Characterizing selection effects and Data systematics in Stage IV spectroscopic surveys

Antonio Farina
Edoardo Maragliano

Sestri

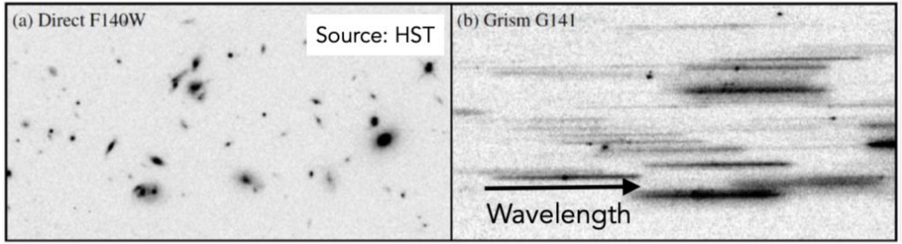
2024

Slitless spectroscopy

- Space observations make impossible using slits
- **Slitless spectroscopy:** each source is spread into its spectrum directly on the focal plane
- **Pro**: blind search for emission-line galaxies, no imaging needed
- **Cons:** challenging characterization of systematic effects. We will discuss in particular:
  - Characterization of the selection function (Antonio)
  - Impact of interlopers on post-recon BAO analysis (Edoardo)

Euclid

Roman



Credit J. Bautista, I. Tutusaus

Random catalogs for early spectroscopic data

Antonio Farina, Ben Granett, Enzo Branchini, Alfonso Veropalumbo

Understanding the Galaxy/Matter Connection in the Era of Large Surveys

Sestri Levante - 17/09/2024

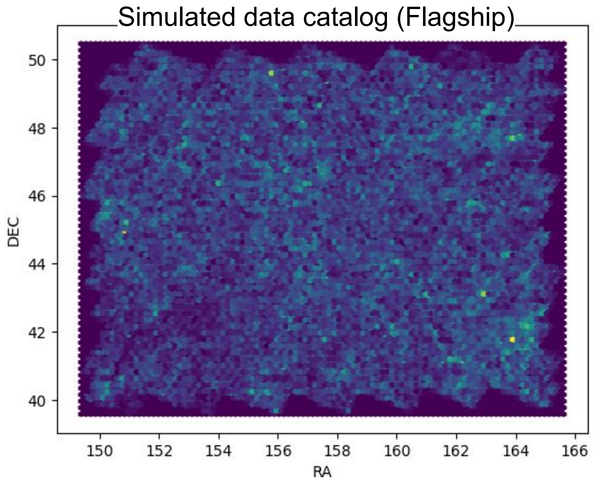
Selection function and random catalog

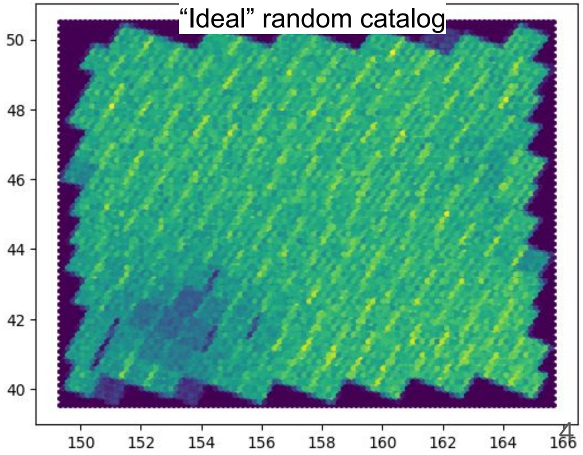
- To effectively utilize spectroscopic cosmological data we need to characterize their **selection function.**

- Characterizing the selection function in presence of a slitless spectrograph is not an easy task**:**
  - Depends on **RA, Dec, redshift, line flux, angular size, …**

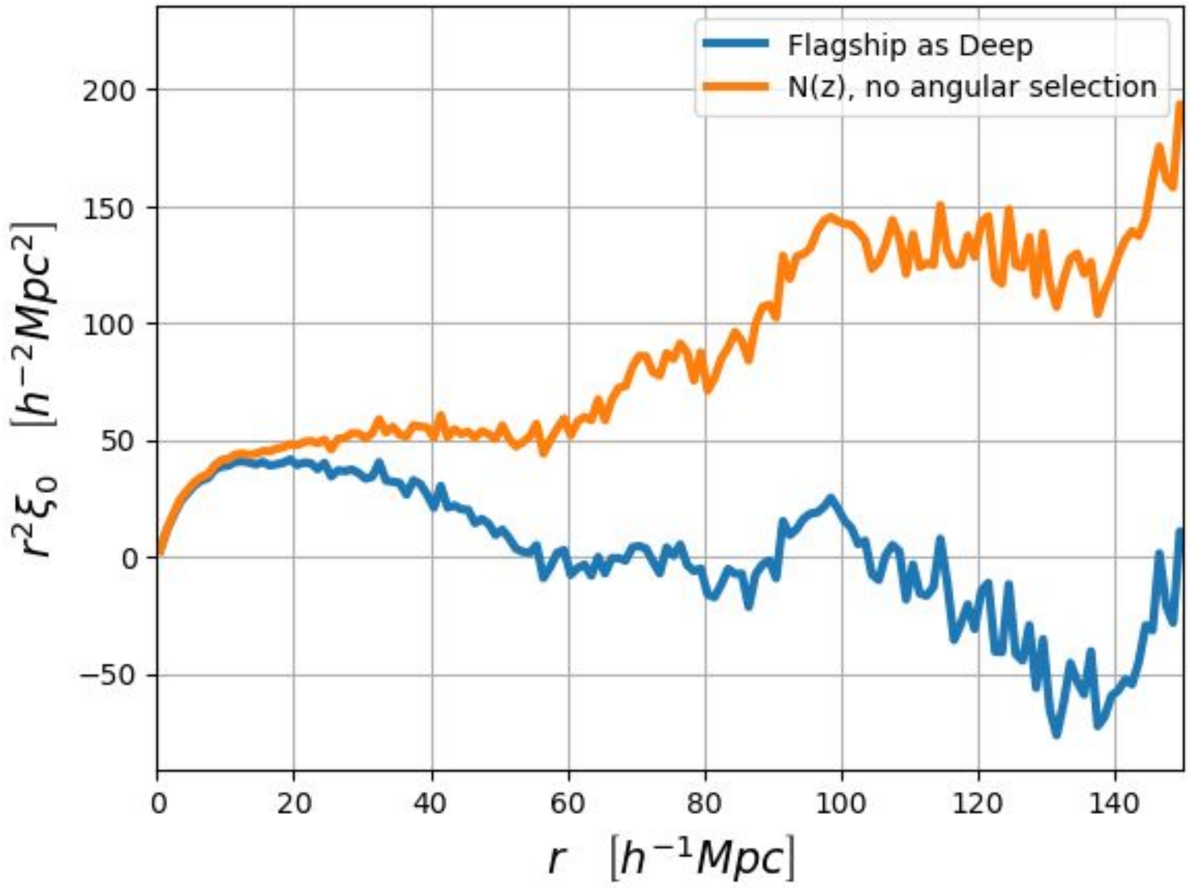
- We characterize the selection function through a **random catalog**:

  - **Traces the mean density** of the spectroscopic sample;
  - Poisson sampling (no clustering signal);
  - Much denser than data sample.





Not taking properly into account the selection effects leads to measurements reflecting the properties of the selection function rather than those of the underlying galaxy distribution



Forward modelling approach

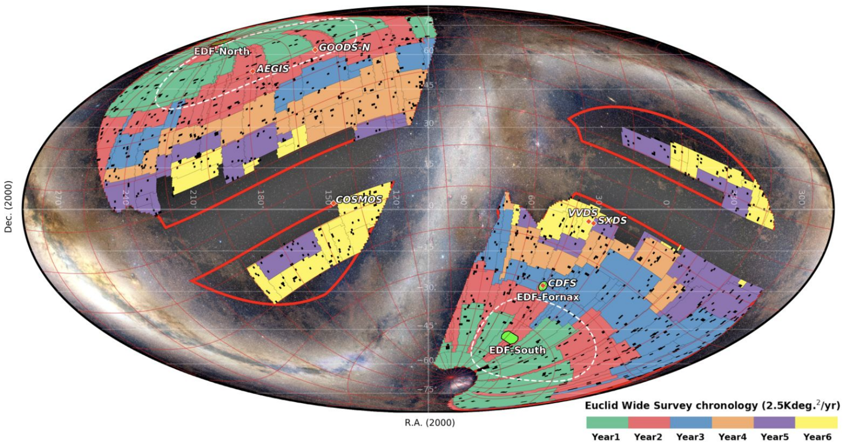
- **Synthetic galaxies** (randoms) are placed over the Wide Survey area and assigned **redshift and properties** sampled from an highly pure and complete reference catalog.

- The **noise at the location of the emission lines** is read from calibrated frames to compute **SNR for the randoms.**

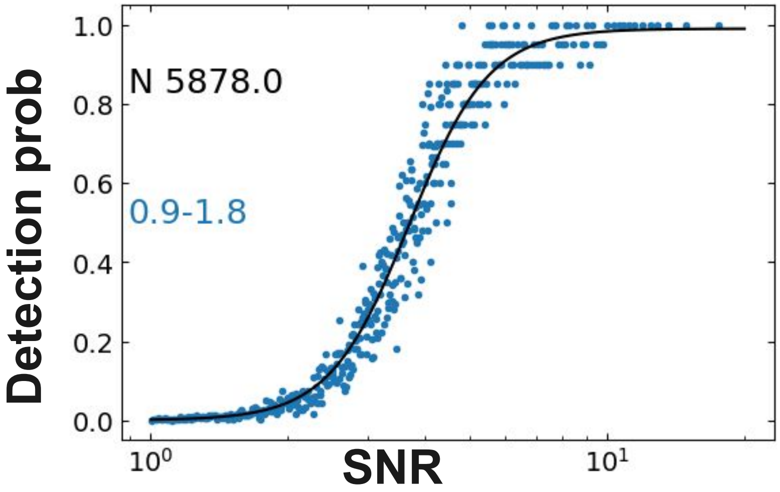
- A **bypass survey simulator** takes redshift, properties and noise as an input and gives back an estimate for the **SNR.**

- The **detection model maps the SNR to detection probability.**

- The random catalog is **downsampled or weighted by the detection probability**.



Scaramella et al. 2022



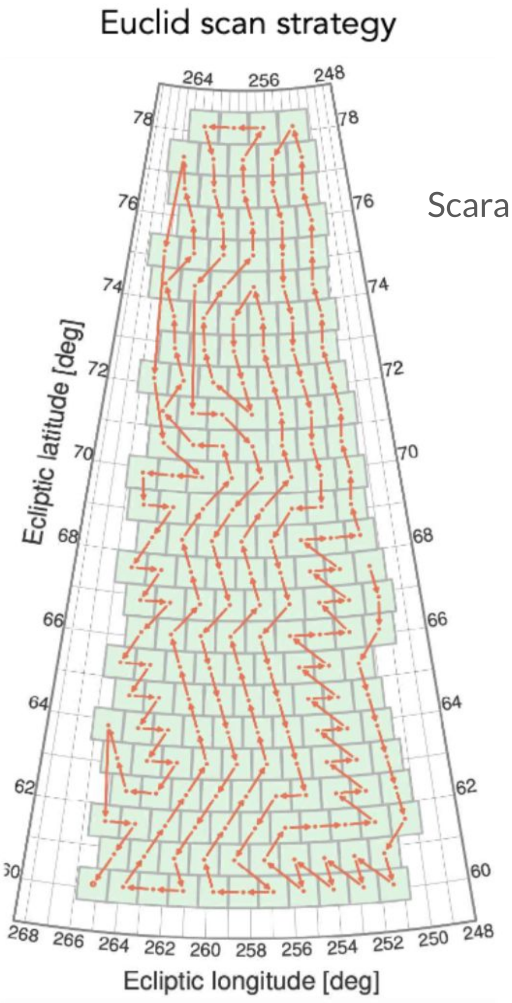
Absence of an external calibrator: a timing issue

- Having a telescope orbiting in L2 poses a lot of technical issues

- The design of the survey strategy is a complex task, crucial for the successful outcome of the mission

- It may happen (e.g. in the Euclid case), that a substantial fraction of the data is collected before having a reliable calibrator (at least not at full depth)

What can we do in the meantime?



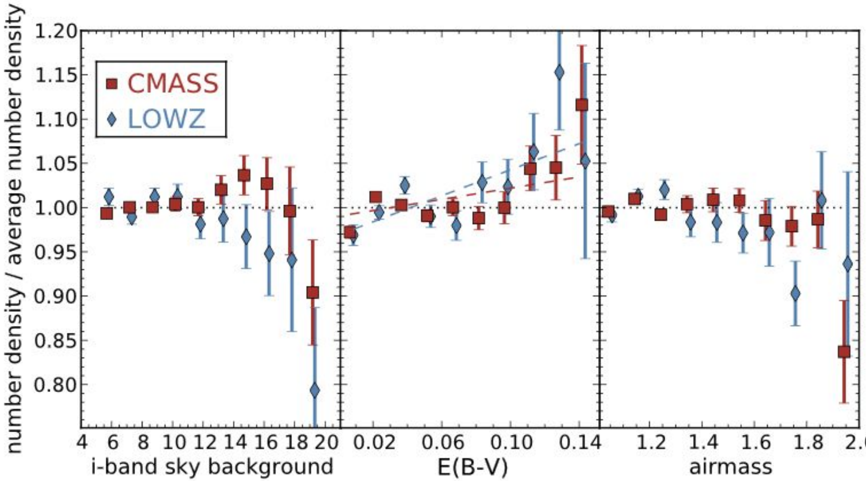
Scaramella et al. 2022

Calibration on Wide Data: regression

- The idea is correcting for systematic variations in number density by regressing the measured galaxy number density against each potential source of systematic.

- The resulting trends can then be imprinted into the random using weights to null-out modes that are thought to be contaminated

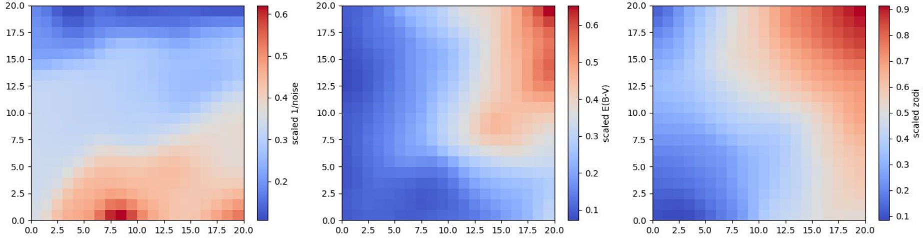
- This strategy is well consolidated as it’s been adopted by several past surveys, e.g. BOSS

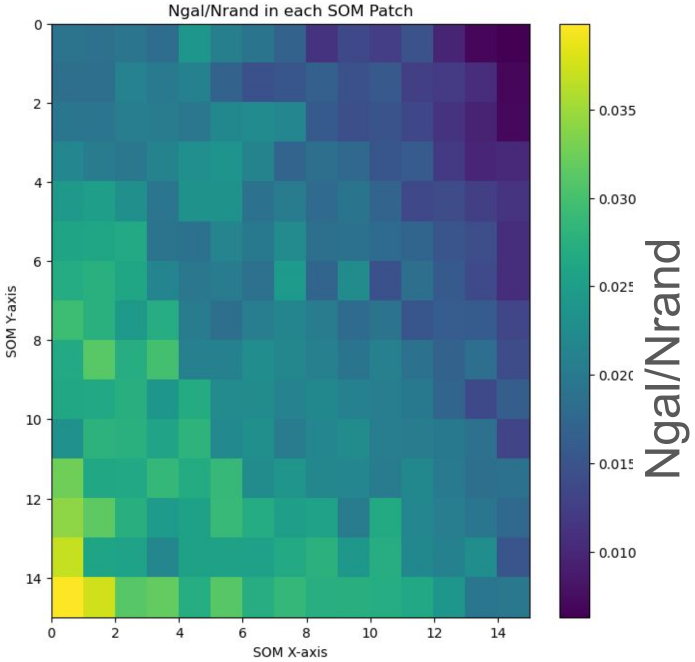


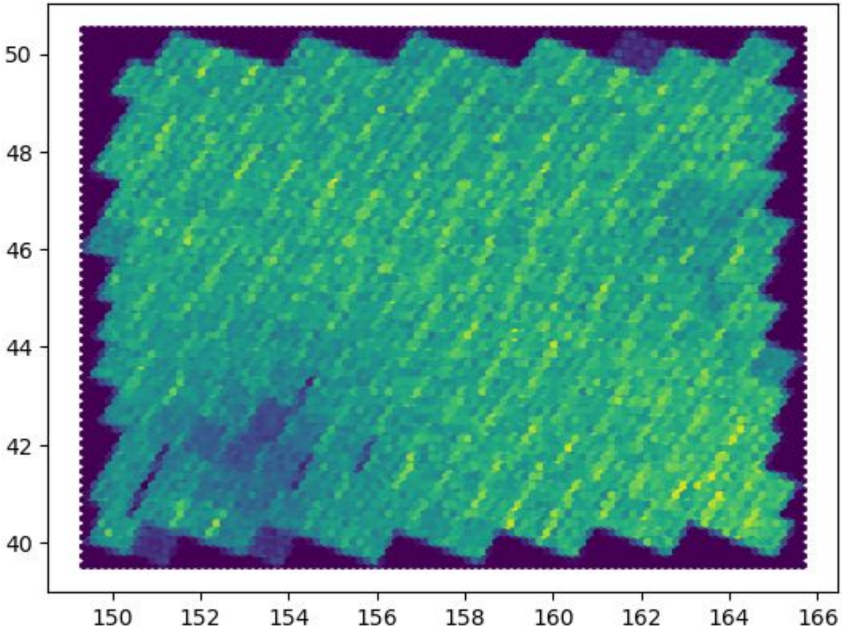
Ross et al. 2017

Regression with Self-Organizing Maps (SOM)

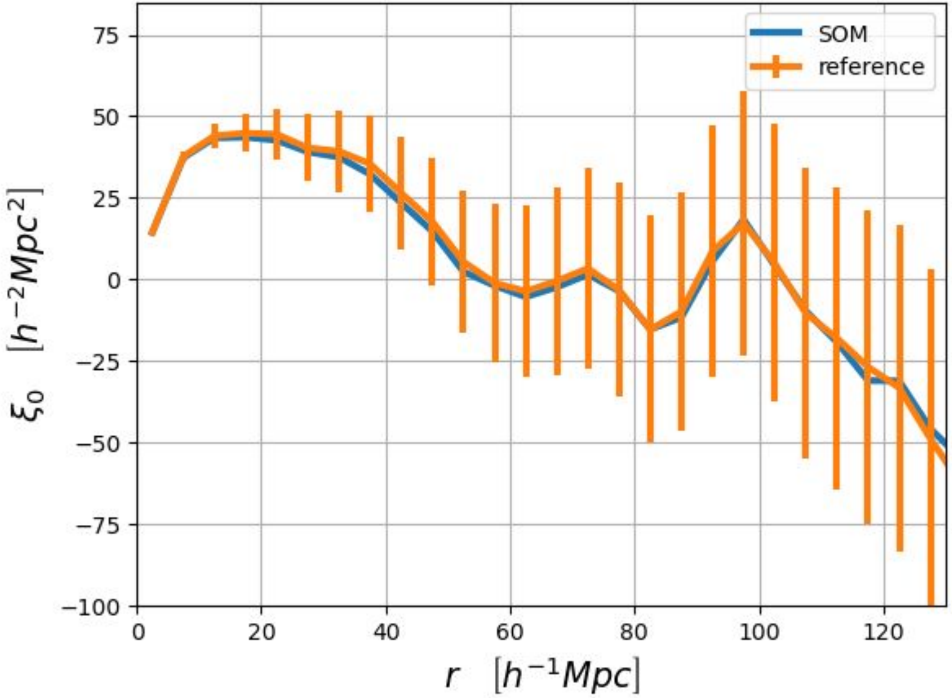
- SOM are unsupervised machine learning algorithms for data compression. They preserve the topological structure of the data
- Construct hyperparameter space of potential selection dependencies
- Get a map from N-dimensions to 2D hidden space training the SOM on data.
- Use the map to project a uniform random into this 2D grid.
- Compute weight Ngal/Nrand in each grid cell and downsample the random accordingly.





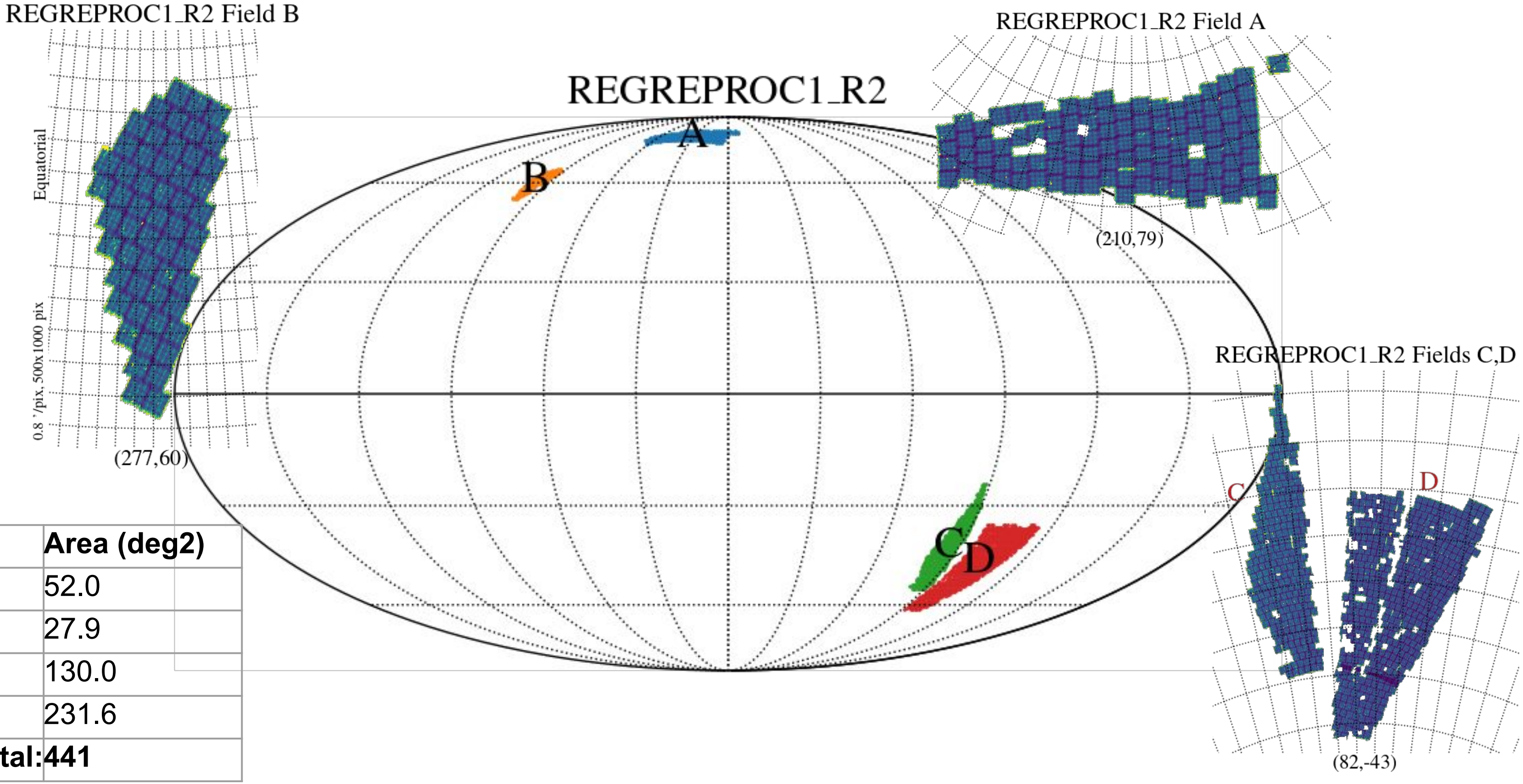


50
48
46
44
42
40
150
152
154
156
158
160
162
164
166



SOM
reference
75
50
25
0
−25
−50
−75
−100
0
20
40
60
80
100
120
$\xi_0 \quad [h^{-2}Mpc^2]$
$r \quad [h^{-1}Mpc]$

Euclid data



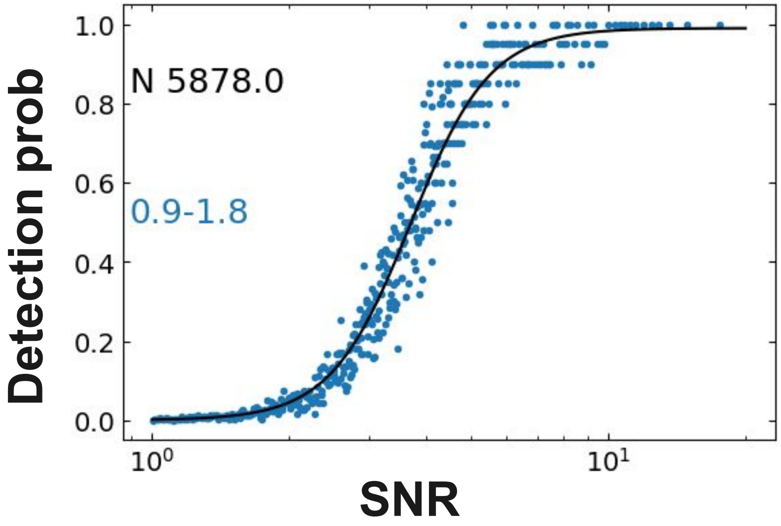
Field	Area (deg2)
A	52.0
B	27.9
C	130.0
D	231.6
Total:	**441**

Calibration of the detection model

$$n(\sigma) = \int ds\, \phi(s)\, C(s/\sigma)$$

- Wide survey **density-noise** relation: measured from data
- $\boldsymbol{\phi}$**(s)** is the distribution function for the signal

We want to find the sigmoid parameters that best reproduce n(σ)



$$C(x) = \frac{c}{1 + \left(\frac{x}{b}\right)^a}.$$

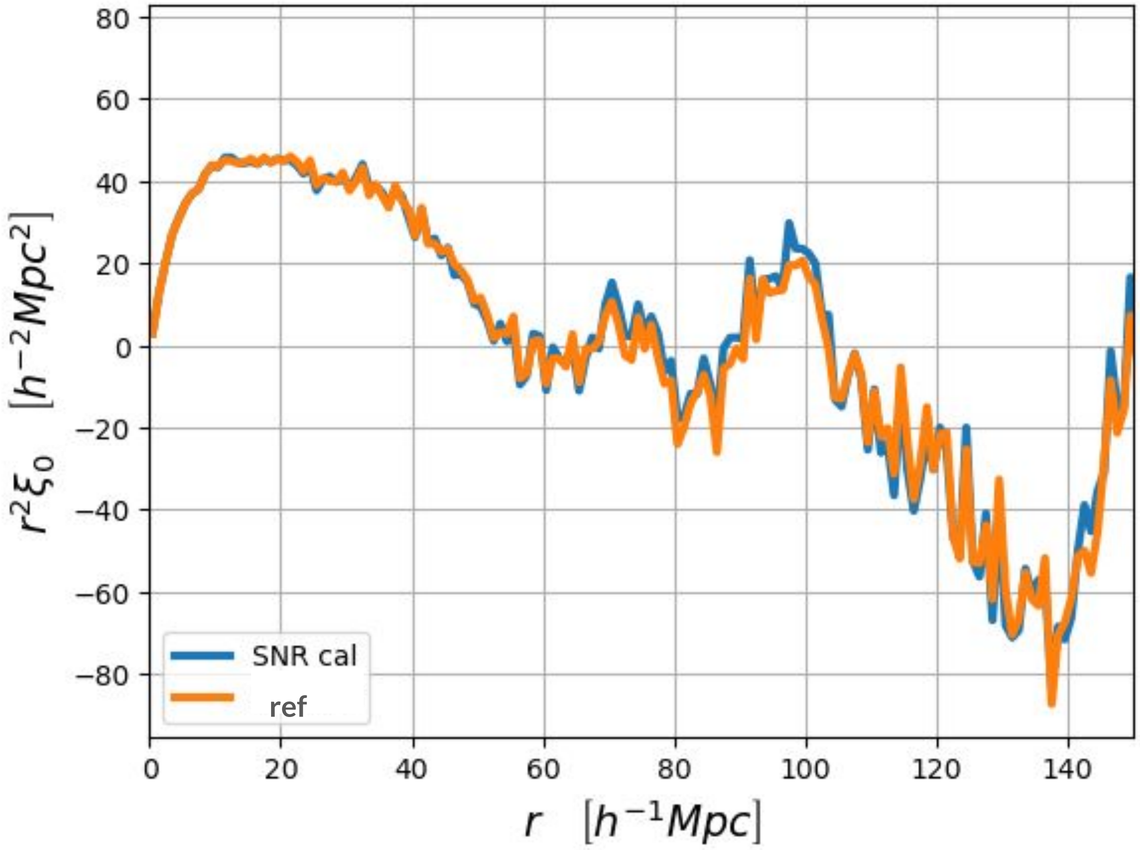
Derived completeness function sigmoid parameters

⚠ $\boldsymbol{\phi}$**(s)** needs to be assumed. We can take it from models or simulations

⚠ **Interloper rates cannot be estimated in this way.**

- The procedure has been tested so far only in the ideal case in which $\boldsymbol{\phi}(\mathbf{s})$ is perfectly known

- Robustness wrt variations and biases on $\boldsymbol{\phi}(\mathbf{s})$ still needs to be investigated



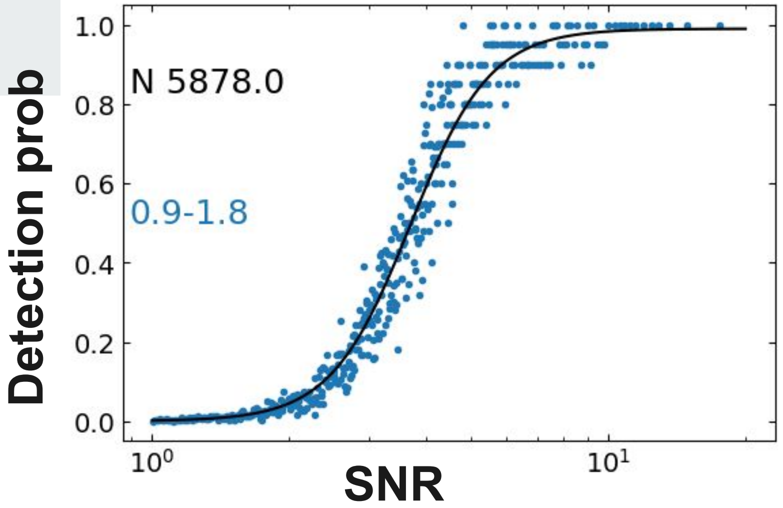
BACKUP SLIDES

Detection model

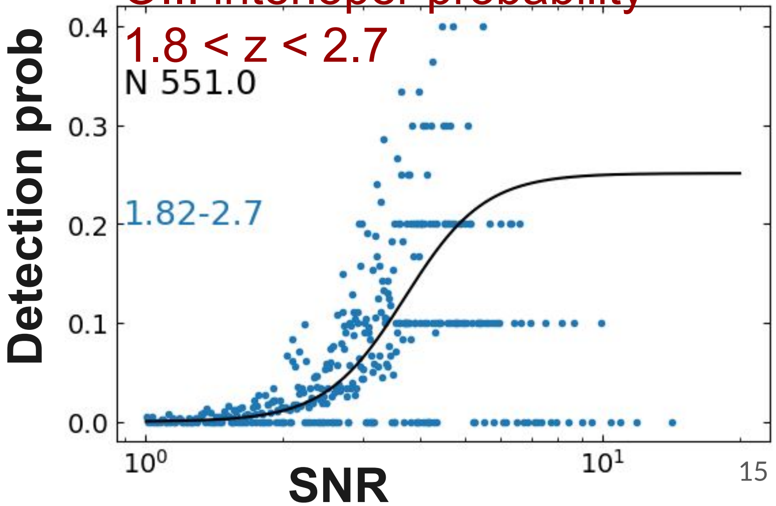
- The detection model maps emission line **SNR to detection probability**.
  Parametrized with **sigmoid function**:

$$C(x) = \frac{c}{1 + \left(\frac{x}{b}\right)^a}.$$

- Simple sigmoid function works on FastSpec+SPE simulations.
- Also model **SNR dependence of interlopers** (line misidentifications, not noise ones).
- **Must be calibrated on simulations and real data run through the pipeline!**



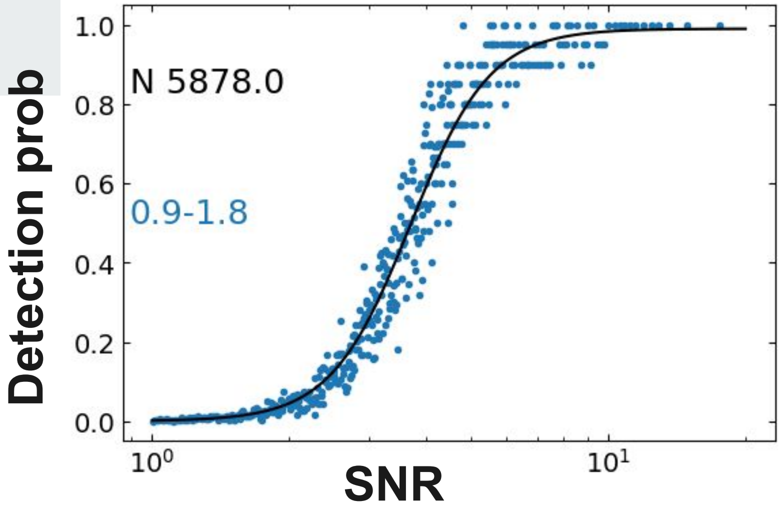
OIII interloper probability

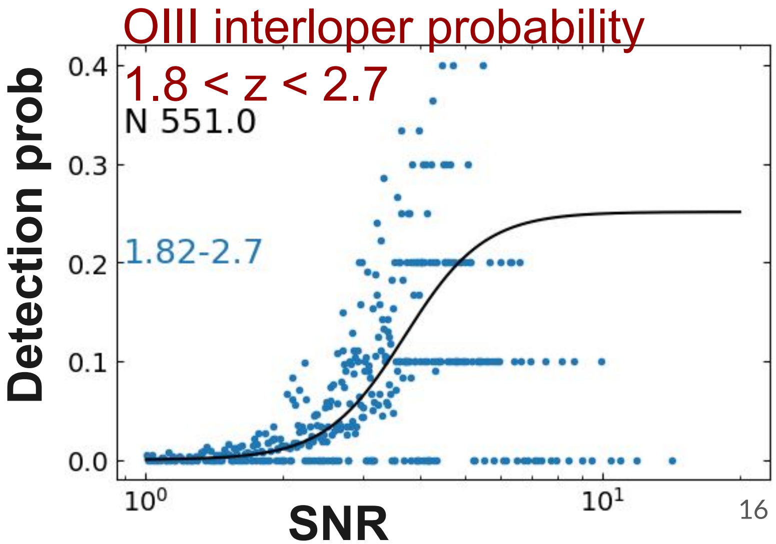


Detection model

- **All selection effects are modeled through SNR:**
  - number of exposures;
  - background noise (zodi, straylight...);
  - overlapping spectra;
  - galactic extinction.

- **Other types of selection systematics must be dealt with in another way** (eg weights).





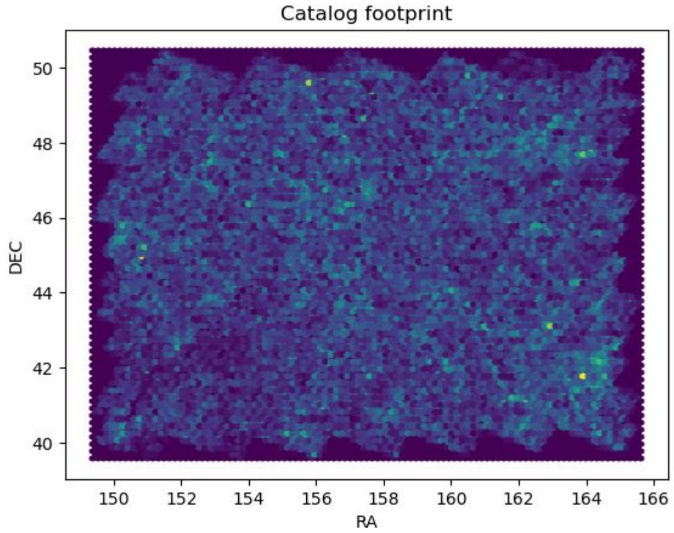
Test on Flagship catalog

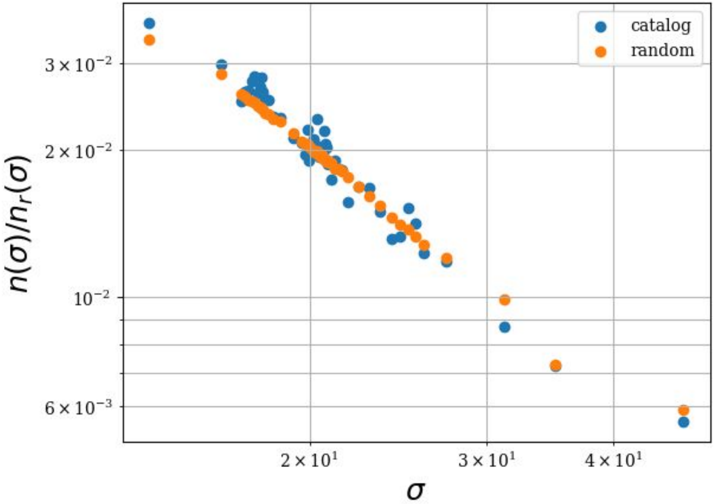
- Observed-like catalog generated with **Pypelid** from the Flagship:
  - we know the underlying $\boldsymbol{\phi}$**(s)**
  - 100 deg^2 with 1 passes

1. We measured n(σ) from the observed-like catalog
2. We measured $\boldsymbol{\phi}$**(s)** from Flagship
3. We fitted sigmoid parameters using

$$n(\sigma) = \int ds\, \phi(s)\, C(s/\sigma)$$

4. We selected random objects according to the selection function obtained in this way





Overview

- BAO stand as one of Euclid's main probes
- Reconstruction techniques are crucial to obtain the most stringent constraints on the Alcock-Paczynski parameters → Expansion history
- Galaxy interlopers introduce systematics in the 2PCF

What is the impact of interlopers galaxies on the reconstructed BAO peak?
Do they introduce significant systematics in the AP parameters?

My work in summary

- Reconstruction on mock catalogues
- Measurement of the 2-point statistics
- Fit of the BAO peak to infer the AP parameters
- Comparison of the pre- and post-rec results

Necessary steps:

- Adapting the reconstruction code to work on lightcones

Interloper galaxies

Galaxies with catastrophic redshift errors.

Two kinds:

- Line interlopers
    - Misidentified genuine emission lines
    - Deterministic shift in redshift
- Noise interlopers
    - Noise misidentified as a genuine line
    - Random shift in redshift

Catalogues

Euclid Large Mocks

- 100 out of the 1000 mocks
- 30 deg lightcone geometry
- Split into 4 redshift bins
- Realistic volumes for DR1
- Realistic contamination

Euclid Large Mocks

	$z1$ 0.9 – 1.1	$z2$ 1.1 – 1.3	$z3$ 1.3 – 1.5	$z4$ 1.5 – 1.8
OIII	0.03	0.12	0.09	0.01
SIII	0.01	0.03	0.08	0.07
noise	0.12	0.08	0.08	0.06

I focus on the first and third redshift bin:

- z1 mainly contaminated by noise interlopers
- z3 features similar fractions of noise, OIII e SIII

From Euclid Collaboration, Risso et al., in preparation

Modelling and inference

Modelling

- Template model as in Ross et al. (2016) BOSS DR12
- Doesn’t account for the presence of interlopers
- I focus on the AP parameters α

$$\alpha_{\parallel} = \frac{H_{fid}(z)}{H_{true}(z)} \qquad \alpha_{\perp} = \frac{d_{true}(z)}{d_{fid}(z)}$$

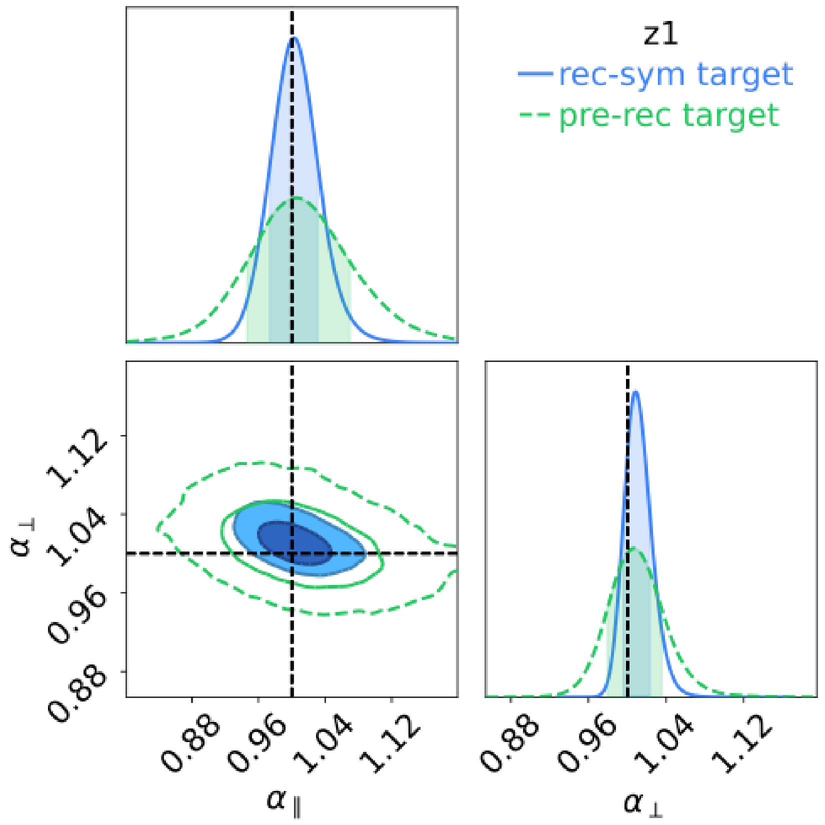
valore atteso = 1

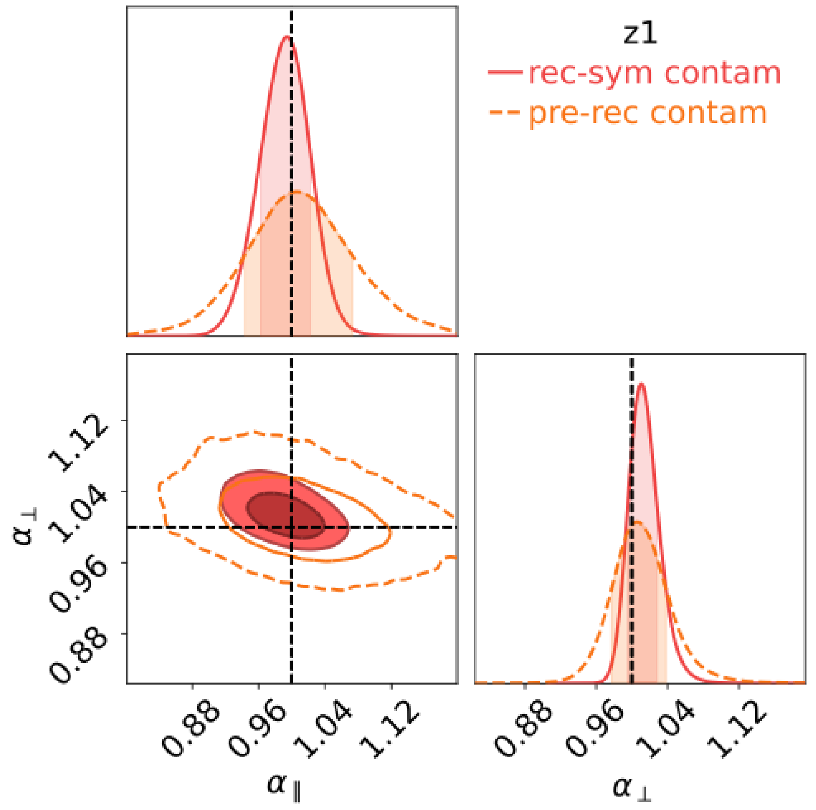
Inference

- Fit of the BAO peak in $s \in [50; 150]\mathrm{h}^{-1}\mathrm{Mpc}$
  → MCMC sampling

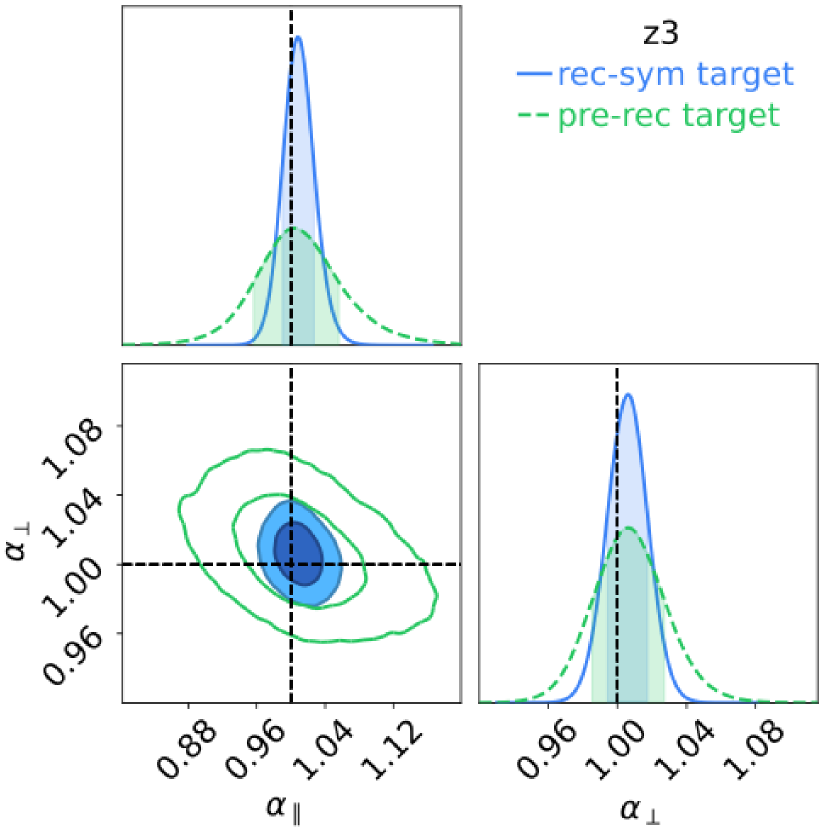
Ross et al. (2016) arXiv:1607.03145 [astro-ph.CO]

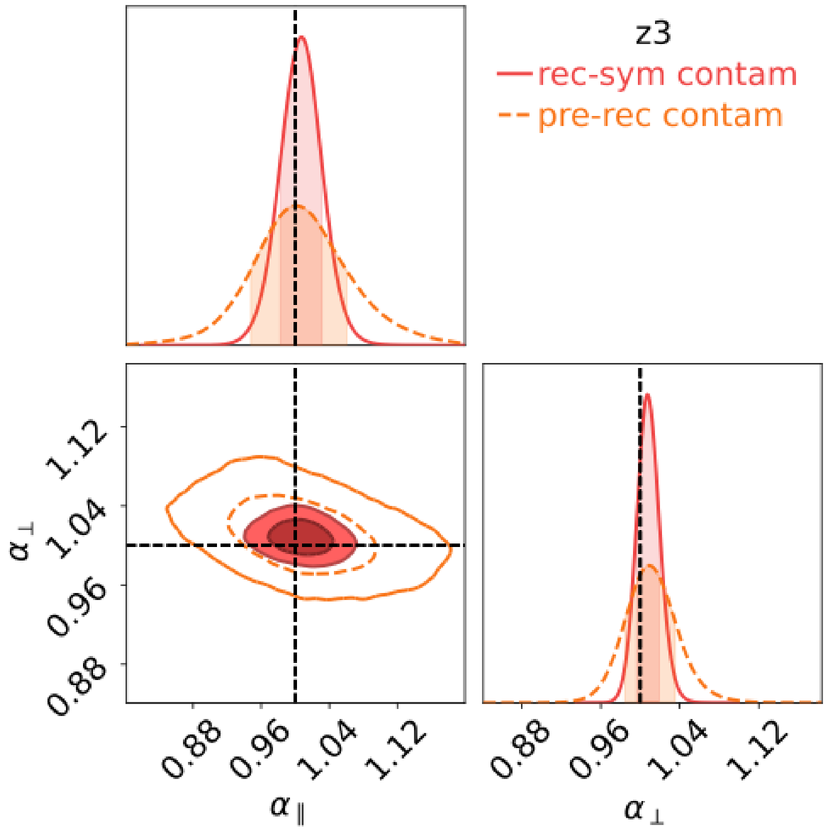
Results: Euclid Large Mocks





Results: Euclid Large Mocks



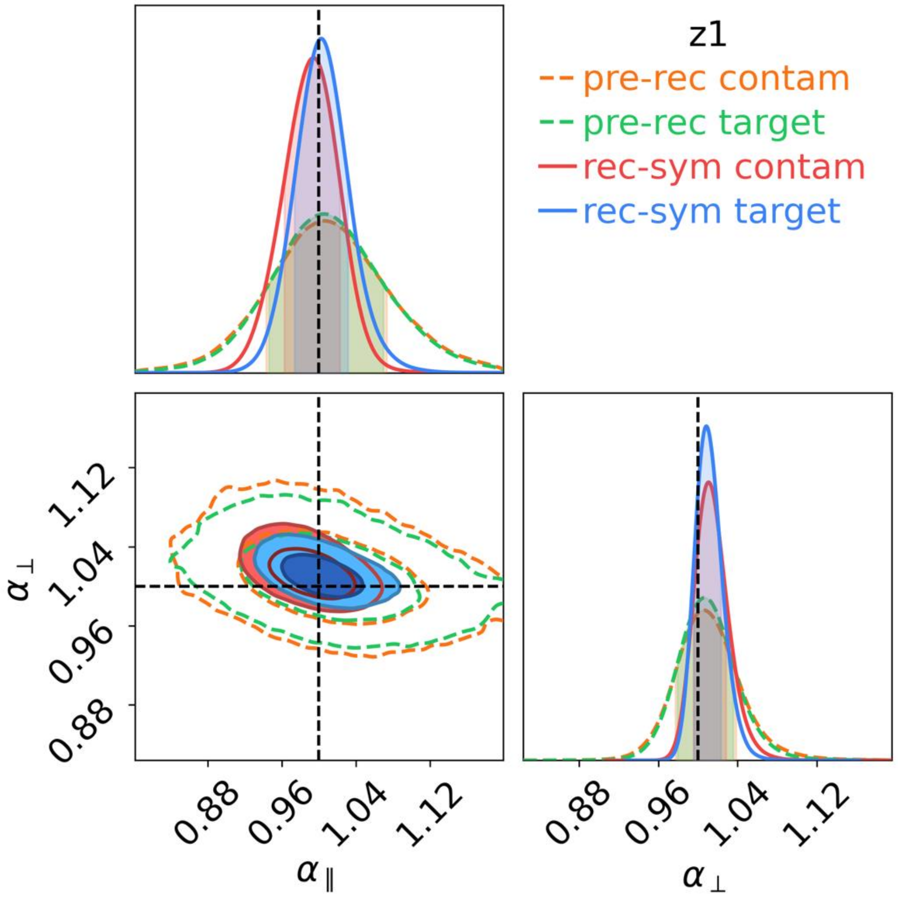


Conclusions

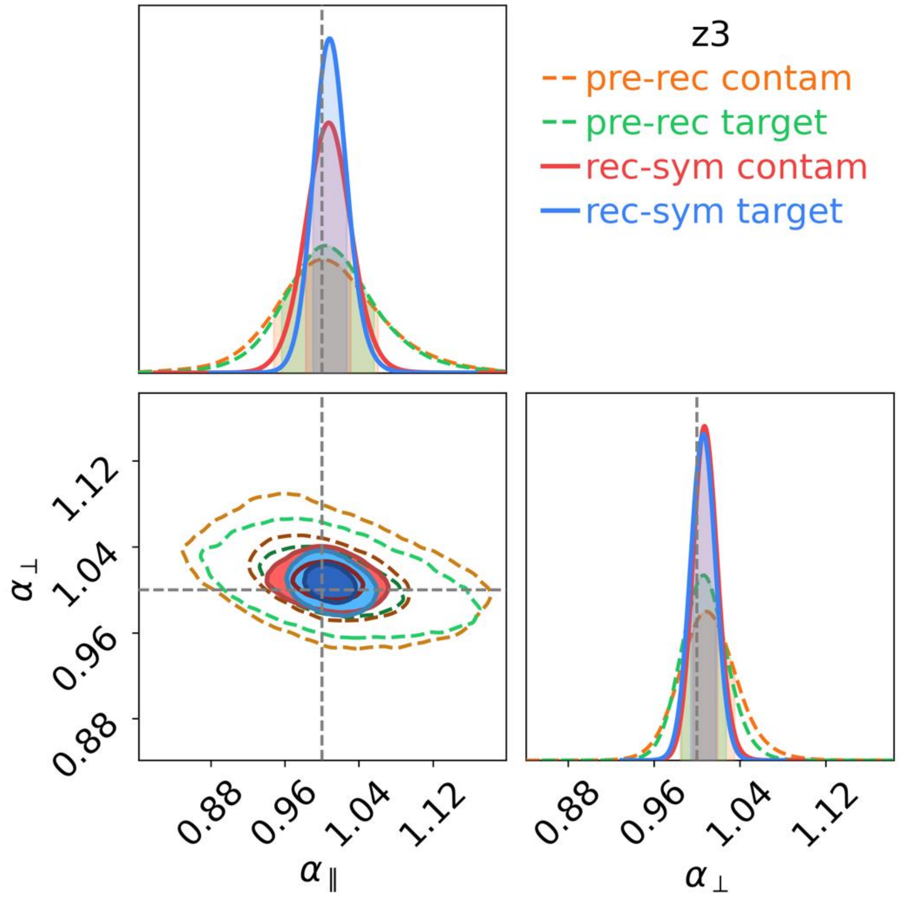
- Reconstruction yields better constraints on AP parameters also on contaminated catalogues
- The improvement in the two cases have the same order of magnitude

- With DR1 statistics:
    - Reconstruction on contaminated catalogue does not introduce significant systematics

Backup



z1
pre-rec contam
pre-rec target
rec-sym contam
rec-sym target
1.12
1.04
0.96
0.88
$\alpha_\perp$
0.88
0.96
1.04
1.12
$\alpha_\parallel$
0.88
0.96
1.04
1.12
$\alpha_\perp$



z3
pre-rec contam
pre-rec target
rec-sym contam
rec-sym target
1.12
1.04
0.96
0.88
$\alpha_\perp$
0.88
0.96
1.04
1.12
$\alpha_\parallel$
0.88
0.96
1.04
1.12
$\alpha_\perp$