessitent d'être raffinées. Il ne faut pas oublier non plus que les parallaxes utilisées sont celles obtenues après seulement un an de données, dont on a vu qu'elles avaient encore quelques petits effets dépendant des caractéristiques des étoiles, ceci devant sans aucun doute changer au fur et à mesure que s'allonge la durée de la mission, et que s'améliorent ainsi les résultats des consortiums de données. Une chose est sûre, cependant : les parallaxes préliminaires sont déjà d'une qualité indiscutable, et nous avons pu le constater tout au long de ces pages. Les solutions des deux consortiums sont proches, et nous avons d'ailleurs vu la meilleure façon de les combiner.

Les prolongements possibles à ce travail sont naturellement la validation des parallaxes définitives Hipparcos, mais également tout ce qui concerne la calibration des magnitudes absolues spectroscopiques, qui ont été utilisées ici, et pour lesquelles les parallaxes Hipparcos joueront un rôle déterminant.

6.7.2 Calibration des magnitudes absolues

Il a été largement question des magnitudes absolues spectroscopiques ou photométriques dans les pages précédentes, en relation avec la parallaxe Hipparcos. Ce qui est naturellement sous-jacent est la question de la calibration future des magnitudes absolues à l'aide des parallaxes Hipparcos.

Du point de vue méthodologique, beaucoup d'articles ont été consacrés à la calibration des magnitudes absolues moyennes d'un groupe homogène d'étoiles à l'aide des parallaxes trigonométriques, et le sujet s'y prête : biais «de Lutz-Kelker» sur la parallaxe observée, biais de Malmquist sur la magnitude absolue moyenne, données censurées, points aberrants, tout concourt à faire de ce problème originellement astronomique un problème

statistique. Il est clair qu'une naïve application de la loi de Pogson (pour déduire la magnitude absolue à partir de la parallaxe) ne suffit pas pour obtenir la magnitude absolue moyenne d'un groupe d'étoiles.

Et ce n'est pas la formidable amélioration des précisions sur la parallaxe qu'apportera Hipparcos qui résoudra ce problème. On pourrait en effet penser que l'on n'utilisera que les parallaxes les plus précises de façon à limiter le biais «de Lutz-Kelker». Il doit être clair que le biais sera toujours présent, à un degré moindre certes, mais comme on cherchera à obtenir une meilleure précision sur les magnitudes absolues (sinon pourquoi demander une bonne précision aux parallaxes trigonométriques), on n'aura fait que repousser le problème. De plus, ignorer le nombre très important de parallaxes avec une précision relative médiocre, non seulement représente une censure de l'information, mais ne sera tout simplement pas possible ; y compris pour les étoiles avec une parallaxe négative, car cette mesure – qui peut sembler à première vue inutilisable – apporte tout de même de l'information. Pour de telles étoiles, comme pour les autres, on pourrait faire usage de toutes les autres informations marginales apportées par les autres observables (mouvements propres, vitesse radiale et les différentes caractéristiques photométriques et spectrales).

Au contraire, le défi à relever est de taille : comment traiter un nombre très important d'étoiles possédant de nombreuses données (éventuellement très bruitées, certaines inexactes) avec la méthodologie adéquate. Ceci implique non seulement la formulation statistique correcte, mais également des moyens de calculs importants. Le temps de calcul n'est cependant que de peu d'importance, quand on le compare aux ≈ 10 ans nécessaires à l'acquisition des parallaxes Hipparcos...

Pour récapituler l'historique des problèmes méthodologiques liés à la calibration des magnitudes absolues, il faut remonter à Malmquist (1920, 1936) qui calcule le biais sur la magnitude absolue moyenne d'un groupe d'étoiles. En ce qui concerne les parallaxes, Trumpler & Weaver (1953, p. 369) mentionnent le fait que leur distribution non uniforme et l'erreur sur la parallaxe observée introduisent un biais lorsque l'on fixe une limite à la parallaxe observée. Ce biais sera formulé sous la forme d'une correction à la magnitude absolue par Lutz & Kelker (1973), et revu dans le cas d'un échantillon limité en magnitude dans Lutz (1979).

Toujours sous l'angle d'une correction, Hakkila (1989) a étudié la magnitude absolue moyenne d'un groupe d'étoiles dans un échantillon limité en distance. Mais, bien que reprenant et corrigeant l'estimation de Lutz-Kelker, Turon & Crézé (1977) préconisent plutôt une approche où l'estimation de la magnitude moyenne est obtenue par maximum de vraisemblance sur un échantillon où aucune censure n'est faite sur la parallaxe observée.

Avec un point de vue *explicitement* bayésien (les corrections mentionnées ci-dessus l'étant implicitement), Smith Jr (1987a-d) apporte une vision cohérente de l'estimation des magnitudes absolues à partir des parallaxes trigonométriques , ainsi que de l'estimation de la parallaxe la plus probable, sachant la parallaxe spectroscopique et la parallaxe trigonométrique [Smith, 1985].

Enfin, récemment, Ratnatunga & Casertano (1991) ont développé un algorithme qu'ils ont appliqué à la calibration de la magnitude absolue comme fonction linéaire de l'indice de couleur $R - I$. Ils avaient déjà utilisé le même type d'approche pour une analyse cinématique des populations galactiques [Casertano *et al.*, 1990]) et on peut la décrire succintement.

En choisissant comme modèle une loi *a priori* paramétrique $M_V = f_\Theta(R-I)$, on calcule pour chaque étoile la densité marginale de la parallaxe observée puis la densité conditionnelle sachant le modèle, et la valeur des paramètres Θ est enfin déterminée par maximum de vraisemblance.

Contrairement à la plupart des estimations faites précédemment, cette approche n'utilise donc que les données *observées* et ne cherche pas à déterminer un meilleur estimateur de la vraie magnitude absolue individuelle, par exemple. Mais les points les plus remarquables de l'algorithme développé sont l'insensibilité aux effets de sélection de l'échantillon utilisé, la prise en compte des censures éventuelles (comme l'utilisation des parallaxes les plus précises), la correction implicite des biais de Lutz-Kelker et de Malmquist, et enfin un test sur les points aberrants.

Cette approche est très intéressante et à titre d'exemple on pourrait l'appliquer de la manière suivante, en utilisant la photométrie $uvby{-}\beta$: trouver un modèle

$$M_V = g_\Theta(T_{\rm eff}, \log g, [{\rm Fe/H}], v\sin i, \dots) = g_\Theta(u, v, b, y, \beta, v\sin i)$$

après correction du rougissement sur les indices observés.

Lorsque les données définitives Hipparcos seront obtenues, nul doute que nous nous intéresserons donc à la calibration du diagramme-HR. D'ici là nous tenterons d'élaborer la méthodologie statistique adéquate.

Quatrième partie

CINÉMATIQUE

Dans cette dernière partie nous aborderons un problème concernant la cinématique stellaire, et plus particulièrement celle des étoiles naines de type spectral A de population I. Cette partie constitue le carrefour de ce que l'on a abordé dans les chapitres précédents et conclut donc cette thèse.

En effet, nous utiliserons des distances obtenues à partir des magnitudes absolues déduites de la photométrie $uvby{-}\beta$ (chap. 2), corrigées de l'absorption. Les données traitées (magnitudes apparentes, positions, mouvements propres, vitesses radiales) proviennent de la base INCA (décrite au chap. 1) ; certaines données complémentaires nous seront cependant utiles. Quant aux méthodes que nous adopterons pour traiter ces données, elles sont essentiellement décrites au chapitre 4 : nous nous attacherons à valider statistiquement la réalité des groupes mis en évidence.

Nous montrerons en effet dans ce chapitre que la distribution des vitesses des étoiles étudiées peut être plus adéquatement expliquée par la contribution de plusieurs groupes cinématiques dont une partie des étoiles de chaque groupe aurait une origine commune. La conséquence en est que, contrairement à ce qui est communément admis, le temps de mélange dynamique est beaucoup plus long qu'une année galactique.

Chapitre 7

Distribution locale des vitesses d'étoiles A

7.1 Les vitesses spatiales

Notre Galaxie est animée d'une rotation différentielle, les étoiles proches du centre galactique tournant plus rapidement que celles des régions périphériques, la période de rotation (année galactique) étant d'environ 2.4 10^8 années. Au voisinage du Soleil, la vitesse linéaire de rotation des étoiles est d'environ $V_0 = 220$ km/s.

Pour repérer la position d'une étoile dans la Galaxie, nous utiliserons les coordonnées cartésiennes (X, Y, Z), où X est dans la direction du centre galactique, Y dans la direction de la rotation galactique, et Z vers le pôle nord galactique. Nous noterons (U, V, W) les composantes de la vitesse spatiale d'une étoile par rapport au soleil, corrigée de la rotation différentielle, les trois axes étant alignés de la même façon que (X, Y, Z).

Si l'on admet que, dans un groupe homogène d'étoiles de population I, les composantes des vitesses par rapport au soleil sont distribuées de façon approximativement gaussienne, de moyennes $\overline{U}, \overline{V}, \overline{W}$ et de dispersions $\sigma_U, \sigma_V, \sigma_W$, alors la distribution tri-dimensionnelle des vitesses est ellipsoïdale, avec $\sigma_U > \sigma_V > \sigma_W$, et définit ce que l'on appelle l'*ellipsoïde des vitesses*, dont la direction du grand axe est nommée direction du *vertex*, et l'angle de cette direction avec la direction du centre galactique est la *déviation du vertex*. On observe, pour les étoiles de population I, que la déviation du vertex est plus grande pour les étoiles les plus jeunes, et que la dispersion des vitesses augmente progressivement avec l'âge. L'augmentation de σ_U avec l'âge a notamment pour conséquence de contribuer au *courant asymétrique*, visible à l'asymétrie de la distribution de la composante V (voir King, 1989, p. 138).

Pour un groupe homogène d'étoiles, le vecteur $(\overline{U}, \overline{V}, \overline{W})$ représente la vitesse moyenne des étoiles du groupe par rapport au soleil, et son opposé $(-\overline{U}, -\overline{V}, -\overline{W})$ est le mouvement solaire ; la vitesse $-\overline{V}$ diffère de la composante $v_\odot$ de la vitesse particulière du soleil par rapport au centre local des vitesses (LSR, point fictif ayant la vitesse circulaire V_0) par le courant asymétrique. Pour que la vitesse particulière du soleil soit bien déterminée, il faut que les vitesses soient bien «mélangées», ce qui n'est pas possible avec les plus jeunes étoiles.

Pour expliquer l'augmentation de la dispersion des vitesses avec l'âge, Wielen (1977) a fait intervenir un processus de diffusion des orbites stellaires, causé par des instabilités

locales du champ gravitationnel dans notre Galaxie. La conséquence en serait une augmentation de la dispersion des vitesses proportionnellement à $t^{\frac{1}{2}}$. Ce mécanisme impliquerait également un temps de mélange de l'ordre de 2 10^8 ans : la vitesse d'une étoile changerait aléatoirement de plus de 10 km/s en une rotation galactique. Par conséquent les étoiles plus vieilles qu'une année galactique environ devraient avoir effacé toute mémoire cinématique du moment de leur formation et ne plus être distinguable des autres étoiles de champ. Lacey (1984) et Paloŭs & Piskunov (1984), entre autres, trouvent également que le temps de mélange est de l'ordre de 2 10^8 ans.

Le but de ce chapitre est d'analyser les distributions des vitesses d'étoiles plus vieilles que 4 10^8 ans afin de mettre en évidence des groupes dont les étoiles partagent le même comportement cinématique. Si l'on voit encore dans leurs caractéristiques cinématiques les reliefs des vitesses qu'elles possédaient lors de leur formation, il est clair que le temps de mélange ci-dessus mentionné est sous-évalué. Plus généralement, on ne saurait mieux formuler l'enjeu qu'en traduisant ce qu'écrivait Eggen (1965) : «*On a pour habitude, en appliquant des procédures statistiques diverses à l'étude des mouvements stellaires, de supposer que ces mouvements sont distribués aléatoirement, avec, au plus, quelques variations mineures. Si, en réalité, les mouvements observés sont essentiellement dûs à ceux de quelques groupes d'étoiles, alors beaucoup de ces procédures peuvent s'avérer invalides.*» Nous allons montrer qu'en effet, dans les échantillons étudiés, les distributions des composantes des vitesses sont beaucoup plus adéquatement représentés par des mélanges de quelques populations gaussiennes, représentant autant de groupes cinématiques différents.

Dans un premier temps, nous allons calculer les composantes de la vitesse spatiale en fonction des distances, mouvements propres et vitesses radiales des étoiles, et obtenir les expressions des erreurs formelles sur ces composantes.

Calcul des vitesses spatiales

À partir des composantes du mouvement propre ($\mu_\alpha \cos\delta$, μ_δ) d'une étoile, de sa vitesse radiale v_r, et de sa distance héliocentrique r, et après correction de l'effet dû à la rotation galactique, on peut montrer que les composantes de la vitesse spatiale de l'étoile relative au soleil s'écrivent sous la forme :

$$\left|\begin{array}{rcllll} U & = & (m_{11}\mu_\alpha \cos\delta + m_{12}\mu_\delta + m_{13}) & r & +m_{14}v_r \\ V & = & (m_{21}\mu_\alpha \cos\delta + m_{22}\mu_\delta + m_{23}) & r & +m_{24}v_r \\ W & = & (m_{31}\mu_\alpha \cos\delta + m_{32}\mu_\delta + m_{33}) & r & +m_{34}v_r \end{array}\right.$$

où les m_{ij} dépendent des coordonnées galactiques l et b de l'étoile, des coordonnées équatoriales du pôle galactique nord et du nœud ascendant du plan galactique ; les termes m_{i3}, dépendant également des constantes de Oort A et B, permettent de corriger de la rotation différentielle.

En écrivant sous cette forme les composantes U, V et W, et en supposant que les erreurs sur le mouvement propre, la distance et la vitesse radiale sont indépendantes, la matrice de variance-covariance des composantes de la vitesse spatiale s'exprime alors facilement en fonction des variances formelles des variables initiales ($s_r^2, s_{\mu_\alpha \cos\delta}^2, s_{\mu_\delta}^2, s_{v_r}^2$)

Nous ferons l'approximation que la distribution des erreurs est gaussienne, ce qui est inexact pour les distances, mais sans doute admissible pour les mouvements propres

et la vitesse radiale. Par conséquent, nous supposerons que les erreurs sur les vitesses spatiales sont également distribuées de façon approximativement gaussienne ; de plus, pour simplifier, nous négligerons les covariances des erreurs.

Grâce au calcul précédent des erreurs formelles sur les composantes de la vitesse, nous pourrons ultérieurement séparer de manière plus précise les différentes populations d'étoiles. L'ordre de grandeur des erreurs formelles pour les échantillons utilisés est $\langle s_U \rangle \approx 3.3$, $\langle s_V \rangle \approx 3.1$, $\langle s_W \rangle \approx 2.9$ km/s.

7.2 Bouffées de formation d'étoiles

L'étude contenue dans l'article ci-joint [Gómez *et al.*, 1990] concerne un échantillon limité en magnitude apparente d'étoiles classées spectroscopiquement naines, de type B5 à F5, et plus proches que 250 pc.

Après avoir déterminé les caractéristiques de l'ellipsoïde des vitesses pour chaque type spectral, on peut constater (fig. 1) la présence de plusieurs modes sur l'histogramme des composantes des vitesses (essentiellement U). L'idée de Gómez *et al.* fut de relier ces modes à des reliefs de formation d'étoiles par bouffées, à chaque bouffée correspondant une vitesse moyenne et une dispersion de l'ordre de grandeur de celle des nuages interstellaires dans lesquels se forment les étoiles.

Puisque, sous cette hypothèse, on continuerait à voir des bouffées de formation pour des étoiles de type F5 ($\leq 2\ 10^9$ ans), on notera que cette interprétation aurait pour conséquence de prohiber l'utilisation d'étoiles plus jeunes – aux vitesses encore moins bien mélangées – pour estimer la vitesse du soleil à l'aide de la relation du courant asymétrique.

De plus, l'hypothèse concernant la mise en lumière de restes de bouffées de formation sur des étoiles assez vieilles a une autre conséquence importante : le temps de mélange serait bien supérieur à 2 années galactiques.

Dans la mesure où les étoiles des échantillons étudiés n'avaient pas d'âge individuel, Gómez *et al.* utilisèrent des âges d'amas – supposant que chaque éventuelle bouffée avait pû créer également des amas – pour montrer que les modes des composantes des vitesses correspondaient aux vitesses moyennes de ces amas.

Astron. Astrophys. 236, 95–98 (1990)

ASTRONOMY AND ASTROPHYSICS

Research Note

Local kinematic properties of Population I (B5–F5)-type stars and galactic disk evolution

A.E. Gómez [1,⋆], J. Delhaye [1,⋆], S. Grenier [1,⋆], C. Jaschek [2,⋆⋆], F. Arenou [1,⋆], and M. Jaschek [2,⋆⋆]

[1] Observatoire de Meudon, Place Jules Janssen, F-92195 Meudon Cedex, France
[2] Centre de Données Stellaires, Observatoire de Strasbourg, 11, rue de l'Université, F-67000 Strasbourg, France

Received January 9, accepted February 5, 1990

Abstract. Using a sample of (B5-F5) V type stars, located up to 250 pc, the basic parameters of the velocity ellipsoid have been derived. The observed distribution function of the peculiar velocities for different subsamples have been decomposed into a sum of three-dimensional gaussians using the SEM algorithm (Celeux and Diebolt, 1985, 1986). Assuming that the stars are formed in bursts, the observed velocity distribution is explained as the sum of several independent (approximately spherical) distributions, each one corresponding to one generation of stars.

Key words: kinematics of (B5-F5) V type stars – velocity ellipsoid – solar velocity – velocity dispersion increase with age – galactic disk evolution

1. Introduction

Absolute magnitude calibrations of (B5-F5) type stars, dwarfs and giants, were derived by Grenier et al. (1985, Paper I) using large and homogeneous samples of stars, carefully selected and treated statistically in the same way. The purpose of our work is a kinematical study of the solar neighbourhood based on a large sample – 1000 stars – from Paper I.

The sample used is the apparent-magnitude limited sample of dwarf stars of Paper I, restricted to the region nearer than 250 pc. The sources for the different spectroscopic and photometric data are given in Paper I. We have used the absolute magnitude calibrations obtained in Paper I to derive the heliocentric distance of each star. Our sample does not contain high velocity stars nor known cluster members.

2. The peculiar velocities distribution function

In Table 1 we present the results concerning the first and second order moments of the distribution functions of the residual velocity components for the different subsets (we call subset any of the subdivisions of our sample). The components of the space velocity with respect to the centroid were calculated for each star, after correcting for galactic rotation. They are expressed in the conventional directions, namely: U, in the direction of the galactic center; V, in the direction of the galactic rotation; W, in the direction perpendicular to the galactic plane. In the Table 1, U_0, V_0, W_0 are the components of the solar velocity with respect to the group centroid expressed in km s^{-1}. Similarly σ_U, σ_V, and σ_W are the semi-major axes of the velocity ellipsoid in the U, V, and W directions, also expressed in km s^{-1}. The vertex deviation ϕ is given in degrees; it is omitted when $\sigma_U \cong \sigma_V$, since then ϕ is almost undetermined. The last column finally provides the logarithm of the average age of each group. The average age ($\bar{t}$) in years was derived using the isochrones of Maeder and Mermilliod (1981), taking into account the dispersion on the absolute magnitude obtained in Paper I.

The obtained results are as expected. Nevertheless, the large spread in the values of U_0 in Table 1 deserves attention. Figure 1 gives the distribution of the U components for each subset. For the sake of comparison we have also indicated (vertical bars) the positions of the open cluster velocity components, but including only clusters nearer than 250 pc (the distance limit of our sample). The average cluster velocity components are given in Table 2, where the clusters are ordered according to their ages. The latter are taken from Mermilliod (1981a). The age dispersion within a group amounts to 0.05 in log t for the younger clusters and to 0.02 for the older. The age given by Mermilliod (1981b) are in good agreement with those given by Palouš et al. (1977), except for IC 2602 for which Palouš et al. quote $(10 \pm 0.5)\ 10^7$ yr.

With the exception of the earliest (and youngest) stars, Fig. 1 suggests a possible mixture of velocity distributions. Moreover, the position of the maximum for the youngest stars agrees well with those of open clusters with ages $< 8\ 10^7$ yr. For the subset A0-A2 there are two observed maxima, one coincides with that of the earliest stars and the other agrees with that of the UMa moving cluster. The latter has an age intermediate between the youngest clusters and the Coma Ber cluster. For A3 and later subsets contributions from several distributions seem present. In particular, the "gap" which appears in the A2 subset is filled up in the A3-A4 subset and coincides very well with the position of Coma

Send offprint requests to: A.E. Gómez

⋆ URA N° 335 (CNRS)
⋆⋆ URA N° 654 (CNRS)

Table 1. Velocity ellipsoid

Type	N	U_0 (km s^{-1})	V_0 (km s^{-1})	W_0 (km s^{-1})	σ_U (km s^{-1})	σ_V (km s^{-1})	σ_W (km s^{-1})	ϕ (deg)	$\log \bar{t}$
B5-B7V	53	11.6±1.5	17.0±1.5	7.8±0.9	10.5±1.0	10.6±1.0	6.8±0.7		7.8
B8V	45	11.8±1.4	14.8±2.2	7.1±1.4	9.2±1.0	14.6±1.5	9.2±1.0		8.2
B9V	55	11.9±1.4	15.1±1.5	7.9±1.4	10.1±1.0	11.1±1.1	10.2±1.0		8.3
B9.5V	32	14.1±2.4	10.7±1.8	8.1±1.4	13.5±1.7	10.2±1.3	7.7±1.0		8.45
A0V	84	13.2±1.9	13.6±1.3	7.6±1.0	17.5±1.4	11.8±0.9	8.8±0.7	19	8.5
A1V	101	8.9±1.7	12.2±1.2	8.3±0.8	16.8±1.2	11.9±0.8	7.7±0.5	27	8.6
A2V	96	9.3±2.0	8.6±1.2	7.0±0.7	19.3±1.4	11.8±0.8	8.2±0.6	24	8.65
A3-A4V	121	6.9±1.7	7.3±1.1	6.8±0.7	18.5±1.2	11.6±0.8	7.9±0.5	25	8.7
A5-A6V	43	9.3±2.4	9.2±1.9	6.7±1.4	16.0±1.7	12.7±1.4	9.0±1.0	30	8.85
A7V	37	13.6±3.0	11.6±2.3	8.3±1.4	18.3±2.1	13.8±1.6	8.7±1.0	29	8.9
A8-A9V	29	14.6±3.4	11.2±2.6	5.8±1.2	18.3±2.4	13.9±1.8	6.5±0.9	32	9.0
F0V	53	12.3±2.8	9.7±1.7	5.3±1.8	20.5±2.0	12.4±1.2	12.9±1.3	22	9.1
F1-F2V	67	11.6±2.8	8.1±1.4	5.9±1.0	22.4±1.9	11.3±1.0	8.0±0.7	10	9.15
F3V	33	11.7±3.8	6.4±2.3	9.8±1.9	21.4±2.6	13.3±1.6	11.0±1.4	21	9.25
F4V	37	10.4±3.8	12.7±2.2	6.5±1.8	22.5±2.6	13.3±1.5	10.9±1.3	11	9.3
F5V	103	7.9±2.5	10.7±1.5	5.7±1.2	25.3±1.8	14.9±1.0	12.1±0.8	12	9.3

Notes: N: number of stars. U_0, V_0,W_0: solar velocity components and their corresponding rms errors. σ_U, σ_V, σ_W: residual velocity dispersions and their corresponding rms errors. ϕ: vertex deviation angle. $\bar{t}$: average age in years.

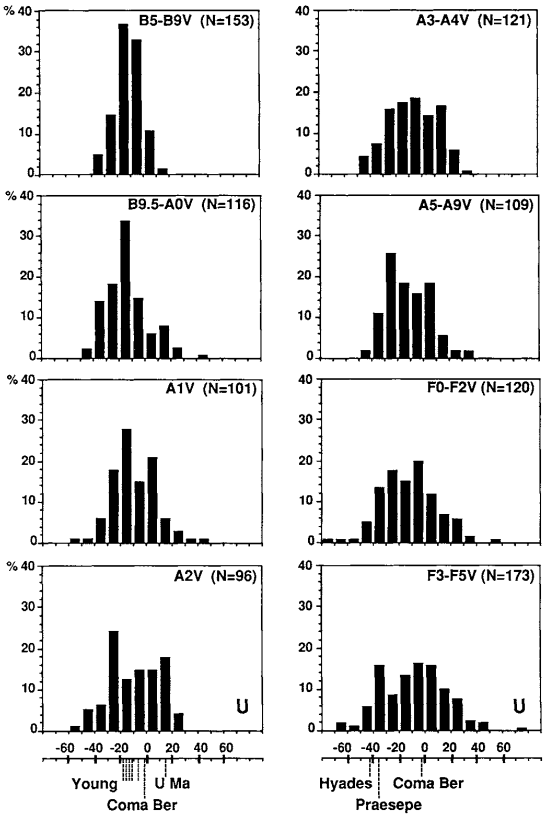


Fig. 1. Histogram of the distribution of the U components (in km s^{-1}). The vertical bars correspond to the open clusters values from Table 2

Ber. Finally for the F3-F5 subset there appears a new contribution, which coincides with the position of the old clusters Hyades and Praesepe.

We intend to interpret the histograms of Fig. 1 assuming that stars are formed in "bursts", i.e. in a discontinuous fashion. If we observe a long time after the "bursts", the velocity histogram is a sum of the distributions corresponding to the different bursts, plus an addition of stars which migrated into the solar vicinity from outside. The contamination as well as the "spreading" of the old star bursts (mainly due to stochastic acceleration processes) will become progressively more effective as time passes, so that after a "long time" any trace of the individual burst will be lost.

In order to isolate possible bursts of star formation and to assert that the "bumps" observed in Fig. 1 are physical associations (and not the result of the human eye image-processing), we have decomposed each sample (B9.5-A4 and F3-F5) into a sum of three-dimensional Gaussians. To do so we have used the SEM (Stochastic, Expectation, Maximization) algorithm developed by Celeux and Diebolt (1985, 1986). The aim of the SEM algorithm is to resolve the finite mixture density estimation problem under the maximum likelihood approach, using a probabilistic teacher step. Full details can be found in the above mentioned papers. Through SEM one can obtain the number of components of the Gaussian mixture (without any assumption on this number), its mean values and dispersions and the percentage of each component with respect to the whole sample. SEM also gives an estimation of the parameters standard-deviations, which allow to measure the degree of overlap of the mixture components. Different statistical procedures and tests were applied in order to verify the results obtained with SEM; EM algorithm with Bootstrap, the Wilks test (Soubiran, 1988; Soubiran et al., 1989) and multivariate data analysis (Bougeard et al., 1989; Bougeard and Arenou, 1989). Finally, the errors in the data used to calculate the velocity components (radial velocity, proper motions and distance) were taken into account (Diebolt and Celeux, 1989). The errors were assumed to be distributed randomly, with null expectation and mean square error obtained from the sources of the corresponding

data. For the distances, a 20% relative error was adopted. The results are given in Table 3. We can isolate four possible bursts of star formation, each one defined by its $\bar{U}$, $\bar{V}$, and $\bar{W}$, which are indicated in the "remarks" column. As for the dispersions σ_U, σ_V, and σ_W we notice that all of them are $\leqq 14\,\mathrm{km\,s^{-1}}$ and of similar magnitude. The obtained distribution functions are approximately spherical, the corresponding mean dispersions $\bar{\sigma}$ are smaller than $11\,\mathrm{km\,s^{-1}}$, which suggest that the bursts kinematically share the same characteristics of the gas in the solar neighbourhood.

We may now ask if these bursts are physically real, or if they are pure random coincidences. We would expect that stars born in the same burst have the same or similar ages. Since we cannot determine individual ages, we can circumvent the difficulty by assuming that each burst also produced a few open clusters; both the kinematics and the ages of these clusters must then be compatible with those of the star bursts.

We have remarked already that Fig. 1 shows a good agreement between the position of the bursts and the open clusters in the U component diagram. We find then that burst I corresponds to young stars with ages less than or equal to $8\ 10^7$ yr, burst II to the stars of the UMA cluster generation, burst III to those of the Coma Ber generation and IV to those of the Hyades generation.

If we consider that these bursts are real (i.e. not chance associations) we must show that:

1. the distribution functions (of all three velocity components) of the bursts coincide well with those of the clusters associated with the bursts;

2. the ages of the stars of each burst and the ages of the clusters coincide.

With regard to the first point, Table 4 provides the data for the bursts and the clusters, the latter being taken from Table 2. In Table 4, we have also included the subset of (B5-B9) V stars which, according to our considerations, are associated to burst I.

Table 2. Velocity components for clusters within 250 pc

Age (years)	Cluster	U ($\mathrm{km\,s^{-1}}$)	V ($\mathrm{km\,s^{-1}}$)	W ($\mathrm{km\,s^{-1}}$)	Ref.
$3.6\ 10^7$	NGC 2451	-15.1 ± 2.1	-19.7 ± 1.2	-16.5 ± 2.7	M
	IC 2391	-18.3 ± 2.1	-13.5 ± 0.8	-5.9 ± 2.1	P
	IC 2602	-0.7 ± 1.8	-25.7 ± 1.1	-1.4 ± 1.8	P
$5.1\ 10^7$	α Per	-10.8 ± 1.4	-20.5 ± 2.0	-0.7 ± 2.2	P
	BL 1	-17.0 ± 2.8	-11.3 ± 2.1	-9.2 ± 1.1	M
$7.8\ 10^7$	Pleiades	-5.8 ± 1.3	-24.0 ± 2.0	-12.4 ± 2.0	P
$3.0\ 10^8$	U Ma	$+14.5\pm0.8$	$+2.5\pm0.6$	-8.5 ± 0.9	E
$4.0\ 10^8$	Coma Ber	-1.8 ± 1.1	-8.2 ± 1.1	-0.7 ± 0.6	P
$6.6\ 10^8$	Hyades	-44.4 ± 0.8	-17.0 ± 1.0	-5.0 ± 1.3	P
	Praesepe	-37.1 ± 1.6	-23.5 ± 2.4	-7.0 ± 1.7	P

References: E: Eggen (1973), M: Mermilliod (1986), P: Palouš et al. (1977).

Table 3. Decomposition of the velocity distribution in gaussian components

Type	N	$\bar{U}$ ($\mathrm{km\,s^{-1}}$)	σ_U ($\mathrm{km\,s^{-1}}$)	$\bar{V}$ ($\mathrm{km\,s^{-1}}$)	σ_V ($\mathrm{km\,s^{-1}}$)	$\bar{W}$ ($\mathrm{km\,s^{-1}}$)	σ_W ($\mathrm{km\,s^{-1}}$)	%	Remark
(B9.5-A0) V	94	−19	11	−16	11	−8	7	81	Burst (I)
$N=116$	22	11	13	−2	4	−5	7	19	Burst (II)
A1 V	74	−16	13.5	−17	9.5	−8	7	73	Burst (I)
$N=101$	27	10	7	2	4	−9	6	27	Burst (II)
A2 V	61	−21	13	−14	10	−7	9	64	Burst (I)
$N=96$	35	12	6	2	5	−7	6	36	Burst (II)
(A3-A4) V	34	−20	11	−19	9	−7	9	28	Burst (I)
$N=121$	36	13	9	1	7	−6	7	30	Burst (II)
	51	−11	14	−7	7	−8	6	42	Burst (III)
(F3-F5) V	45	−36	13.5	−16	14	−7	12	26	Burst (IV)
$N=173$									

Notes: N: number of stars. $\bar{U}$, $\bar{V}$, $\bar{W}$: mean velocity components. σ_U, σ_V, σ_W: residual velocity dispersions. SEM standard deviations lie between 1 and $5\,\mathrm{km\,s^{-1}}$ for mean velocities, 1 and $3\,\mathrm{km\,s^{-1}}$ for dispersions, 5 and 10% for percentages of stars belonging to each burst.

Table 4. Velocity distribution function of stars in the detected star formation bursts

Burst	Object	N	$\bar{U}$ (km s^{-1})	$\bar{V}$ (km s^{-1})	$\bar{W}$ (km s^{-1})	$\bar{\sigma}$ (km s^{-1})
I	(B5-A4) V	416	−16	−16	−8	10
	Clusters	6	−12	−19	−8	
II	(B9.5-A4) V	120	12	1	−7	7
	Cluster	1	15	3	−9	
III	(A3-A4) V	51	−11	− 7	−8	10
	Cluster	1	− 2	− 8	−1	
IV	(F3-F5) V	45	−36	−16	−7	13
	Clusters	2	−41	−20	−6	

Notes: N: number of objects. $\bar{U}$, $\bar{V}$, $\bar{W}$: Mean velocity components[a]. $\bar{\sigma}$: Mean residual velocity dispersion[a].
[a] Averaged rms errors lie between 1 and 3 km s^{-1}.

Concerning the second point, we may reason as follows. Group II is not present in the subset (B5-B9) V; this implies that stars of this burst must be older than the age of a B9V star, i.e. 2.7 10^8 yr. For group III a similar reasoning implies a minimum age of 4 10^8 yr (corresponding to an A0 star which left the dwarf stage), and for group IV, 6 10^8 yr (age of an A4 star which left the main sequence stage). If we compare these minimum ages with the ages of the open clusters which we considered to be characteristic for each group, we find (see Table 2) respectively 3 10^8, 4 10^8, and 6.6 10^8 yr – in each case the agreement is excellent.

Scalo (1987) studying the present-day mass function (PDMF) of the solar neighbourhood main sequence stars, observed two peaks, one at spectral type F2-F5 and the other around A0, and a dip between types F0 and A5. He interpreted each "knee" in the PDMF as representing the effect of one past burst of star formation, one burst may consist of more than one burst, since the resolution in mass of the PDMF is poor. The bursts found here are in good agreement with those suggested by the Scalo's results.

3. Concluding remarks

We may conclude that the observed distribution of residual velocities is the sum of several independent (spherical) distributions, each one corresponding to one generation ("burst") of stars. This can be seen easily for early A-type stars; it becomes less visible for F-type stars.

Our results show that the kinematic characteristics of the bursts are still observable after about 2 to 3 10^9 yr, suggesting that the galactic disk is neither well mixed nor relaxed in 10 galactic years.

In particular, the determination of the solar velocity with respect to the circular velocity via the Strömberg relation (see Delhaye, 1965) needs the use of well mixed samples and, in the light of the present work, this implies ages larger than 2 to 3 10^9 yr.

Finally, the estimation of the distribution function of residual velocities is expected to depend on the investigated space volume, since larger volumes should include new bursts.

Acknowledgements. We thank Dr. G. Celeux for his help in the use of the SEM algorithm, M. Bougeard for helpful discussions and A. Sellier and M. Boumghar for their technical assistance. We also thank Dr. J. Lequeux for his useful comments.

References

Bougeard, M., Arenou, F.: 1989, in *Errors, Bias and Uncertainties in Astronomy*, eds. F. Murtagh and C. Jaschek, Strasbourg, 11–14 September
Bougeard, M., Arenou, F., Gómez, A.E.: 1989, Bull. 47th International Statistical Institute, Vol. 11, p. 161, Paris
Celeux, G., Diebolt, J.: 1985, *Computational Statistics Quarterly*, Vol. **2**, 73
Celeux, G., Diebolt, J.: 1986, Revue de Statistiques Appliquées, Vol. XXXIV, n° 2
Delhaye, J.: 1965, in *Stars and Stellar Systems*, 5, ed. A. Blaauw, M. Schmidt, Univ. Chicago Press, Chicago, p. 61
Diebolt, J., Celeux, G.: 1989 (private communication)
Eggen, O.: 1973, *Publ. Astron. Soc. Pacific* **85**, 381
Grenier, S., Gómez, A.E., Jaschek, C., Jaschek, M., Heck, A.: 1985, *Astron. Astrophys.* **145**, 331 (Paper I)
Mermilliod, J.C.: 1981a, *Astron. Astrophys. Suppl. Ser.* **44**, 467
Mermilliod, J.C.: 1981b, *Astron. Astrophys.* **97**, 235
Mermilliod, J.C.: 1986 (private communication)
Palouš, J., Ruprecht, J., Dluzhnevskaya, O.B., Piskunov, T.: 1977, *Astron. Astrophys.* **61**, 27
Scalo, J.M.: 1987, in *Starbursts and Galaxy Evolution*, eds. T.X. Thuan, T. Montmerle, J. Tran Thanh Van, p. 445
Soubiran, C.: 1988, Stage de D.E.A., Observatoire de Paris
Soubiran, C., Gómez, A.E., Arenou, F., Bougeard, M.: 1989, in *Errors, Bias and Uncertainties in Astronomy*, eds. F. Murtagh, C. Jaschek, Strasbourg, 11–14 September

Validation de la séparation en sous-populations

La première question qui vient à l'esprit est la réalité des modes de la distribution des vitesses, que l'on voit en particulier pour la composante U sur la fig. 1 de l'article précédent. On sait bien en effet qu'il est toujours possible de choisir le pas d'un histogramme pour changer son apparence. Comme la présence de plusieurs sous-populations est suggérée par la présence de plusieurs modes, il nous faut donc montrer la validité statistique de ceux-ci.

Prenons par exemple le groupe des étoiles A1 V. Deux modes apparaissent : l'un compris entre -20 et -10 km/s (28 étoiles), l'autre entre 0 et 10 km/s (21 étoiles). Supposons que les étoiles puissent avoir n'importe quelle vitesse entre -60 et 45 km/s, si bien que l'arrivée d'un certain nombre d'étoiles sur un intervalle de 10 km/s soit «poissonnien». La probabilité d'avoir exactement n étoiles sur l'intervalle I est $p_n(I) = e^{-(\lambda I)} \frac{(\lambda I)^n}{n!}$ où $\mu = \lambda I$ est le nombre moyen d'étoiles sur cet intervalle, soit en l'occurence $10\frac{101}{45-(-60)} = 9.62$ étoiles/km/s. La probabilité d'avoir plus de 20 étoiles est $1-\sum_{i=0}^{20} p_i(I) \approx 0.0001$. Par conséquent les deux modes sont significatifs.

De façon plus intéressante, on peut justifier le nombre de sous-populations gaussiennes par un test de Wilks (1963) : on teste l'hypothèse nulle de K composants contre l'hypothèse alternative de $K' > K$ composants. On peut montrer, par exemple, que le nombre significatif de composants gaussiens pour les échantillons A1 V et A2 V est de 2 [Soubiran *et al.*, 1989].

Plus généralement, un certain nombre d'articles (voir par exemple page 86) ont été consacrés à cette séparation de composants gaussiens : Arenou (1990), Arenou & Bougeard (1992), Bougeard & Arenou (1989), Bougeard *et al.* (1989a), Bougeard *et al.* (1989b), Robert & Soubiran (1991), Soubiran *et al.* (1989). En s'attachant plus particulièrement à l'échantillon d'étoiles A2 V, nous avons notamment montré que la séparation des groupes donnait quasiment les mêmes résultats, que l'on utilise des méthodes paramétriques (composants gaussiens) ou non. La présence de deux groupes au comportement cinématique distinct, tout au moins dans cet échantillon, est donc fortement probable.

Néanmoins, cette analyse a été faite en n'utilisant que les données cinématiques et une indication indirecte de l'âge. Nous avons donc décidé de compléter cette étude, d'une part à l'aide de données complémentaires, d'autre part avec une méthode plus discriminante.

7.3 Âge, métallicité et propriétés cinématiques

Les données utilisées

Si l'on reprend la définition de Norris *et al.* (1985), «*une population stellaire est caractérisée par la fonction trivariée, décrivant la distribution des étoiles qui la composent, en ce qui concerne l'âge, la composition chimique et la cinématique*».

Ce qui est vrai pour les populations stellaires de notre Galaxie, l'est également pour des éventuels groupes issus de bouffées de formation : on s'attend à ce que les étoiles d'un même groupe partagent les mêmes caractéristiques cinématiques mais également la composition chimique du gaz dont elles sont toutes issues, et qu'elles aient un âge voisin.

De l'étude précédente il ressort d'ailleurs clairement que l'âge des étoiles est un paramètre discriminant qui permettrait de valider la séparation cinématique effectuée ; pour l'article joint, nous ne disposions que d'une valeur supérieure, liée à la classe spectrale.

Dans le cadre des Actions Intégrées Franco-Espagnole, l'équipe de J. Torra de l'Université de Barcelone s'est chargée de déterminer des âges individuels, lorsque c'était possible [Figueras *et al.*, 1992]. Les paramètres physiques fondamentaux ont été acquis à partir de la photométrie $uvby{-}\beta$, à l'aide de nouvelles observations [Figueras *et al.*, 1991]. Ils ont ensuite déterminé les âges grâce aux séquences évolutives de Maeder et Meynet (1988) réactualisées, correspondant à une composition chimique solaire. Les âges n'étant pas bien déterminés pour les étoiles proches de la ZAMS, nous n'avons qu'une valeur supérieure de l'âge pour ces plus jeunes étoiles et donc une erreur standard élevée.

Comme dans l'échantillon ci-dessus le nombre d'étoiles brillantes qui sont d'un type spectral donné, qui possèdent des données cinématiques et pour lesquelles on peut déterminer un âge est trop faible, nous avons constitué un nouvel échantillon d'étoiles en allant un peu plus loin en magnitude apparente.

L'échantillon que nous allons utiliser dans ce qui suit a été créé à partir des étoiles de la base de données INCA, de type A0-A4, classées spectroscopiquement naines, plus brillantes que $m_V = 8.5$, qui possèdent un mouvement propre, une vitesse radiale, de la photométrie $uvby{-}\beta$, et, bien sûr, une estimation de l'âge. De cet échantillon nous avons naturellement soustrait les étoiles d'amas, pour ne pas biaiser les résultats. Il nous reste 369 étoiles qui répondent à l'ensemble de ces critères.

La photométrie $uvby{-}\beta$ nous permet de déterminer une magnitude absolue, donc une distance photométrique, qui est utilisée pour calculer les vitesses spatiales. Elle nous permet également de disposer d'un indicateur de la métallicité, $\delta m_0 = m_0^{\mathrm{Hyades}} - m_0$ (voir page 24). Avec cet indicateur, δm_0 est nul en moyenne pour les Hyades, positif pour les étoiles déficientes, et vaut environ 0.018 pour le soleil [Crawford, 1975]. À titre indicatif, nous indiquons ci-dessous différentes populations stellaires, telles qu'elles étaient discriminées approximativement en fonction de δm_0, d'après Moon (1985) :

	δm_0	< -0.010	étoiles Am et Ap
$-0.010 <$	δm_0	$< +0.025$	étoiles normales de population I
$+0.025 <$	δm_0	$< +0.045$	étoiles du vieux disque
$+0.045 <$	δm_0	$< +0.090$	étoiles de population II intermédiaire
$+0.090 <$	δm_0		étoiles de population II extrême

Les distributions suivant les composantes U et V de la vitesse ainsi qu'en âge et en δm_0 des étoiles de notre échantillon sont indiquées figure 7.1. L'essentiel des étoiles a un âge compris entre 4×10^8 et 6×10^8 années.

Les distributions des composantes X, Y, Z des positions relatives au Soleil, et la distribution en magnitude apparente m_V de ces étoiles sont représentées sur la figure 7.2. L'échantillon est essentiellement limité en magnitude apparente et les étoiles sont en général plus proches que 200 pc.

Avec l'échantillon que nous venons de définir, et qui est différent de celui qui avait été utilisé pour l'article précédent, nous voulons mettre en évidence plusieurs groupes, et montrer que les vitesses ne sont pas bien mélangées pour les étoiles les plus vieilles de notre échantillon. Si l'on conserve uniquement celles qui ont plus de deux années galactiques

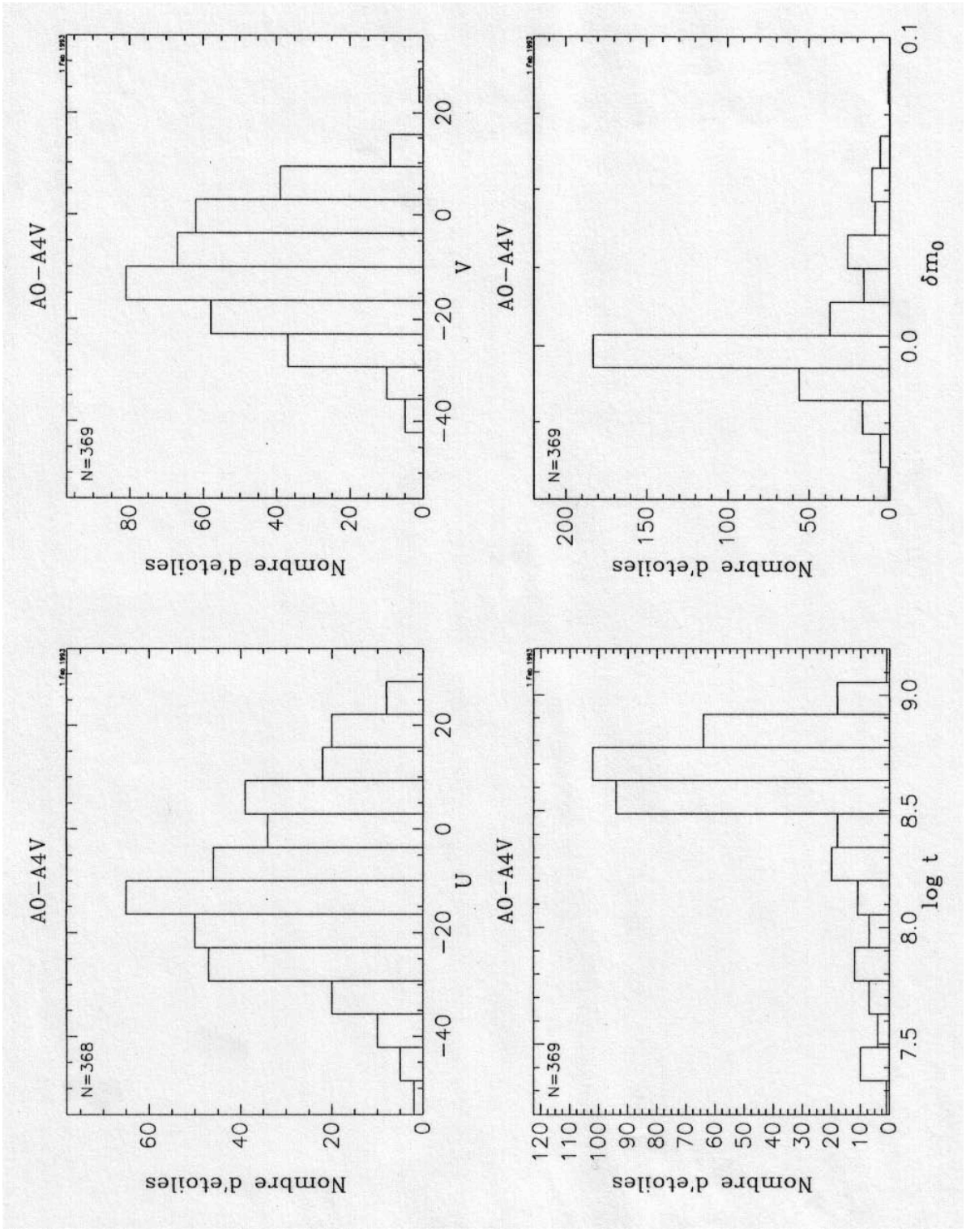


FIG. 7.1: *Distribution dans l'échantillon d'étoiles A0-A4 V des composantes U et V de la vitesse, de l'âge* $\log t$ *et de* δm_0.

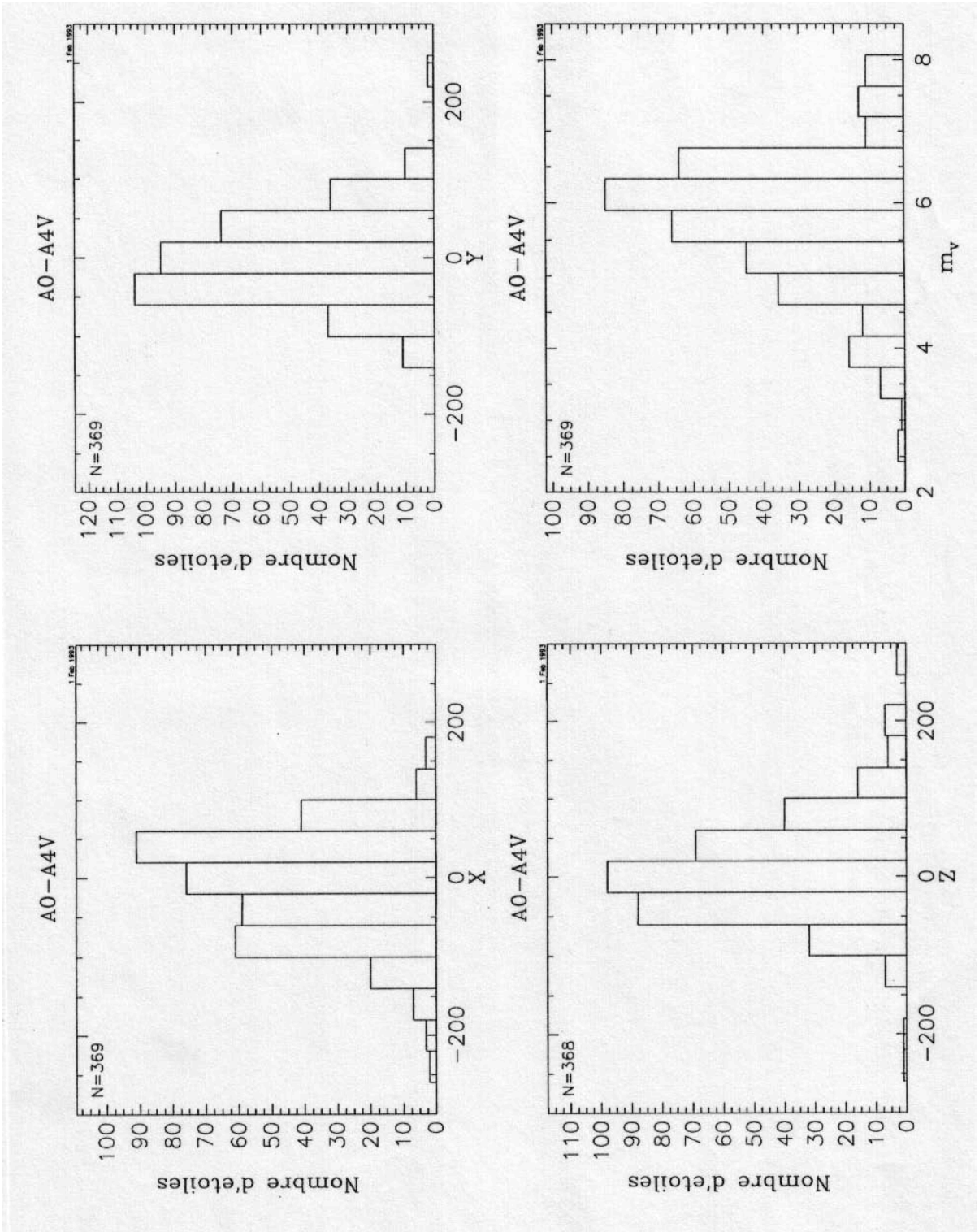


FIG. 7.2: *Distribution des composantes X, Y, Z des positions relatives au Soleil et distribution en magnitude apparente.*

($\log t > 8.7$), on peut montrer que les vitesses ne sont pas distribuées de façon gaussienne : tous les tests de normalité basés sur l'asymétrie rejettent l'hypothèse nulle pour chaque variable U, V et W. En utilisant le test de Lilliefors, l'hypothèse de normalité est rejetée pour U (proba < 0.01) et pour V (proba < 0.006). Ceci indique que la distribution des vitesses n'est pas ellipsoïdale et nous autorise donc à rechercher plusieurs groupes dans notre échantillon.

Variation avec l'âge

Nous pouvons d'abord nous demander si les propriétés cinématiques sont liées à l'âge, ou, en d'autres termes s'il existe une dépendance entre les distributions des vitesses et celle de l'âge, dans l'échantillon étudié.

Nous pouvons répondre positivement à cette question ; en effet, un test bilatéral de Kendall entre la composante V de la vitesse et l'âge t rejette l'hypothèse nulle d'indépendance (probabilité $< 2\ 10^{-6}$). Cette dépendance que l'on vient de mettre en évidence est positive, c'est-à-dire, grossièrement, qu'aux valeurs les plus petites (les plus négatives) de V correspondent les étoiles les plus jeunes. Il n'y a pas de dépendance (monotone) évidente de U et W avec l'âge.

Avant de voir plus précisément comment se relie l'âge, ainsi que la métallicité, aux sous-populations, nous allons introduire les intégrales du mouvement de ces étoiles, et voir ce qu'elles impliquent en terme de survivance des groupes.

7.3.1 Intégrales du mouvement

Pour que les étoiles d'une même bouffée de formation se retrouvent ensemble maintenant, il faut que ces étoiles possèdent des intégrales du mouvement qui soient voisines ; dans le cas contraire, en effet, la cohésion du groupe ne se serait pas maintenue quelques années galactiques plus tard, et les étoiles seraient devenues des étoiles de champ, dispersées tout au long de la rotation galactique.

Nous allons nous placer dans un cadre très simple en supposant que notre Galaxie est un système stationnaire axisymétrique ; utilisant les coordonnées cylindriques (R, θ, Z), cela signifie que la densité dans l'espace des phases $f(x, y, z, u, v, w, t)$ est indépendante du temps et de θ, angle de révolution autour de l'axe Z. La densité de phase f est fonction des intégrales premières isolantes du mouvement [King, 1989, p. 122].

La première intégrale du mouvement est alors celle de l'énergie

$$H = \frac{1}{2}(u^2 + (v + V_0)^2 + w^2) + \phi(R, Z)$$

où $(u, v, w) = (U + u_\odot, V + v_\odot, W + w_\odot)$ désignent les vitesses particulières des étoiles (c'est-à-dire corrigées de la vitesse particulière du soleil) relatives au LSR, et V_0 la vitesse linéaire de rotation du LSR autour du centre galactique. On note R_0 la distance du soleil au centre galactique, R le rayon galactocentrique de l'étoile considérée et $\phi(R, Z)$ le potentiel gravitationnel auquel elle est soumise.

La seconde intégrale du mouvement concerne le moment angulaire

$$h = R(v + V_0) = R(V + v_\odot + V_0).$$

Comme le montrent les observations dans notre Galaxie, il peut exister une troisième intégrale du mouvement [King, 1989], mais la complexité de sa mise en évidence, en

particulier le fait que, dans le cas considéré, on n'en ait pas d'expression analytique, dépasse notre propos.

Le potentiel que nous utilisons est celui de Carlberg & Innanen (1987) s'écrivant à l'aide des coordonnées cylindriques :

$$\phi(R,Z) = -\sum_{j=1}^{4} \frac{\mathcal{M}_j}{\sqrt{(a_j + \sum_{i=1}^{3} \beta_{i,j}\sqrt{Z^2+h_i^2})^2 + b_j^2 + R^2}}$$

$\mathcal{M}_j$ étant la masse de la composante considérée (disque-halo 6.34 10^5, bulbe 2 10^5, noyau 4 10^4, halo obscur 3.205 10^6 km^2.s^{-2}.kpc), b_j le rayon de cette composante (8, 3, 0.25, 35 kpc), a_j étant l'échelle de longueur du disque ($a_1 = 3$ kpc, $a_{j\neq 1} = 0$) et la somme au dénominateur correspondant à trois composantes du disque de différentes échelles de hauteur h_i (vieux disque 0.325, matière obscure 0.09, et jeune disque 0.125 kpc) ; $\beta_{1,1} = 0.4$, $\beta_{1,2} = 0.5$ et $\beta_{1,3} = 0.1$ sont les pondérations pour ces trois composantes du disque-halo, et les autres $\beta_{i,j}$ sont nuls.

Nous adoptons comme vitesse du soleil $(u_\odot, v_\odot, w_\odot) = (9, 12, 7)$ km/s [Delhaye, 1965], $R_0 = 8.5$ kpc comme distance du soleil au centre galactique et $V_0 = 235$ km/s comme vitesse de rotation du LSR [Carlberg & Innanen, 1987].

Le graphique (H, h) (diagramme de Lindblad) pour l'échantillon étudié est tracé figure 7.3. En abscisse l'énergie $-H$ a pour unité 10^3 km^2.s^{-2}, et en ordonnée le moment angulaire est exprimé en 10^2 kpc.km.s^{-1}. La forme caractéristique du diagramme est due au fait que les étoiles n'ont qu'une petite variation autour de la courbe $H = \frac{1}{2R_0^2}h^2 + \phi(R_0, 0)$, puisqu'elles sont proches du soleil et que les vitesses particulières sont petites par rapport à V_0. Les étoiles sont sur des orbites presque circulaires (en fait épicycliques) et c'est essentiellement la vitesse $v + V_0$ qui est discriminante et qui permet que les étoiles se séparent et se retrouvent périodiquement.

La persistance d'un groupe d'étoiles a comme condition nécessaire que les étoiles qui composent ce groupe partagent une énergie et un moment angulaire voisin, et qu'elles soient donc proches sur le diagramme de Lindblad.

La question qui se pose maintenant est : quelle est la dimension critique d'un groupe dans un tel diagramme, dimension au-delà de laquelle les étoiles n'auraient pu rester ensemble ? Il est difficile de répondre directement à une telle question, mais il est au moins possible d'en trouver une limite supérieure. En effet, nous observons déjà de tels groupes : les amas ouverts ; certes, les étoiles qui les composent sont maintenues gravitationnellement, et c'est pourquoi la dimension d'un amas sur le diagramme de Lindblad représente une limite supérieure.

Par ordre d'âge croissant, on a donc tracé sur les figures 7.4, 7.5 et 7.6 les diagrammes de Lindblad correspondant respectivement aux Pléiades, à Coma Ber et aux Hyades. On notera que l'on a sur chaque diagramme un groupe serré et des étoiles un peu dispersées qui peuvent s'interpréter soit comme des étoiles qui sont en train de s'échapper de l'amas, soit comme des étoiles de champ n'appartenant pas à l'amas considéré. Bien qu'établi sous des hypothèses restrictives, ce graphique pourrait s'avérer utile à la fois pour détecter des étoiles non-membres et pour rechercher des étoiles de champ ex-membres de l'amas.

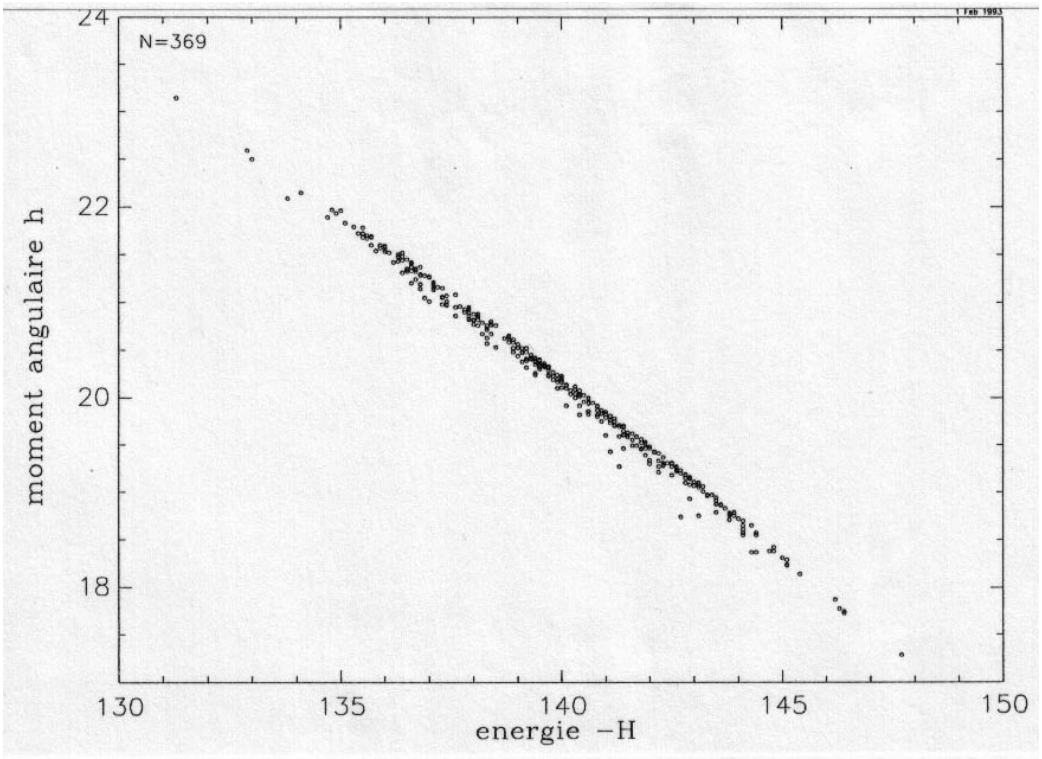


FIG. 7.3: *Diagramme de Lindblad de l'échantillon A0-A4 V.*

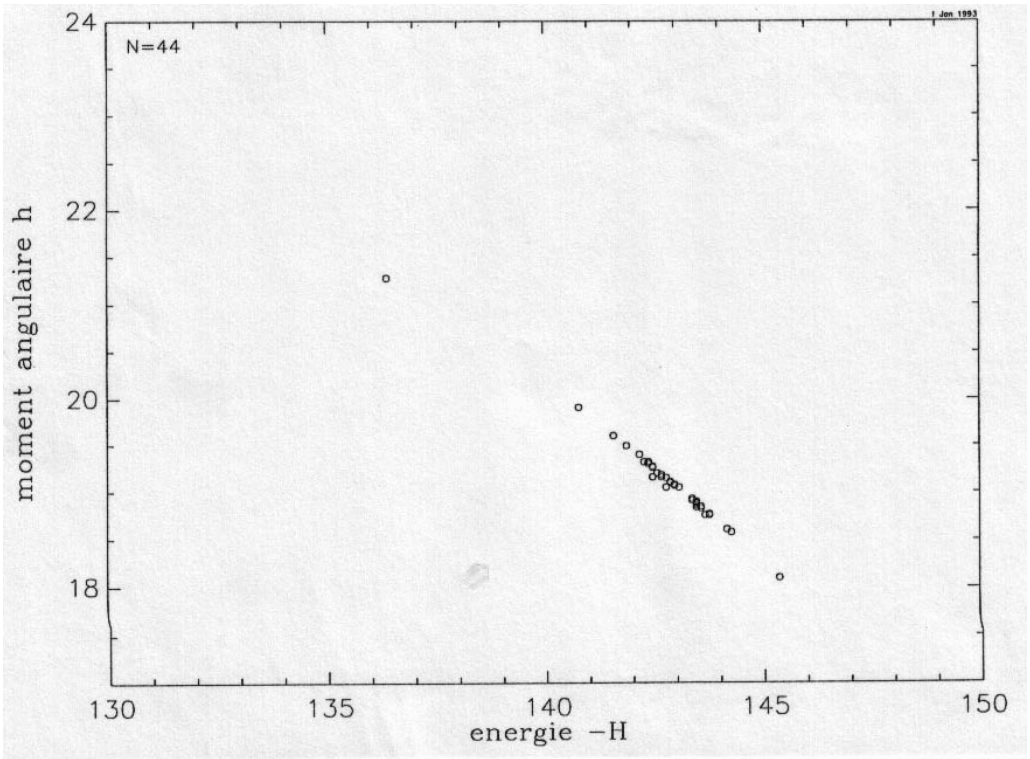


FIG. 7.4: *Diagramme de Lindblad des Pléiades.*

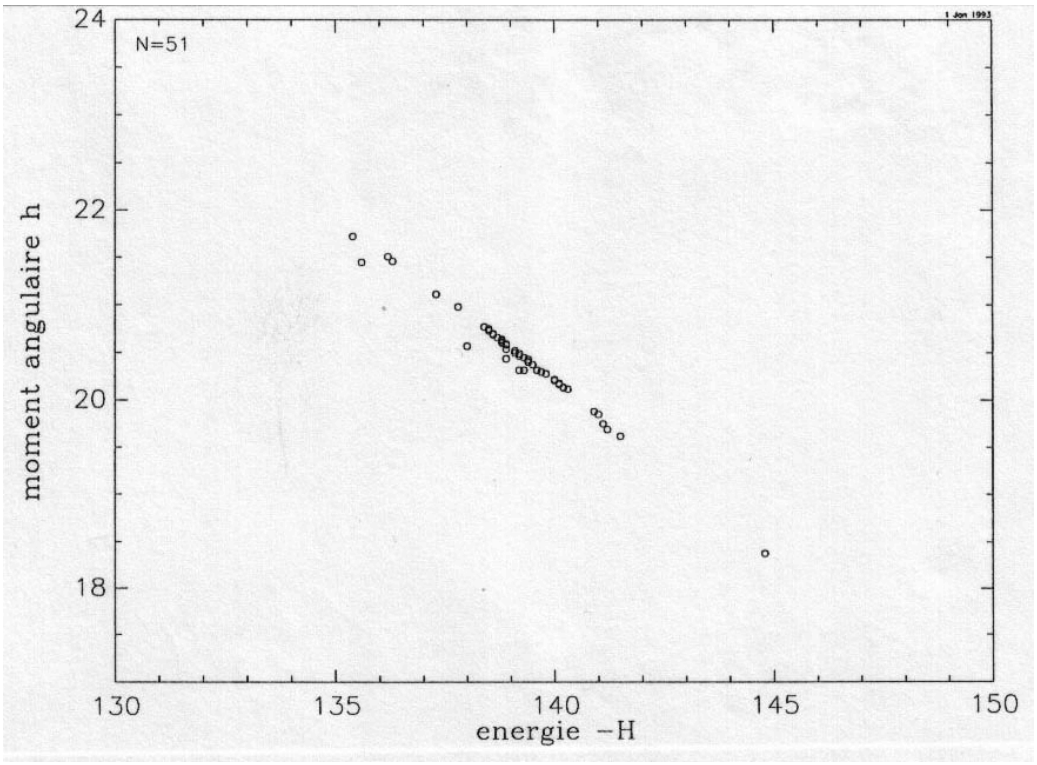


FIG. 7.5: *Diagramme de Lindblad de Coma.*

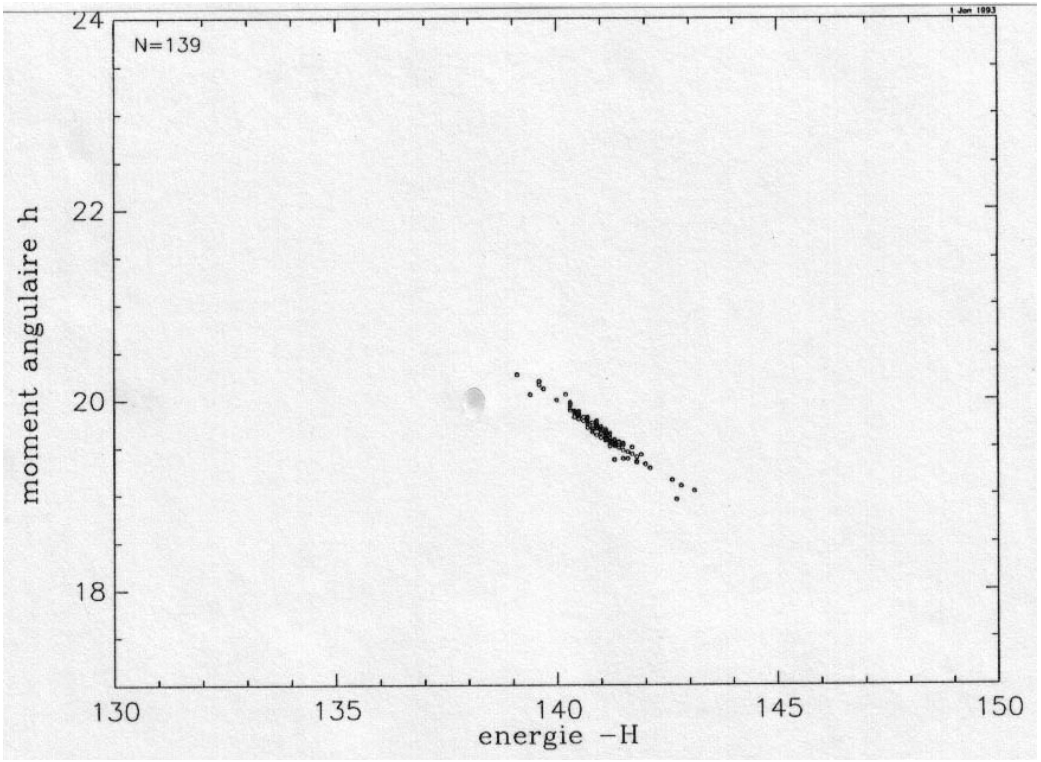


FIG. 7.6: *Diagramme de Lindblad des Hyades.*

Les tailles respectives des amas fournissent une contrainte sur la taille d'un groupe ayant une origine commune. Si l'on supposait que chaque étoile de notre échantillon appartient à un tel groupe, il est clair, à la vision des différentes figures, qu'il faudrait un mimimum de trois groupes dans notre échantillon.

Si l'on raisonne uniquement sur le moment angulaire, on peut noter également que si les étoiles sont ensemble maintenant, en étant séparées par une distance inférieure à $\delta R = \pm 250$ pc près, cela représente une différence de moment angulaire $\delta h \approx V_0 \delta R \approx \pm 0.6\ 10^5\ \mathrm{pc}^2/(10^6\ \mathrm{ans})$. Si l'on observe la plage de variation de h, cela implique également un minimum de trois groupes dans l'échantillon.

7.3.2 Séparation des groupes

Nous allons discriminer des groupes d'étoiles qui possèdent des caractéristiques cinématiques voisines, un âge proche et une indication de la métallicité semblable. Nous allons donc nous servir des variables U, V, $\log t$ et δm_0 pour caractériser ces groupes. Nous n'utiliserons pas W à cause de son pouvoir trop peu discriminant : on a vu en effet plus haut qu'elle était de distribution gaussienne et indépendante de l'âge.

Pour séparer les groupes, nous avons tout d'abord utilisé le programme SEMMUL de séparation de composants gaussiens, en tenant compte des erreurs de mesures, décrit et testé au §4.4.1, et nous avons obtenu quatre composants. On peut objecter que, si l'on s'attend à des groupes distribués de façon gaussienne (sphérique) en ce qui concerne les composantes des vitesses, c'est beaucoup plus contestable pour l'indicateur de la métallicité δm_0 et surtout pour l'âge, $\log t$. De plus, pour quatre composants en dimension quatre, nous devons déterminer 59 paramètres à l'aide de 369 points, ce qui conduit à des solutions peu stables.

Compte-tenu de ces remarques, nous n'utiliserons pas cette méthode mais plutôt une méthode de classification classique non (explicitement) paramétrique, une classification ascendante hiérarchique. Les classes sont obtenues en minimisant la variance intra-groupe [Murtagh & Heck, 1987]. Nous avons conservé les quatre premiers groupes de la hiérarchie, de façon à conserver un nombre significatif d'étoiles dans chaque groupe.

Le groupe n°1 est séparé des groupes n°2 et n°3 par U et dans une moindre mesure par V. Mais le groupe 3 est principalement déterminé à partir de $\log t$: il s'agit des étoiles les plus jeunes. Le groupe n°4 est dû à δm_0, et ce sont les étoiles les plus déficientes. Ces groupes sont représentés pour chaque couple de variables sur la figure 7.7 ; les unités des axes sont des nombres d'écart-type car les variables ont été centrées et réduites avant classification.

De plus, une analyse en composantes principales (fig. 7.8) indique que toutes les variables sont bien représentées sur le premier axe (41% de la variance) et contribuent de manière égale à sa formation. Le second axe (31%) oppose U et V d'une part à $\log t$ et δm_0 d'autre part ; toutes les variables y sont également bien représentées. On peut noter que les variables U et V sont corrélées ($\rho = 0.45$) et que les variables $\log t$ et δm_0 sont également corrélées ($\rho = 0.45$). Le premier axe s'est positionné entre ces deux groupes de variables, ce qui suggère une rotation des axes : la seconde diagonale représente en effet la séparation entre un groupe jeune à composition normale (n°3) et un groupe ancien déficient (n°4) ; quant à la première diagonale, elle représente la séparation de notre échantillon en deux groupes (n°1 et n°2) de caractéristiques cinématiques différentes.

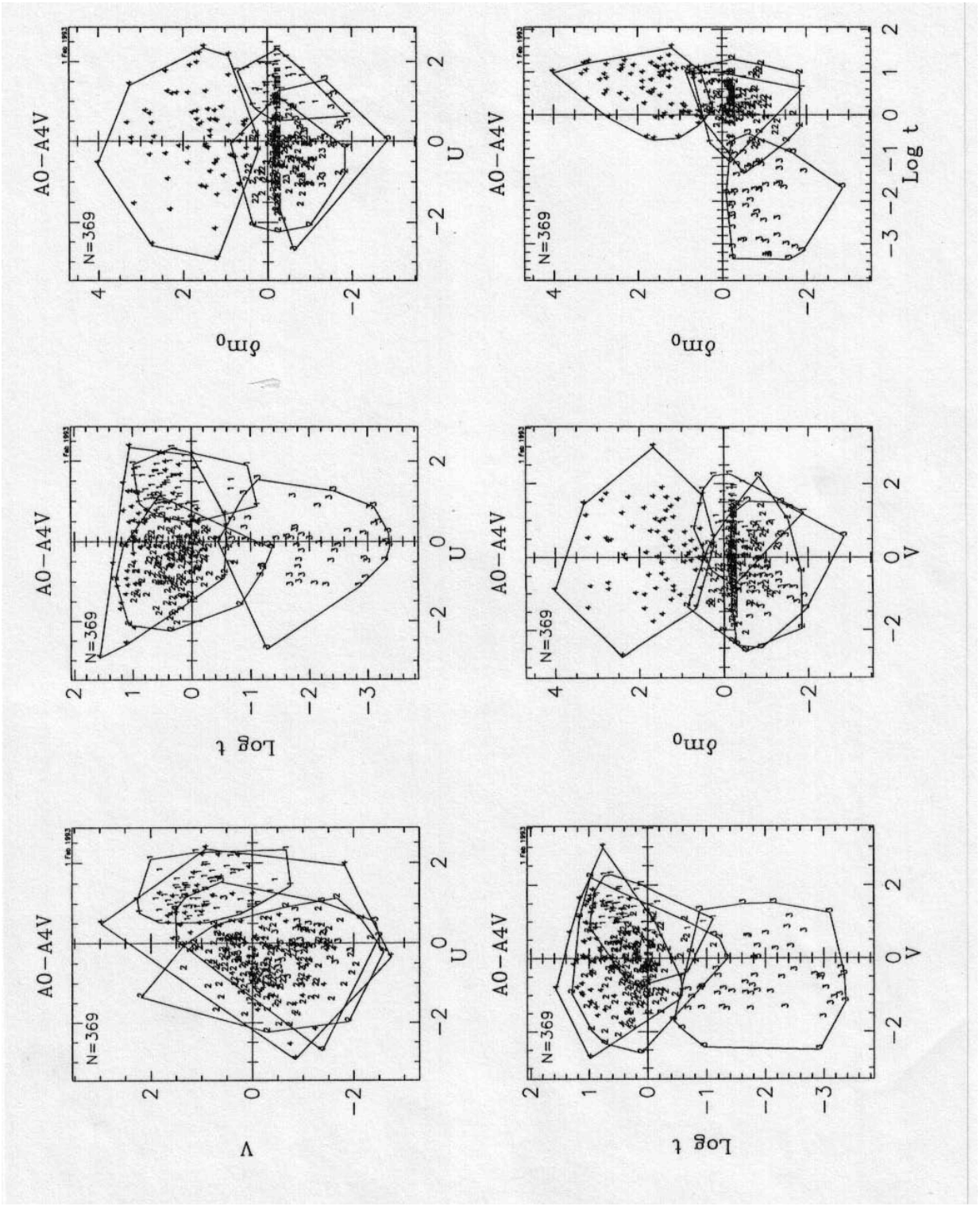


FIG. 7.7: *Position des 4 groupes déterminés, pour chaque couple des variables* U, V, $\log t$ *et* δm_0.

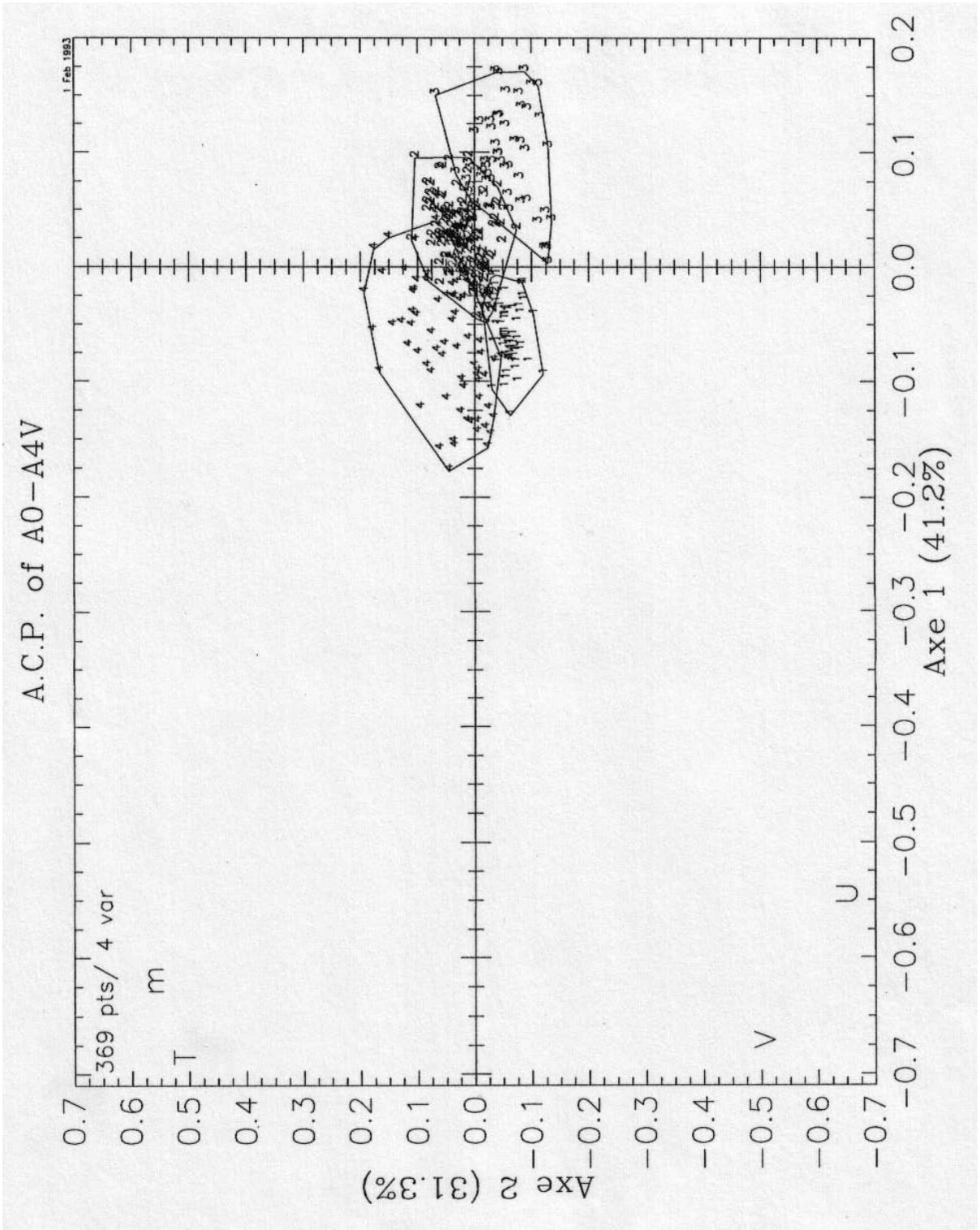


FIG. 7.8: *Projection suivant les deux premiers axes de l'analyse en composantes principales de l'échantillon A0-A4 V avec les variables U, V, $T = \log t$ et $m = \delta m_0$.*

Le groupe le plus particulier est le quatrième. En effet, il semble le plus agé et le plus déficient, et les vitesses semblent mélangées. En réalité, ce groupe a des δm_0 beaucoup trop élevés. Si on l'étudie plus attentivement, on s'aperçoit que la projection de la vitesse de rotation suivant la ligne de visée ($V \sin i$) est en moyenne 177 ± 10 km/s pour ce groupe, significativement différente de 136 ± 6 km/s, valeur moyenne pour les étoiles des autres groupes. Cette dernière valeur est tout à fait ordinaire pour des étoiles A V, alors que la vitesse est beaucoup trop grande pour le groupe 4, et correspond à celle d'étoiles B5 V [Jaschek & Jaschek, 1987].

Ceci tendrait à faire penser que les étoiles de ce groupe sont en réalité plus jeunes. À ceci s'ajoute le fait que l'indice δc_0 est corrélé avec $V \sin i$ (voir §2.4.5) et que, si l'on n'en tient pas compte pour la calibration de la magnitude absolue à l'aide de la photométrie $uvby-\beta$, on a tendance à donner une magnitude absolue trop brillante, dans notre cas sans doute de quelques dizièmes de magnitudes. Il se trouve que les valeurs des $V \sin i$ n'ont pas été utilisées pour le calcul des âges des étoiles de notre échantillon, et ceci implique que les âges des étoiles du groupe n°4 sont plus petits que ceux calculés [Figueras *et al.*, 1992].

Enfin, un dernier indice d'anormalité dans ce groupe est la présence de plusieurs étoiles de type λ Boo (au moins 4 étoiles). Ce type d'étoiles, caractérisé par une faible métallicité, est passé inaperçu jusqu'à une période récente, et ces étoiles pourraient bien être plus nombreuses qu'on ne le pense. Jaschek & Jaschek (1987) notent que ces étoiles ont une petite vitesse radiale et une vitesse rotationnelle élevée. Selon Gerbaldi *et al.* (1992), ces étoiles sont au-dessus de la séquence principale non pas parce qu'elles en viennent, mais parce qu'elles y arrivent Toujours en ce qui concerne l'estimation de l'âge, les trajets évolutifs qui ont été utilisés sont valables pour une composition chimique solaire, alors que ce groupe est apparemment déficient. Les âges que nous utilisons pour ces étoiles pourraient alors être totalement erronés.

Clairement, il y a là un sujet de recherche à explorer, dont nous nous préoccuperons dans le futur. Pour le moment, contentons-nous de noter que les âges et les vitesses spatiales calculées (qui utilisent la distance photométrique) sont sans doute inadéquats, et qu'il y a peut être un mélange de populations aux caractéristiques cinématiques différentes. Ceci implique que le groupe 4 est bien particulier et que l'on ne peut pas s'en servir pour en tirer des conséquences au sujet de la distribution des vitesses avec l'âge.

Le tableau 7.1 indique pour chaque groupe, le numéro du groupe, le nombre d'étoiles, la valeur la plus probable de la distribution des $\log t$, la moyenne des δm_0, suivis des moyennes et des dispersions des composantes U et V de la vitesse, et du rapport de ces dispersions. Le rapport $\frac{\sigma_V}{\sigma_U}$ montre que les distributions sont approximativement sphériques, sauf pour le dernier groupe dont on a déjà indiqué la particularité.

Nous laisserons de côté le groupe 3, trop jeune pour ce qui nous intéresse, et le groupe 4 pour les raisons mentionnées plus haut. La classification que nous avons effectuée nous permet donc d'avoir les groupes 1 et 2, mieux délimités en âge et en métallicité. Ces deux groupes ont des caractéristiques cinématiques très nettement différentes.

La cinématique du groupe 1 est très voisine de celle du groupe U Ma, mais avec un âge plus élevé. Il est vrai que le cas de U Ma est un peu particulier dans la mesure où son âge varie suivant les auteurs, de $2.7\ 10^8$ années [Eggen, 1983] à $4.9\ 10^8$ années [Paloŭs & Hauck, 1986]. Nous trouvons pour notre groupe 1 un âge proche de cette dernière estimation ($\log t \approx 8.71$ contre 8.69). La différence d'âge entre Eggen et

TAB. 7.1: *Caractéristiques des groupes trouvés dans l'échantillon A0-A4 V.*

gr	N	$\langle \log t \rangle$	$\overline{\delta m_0}$	$\overline{U}$	σ_U	$\overline{V}$	σ_V	$\frac{\sigma_V}{\sigma_U}$
3	57	7.87	$-.012_{\pm.001}$	$-11.8_{\pm1.6}$	$11.6_{\pm2.1}$	$-16.2_{\pm1.3}$	$11.1_{\pm2.0}$	$0.96_{\pm.24}$
2	167	8.62	$-.003_{\pm.001}$	$-19.2_{\pm0.8}$	$11.0_{\pm0.6}$	$-13.7_{\pm0.7}$	$9.3_{\pm0.5}$	$0.85_{\pm.07}$
1	70	8.71	$-.001_{\pm.001}$	$+11.0_{\pm0.8}$	$7.2_{\pm0.6}$	$+\ 0.6_{\pm0.8}$	$6.7_{\pm1.1}$	$0.93_{\pm.11}$
4	75	? 8.88	$+.034_{\pm.002}$	$-8.6_{\pm2.0}$	$17.7_{\pm1.4}$	$-8.6_{\pm1.7}$	$11.7_{\pm0.96}$	$0.66_{\pm.17}$

Paloŭs et Hauck peut provenir du fait que le premier a considéré [Fe/H]$= -0.1$ pour U Ma et les seconds ont utilisés des trajets évolutifs différents et valables pour des étoiles de composition solaire.

La réalité de groupes avec des caractéristiques cinématiques différentes est bien établie. Nous avons en effet trouvé quasiment les mêmes groupes en utilisant la méthode paramétrique SEMMUL avec gestion des erreurs, ci-dessus mentionnée, sauf le groupe le plus jeune, parce que l'erreur de mesure sur ces étoiles est très grande. De plus, nous avons également effectué une classification à l'aide des trois variables $-H$, $\log t$ et δm_0, et nous avons obtenu quatre groupes de composition semblables à ceux mis en évidence ci-dessus. Ceci dit, d'une part les variables $\log t$ et δm_0 ne sont sans doute pas assez discriminantes, et d'autre part, les données cinématiques ont encore des erreurs de mesure importantes, et ceci ne permet pas de séparer plus adéquatement les groupes.

Cependant, les résultats obtenus ne permettent pas de valider ou de rejeter l'hypothèse que ces groupes représentent la signature d'anciennes bouffées de formation. En fait, tels qu'ils sont définis, les groupes sont un peu trop étalés sur le diagramme de Lindblad pour pouvoir répondre aux critères mentionnés au §7.3.1 ; ils montrent néanmoins des dispersions de vitesse assez faibles, spécialement pour le groupe 1, constitué vraisemblablement pour partie d'ex-membres d'U Ma.

Pour pouvoir réellement conclure, il faudra attendre les parallaxes et les mouvements propres d'Hipparcos ; des programmes d'observations sont également en cours, qui permettront d'obtenir des nouvelles vitesses radiales. Tout ceci devrait permettre d'ici quelques années non seulement d'augmenter la taille des échantillons mais également la précision des données cinématiques.

On notera que les restes de quelques groupes distincts aux caractéristiques cinématiques différentes fournissent une contribution certaine à la déviation du vertex : la superposition des groupes 1 et 2 suffit en effet à empêcher l'ellipsoïde des vitesses de notre échantillon d'être aligné en direction du centre galactique.

L'utilisation de données complémentaires, ainsi que les méthodes d'analyse multivariée nous ont donc permis de mettre en évidence la présence de groupes homogènes en âge et

composition, mais aux caractéristiques cinématiques différentes, dont semble être formé notre échantillon. La réalité du groupe 1 étant claire, nous pouvons en déduire qu'au terme de deux années galactiques les vitesses ne sont pas encore bien mélangées dans le voisinage solaire. Ceci implique qu'il n'est pas possible d'appliquer la relation du courant asymétrique, pour déterminer la vitesse particulière du soleil, sans un critère d'âge plus précis que la simple classification spectrale : un minimum de 10^9 ans semble être requis pour les étoiles servant à déterminer la vitesse particulière du soleil.

Conclusion

Cette thèse a été consacrée aux nombreuses données stellaires qui ont servi au Catalogue d'Entrée d'Hipparcos, et aux données préliminaires du satellite. Nous avons décrit des méthodologies et les avons appliquées sur des domaines très vastes, allant de l'estimation de l'absorption interstellaire à la cinématique des étoiles proches, en passant par la validation des parallaxes d'Hipparcos.

Après avoir décrit l'élaboration et la gestion de la base de données INCA, nous avons détaillé les calibrations photométriques et spectroscopiques que nous avons implantées. Ceci nous a notamment servi à estimer les parallaxes photométriques et spectroscopiques après avoir corrigé les biais qui les entachent.

Pour cela nous avons d'abord développé une partie statistique, où nous avons essentiellement étudié l'influence des erreurs de mesure sur les données. Nous en avons déduit des estimations de densité de probabilité, et des estimations optimales (non biaisées et de variance minimum) de paramètres de distributions dont les variables sont affectées d'erreur ; nous avons également tiré profit des possibilités de l'estimation conditionnelle. Nous avons utilisé ces estimations pour l'étude des parallaxes d'Hipparcos et également pour séparer des mélanges gaussiens. Nous avons, de plus, décrit les simulations effectuées et les tests statistiques que nous avons souvent été amenés à utiliser.

Nous avons ainsi développé ou implanté un certain nombre d'outils logiciels portant à la fois sur les domaines astronomiques, statistiques et graphiques.

Nous avons ensuite réalisé un modèle de l'absorption interstellaire dans le visible qui permet d'avoir une approximation réaliste de l'extinction, de manière tridimensionnelle, jusqu'à plusieurs centaines de parsecs du Soleil. Cette modélisation s'avère utile pour beaucoup d'études statistiques qui concernent les étoiles dans notre Galaxie ; à titre d'exemples d'application, citons l'élaboration d'un modèle de la Galaxie ou la calibration de magnitudes absolues.

Il faut noter que, jusqu'à présent, il fallait pour cela se contenter de relations très approximatives, qui ne tenaient pas compte des irrégularités de la distribution de la matière interstellaire dans notre Galaxie.

Ayant participé à la création du Catalogue d'Entrée d'Hipparcos, nous avons été naturellement amené à comparer les premières données astrométriques et photométriques du satellite avec le contenu du Catalogue d'Entrée. Ces comparaisons ont permis de mettre en évidence la bonne qualité des données préliminaires du satellite et du Catalogue d'Entrée. Nous avons notamment démontré que, mis à part quelques problèmes mineurs sur les données au sol, qu'il faudra naturellement explorer dans le futur, la précision des positions

et des magnitudes du Catalogue d'Entrée est très supérieure aux spécifications initiales de l'Agence Spatiale Européenne.

Nous avons également proposé des méthodologies qui pourront permettre de valider les résultats futurs de la mission Hipparcos, en ce qui concerne les parallaxes. Nous avons appliqué ces méthodes aux résultats préliminaires du satellite et fait la preuve de la qualité de ces résultats. Sur ces données préliminaires, nous avons utilisé plusieurs méthodes d'estimation des erreurs externes des parallaxes des deux consortiums de réduction des données, une estimation du point-zéro global de ces parallaxes, et attiré l'attention sur les différents écueils auxquels il faudra s'attendre lorsqu'il s'agira de déterminer les parallaxes définitives. Nous avons, de plus, étudié la variation des erreurs systématiques en fonction de divers paramètres. L'ensemble forme une chaîne de programmes qui permettent de vérifier rapidement la qualité des parallaxes à chaque solution intermédiaire.

Tout ceci nous a permis d'indiquer comment les parallaxes des deux consortiums pourront être combinées de manière à obtenir la meilleure estimation de chaque parallaxe définitive, ainsi que son erreur standard.

Finalement, nous avons étudié la cinématique des étoiles naines A au voisinage du Soleil. Nous avons montré, en utilisant des méthodes d'analyse multivariée, que, contrairement à ce qu'il est courant de supposer, le temps de mélange des vitesses spatiales est nettement supérieur à deux rotations galactiques. Pour cela nous avons mis en évidence le fait que la fonction de distribution locale des vitesses pouvait plus adéquatement être expliquée par la contribution de quelques groupes d'étoiles. Ceci a notamment pour conséquence de prohiber l'utilisation d'étoiles plus jeunes que 10^9 ans environ pour le calcul de la vitesse particulière du Soleil.

Tous les sujets que nous avons traités dans cette thèse seront des grands bénéficiaires des futures données d'Hipparcos. La cartographie de l'absorption pourra être nettement mieux connue à l'aide des données photométriques de l'expérience Tycho, et des parallaxes d'Hipparcos. Ces dernières permettront également d'obtenir des magnitudes absolues individuelles avec une excellente précision et d'améliorer, avec de meilleures données mais également avec des méthodes statistiques adéquates, les calibrations des magnitudes absolues. Enfin, les mouvements propres fournis par le satellite contribueront à l'augmentation et l'amélioration des données cinématiques des étoiles, et permettront de mieux connaître la dynamique de notre Galaxie.

C'est donc l'ensemble des thèmes abordés dans cette thèse qui pourront bientôt profiter de prolongements marquants, grâce au satellite Hipparcos. Nous serons amené à y travailler dans le futur, puisque nous sommes impliqué dans différents programmes de recherche proposés pour l'exploitation scientifique des données préliminaires et définitives.

Puissent au moins ces quelques pages, jetant un pont entre le passé (le Catalogue d'Entrée) et le futur (les résultats d'Hipparcos), avoir témoigné un tant soit peu du succès de la mission et de l'enthousiasme d'y participer.

Cinquième partie

ANNEXES, TABLES

Annexe A

Bibliothèques de programmes informatiques

C'est une évidence de dire que l'outil informatique est indispensable pour le moindre calcul, la moindre simulation, le moindre traitement de données.

Bien que je sois informaticien, il a été peu question de logiciels dans les pages précédentes, puisqu'il s'agit d'une thèse d'Astronomie. Il fallait pourtant trouver un endroit pour placer les mots «langage C», «Unix», et autres mots à la mode.

Cette annexe présente donc deux librairies de logiciels qui ont été écrits pour l'occasion puis utilisés régulièrement et qui peuvent servir à nouveau. De nombreux autres programmes ont naturellement été élaborés pour des besoins plus ponctuels, et qui ne sont donc pas cités ici.

Il faut donc insister sur le fait que les traitements de données et les développements logiciels ont été grandement facilités par l'utilisation de stations de travail (pour la rapidité) sous Unix (pour la richesse du système d'exploitation).

A.1 Bibliothèque astronomique

L'ensemble de ces fonctions et programmes représente environ 3500 lignes de code C.

- Positions et vitesses :
  - Conversion d'équatorial en écliptique ou galactique et réciproquement, calcul de la précession ;
  - Calcul des positions ou des vitesses spatiales et de leurs erreurs associées ;
  - Format des coordonnées : interne-externe ;
  - Photocentre, géocentre, séparation et angle de position de deux étoiles doubles ;
- Magnitudes et couleurs :
  - Calcul d'absorption, distance, $(B-V)$,... par modélisation de l'absorption ;
  - Conversion de magnitudes et couleurs de Johnson en Tycho ;
  - Paramètres fondamentaux par la photométrie $uvby{-}\beta$;
  - Magnitude intégrée de plusieurs étoiles ;

- Types spectraux :
  - Magnitude absolue en fonction du type spectral ;
  - $(B - V)$ intrinsèque en fonction du type spectral ;
  - Codage interne-externe ;

A.2 Bibliothèque statistique

L'ensemble de ces fonctions et programmes représente environ 5000 lignes de code C.

- Calculs sur des distributions : loi de Cauchy, de Fisher-Snédécor, log-normale, normale, de Poisson, de Student et uniforme :
  - Densités de probabilité ;
  - Probabilités ;
  - Quantiles ;
  - Tirages de nombres pseudo-aléatoires ;
- Tests statistiques :
  - Test de Kolmogorov-Smirnov et du χ^2 entre un échantillon et une distribution théorique ou entre deux échantillons ;
  - Tests de normalité de Lilliefors, basé sur l'aplatissement ou sur l'asymétrie ;
- Estimations :
  - De centre : empiriques avec ou sans bootstrap, mode, médiane, tronqué à $< 3\sigma$, à $n\%$, pondéré ;
  - De dispersion : empirique, valeur absolue, largeur à base de quantiles, pondéré ;
  - De densités de probabilité ;
  - De variables sans erreurs ;
  - Séparation (méthode EM, SEM, ...) de composants gaussiens avec gestion des erreurs ;
- Analyse multivariée :
  - Analyse en composantes principales avec ou sans gestion des erreurs ;
  - Classification ascendante hiérarchique ;

Annexe B

Publications

Multivariate mixture distributions in stellar kinematics : Statistical and numerical stability of the SEM algorithm
Bougeard M.L., Arenou F., 1989, dans : *Errors, bias and uncertainties in Astronomy*, eds. F.Murtagh et C.Jaschek, Cambridge University Press, p. 277.

Séparation de mélanges gaussiens unidimensionnels en statistique stellaire : les méthodes graphiques
Bougeard M.L., Arenou F., Soubiran C., Grenier S., 1989, dans : *Errors, bias and uncertainties in Astronomy*, eds. F.Murtagh et C.Jaschek, Cambridge University Press, p. 281.

Analysis of the mixture of normal distributions in stellar kinematics : maximum likelihood, bootstrap and wilks test
Soubiran C., Bougeard M.L., Gómez A.E., Arenou F., 1989, dans : *Errors, bias and uncertainties in Astronomy*, eds. F.Murtagh et C.Jaschek, Cambridge University Press, p. 407.

Comparative approach parametric and nonparametric classification methods for small samples in stellar kinematics
Bougeard M.L., Arenou F., Gómez A.E., 1989, Bulletin 47$^{\text{ème}}$ International Statistical Institute, contrib. papers, vol I, p. 161, Paris.

The INCA Database
Gómez A., Morin D., Arenou F., 1989, *The Hipparcos Mission*, ESA–SP 1111, vol.II, 23.

Local kinematic properties of Population I (B5-F5)-type stars and galactic disk evolution
Gómez A.E., Delhaye J., Grenier S., Jaschek C., Arenou F., Jaschek M., 1990, *Astron. Astrophys. Lett.* **236**, 95.

Méthodes d'analyse multivariée
Arenou F., 1990, *Journée bisontine d'astrométrie et de photométrie*, 6 février 1990, Besançon.

Modèle tridimensionnel de l'extinction interstellaire
Arenou F., 1991, *Des données de Hipparcos à la détermination des paramètres stellaires fondamentaux*, 7-8 février 1991, Strasbourg.

The Hipparcos INCA Database
Turon C., Arenou F., Baylac M.-O., Boumghar D., Crifo F., Gómez A., Marouard M., Mekkas M., Morin D., Sellier A., 1991, dans *Databases & On-line Data in Astronomy*, eds. M.A. Albrecht et D. Egret, p. 67.

The Hipparcos Input Catalogue: I. Star selection
Turon C., Gómez A., Crifo F., Crézé M., Perryman M. A. C., Morin D., Arenou F., Nicolet B., Chareton M., Egret D., 1992, *Astron. Astrophys.* **258**, 74.

The observing programme. Performances of the Input Catalogue
Turon C., Arenou F., Crifo F., Gómez A., Morin D., 1992, dans *Highlights of Astronomy*, Vol. 9, ed. J. Bergeron, p. 388.

The printed version of the Hipparcos Input Catalogue
Turon C., Arenou F., Crifo F., Gómez A., Marouard M., Morin D., Sellier A., 1992, dans *Highlights of Astronomy*, Vol. 9, ed. J. Bergeron, p. 397.

Reliability of the Hipparcos Input Catalogue tested by the 'First look'
Crifo F., Gómez A., Arenou F., Morin D., Schriver H., 1992, *Astron. Astrophys.* **258**, 116.

Comparison of the first results from the Hipparcos star mappers with the Hipparcos Input Catalogue
Turon C., Arenou F., Evans D.W., van Leeuwen F., 1992, *Astron. Astrophys.* **258**, 125.

A tridimensional model of the galactic interstellar extinction
Arenou F., Grenon M., Gómez A., 1992, *Astron. Astrophys.* **258**, 104.

The Hipparcos Input Catalogue
Turon C., Crézé M., Egret D., Gómez A., Grenon M., Jahreiß H., Réquième Y., Argue A.N., Bec-Borsenberger A., Dommanget J., Mennessier M.O., Arenou F., Chareton M., Crifo F., Mermilliod J.C., Morin D., Nicolet B., Nys O., Prévot L., Rousseau M., Perryman M.A.C., 1992, 1992, *The Hipparcos Input Catalogue*, ESA–SP 1136, 7 volumes.

Hipparcos et les distances galactiques
Arenou F., Gómez A., 1992, *Distancia'92*, Congrès international de statistique sur l'analyse en distance, eds. S. Joly et G. Le Calve, p. 313.

The performance of the Hipparcos Input Catalogue as compared with the first results of the Hipparcos mission
Turon C., Arenou F., Froeschlé M., van Leeuwen F., Lindegren L., Mignard F., Morin D., Perryman M.A.C., 1992, dans *Astronomy from Large Databases II*, eds A. Heck & F. Murtagh, 135.

Comparing Parametric and Nonparametric Statistical Methods for studying the Velocity Distributions of Population I Stars
Arenou F., Bougeard M.L., 1992, *Developments in Astrometry and Their Impact on Astrophysics and Geophysics*, I.A.U. symp. **156**, eds. Mueller & Kolaczek.

Bibliographie

[Aïvazian *et al.*, 1986] Aïvazian, S., Énukov, I., Méchalkine, L., 1986, *Eléments de modélisation et traitement primaire des données*, traduction française, ed. Mir, Moscou.

[Arellano *et al.*, 1990] Arellano Ferro, A., Parrao, L., 1990, *Astron. Astrophys.* **239**, 205.

[Arenou, 1990] Arenou, F., 1990, *Journée bisontine d'astrométrie et de photométrie*, 6 février 1990, Besançon.

[Arenou & Bougeard, 1992] Arenou, F., Bougeard M.L., 1992, *Developments in Astrometry and Their Impact on Astrophysics and Geophysics*, I.A.U. symp. **156**, eds. Mueller & Kolaczek.

[Arenou *et al.*, 1992] Arenou, F., Grenon, M., Gómez, A., 1992, *Astron. Astrophys.* **258**, 104.

[Arenou & Morin, 1988] Arenou, F., Morin, D., 1988, *Astronomy from large databases*, ESO Conference, ed. Murtagh F. & Heck A., 269.

[Balona *et al.*, 1984] Balona, L.A., Shobbrook, R.R., 1984, *Mon. Not. R. Astron. Soc.* **211**, 375.

[Barbier, 1989] Barbier-Brossat, M., 1989, *Astron. Astrophys. Suppl. Ser.* **80**, 67.

[Blauw, 1963] Blauw, A., 1963, *Basic Astronomical data*, ed. K.A. Strand, Univ. of Chicago Press, p. 383.

[Bosq & Lecoutre, 1987] Bosq, D., Lecoutre, J.-P., 1987, *Théorie de l'estimation fonctionnelle*, ed. Economica, Paris.

[Bougeard & Arenou, 1989] Bougeard, M.L., Arenou, F., 1989, dans : *Errors, bias and uncertainties in Astronomy*, eds. F.Murtagh et C.Jaschek, Cambridge University Press, p. 277.

[Bougeard *et al.*, 1989a] Bougeard, M.L., Arenou, F., Soubiran, C., Grenier, S., 1989, dans : *Errors, bias and uncertainties in Astronomy*, eds. F.Murtagh et C.Jaschek, Cambridge University Press, p. 281.

[Bougeard *et al.*, 1989b] Bougeard, M.L., Arenou, F., Gómez, A. E., 1989, 47ème session de l'Institut International de Statistiques 29 août-6 septembre 1989, Paris.

[Brotzman & Gessner, 1991] Brotzman, L. E., Gessner, S. E., 1991, *Selected Astronomical Catalogs*, publié sur CD-ROM, Vol. 1.

[Carlberg & Innanen, 1987] Carlberg, R. G., Innanen, K. A., 1987, *Astron. J.* **94**, 666.

[Casertano *et al.*, 1990] Casertano, S., Ratnatunga, K. U., Bahcall, J. N., 1990, *Astrophys. J.* **357**, 435.

[Celeux & Diebolt, 1986] Celeux, G., Diebolt, J., 1986, *R.S.A.*, **34**, n2, 35.

[Celeux & Diebolt, 1989] Celeux, G., Diebolt, J., 1989, Une version de type recuit-simulé de l'algorithme EM, *Rapport de recherche INRIA*, **1123**.

[Corbally & Garrison, 1984] Corbally, C. J., Garrison, R. F., 1984, in *The MK process and Stellar Classification*, proceedings of a workshop, Toronto, june 1983, R. F. Garrison ed., p. 277.

[Crawford, 1975] Crawford, D. L., 1975, *Astron. J.* **80**, 955.

[Crawford & Mandewala, 1976] Crawford, D. L., Mandewala, N., 1976, *Publ. Astron. Soc. Pac.* **88**, 917.

[Crawford, 1978] Crawford, D. L., 1978, *Astron. J.* **83**, 48.

[Crawford, 1979] Crawford, D. L., 1979, *Astron. J.* **84**, 1858.

[Crézé *et al.*, 1989] Crézé, M., Nicolet, B., Chareton, M., 1989, 1989, *The Hipparcos Mission*, ESA–SP 1111, vol.II, 47.

[Crifo, 1988] Crifo, F., 1988, Proc. Sitges Coll. "Scientific Aspects of the Input Catalogue Preparation II", 3-7 June 1985, Turon, C., & Perryman, M. A. C., 79.

[Crifo *et al.*, 1992] Crifo, F., Gómez A., Arenou F., Morin D., Schriver H., 1992, *Astron. Astrophys.* **258**, 116.

[Dawson, 1990] Dawson, P.C., 1990, *J. Roy. Astron. Soc. Can.* **84**, 3.

[Delhaye, 1965] Delhaye, J., 1965, *Galactic Structure*, eds. A. Blauw & M. Schmidt, Univ. of Chicago Press, p. 61.

[Diebolt & Celeux, 1989a] Diebolt, J., Celeux, G., 1989, communication privée.

[Diebolt & Celeux, 1989b] Diebolt, J., Celeux, G., 1989, communication privée.

[Diebolt & Robert, 1990] Diebolt, J., Robert, C., 1990, Bayesian estimation of finite mixture distributions, Rapports techniques 110 et 111, LSTA, Universite Paris VI.

[Dommanget, 1967] Dommanget, J., 1967, *Catalogue d'Ephémérides*, Comm. de l'Obs. Royal de Belgique, Série B, n° 15.

[Dommanget, 1989] Dommanget, J., 1989, 1989, *The Hipparcos Mission*, ESA–SP 1111, vol.II, 149.

[Dommanget, 1992] Dommanget, J., 1992, communication privée.

[Dommanget & Nys, 1982] Dommanget, J., Nys, O., 1982, *Second Catalogue d'Ephémérides*, Comm. Obs. Royal de Belgique, Série B, n° 124.

[Donati *et al.*, 1989] Donati, F., Canuto, E., Belforte, P., Carlucci, D., van Leeuwen, F., 1989, *The Hipparcos Mission*, ESA–SP 1111, Vol. III, 73.

[Dudewicz & Mishra, 1988] Dudewicz, E. J., Mishra, S. N., 1988, *Modern Mathematical Statistics*, ed. John Wiley & sons, New York.

[Egret *et al.*, 1992] Egret ,D., Didelon, P., McLean, B.J., Russell, J.L., Turon, C., 1992, *Astron. Astrophys.* **258**, 217.

[Eggen, 1965] Eggen, O. J., 1965, *Galactic Structure*, eds. Blauw & Schmidt, Univ. Chicago press, p 111.

[Eggen, 1983] Eggen, O. J., 1983, *Astron. J.* **88**, 642.

[Evans, 1992] Evans, D. W., 1992, *Background variations in the IDT Photometric Reductions*, document interne RGO/NDAC 92.01.

[Figueras *et al.*, 1991] Figueras, F., Torra, J., Jordi, C., 1991, *Astron. Astrophys. Suppl. Ser.* **87**, 319.

[Figueras *et al.*, 1992] Figueras, F., Torra, J., Jordi, C., Asiain, R., 1992, *IAU Colloquium* **138**.

[Froeschlé, 1992a] Froeschlé, M., 1992, communication privée.

[Froeschlé, 1992b] Froeschlé, M., 1992, *Comparison between sphere solutions by FAST and NDAC*, rapport interne FAST.

[Froeschlé, 1992c] Froeschlé, M., 1992, *Hipparcos : le géomètre du ciel*, L'Astronomie, Juin-Juillet-Août 1992.

[Gerbaldi *et al.*, 1992] Gerbaldi, M., Zorec, J., Castelli, F., Faraggiana, R., 1992, *IAU Colloquium* **138**.

[Gliese, 1969] Gliese, W., 1969, *Veröff. Astron. Rechen-Inst. Heidelberg* **22**.

[Gliese *et al.*, 1986] Gliese, W., Jahreiß, H., Upgren, A. R., 1986, *The Galaxy and the solar system*, eds. Smoluchowski, Bahcall & Matthews, the University of Arizona Press.

[Gómez *et al.*, 1990] Gómez, A. E., 1990, Delhaye, J., Grenier, S., Jaschek, C., Arenou, F., Jaschek, M., 1990, *Astron. Astrophys. Lett.* **236**, 95.

[Gómez, 1992] Gómez, A. E., 1992, *IAU Colloquium* **137**, April 92, Vienna.

[Gray, 1991] Gray, R.O., 1991, *Astron. Astrophys. Suppl. Ser.* **252**, 237.

[Grenier *et al.*, 1985] Grenier, S., Gómez, A. E., Jaschek, C., Jaschek, M., Heck, A., 1985, *Astron. Astrophys.* **145**, 331.

[Grenon, 1989] Grenon, M., 1989, *The Hipparcos Mission*, ESA–SP 1111, Vol. III, p. 205.

[Grenon *et al.*, 1992] Grenon, M., Mermilliod, M., Mermilliod, J. C., 1992, *Astron. Astrophys.* **258**, 88.

[Guarinos, 1991] Guarinos, J., Thèse de l'Observatoire de Paris, 1991.

[Guthrie, 1987] Guthrie, B.N.G., 1987, *Mon. Not. R. Astron. Soc.* **226**, 361.

[Hakkila, 1989] Hakkila, J.,1989, *Astrophys. J.* **346**, 932.

[Hauck & Mermilliod, 1990] Hauck, B., Mermilliod, M., 1990, *Astron. Astrophys. Suppl. Ser.* **86**, 107.

[Hilditch *et al.*, 1983] Hilditch, R.W., Hill, G., Barnes, J.V., 1983, *Mon. Not. R. Astron. Soc.* **204**, 241.

[Høg *et al.*, 1992] Høg, E., Bastian, U., Grewing, M., Halbwachs, J.L., Wicenec, A., Bässgen, Bernacca, P.L., Donati, F., Kovalevsky, J., van Leuwen, F., Lindegren, L., Pedersen, C., Scales, D. R., Snijders, M.A.J., Wesselius, P.R., 1992, *Astron. Astrophys.* **258**, 177.

[Houk, 1988] Houk, N., 1988, *Michigan Catalog of Two Dimensional Spectral Types for the HD stars*, vol 4, Ann Arbor.

[Huber, 1981] Huber, P., 1981, *Robust Statistics*, ed. Wiley, New York.

[Jahreiß *et al.*, 1992] Jahreiß, H., Réquième, Y., Argue, A. N., Dommanget, J., Rousseau, M., Lederle, T., Le Poole, R. S., Mazurier, J. M., Morrison, L. V., Nys, O., Penston, M. J., Périé, J. P., Prévot, L., Tucholke, H. J., de Vegt, C., 1992, *Astron. Astrophys.* **258**, 82.

[Jaschek & Gómez, 1985] Jaschek, C., Gómez, A. E., 1985, *Astron. Astrophys.* **146**, 387.

[Jaschek & Mermilliod, 1984] Jaschek, C., Mermilliod, J. C., 1984, *Astron. Astrophys.* **137**, 358.

[Jaschek & Jaschek, 1987] Jaschek, C., Jaschek, M., 1987, *The classification of stars*, Cambridge University Press.

[Johnson & Morgan, 1953] Johnson, H. L., Morgan, W. W., 1953, *Astron. J.* **117**, 313.

[King, 1989] King, I. R., 1989, *The milky way as a galaxy*, ed. Buser & King, Geneva Observatory, p. 117.

[Lacey, 1984] Lacey, G. C., 1984, *Mon. Not. R. Astron. Soc.* **208**, 687.

[Lebeaux, 1986] Lebeaux, M. - O., 1986, *Manuel de référence SAS - ADDAD*, CIRCE, Orsay.

[Lecoutre & Tassi, 1987] Lecoutre, J.-P., Tassi, P., 1987, *Statistique non paramétrique et robustesse*, ed. Economica, Paris.

[Lindegren, 1985] Lindegren, L., 1985, "Scientific Aspects of the Input Catalogue Preparation", ESA SP-234, 31.

[Lindegren, 1989] Lindegren, L., 1989, *The Hipparcos Mission*, ESA–SP 1111, Vol. III, 311.

[Lindegren, 1992a] Lindegren, L., 1992, rapport interne NDAC, 13/07/1992.

[Lindegren, 1992b] Lindegren, L., van Leeuwen F., Petersen, C., Perryman M. A. C., Söderhjelm, S., 1992, *Astron. Astrophys.* **258**, 138.

[Lindegren, 1992c] Lindegren, L., 1992, communication privée.

[Lindegren & Kovalevsky, 1992] Lindegren, L., and Kovalevsky, J. 1992, dans *Highlights of Astronomy*, Vol. 9, ed. J. Bergeron, p.???.

[Luri *et al.*, 1992] Luri, X., Mennessier, M.O., Torra, J., Figueras, F., 1992, *Distancia'92*, Congrès international sur analyse en distance, eds. S. Joly et G. Le Calve, p. 123.

[Lutz, 1979] Lutz, T. E.,1979, *Mon. Not. R. Astron. Soc.* **189**, 273.

[Lutz & Kelker, 1973] Lutz, T. E., Kelker, D. H., 1973, *Publ. Astron. Soc. Pac.* **85**, 573.

[Lyngå, 1987] Lingå, G., 1987, *Catalogue of open cluster data*, available through CDS, Strasbourg, France.

[Maeder & Meynet, 1988] Maeder, A., Meynet, G., 1988, *Astron. Astrophys. Suppl. Ser.* **76**, 411.

[Malmquist, 1920] Malmquist, K.G., 1920, *Lund Astron. Obs. Medd.*, Ser. II, **22**.

[Malmquist, 1936] Malmquist, K.G., 1936, *Meddel. Stockholm Obs.* **26**.

[Mennessier *et al.*, 1992] Mennessier, M-O., Barthès, D., Boughaleb, H., Figueras, F., Mattei, J. A., 1992, *Astron. Astrophys.* **258**, 99.

[Mermilliod, 1992] Mermilliod, J-C., 1992, *Astronomy from large databases II*, Haguenau, ed. Murtagh F. & Heck A., 373.

[Mignard *et al.*, 1992] Mignard, F., Froeschlé, M., Falin, J.L., 1992, *Astron. Astrophys.* **258**, 142.

[Moon, 1985] Moon, T.T., 1985, *Communications from the University of London Observatory* n° **78**.

[Murtagh & Heck, 1987] Murtagh,F., Heck, A., 1987, *Multivariate Data Analysis*, Reidel.

[Norris *et al.*, 1985] Norris, J., Bessell, M.S., Pickles, A.J., 1985, *Astrophys. J. Suppl. Ser.* **58**, 463.

[Oblak *et al.*, 1976] Oblak, E., Considère, S., Chareton, M., 1976, *Astron. Astrophys. Suppl. Ser.* **24**, 69.

[Olsen, 1984] Olsen, E.H., 1984, *Astron. Astrophys. Suppl. Ser.* **57**, 443.

[Olsen, 1988] Olsen, E.H., 1988, *Astron. Astrophys.* **189**, 173.

[Paloŭs & Piskunov, 1984] Paloŭs, J., Piskunov, A. E., 1984, *Astron. Astrophys.* **143**, 102.

[Paloŭs & Hauck, 1986] Paloŭs, J., Hauck, B., 1986, *Astron. Astrophys.* **162**, 54.

[Pelat, 1989] Pelat, D., 1989, *Cours Bruit et Signaux*, DEA d'Astrophysique et Techniques Spatiales, Université Paris 7.

[Perryman & Hassan, 1989] Perryman, M. A. C., Hassan, H., 1989, *The Hipparcos Mission*, ESA–SP 1111, Vol. I, *The Hipparcos satellite*.

[Perryman & Turon, 1989] Perryman, M. A. C., Turon, C., 1989, *The Hipparcos Mission*, ESA–SP 1111, Vol. II, *The Input Catalogue*.

[Perryman *et al.*, 1989] Perryman, M. A. C., Lindegren, L., Murray, C. A., Høg, E., Kovalevsky, J., 1989, *The Hipparcos Mission*, ESA–SP 1111, Vol. III, *The Data Reductions*.

[Perryman *et al.*, 1992] Perryman, M. A. C., Høg, E., Kovalevsky, J., Lindegren, L., Turon, C., Bernacca, P. L., Crézé, M., Donati, F., Grenon, M., Grewing, M., van Leeuwen, F., van der Marel, H., Murray, C. A., Le Poole, R. S., Schrijver, H., *Astron. Astrophys.* **258**, 1.

[Press, 1990] Press, W. H., et al., 1990, *Numerical Recipes in C, The Art of Scientific Computing*, Cambridge University Press, Cambridge.

[Prévot, 1992] Prévot, L., 1992, 1992, *The Hipparcos Input Catalogue*, ESA–SP 1136, Vol. 7, Annex 4, 1.

[Ramberg *et al.*, 1979] Ramberg, J. S., Dudewicz, E. J., Tadikamalla, P. R., Mykytka, E. F., 1979, *Technometrics* **21**, 201.

[Ratnatunga & Casertano, 1991] Ratnatunga, K. U., Casertano, S., 1991, *Astron. J.* **101**, 1075.

[Redner & Walker, 1984] Redner, R., Walker, H., 1984, Mixture densities, maximum likelihood and the EM algorithm, *SIAM Rev*, **26**, 195.

[Robert & Soubiran, 1991] Robert, C., Soubiran, C., 1991, Estimation of a mixture model through Bayesian sampling and Prior Feedback, Rapport technique 138, LSTA, Universite Paris VI.

[Robert, 1992] Robert, C., 1992, *L'analyse statistique bayésienne*, Économica, Paris.

[Scales *et al.*, 1992] Scales, D. R., Snijders, M. A. J., Andreasen, G. K., Grenon, M., Grewing, M., Høg, E., van Leuwen, F., Lindegren, L., Mauder, H., 1992, *Astron. Astrophys.* **258**, 217.

[Schmidt-Kaler, 1965] Schmidt-Kaler, T., 1965, in *Landolt-Börnstein*, New Ser., Group 6, K. H. Hellwege ed., p. 284.

[Schmidt-Kaler, 1982] Schmidt-Kaler, T., 1982, in *Landolt-Börnstein*, vol. VI/2b, K. Schaifers & H. H. Voigt eds., p. 1.

[Schuster *et al.*, 1989] Schuster, W.J., Nissen, P.E., 1991, *Astron. Astrophys.* **221**, 65.

[Smart, 1968] Smart, W.M., 1968, *Stellar Kinematics*, ed. Longmans.

[Smith, 1985] Smith, H., 1985, *Astron. Astrophys.* **152**, 413.

[Smith, 1987a] Smith, H., 1987, *Astron. Astrophys.* **171**, 336.

[Smith, 1987b] Smith, H., 1987, *Astron. Astrophys.* **171**, 342.

[Smith, 1987c] Smith, H., 1987, *Astron. Astrophys.* **188**, 391.

[Smith, 1987d] Smith, H., 1987, *Astron. Astrophys.* **188**, 233.

[Smith, 1988] Smith, H., 1988, *Astron. Astrophys.* **198**, 365.

[Soubiran, 1988] Soubiran, C., 1988, *Analyse et séparation de mélanges de distributions gaussiennes - Application à la cinématique stellaire*, Mémoire de D.E.A d'Astronomie Statistique et Dynamique de l'Observatoire de Paris.

[Soubiran *et al.*, 1989] Soubiran, C., Bougeard, M.L., Gómez, A., Arenou, F., 1989, dans : *Errors, bias and uncertainties in Astronomy*, eds. F.Murtagh et C.Jaschek, Cambridge University Press, p. 407.

[Strömgren, 1966] Strömgren, B., 1966, *Ann. Rev. Astron. Astrophys.* **4**, 433.

[Tassi, 1989] Tassi, P., 1989, *Méthodes statistiques*, ed. Économica, Paris, 2ème édition.

[Trumpler & Weaver, 1953] Trumpler, R. J., Weaver, H. F., 1953, *Statistical Astronomy*, Dover publications, New York.

[Turon Lacarrieu & Crézé, 1977] Turon Lacarrieu, C., Crézé, M.,1977, *Astron. Astrophys.* **56**, 273.

[Turon *et al.*, 1989] Turon, C., Gómez, A., Crifo, F., Grenon, M., 1989, 1989, *The Hipparcos Mission*, ESA–SP 1111, vol II, p.7.

[Turon, 1992] Turon, C., 1992, ESA Bulletin **69**, p.36.

[Turon *et al.*, 1992] Turon, C., Arenou, F., Evans, D.W., van Leeuwen, F., 1992, *Astron. Astrophys.* **258**, 125.

[Upgren & Carpenter, 1977] Upgren, A. R., Carpenter, K. G., 1977, *Astron. J.* **82**, 227.

[van Altena *et al.*, 1991] van Altena, W. F., Truen-Liang Lee, J., Hoffleit, E. D., 1991, *The General Catalogue of Trigonometric Stellar Parallaxes*, preliminary version, voir Brotzman et Gessner.

[Westerlund *et al.*, 1988] Westerlund, B.E., Garnier, R., Lundgren, K., Pettersson, B., Breysacher, J., 1988, *Astron. Astrophys. Suppl. Ser.* **76**, 101.

[Wielen, 1977] Wielen, R., 1977, *Astron. Astrophys.* **60**, 263.

[Wilks, 1963] Wilks, S., 1963, *Mathematical Statistics*, Wiley.

[Wooley *et al.*, 1970] Wooley, R., Sir, Epps, E.A., Penston, M.J., and Pocock, S.B., 1970, Herstmonceux R. Obs. Ann., **5**.

[Zhang, 1983] Zhang, E.H., 1983, *Astron. J.* **88**, 825.

Liste des acronymes

ACP	Analyse en Composantes Principales	83
CCDM	Catalogue des Composantes d'étoiles doubles et Multiples	19
CDS	Centre de Données astronomiques de Strasbourg	10
DRC	Data Reduction Consortia	5
EM	Estimation - Maximisation	80
ESA	European Space Agency	8
ESOC	European Space Operations Centre	21
FAST	Fundamental Astronomy by Space Techniques	20
GCTSP	General Catalog of Trigonometric Stellar Parallaxes	137
HD	Henry Draper Catalogue	9
HIPPARCOS	HIgh Precision PARallax COllecting Satellite	7
H-R	Hertzsprung-Russel	25
INCA	INput CAtalogue	8
mas	millième de secondes d'arc	7
LSR	Local Standard of Rest	183
MK	Morgan-Keenan	9
NDAC	Northern Data Analysis Consortium	20
SEM	Stochastique-Estimation-Maximisation	80
SEMMUL	Stochastique-Estimation-Maximisation-Multidimensionnel	80
SGBD	Système de Gestion de Base de Données	10
SIMBAD	Set of Identifications, Measurements and Biblio. of Astron. Data	10
TCL	Théorème Central Limite	62
UBV	Ultraviolet-Bleu-Visible	29
ZAMS	Zero Age Main Sequence	25

Table des figures

Liste des tableaux

Version 1.2 du 21 Juin 1993 – Écrit avec LaTeX
Version 1.3 du 10 octobre 2005: les images et articles ont été scannés et incorporés, puis pdftex+hyperref ont été utilisés pour produire un document pdf.

Sixième partie

ENGLISH SUMMARY

Annexe E

Contribution to the statistical validation of Hipparcos data: the Input Catalogue and the preliminary data

Note: the part, chapter, section, table and figure numbers refer to the French version. The French version is also needed for the references and the annexes.

Introduction

In order to get ready for the arrival of the Hipparcos data, it is important to learn some methodologies, in particular statistical ones.

Since we were involved in the preparation of the Hipparcos Catalogue d'Entrée, it was natural to take part in the validation of the preliminary Hipparcos data. A good knowledge of the ground-based data and of the preliminary data was also an advantage to prepare the work on the observation proposals.

This thesis is divided into four parts. The first one deals with the data at our disposal, coming mainly from the INCA Data Base. We then describe the $uvby{-}\beta$ photometric calibrations we use and finally we describe the interstellar extinction model which we have built.

The second part is devoted to some statistical tools, for instance how to avoid biases arising from the measurement errors. These tools are used in the third part to compare the ground-based data with the preliminary data coming from the Hipparcos satellite. In particular, we develop a method allowing to obtain the external errors and the instrumental zero-point when the final parallaxes are available.

Finally the last part is devoted to a kinematic study of dwarf A-type stars in the solar neighbourhood.

FROM OBSERVATIONAL DATA TO PHYSICAL PARAMETERS

Chapter 1 briefly describes the ground-based data (included into the INCA Data Base) and the data coming from the satellite.

Chapter 2 concerns calibrations of $uvby{-}\beta$ photometry which we implemented and which give us individual absolute magnitude and colour excess. The photometric absolute magnitudes are then compared with the spectroscopic mean absolute magnitudes in order to know the dispersion of the latter.

Chapter 3 presents a tridimensional model of interstellar extinction which allows us to obtain the absorption and the colour excess of the stars with reasonable accuracy. Although the model was built for prediction of colour excess for some thousands of stars for the Hipparcos mission, it is in fact of a more general use.

E.1 Observational data

E.1.1 The Hipparcos mission

Due to the Hipparcos satellite operation work, the list of stars needed to be known in advance, with the correct observing time for each star. As a consequence, the positions needed to be more accurate than 1.5 arcsec and the H_p magnitudes more precise than 0.5 mag.

E.1.2 The stars observed by Hipparcos

The stars observed by Hipparcos were selected on scientific criteria. From the 600 000 stars received from the astronomical community, the work of the INCA Consortium was to build a Catalogue of about 120 000 stars with their accurate positions and magnitudes.

The proposals

The data given by Hipparcos will have a lot of consequences: the satellite will provide positions accurate enough to build a precise reference system. Proper motions will be useful for dynamical and kinematical studies in our Galaxy. The major astrophysical results will probably come from the parallaxes, not only because the improvement of the cosmic distance scale, but also because of direct distance determination will be available for stars for which, up to now, only indirect estimations were available. This will improve our knowledge of luminosities, radii, ages and masses.

This is why so many types of stars were proposed to be observed by Hipparcos (table 1.1).

Besides the 214 proposals, a supplementary list of stars was needed for the attitude control of the satellite and the Data Reduction task: the "Survey' '.

The Survey

The "Survey' ' is a list of stars brighter than magnitude

$$\left| \begin{array}{lcll} V_{\rm lim} & = & 7.3 + 1.1 \sin |b| & \text{if spectral type is later than G5} \\ V_{\rm lim} & = & 7.9 + 1.1 \sin |b| & \text{otherwise} \end{array} \right. \tag{E.1}$$

This list is not really complete, because of systematic errors of visual apparent magnitudes, random errors (which scatter into the sample more stars than will be scattered out), errors or inhomogeneities of their spectral type; 6% of the stars were not observed because of observational constraints.

The way the Survey was built may be explained: the cut-off in spectral type was used to exclude red giants, in favour of nearby stars which will have a more precise parallax. The dependence with the galactic latitude is used to have a somewhat uniform distribution of stars on the sky.

E.1.3 The INCA database

The data needed for the Catalogue d'Entrée were managed in a very efficient way by the INCA Data Base (DB), which uses the SIMBAD structure and software.

The INCA DB was created as a sub-base of SIMBAD, and then evolved independently. A lot of new measurements were added, concerning astrometry, photometry, variability or multiplicity of stars.

Some Catalogues introduced into the INCA DB were not directly used for the Catalogue d'Entrée, but for our work: the 4th volume of the Michigan Spectral Survey was used for our interstellar extinction model (chap. 3); Hauck & Mermilliod $uvby{-}\beta$ Catalogue was used to obtain photometric absolute magnitudes (chap. 2); radial velocities were used in the part IV to compute the space velocities of the stars.

The included paper [Arenou & Morin, 1988] describes in detail the data and the software needed for the Catalogue d'Entrée preparation.

E.1.4 The Hipparcos Catalogue d'Entrée

The Catalogue d'Entrée was created after some iterations using the INCA DB. Some specific attention was paid to double or multiple systems. These objects, although astronomically interesting, were sources of difficulties for the Catalogue d'Entrée task, for the observation by the satellite and for the Data Reduction task.

We may notice that, excluding the Survey stars, the stars in the Catalogue d'Entrée do not form an homogeneous sample; they only represent the best compromise between different scientific interests and observational constraints.

E.1.5 The results of the Data Reduction Consortia

The sphere solution gives the 5 astrometric parameters, with their associated formal covariances. After a one-year mission, the two Consortia, FAST and NDAC, obtained

a sphere solution, and compared their solutions. In order to validate independently the parallaxes, an external comparison is clearly also needed.

Two solutions were obtained, one (5P) giving the 5 astrometric parameters for each star, the other (3P) using the proper motions coming from the Catalogue d'Entrée.

Some external comparisons were done (chap. 6), keeping only one solution for each Consortium: FAST-3P and NDAC-5P.

E.2 The fundamental parameters of the stars through $uvby{-}\beta$ photometry

The $uvby{-}\beta$ photometry allows us to obtain the physical parameters of the stars. We use it to obtain the reddening and the visual absolute magnitude, in order to evaluate photometric parallaxes. In this case the precision is better than what is obtained with spectroscopic parallaxes, because photometric absolute magnitudes take into account effects of evolution, metallicity, and in some cases rotation of the stars.

We chose this photometry (instead of other photometric systems) because of the high number of measurements available and because it is well fitted to the stars we use to compare the preliminary Hipparcos parallaxes (dwarf, giants and supergiants B, A, F).

E.2.1 The $uvby{-}\beta$ photometry

There are a lot of existing calibrations of physical parameters from $uvby{-}\beta$ photometry, corresponding to different groups in the HR diagram. What we have done is to choose the "best' " calibrations, implement the calibrations, resolve automatically the conflicts at the boundaries of the different calibrations and get an estimation of the standard errors on the resulting parameters.

Calibrations are described in detail in the annex, page 33.

E.2.2 Tests of the calibrations

As a first test of the implementation of these calibrations, we represent an HR diagram of the stars having $uvby{-}\beta$ photometry in the INCA DB (fig 2.1).

We also compare the absolute magnitudes obtained with these calibrations with the absolute magnitudes deduced from distance moduli of some open clusters (fig. 2.2).

E.2.3 Photometric and spectroscopic absolute magnitudes

We also could compare the absolute magnitude with the mean absolute magnitudes coming from spectral classifications. In fact, if the sample used is not perfectly defined (ie complete in distance or in magnitude), no useful conclusion can be drawn from such a comparison, with the exception of the dispersions around the mean absolute magnitudes of the different spectral types. These dispersions will be useful to determine the size of the Malmquist bias, §6.3.

Spectroscopic absolute magnitudes

Several absolute magnitude calibrations can be used. We denote here M^S, M^C, M^G the calibrations of M_V as a function of spectral type and luminosity class from Schmidt-Kaler (1982), Corbally & Garrison (1984), and Grenier *et al.* (1985), respectively. In the last one the Malmquist bias is explicitely taken into account.

It is difficult to choose within these calibrations which is the "best' '. Using another calibration of Schmidt-Kaler (1982) giving the absolute magnitude (which we denote M^{Sc}) as a function of $(B-V)_0$ and luminosity class, Guarinos (1991) compared the different calibrations. As, for a given spectral type,

$$\sigma_{M^G-M^{Sc}} = \sigma_{M^C-M^{Sc}} = \sigma_{M^S-M^{Sc}} = \sigma_{M^{Sc}}$$

the dispersions of the differences are not useful to find the "best' ' (minimum variance) calibration. The mean values of the differences are also difficult to use: a) because of the Malmquist bias, if the sample used for the comparison is not well defined; b) because if we don't take into account some criteria related to the age of the stars, we may have a width of about 2 magnitudes for the earlier stars, and the mean value may lie almost anywhere in this interval.

It is not the place here to address the calibration problem (which will be studied when the Hipparcos data will be available); we use in what follows the Schmidt-Kaler calibration (1982) because it covers almost the whole HR diagram.

Absolute magnitude comparison

We compare the absolute magnitudes deduced from this calibration with the absolute magnitudes coming from the $uvby-\beta$ photometry. The sample used is made up of all non peculiar stars having $uvby-\beta$ photometry and MK spectral type of the INCA DB (8 000 stars). Table 2.2 shows the mean differences between absolute magnitudes and the residual dispersion (evaluated using the total dispersion of differences and the standard error on photometric absolute magnitudes).

For giant stars, the mean differences are too important. For dwarf stars, if we take into account a Malmquist bias of about 0.4 mag, the mean differences are acceptable.

Later we use these dispersions with the following assumptions: a) errors on the absolute magnitudes due to the spectral classification and errors on the absolute magnitudes coming from photometry are normally distributed and uncorrelated; b) the sample used above is representative of the Catalogue d'Entrée.

E.3 A model of interstellar extinction in the solar neighbourhood

E.3.1 Why study extinction?

The study of the interstellar matter is interesting in itself, in order to assess the structure and the evolution of galaxies. But it is also a source of problems: light is selectively absorbed, and then the apparent colour of the stars is redder than their intrinsic colour. Because of absorption, the distances of the stars are also systematically overestimated.

E.3.2 An empirical model of interstellar extinction

In order to prepare the Hipparcos mission, the colour of the stars needed to be known. For about 130 000 stars among the 214 000 stars of the INCA DB, there was no photoelectric photometry and the accuracy of the colour index (when it was available) was worse than 0.3 mag. It was then decided to obtain a colour for these stars, using an intrinsic colour coming from the spectral type, and a colour excess coming from an extinction model.

Up to now there was no complete mapping of the interstellar extinction, and no realistic model either. Given the number of stars to which a colour was requested, an analytical model needed to be built.

The included paper [Arenou *et al.*, 1992] describes this model.

E.3.3 Perspectives

In the years to come, a lot of data will be available:

- Hipparcos parallaxes and proper motions will give better absolute magnitude calibrations and, as a consequence, a better accuracy for the absorption;
- Tycho experiment will give both B_T and V_T magnitudes for about 500 000 stars, and only one magnitude for each of the 500 000 other stars;
- Tycho will also give parallaxes less accurate than Hipparcos parallaxes for 500 000 stars;
- there will be more Geneva and $uvby{-}\beta$ photometry available;
- the Michigan Spectral Survey will give spectral types for 225 000 stars when the 7th volume will be published;

We could imagine – as an upgrade of the existing model – a project which would give simultaneously:

- calibration of the intrinsic colours;
- calibration of the mean absolute magnitudes;
- colour excess;
- a tridimensional model up to 2kpc on bins of 1 square degree of the interstellar absorption, with an accuracy better than 0.2 mag;
- the coefficient R= $\frac{A_v}{E(B-V)}$ in each of these regions;

which would be also a contribution to the knowledge of circumstellar extinction, of star formation regions, etc.

STATISTICAL TOOLS

In different parts of this thesis we needed some statistical analysis. The next chapter is devoted to the obtention of estimators of "error-free' " variables when the initial variables have measurement errors and, more generally, this chapter describes all the statistical tools we used in this thesis.

E.4 Study of distributions with measurement errors

E.4.1 Generalities about estimation

Assuming X and Y to be random variables, we denote $f(x)$ its probability density function (pdf), $f(x|y)$ the conditional pdf, $F(x)$ its distribution function (df), and $E[X]$ its expectation.

Properties of estimators

In what follows, we use some properties of estimators: unbiasness, efficiency, robustness. They are defined in the (French) text.

Bayesian estimation

In the Bayesian approach of estimation, the parameter to find is considered as a random variable, with an *a priori* law. A Bayesian estimator is obtained with the mode, the median or the expectation of the *a posteriori* law

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int_\Theta f(x|\theta)f(\theta)d\theta}$$

It can be shown [Aïvazian *et al.*, 1986] that the mode and the expectation converge to the maximum likelihood estimator when the sample size $n \to \infty$, independently of the choice of $f(\theta)$.

E.4.2 Estimation taking errors into account

The distributions that we study, often come from variables with measurement errors. We firstly make the hypothesis that the variables y without measurement errors follow a Gaussian law and that measurement errors are normally distributed.

Simple Gaussian model

We show in annex (p. 92) that, if the variables without measurement errors $y_i \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$, and if $x_i = y_i + \varepsilon_{x_i}$ with $\varepsilon_{x_i} \rightsquigarrow \mathcal{N}(0, (\sigma_{x_i})^2)$ are the variables with measurement errors, then, by maximum likelihood, the estimators $\widehat{m}$ and $\widehat{s}$ of the unknown parameters μ and σ are found to be $\widehat{m} = \frac{\sum_{i=1}^{n} \frac{x_i}{\sigma^2+\sigma_{x_i}^2}}{\sum_{i=1}^{n} \frac{1}{\sigma^2+\sigma_{x_i}^2}}$ (with the associated variance $s_{\widehat{m}}^2 = \frac{1}{\sum_{i=1}^{n} \frac{1}{\sigma^2+\sigma_{x_i}^2}}$), $\widehat{s}$ being the solution of $\widehat{s}\sum_{i=1}^{n} p_i(p_i(x_i - \mu)^2 - 1) = 0$. We also show that $\widehat{m}$ is a minimum variance unbiased estimator of μ.

We use this estimator $\widehat{m}$ for the Hipparcos parallaxes and we use $\widehat{s}$ in order to compute the dispersions of the spectroscopic mean absolute magnitudes.

Solution of likelihood equations – The likelihood equations written above must be resolved by iteration. The solution is quickly found using the Newton method.

Statistical tests

In the previous section, we made an hypothesis of normality on the used distribution. We need tests to accept or reject this hypothesis.

More generally, we describe the tests we use elsewhere. When we mention that a test has been applied this means that the null hypothesis has been accepted (or rejected) with a two-sided test at a 5% threshold.

Normality tests – As normality tests, we use:

- Kolmogorov test when mean and variance are known before;
- Lilliefors test when mean and variance are computed from the sample;
- Test on skewness;
- Test on kurtosis;

Other adequation tests – To compare a sample distribution with a known distribution we use a Kolmogorov test. To compare two sample distributions we use a χ^2 test or, preferably, a Kolmogorov test because it does not need to bin data (which causes a loss of information).

Correlation and independence test – When we want to detect monotonic associations between variables, we use the Kendall's τ independence test because it is non parametric, non linear and robust.

Small departure from normality – robustness

We may ask whether the estimators found in §4.2.1 are robust. By robustness we mean here the problem of outliers and not the case where we know that the *a priori* law is not

Gaussian (if there is in fact a mixture of populations, we study in §4.4.1 how to find the parameters of the mixture).

It is well known that the arithmetic mean is sensible to outliers and that the empirical variance is very sensible to outliers. There exists much more robust estimators of the centre or width of a normal distribution: for instance, the symmetrically truncated mean or median for the centre and the mean deviation or the semi-interquartile range for the width; on the other hand, they have a greater variance (table 4.1). In this table, the coefficient $\sqrt{\frac{\pi}{2}}$ and 0.741 come from the fact that, otherwise, these estimators are biased estimators of the standard deviation.

In what follows we use as robust estimators of the centre and the width of a distribution the median and $\frac{1}{2}(x_{(0.8415)} - x_{(0.1585)})$, respectively, which we call "estimators based on quantile' '.

For the Gaussian model described in §4.2.1, we could suppose that the estimators quoted before (§4.2.1) are not robust. In order to find robust estimators, we use the "normalized' ' data $x_i' = \frac{x_i - m}{\sqrt{\sigma^2 + \sigma_{x_i}^2}}$ which should have a zero mean and unit variance: we change then by iteration $\widehat{m}$ and $\widehat{s}$ until x_i' has a mean around 0 and a variance near 1.

We compare all these estimators on page 66.

Simulations

Now we show how to perform simulations of a given distribution.

- When the pdf and the reciprocal of the df ($F^{-1}(z)$) are known analytically, we just have to generate uniform "random' ' numbers z;
- When the pdf is not known analytically, we may use the generalized lambda distribution defined by
$$F^{-1}(z) = \lambda_1 + \frac{z^{\lambda_3} - (1-z)^{\lambda_4}}{\lambda_2}$$
where $z \in [0, 1]$. As the 4 first moments may be written as a function of $\lambda_1, ... \lambda_4$, we compute the empirical 4 first moments and the corresponding λ_i; then z is generated uniformly over $[0, 1]$.
- Another way is to use the properties of the distribution which we want to simulate, when these properties are known: this may be done for instance for a normal law, a Cauchy law or a χ^2 law.

For all these methods, we just need in fact to generate uniform numbers on $[0, 1]$.

Comparison between estimators – To compare the quality of different estimators of the centre of a distribution (resp. the width), we use the asymptotic relative efficiency (ARE), taking the variance of the empirical mean (or the empirical standard deviation) as a reference.

For the estimators of the centre, we use:

1. the mean of the distribution truncated at $[-3\sigma, +3\sigma]$;
2. the weigthed mean (eq. 4.1);

3. the median;

4. the mean of the distribution truncated at 38%;

5. the robust weigthed mean (§4.2.3);

For the estimators of the width, we use:

1. the std deviation of the distribution truncated at $[-3\sigma, +3\sigma]$;

2. the weigthed std deviation (eq. 4.2);

3. the quantile-based width (cf p. 64);

4. the absolute deviation;

5. the robust weigthed std deviation (§4.2.3);

We simulate a Gaussian $\mathcal{N}(\mu, \sigma^2)$, the adopted average of measurement standard errors is $k\sigma$. Varying k, we may see the ARE of the estimators of the centre (fig. 4.1) and of the width (fig. 4.2).

The weigthed estimators are clearly always more efficient than the others and the robust estimators very sensitive to the size of the measurement errors.

Now, we fix the size of the measurement standard errors at 0.5σ on average and we add some outliers. This is done by combining a mixture of the preceding Gaussian $\mathcal{N}(\mu, \sigma^2)$ with a Gaussian $\mathcal{N}(\mu, (4\sigma)^2)$ in a variable proportion. The ARE of the estimators of the centre and of the width are presented fig. 4.3 and 4.4, respectively.

The robust weigthed estimators are not the more efficient. Concerning the centre, the weighted mean is still efficient, even if there is an important pollution. The robust estimators lose their efficiency when there are measurement errors. Concerning the width, the robust weigthed estimator is not always the best one, but its efficiency does not vary much. However, this situation is subject to changes when the pollution is asymmetrical, when the sample is small or when the size of the measurement errors is small.

E.4.3 Deconvolution of errors

In what follows, we are going to look for estimators of "error-free' ' variables when the variables at our disposal have measurement errors. Although this is applied to observed parallaxes, this is more general.

We first suppose that the initial variables y follow a Gaussian $\mathcal{N}(\mu, \sigma^2)$ law.

Bayesian point of view in the Gaussian model

x are the variables with a Gaussian measurement error. Maximizing the conditional *a posteriori* pdf, we find the Bayesian estimator of y given x: $\widehat{y} = x + (\mu - x)\frac{\sigma_x^2}{\sigma^2+\sigma_x^2}$ and its associated variance $\mathrm{Var}(\widehat{y}) = \frac{\sigma^2\sigma_x^2}{\sigma^2+\sigma_x^2}$

Bias due to measurement errors

We now take a general distribution for the y, keeping a Gaussian law for the measurement errors.

Throughout the next sections, we take an example of a distribution represented fig. 4.5. We choose the distribution of pdf

$$f(y) = \frac{1}{3\pi}\left(\frac{1}{1+(y-4)^2} + \frac{1}{1+(y/2)^2}\right)$$

as an example of a bimodal distribution.

We now take a measurement standard error of 0.5, and the simulated distribution of x is shown on fig. 4.6. If we look at the variation of $x - y$ as a function of x, we have an important bias. (To obtain this figure, we have applied a robust linear filter, taking for each point x on abscissa the median of the ordinate of the 500 points on each side of x; the curve is not represented on both extremities of the abscissa, as there are not enough points. The standard error on each point of the curve is $\frac{\sigma}{\sqrt{n}}\sqrt{\frac{\Pi}{2}}$, where $n = 500$ and $\sigma = 0.5$ in this example).

At a first glance the bias could seem unexpected since $E[X] = Y$. This is due to the non-uniform distribution of y and to the measurement errors. The bias would not have occurred if we had plotted $x - y$ as a function of y: we have $E[X - Y|Y] = 0$ but $E[X - Y|X] \neq 0$.

This is the bias we obtain if we truncate an observed distribution and if we compute a statistic on the truncated distribution. This is the case for instance when we only keep the nearby stars from a distribution of parallaxes; the resulting average of the parallaxes (or of the absolute magnitudes) is biased.

Bias calculation – Fortunately, it is possible to compute analytically the bias.

We suppose that we know the *a priori* law of y, which pdf is $f(y)$ and the conditional pdf (Gaussian or not) $f(y|x)$. The Bayesian estimator of y given x is then

$$\widehat{y} = E[Y|X] = \frac{\int y f(x|y) f(y) dy}{\int f(x|y) f(y) dy}$$

The figure 4.8 shows the bias of $x - \widehat{y}$ as a function of x. The estimator $\widehat{y}$ is not only a way to compute the bias but does represent an estimator better than x in a risk sense.

With this estimator, the problem is: which *a priori* law should we take? Fortunately, there exists a special case in which we don't have to find an *a priori* law.

Estimation without an *a priori* law

We now only assume that the conditional pdf is Gaussian. After some calculations we find that $\widehat{y} = x + s^2 \frac{f'(x)}{f(x)}$

This formula allows us to find the conditional estimator without using an *a priori* law, as we only use the observed pdf $f(x)$ and its first derivative $f'(x)$.

This is not a new result but, with this conditional point of view, we may also find the precision of this estimator:

$$\mathrm{Var}(\widehat{y}|x) = s^2\left(1 + s^2\left(\frac{f'(x)}{f(x)}\right)'\right)$$

using only the two first derivatives of the observed pdf.

Our next problem is then how to obtain good estimators of an observed pdf and of its derivatives.

Empirical estimation of the observed probability density

We have a n-points sample; in order to find an estimator of the pdf, there are some methods, among others:

- We could use a continuous df $F_n(t)$ which is differentiable, except in n points, and defined by
$$F_n(\frac{x_i + x_{i+1}}{2}) = F_n(\frac{x_i + x_{i-1}}{2}) + 1$$
the pdf is then found by derivation. However, this pdf would need smoothing.

- The Generalized Lambda Distribution defined at §4.2.4 may be used. The parameters $\lambda_1, ... \lambda_4$ are determined using the 4 first moments. The pdf is then found by
$$f(x) = f(F^{-1}(z)) = \frac{\lambda_2}{\lambda_3 z^{\lambda_3 - 1} + \lambda_4 (1-z)^{\lambda_4 - 1}}$$
but this estimator of the df is more convenient for unimodal and simple distributions.

- There is a way to obtain a smooth and differentiable estimator of the pdf, using the satistic $\widehat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} k(\frac{x - x_i}{h})$ where $k(t)$ is a symmetric, positive function which integral is equal to 1, and leading to 0 when $|t| \to \infty$.

We used this last estimator with a Gaussian for the function $k(.)$ (named a convolution kernel) and a variable h depending on n (which allows the convergence of the estimator towards the real pdf), on the width of the distribution and on the size of the measurement error; this is in fact a "window' ' through which we count the weight of each observation.

In the case of the example of §4.3.2, we use this estimator of the pdf (fig. 4.9 to be compared to the histogram fig. 4.6). With this pdf and the estimator of the "error-free' ' variable (eq 4.6), we show on fig. 4.9 the bias corrected (not significantly different from 0).

When possible, we use this method to find the "error-free' ' variables, instead of an *a priori* law, because we are frequently faced to (unknown) selection biases in our data.

E.4.4 Multivariate estimations

Mixture of Gaussian populations

Introduction – In part IV, we address the problem of separating a mixture of Gaussian components. Up to now, this had been addressed without taking into account the measurement errors.

Statistically, this is a problem of incomplete data; the problem is to find the parameters of each of the populations which are assumed Gaussian (proportions, means and variances). In dimension k, for m components of the mixture, the number of parameters to estimate is $m\frac{(k+1)(k+2)}{2} - 1$. This can then be a problem for small samples.

The most powerful method is the EM (Estimation-Maximization) method, the principle of which we describe briefly in the included paper [Bougeard & Arenou, 1989]. However, this algorithm needs an initial solution and may badly deviate from the true solution if the components are too overlapped or if one component is too small.

A recent method, SEM (SEMMUL in the multidimensional case), with a stochastic step, permits us to find the number of components, without needing an initial solution and has better convergence properties.

The classical Bayesian approach is useless as all the partitions of the sample should be taken into account. However the Bayesian sampling allows us to obtain the estimates of the parameters.

We implemented a version developed from SEM which takes into account the measurement errors. We show, in next section, two examples of the algorithm behaviour.

Simulations – We generate two overlapping populations (proportion 50%/50%) with a total of only 100 points. The means of the populations are -20 and 20, respectively; the measurement standard errors are 7 on average; the standard deviation is 10 in the first test and 15 in the second test (fig. 4.11 and 4.12).

The results of the first test are in table 4.2. The multidimensional version where measurement errors are taken into account gives the best result, close to the true solution (50,-20,10 ; 50,20,10). For the second test, the results are not so good.

Stability of the algorithms – The stability of the results obtained with SEMMUL is studied in the paper [Bougeard & Arenou, 1989].

We also compared these results with those obtained with non-parametrical methods; in another paper [Arenou, 1990], with a clustering around moving centers, we find that 97% of the stars are in the classes found by SEMMUL. Whe also show that the results of a hierarchical ascending classification agree with the classes found by SEMMUL for 92% of the stars [Arenou & Bougeard, 1992].

Estimations by least squares

In chapter 3, we use a powerful software developed for the Space Telescope preparation: Gaussfit. It gives a robust solution and it takes into account the measurement errors on all the variables.

E.4.5 Conclusion

We have described:

- how to find the better estimators in the Gaussian case,
- conditional estimators of the "error-free' " variables, in the Gaussian or without an *a priori* law,
- estimators of the observed pdf,
- how to correct the bias coming from measurement errors,
- how to separate Gaussian populations in a mixture,

and we will use these results in other parts of this thesis.

VALIDATION OF THE HIPPARCOS INPUT CATALOGUE AND OF THE PRELIMINARY DATA FROM HIPPARCOS

This part is devoted to the comparisons between ground-based data and the first results of the satellite. In chapter 5 we compare positions and magnitudes with the data of a six month mission obtained with the Hipparcos star mappers. In chapter 6 we will address the problem of validating the parallaxes obtained after a one-year sphere solution by both Data Reduction Consortia (DRC). The basic idea is to study these parallaxes from a methodological point of view.

Due to the very preliminary nature of these data, no general conclusion should be drawn from these data, although they already seem of high quality.

E.5 Hipparcos positions and magnitudes

For everybody who worked on the elaboration of the Hipparcos Catalogue d'Entrée, the first question was: would the satellite find the right stars at the right places, with the right magnitudes?

Among the 118 000 stars of the Catalogue d'Entrée, less than 0.1% were not found, and this is a rather good result. The accuracy of the ground-based measurements were important for the optimization of the observing time and for the Data Reduction task. We may also imagine that the Data Reduction Consortia needed to validate their first results.

The included paper [Turon *et al.*, 1992] is the first comparison between the data from the Catalogue d'Entrée and the positions and magnitudes obtained with the Hipparcos star mappers (for the first 6 months of mission).

This comparison is satisfactory, and the only problems encountered concern:

- biases in some astrometric southern Catalogues;
- small variations (with the position) of the differences between ground-based positions and positions deduced from the star mapper data;
- small deviations in the photometric data;

E.6 Study of preliminary Hipparcos parallaxes

E.6.1 Introduction

The Hipparcos parallaxes should be absolute, however it may exist a (small) global zero-point shift due to periodic basic angle variations. If it exists, even if it is small, it must be found and shown independant of the characteristics of the stars.

We have at our disposal the preliminary parallaxes obtained by each Data Reduction Consortium after a one year mission duration. We study the FAST 3 parameters and the NDAC 5 parameters solutions making internal and external comparisons. We show a method allowing to find the instrumental zero-point shift and the external errors, when the final Hipparcos will be available.

Due to the preliminary nature of these plxs and to the data rigths of the proposals, we use them only for a validation, and never to obtain early scientific results. For instance, we only use the existing absolute magnitude calibrations although the absolute magnitude deduced from Hipparcos parallaxes would have help us.

Let us make a small remark concerning the zero-point z. Assuming that the Hipparcos parallaxes have an standard error of about 2 mas, we would need about 40 000 distant stars to get a precision of 0.01 mas on z; this is not possible given the content of the Catalogue d'Entrée 6.1.

The distant stars may be selected using the spectroscopic parallax estimate. For these distant stars the errors on the spectroscopic parallaxes are small (this is true for the random errors). However, if we average these estimates of the spectroscopic parallax, we show that it produces a systematic error on z. In this case and in other cases, we study in details the possibility of systematic errors.

Before doing that, we study in the next section the comparisons between Consortia parallax data.

E.6.2 FAST and NDAC parallaxes - Internal comparisons

First look on preliminary parallaxes

We have chosen to use the FAST one-year sphere solution with 3 parameters and the NDAC sphere solution with 5 parameters because there would not be enough place here to consider all the solutions already obtained. The distribution of these parallaxes are given on fig. 6.1 and 6.2 for FAST and NDAC, respectively. There are 5632 stars with a negative parallax in the first case and 3982 stars in the second case.

The standard errors on the parallaxes depend on the magnitude (table 6.2) and on the ecliptic latitude (table 6.3). The DRC have thus computed a formal error associated to each parallax. The distributions of these formal errors are given on fig. 6.3 and 6.4 for FAST and NDAC, respectively. The mean values of these errors are $\langle s_{\rm F} \rangle = 2.217$ mas and $\langle s_{\rm N} \rangle = 2.158$ mas[1]. These results show that the two distributions of parallaxes are comparable.

As the measurement standard error varies from star to star, a distribution of distant stars is not Gaussian; however the distribution of the "normalized' ' data should ideally

1. We use everywhere the subscript F for FAST-3P, N for NDAC-5P, H for Hipparcos (FAST or NDAC), P for photometric, S for spectroscopic ; $s_{\rm H}$ for the formal error on one parallax, and $\sigma_{\pi_{\rm H}}$ for the external error on one parallax.

be a centered Gaussian with unit variance.

Concerning all the stars common to FAST-3P and NDAC-5P solutions, we may now represent the distribution of the differences of parallaxes (fig. 6.5), and the distribution of the "normalized' ' differences (fig. 6.6). In this last case, the differences are divided by the formal error on the differences, taken as $\sqrt{(0.95 s_{\rm F})^2 + (0.8 s_{\rm N})^2}$; the coefficients 0.95 and 0.8 are due to the fact that FAST-3P and NDAC-5P parallaxes are correlated. We justify these coefficients later.

As it may be seen on the first figure, the average of the differences is not significantly different from 0 (-0.04 ± 0.018), and the dispersion is 2.713 mas. Although the number of points is too important (and then some normality tests reject the null hypothesis), the distribution of the errors on the "normalized' ' differences seems Gaussian; this may be seen on the second figure.

Errors on Hipparcos parallaxes

Comparing both distributions of parallaxes (fig. 6.1 and 6.2), it seems that the errors on NDAC-5P are smaller than the errors on FAST-3P; this is also confirmed by the number of negative parallaxes.

We could use these negative parallaxes in the following way: taking negative FAST-3P parallaxes (then corresponding to distant stars), we may compute the dispersion of NDAC-5P parallaxes (2.59 mas). With the negative NDAC-5P parallaxes, the dispersion of FAST-3P parallaxes is 2.87 mas. These estimates are not very useful in fact because the errors on FAST and NDAC parallaxes are correlated. If the parallaxes were not correlated, the dispersion on the difference would be $\sqrt{2.59^2 + 2.87^2} = 3.87$ instead of 2.713 mas.

Bias as a function of the parallax

Now, we may ask if the differences between Consortia parallaxes vary with the parallax itself. These differences $\pi_{\rm FAST} - \pi_{\rm NDAC}$ versus $\pi_{\rm FAST}$ are represented fig. 6.7 and versus $\pi_{\rm NDAC}$ fig. 6.8, respectively. There is clearly a bias (the standard error on each ordinate is about 0.09 mas), slightly smaller in the second case.

This could be an important problem as the zero-point would be dependent on the parallaxes. In fact, we have shown in §4.3.2 that it is due to the measurement errors.

We must then use the conditional estimator of the parallax: we denote $\pi_{\rm H}$ the parallax obtained by one of the DRC, $f(\pi_{\rm H})$ the pdf of these parallaxes, then $\widehat{\pi} = \pi_{\rm H} + \sigma^2_{\pi_{\rm H}} \frac{f'(\pi_{\rm H})}{f(\pi_{\rm H})}$.

Estimation of external errors

It is possible not only to suppress the bias, but also to *use* it. We will see that it allows us to find the estimates of the external errors of the parallaxes, the errors during each reduction process, and the common errors (instrumental error).

We adopt a simple additive model for the errors:

$$\begin{aligned} \pi_{\rm FAST} &= \pi + \varepsilon_C + \varepsilon_F \\ \pi_{\rm NDAC} &= \pi + \varepsilon_C + \varepsilon_N \end{aligned}$$

where ε_C, ε_F and ε_N are uncorrelated, with variance σ^2_C, σ^2_F and σ^2_N, respectively. We may assume that the expectation of ε_F and the expectation of ε_N are zero (the mean difference

between DRC parallaxes is almost 0); $E[\varepsilon_C]$ is then the zero-point of the parallaxes. The "external' ' errors of FAST and NDAC parallaxes are $\sigma_{\pi_\mathrm{F}} = \sqrt{\sigma_C^2 + \sigma_F^2}$ and $\sigma_{\pi_\mathrm{N}} = \sqrt{\sigma_C^2 + \sigma_N^2}$, respectively; σ_C^2, the common error, is the ultimate precision which could be obtained after a one-year mission, if there were no approximations during the sphere solution process, and no errors in the Catalogue d'Entrée proper motions (for the 3P solution) or if there were no contamination between parallaxes and proper motions (for the 5P solution). We call σ_F and σ_N the standard errors on one parallax due to the reduction process.

The total standard error on Hipparcos parallax may be reduced if we take the appropriate weighting between FAST and NDAC parallaxes. We address this problem in §6.2.5.

Errors in the reduction process – It is easy to show that the bias shown in fig. 6.7 is

$$E[\pi_\mathrm{F} - \pi_\mathrm{N}|\pi_\mathrm{F}] = -\sigma_F^2 \frac{f'(\pi_\mathrm{F})}{f(\pi_\mathrm{F})}$$

and, similarly, in fig. 6.8, the bias is for each parallax:

$$E[\pi_\mathrm{F} - \pi_\mathrm{N}|\pi_\mathrm{N}] = \sigma_N^2 \frac{f'(\pi_\mathrm{N})}{f(\pi_\mathrm{N})}$$

In order to compute the above expectations, we need to find the observed pdfs and their derivatives. For this purpose, we use the estimate found in §4.3.4.

We compute σ_F and σ_N for each star, assuming that these dispersions have been taken into account by each DRC when they computed the formal error for each parallax; we then write $\sigma_F \approx k_F s_\mathrm{F}$ and $\sigma_N \approx k_N s_\mathrm{N}$, respectively, and we compute the mean value of k_F and k_N.

This is simply evaluated using the quantiles of the distribution of the observed parallaxes (table 6.4). k_F and k_N do not vary very much, except near the mode of the distribution of the parallaxes, where the sign of the derivative of the pdf changes.

Thus, we find that $k_F \approx 0.95$ and $k_N \approx 0.8$ and we may verify it, in the figures 6.9 and 6.10, where we substracted $E[\pi_\mathrm{F} - \pi_\mathrm{N}|\pi_\mathrm{F}]$ and $E[\pi_\mathrm{F} - \pi_\mathrm{N}|\pi_\mathrm{N}]$, respectively, to each difference $\pi_\mathrm{F} - \pi_\mathrm{N}$. These figures also show that the differences of the parallaxes do not vary with the parallax itself.

Given that, on the average, the internal errors are $\langle s_\mathrm{F} \rangle = 2.217$ mas and $\langle s_\mathrm{N} \rangle = 2.158$ mas, we then find that the global contribution of the errors due to the reduction process to the external errors are $\langle \sigma_F \rangle \approx 2.106$ mas and $\langle \sigma_N \rangle \approx 1.726$ mas on the average, for FAST-3P and NDAC-5P, respectively.

We may verify this result as $\sigma_{\pi_\mathrm{F}-\pi_\mathrm{N}} = 2.713$ mas while we find $\sigma_{\pi_\mathrm{F}-\pi_\mathrm{N}} = \sqrt{\sigma_F^2 + \sigma_N^2} \approx 2.723$ mas. Another way to verify it is to compute the total dispersion of the distribution of FAST parallaxes (7.000 mas) and of NDAC parallaxes (7.100 mas) for the stars belonging to the two solutions. The quadratic difference between these dispersions is 1.19 mas while the quadratic difference between σ_F and σ_N is 1.21 mas. This last calculation probably means that there are few outliers in these distributions.

Common error – We now wish to find the estimation of the common dispersion σ_C. This is not straightforward without using external data. We could obtain the observed

distribution with a convolution of an *a priori* distribution with the error law. However the size of the errors could be underestimated if the variance of the *a priori* distribution is overestimated.

We then used the fact that

$$\mathrm{Var}(\pi) = E[\mathrm{Var}(\pi|\pi_\mathrm{N})] + \mathrm{Var}(E[\pi|\pi_\mathrm{N}])$$

π being the true parallax, and $E[\pi|\pi_\mathrm{N}] = \pi_\mathrm{N} + \sigma^2_{\pi_\mathrm{N}} \frac{f'(\pi_\mathrm{N})}{f(\pi_\mathrm{N})}$ and $\mathrm{Var}(\pi|\pi_\mathrm{N}) = \sigma^2_{\pi_\mathrm{N}}(1 + \sigma^2_{\pi_\mathrm{N}}(\frac{f'(\pi_\mathrm{N})}{f(\pi_\mathrm{N})})')$

We then write $\sigma_C \approx k s_\mathrm{N}$, thus $\sigma_{\pi_\mathrm{N}} = \sqrt{k^2 + 0.8^2} s_\mathrm{N}$ and we find the mean value of k which minimizes the right part of eq 6.1. Simulations showed us that it could give the expected result at least for NDAC-5P parallaxes. Why we may have a minimum may be explained if we notice that $E[pi|pi_N]$ is a correction to the observed parallax: as the variance increases, $E[pi|pi_N]$ goes closer to the mode of the parallax distribution (thus $\mathrm{Var}(E[pi|pi_N])$ decreases). Now, if the variance is too important a star is rejected on the other side of the mode, thus increasing the variance $\mathrm{Var}(E[pi|pi_N])$. Given the fact that we benefit by an asymetrical, leptokurtic distribution and by measurement errors of different sizes but small on the average, $\mathrm{Var}(\pi)$ has a minimum in our case. Still, we must point out that this method is conjectural. We may see on second column of table 6.5 that the minimum is found with $k \approx 0.88$. The same procedure was in fact not possible with FAST-3P parallaxes probably because the described method is not valid in a general case.

All these results obtained from the internal comparisons between DRC parallaxes (22820 stars) are summarized in table 6.6. For each parallax, we may write, with a sufficient accuracy, that the "external errors' " on each parallax are $\sigma_{\pi_\mathrm{N}} \approx \sqrt{0.8^2 + 0.88^2} s_\mathrm{N} = 1.19 s_\mathrm{N}$ and $\sigma_{\pi_\mathrm{F}} \approx \sqrt{0.95^2 + (0.88\frac{s_\mathrm{N}}{s_\mathrm{F}})^2} s_\mathrm{F} \approx 1.32 s_\mathrm{F}$, for NDAC-5P and FAST-3P, respectively.

Best estimate of Hipparcos parallax

In order to verify these results which permit us to obtain an estimate of the "external' " errors for *each* parallax, we may now ask what would be the final estimator of Hipparcos parallaxes.

The answer is described in §4.2.1. Using the FAST and the NDAC parallaxes, the best unbiased estimator of the Hipparcos parallax for one star is

$$\widehat{\pi_\mathrm{H}} = \frac{\frac{\pi_\mathrm{F}}{\sigma_C^2+\sigma_F^2} + \frac{\pi_\mathrm{N}}{\sigma_C^2+\sigma_N^2}}{\frac{1}{\sigma_C^2+\sigma_F^2} + \frac{1}{\sigma_C^2+\sigma_N^2}}$$

or, using the results above, as a function of the internal errors,

$$\widehat{\pi_\mathrm{H}} = \frac{\frac{\pi_\mathrm{F}}{(1.32 s_\mathrm{F})^2} + \frac{\pi_\mathrm{N}}{(1.19 s_\mathrm{N})^2}}{\frac{1}{(1.32 s_\mathrm{F})^2} + \frac{1}{(1.19 s_\mathrm{N})^2}}$$

the variance of this estimator is $\frac{1}{\frac{1}{(1.32 s_\mathrm{F})^2} + \frac{1}{(1.19 s_\mathrm{N})^2}}$ (1.93 mas on the average).

Note, however, that this is just given as an example, since we used a 5-parameters solution with a 3-parameters solution.

Since this "Hipparcos' " estimator has a minimum variance, it should then give the smaller total dispersion of the parallaxes obtained by this way. This is approximately true (third column of table 6.5), within an 0.035 mas standard error.

We may notice that, for the 5P solution, the errors during the reduction process are not high, perhaps because of the use of the gyroscopes. It is also noticeable that the proper motions (not well defined after a one-year mission) have not contaminated too much the parallax solution.

L. Lindegren (1992a) made an estimation of the global external errors with both 3-parameters solutions, using an *a priori* law for the parallaxes. He found $\langle\sigma_C\rangle = 2.07$ mas and $\langle\sigma_F\rangle = 1.99$. Our estimates are consistent with these results; we find a smaller global common error and this may be explained by the fact that, when comparing two 3P solutions, the error of the ground-based proper motions should increase the common error.

We are somehow reluctant to use an *a priori* law, as the way the Catalogue d'Entrée was built do not enable us to find it easily. Considering the pdf of spectroscopic or photometric parallaxes, fig. 6.30 and 6.31, respectively, it would have been difficult to imagine them.

Although our method would need some further refinements, our estimates allow us to obtain, in a non-parametric way, the individual external error on each parallax and then the best estimate of the resultant Hipparcos parallax; the only hypothesis is the Gaussian nature of the error on the parallaxes.

Since we only used the preliminary parallaxes, we now clearly need to make some comparisons with external data. This is done in the next sections. Among the external data, we use the spectroscopic parallaxes, and we start studying their estimates.

E.6.3 Estimates of the spectroscopic parallaxes

Given the number of stars with spectral types and luminosity classes, the spectroscopic parallaxes should be an appropriate source of external data to validate the preliminary parallaxes.

One could think that a spectroscopic parallax is easily obtained, using the Pogson law. However, Smith Jr (1985) used a particular estimator of the spectroscopic parallax when combining this parallax with a trigonometric parallax in order to obtain the best parallax. This is mainly due to the Malmquist bias.

Malmquist bias

Having a group of stars with the same spectral type and luminosity class, it seems natural to take the average of the indiviual absolute magnitudes if we need the mean absolute magnitude corresponding to this spectral type and luminosity class. But if the stars of this group come from a magnitude-limited sample, we are faced to a sampling bias. This is the Malmquist bias.

Suppose the distribution of absolute magnitudes M be Gaussian for each spectral type and luminosity class, hence

$$\phi(M) = \frac{\phi_0}{\sigma_M\sqrt{2\Pi}} e^{-\frac{1}{2}\frac{(M-M_0)^2}{\sigma_M^2}}$$

where M_0 denotes the (true) mean value, σ_M the dispersion and ϕ_0 the spatial density for the stars of this type. Thus,

$$\left| \begin{array}{rcl} \overline{M}_m & = & M_0 - \sigma_M^2 \frac{d \ln n(m)}{dm} \\ \sigma_m^2 & = & \sigma_M^2 \big(1 + \sigma_M^2 \frac{d^2 \ln n(m)}{dm^2}\big) \end{array} \right. \tag{E.2}$$

where $\overline{M}_m$, σ_m and $n(m)$ denote the mean absolute magnitude, the dispersion and the number of stars having this apparent magnitude m, respectively. Changing accordingly the definitions above, this result holds true for a sample of $N(m)$ stars brighter than a magnitude m.

With uniform spatial distribution

$$\left| \begin{array}{rcl} \overline{M}_m & = & M_0 - 1.38\sigma_M^2 \\ \sigma_m^2 & = & \sigma_M^2 \end{array} \right. \tag{E.3}$$

As we can see, this result assumes that the absolute magnitude is normally distributed. From a physically point of view this is not really the case: for main sequence stars it should be truncated by the ZAMS and it should take into account the fact that the lifetime spent on the main sequence varies with the position on it. A more appropriate distribution should then probably be truncated and non-symmetrical.

Calculation of the spectroscopic parallaxes

Reciprocally, we suppose we know the (true) mean absolute magnitude M_0 of a group of stars with apparent magnitude m within a given spectral type and luminosity class. Since we have the Malmquist bias, the distribution of the absolute magnitudes is

$$\Phi'_m(M) = \frac{\Phi_0}{\sigma_M \sqrt{2\Pi}} e^{-\frac{1}{2}\frac{(M-\overline{M}_m)^2}{\sigma_M^2}}$$

Smith Jr (1985) found the estimator of the most probable spectroscopic parallax $\pi^* = \pi_0 \cdot 10^{-0.368\sigma_M^2}$ given that

$$\phi_m(\pi) = k(m, M_0, \sigma_M) e^{-\frac{1}{2}\frac{25(\pi-\pi^*)^2}{\sigma_M^2}}$$

where $\pi_0 = 10^{-\frac{(m-M_0+5)}{5}}$ is the classical estimator of the spectroscopic parallax.

The difference between these two estimators is due to the Malmquist bias, and, to a less extent, due to use of the the most probable estimator.

Concerning this last estimator, denoting $\langle M \rangle$ the mean absolute magnitude (M_0 or $\overline{M}_m$): we could equally use the median $\pi_{\text{med}} = 10^{-\frac{(m-\langle M \rangle+5)}{5}}$, the expectation $\pi_{\text{med}} \times 10^{\frac{1}{2}\frac{\ln 10}{25}\sigma_M^2}$, or the mode $\pi_{\text{med}} \times 10^{-\frac{\ln 10}{25}\sigma_M^2}$ of the lognormal distribution of the spectroscopic parallaxes whose variance is ${\pi_{\text{med}}}^2 \times 10^{\frac{\ln 10}{25}\sigma_M^2}(10^{\frac{\ln 10}{25}\sigma_M^2} - 1) \approx (0.4605\pi_{\text{med}}\sigma_M)^2$

Simulation – As we want to use the mean value of spectroscopic parallaxes later, we would like to know which estimator is unbiased; firstly we want to know the size of the bias. We then perform a simulation using stars of a given type, in a volume-limited sample and in a magnitude-limited sample, and compute three different estimators of the spectroscopic parallax:

1. the usual estimator: $\pi_0 = 10^{-\frac{(m-M_0+5)}{5}}$

2. the estimator given by Smith Jr: $\pi^* = \pi_0 \cdot 10^{-0.368\sigma_M^2}$

3. the expectation: $\widetilde{\pi} = \pi_0 \cdot 10^{0.046\sigma_M^2}$ for the volume-limited sample and $\widetilde{\pi} = \pi_0 \cdot 10^{-0.230\sigma_M^2}$ for the magnitude-limited sample.

In order to compare these three estimators with the true parallax π, we study the relative variation $\delta = \frac{\text{estimator}-\pi}{\pi}$ whose distributions are on figure 6.11. With a dispersion σ_M of only 0.5 mag, the bias can reach 20%.

Furthermore, if we want to know $\langle\delta\rangle$ more quantitatively, we must notice that the distribution of δ is lognormal. The mean, median and mode of δ are then shown on table 6.7 and table 6.8 for the distance-limited sample and for the magnitude-limited sample, respectively.

We may notice that if we want to use the mean of the spectroscopic parallaxes, no estimator gives $\overline{\delta} \approx 0$. We show that this is due to the fact that

$$\overline{\delta} \approx \frac{1}{2}(\frac{\ln 10}{5})^2\sigma_M^2$$

As a consequence, we choose:

$$\pi_S = \begin{cases} \pi_0.10^{-\frac{1}{2}\frac{\ln 10}{25}\sigma_M^2} & \approx \pi_0.10^{-0.046\sigma_M^2} & \text{(volume-limited sample)} \\ \pi_0.10^{-(\frac{\ln 10}{5}\frac{d\log N(m)}{dm}+\frac{1}{2}\frac{\ln 10}{25})\sigma_M^2} & \approx \pi_0.10^{-0.322\sigma_M^2} & \text{(magnitude-limited sample)} \end{cases} \quad \text{(E.4)}$$

as the estimator of the spectroscopic parallax when we want to get an unbiased estimator of the mean of spectroscopic parallaxes.

Completeness of the samples

The estimator quoted before implies that we must know the selection biases in the Catalogue d'Entrée. There is a limiting magnitude for the stars observed by Hipparcos ($H_p = 12.4$ mag). As may be seen on figure 6.12, the number of stars fainter than $V = 9$ is relatively small. For stars having spectroscopic distances greater than 500 pc, the slope of the cumulative distribution of the number of stars versus apparent magnitude (fig. 6.13) is not different, except for fainter stars.

With the presence of the "Survey' ', we know that the Catalogue d'Entrée is almost complete up to magnitude $V_{\lim}$. For stars fainter, it is difficult to find the selection biases, given the way the Catalogue d'Entrée was built. For a further use of the spectroscopic parallaxes, we will then use a Malmquist correction but using a slope $\frac{d\log N(m)}{dm}$ dependent on m.

We may now directly compare the preliminary parallaxes with external data.

E.6.4 Comparison with external estimations

We now study the differences between the preliminary parallaxes and external data: spectroscopic parallaxes, photometric parallaxes, dynamical parallaxes, parallaxes deduced from cluster distance moduli, with stars in the Magellanic clouds, and, of course, trigonometric parallaxes.

As a first step, we study the first two moments of the distribution of these differences. Then, we see how the systematic differences[2] between preliminary parallaxes and other estimates of the parallaxes vary with positions, proper motions, magnitudes and colours of the stars observed, and with the parallaxes themselves. Finally, we find the external error and the zero-point of the parallaxes, from a methodological point of view, applied to the preliminary parallaxes.

Trigonometric parallaxes

In order to compare the ground-based parallaxes with the preliminary parallaxes, we use the trigonometric parallaxes from the General Catalogue of Stellar Trigonometric Parallaxes (GCTSP) [van Altena *et al.*, 1991]. Using the FAST-3P parallaxes, we may see on figure 6.14 that the ground-based parallaxes are much more scattered, with some outliers. The differences $\pi_{\rm FAST} - \pi_{\rm GCTSP}$ are shown in fig. 6.15.

On the top-left of the figure, the number of stars, the mean/median and the standard deviation/a robust width are given. The robust estimators (median and width) are used to compute the Gaussian which is plotted only to show the skewness and the kurtosis of the distribution. This does not mean that the distribution is supposed to be normally distributed.

The width is about 12 mas and the mean difference is about 2 mas. If we suppose that it is due to GCTSP parallaxes, this bias comes from distant stars, farther than 30 mas.

Spectroscopic parallaxes

The second comparison concerns the difference between preliminary parallaxes and spectroscopic parallaxes. For the computation of spectroscopic parallaxes we use only photoelectric magnitudes; absorption is computed from the colour excess, using a photoelectric $B-V$, and an intrinsic colour coming from the Schmidt-Kaler (1982) calibration. We keep only non-peculiar, non-binary stars with MK spectral type. The formal error on spectroscopic parallaxes is computed using the dispersion of the absolute magnitude found in §2.3.2.

The distribution of differences (fig. 6.16) is not symmetric, probably due to giant stars wrongly classified as dwarfs. Keeping only the distant stars ($\pi_{\rm S} < 2$ mas), we may notice (fig 6.17) that the dispersion is reduced from 3.5 mas to 2.1 mas. This is our first estimate of the external error of the preliminary parallaxes.

We may also notice that the systematic difference is negative in the first case, whereas it is positive when only distant stars are taken into account. This is because we truncated the observed distribution, and we will study this statistical problem in §6.5.1.

Photometric parallaxes

In what follows we call photometric parallax the parallax which is obtained with the absolute magnitude coming from $uvby{-}\beta$ photometric calibrations, described on chapter 2.

2. which we write sdp in what follows

Compared to spectroscopic parallaxes, the dispersions of photometric parallaxes are smaller, and this may be seen on fig. 6.18. With the more distant stars, the dispersion is 2.1 mas, as was obtained in the case of the comparison of the preliminary parallaxes with the spectroscopic parallaxes.

However, this distribution of the differences is leptokurtic, perhaps because we used different photometric calibrations, producing different uncertainties on the absolute magnitudes. If this is the case, then it means that the errors of photometric parallaxes have also an influence, even for distant stars, and that the external errors of the preliminary parallaxes are smaller than 2.1 mas in this sample. This is possible since the stars having $uvby{-}\beta$ photometry are generally bright, and bright stars have a smaller error on the Hipparcos parallax.

Parallaxes of stars in open clusters

For clusters far enough, we may assume that all the cluster stars have the same distance. Given the uncertainty on the distance moduli of clusters, we have an internal precision on the parallaxes of these stars probably around 5%-10%.

Using BDA cluster data base, we automatically select Catalogue d'Entrée stars with a cluster identifier, suppress the suspected non-members, and give to the rest of the stars the parallax of the cluster, using the distance moduli quoted by Lyngå(1987). Although we suppressed the known non-members, we still may have stars in our sample which are not physically members of the cluster.

The difference between NDAC preliminary parallaxes and these cluster parallaxes is shown fig. 6.20. The distribution is skew and the long tail of the distribution is probably due to non-members.

However, given their apparent proximity on the sky, stars from a given cluster allow us to test the behaviour of preliminary parallaxes on a small sky zone. We already know [Lindegren, 1989] that the precision on the mean parallax of a group of adjacent stars is about $\frac{\sigma_\pi}{n^{0.35}}$ instead of $\frac{\sigma_\pi}{\sqrt{n}}$; this is due to correlations.

Comparing the mean preliminary parallaxes in each cluster (26 clusters with at least 5 stars) to the known parallax of the cluster, we do not find any global problem. However we may notice that a cluster (CL Blanco 1, 12 stars) has a NDAC-5P mean parallax of 0.15 ± 0.73 mas, significantly different from the parallax, 4.11 ± 0.11 mas, deduced from its distance modulus [Westerlund *et al.*, 1988]; there is no indication that this problem could come from non-member stars. This probably illustrates the preliminary character of Hipparcos parallaxes as this cluster is near the ecliptic, where the parallaxes are not well measured after a one-year mission. The situation is subject to change, and has probably yet changed.

Magellanic clouds stars

Given the distance of the Magellanic clouds (about 50 & 65 kpc), the star parallaxes are negligible. 46 stars are however observed by Hipparcos in order to get an upper limit on their proper motion. 32 stars have a preliminary NDAC parallax.

With a (true) parallax close to 0, these few stars allow us to see directly the error on the preliminary parallaxes (fig. 6.22). The Gaussian hypothesis is not rejected for this

distribution and its mean is not significantly different from 0.

Dynamical parallaxes

If the orbital elements of a binary system are known, and using the mass-luminosity relationship, we obtain an estimate of its dynamical parallax. There are 369 dynamical parallaxes in the INCA DB; their relative error is supposed to be smaller than 5% [Dommanget, 1992]. We only keep the stars with a dynamical parallax smaller than 20 mas, and the distribution of $(\pi_{\rm NDAC} - \pi_{\rm dynamical})$ is shown in fig. 6.23. Two obvious outliers were rejected. The median is -0.8 ± 0.45 mas and the width 3.15 ± 0.39 mas.

We may suspect the presence of outliers in the tails of the distribution; these stars are close binaries. In each system we computed for the brighter star other estimates of the parallax, they were consistent with the dynamical parallax. We cannot reject the possibility that some problems arised during the reduction process for these few binary stars.

The width of distribution is larger than expected. However, after rejecting outliers, the normalized difference is found to be Gaussian $\mathcal{N}(0, 1^2)$ by a Kolmogorov test.

This indicates that there are no important systematic effects on both parallaxes and that the standard errors are correctly estimated. However, given the uncertainties on orbital elements and the mass-luminosity relationship, we may not rely on dynamical parallaxes to validate the preliminary parallaxes.

E.6.5 Independence of the sdp of the preliminary parallaxes

We already noticed that it was important to show that a possible systematic error (estimated by the sdp) of the Hipparcos parallaxes should not depend on the characteristics of the stars. We would then conclude that there were no problems during the observation by Hipparcos and during the reduction process, and we could find the global zero-point of Hipparcos parallaxes using the more distant stars only.

Independence of sdp of the parallaxes from the parallax itself

In order to study the variation of the sdp with the parallax itself, we use the only estimates (with enough stars) that we have at our disposal: spectroscopic and photometric parallaxes.

Bias on the parallax – Firstly we plot the differences $\pi_{\rm N} - \pi_{\rm S}$ versus the spectroscopic parallax $\pi_{\rm S}$ (fig. 6.25) and $\pi_{\rm N} - \pi_{\rm P}$ versus the photometric parallax $\pi_{\rm P}$ (fig. 6.24). We notice a positive bias for the smallest parallaxes and a negative bias for the greater parallaxes. Since these biases reach some mas, we must find an explanation.

If we suppose that the bias is due to the fact that the error on $\pi_{\rm P}$ is correlated with the error on $\pi_{\rm N} - \pi_{\rm P}$, we may suppose that the bias should disapear if we study $(\pi_{\rm H} - \pi_{\rm S})$ vs $\pi_{\rm P}$ or $(\pi_{\rm H} - \pi_{\rm P})$ vs $\pi_{\rm S}$, as errors on spectral classification and errors on photometry should be uncorrelated. We may notice on fig. 6.26 and 6.27 that the bias is still present, although reduced.

In order to verify if there is a correlation between the spectroscopic and photometric parallax errors, we keep the stars having simultaneously a spectroscopic parallax, a

photometric parallax and a NDAC parallax greater than 20 mas (294 stars). With these last parallaxes, we may compute an absolute magnitude with an uncertainty smaller than 0.22 mag; the absolute magnitude is corrected from the "Lutz-Kelker' " bias with the analytical formula given by Smith Jr (1987d). $M_{\rm spectro} - M_{\rm NDAC}$ is clearly correlated with $M_{uvby-\beta} - M_{\rm NDAC}$ (fig. 6.28): Kendall's τ is 0.3 of probability almost zero.

The reason of this correlation is not clear: no new binaries (from FAST early results), the absorption is small for these stars and the apparent magnitudes used are photoelectric. These stars are F, G or K dwarfs whose absolute magnitudes come from different calibrations.

We may then verify, for instance, if the absolute magnitudes used by Crawford (1975) for its calibration of absolute magnitude from $uvby{-}\beta$ photometry for F stars were not biased. The mean difference between the absolute magnitudes deduced from NDAC parallaxes and those used by Crawford for his calibration is -0.03 ± 0.06 mag (18 stars). Thus, we reject the hypothesis of a global shift of the absolute magnitudes.

It is hard to see in more details this problem as it would lead us to compute individual absolute magnitudes and to verify the absolute magnitudes calibrations.

Nevertheless, we may obtain an estimate of the dispersions of the absolute magnitude differences we used

	median		width	
$M_{uvby-\beta} - M_{\rm NDAC}$	0.02	$\pm$0.03	0.38	$\pm$0.03
$M_{\rm spectro} - M_{\rm NDAC}$	$-$0.13	$\pm$0.03	0.55	$\pm$0.04

The negative offset of spectroscopic absolute magnitudes is probably simply due to the Malmquist bias.

We then assume that the bias seen on fig. 6.24 and 6.25 is simply due to a statistical artefact which we compute in the next section.

Calculation of the bias – The bias was first calculated by L. Lindegren (1992c). We write it with a Bayesian approach. If π is the true parallax, we may show that the bias shown in fig. 6.24 is

$$E[\pi|\pi_{\rm P}] - \pi_{\rm P} = \frac{\int f(\pi_{\rm P}|\pi) f(\pi) \pi d\pi}{\int f(\pi_{\rm P}|\pi) f(\pi) d\pi} - \pi_{\rm P}$$

with $f(\pi_{\rm P}|\pi) = \alpha e^{-\frac{1}{2}\frac{25(\log \pi_{\rm P} - \log \pi)^2}{\sigma_M^2}}$ where α is a term independent of π, then vanishing in the equation above.

We then need an *a priori* pdf $f(\pi)$. For this purpose, L. Lindegren used a beta distribution of the second kind; using this pdf, the bias is shown in fig. 6.29). We use in the next section an *a priori* pdf obtained from external data.

Distribution of the true parallaxes – If we had an infinite population of stars, or a magnitude-limited sample, the pdf of the parallaxes would be straigthforward. Given the content of the Catalogue d'Entrée, an *a priori* law for the distribution of the true parallaxes is, in fact, difficult to find.

Smooth estimates of the pdf of observed spectroscopic and photometric parallaxes are plotted on fig. 6.30 and fig. 6.31, respectively. These two pdf are not supposed to be

the same, as the stars with $uvby{-}\beta$ photometry are probably brighter than stars with a spectral classification. The shape of these pdf shows that it is difficult to replace the distribution of the true parallaxes by a theoretical one.

We then use an estimate of the distance moduli using the spectroscopic and photometric absolute magnitudes. The errors on these distance moduli are supposed to be Gaussian. By trial and error we find an *a priori* law of the true distance moduli, we then do a convolution of this law with a Gaussian until we get a reasonable fit to the observed spectroscopic (or photometric) distance moduli distribution.

When this is done, we may compute the bias as explained in previous section and substract it from the differences $(\pi_{\rm H}-\pi_{\rm P})$ (and $(\pi_{\rm H}-\pi_{\rm S})$). The result is shown on fig. 6.32 and fig. 6.33. The mean differences $(\pi_{\rm H}-\pi_{\rm P})$ are not really different from zero, although still negative. The differences $(\pi_{\rm H}-\pi_{\rm S})$ are still too important, and this may be explained by an underestimation of the dispersions of the spectroscopic absolute magnitudes; in this case, we should multiply these dispersions by a factor of 2, which seems too important. The Malmquist correction we did is also perhaps underestimated.

If we really want to completely explain this effect, we must carefully study the absolute magnitude calibrations, which is not yet possible since we restrict the use of the preliminary parallaxes to validations only, and not to a preliminary scientific use.

Fortunately, we may still show that sdp of the preliminary parallax are independent of the parallax. This is done using the preliminary parallax as the reference parallax; we must correct from the bias coming from the errors on these preliminary parallaxes. We thus study $(\pi_{\rm N}-\pi)-E[\pi_{\rm N}-\pi|\pi_{\rm N}]=E[\pi|\pi_{\rm N}]-\pi=\pi_{\rm N}+\sigma^2_{\pi_{\rm N}}\frac{f'(\pi_{\rm N})}{f(\pi_{\rm N})}-\pi$ as a function of $\pi_{\rm N}$, using for π the photometric parallax (fig. 6.34). This may also be done using spectroscopic parallaxes (fig. 6.35).

Although there is a very small effect near 10 mas, we must admit that there are no variations of the sdp of preliminary parallaxes with the parallax itself. We may also notice that the correction of the statistical biases is much more easily done using preliminary parallaxes than using spectroscopic or photometric parallaxes. This is probably due to their good statistical properties (Gaussian measurement errors, few outliers). Another point is that, if we want to use spectroscopic or photometric parallaxes, we will have to carefully study their errors.

As a summary of this section, we may conclude that the sdp of preliminary parallaxes are independent of the parallax and that there is no visible global shift of the spectroscopic or photometric absolute magnitudes for stars closer than 50 pc; for these stars the dispersion of spectroscopic absolute magnitudes is about 0.5 mag (0.35 mag for the absolute magnitudes obtained from $uvby{-}\beta$ photometry).

Independence of sdp of the parallaxes from astrometric and photometric data

This section addresses an important problem: we must show that the sdp of the preliminary parallaxes are independent of the characteristics of the stars. In order to avoid any circular reasonning, we obviously use data coming from an external source of data (the Catalogue d'Entrée).

The errors on preliminary parallaxes are computed using spectroscopic parallaxes smaller than 2 mas. This choice is simply due to the number of stars available. Since we truncate the observed spectroscopic parallax distribution, we know that we will have a bias on $\pi_{\rm H}-\pi_{\rm S}$ of about 0.4 mas. This offset (which must be substracted from the

comparisons to come) is not important because we study here the variations of the sdp and not the sdp themselves.

As a function of positions – Figure 6.36 points out the non-uniform distribution of sdp of NDAC-5P parallaxes, mainly in the ecliptic region. For FAST-3P parallaxes this is not the case but the sdp are oscillating around their average (fig. 6.37d).

There is clearly no independence between the sdp of the parallaxes and the position. More quantitatively, we only need to find two bins where the sdp are significantly different. This is the case for $40° < \delta < 50°$ and $50° < \delta < 60°$ on fig. 6.36b. This is also the case for $-60° < \delta < -50°$ and $50° < \delta < 60°$ on fig. 6.37b.

As a function of proper motions – Variations of the sdp of the preliminary parallaxes with the proper motion are interesting for both the 3P and the 5P sphere solutions. In the first case, this may show if the ground-based proper motions do contaminate the 3P parallaxes; to verify the result, we also use the NDAC-3P solution. In the second case, we may think that the one-year proper motions contaminate the 5P parallaxes.

The figures 6.38, 6.39 and 6.40 concern the NDAC-5P, FAST-3P and NDAC-3P parallaxes, respectively. For stars with a proper motion close to 0 (± 0.005"/year), the sdp of the preliminary parallaxes are close to their average (≈ 0.5 mas). For the other stars, there are important variations.

We may then ask whether these variations are due to the spectroscopic parallaxes: a nearby star, wrongly classified, should have a noticeable proper motion, and thus the differences $\pi_{\rm H} - \pi_{\rm S}$ should be positive. Given that a) we notice on the graphs some negative mean differences and b) we rejected the differences greater in modulus than 3 times the formal error, we may assume that it is due to the preliminary parallaxes.

Concerning the 5P solution, we notice on fig. 6.38 a parabolic shape; this may be explained in the following way: the greater proper motions were not accurately found after one year, increased the parallax, thus creating a positive $\pi_{\rm H} - \pi_{\rm S}$ difference. A Kendall's τ between the error distribution and the proper motion (in modulus) lead us to reject the independence hypothesis.

Both 3P solutions are similar; in right ascension the sdp are slightly negative for negative proper motion. There is a dependence between the sdp and proper motion; for example, into the bin $-15 < \mu_\delta < -10$ mas/year, the sdp are significantly different from their global average.

As a function of magnitudes and colours – Finally, we must verify whether the sdp of the preliminary parallaxes vary with magnitude and colour, since a chromaticity problem could lead to problems concerning the determination of the astrometric parameters.

The error bars are important and there are great variations with the colour, the sdp reaching 2 mas when $B - V \approx 0.9$. These problems are perhaps also related to the way each DRC reduced the data, because the two NDAC solutions show similar variations. The size of the error bars in each bin prevents us from showing a significative dependence, although it probably exists.

Concluding remarks about these comparisons – As we showed before, there is no independence between the sdp of the preliminary parallaxes and the astrometric or photometric data, although the level of dependence is not too high, given that the parallaxes were obtained after only a one-year sphere solution.

Many other comparisons could be done and should be done when the one year and a half sphere solution will be available: we only did the comparisons with the parallaxes, as the proper motions were not at our disposal. With the future sphere solutions, with more stars, it will also be possible to use other data (photometric one) which will enable us to analyse more precisely the variations of the sdp.

E.6.6 Zero-point of the preliminary parallaxes

Although it is not yet true, we assume in what follows that the sdp of the preliminary parallaxes are independent of the characteristics of the stars, and that the sdp represent in fact the global zero-point.

Direct estimation

The sdp being independent of the parallax, we only need to study the more distant stars. These stars must be numerous enough to get the best accuracy on the global zero-point.

Concerning the number of stars, we have not enough Magellanic cloud stars; in the case of cluster stars, we have the problem of non-members. We then only use the spectroscopic and photometric parallaxes for the purpose of finding the zero-point.

We know that we would have a bias if we take the more distant stars because of the truncation of the observed distribution. In order to suppress or, at least, to have a smaller bias, we replace the estimator of the spectroscopic (or photometric) parallax by the Bayesian estimator $E[\pi|\pi_{\rm S}]$ (or $E[\pi|\pi_{\rm P}]$); we thus define the zero-point by $z_H = \pi_{\rm H} - E[\pi|\pi_{\rm S}]$ for $\pi_{\rm S} < 2$ mas. We also denote k_H the average of the ratio of the external error over the formal error; for example $k_N = \langle \frac{\sigma(\pi_{\rm N} - E[\pi|\pi_{\rm S}])}{s_{\rm N}} \rangle$. This quantity is a kind of square root of a unit weight variance which should be multiplied by the formal error on one parallax in order to get the external error. In the best case, we would have $z_H \approx 0$ and $k_H \approx 1$.

However, we must verify two points: the zero-point should be independent of a) the limit on the observed parallax (2 mas here), b) the formal errors on preliminary parallaxes.

Concerning the first point, we study the sdp with the deciles (195 stars) of the spectroscopic parallax up to 2 mas. Table 6.10 (column 1-2) shows that the zero point increases then decreases. This means that we could not succeed to suppress the bias (we have the same behaviour as in fig. 6.29).

Concerning the second point, we study the sdp with the deciles of the formal errors on the preliminary parallaxes. Table 6.11 (column 1-2) shows that there is also a variation of the zero-point with the formal error.

The situation could then seem desesperate. Fortunately, we may use spectroscopic and photometric parallaxes in another way. We define the zero-point as the difference $z_H = \pi_{\rm H} - \pi_{\rm P}$ for $\pi_{\rm S} < 2$ mas (or $z_H = \pi_{\rm H} - \pi_{\rm S}$ for $\pi_{\rm P} < 2$ mas). This method may be used if the errors on spectroscopic and photometric parallax are uncorrelated. We showed

that it was not the case for nearby stars, but we assume that it may be the case for distant stars.

Table 6.10 (column 5) represents the variation of the zero-point with the limit on the spectroscopic parallaxes. ($\sigma_z \approx 0.37$ mas for 30 stars in each decile). Although there are some significatively different zero-points (probably due to outliers), there is at least no statistical bias due to the errors.

The drawback of this method is that we only have less stars at our disposal. We may notice that there are no more variations of the zero-point with the formal errors of the preliminary parallaxes (table 6.11, column 5).

Summary of the comparisons – Concerning the estimates of the zero-point and of the external errors, obtained with the two first moments of the distribution of the errors, we may now conclude. Table 6.12 and 6.13 give the estimates of (z_F,k_F) and (z_N,k_N) for FAST-3P and NDAC-5P, respectively.

Given the different problems we discussed before, we may not have a definitive answer for the errors on these preliminary parallaxes... However, we may notice that the zero-point is close to 0 and that the external errors are closer to the internal errors in the case of the 5P sphere solution than in the case of the 3P solution.

Finally, we look for another method to find z_H et k_H in the next section.

Estimation with distribution functions

Given that the distributions of spectroscopic and photometric parallax errors are lognormal, we know that for distant stars, the distribution of these parallaxes must be close to the distribution of the true parallaxes.

This means that, for the most distant stars, we may use the pdf of spectroscopic or photometric parallaxes as an *a priori* pdf of the true parallaxes, from which we may compute the observed density (of FAST or NDAC parallaxes).

It may seem surprising to use a marginal pdf (the spectroscopic or photometric pdf) as an *a priori* pdf. We made in fact a simulation with the photometric pdf which showed that the initial pdf and the marginal pdf were almost identical up to about 3 mas, becoming to be very different after 5 mas.

In practice, we work with the df instead of the pdf and we show in this section that the comparison for the smallest parallaxes between the theoretical df and the observed df may give an estimate of the global zero-point and of the external errors simultaneously.

Compared to the estimations we did in last section, this method seems more robust, as we use the pdf of spectroscopic (or photometric) parallaxes, not the individual values of these parallaxes. Furthermore, there are probably no biases to expect from such a comparison. Finally, the observed df of the preliminary parallaxes is easily found by computing the number of stars, without transforming the preliminary parallaxes.

Given the pdf of the true parallaxes $f_\pi(\pi)$, and the pdf of the standard errors on the observed parallax $f_{\sigma_\pi}(\sigma_\pi)$, and assuming that the errors are independent of the parallax, the distribution function of the preliminary parallaxes is:

$$F(\pi) = \int_{-\infty}^{\pi} \int_{0}^{+\infty} \int_{0}^{+\infty} f(\pi_{\rm H}|\pi,\sigma) f_\pi(\pi) f_\sigma(\sigma) d\pi d\sigma d\pi_{\rm H}$$

This theoretical df may be compared to the observed df:

$$F_n(\pi) = \sum_{i:\pi_{\mathrm{N}i}<\pi} \frac{1}{n}$$

We thus compute $F(\pi)$ using the pdf of the spectroscopic (or photometric) parallaxes for $f_\pi(\pi)$, and the pdf of the standard errors on NDAC (or FAST) parallaxes for $f_\sigma(\sigma)$. The computation is done numerically with the Gauß method; the computation of the observed df is much more easier...

The comparison between theoretical df and the NDAC observed df is shown in fig. 6.41 and 6.42, using the spectroscopic and photometric pdfs, respectively. The same comparison is done for FAST df, in fig. 6.43 and fig. 6.44.

The theoretical df should have a slight offset from the observed df if the global zero-point is different from 0. We may see such an offset on fig. 6.43 but it may be due to the spectroscopic parallaxes, less precise than the photometric parallaxes. This is why we use the latter first.

The two global parameters z and k are then introduced into the conditional pdf of the errors:

$$f(\pi_{\mathrm{H}}|\pi,\sigma) = \frac{1}{k\sigma\sqrt{2\Pi}} e^{-\frac{1}{2}(\frac{\pi_{\mathrm{H}}-(\pi+z)}{k\sigma})^2}$$

In order to compare two distributions, it is natural to use the Kolmogorov test, more powerful than the χ^2 test. This is why we choose the statistic $K_n = \sup_\pi |F_n(\pi) - F(\pi)|$ as criteria of goodness of fit between $F(\pi)$ and $F_n(\pi)$.

One could think that we may have an erroneous solution if we use the greater parallaxes, as the pdf of photometric parallaxes should more and more deviate from the true pdf. This is why we show on table 6.14 the different solutions obtained with different limits on the FAST and NDAC parallaxes. There is clearly no variation.

We did also some simulations from which we found that we could obtain an uncertainty of about 0.03 on k and z. This is quite small, but it does not take into account the uncertainty on the *a priori* pdf. However, we use here about 500 stars (up to 2 mas) while this number will increase when all the Hipparcos parallaxes are available.

We then obtain

$$\begin{array}{rcrcl rcrcl} z_F & = & -0.03 & \pm 0.027 & \text{mas}, & k_F & = & 1.31 & \pm 0.028 & \text{for FAST-3P} \\ z_N & = & -0.04 & \pm 0.027 & \text{mas}, & k_N & = & 1.10 & \pm 0.028 & \text{for NDAC-5P} \end{array}$$

and with these values, we may see on figs. 6.45 and 6.46 that the fit is very satistactory.

We may notice that the zero-point is not significantly different from 0, for FAST-3P and NDAC-5P preliminary parallaxes, and also that the difference between the zero-points found ($z_F - z_N \approx 0.01 \pm 0.038$) is not different from the mean difference between FAST and NDAC parallaxes ($\langle \pi_{\mathrm{F}} - \pi_{\mathrm{N}} \rangle \approx -0.04 \pm 0.02$ mas).

Applying the same method, but with the pdf of spectroscopic parallaxes as the *a priori* pdf, we obtain

$$\begin{array}{rcrcl rcrcl} z_F & = & 0.09 & \text{mas}, & k_F & = & 1.27 & \text{for FAST-3P} \\ z_N & = & 0.07 & \text{mas}, & k_N & = & 1.02 & \text{for NDAC-5P} \end{array}$$

using the $\approx 2\,200$ parallaxes smaller than 2 mas, but the fit is not as good as using the photometric pdf.

E.6.7 Conclusions and prospects

Hipparcos parallaxes

This work is only tentative since it is applied to the preliminary parallaxes. However, we have found different methods in order to obtain the external error on each parallax and the global zero-point, trying to use primarily the observed data, not models which could become obsolete in the future.

We have shown that the two parallax distributions of the Consortia allow us to find an estimate of their "external' " errors and of the correlation between their errors. This has been done with internal comparisons, using the hypothesis that errors are normally distributed only. We have done also real external comparisons.

The first external method uses the distant stars, and, after correcting for various biases, computes the zero-point with the mean difference between preliminary parallaxes and other parallax estimates, and the square root of the unit weight variance which gives us the external error on each parallax. The second one is a fit betwen the theoretical distribution function and the observed df. It gives us other estimates of the same quantities k and z as above, using an *a priori* law of the true parallaxes. We have also shown how to obtain the best estimate of an Hipparcos parallax given the FAST parallax and the NDAC parallax.

Concerning the obtention of the final Hipparcos parallaxes, we could then proceed as follows:

1. We may use the internal comparison and the first external method in order to check and suppress outliers and then obtain the external errors; the variation of these errors could be studied as a function of parameters such as magnitude, latitude;

2. The ratio between external and formal error being fixed, the global zero-point could be obtained with the fit of the dfs;

3. Once the parallaxes of each Consortium have been corrected from their global zero-point (which would be the same in the best case), the final individual Hipparcos parallaxes would be found with the best estimator.

One would have noticed that we did not exactly obtain the same results with the different methods, although these results are not contradictory. We believe it is due to the preliminary nature of the parallaxes we used: we have shown that there were remaining variations with the astrometric or photometric data. The methods we used probably need also to be refined.

One thing is certain, however: the preliminary parallaxes are of a very high quality and the solutions found by the two Consortia are close.

The prospects of the work we did are the validation of the final Hipparcos parallaxes but also the calibration of the spectroscopic mean absolute magnitudes in which the Hipparcos parallaxes will play a major role.

Calibration of absolute magnitudes

We largely used the spectroscopic or photometric absolute magnitudes in the preceding pages.

From a methodological point of view, many papers have been devoted to the calibration of an homogeneous group of stars with the trigonometric parallaxes: "Lutz-Kelker' ' bias on the observed parallax, Malmquist bias on the mean absolute magnitude, censured data, outliers: this originally astronomical problem becomes a statistical problem.

The precision of Hipparcos parallaxes will not change the problem: compared to the existing parallaxes, the biases due to measurement errors will decrease for a given star but we will call for a better precision on absolute magnitudes; although smaller in size, the biases will be still present.

On the contrary, we will be faced to the problem of dealing with numerous data (possibly noisy or with outliers) with the appropriate methodology. This implies not only the use of the correct statistical methods but also much computing time. This last point is not really important if we think about the time it took to obtain the Hipparcos parallaxes (about 10 years).

In order to make a review of the problems concerning the absolute magnitude calibrations, we may begin with Malmquist (1920, 1936) who calculated the bias on the mean absolute magnitude of a group of stars. Concerning the parallaxes, Trumpler & Weaver (1953, p. 369) mention the fact that the non-uniform distribution of the parallaxes combined with their measurement errors introduce a bias when the observed distribution is truncated. The implications for the absolute magnitude calibrations are studied by Lutz & Kelker (1973), who calculated a correction to the absolute magnitude. This is completed by Lutz (1979) for a magnitude-limited sample. Concerning the corrections, Hakkila (1989) studied the mean absolute magnitude of a distance-limited sample.

Turon & Crézé (1977) recommended to use a maximum likelihood estimate on the whole distribution of the observed parallaxes, without truncating this distribution.

With a Bayesian point of view (the above mentioned corrections being implicitly Bayesian), Smith Jr (1987a-d) gave a consistent picture of the estimate of the absolute magnitudes using the trigonometric parallaxes, and also showed how to get the estimate of the most probable parallax given the trigonometric parallax and the spectroscopic parallax.

Recently, Ratnatunga & Casertano (1991) developed an algorithm which was applied to the calibration of the absolute magnitude using the $R-I$ colour index. They already used this approach for a kinematic study of galactic populations, and we may briefly describe it.

Using a parametric *a priori* model $M_V = f_\Theta(R-I)$, the marginal density of the observed parallax is computed for each star, then the conditional density given the model is computed, and finally the value of the parameters Θ is obtained by maximum likelihood.

Unlike the other estimates mentioned above, this approach uses only the observed data and, for instance, does not try to obtain the individual absolute magnitudes. The selection biases are taken into account, and also are the censured data; there is also a test of outliers. This very interesting approach could be used for instance to obtain the absolute calibration of $uvby{-}\beta$ photometry, using the various indices, the rotation, etc.

When the final Hipparcos data will be obtained, there is little doubt that we will be concerned with the HR diagram calibration. While waiting for these parallaxes, we will study the appropriate statistical methodology.

KINEMATICS

In this part, we tackle a problem in stellar kinematics of population I A-type stars. This last part uses also the data and part of the results we obtained in the preceding chapters.

We use the distances obtained with the calibrations of $uvby-\beta$ photometry (chap. 2); the other data (apparent magnitudes corrected from interstellar extinction, proper motions, radial velocities) come from the INCA data base.

We show in this part that the spatial velocity distributions may be more adequately explained as the sum of independent approximately spherical distributions, part of these groups being made of stars with a common origin. The consequence is that the dynamical mixing time is greater than 2 10^8 years, value usually adopted. Concerning the methods used to validate the separation between groups, they are described in chapter 4.

E.7 Local velocity distribution of A V-type stars

E.7.1 Space velocities

Our Galaxy is a differential rotational system, the rotation period being about 2.4 10^8 years. The linear rotational velocity near the Sun is about $V_0 = 220$ km/s.

We use the Cartesian coordinates (X, Y, Z) to describe the position of a star with respect to the Sun, where X is in the direction of the galactic center, Y in the direction of the galactic rotation, and Z towards the north galactic pole. We denote (U, V, W) the corresponding components of the space velocity with respect to the Sun, corrected from differential rotation.

If we consider an homogeneous group of population I stars, the U, V and W components are usually considered as normally distributed, with means $\overline{U}$, $\overline{V}$, $\overline{W}$ and dispersions $\sigma_U, \sigma_V, \sigma_W$, respectively. This defines the so-called velocity ellipsoid, with $\sigma_U > \sigma_V > \sigma_W$, the direction of the major axis being the vertex direction, and the angle between this direction and the direction of the galactic centre being the vertex deviation. One may notice that the vertex deviation is greater for the youngest stars and that the velocity dispersion increases with age. The increasing of σ_U contributes to the asymmetric drift: the stars from groups with high σ_U have the tendency to lag behind the LSR.

For a given group of stars, $(\overline{U}, \overline{V}, \overline{W})$ represents the mean velocity of the group and $(-\overline{U}, -\overline{V}, -\overline{W})$ is the solar motion. The $v_\odot$ peculiar velocity component of the Sun with respect to the Local Standard of Rest (LSR, fictitious point having the circular speed V_0) differs from $-\overline{V}$ by an amount given by the asymmetric-drift. In order to be able to compute the Sun velocity components, the velocities must be well "mixed' ', which is simply not possible with the youngest stars.

In order to explain the increase of the velocity dispersion with age, Wielen (1977) supposes the existence of local fluctuations of the gravitational field in our Galaxy, causing a diffusion of stellar orbits in phase space; the consequence is that a disk star would change its space velocity at random by more than 10 km/s per galactic revolution and that the age-dependence of the velocity dispersion would be $\propto t^{\frac{1}{2}}$. The mixing time of disk stars would be about 2 10^8 years. Lacey (1984) and Paloŭs & Piskunov (1984) find also a mixing time of about 2 10^8 years.

The goal in this chapter is to analyze the velocity distributions of stars older than two galactic years ($> 4\ 10^8$) in order to find groups of stars which share the same kinematical behaviour. More generally, we could not say better than Eggen (1965): «*It is usual in applying the various statistical procedures used in the study of stellar motions to assume that these motions are randomly distributed with, at most, only minor variations. If in fact the observed motions are dominated by those of a relatively few stellar groups, then many of these procedures may be invalid.*» We show in what follows that the velocity distributions are described more simply as the mixture of some different groups whose velocities are spherically distributed.

We first compute the space velocity components and their formal errors.

Calculation of space velocities – From components of the proper motion ($\mu_\alpha \cos\delta$, μ_δ) of a star, its radial velocity v_r, and its heliocentric distance r, and after correcting from differential rotation, the components of the space velocity may be written:

$$\left|\begin{array}{lcllll} U & = & (m_{11}\mu_\alpha \cos\delta + m_{12}\mu_\delta + m_{13}) & r & +m_{14}v_r \\ V & = & (m_{21}\mu_\alpha \cos\delta + m_{22}\mu_\delta + m_{23}) & r & +m_{24}v_r \\ W & = & (m_{31}\mu_\alpha \cos\delta + m_{32}\mu_\delta + m_{33}) & r & +m_{34}v_r \end{array}\right.$$

where m_{ij} depends upon the galactic coordinates of the star, the coordinates of the galactic center; the m_{i3} terms depend also on the Oort constants and allow us to correct the velocities from differential rotation.

We compute the formal errors on U, V and W, as a function of s_r^2, $s_{\mu_\alpha \cos\delta}^2$, $s_{\mu_\delta}^2$, $s_{v_r}^2$, assuming that errors on proper motion, radial velocity and distance are independent.

We also assume that the distribution of the errors on U, V and W are normally distributed. In our sample, the averages of the formal errors are $\langle s_U \rangle \approx 3.3$, $\langle s_V \rangle \approx 3.1$ and $\langle s_W \rangle \approx 2.9$ km/s, respectively.

E.7.2 Star formation bursts

The included paper is devoted to the study of the kinematical properties of a magnitude-limited sample of dwarfs B5 to F5, closer than about 250 pc. The velocity ellipsoid is explained as a sum of some spherical distributions corresponding to star formation bursts.

Validation of the separation into subgroups – The first question is whether the modes observed in the U distribution are real or purely random. Using for instance the sample of A1 V stars, there are two modes, one between -20 and -10 km/s (28 stars), and the other between 0 and 10 km/s (21 stars). If these numbers of stars in a given interval

$I = 10$ km/s are considered as Poissonian, the probability to get more than 20 stars is $1 - \sum_{i=0}^{20} e^{-(\lambda I)} \frac{(\lambda I)^i}{i!} \approx 0.0001$. As a consequence, the two modes are significant.

Using a more interesting approach, if we consider the subgroups to be normally distributed, the number of components may be tested with a Wilks test: (H0)=K components against (H1)=$K' > K$ components. For the A1 V and A2 V samples, the significant number of components is 2.

Different papers (see for instance page 86) have been devoted to the separation into subgroups: Arenou (1990), Arenou & Bougeard (1992), Bougeard & Arenou (1989), Bougeard *et al.* (1989a), Bougeard *et al.* (1989b), Robert & Soubiran (1991), Soubiran *et al.* (1989). For the A2 V sample, the different methods (parametrical or not) give about the same results, leading to the conclusion that there are surely two well-defined subgroups.

The problem is that the analysis has been done using kinematical data only, and a very indirect age determination. We thus decided to complete this study with complementary data.

E.7.3 Age, metallicity and kinematics

The data – Let us we follow the definition given by Norris *et al.* (1985): «*a stellar population is characterized by the trivariate function describing the distribution of its component stars with respect to age, composition, and kinematics*».

This must be also true for star bursts: within one burst, we expect the stars to have the same kinematical behaviour, the same chemical composition of the interstellar medium where they formed, and a similar age.

In the joint paper, the age of the stars is clearly a discriminant parameter; mean ages were deduced only from the spectral class. The astronomical team of Barcelona University computed ages for our stars, using $uvby{-}\beta$ photometry and Maeder & Meynet evolutionary tracks. For younger stars near the ZAMS, we only have an upper limit for the age.

The number of stars with a given spectral type, having kinematical data and ages was not very high; consequently, we built a new sample from the INCA data base. This sample contains 369 stars of A0-A4 spectral type, spectroscopically classified as dwarfs, brighter than $m_V = 8.5$, having known proper motion, radial velocity, $uvby{-}\beta$ photometry and age.

$uvby{-}\beta$ photometry allows us to get an absolute magnitude, and then a photometric distance, used to compute the space velocities. It also allows us to obtain an indicator of metallicity $\delta m_0 = m_0^{\text{Hyades}} - m_0$ (see page 24). The mean value of δm_0 is thus 0 for the Hyades, positive for deficient stars, and about 0.018 for the Sun. Moon (1985) gives the following classification:

	δm_0	< -0.010	Am and Ap stars
$-0.010 <$	δm_0	$< +0.025$	Normal population I stars
$+0.025 <$	δm_0	$< +0.045$	Older population I and old disk stars
$+0.045 <$	δm_0	$< +0.090$	Intermediate population II stars
$+0.090 <$	δm_0		Extreme population II stars

The distributions of the sample stars in U, V, $\log t$ and δm_0 are indicated on fig.7.1. Most of the stars have an age between 4×10^8 and 6×10^8 years.

The distributions of the sample stars in the components X, Y, Z of the position relative to the Sun, and in apparent magnitude m_V are indicated on fig. 7.1. The sample is basically magnitude-limited, and the stars are closer than about 200 pc.

Using this sample, which does not exactly coincide with the sample used in the joint paper, we study whether the velocities are well mixed or not for the older stars. If we keep the stars older than two galactic years ($\log t > 8.7$), we may show that the velocity components are not normally distributed: all the normality tests based upon the skewness reject the null hypothesis. Using Lilliefors normality test, the normality hypothesis is rejected for U ($P < 0.01$) and for V ($P < 0.006$). This, and the fact that there are two modes in the U distribution, allow us to look for different groups in our sample.

Variation with age – We first may ask whether the kinematical properties are independent of the age. The answer is yes, a Kendall test reject the independence hypothesis concerning V ($P < 2\ 10^{-6}$). The younger stars correspond to the most negative values of V. Before we study how age and metallicity are related to the kinematical behaviour, we introduce the integrals of motion in order to find what they imply for the survival time of a given group.

Integrals of motion

If the stars of a given burst are now together, these stars should have integrals of motion which are nearly the same; if this was not the case, the stars would have become field stars, lost along the galactic rotation.

We use a simple model, asuming that our Galaxy is stationary and axisymmetrical. The density phase function is a function of the isolating integrals of motion.

The first integral of motion is the energy integral:

$$H = \frac{1}{2}(u^2 + (v + V_0)^2 + w^2) + \phi(R, Z)$$

where $(u, v, w) = (U+u_\odot, V+v_\odot, W+w_\odot)$ are the components of the peculiar velocity with respect to the LSR, and V_0 the circular velocity. We denote R_0 and R the galactocentric distances of the Sun and of the considered star, respectively, and $\phi(R, Z)$ the galactic potential in cylindrical coordinates.

The second integral of motion is the angular-momentum integral:

$$h = R(v + V_0) = R(V + v_\odot + V_0).$$

It could exist a third integral but we don't have its analytical expression in our case and it is beyond the scope of this work.

We use the potential of Carlberg & Innanen (1987):

$$\phi(R, Z) = -\sum_{j=1}^{4} \frac{\mathcal{M}_j}{\sqrt{(a_j + \sum_{i=1}^{3} \beta_{i,j} \sqrt{Z^2 + h_i^2})^2 + b_j^2 + R^2}}$$

$\mathcal{M}_j$ being the mass of the considered component (disk-halo, bulge, nucleus, dark halo), b_j the core radius of each halo component, a_j are the scale length of the disk ($a_1 = 3$ kpc, $a_{j\neq 1} = 0$), the sum corresponding to three disk components (old disk, dark matter, young disk) of different scale height h_i ; $\beta_{i,j}$ are weights which sum to one concerning the disk and which are zero otherwise.

We use $(u_\odot, v_\odot, w_\odot) = (9, 12, 7)$ km/s [Delhaye, 1965] for the peculiar velocity of the Sun, $R_0 = 8.5$ kpc for its distance from the Galactic center and $V_0 = 235$ km/s for the circular velocity of the LSR [Carlberg & Innanen, 1987].

The (H, h) diagram (Lindblad diagram) of our sample is drawn on figure 7.3. The units are 10^3 km^2.s^{-2} for the energy $-H$ and 10^2 kpc.km.s^{-1} for the angular momentum h. The form of the diagram is due to the fact that there is only a small variation about $H = \frac{1}{2R_0^2}h^2 + \phi(R_0, 0)$, because the stars are close to the Sun and the peculiar velocities are small compared to the circular velocity. The stars are on epicyclic orbits and $v + V_0$ is the most discriminant variable which allows the stars to separe one from another, or to encounter periodically.

In order for stars in a group to stay together, their energy and angular momentum should be close together and thus close on a Lindblad diagram. It is difficult to find a critical size for a group on this diagram, but we may find an upper bound, using the open clusters of the same age range.

In increasing order of age, the figures 7.4, 7.5 and 7.5 show the Lindblad diagrams of the Pleiades, Coma Ber and the Hyades, respectively. On each graph we have a small area surrounded by some stars which could possibly be escaping from the cluster, or be non-members. These diagrams do not take into account the gravitational forces between cluster members; they may still be interesting in order to find non-members or to discover former members among the field stars.

Comparing the size of a cluster on the diagram and the total size of our sample, we then find that there could be at least three non-overlapping subgroups in our sample in order that each subgroup may have survived.

We arrive to the same result if we consider only the angular momentum: if the stars are still together, separated by a distance smaller than $\delta R = \pm 250$ pc, the variation of angular momentum must be less than $\delta h \approx V_0 \delta R \approx \pm 0.6\ 10^5$ pc^2/(10^6 years); considering the range of h in our sample, this also implies a minimum of three groups.

Separation between subgroups

In order to find subgroups in our sample with small differencies in kinematics, age and chemical composition; we use U, V, $\log t$ and δm_0 data for this purpose.

We first used the SEMMUL algorithm, taking measurement errors into account (cf §4.4.1), and we found 4 subgroups. One may object that, although the distribution of velocity components into one subgroup is expected to be Gaussian, the distribution of ages has no reason to be normally distributed. That is why we used another classification method, a hierarchical clustering, where the clusters were obtained with a minimum variance approach. We kept the first 4 levels of the hierarchy.

The subgroup n°1 is separated from subgroups n°2 and n°3 in U. Subgroup 3 is created by $\log t$: these are the youngest stars. Subgroup n°4 is created by δm_0, and is made of the more deficient stars. The four groups are shown, for each couple of variables, on figure 7.7 ;

the units of the axis are number of standard deviations because the data were normalized before clustering.

A principal component analysis (figure 7.8) shows that the whole variables are well represented on both axes, the first (41% of the variance) and the second (31%). The second axis separates U and V (correlation coefficient $\rho = 0.45$) from $\log t$ and δm_0 ($\rho = 0.45$). This suggests a rotation of the axis: the second diagonal is an age axis and it shows the separation between a young subgroup with a normal chemical composition (n°3) and an old, deficient group (n°4); the first diagonal is a kinematical axis, discriminating the two groups n°1 and n°2.

The most peculiar group is the forth one: it seems the older, the δm_0 are too high; but we may note that the average $V \sin i$ is 177 ± 10 km/s, significantly different from 136 ± 6 km/s (mean value of the other groups) and is close to the value corresponding to B5 V stars.

This would lead us to presume that this group is younger than expected. If we also notice that the ages were computed without taking the $V \sin i$ into account, this confirms that there is probably a problem with the ages determination (we recall, see §2.4.5, that the absolute magnitude computed without taking the $V \sin i$ into account gives an absolute magnitude too bright).

Finally, the last problem in this group is the presence of several λ Boo-type stars. These stars are characterized by weak metallic lines. The observations of these stars show that they have a low radial velocity and a high rotational velocity. Gerbaldi *et al.* (1992) argue that these stars are just arriving at the main sequence, not departing from it. Another point is that the evolutionary tracks used to estimate the ages were based on a solar chemical composition. Consequently, the ages we used may be wrong.

We will study this problem in the future. For the moment, we just note that the ages and the space velocities (computed using the photometric absolute magnitude) are doubtful. This group itself probably is a mixture of populations with different kinematical behaviours. Fortunately, the clustering helped us to find this group, and we will not take it into account in what follows.

Table 7.1 shows, for each group, the group number, the number of stars, the most probable value of the age, the average of δm_0, and the mean value and the dispersions of U and V velocity components. The last column shows that the distributions are approximately spherical, except for the last group.

In what follows, we will not consider the group n°3, because it is too young, and the group n°4, for the reasons mentioned above. According to the clustering, the groups 1 and 2 are approximately homogeneous as far as the age and chemical composition are concerned, but these two groups have clearly different kinematical properties.

The kinematical properties of subgroup n°1 are close to those of U Ma group, but with a higher dispersion and an older age. This age is, in fact, not well established, Eggen (1983) gives 2.7 10^8 years and Paloŭs & Hauck (1986) give 4.9 10^8 years. Our group n°1 has an age close to this latter estimate ($\log t \approx 8.71$ vs 8.69). The age difference between Eggen and Paloŭs & Hauck may be explained by the fact that the former considered [Fe/H]=−0.1 for U Ma and the latter used evolutionary tracks corresponding to a solar composition.

The existence of groups with different kinematical characteristics is well established. We found nearly the same groups using SEMMUL parametric algorithm mentioned above (except for the youngest group, because of the large standard errors on $\log t$). Moreover, we also did a classification using $-H$, $\log t$ and δm_0 as variables and we found the same groups. The variables $\log t$ and δm_0 are probably not discriminant enough, and the kinematical data still have too large measurement errors.

Nevertheless, our results do not enable us to confirm or reject the hypothesis that these groups are the signature of bursts: they spread a little too much on a Lindblad diagram, but the velocity dispersions are small: the first group is probably made of former members of U Ma.

In order to have a more precise conclusion about these groups, we have to wait the parallaxes and the proper motions given by Hipparcos, and the radial velocities currently measured. This will allow us not only to increase the size of the samples but also the precision of the kinematical data.

We may note that the remnants of distinct groups with different kinematics contributes to the vertex deviation: the superposition of groups n°1 and n°2 is sufficient to prevent the velocity ellipsoid to be in the direction of the Galactic center.

The use of complementary data and of multivariate data analysis allows us to find in our sample homogenous groups with respect to age and composition, but with different kinematical properties. The existence of group n°1 being clear, we may conclude that the velocities are not well mixed after two Galactic years in the solar neighbourhood. This implies that it is not possible to use the asymmetric-drift relationship in order to compute the peculiar velocity of the Sun without a precise age criterion: a minimum of 10^9 years seems to be required in order to compute the Sun velocity.

Conclusion

This thesis has been devoted to the stellar data which were used for the preparation of the Hipparcos Catalogue d'Entrée and to the preliminary data given by the satellite. We described different methods and we applied them on various areas: estimation of the interstellar extinction, study of the kinematics of nearby A-type stars, validation of preliminary Hipparcos parallaxes.

We described the data and the software which were necessary to use the INCA database. We described also the photometric and spectroscopic calibrations which were implemented. This was useful to obtain the photometric and spectroscopic parallaxes after having corrected them from various statistical biases.

For this purpose, we studied the influence of measurement errors on data. We obtained estimations of pdf, of minimum variance unbiased estimates, and conditional estimates. These estimates were used for studying Hipparcos parallaxes and also to obtain the parameters of mixtures of Gaussian components. We also described the simulations we did and the statistical tests we applied.

We thus developed or implemented various softwares, dealing with astronomy, statistics or data analysis.

We then built a model of interstellar extinction which allows us to obtain a realistic approximation of extinction in the visible band, up to several hundreds of pc. This model could be useful for all the statistical studies, for instance a galactic model or the absolute magnitude calibration.

Up to now, the existing models were only rough estimates, which did not take into account the variations of the distribution of interstellar matter.

As we contributed to the Hipparcos Catalogue d'Entrée preparation, we were naturally induced to compare the first astrometric and photometric data from the satellite with the Catalogue d'Entrée content. These comparisons showed the good quality of both space and ground-based data. We showed in particular that – apart from minor effects which will need to be explained – the precision of the Catalogue d'Entrée positions and magnitudes were clearly better than the initial ESA specifications.

We also proposed, from a methodological point of view, different analysis which may be useful to validate the future Hipparcos parallaxes. We applied these methods to the preliminary results and we showed their quality. Using these preliminary data, we obtained several methods to find estimates of the external errors on the parallaxes from both DRC, estimates of the global zero-point shift of the parallaxes, and we have pointed out the problems we are faced when the photometric or spectroscopic parallaxes are used. We also studied the variation of the systematic errors as a function of various parameters. All these comparisons and estimations are implemented in a software which allows us to verify quickly the quality of each intermediate parallax solution.

We also indicated the best way to combine the parallaxes of both Consortia in order to obtain the best estimate of each final parallax together with its formal error.

Finally we studied the kinematics of A V in the solar neighbourhood. We showed, using multivariate analysis methods, that the mixing time of space velocities is greater than two galactic years, contrary to what is usually assumed. We showed that the local velocity distribution could be explained as a contribution of some groups of stars with different kinematical behaviours. As a consequence, stars younger than about 10^9 years may not be used to estimate the sun peculiar velocity.

The various themes we studied in this thesis will greatly benefit from future Hipparcos data. The extinction will be known more accurately, using the photometric data from Tycho experiment and the Hipparcos parallaxes. With these parallaxes, individual absolute magnitudes will be obtained with a high precision and the absolute magnitude calibrations will be improved. Finally the proper motions will contribute to increase the number and the quality of kinematical data, thus of the dynamics of our Galaxy.

We will work on these subjects as we are involved into different research proposals dealing with the scientific use of the preliminary and final Hipparcos data.

We hope that these pages, bridging a gap between the past (the Catalogue d'Entrée) and the future (the Hipparcos results), have attested the success of the Hipparcos mission and the enthusiasm to have taken part in it.

Version 1.1 of March 3, 1993 – Written with LaTeX
Version 1.2 of Oct. 10, 2005: included as an appendix of the french version.