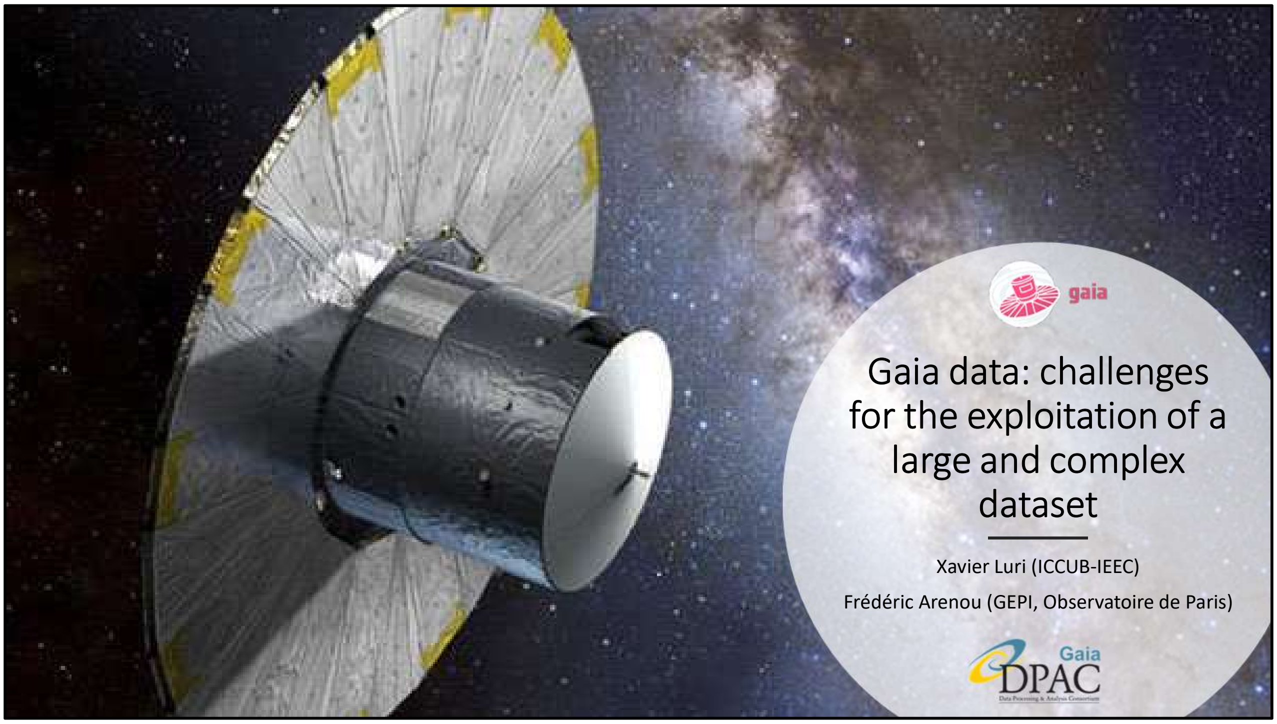

gaia
Gaia data: challenges for the exploitation of a large and complex dataset
Xavier Luri (ICCUB-IEEC)
Frédéric Arenou (GEPI, Observatoire de Paris)
Gaia
DPAC

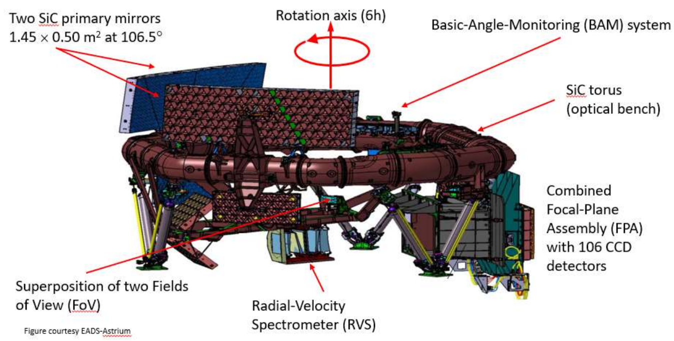
Abstract

In recent years it has become very common to hear statements on how Big Data, the availability of very large datasets, is revolutionizing science. It is applicable to a wide variety of area, but it is often forgotten that the breakthroughs achieved with these data do not only come from its volume, but specially from the capability to do a meaningful data analysis with them. This capability requires the large processing capability of computers but also, and more critically, a proper understanding of the statistical properties of these samples and the ability to design statistical analysis tools to extract knowledge from the data. A clear example of this is the datasets produced by the Gaia mission of the European Space Agency. It is generating very large astrometric catalogues (two billion objects) with unprecedented accuracy, and in this talk I will discuss the challenges faced by the astronomical community to fully exploit its scientific potential. These challenges range from the basic need to understand the properties of the data (data censorships, variable transformation, random errors, systematics) to the design and implementation of analysis tools appropriate to handle them.

The Gaia mission

Gaia – JSM 2020

The spacecraft



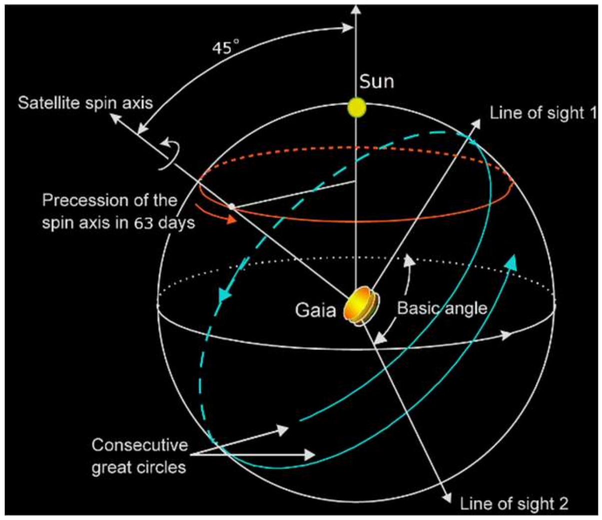
Images courtesy EADS-Astrium

Located at the Earth-Sun L2 Lagrange point.
Completed 5 years of mission, expected 5-6 years of extended mission.

Scanning law

The spacecraft is slowly rotating at a constant **angular rate of 1° per minute around an axis** perpendicular to those two fields of view

Objects transit in the second field of view 106.5 minutes after crossing the first one.



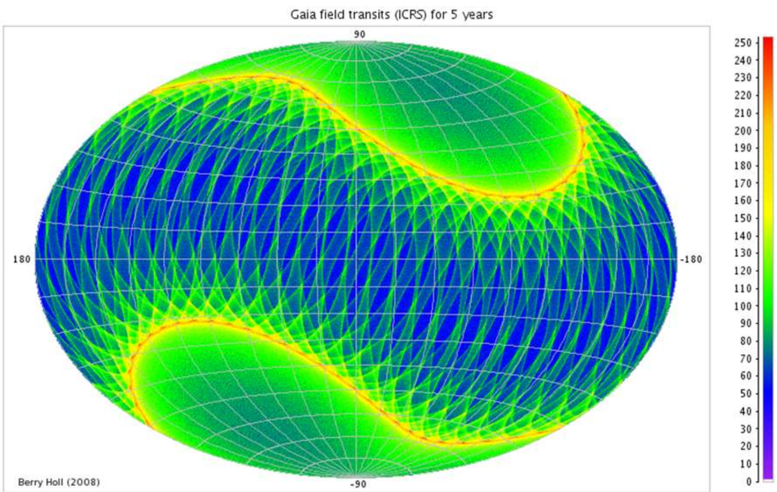
This scan axis describes a **slow precession motion around the Sun-to-Earth direction, with an average period of 63 days.**

Scanning Space Astrometry: to transform positional information into timing data

Gaia builds on the global astrometry concept successfully demonstrated by the Hipparcos mission. This measurement principle relies on the systematic and repeating observation of the star positions in two fields of view. For this purpose, the spacecraft is slowly rotating at a constant angular rate of 1° per minute around an axis perpendicular to those two fields of view, which thus describe a circle in the sky in 6 hours. With a basic angle of 106.5° separating the astrometric fields of view, objects transit in the second field of view 106.5 minutes after crossing the first one.

The spacecraft rotation axis makes an angle of 45° with the Sun direction. This represents the optimal point between astrometry requirements - that call for a large angle - and implementation constraints - such as payload shading and solar array efficiency. This scan axis further describes a slow precession motion around the Sun-to-Earth direction, with an average period of 63 days. This allows the scanning law definition to be independent from the orbital position around L2.

Number of Gaia sky transits (5 years)



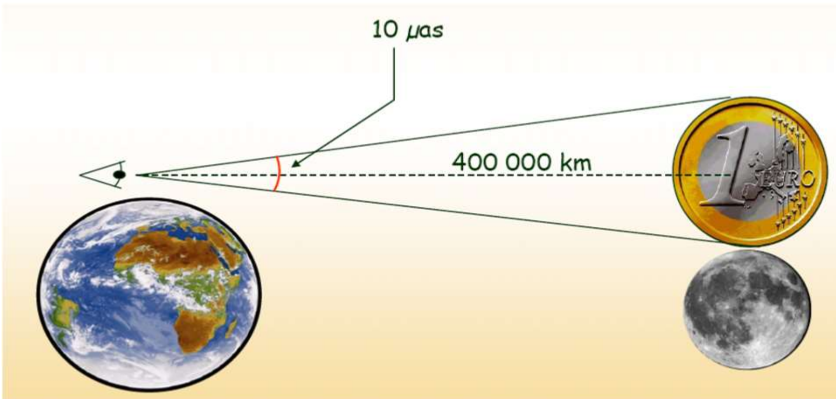
This peculiar coverage of the sky has important consequences in the properties of the catalogue (completeness, precision) that can make its analysis a complex task

Equatorial coordinates, RA running from -180° to +180° right-to-left.

Gaia – JSM 2020

These pictures show the expected number of field-of-view transits experienced by sources at different celestial positions due to the Gaia nominal scanning law. In the top six snapshots the location of the Sun is indicated by the yellow circle and the spin axis by the open circle. The projection uses equatorial coordinates, with right ascension running from -180° to +180° right-to-left. The blue line is the ecliptic (the plane in which Gaia orbits around the Sun together with the Earth). The average number is 88 field-of-view transits, although normally an average value of 72 transits is quoted (accounting for dead time). An over-abundance of transits occurs at 45° from the ecliptic due to the difference between the 45° spin axis angle with respect to the Sun and the 90° angle between spin axis and the fields of view.

Gaia and the astrometric accuracy



Note that this is in fact a statistic statement. What does it really mean?

The accuracy of the Gaia measurements can not really be summarized in a single number

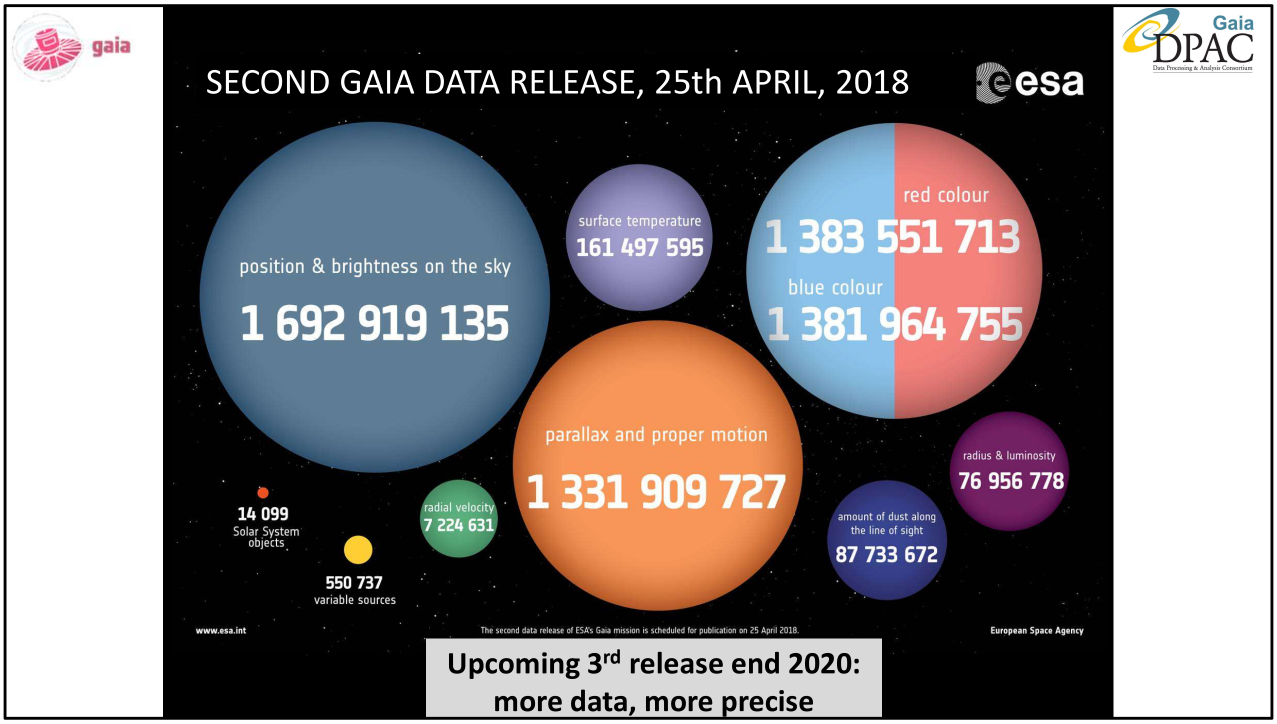
The catalogue(s)

MISSION STATUS NUMBERS

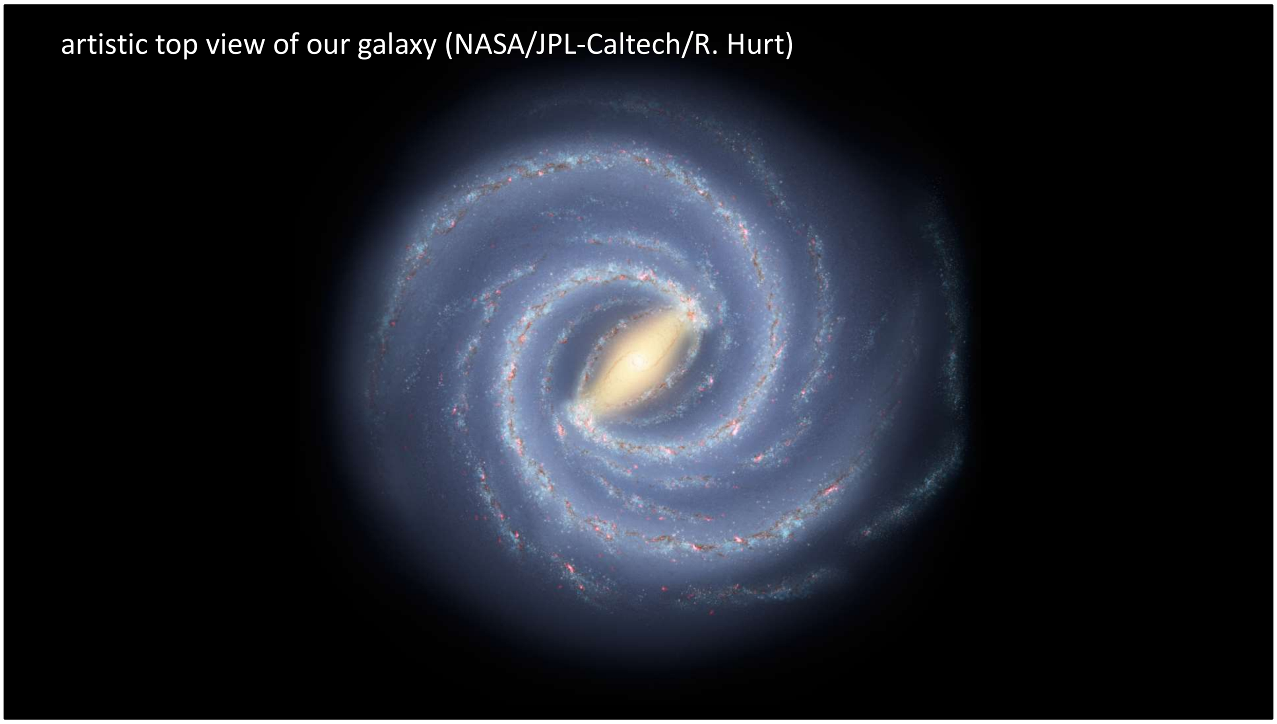
https://www.cosmos.esa.int/web/gaia/mission-numbers

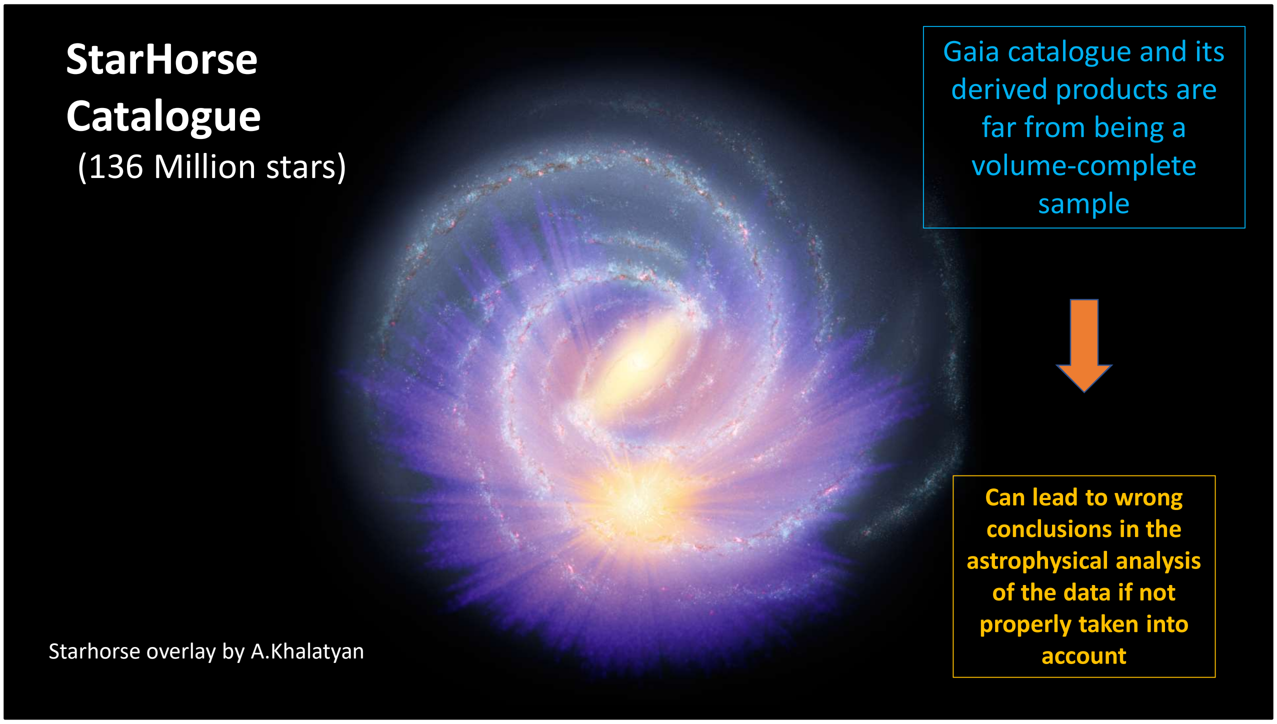
CURRENT DATE AND TIME	2020-05-13T07:30:16 (TCB)
MISSION STATUS	
Satellite distance from Earth (in km)	1,614,802
Number of days having passed since 25 July 2014	2119
Number of days in mission extension	302
OPERATIONS DATA (collected since 2014/07/25)	
Volume of science data collected (in GB)	79,129
Number of object transits through the focal plane	150,329,034,331
Number of astrometric CCD measurements	1,481,814,766,972
Number of photometric CCD measurements	299,212,364,190
Number of spectroscopic CCD measurements	29,185,344,753
Number of object transits through the RVS instrument	9,776,081,754

https://www.cosmos.esa.int/web/gaia/mission-numbers

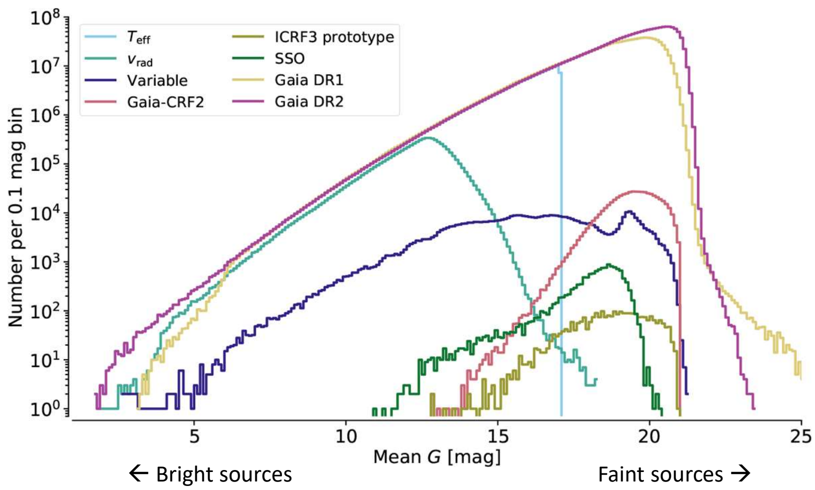

gaia
Gaia DPAC
SECOND GAIA DATA RELEASE, 25th APRIL, 2018
esa
position & brightness on the sky
1 692 919 135
surface temperature
161 497 595
red colour
1 383 551 713
blue colour
1 381 964 755
parallax and proper motion
1 331 909 727
radius & luminosity
76 956 778
amount of dust along the line of sight
87 733 672
radial velocity
7 224 631
14 099
Solar System objects
550 737
variable sources
www.esa.int
The second data release of ESA's Gaia mission is scheduled for publication on 25 April 2018.
European Space Agency
Upcoming 3rd release end 2020:
more data, more precise

The selection function


artistic top view of our galaxy (NASA/JPL-Caltech/R. Hurt)


StarHorse Catalogue
(136 Million stars)
Gaia catalogue and its derived products are far from being a volume-complete sample
Can lead to wrong conclusions in the astrophysical analysis of the data if not properly taken into account
Starhorse overlay by A.Khalatyan

Number of sources vs brightness

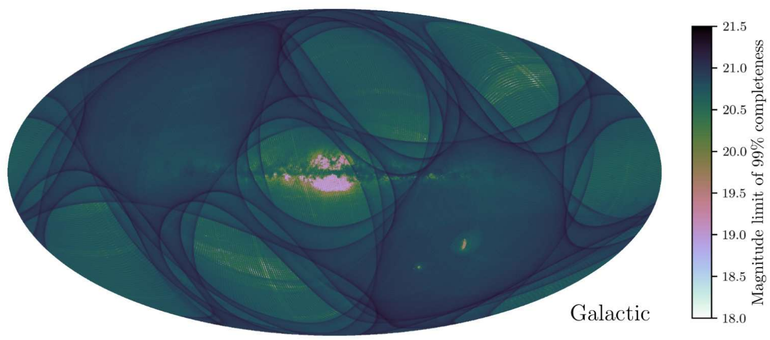


https://www.aanda.org/articles/aa/full_html/2018/08/aa33051-18/F1.html

Even the main catalogue is, at best, (globally) complete to G~20.

But the several sub-products have very (wildly) different completeness

→
Leads to non-representative samples, and therefore biased results (e.g. Malmquist bias), if not properly taken into account



Boubert and Everall (submitted, 2020b)

Modelling completeness

Reality is even more complex, with many regional effects

Even if global properties are taken into account, use of regional samples can still lead to biased results

The data: statistical properties

Data reduction (DPAC)

The DPAC consortium is making a huge effort in producing a data set where:

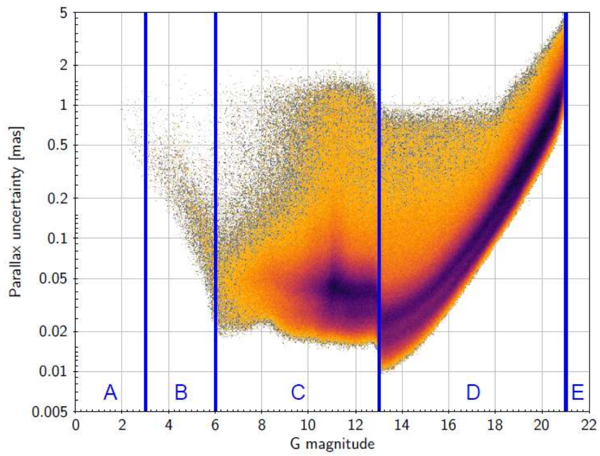
- The formal uncertainties for each piece of data are as realistic as possible
- Systematics in the data are reduced and checked against independent data as far as possible
- The datasets are as complete as possible and their selection function is well defined

However, the scientific exploitation of the Gaia data pushes it to its limits, so **every effect in it, no matter how minor, becomes significant**

Gaia DPAC

Formal uncertainty in parallax

gaia



Regimes of G:

- A: Too bright
- B: Partly saturated (unreliable)
- C: Detector and calibration limited
- D: Photon limited
- E: Too faint (not published)

Formal uncertainties in *Gaia* DR2 were estimated from the internal consistency of measurements and do not represent the total error

Lindegren et al., 2018 Aug 27 Gaia DR2 astrometry, slide 4 of 54

Gaia DPAC Data Processing & Analysis Consortium

Different error regimes, requires individualized treatment of objects rather than a global recipes.

$$\varpi_i^{\mathrm{DR2}} - \varpi_i^{\mathrm{true}} = r_i + s(\alpha, \delta, G, C, \dots)$$

Random error   Systematic error

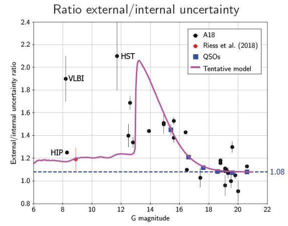
$$\sigma_{\mathrm{ext}} = \sqrt{k^2\sigma_i^2 + \sigma_s^2}$$

To estimate this external ("real") error we need external data to compare, but none matches by far the quality and amount of Gaia data!

- For *Gaia* DR1 (TGAS) the published uncertainties correspond to σ_{ext}
- For *Gaia* DR2 the published uncertainties correspond to σ_i

A tentative external "calibration"

Ratio external/internal uncertainty



$\sigma_{\mathrm{ext}} = \sqrt{k^2\sigma_i^2 + \sigma_s^2}$

Faint ($G \gtrsim 13$):
$k = 1.08$
$\sigma_s = 0.043$ mas

Bright ($G \lesssim 13$):
$k = 1.08$
$\sigma_s = 0.021$ mas

The model may be too pessimistic for $G \simeq 13$ to 15

Lindegren et al., 2018 Aug 27   Gaia DR2 astrometry, slide 16 of 54

Parallax zero-point

The zero point ϖ_0 is the expected measured parallax for a source at infinity; it should thus be *subtracted* from the catalogue value.

As a global average, $\varpi_0 \equiv \langle s \rangle \simeq -0.03$ mas, but:

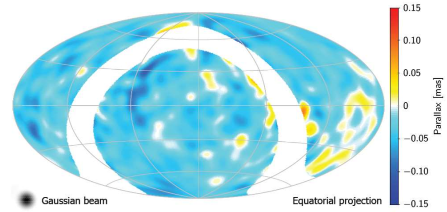
- s definitely depends on (α, δ)
- s probably depends of G
- s may depend of $C = G_{BP} - G_{RP}$
- the dependence is probably multivariate, $s(\alpha, \delta, G, C, \dots)$

No general recipe can be given
for the correction of the zero point

Lindegren et al., 2018 Aug 27 Gaia DR2 astrometry, slide 8 of 54

Systematics $s(\alpha, \delta)$ on large scales

QSO parallaxes smoothed by a Gaussian beam ($\sigma = 3.7°$)
(only $|\sin b| > 0.2$ shown)

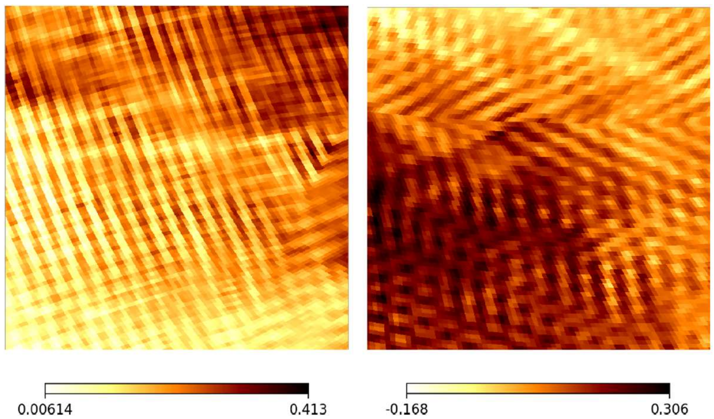


Mean value $= -0.030$ mas, RMS of smoothed values $= 0.020$ mas

Lindegren et al., 2018 Aug 27 Gaia DR2 astrometry, slide 9 of 54

A critical issue for the determination of distances of astronomical objects. Difficult to assess but under control. However, needs to be taken into account when exploiting the Gaia data (e.g. include it in a Bayesian model fit)

Correlations



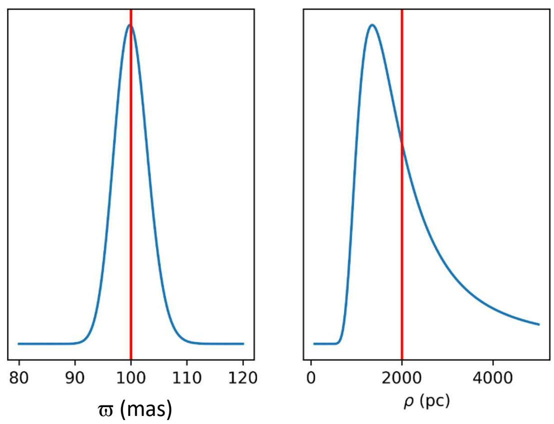
The five (six) astrometric parameters are correlated. Correlation matrices are provided in the catalogue.

These correlations have been strongly reduced from DR1 to DR2, but still need to be taken into account for astrophysical applications. Otherwise it leads to biased results.

Correlation $\rho(\varpi, \mu_\delta)$ towards the bulge (*left panel*) and $\rho(\alpha, \delta)$ towards the Large Magellanic Cloud

https://www.aanda.org/articles/aa/full_html/2018/08/aa33234-18/aa33234-18.html

Intrinsic properties of the data



PDF of parallaxes vs. PDF of distances

https://www.aanda.org/articles/aa/full_html/2018/08/aa32964-18/aa32964-18.html

Gaia directly measures star parallaxes. However, the astrophysically relevant parameters is the distance. For the true parallax, the relation is:

$$r = \frac{1}{\varpi}$$

The PDF of parallaxes is quasi-normal, well behaved, but leads to a PDF of the estimated distances with undesired properties (biased mean, skewness). Use of stellar parallaxes to estimate distances is a difficult task! Requires very accurate statistical handling.

gaia

Plus issues found after publication…

Gaia DPAC
Data Processing & Analysis Consortium

esa

e 2 » Gaia DR2 known issues

KNOWN ISSUES WITH THE GAIA DR2 DATA

This page lists the issues in our second data release that have been discovered after the release of the Gaia data and related documentation. The Gaia DR2 contents page contains a summary of limitations that were known, and documented, already at the release date. Tips on how to better make use of the Gaia Archive can be found here.

Overview:

- Astrometry: 2- versus 5-parameter solutions
- Astrometry: Considerations for the use of DR2 astrometry
- Astrometry: Systematic effects in Gaia DR2 parallaxes for very bright stars
- Crossmatch: Hipparcos2
- Radial Velocities: Potential contamination in crowded regions
- Photometry: Systematic effects and response curves

Issues discovered after the release of the Gaia Data

Astrometry: Considerations for the use of DR2 astrometry (Lindegren, Vienna 2018)
https://iopscience.iop.org/article/10.3847/2515-5172/ab2632

Systematic effects in Gaia DR2 parallaxes from very bright stars (i.e. G < 5) stars
(they may have additional systematic errors due to calibration issues), June 2019

Data exploitation: methodology

Conclusions

The users of Gaia data need to :

- Abandon any naïve or over-simplistic use of the Gaia data.
- Large datasets come with the benefit of an abundance of data, and with the curse of complex properties that can affect its exploitation
- Sound statistical treatment, taking into account all the properties of the catalogue and the information provided, is needed for a correct scientific exploitation
- The many relevant effects are documented in the Gaia papers and the archive documentation, but given its complexity they should be preferably re-evaluated in each case
- Methodologies in statistics for a correct exploitation are available (Bayesian modelling, Monte-Carlo techniques, time-series analysis, etc.); we need to properly understand them and to correctly apply them. Beware of the "not invented here" and "re-invent the wheel" syndromes.
- The use of simulations implementing the effects in the data is a good tool to understand the consequences of these effects in the data analysis. Several Gaia mock-ups are available.

gaia

Gaia DPAC

Data Processing & Analysis Consortium

Gaia Data Release 2. Using Gaia parallaxes

Show affiliations

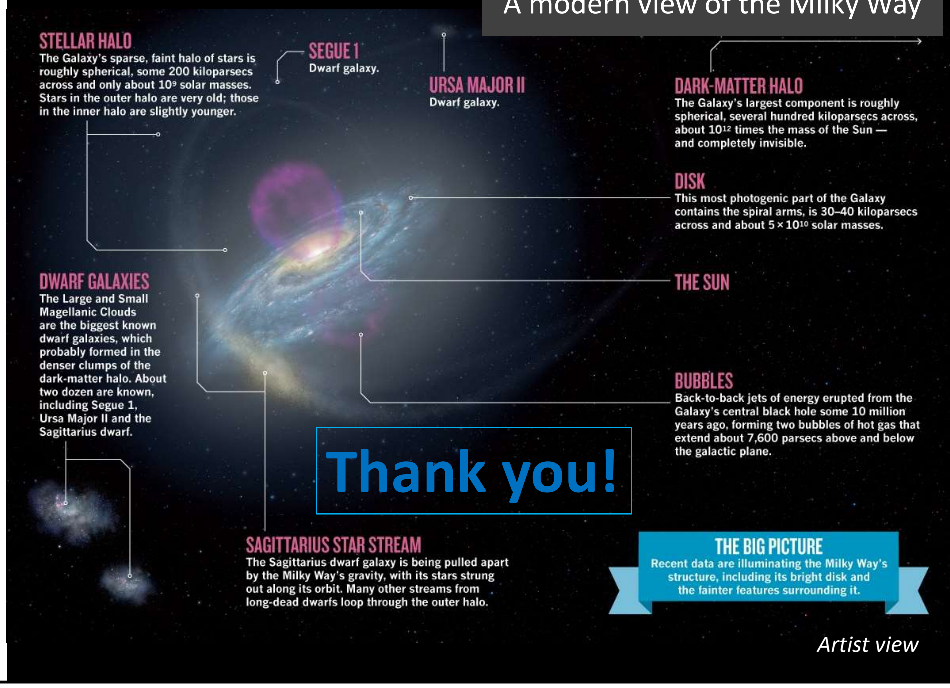
Luri, X.; Brown, A. G. A.; Sarro, L. M.; Arenou, F.; Bailer-Jones, C. A. L.; Castro-Ginard, A.; de Bruijne, J.; Prusti, T.; Babusiaux, C.; Delgado, H. E.

Context. The second Gaia data release (Gaia DR2) provides precise five-parameter astrometric data (positions, proper motions, and parallaxes) for an unprecedented number of sources (more than 1.3 billion, mostly stars). This new wealth of data will enable the undertaking of statistical analysis of many astrophysical problems that were previously infeasible for lack of reliable astrometry, and in particular because of the lack of parallaxes. However, the use of this wealth of astrometric data comes with a specific challenge: how can the astrophysical parameters of interest be properly inferred from these data?
Aims: The main focus of this paper, but not the only focus, is the issue of the estimation of distances from parallaxes, possibly combined with other information. We start with a critical review of the methods traditionally used to obtain distances from parallaxes and their shortcomings. Then we provide guidelines on how to use parallaxes more efficiently to estimate distances by using Bayesian methods. In particular we also show that negative parallaxes, or parallaxes with relatively large uncertainties still contain valuable information. Finally, we provide examples that show more generally how to use astrometric data for parameter estimation, including the combination of proper motions and parallaxes and the handling of covariances in the uncertainties.
Methods: The paper contains examples based on simulated Gaia data to illustrate the problems and the solutions proposed. Furthermore, the developments and methods proposed in the paper are linked to a set of tutorials included in the Gaia archive documentation that provide practical examples and a good starting point for the application of the recommendations to actual problems.

Example: a summary of recommendations for the usage of the Gaia parallaxes

Luri et al.
A&A volume 616 (2018)

A modern view of the Milky Way



Thank you!

Artist view