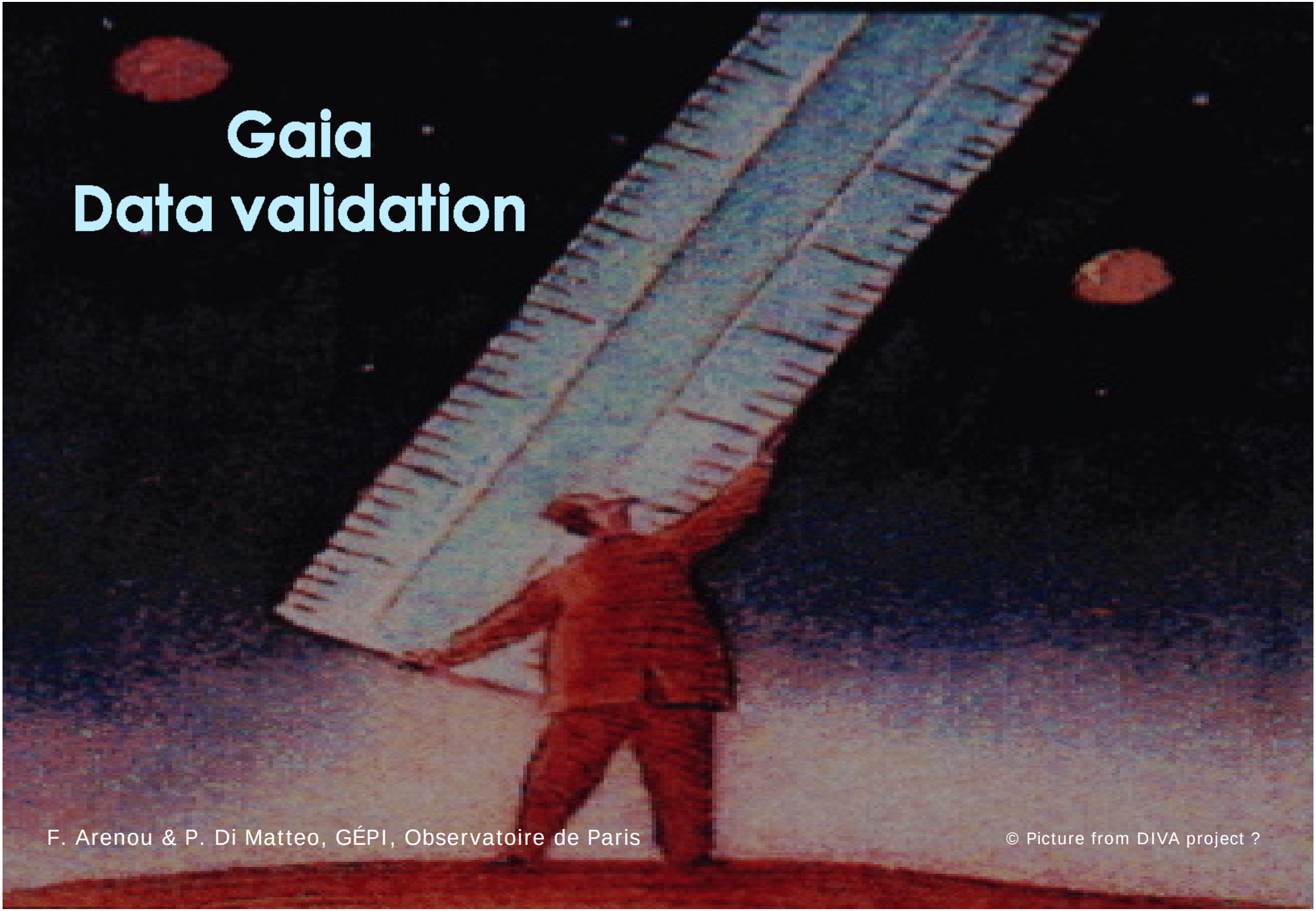
Gaia
Data validation



F. Arenou & P. Di Matteo, GÉPI, Observatoire de Paris

Why a validation?

Background

- Gaia is a very complex mission
  - The satellite is a complex engine measuring a complex sky!
  - Obtaining the billions of parameters is a complex process
    - There are many ways to get systematic errors!
- DPAC is responsible of the quality of the Catalogue
  - 400+ scientists/engineers... hundreds of person-years
  - The Gaia Catalogue should not be a quick and dirty work
  - Pressure from outside should not impose the agenda
    - Some form of validation before publication is needed!
- Experience from Hipparcos
  - Users easily misinterpret the (statistical by nature) data
  - Some effort was put in data validation (1PhD, 2 papers, 3 chapters)

The Hipparcos and Tycho Catalogues

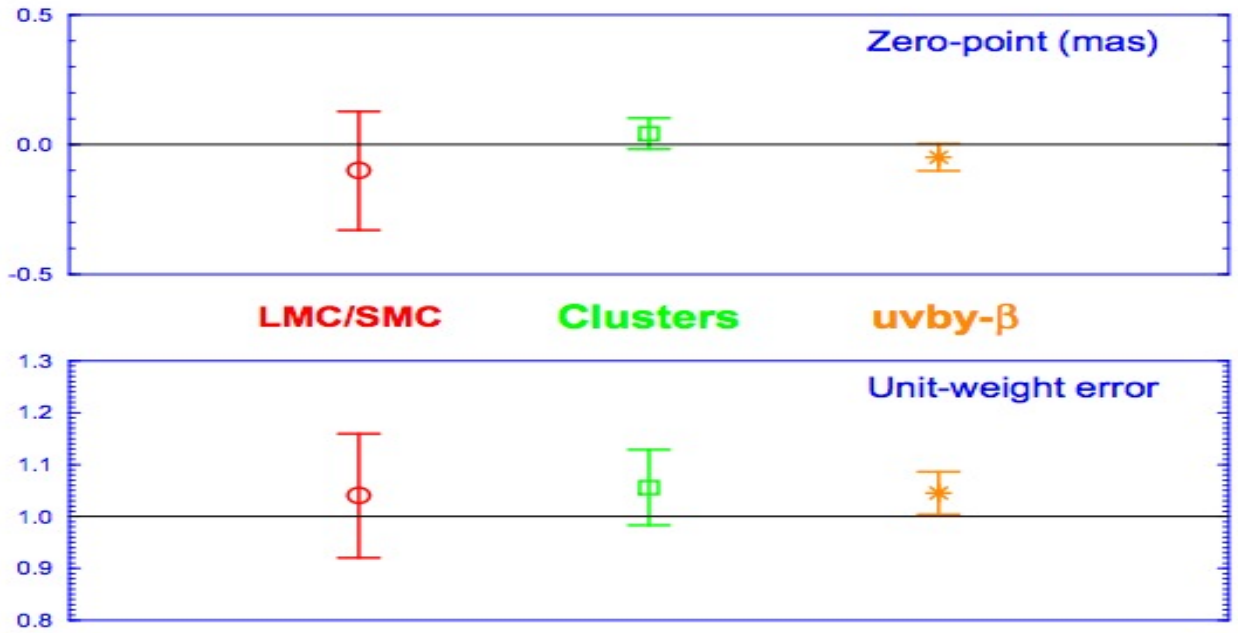


Figure 20.5. *Zero-point and unit-weight of Hipparcos parallaxes, from external comparisons using distant stars.*

THE ASTRONOMICAL JOURNAL, 129:1616–1624, 2005 March

CONFIRMATION OF ERRORS IN *HIPPARCOS* PARALLAXES FROM *HUBBLE SPACE TELESCOPE* FINE GUIDANCE SENSOR ASTROMETRY OF THE PLEIADES[1]

DAVID R. SODERBLOM AND ED NELAN
Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218; soderblom@stsci.edu, nelan@stsci.edu

G. FRITZ BENEDICT, BARBARA MCARTHUR, IVAN RAMIREZ, AND WILLIAM SPIESMAN
McDonald Observatory, University of Texas, Austin, TX 78712; fritz@astro.as.utexas.edu,
mca@astro.as.utexas.edu, ivan@astro.as.utexas.edu, spies@astro.as.utexas.edu

A&A 439, 805–822 (2005)
DOI: 10.1051/0004-6361:20053192

Rights and wrongs of the Hipparcos data

A critical quality assessment of the Hipparcos catalogue

F. van Leeuwen

Validation / verification

- Each Gaia Coord. Unit (C.U.) yet implemented its own tests
  - Junit unitary test
  - Integration tests
    - Include sometimes comparisons to external data (e.g. RV standards)

- Validation ≠ verification
  - Verification: "*Are we building the Catalogue right?*"
  - Validation: "*Did we build the right Catalogue?*"
    - Change of perspective from what is being done in the DPAC CU3-8s
    - Starting sometimes from scratch

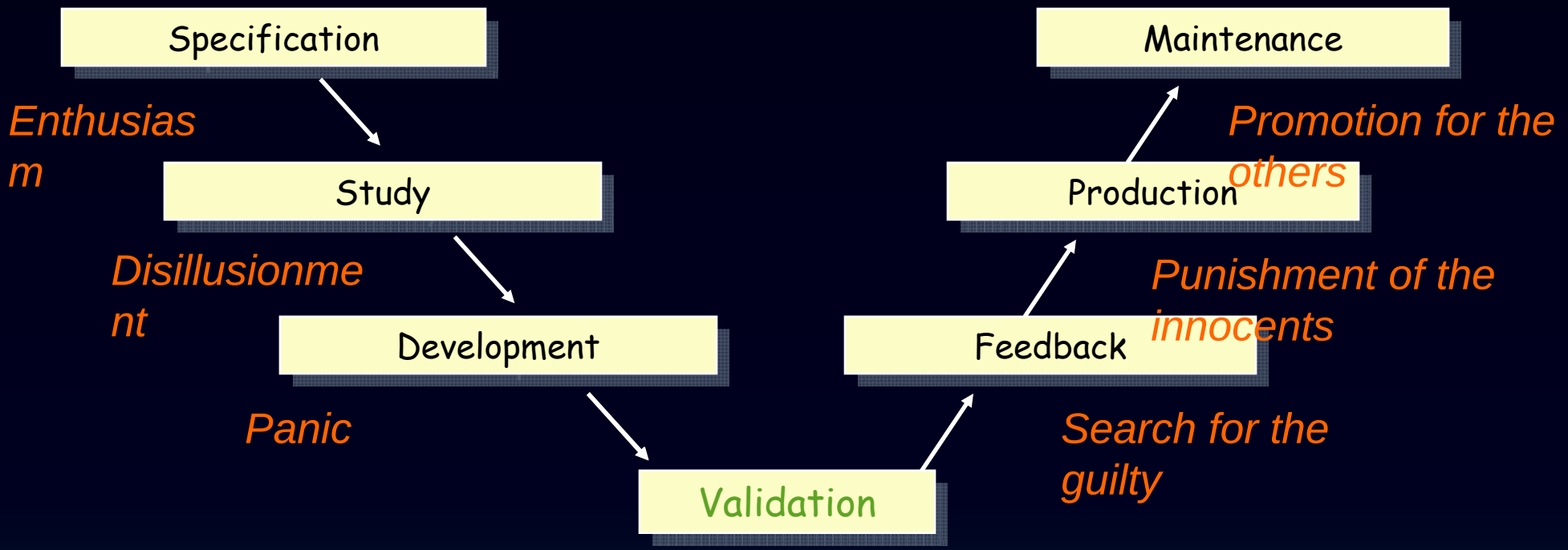
- Will be based on some external prior data knowledge
  - Not being too much dependent on it
  - Priors which should not (too much) be present in the DPAC chain

Validation goals

- Check and ensure the quality of the Catalogue
  - Have a critical look at the output
    - *"The* (wo)*man of science has learned to believe in justification, not by faith, but by verification"* - Thomas Huxley
  - Do not leave gross errors undetected before publication
  - And correct mistakes as soon as possible!
    - Feedback to C.U. between intermediate Catalogue releases

- Assess the statistical properties
  - Unbiased parameters (systematics)
  - Unbiased parameter standard errors (random)
  - Possibly indicate the level of systematics (or data correlation)
  - Validation is also part of the documentation (Catalogue properties)

Cycle of life ?



How to proceed with a validation?

How

- It is assumed validation occurs at each intermediate release
  - Or at least some basic validations occur in the release process
  - Should not slow down the publication though
- A lot of routine scenarios may have to be implemented
  - Indicating what to test and what to do when tests fail
  - Running routinely or on demand
- Validation approach should be transversal
  - *Instruments* already handled by Coord Units (astro/photo/spectro)
  - *Objects* sometimes handled by C.U. too (CU4, CU8)
  - Validation will thus mostly be based on scientific topics with data being the *combination* of individual C.U. data

Gaialeaks

- Validation tests are scientific by nature
  - Caveat: no science should be done with that!
  - Not before the official publication release
- What to do with data deviating from what was assumed before?
  - Either coming from e.g. calibration errors
    - Back to C.Us for handling
  - Or from some possibly yet unknown scientific phenomena
    - The correct definition of outliers may also be: the future science
    - … nothing special should be done before publication!
- Some precautions should be taken
  - To avoid dissemination
  - No more tests than what is needed
  - To make clear that the validation job is for the Gaia quality only!

What validation items ?

Typical Work Packages

- Tests on internal consistency
- Problem-based tests
- Comparison with a Galaxy (Besançon) model
- Comparison with external catalogues
- Special objects: SSO, DMS, variables
- Statistical & graphical analysis

WP: Internal consistency

- Basic checkings: formal validation
  - Parameter content (check NaN, types, etc.)
  - Subfields present as indicated, e.g.:
    - epoch data present (when and only when indicated)
    - RVS data present as indicated
  - All fields are within valid ranges
  - Check for outliers

- Internal consistency
  - Use assumed properties of parameters (e.g. positivity)
    - No large proper motions for distant stars
  - Exploit intrinsic redundancy between instrument data
    - E.g. photometry should be consistent with spectroscopy
    - Gaia is an complete observatory in orbit!

WP: Problem-based tests

- Build tests based on what is known to produce effects on given parameters
  - Instrumental or calibration problems
  - Classification errors
  - Processing shortcuts, rough models
- Examples, to be more specific
  - Analysis of the variability properties both spatially and in time
    - as photometric calibration problems introduce a spurious variability
  - Check the distribution of parallaxes
    - Annual thermal or calibration effects would introduce a parallax bias
  - Compute distributions of distance to nearest neighbour
    - Components only (possibly redundancies?)
    - Components + sources (possibly redundancies?)
    - From SSO observations to nearest non-SSO (redundancies?)

WP: Model-based tests

- Develop code on Gaia *simulated* data
  - Extract “truth” for all observables
    - Compute the distribution, confidence intervals, ranges for all parameters
    - Correlations between these observables
  - Understand and explain the main structures (see e.g. Hipp Vol 1)
- Apply this code on actual Gaia catalogue data
  - Apply statistical tests
  - Checking whether the large, expected structures are present
  - Not going into details

WP: Model-based tests

- A large work yet done !
  - By (Barcelona-led) CU2
- CU2 output
  - Universe model
    - Based on Besançon Galaxy model
    - With large add-ons (variable, binaries)
  - Gaia Analysis Tool - GUMS
    - Produces statistics (numbers) or tables to which data can be compared…
  - Add to this: specific models
    - E.g. for solar system objects

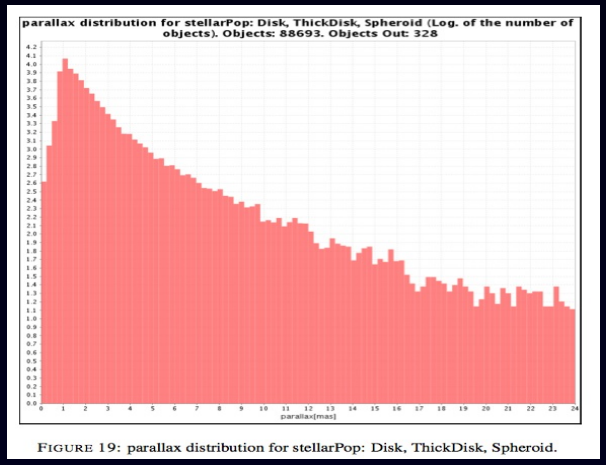


FIGURE 19: parallax distribution for stellarPop: Disk, ThickDisk, Spheroid.

Parallaxes (top)
[Fe|H] vs velocity
(bottom)

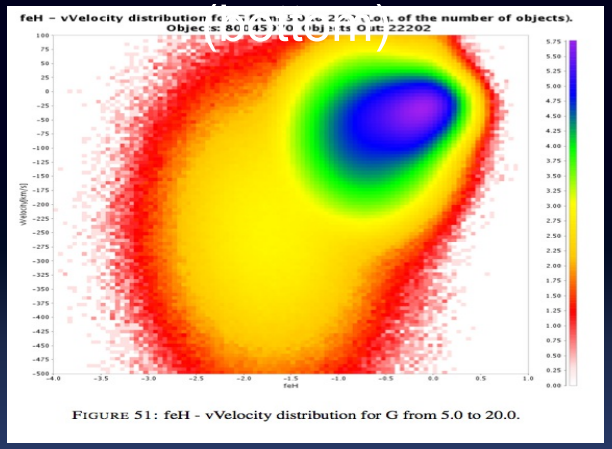


FIGURE 51: feH - vVelocity distribution for G from 5.0 to 20.0.

WP: External tests

- A very simple recipe
  - Get external data
  - Make cross-matching
  - Compare data

- More complicated in practice !
  - Difficulties to find equivalent data
    - E.g. for astrometry, lack of precision, high level of systematics
    - One reason why Gaia will be launched!
  - Difficulties to X-match
    - No other all-sky survey with a comparable angular resolution and similar multiple star discovering power
  - Difficulties to compare
    - Should not attribute to Gaia, errors coming from comparison data!

WP: Statistics & Visualisation

- Tests will be statistical
  - Blind tests : e.g. testing systematically ranges of observables
- An effort of fast visualisation is needed
  - All CU2 GAT graphs
  - By epoch or temporal variations
- Comparisons will be far from obvious
  - Beyond scientific competence, statistical analysis skills are needed
  - E.g. working with truncated, censored or correlated data
    - Limited magnitude range, relative precision censorship

Typical validation scenarios (not exhaustive)

Basic checks (examples)

- Subfields present as indicated, e.g.:
  - epoch data present (when and only when indicated)
  - RVS data present as indicated
- Distributions of distance to nearest neighbour, e.g.:
  - components only (possibly redundancies?)
  - components + sources (possibly redundancies?)
  - from SSO observations to nearest non-SSO (redundancies?)
- Fields
  - all fields are within valid ranges
  - all fields have "reasonable" distributions
  - check for outliers
  - for some fields checks may have to be made separately for different classes of sources

Global checks (examples)

- Sky distributions, e.g.:
  - all sources, except components
  - sources with G<20^{m}, except components
  - median errors for various quantities for various groups of sources
  - distributions of significantly negative parallaxes
- Characterisation of the bright limit
  - which bright stars are missing
  - check surroundings of bright sources for artifacts
- Characterisation of the faint limit
  - will depend e.g. on the number of transits
- Proper motions
  - High proper motion stars are successfully recovered
  - Proper motions for sources with very small parallaxes

Parallax comparisons

- What has been done two decades ago for Hipparcos++
  - Mostly based on positivity
  - Existing ground-based data otherwise very poor
  - Photometric parallaxes + statistical ML model (truncated data)
  - Distant stars

- From that we get a confidence in the data (on a global scale)
  - Parallax systematics + standard errors correctly estimated
    - Now the correlation at small angular scales will be more scrutinized!
  - Need for systematics < 0.1 μas
    - Because data will be averaged, hoping to improve with $1/\sqrt{N}$
  - Checking systematics at the 0.1 μas level yet difficult to achieve
    - Need 5000 bright stars… or 10 million 20^m stars (σ=0.3 mas/star)
    - Using all detected quasars < 20^m I expect a 0.4 μas level only

Typical comparison data used

- Stellar kinematics
  - Which contains both astrometric and spectroscopic data
- Rough consistency for main galactic populations
  - Between position / kinematics / chemical composition
- HR diagram for special populations
  - mixing astrometry + photometry
- Cepheids and other distance indicators
  - Astrometry+photometry+variability

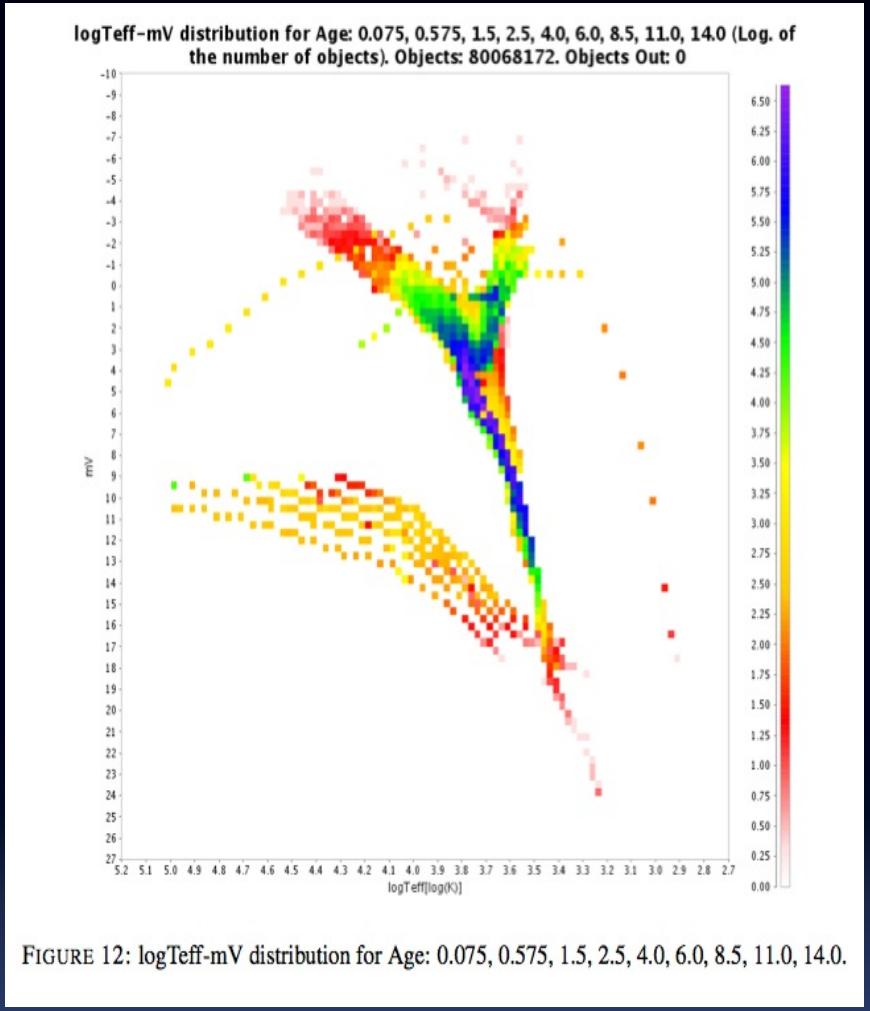


FIGURE 12: logTeff-mV distribution for Age: 0.075, 0.575, 1.5, 2.5, 4.0, 6.0, 8.5, 11.0, 14.0.

Spatial tests (e.g.)

- Production of 3D spatial maps
- Analysis of the on-sky (2D) spatial distribution of poorly classified objects or non-classified objects
  - e.g. low DSC probability
  - and their spatial neighbourhoods,
  - to see whether photometric delending/crowding problems may be an issue (this could feed back into improving BP/RP extraction)
- Analysis of the 3D interstellar extinction distribution
  - Compared to our current understanding of gas and dust distributions from infrared surveys (colours)
- Analysis of the Galactic metallicity distribution
  - both spatially and as a function of stellar kinematics and ages (produced by CU8)

Luminosity tests (example)

- H-R diagrams for selected stellar populations
  - e.g. known globular/open clusters, compared to current knowledge
- G-band absolute magnitude function
  - perhaps the luminosity function too, with the APs calculated in CU8
  - For various samples of stars + compare with current knowledge.
- QSO redshift and luminosity distribution
  - Compared to results from SDSS, Pan-STARRS and other surveys,
  - Taking into account the selection effects.
  - This will help understand the type I and type II errors in the CU8 QSO/star classification

(Many) open questions

- What are the obvious needs for comparison data ?
  - Existing Catalogues
    - E.g. for star clusters or other homogeneous populations
  - Special objects
    - Special stars, variables, multiple stars, solar system objects
  - Other data ?
    - Beside the existing data acquired by some CU for validation purposes
    - Refer to C. Soubiran's talk
- When a problem is detected, how to solve it ?
  - External follow-up data may be needed on a case by case basis
  - Does not always settle the problem
    - For example, the Hipparcos (possible) problem in the Pleiades is still not fully solved 15 years later

Technical points open (a lot)

- Data queries for validation purposes
  - Will need all kind of access (sequential, spatial, etc) !
  - Queries may be complicated conditional queries (software!)
  - By object & transit
- How to compare Gaia to external data ?
  - How to X-match ?
  - How to handle missing values ? etc. ?
    - Choices will have an obvious impact on the architecture/processing
- How to be robust enough ?
  - Outliers, resolved vs unresolved multiple stars, dense areas
  - And truncated, censored or correlated data

Next steps (organisation)

Organisation (GAP/CU9)

- The publication of Gaia will be the task of a new CU9 Unit
  - For the moment under a GAP
- The various CU9 work areas are not independent:
  - E.g., the validation will need the tools developed within CU9
  - These tools depend on the Operations and Support area.
- Noting that data validation can also indirectly be a validation of the analysis tools.
  - This has to be accounted for in the work package definition and in the timescales (e.g. validation will then need its own tools before)

Organisation (work and FTEs)

- Topical meetings will be needed
  - In order to avoid multiple meetings, should preferably be done within a full GAP or other DPAC meeting
- Yet several volunteers from various European Institutions
  - Barcelona, Besançon, Bruxelles, Geneva, Heidelberg, Nice, Paris
  - FTEs can be:
    - Scientists with a small FTE implication, acting for consultancy
    - Engineers for tool developments in V.O. env. with large FTEs
  - Not a private club
  - And, again, actual work for Gaia, no personal scientific return
- List of people concerned
  - http://www.rssd.esa.int/wikiSI/index.php?title=GAP:Data_Validation&instance=Gaia

Organisation (areas)

- By scientific area
  - Solar system
  - Stellar physics
  - Galactic structure
  - Reference system and relativity
- By themes
  - Photometry, spectroscopy, kinematics
  - Multiplicity
  - Variability
- A lot of scientific areas already covered


Thank you for
your attention