bayesian treatment of distance errors in the cosmic distance ladder

Germán Chaparro Molano
UNIVERSIDAD ECCI, COLOMBIA

Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS 00, 1 (2019)
Advance Access publication 2019 March 02

doi:10.1093/mnras/stz615

Predicting extragalactic distance errors using Bayesian inference in multimeasurement catalogues

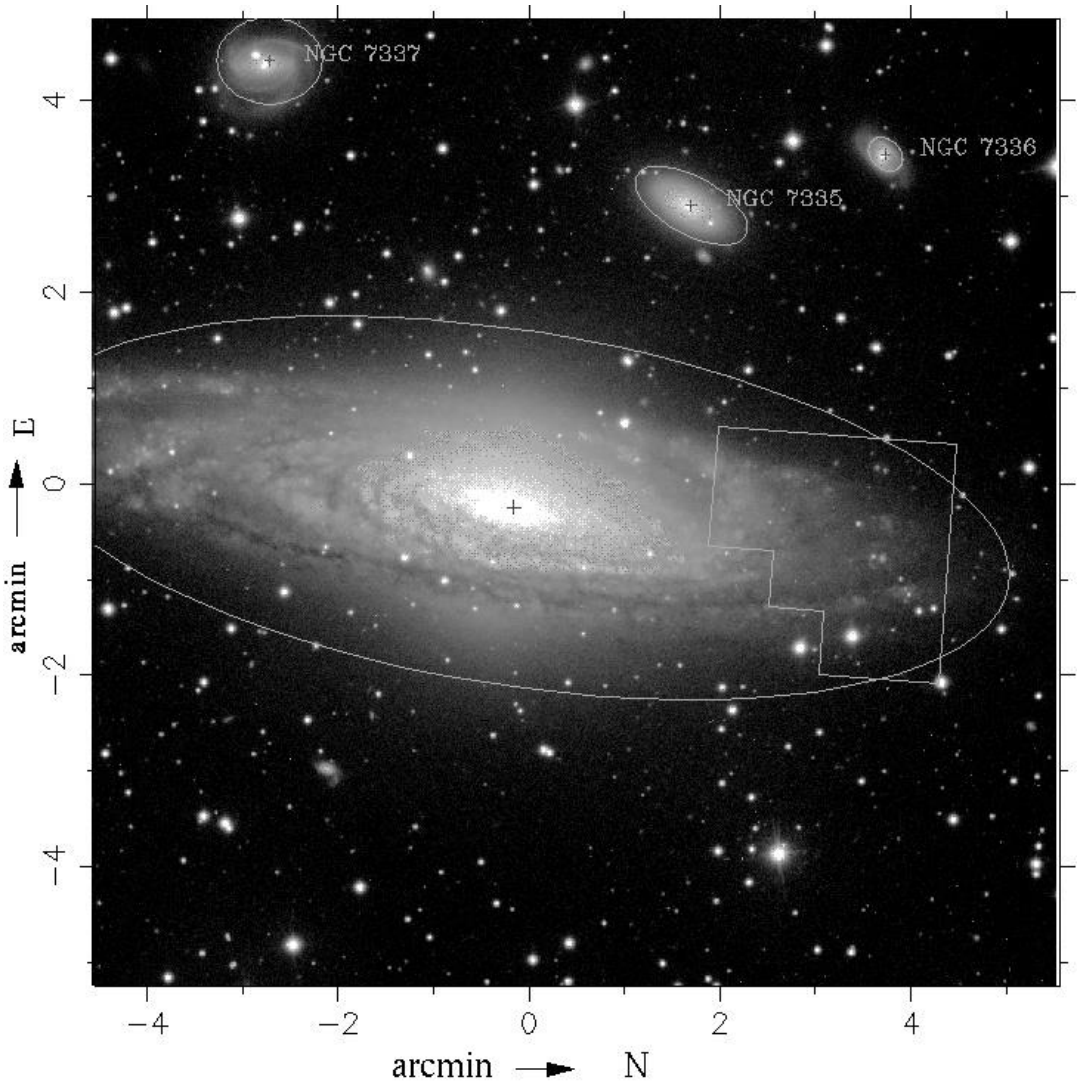
Germán Chaparro-Molano [1★] Juan Carlos Cuervo,[1] Oscar Alberto Restrepo Gaitán[1,2] and Sergio Torres Arzayús[3]

[1] *Vicerrectoría de Investigación, Universidad ECCI, Bogotá, Colombia*
[2] *Radio Astronomy Instrumentation Group, Universidad de Chile, Santiago de Chile, Chile*
[3] *Centro Internacional de Física, Bogotá, Colombia*

Accepted 2019 February 26. Received 2019 January 17; in original form 2018 May 3

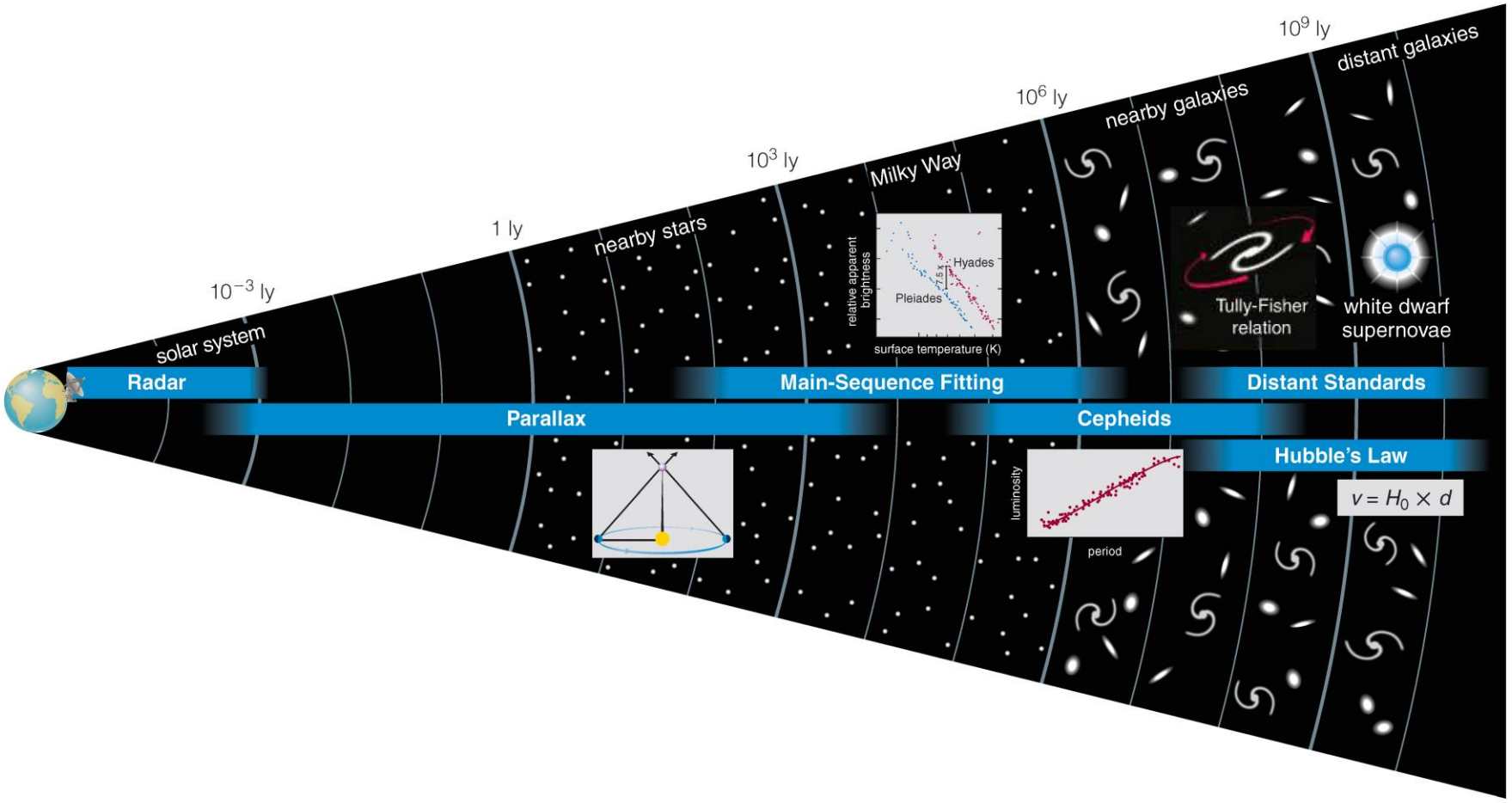
Chaparro Molano+2019 astro-ph:1805.02578
https://github.com/saint-germain/errorprediction

NGC 7331 (Group)

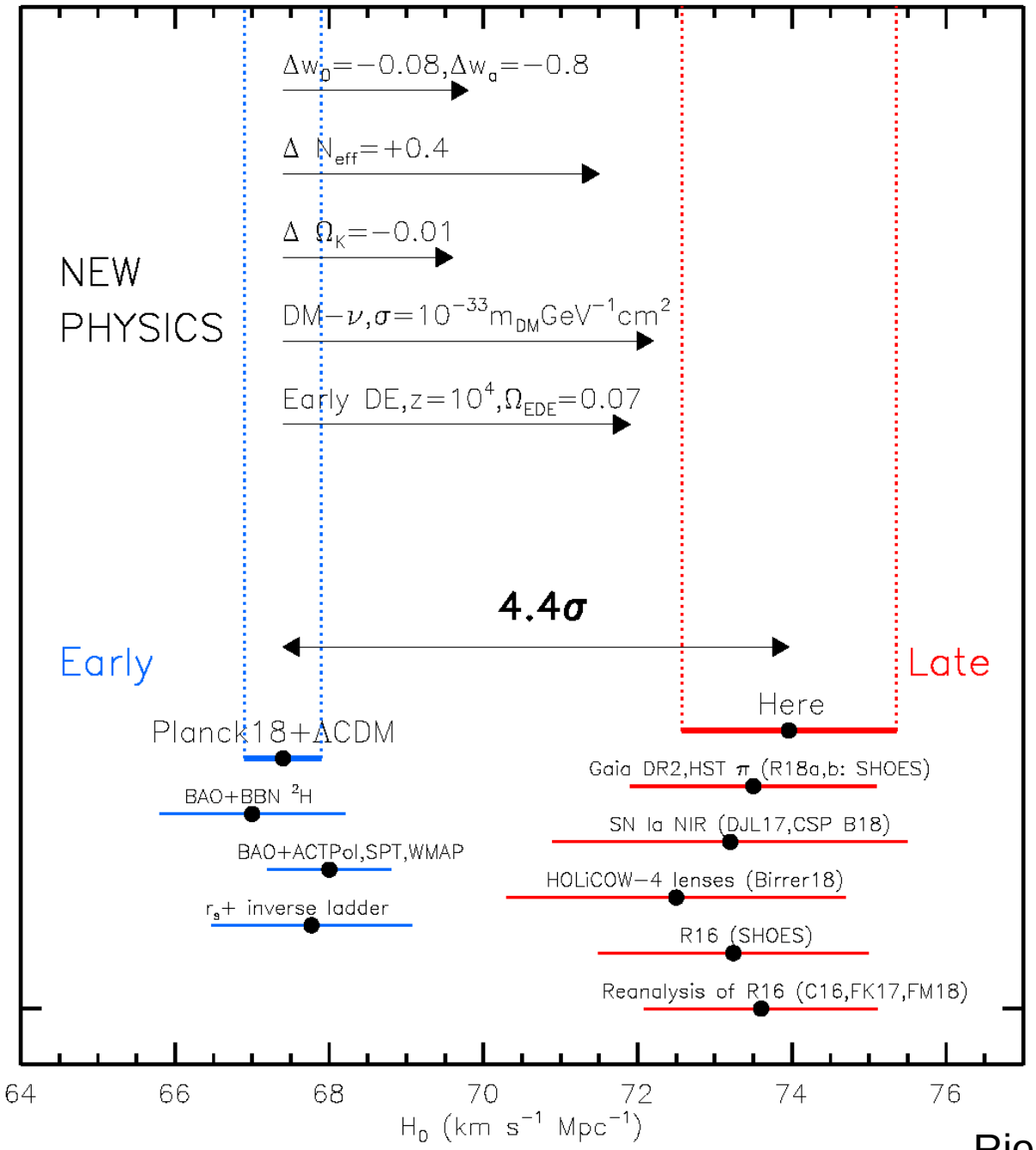


Hughes et al 1998

The Cosmic Distance Ladder



From The Essential Cosmic Perspective, 6th Ed., Bennet+2012

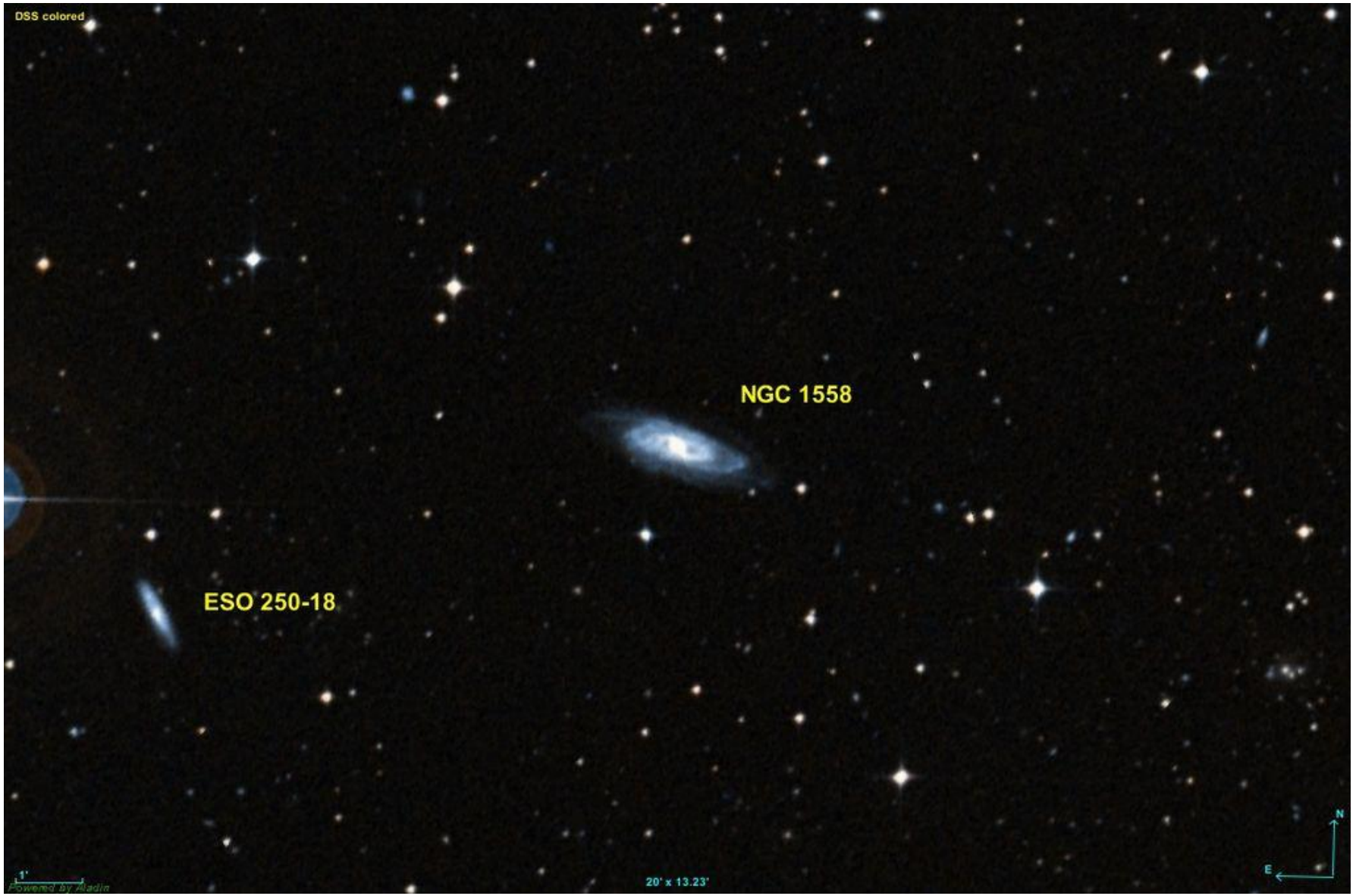


Riess et al 2019

Redshift-Independent Extragalactic Distance Catalogs

- NED-D (NASA/IPAC)
  ~300k measurements for ~180k galaxies
- HyperLEDA A007 (Lyon/SAO)
  ~12k measurements for ~4000 galaxies
- Cosmicflows-3 (Tully+2016)
  ~10k galaxies w/ up to 4 distance methods ea.

NGC 1558


DSS colored
NGC 1558
ESO 250-18
1'
Powered by Aladin
20' x 13.23'
N
E

NGC 1558

Distance modulus	Distance modulus error	Distance (Mpc)	Distance error (Mpc)	Method	Year
34.12	0.32	67	10	Tully-Fisher	1992
34.2	0.32	69	10	Tully-Fisher	1992
33.92	0.43	61	12	Tully-Fisher	1997
34.15	0.43	68	13	Tully-Fisher	1997
34.25	0.43	71	14	Tully-Fisher	1997
33.88	0.8	60	22	IRAS	1997
33.99	0.07	63	2	Sosies	2002
33.86	0.46	59	13	Tully-Fisher	2007
33.88	0.47	60	13	Tully-Fisher	2007
33.98	0.4	63	12	Tully-Fisher	2007
33.99	0.45	63	13	Tully-Fisher	2007
29.08	0.36	7	1	Tully-Fisher	2009
33.95	0.29	62	8	Tully-Fisher	2009
34.12	0.27	67	8	Tully-Fisher	2009
34.16	0.2	68	6	Tully-Fisher	2013

NED-D: A Master List of Redshift-Independent Extragalactic Distances

https://ned.ipac.caltech.edu/Library/Distances/

NGC 1558

“Frequentist” methods:

P: Propagation of errors (Cosmicflows-3, Tully et al 2016)

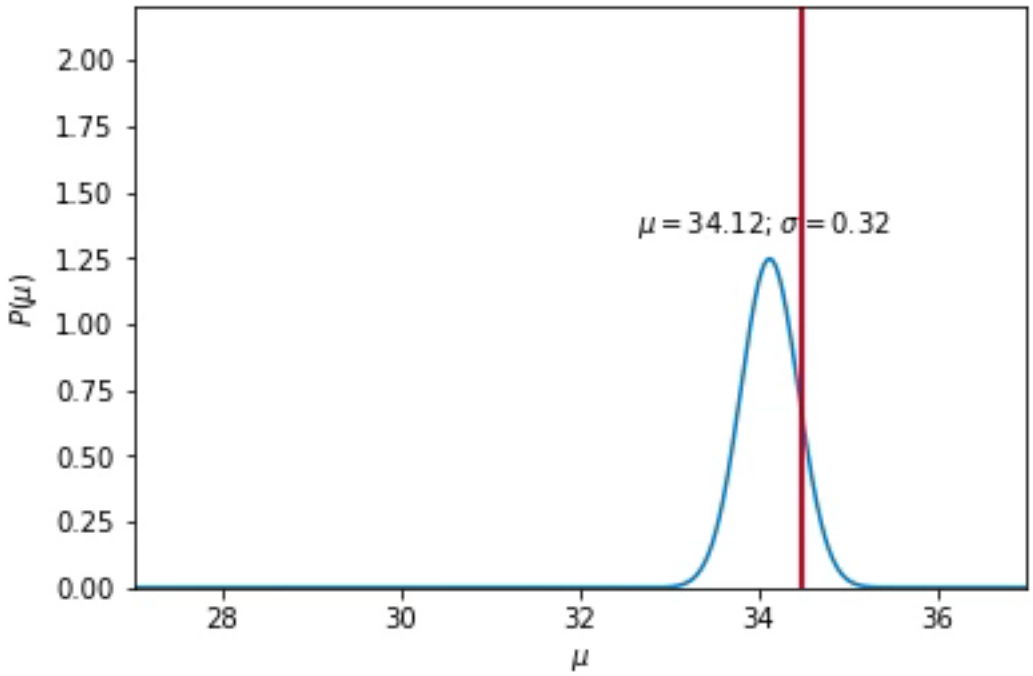
$\sigma_P = 1.6$ Mpc

Q: Quadrature sum of P error and weighted stdev (to account for spread between measurements)

$\sigma_Q = 22$ Mpc

NGC 1558

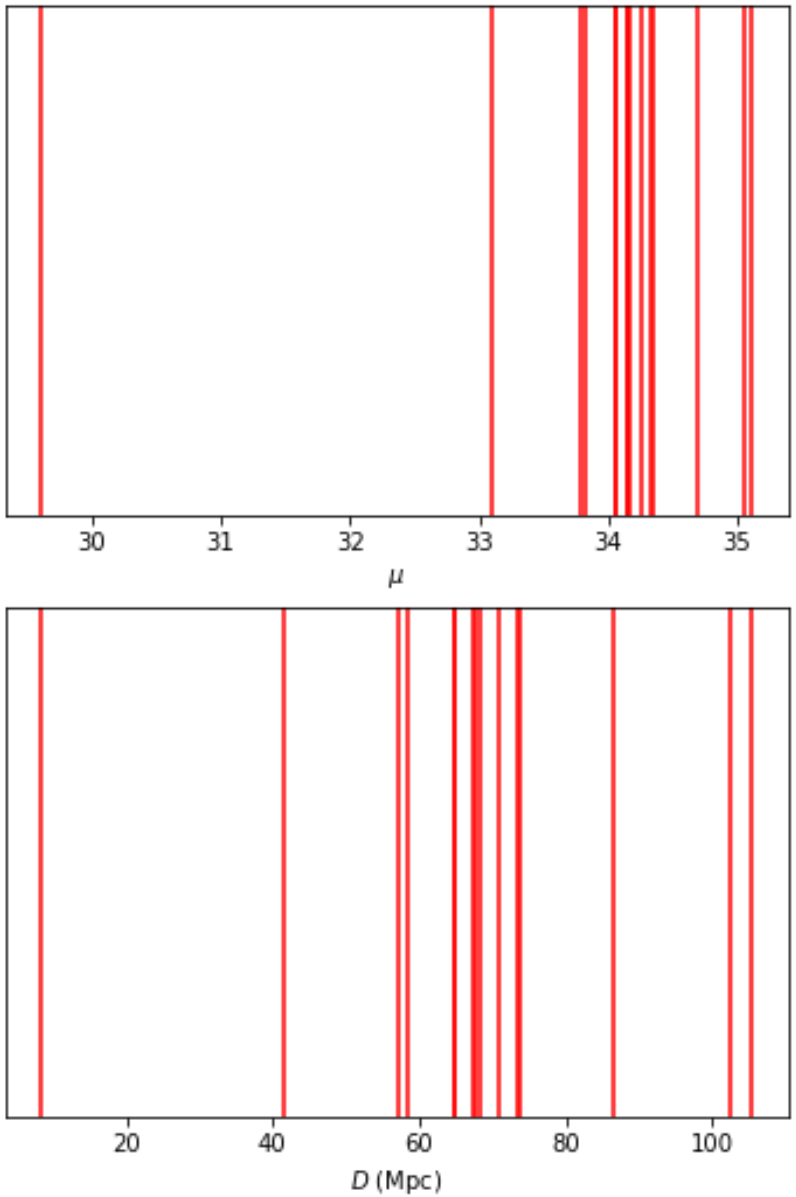
Distance modulus	Distance modulus error
34.12	0.32
34.2	0.32
33.92	0.43
34.15	0.43
34.25	0.43
33.88	0.8
33.99	0.07
33.86	0.46
33.88	0.47
33.98	0.4
33.99	0.45
29.08	0.36
33.95	0.29
34.12	0.27
34.16	0.2



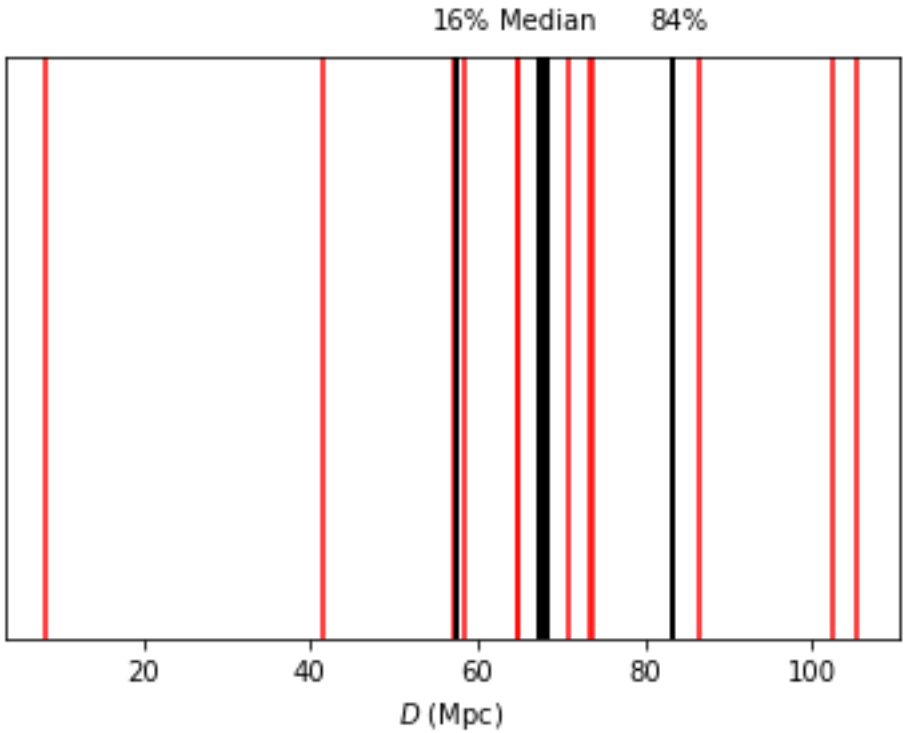
NED-D: A Master List of Redshift-Independent Extragalactic Distances
https://ned.ipac.caltech.edu/Library/Distances/

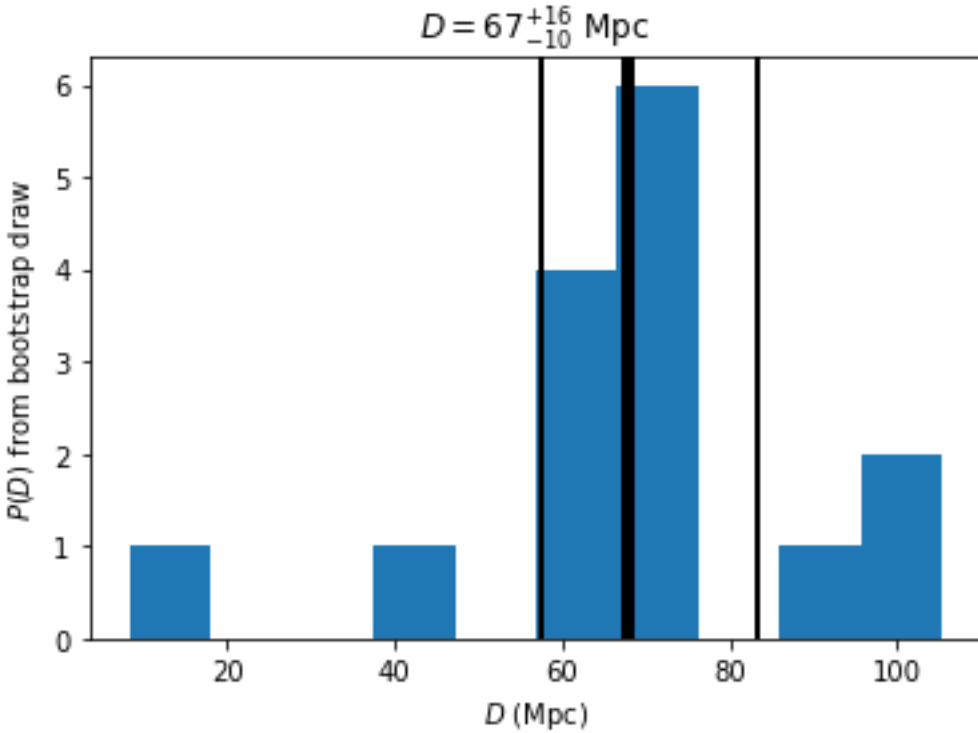
NGC 1558

Distance modulus	Distance modulus error
34.12	0.32
34.2	0.32
33.92	0.43
34.15	0.43
34.25	0.43
33.88	0.8
33.99	0.07
33.86	0.46
33.88	0.47
33.98	0.4
33.99	0.45
29.08	0.36
33.95	0.29
34.12	0.27
34.16	0.2



NGC 1558


16% Median
84%
20
40
60
80
100
D (Mpc)


$D = 67^{+16}_{-10}$ Mpc
P(D) from bootstrap draw
0
1
2
3
4
5
6
20
40
60
80
100
D (Mpc)

NGC 1558

“Frequentist” methods:

P: Propagation of errors (Cosmicflows-3, Tully et al 2016)

$\sigma_P = 1.6$ Mpc

Q: Quadrature sum of P error and weighted stdev (to account for spread between measurements)

$\sigma_Q = 22$ Mpc

Proposed (bootstrap + robust)

H: Half-distance between the 84th and 16th percentiles

$\sigma_H = 11$ Mpc

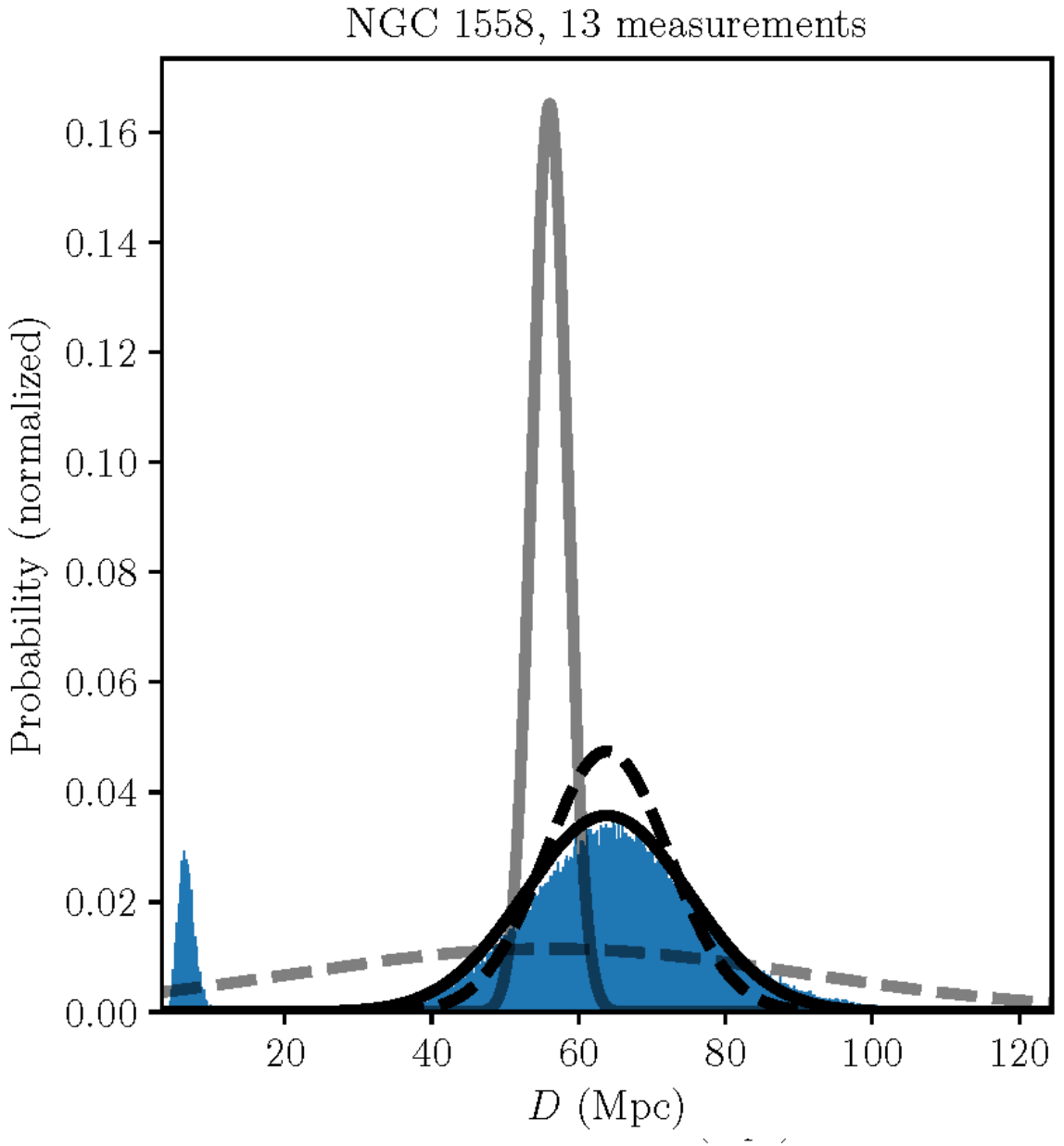
M: Median Absolute Deviation (outlier-proof)

$\sigma_M = 8$ Mpc

If outliers did not affect uncertainty, then H=M

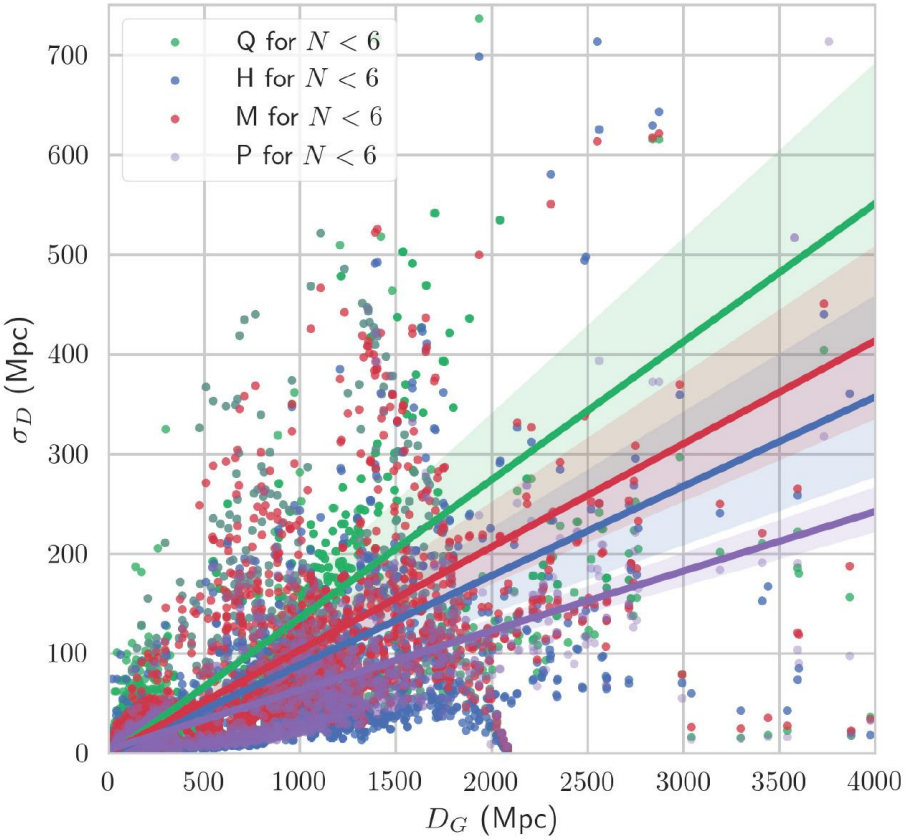
If P and/or Q are good measures of spread, then P and or Q=H

$$D_{\rm G} \sim \mathcal{N}(D, \sigma_{\rm D}^2)$$

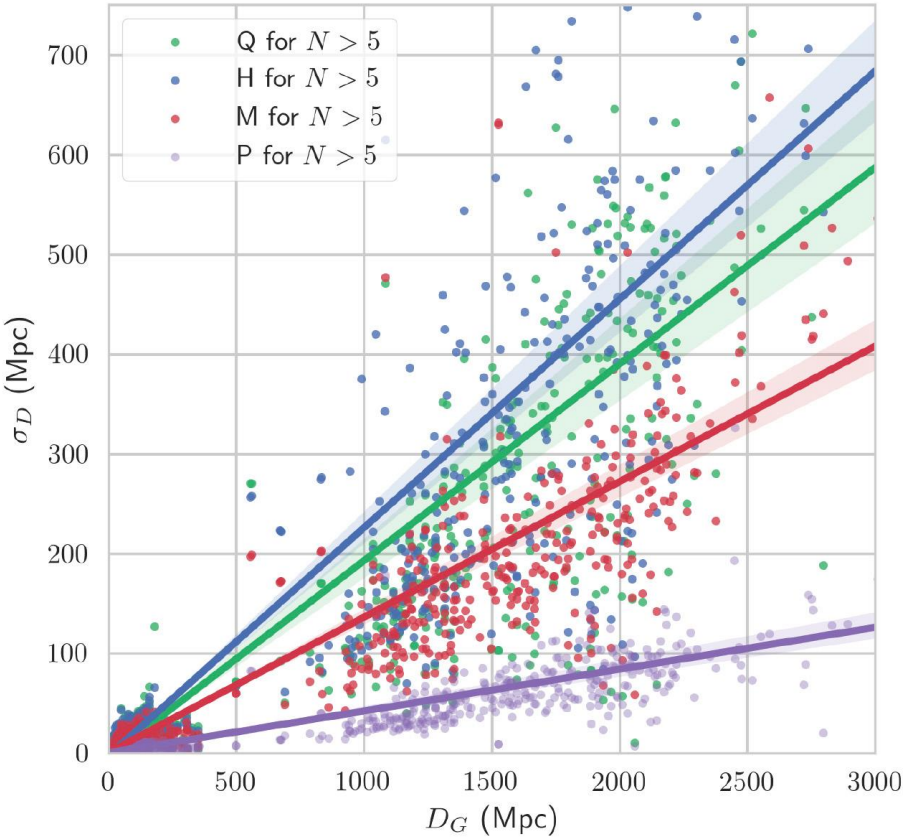


Chaparro Molano+2019 – astro-ph:1805.02578

Errors in NED-D Extragalactic Distances



Few measurements



More than a few measurements

Chaparro Molano+2019 – astro-ph:1805.02578

Pre-Computed Error Tables

GitHub, Inc. [US] | https://github.com/saint-germain/errorprediction

Repository for "Predicting extragalactic distance errors using Bayesian inference in multi-measurement catalogs" paper

Germán Chaparro-Molano, Juan Carlos Cuervo, Oscar Alberto Restrepo Gaitán, Sergio Torres Arzayús; Predicting extragalactic distance errors using Bayesian inference in multi-measurement catalogs, Monthly Notices of the Royal Astronomical Society, , stz615, https://doi.org/10.1093/mnras/stz615

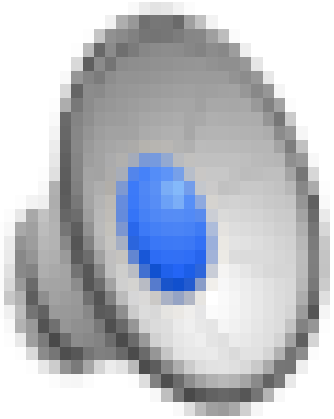
- Published Version
- Preprint: https://arxiv.org/abs/1805.02578

Data tables

Pre-computed error tables for the redshift indepentent extragalactic distance catalogs in HyperLEDA, NED-D, and Cosmicflows-3.

- HyperLEDA
- NED-D
- Cosmicflows-3

M87



- NED-D:
  115 measurements
  M-error = 1.6 Mpc
- HyperLEDA:
  5 measurements
  M-error = 1.2 Mpc
- EHT Collab+2019
  3 measurements
  Error = ~1 Mpc

When to use each error table

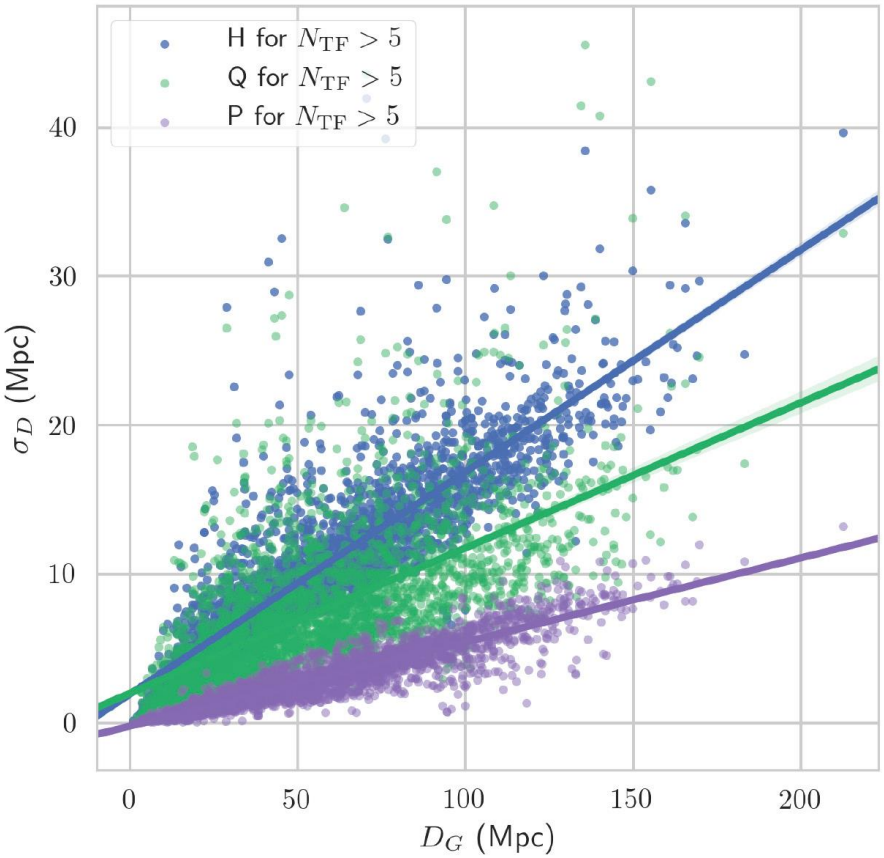
- Comprehensive distances: NED-D
- Comprehensive distances, calibrated: HyperLEDA
- Latest, most trustworthy distances: Cosmicflows-3

Unknown Unknowns

- Missing errors galore!

- NED-D (NASA/IPAC)
  - ~12k measurements without reported errors
  - ~1k galaxies without reported TFR errors
- HyperLEDA A007 (Lyon/SAO)
  - ~1000 measurements without reported errors
  - ~200 galaxies without reported TFR errors

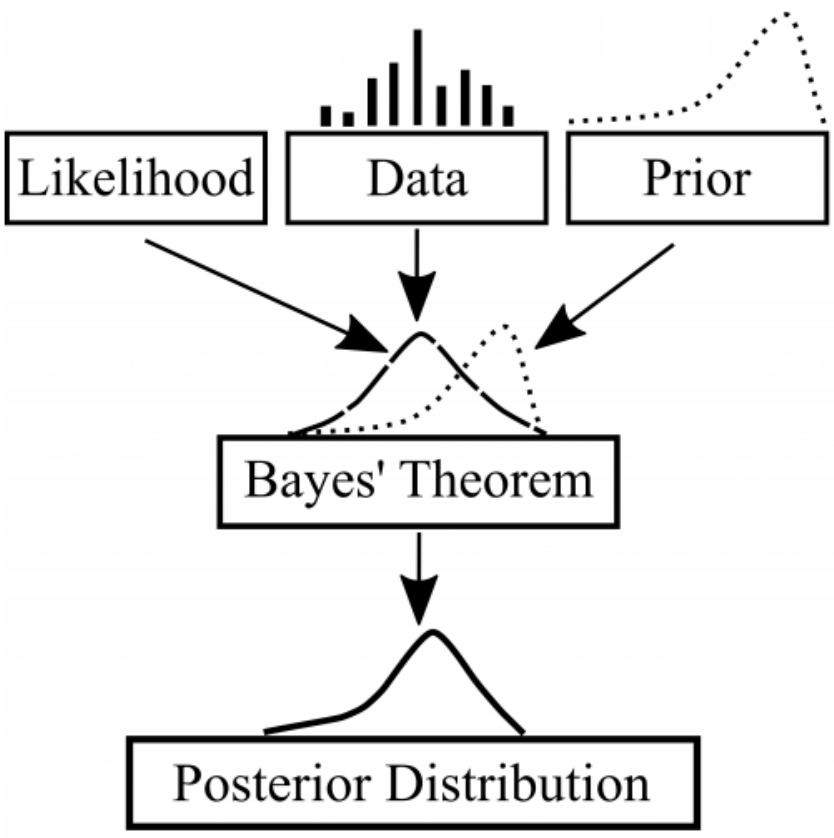
Tully-Fisher Errors in NED-D Extragalactic Distances



Can we extract information on the systematics of the TFR method?
Can we predict missing errors from these systematics?

Chaparro Molano+2019 – astro-ph:1805.02578

Bayesian Inference



Building a Bayesian Predictive Model

Distance Errors take Normally Distributed Values

$$P(\sigma_{\mathrm{D}j}|\hat{\sigma}_{\mathrm{D}j}, \sigma_{\sigma j}) = \mathcal{N}(\hat{\sigma}_{\mathrm{D}j}, \sigma^2_{\sigma j})$$

Likelihood = Probability that all distance errors in a dataset are generated by a Gaussian

$$P(\sigma_{\mathrm{D}}|\hat{\sigma}_{\mathrm{D}}, \sigma_{\sigma}) = \prod_{j}^{m} P(\sigma_{\mathrm{D}j}|\hat{\sigma}_{\mathrm{D}j}, \sigma_{\sigma j})$$

Test whether distance errors can be estimated from D_{G} under a single model depending on a few parameters ($\boldsymbol{\theta}$)

$$P(\sigma_{\mathrm{D}}|D_{\mathrm{G}}, \boldsymbol{\theta}) = \prod_{j}^{m} P(\sigma_{\mathrm{D}j}|D_{\mathrm{G}j}, \boldsymbol{\theta})$$

Posterior can be estimated up to a constant (get values for $\boldsymbol{\theta}$ given the D_{G}data)

$$P(\boldsymbol{\theta}|D_{\mathrm{G}}, \sigma_{\mathrm{D}}) \propto P(\boldsymbol{\theta})P(\sigma_{\mathrm{D}}|D_{\mathrm{G}}, \boldsymbol{\theta})$$

Chaparro Molano+2019 – astro-ph:1805.02578

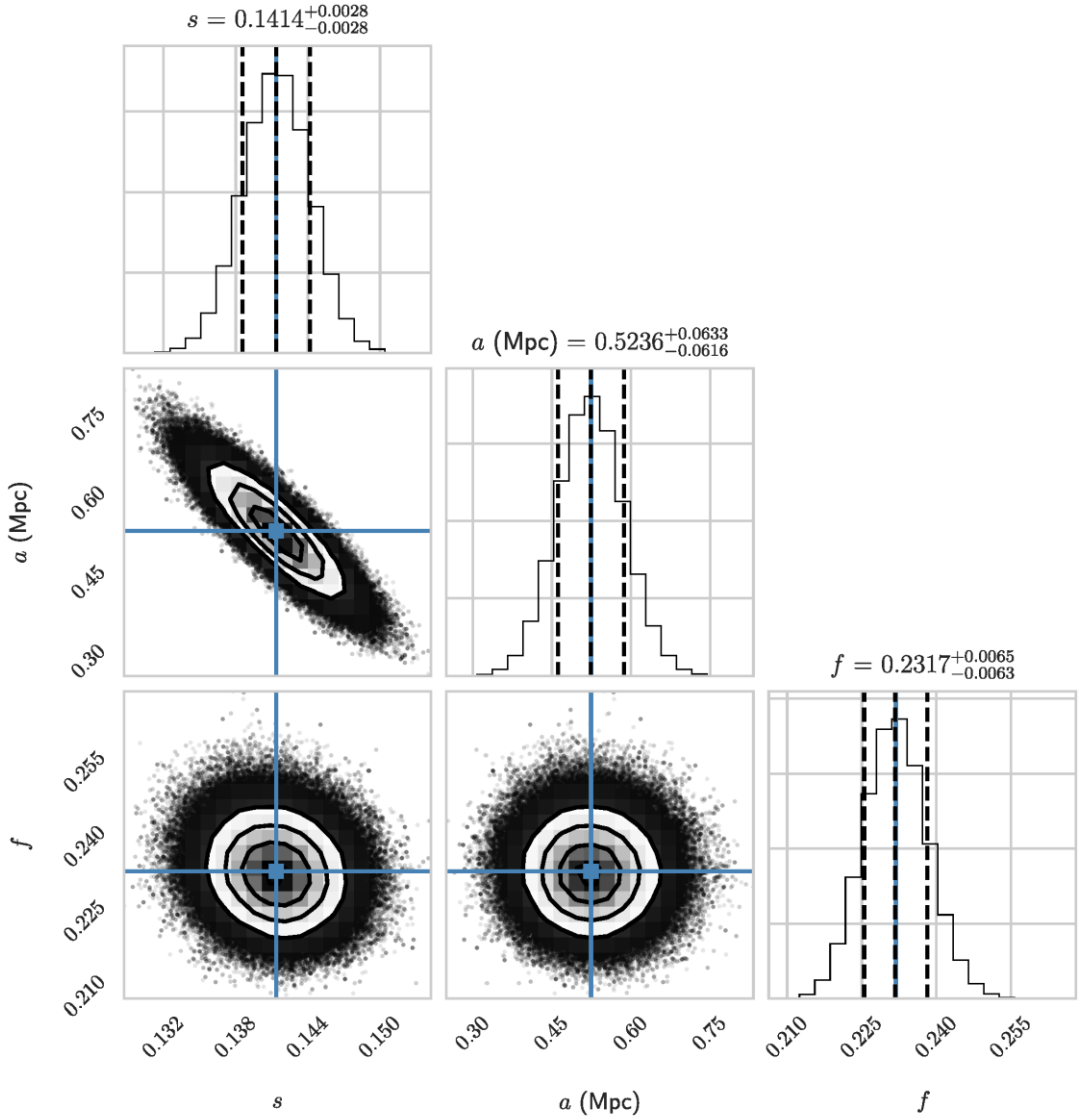
Sampling the Posterior

$$\sigma_{\mathrm{D}} = \sigma_{\mathrm{s}} = s\,D + a$$

$$\sigma_{\sigma} = f\,\hat{\sigma}_{\mathrm{D}}$$

`emcee`

Affine Invariant
MCMC Ensemble Sampler
(Foreman-Mackey 2012)

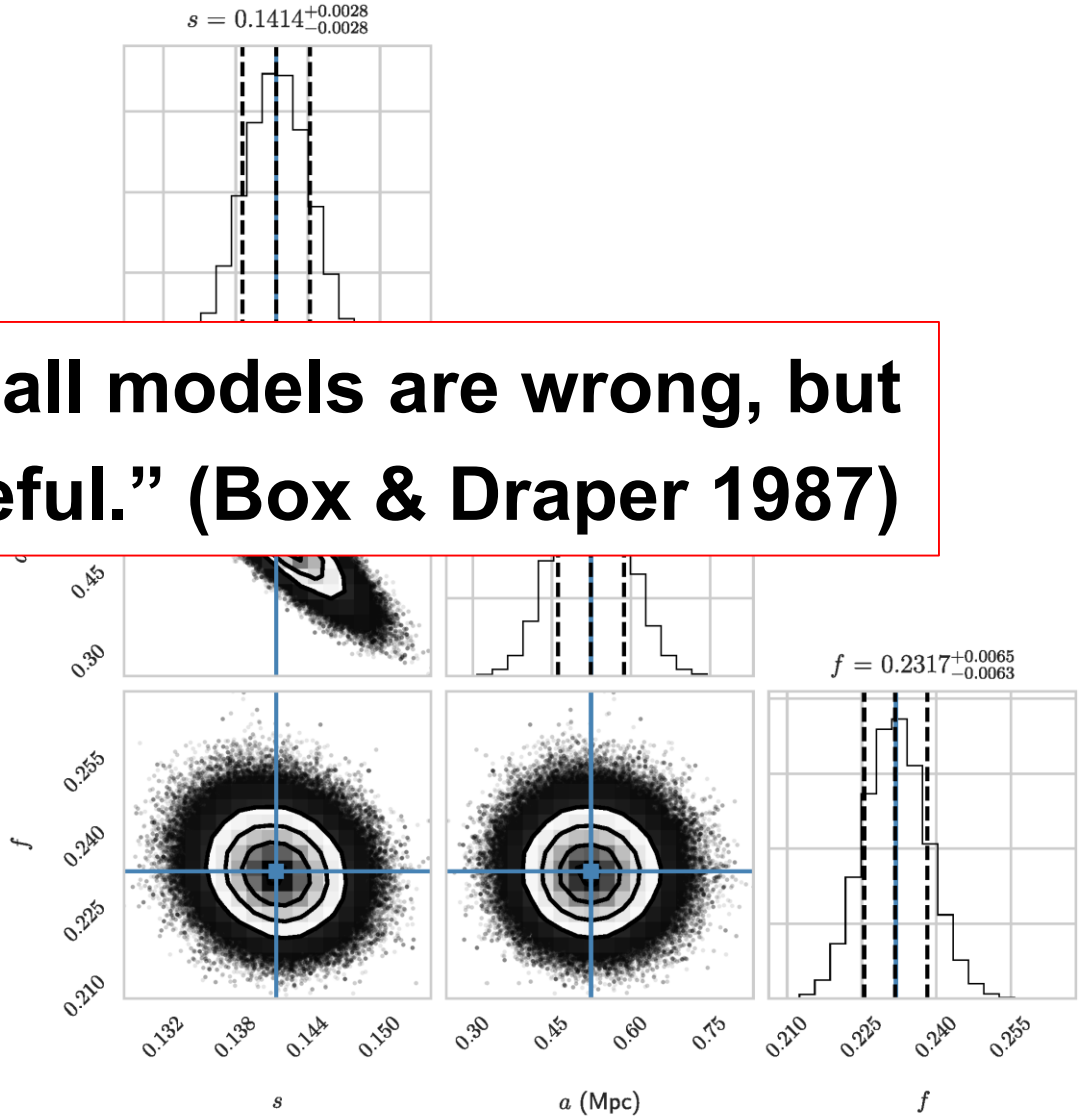


Chaparro Molano+2019 – astro-ph:1805.02578

Sampling the Posterior

$$\sigma_D = \sigma_{\ldots} - sD + a$$

“Essentially, all models are wrong, but some are useful.” (Box & Draper 1987)

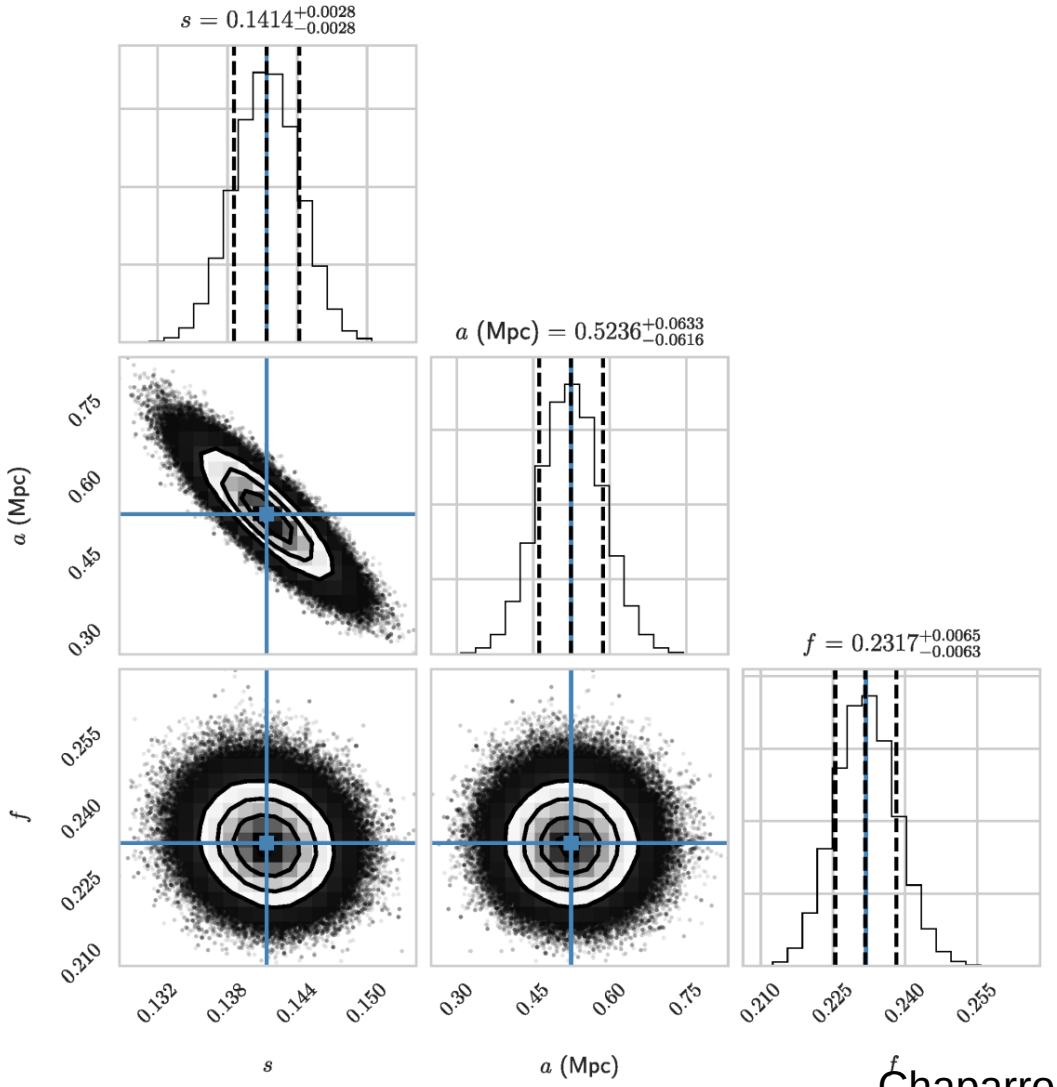


`emcee`

Affine Invariant
MCMC Ensemble Sampler
(Foreman-Mackey 2012)

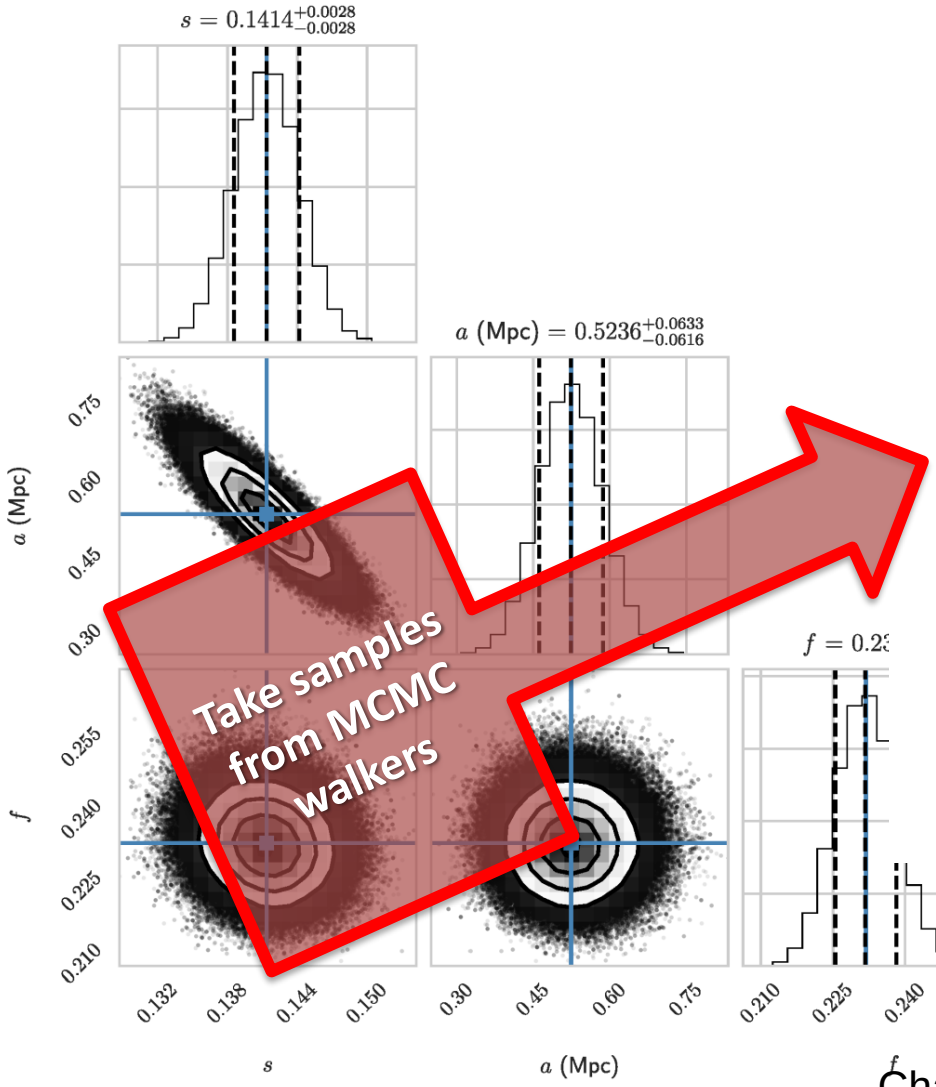
Chaparro Molano+2019 – astro-ph:1805.02578

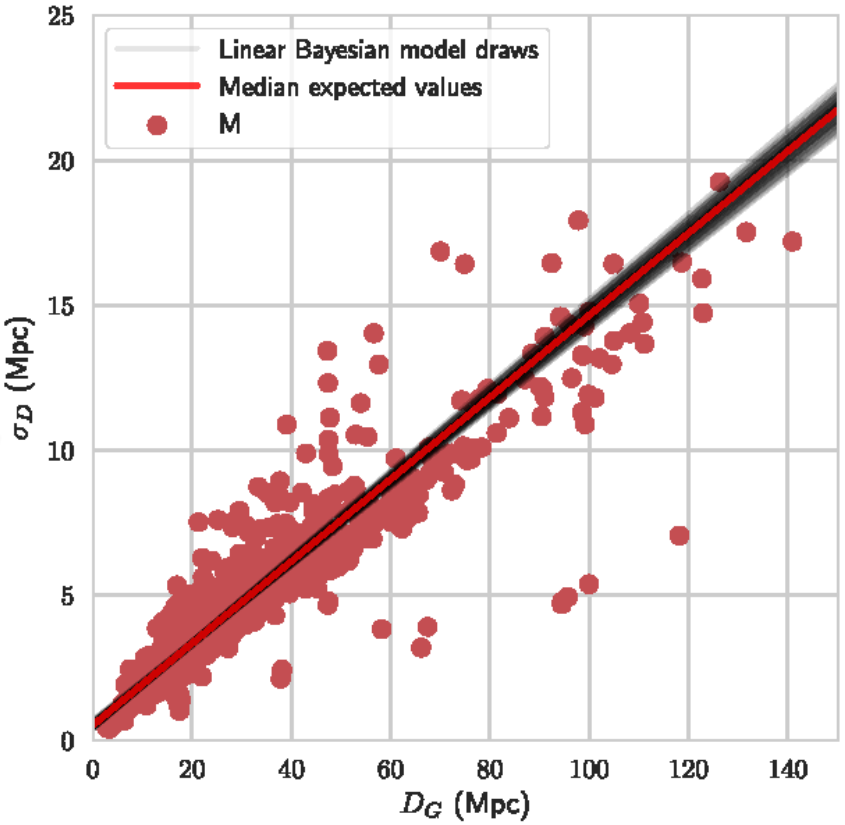
Posterior Predictive Checks



Chaparro Molano+2019 – astro-ph:1805.02578

Posterior Predictive Checks





Chaparro Molano+2019 – astro-ph:1805.02578

Model Comparison: Discrepancy Measure (Gelman+1996, Gelman 2003)

$$\mathcal{D}(\sigma_{\mathrm{D}}|\boldsymbol{\theta}_k)$$

Discrepancy between Data and Expected Values (from model)

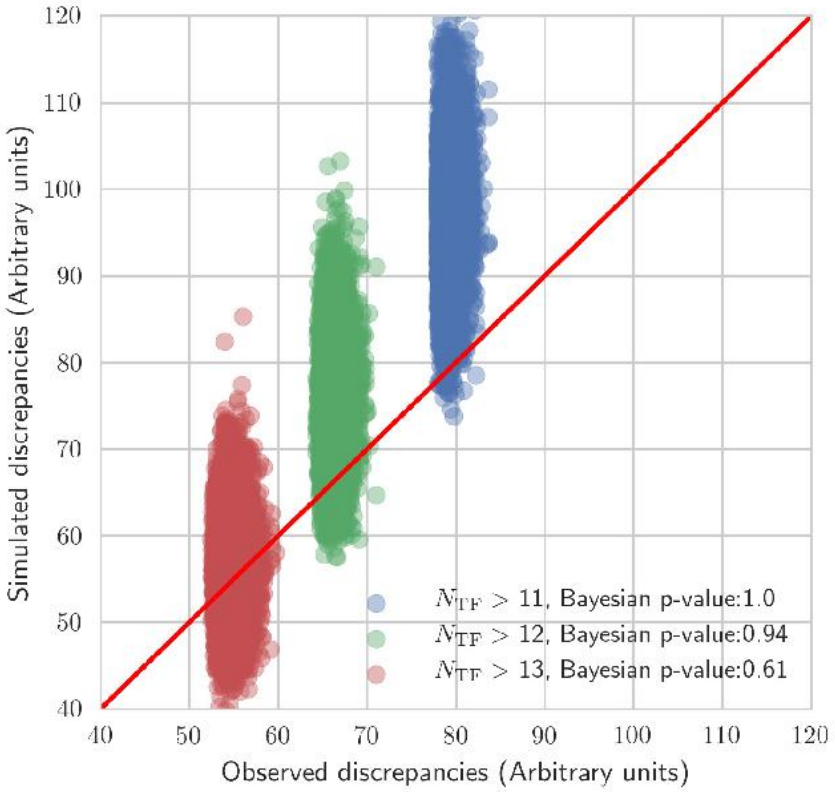
For each MCMC draw $\boldsymbol{\theta}_k$:

Generate synthetic data (from model)

Calculate discrepancies for observed and synthetic data

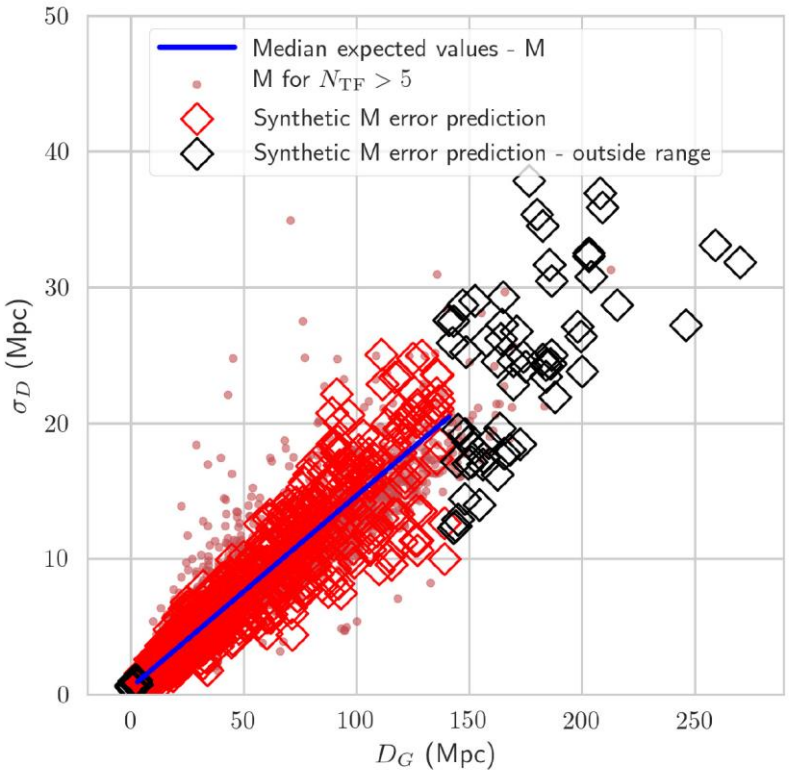
Compare observed and synthetic discrepancies

Get a Bayesian “p-value” = ratio of ‘draws when the observed discrepancies are larger than the synthetic discrepancies’ to ‘total draws’



Chaparro Molano+2019 – astro-ph:1805.02578

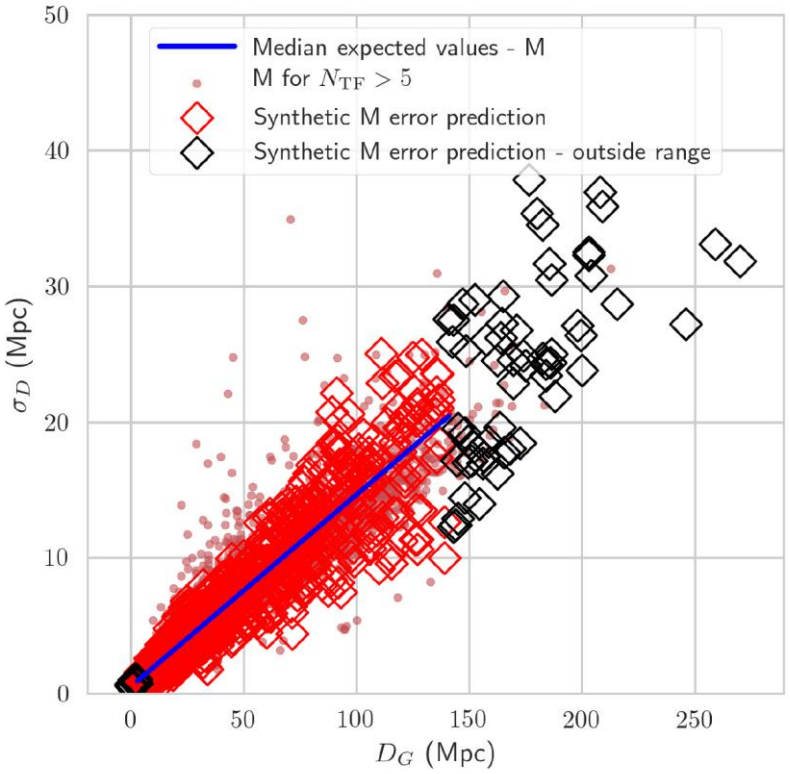
Prediction of Missing Errors



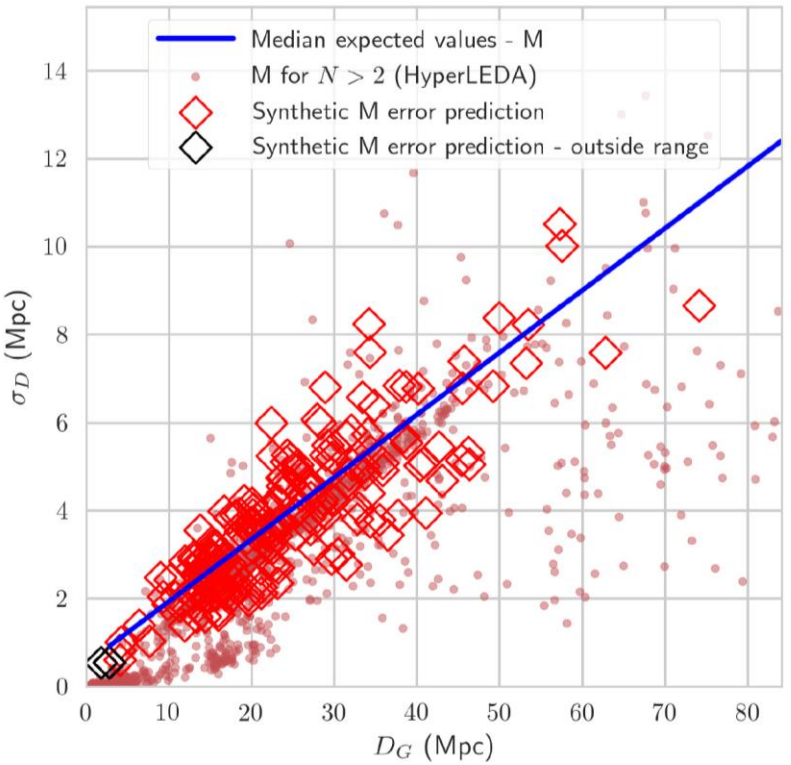
Predictions for 1k galaxies
in NED-D

Chaparro Molano+2019 – astro-ph:1805.02578

Prediction of Missing Errors



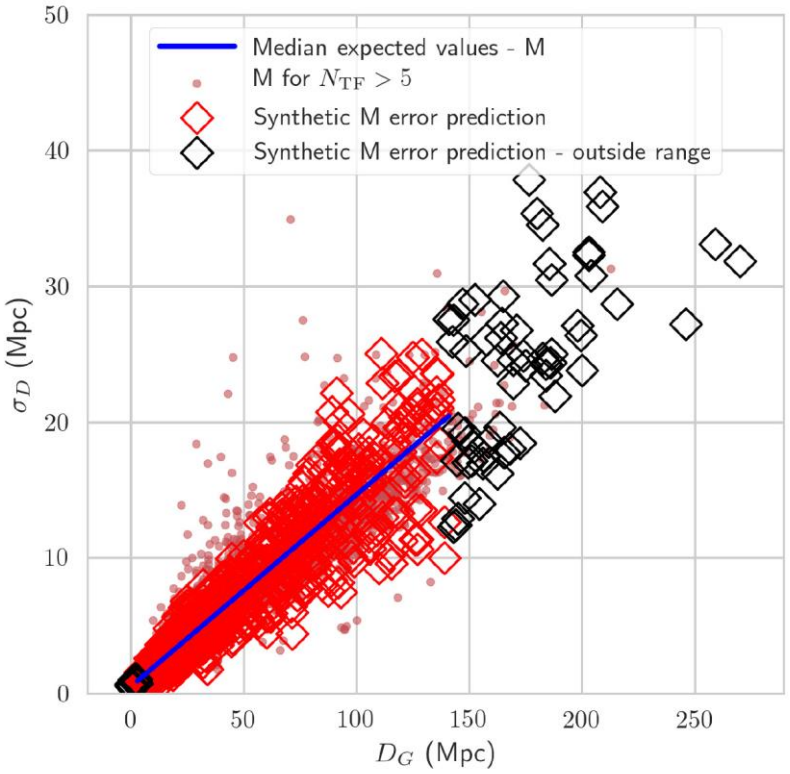
Predictions for 1k galaxies
in NED-D



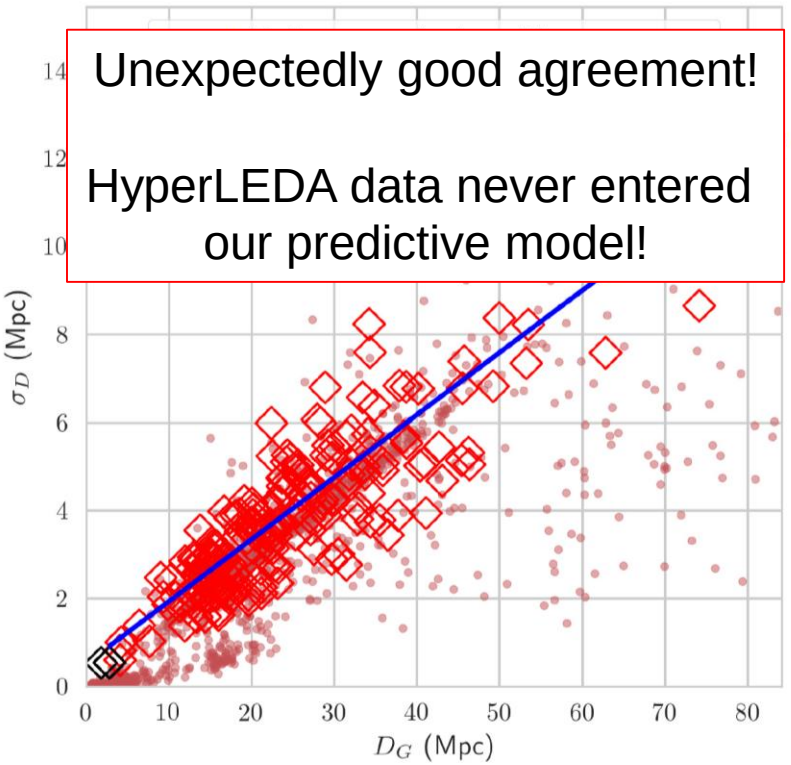
Predictions for 200 galaxies
in HyperLEDA

Chaparro Molano+2019 – astro-ph:1805.02578

Prediction of Missing Errors



Predictions for 1k galaxies
in NED-D



Unexpectedly good agreement!

HyperLEDA data never entered our predictive model!

Predictions for 200 galaxies
in HyperLEDA

Chaparro Molano+2019 – astro-ph:1805.02578

Summary

- We report error data tables using the robust (H, M) and frequentist (P, Q) methods for NED-D, HyperLEDA, and Cosmicflows-3, along with the 16th, 50th, and 84th percentiles of the bootstrap sampled distance distribution for each galaxy.

- These error tables should be a fundamental tool for future precision cosmology, catalog-wide studies.

- We create a Bayesian predictive model for TFR distance errors in the NED-D catalogue and we use it to predict missing errors for 884 galaxies.

- This predictive model was independently validated against HyperLEDA for 203 galaxies with missing TFR distance errors.

- We want to advocate for the use of discrepancy measures for Bayesian model checking, as it checks the model's ability to reproduce observed data scatter.

Error data tables, code, posterior samples at:

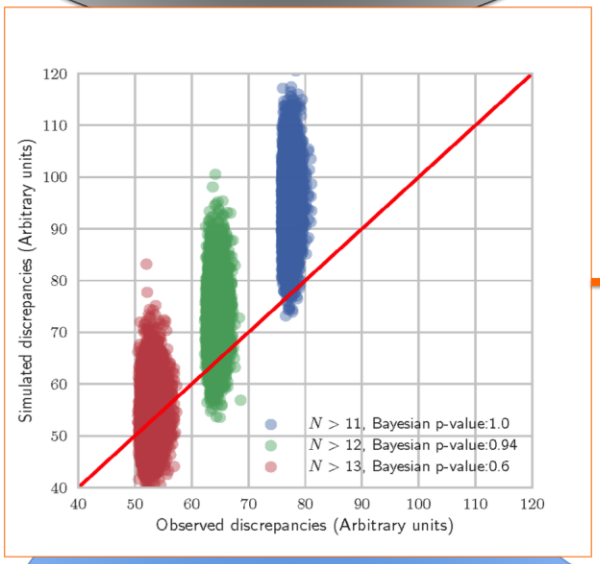
https://github.com/saint-germain/errorprediction

Riess et al. (2019) report Ho = 74 km/s/Mpc (distance ladder method). Planck (2018) reports Ho = 67 (cosmological).

Difference ok 5.8 Km/s/Mpc is (> ~ 3-sigma)

NED-D TF H/M distance errors for galaxies with $N_{meas} > n$

Posterior predictive check using a discrepancy measure and a Bayesian ·p· value (Gelman et al 1996, Brooks et al 2000)



Validation

Bayesian predictive model (MCMC) dependent on *n* for H/M errors

Synthetic H/M distance errors for the same galaxies