DATA PROCESSING: TESTING OF CORE TASKS

F. Figueras[1], **B. López Martí**[1], **C. Fabricius**[2], **J. Torra**[1], **C. Jordi**[1], **P. Llimona**[1], **E. Masana**[1], **X. Luri**[1]

[1]University of Barcelona, 08028 Barcelona, Spain
[2]University of Copenhagen, 2100 Copenhagen, Denmark

ABSTRACT

The implementation and testing of the core processing of Gaia astrometric data has been the main goal of the GDAAS ESA contract (2000-2004). The reduction process, based on the principe of Global Iterative Solution (GIS), involves all the observations of 'well behaved' sources and, by solving a minimisation problem in a linearized system of equations, looks for the attitude reconstruction, the derivation of the geometric and photometric calibration, and astrometric and global parameters determination.

A first fully operational system has been used to process, in a distributed environment, eighteen months of mission data generated by the Gaia simulator (GASS). Results from several tests are presented, all of them devoted to check if the designed implementation ensures good convergence of the global iterations. We conclude that the system is converging, although at a very slow rate, mainly due to the strong correlation between parallax and light deflection.

Key words: Gaia; Data reduction; Global Iterative Solution.

1. INTRODUCTION

Core processing should provide the best representation possible of the Gaia observations in terms of the object model, satellite attitude and instrument calibration. This is a critical and very complex task, since stellar, global, instrument and attitude parameters must be derived simultaneously and at an accuracy enough to achieve the projected science goals. This issue is addressed by means of a 'global iterative solution' (GIS), which iteratively solves a linear system of equations over the source, global, attitude and calibration parameters. Since it allows the independent processing of many astronomical objects or many time intervals, this processing scheme is well suited for parallel computing; however, the proper sequencing and management of the processes is not trivial.

The study of the complexities of the Gaia data analysis and the feasibility of the proposed approach is the main goal of the GDAAS ESA contract (2000-2004) funded by ESA and carried out by the consortium formed by GMV, the University of Barcelona and the Center for Supercomputation of Catalonia (CESCA). Several tests have been performed using the first design of the instrumentation (Gaia-1). The purpose of this contribution is to review the status of the present GIS implementation and to report on the most recent test results.

2. THE GLOBAL ITERATIVE SOLUTION

The global iterative solution (GIS) aims at the attitude reconstruction, the derivation of the geometric and photometric calibrations, and the determination of the astrometric and global parameters. Starting with an initial setup provided by the telemetry data, it solves iteratively the system of linear equations until convergence is reached (Torra et al. 2005). Each GIS iteration is composed of four steps, each of them iterative itself, to update the attitude, calibration, source and global parameters, respectively. These sets of parameters are modelled as follows (Lindegren 2001a):

- **Attitude modelling:** The attitude is represented by means of cubic splines followed by normalization. The spline is written as a linear combination of non-negative local basis functions, called *B-splines*.
- **Geometric and photometric calibration modelling:** The geometric and photometric characteristics of the instruments are described by the two angular coordinates (along and across-scan directions) and a magnitude zero point for each CCD. Chromaticity was not considered in the present implementation.
- **Source astrometric modelling:** This process considers six astrometric parameters, viz.: α, δ, π, $\mu_\alpha cos(\delta)$, μ_δ and $\mu_r = v_R\pi/A$ (Perryman et al. 2000).
- **Global parameters modelling:** The present model only contains one element, the space curvature γ, the dimensionless parameter that parametrizes the gravitational deflection.

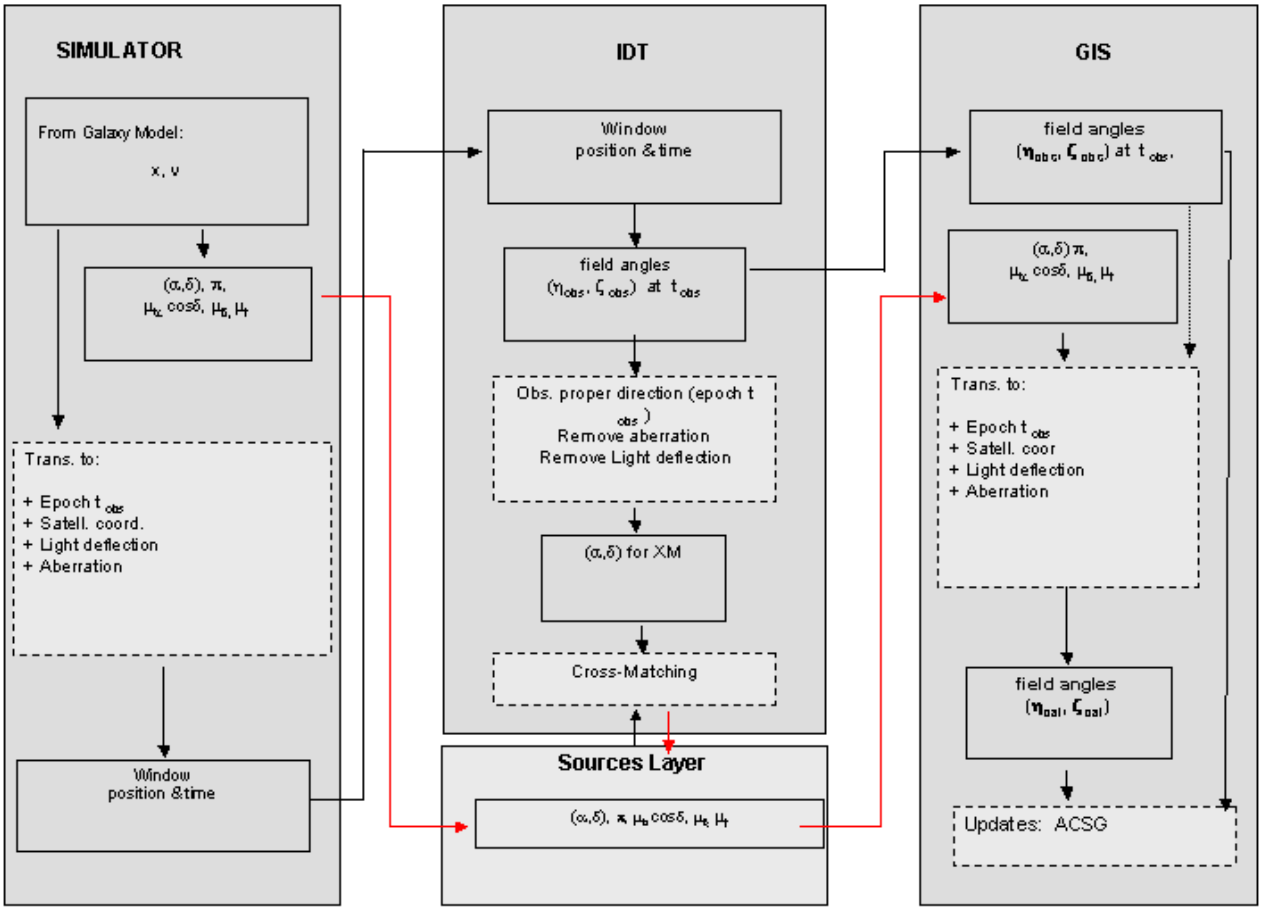

Figure 1. Scheme of the data involved in every step (simulation, initial data treatment and GIS processing) of the reduction process.

3. GIS IMPLEMENTATION

The telemetry data are provided by the Gaia System Simulator (GASS) and ingested into the data base. Some sources (a given percentage of the total sample) may have their nominal parameters (i.e., those from the simulation) entered 'directly' into the data base. These *primary sources* are used to drive the system, to accelerate the convergence and for testing purposes. After the ingestion, a cross-matching process links a given source with all its available observations. All these preliminary processes form the so-called *Initial Data Treatment (IDT)*.

Before starting the GIS computation, the parameters for the attitude and the geometrical and photometrical calibration of the CCDs, the model parameters, and the initial astrometric and photometric parameters for source objects shall be created and initialised in the data base. After this step, the system is ready to start the GIS processing.

A schematic representation of the whole procedure is presented in Figure 1.

We define a *GIS iteration 'N'* as a single run of the full set of algorithms composing GIS, namely source, attitude, calibration and global updating. A run of each of these individual algorithms, an iterative process too, is called *an update*. Hence, a GIS iteration is composed of four updates.

GIS runs in a distributed environment, with individual tasks running on different processing nodes under the control of a master node. As described in (Torra et al. 2005), the attitude and global updating algorithms require access to the all the observations for a given time interval. Therefore, the least-square process was adapted to allow a distributed processing, in which each processor takes care of a subset of observations according to a sky partitioning (Lindegren 2001b). On the other hand, the source updating algorithm is intrinsically a distributed process: For a given source, the least-square process runs in a given processor. Finally, for the calibration updating, a robust weighted mean is computed in the master node considering all the observations of a given calibration unit (a given CCD in a given time interval). These data have been previously pre-processed in a distributed manner and stored as intermediate data in the data base.

4. SECOND CAMPAIGN OF GIS TESTING

The goal of this GIS testing campaign, starting in November 2003, is to check the full GIS feasibility in both an ideal and a realistic case. The environment used for this campaign is the old Gaia instrument design and simulator (Gaia-1 and GASS-1).

A Compaq AlphaServer HPC320 computer available at the CESCA has been used in this testing campaign. The system has eight ES40 nodes (with four processors per node), 20 GB of main memory, 1128 GB of disk space and a peak performance of 53.31 Gflops/s, interconnected by a 100 MB/s Memory Channel II.

Up until September 2004, three tests had been performed:

- **Test 0:** A preliminary test to check the GIS implementation. Input and output data for all the algorithms composing the astrometric reduction process were tested at the μas accuracy level. Portability to other hardware and software platforms was also tested and succeeded. The result was a fully operational system ready to process GIS (Figueras et al. 2004a).
- **Test 1:** A test to check the GIS convergence in an ideal case, using six months of mission data. Only the astrometric parameters (α, δ) were updated during the least-squares process. Parallaxes and proper motions were not updated to avoid the singularity of the matrix due to the short mission time interval processed. The most important conclusions from this test were the following (Figueras et al. 2004b):
  - The weighting system was critical and needed revision. The Huber function used for the computation of the weight reduction factor was shown to decrease too slowly to reject outliers.
  - In addition to the (α, δ) coordinates, the parallax π and the proper motion μ should be updated to achieve a good solution. To that purpose, at least eighteen months of mission data should be processed.
- **Test 2:** A test to check the GIS convergence in a realistic case. This test and its results are presented in Section 5 and 6 (see Figueras et al. 2004c,d, for details).

5. TEST 2: DESCRIPTION

Test 2 was performed using 18 months of mission data for about 200 000 stars with V <13 mag generated using the Barcelona Galaxy Model (Torra et al. 1998). Of them, 10% were considered primary sources, with astrometric and photometric parameters from the simulator ingested 'directly' into the data base prior to GIS processing. Figures 2 and 3 show the distribution of sources and observations for this test.

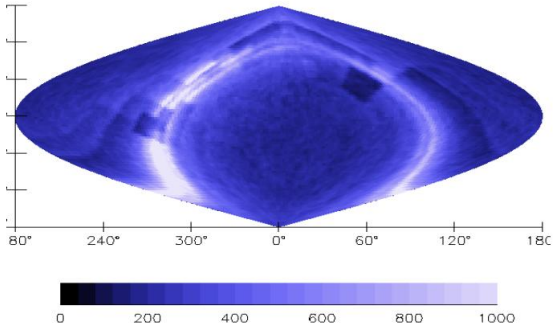

Figure 2. Source spatial distribution in Test 2 (in stars per square degree).

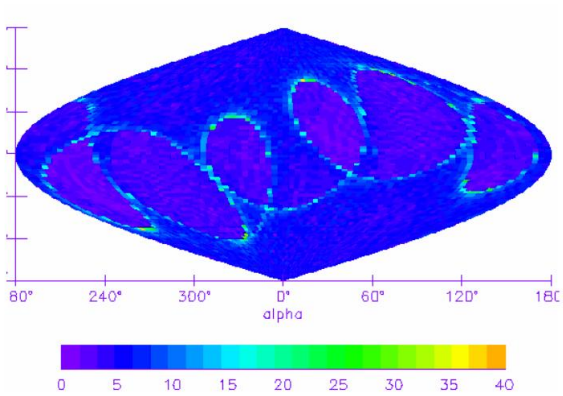

Figure 3. Average number of observations per source in Test 2.

During the eighteen months of mission about 11 $\times$ 10^6 transits have been observed. We recall that in the Gaia-1 instrument design each transit has 16 *elementary observations* corresponding to the passage of the source through the 16 CCDs of the astrometric focal plane. Every elementary observation is the result of reading out a single window from a single CCD.

The source astrometric modelling considers six astrometric parameters, viz.: α, δ, π, $\mu_\alpha \cos\delta$, μ_δ and $\mu_r = v_R\,\pi/A$. In the present case proper motion in the radial direction (μ_r) is not being updated to avoid the singularity of the matrix in the weighted least-squares equations. The update of this parameter is expected to be small, of the order of 0.01 μas yr^{-1}.

5.1. Observational Errors

Standard errors resulting from the IDT process include:

- The standard error in the observing time, which has a near-Gaussian distribution with a standard devia-

tion of about 63 μas. This includes the error resulting from the centroiding algorithm.

- The error in the across-scan position has been set to 1 mas following the simulations of the on-board detection process (Arenou et al. 2003)

The science telemetry data are generated considering the nominal scanning law, the nominal geometric and photometric calibration parameters, the nominal astrometric and photometric parameters of the sources, and light deflection effects with $\gamma = 1.000$.

During the initialization of the data base, a random scatter was added to the nominal values of the GIS parameters:

- A Gaussian random scatter was added to each of the (x, y, z) components of the pointing direction of the satellite (the nominal attitude quaternion data generated by GASS).
- A Gaussian scatter of $\sigma = 1$ mas was added to the along and across-scan geometric calibration parameters. No scatter in the zero point magnitude was considered.
- For the primary source parameters, errors of 1 mas in α (J2000) and δ (J2000) and of 1 mas yr^{-1} in μ_α, μ_δ and μ_r were introduced. This translates into a scatter of about 15 mas in the source position at epoch of observation (2010.0 to 2011.5). In addition, primary sources were ingested with null values of parallax ($\pi = 0$). In this way, we intended to test whether the system was able to derive absolute parallaxes, that is, to recover the true values of parallax for the sources.
- The global parameter γ was set to 1.1, implying a 10% offset from the nominal value (γ =1). This translates into an error of about 1 mas in the field angles.

5.2. Weighting System

Test-0 and Test-1 showed that the GIS weighting scheme is critical to avoid convergence problems. The weighting system includes factors contributing to the error in the field angles differences, as well as an appropriate function for the computation of the weight reduction factor.

The field angles differences have two contributions:

- $\sigma_{\rm mes}$: Standard deviation due to measurement errors (those included in the simulated data from GASS-1).
- $\sigma_{\rm param}$: Standard deviation due to the scatter in the attitude, calibration, global and source parameters used to compute a given set of calculated field angles.

The standard deviation in the field angles differences is calculated as:

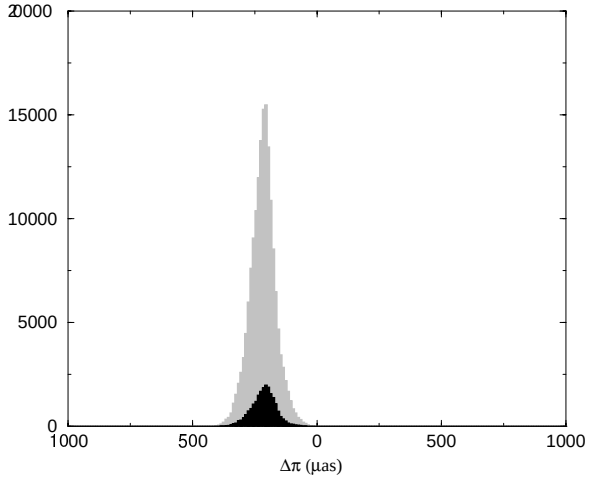

Figure 4. Differences between the updated and the simulated values for the parallax π after the third GIS iteration for primary sources and sources created during cross-matching (black and grey histograms, respectively). A clear shift from zero is seen.

$$\sigma_{\Delta f} = \sqrt{\sigma^2_{\rm mes} + \sigma^2_{\rm param}} \qquad (1)$$

The *wdecay* function, based on a 'blunder' detection method, has demonstrated to be more efficient than a *whuber* function, based on a Huber metric. This latter function was shown to decrease too slowly, whereas *wdecay* works better when some errors are very large (several tens of σ away): it is more effective giving negligible weight to such elementary observations and hence a better rejection of the outliers.

6. TEST 2: RESULTS

6.1. First GIS iterations: Correlation between Parallax and γ

After three GIS iterations, following the updating sequence *source-attitude-calibration-global* (SACG), we noticed that the system was converging very slowly and that systematic residuals were present in the updated parallax values (see Figure 4). The main reason was the strong correlation between the parallax π and gravitational deflection effects (Figueras et al. 2004d).

Let SZA be the spherical triangle with S at the Sun, Z at the spin axis and A at the source. We have $SZ = \xi$ and $ZA \simeq 90^\circ$. Let us introduce $\vartheta = SA$, $\nu = SZA$ and $\varphi = ZAS$. ν is the 'abscissa' of the star respect to the Sun, and the along-scan measurement is essentially this angle (see Figure 5).

It is well known that the parallax π causes a shift of the star by $\pi\, R\, \sin\theta$ along the great circle from the star towards the Sun, where R is the distance from the Sun to

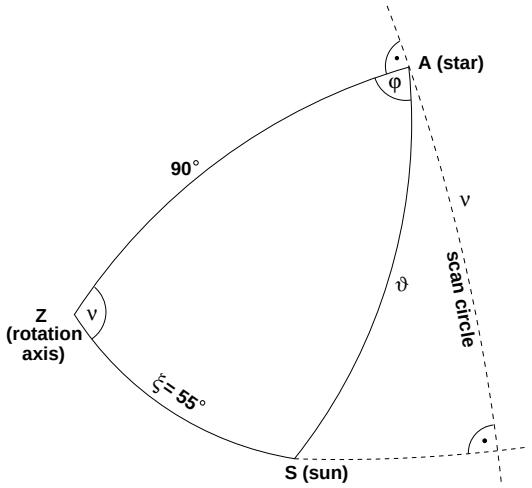

Figure 5. Definition of axes and angles (see Section 6.1).

observer in astronomical units (for Gaia, $R = 1.01$ on average). The projection of this shift on the scanning great circle (normal to Z) is

$$\Delta\nu_\pi = -\pi\, R\, \sin\theta\, \sin\varphi \tag{2}$$

Similarly, the gravitational deflection causes a shift of the star away from the Sun by the amount $(1 + \gamma)/2D\cot(\theta/2)$, where $D = 2\,G_S/c^2\,R$ is the deflection perpendicular to the Sun (G_S is the heliocentric gravitational constant). The projection of this shift on the scanning great circle is

$$\Delta\nu_\gamma = (1+\gamma)\,\frac{D}{2}\,\frac{\sin\theta\,\sin\varphi}{1-\cos\theta} \tag{3}$$

It can be demonstrated that both effects vary as $\sin\nu$. If there is an error $\Delta\gamma$ in the global parameter, then the corresponding error in parallax, $\Delta\pi$, can be evaluated by setting the total shift to zero. Applying a least-squares solution for $\Delta\pi$ and averaging over all observations, assuming that the observations are uniformly distributed in ν we have:

$$\Delta\pi = \Delta\gamma\,\frac{D}{R}\,\frac{1}{1+\cos\xi} \tag{4}$$

With $D/R = 2\,G_S/(c\,R^2) = 3992\ \mu$as and $\xi = 55°$, we can calculate the theoretical parallax shift for a given $\Delta\gamma$. The observed parallax shift was in complete agreement with the theoretical predictions, if the obtained values of the global parameter γ were used in the calculations (L. Lindegren, private communication).

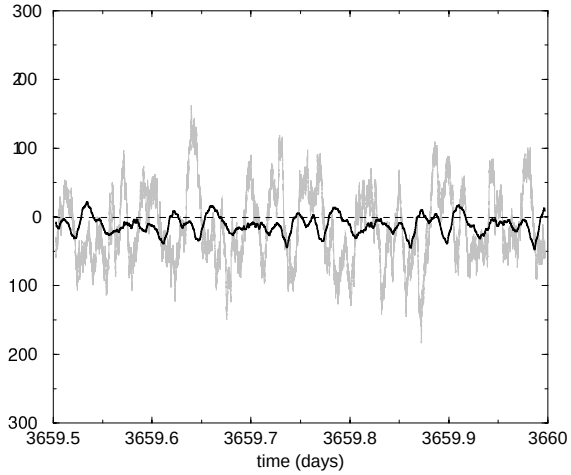

Figure 6. A running average of the differences (updated minus nominal) $< [\Delta\eta\,\cos\zeta]_x >$ after the first (gray) and thirteenth (black) GIS iterations for the 12 h interval between days 3659.5 and 3660.0 (from 2000.0).

6.2. GIS Iterations Using an Overcorrection Factor

To accelerate the convergence of the system, an overcorrection factor Q and a new updating sequence (SG+SGA) for the GIS iterations was introduced. This factor was applied to correct the obtained value of γ after each global iteration. The value of Q was initially set to a constant. After several subsequent iterations it was clear, though, that we should allow for this overcorrection factor to vary with each new update.

Results from first sequence of updates (SG+SGA) with one internal iteration per update indicated that the change in γ was bigger when attitude parameters were previously indicated after each source and global updating. In fact, the major part of the gravitational deflection effect is in the across-scan direction, and data along this direction are only taken into account in the attitude updating. The systematic trends in the attitude update are clearly seen in Figure 6. The residuals show a period of about 3 hours (one Gaia-1 spin period), highly correlated with light deflection effects.

6.3. Final Results

In total, up to thirteen GIS iterations were run, with varying sequences of the internal updating processes. The residuals of the astrometric and global parameters can be seen in Figures 7 to 10. It can be seen that both the global parameter γ (Figure 7) and the parallax π (Figure 8) have remarkably approached their nominal values.

The remaining astrometric parameters seem to have approached a convergence value after a few GIS iterations, which has not varied significantly in spite of the notable improvement in γ and π and the decrease in the scatter of the residuals (see Figure 9 and Figure 10). This is caused by the orientation and rotation errors of the system (see

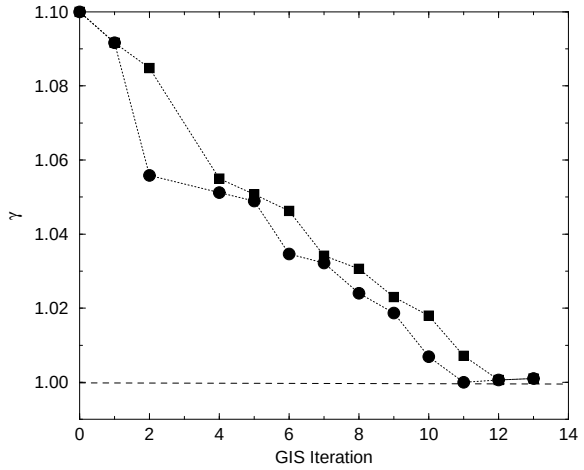

Figure 7. Value of the global parameter γ after each GIS iteration, with and without overcorrection (circle and square symbols, respectively).

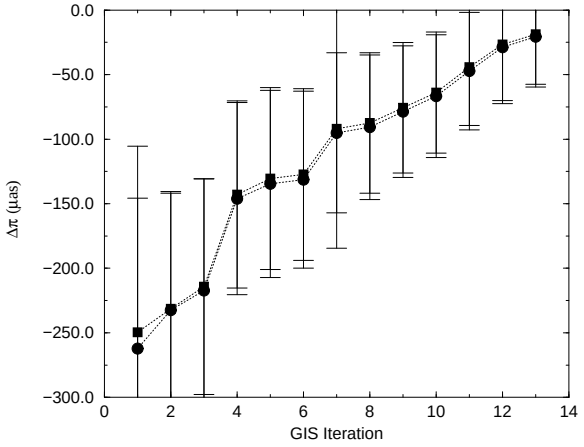

Figure 8. Mean difference between the updated and the nominal values for the parallax π after each GIS iteration for primary and non-primary sources (square and circle symbols, respectively).

Section 6.4). Despite these effects, the final position of the sources is on average only about 17 μas from their expected position in the epoch 2010.75 (mean epoch of observation).

6.4. Rotation of the System

In a general solution for the five astrometric parameters $(\alpha, \delta, \pi, \mu_\alpha, \mu_\delta)$ based on space observations of sources only, the system of positions (at the reference epoch) is in principle undetermined with respect to the exact orientation of the coordinate system, corresponding to three degrees of freedom (one small rotation around each of the principal axes: ε_x, ε_y, ε_z). We call this the *orientation error* of the system. Similarly, the proper motions are undetermined with respect to a small uniform rotation, with components $(\omega_x, \omega_y, \omega_z)$ about the principal axes, called the *rotation error* of the system (Lindegren, private communication).

Table 1. Solution for the system rotation after GIS Iteration 13.

	ε_x	ε_y	ε_z	ω_x	ω_y	ω_z
Value (μas)	0.11	−1.83	−6.28	0.45	3.39	17.85
σ (μas)	0.03	0.04	0.04	0.06	0.07	0.07

When simulating the astrometric solution, the orientation and rotation errors in the resulting system show up as a systematic pattern in the position and proper motion differences between the computed (solved) and simulated ('true') parameters. The expected pattern in these differences is:

$$\begin{pmatrix} -\sin\delta\cos\alpha & -\sin\delta\sin\alpha & \cos\delta \\ \sin\alpha & -\cos\alpha & 0 \end{pmatrix} \varepsilon = \begin{pmatrix} \Delta\alpha \\ \Delta\delta \end{pmatrix} \quad (5)$$

$$\begin{pmatrix} -\sin\delta\cos\alpha & -\sin\delta\sin\alpha & \cos\delta \\ \sin\alpha & -\cos\alpha & 0 \end{pmatrix} \omega = \begin{pmatrix} \Delta\mu_\alpha \\ \Delta\mu_\delta \end{pmatrix} \quad (6)$$

where $\Delta\alpha$, $\Delta\delta$, $\Delta\mu_\alpha$, $\Delta\mu_\delta$ denote the parameter differences (computed minus nominal). Note that $\Delta\alpha$ and $\Delta\mu_\alpha$ include the $\cos\delta$ factor.

Since (approximately) α and δ are known for each star, one can obtain ε and ω from a least-squares solution of the above equations with downweighting of large residuals using the *wdecay* function. The RMS of all positional and proper motion residuals are also estimated using a robust method.

We analysed the rotation of the system after the 11th and 13th GIS iterations. Table 1 show the results for the 13th GIS iteration. The solution, computed using coordinates at the epoch 2010.75, showed that the orientation and rotation errors are very small in the x and y directions, which translates into a very small correction in δ and μ_δ, respectively. The orientation and rotation errors in the z direction are of the same order as the values of the residuals in α and μ_α, respectively (see Figure 9 and Figure 10).

7. SCALING TO THE REAL MISSION

The process of Test 2, dealing with 18 months of mission and about 200 000 sources, used 16 processors at CESCA, and consumed about 1008 hours of elapsed time and 1310 hours of CPU time. More than 80% of the elapsed time was devoted to data retrieval and storage. Raw, intermediate and final data generated a data base with 164.5 GB of data. These figures can be scaled to the real mission as shown in Table 2. Scaling factors for the spin period and number of CCD strips correspond to the

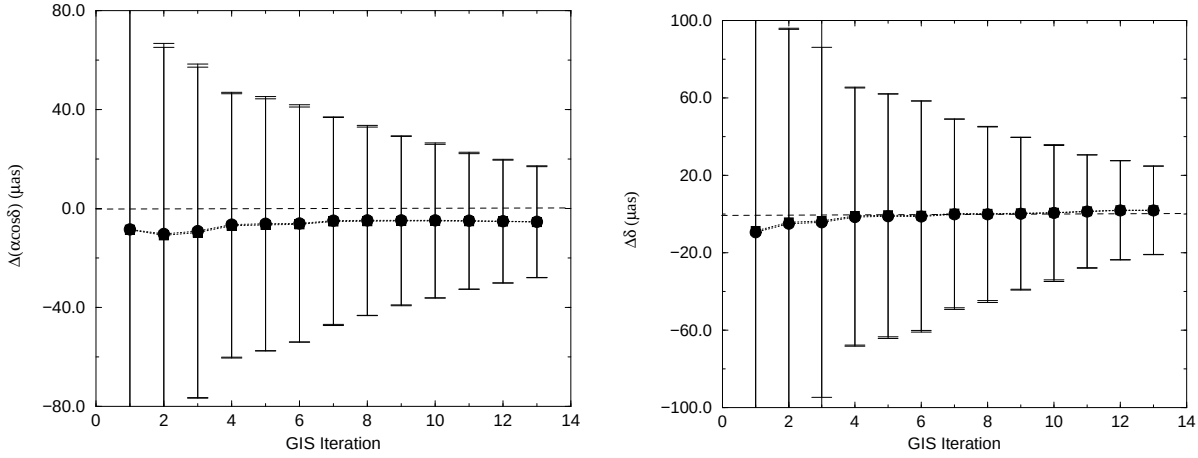

Figure 9. Mean difference between the updated and the nominal values for $\alpha \cos\delta$ (left panel) and δ (right panel) after each GIS iteration for primary and non-primary sources (square and circle symbols, respectively). Epoch: 2010.75.

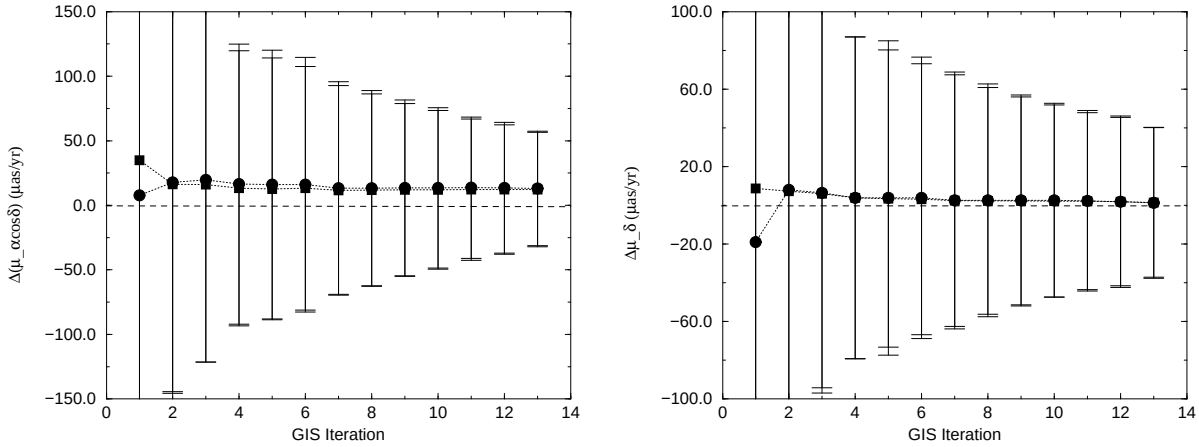

Figure 10. Mean difference between the updated and the nominal values for $\mu_\alpha \cos\delta$ (left panel) and μ_δ (right panel) after each GIS iteration for primary and non-primary sources (square and circle symbols, respectively).

Table 2. Summary of scaling factors

	Test	Mission	Scaling Factor
Observing Period	18 months	5 years	3.33
Observed Objects	200 000	10^9	5000
GIS Objects	200 000	10^8	500
Spin period[a]	3^h	6^h	0.5
Number of CCD Strips	16	10	0.625

[a] This parameter determines the number of transits.

change from the old to the current Gaia instrument design (Gaia-2).

According to these factors, the telemetry ingestion, data base initialisation, cross-matching, and GIS processing of the full mission will require about 4×10^{18} FLOPs, and a data base of about 1–2 Pb. These figures do not include raw, intermediate and final reduction data resulting from the Spectro instrument. On the other hand, the FLOPs consumption provided here has to be considered as a rough estimation of the needs of the actual astrometric reduction process. For one thing, the time consumption from the Spectro instrument should be added. Moreover, to correctly size the hardware/software system to be used for the nominal mission, other factors can be even more important than the hardware FLOP capacity. In particular, the data base input/output capacity seems to be the main limiting factor in our present tests, highly depending on the disk system throughtput and network speed/latency.

8. CONCLUSIONS AND FOLLOW-UP

We have reported on the recent GIS testing campaign performed by the GDAAS consortium during the period from November 2003 to September 2004. The results show that the system converges, although at an insufficient rate. This slow convergence is partially due to the strong correlation between the parallax π, light deflection effects (and hence, the global parameter γ), and attitude

parameters. It has been demonstrated that the final residual values for the rest of the position and proper motion parameters are conditioned by the orientation and rotation of the final system, and can be corrected.

Several follow-up tests have been designed and are being run to better understand the origin of the convergence problems observed during Test-2. The goal is to test whether these problems could be due to initial GIS parameter values too far from the correct solution, so the system was having trouble in correcting the 'memory' of this initial big offset. This third testing campaign should be finished by the end of the running year (Figueras et al. 2004e)

After these tests, a system considering the new Gaia instrument design (Gaia-2) and a set of more complex algorithms (Lindegren 2003), much more representative of the final mission, will be ready to carry on more detailed investigations on the scientific basis. This system has been implemented during the second phase of the Gaia Data Access and Analysis Study contract (GDAAS-2).

The GDAAS-2 GIS testing campaign will provide complete confidence in the overall Gaia data core processing approach. From that, a realistic scalability to a final data processing system will be possible.

ACKNOWLEDGMENTS

This work has been developed during the execution of Phase II of the Gaia Data Access and Analysis Study ESA contract (GDAAS-2, Reference 16439/02/NL/VD), the Spanish Software company GMV acting as main contractor, the University of Barcelona and the Copenhagen University as scientific personnel and CESCA for the hardware equipment and support personnel. We distinctly thank Dr. L. Lindegren and Dr. U. Bastian for their valuable scientific support and useful discussions during the scientific evaluation of the results. This study has been partially supported by the Spanish Ministry for Science and Technology under contract PNE 2003-04352.

REFERENCES

Arenou, F., et al., 2003, *Simulation of the on-board detection*, Gaia technical report OBD-FAJCL-01

Figueras, F., Masana, E., Torra, J. et al., 2004a, *Second campaign of GIS testing: Test-0 report*, Gaia technical note UB-GDAAS2-TN-002

Figueras, F., Jordi, C., Torra, J. et al., 2004b, *Second campaign of GIS testing: Test-1 report, 100 % of primary sources*, Gaia technical note UB-GDAAS2-TN-013

Figueras, F., López Martí, B., Fabricius, C. et al., 2004c, *Second campaign of GIS testing: Test-2 report (Part A)*, Gaia technical note UB-GDAAS2-TN-019

Figueras, F., López Martí, B., Fabricius, C. et al., 2004d, *Second campaign of GIS testing: Test-2 report (Part B)*, Gaia technical note UB-GDAAS2-TN-020

Figueras, F., López Martí, B., Fabricius, C. et al., 2004e, *Third campaign of GIS testing: Test-3 report* , Gaia technical note UB-GDAAS2-TN-022 (in preparation)

Lindegren, L., 2001a, *Proposed prototype processes for the Gaia Global Iterative Solution* Gaia technical report GAIA-LL-34

Lindegren, L., 2001b, *Global Iterative Solution. Distributed processing of the attitude updating* Gaia technical report GAIA-LL-37

Lindegren, L., 2003, *Algorithms for GDAAS Phase II: Definition* Gaia technical report GAIA-LL-44

Perryman, M.A.C., et al., 2000, *GAIA: Composition, Formation and Evolution of the Galaxy*, ESA SCI(2000)4

Torra, J., Chen, B., Figueras, F., et al., 1998, Baltic Astronomy 8, 171

Torra, J., Figueras, F., Jordi, C., et al., 2005, ESA SP-576, this volume