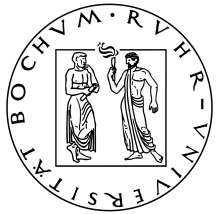
Angular galaxy clustering from deep surveys with complex selection effects

Ziang Yan

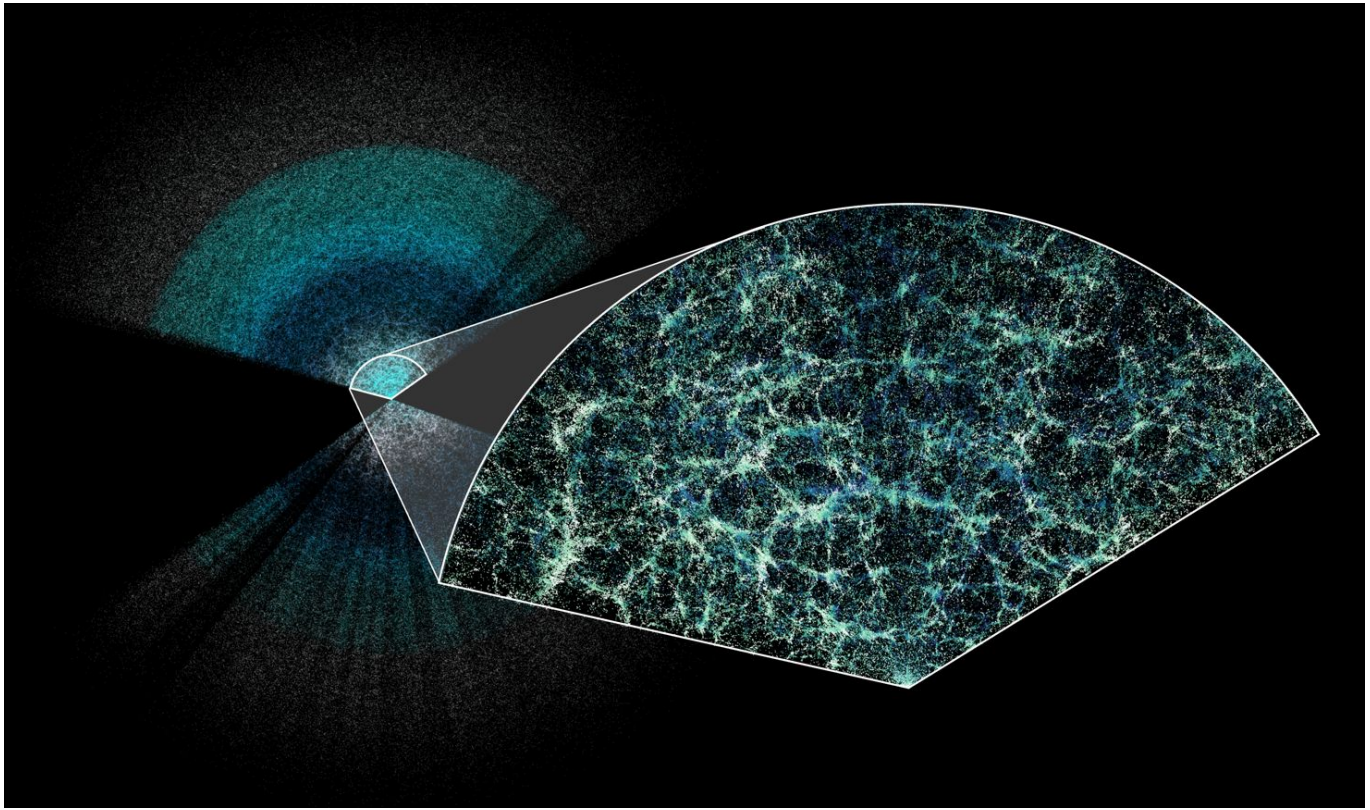
German Centre for Cosmological Lensing
Ruhr University Bochum



GCCL
GERMAN CENTRE FOR COSMOLOGICAL LENSING

Background: galaxy two-point functions

Galaxy distribution as a tracer of the LSS



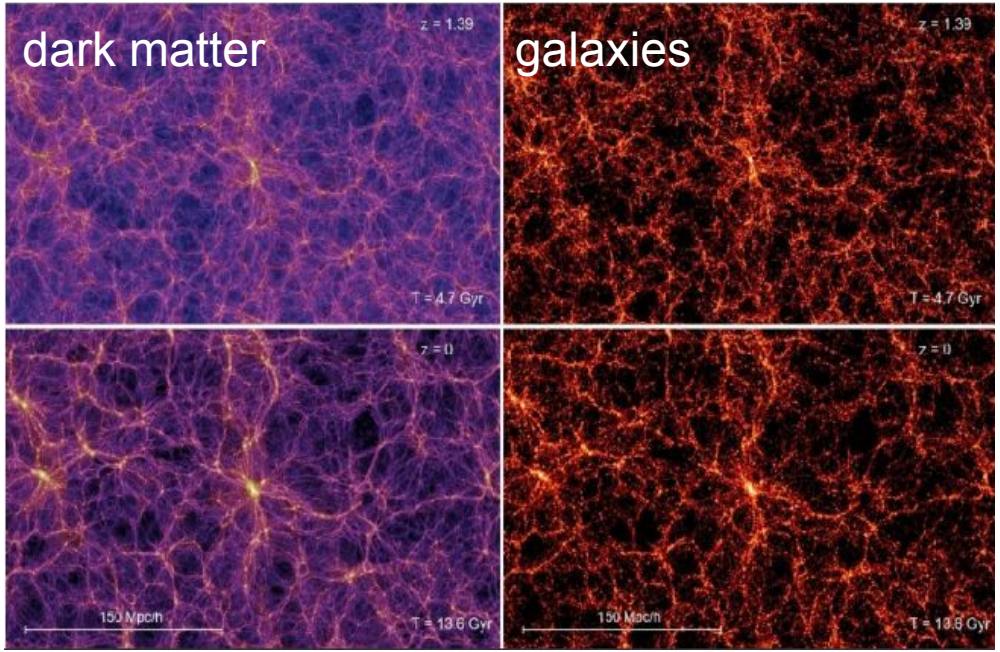
DESI Y1 results:
https://www.desi.lbl.gov/2024/04/04/desi-y1-results-april-4-guide/

Galaxy distribution as a tracer of the LSS

Galaxy distributions follow the distribution of matter.

- on large scales: galaxy overdensity is **linearly biased** matter overdensity
- on small scales: connection between galaxy number and dark matter halo mass is described by **halo occupation distribution**

Galaxy distributions can tell us the evolution of the LSS of the Universe



Dark matter and galaxy distribution from Millenium simulation. Image from https://www.mpi-hd.mpg.de/lin/events/isapp2011/pages/lectures/Springel-1_2.pdf

Galaxy two point correlation function (2PCF)

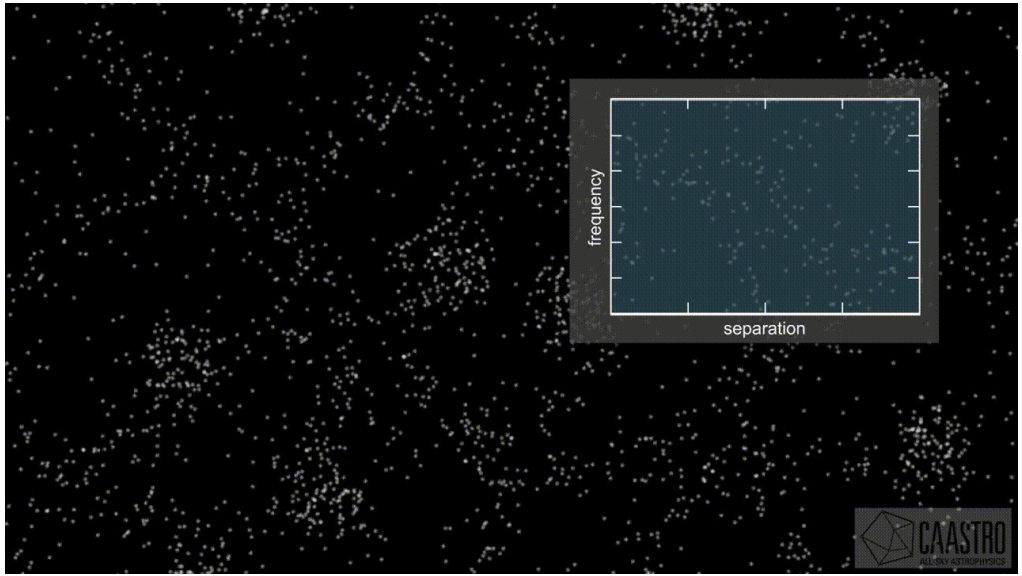
2PCF of general LSS fluctuations:

$$\xi(r) \equiv \left\langle \delta(\boldsymbol{r'} - \boldsymbol{r})\delta(\boldsymbol{r'}) \right\rangle_{\boldsymbol{r'}}$$

Galaxy 2PCF describes the possibility of finding galaxy pairs with certain separations.

$$1 + \xi(r) = \frac{N^{\mathrm{g}}_{\mathrm{pair}}(r)}{N^{\mathrm{r}}_{\mathrm{pair}}(r)}$$

It shows us how galaxies are clustered.



https://en.wikipedia.org/wiki/File:Two-point-correlation-function-astronomy.webm

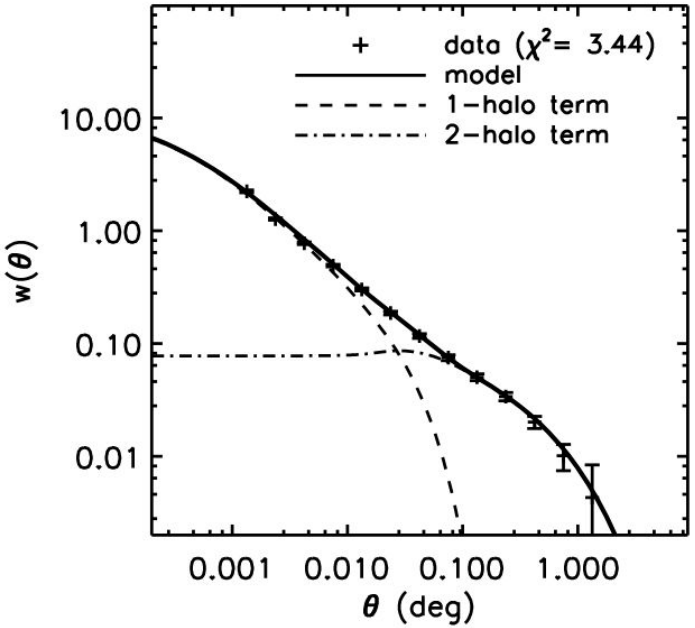
Angular two-point Correlation function

If the distances (redshifts) of galaxies are not available (galaxies from photometric surveys, for example), one can measure the *angular* 2PCF defined similarly as:

$$w(\theta) \equiv \left\langle \delta(\boldsymbol{\theta'} - \boldsymbol{\theta})\delta(\boldsymbol{\theta'}) \right\rangle_{\boldsymbol{\theta'}}$$

- Theoretical model (Limber+linear bias approximations, halo model, HOD model, etc)
- Information gained from 2PCF: **cosmological parameters, structure formation, galaxy formation, etc**
- Refraining constraints by combining it with other statistics **(nx2pt)**

$$w^{ij}(\vartheta) = 2\pi\, b_{\rm g}^i b_{\rm g}^j \times \int_0^\infty \frac{{\rm d}\ell}{\ell} J_0(\ell\vartheta) \int_0^{\chi_{\rm h}} {\rm d}\chi \frac{p^i(\chi) p^j(\chi)}{\chi^2} P_\delta\left(\frac{\ell + 1/2}{\chi}, \chi\right)$$



(Coupon et al.2012)

Measuring 2PCF: the Landy-Szalay estimator

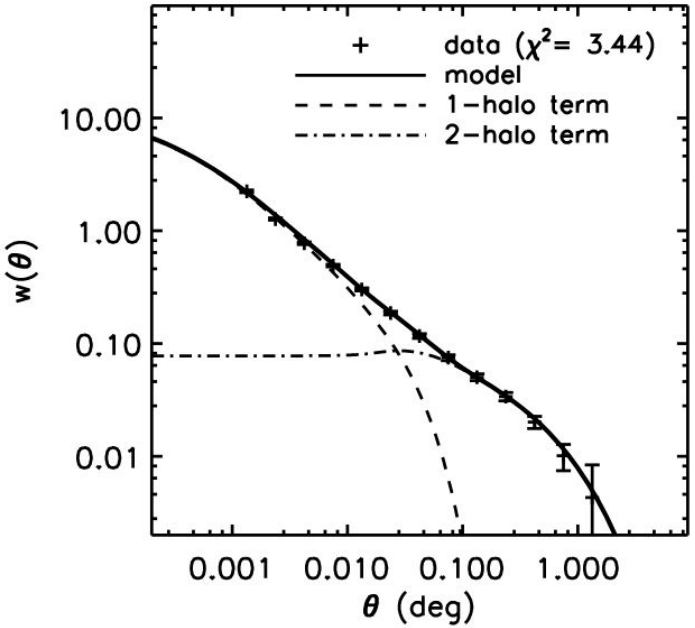
- Practical measurement: Landy-Szalay estimator

$$\hat{w}^{ij}(\vartheta) = \frac{\Theta(\vartheta)\left\{D_iD_j - D_iR_j - R_iD_j + R_iR_j\right\}}{\Theta(\vartheta)\left\{R_iR_j\right\}}$$

Pair counts from a **random** catalogue

Usually a **uniform random (UR)** catalogue within the observation footprint with the same redshift distribution.

UR reflects the random, un-clustered galaxy distribution.



(Coupon et al.2012)

Variable depth in 2PCF measurements: using KiDS-Legacy catalogue as an example

The KiDS-Legacy catalogue

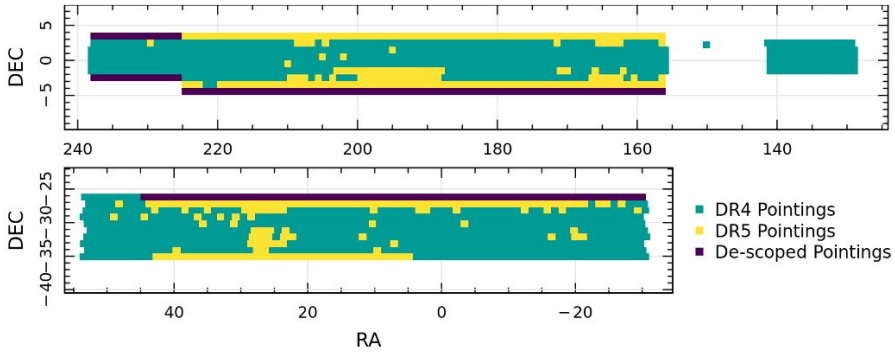
Constructed from the final data release (DR5) of Kilo-degree Survey (KiDS);

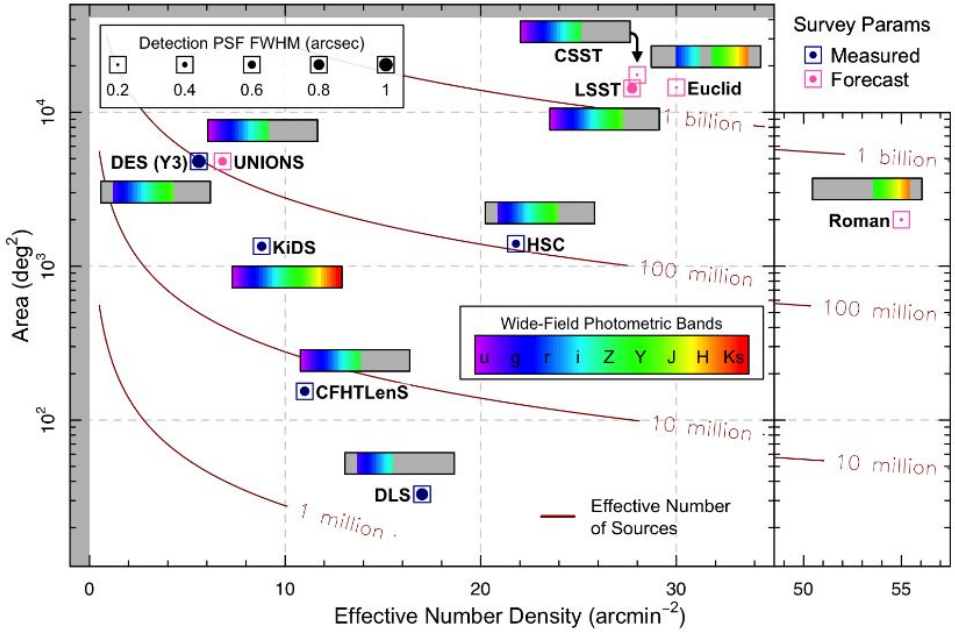
Sky coverage: ~1350 deg^2 (1014 deg^2 after masking);

Number density: ~10 arcmin^-2;

Bands: u, g, r, i1, i2, Z, Y, J, H, Ks;

Science: cosmic shear, 3x2pt, etc;



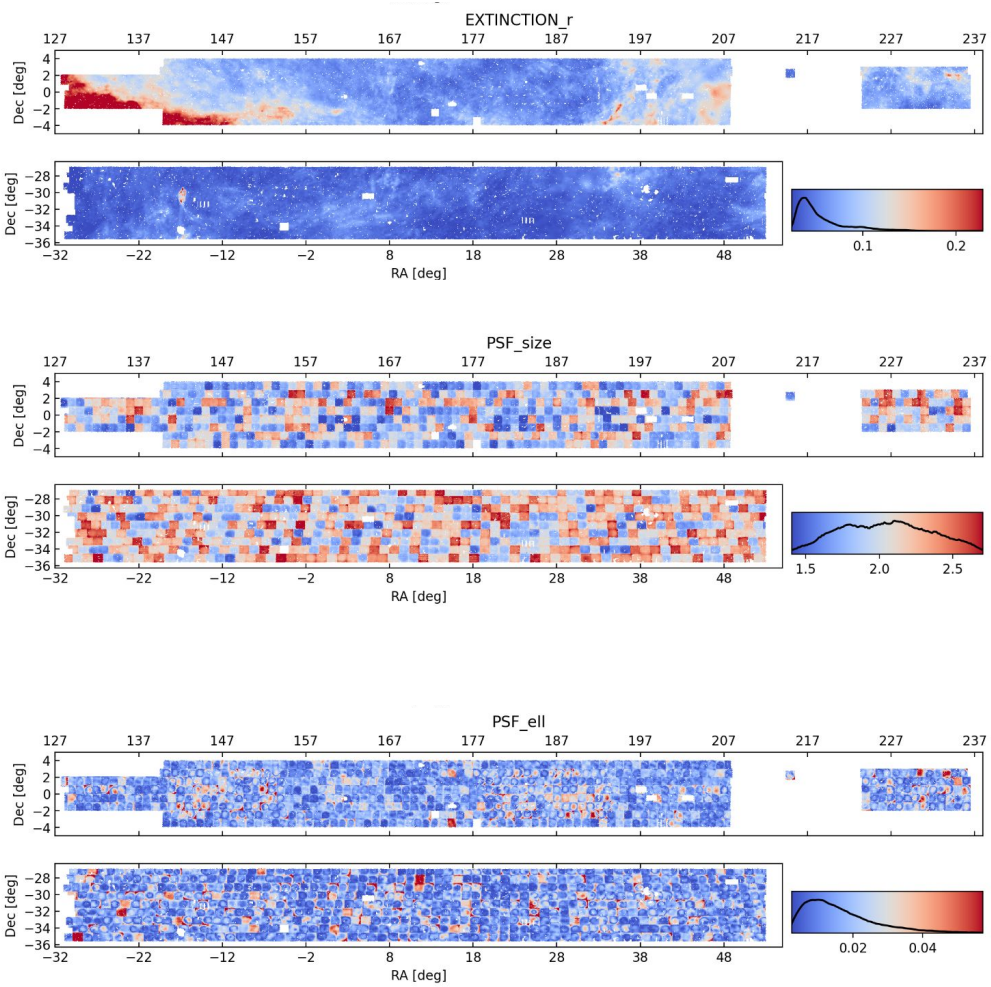


(Wright et al., in press)

Variable depth

- Observational conditions (systematics): seeing, or Galactic extinction, or Milky Way stars, or ... **can cause systematic failure to detect galaxies**

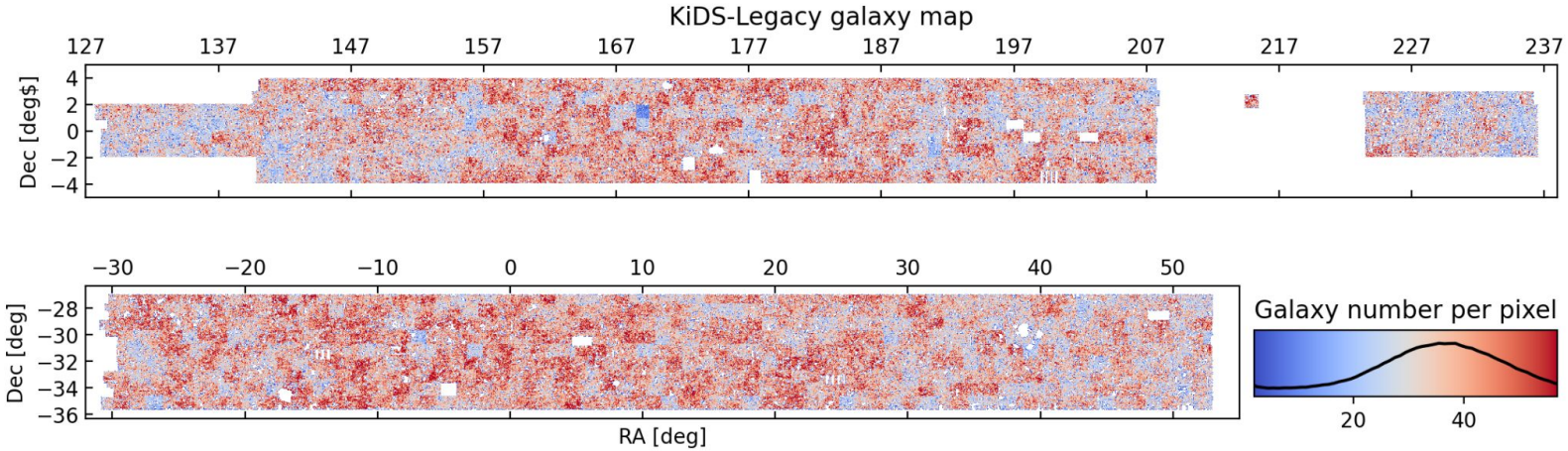
- **KiDS observation strategy causes anisotropic systematics**: the footprint is divided into 1 deg x 1 deg **tiles**. Each tile is observed **only once**.



Variable depth

Inter-tile (e.g. PSF size), intra-tile (e.g. PSF shape) and tile-independent (e.g. extinction) **anisotropic systematics** select galaxies and will induce **“variable depth”** of the galaxy distribution

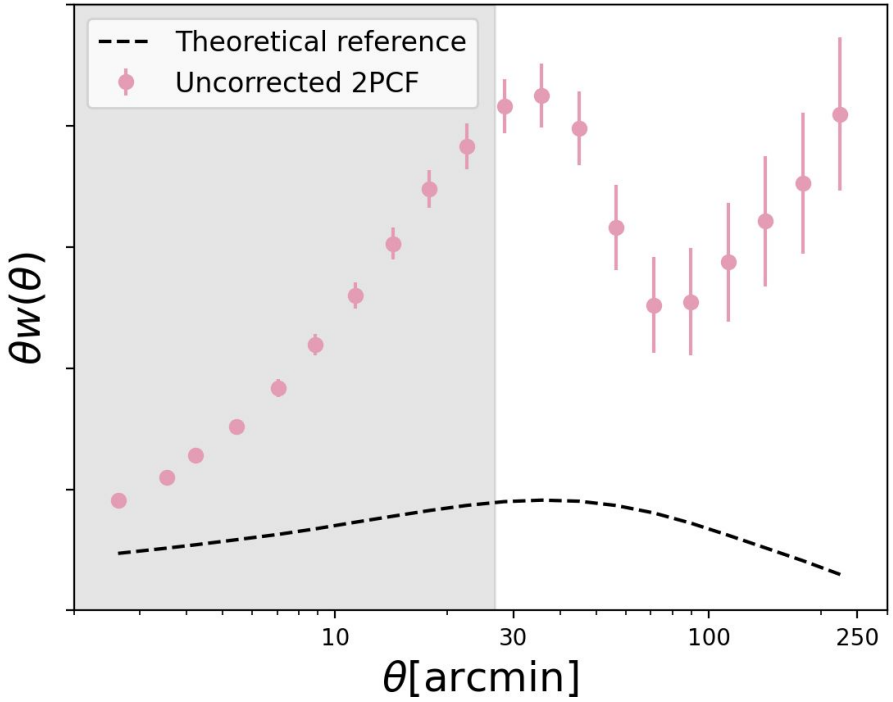
$$\tilde{N}(\boldsymbol{\theta}) = P_{\rm keep}(\boldsymbol{\theta}, \{s^a(\boldsymbol{\theta})\}) \times N(\boldsymbol{\theta})$$



Biased 2PCF due to variable depth

- Variable depth can cause **bias** in galaxy 2PCF measurement.

- We must attempt to mitigate these biases to prevent contamination in our cosmological inference.



2PCF measured from the KiDS-Legacy data; black dashed line is a theoretical reference with KiDS n(z) and reasonable cosmology

Methods to mitigate variable depth

- template-based method: this method parametrize the selection function of systematics and subtract the contaminations (Ho et al. 2012; Ross et al. 2011) or marginalise them out (Leistedt et al. 2013; Leistedt & Peiris 2014; Nicola et al. 2020; Berlfein et al. 2024)
- forward modeling: mock sources are injected into real images after passing through realistic measurement pipeline, resulting in a random catalogue mimicking realistic selections (Bergé et al. 2012; Suchyta et al. 2016)
- **Machine-learning based methods:** aims at probing the relationship between observed galaxy number densities and multivariate systematics vectors, thus recovering the selection functions (Morrison & Hildebrandt 2015; Rezaie et al. 2020)

Method: organized randoms recovered by SOM+HC

Construct “organised randoms” to correct variable depth

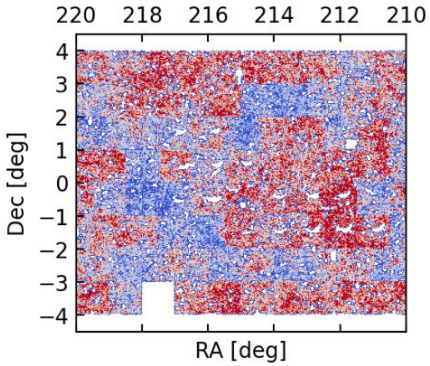
- Landy-Szalay estimator

$$\hat{w}^{ij}(\vartheta) = \frac{\Theta(\vartheta)\left\{D_iD_j - D_iR_j - R_iD_j + R_iR_j\right\}}{\Theta(\vartheta)\left\{R_iR_j\right\}}$$

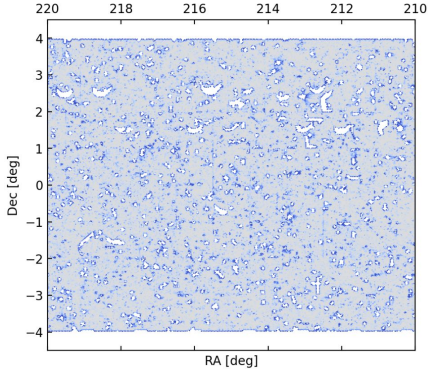
Pair counts from a **organised random (OR)** catalogue instead of a uniform random (UR)

The organised random should reflect the variable depth.

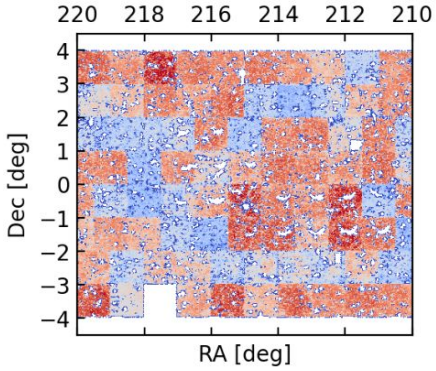
galaxy number per pixel



uniform random per pixel

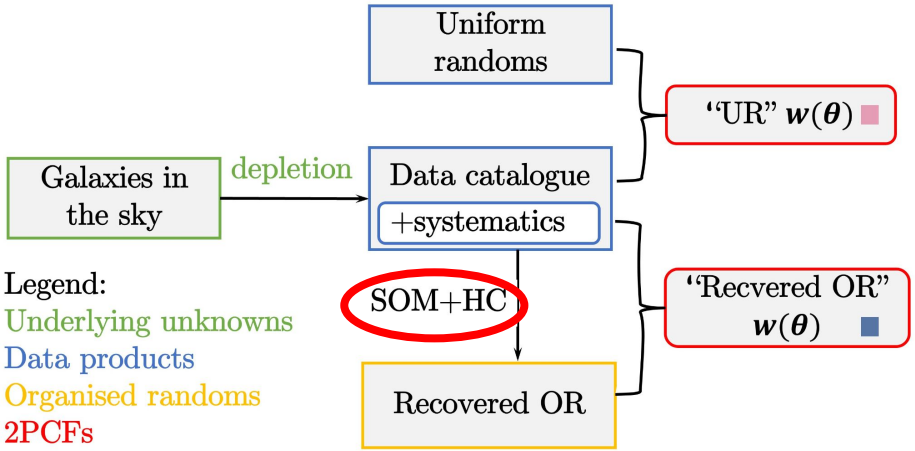


organised random per pixel



Construct “organised randoms” to correct variable depth

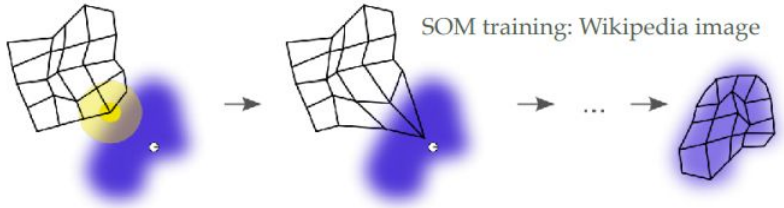
- Organised randoms: a random catalogue that reflects the variable depth in the galaxy catalogue;
- Basic idea: try to find the **LSS-free clustering** of galaxies in the systematics space, then re-distribute galaxies randomly on sky patches of each cluster
- method: self-organising map + hierarchical clustering (Johnston et al. 2021, Yan et al. in prep)

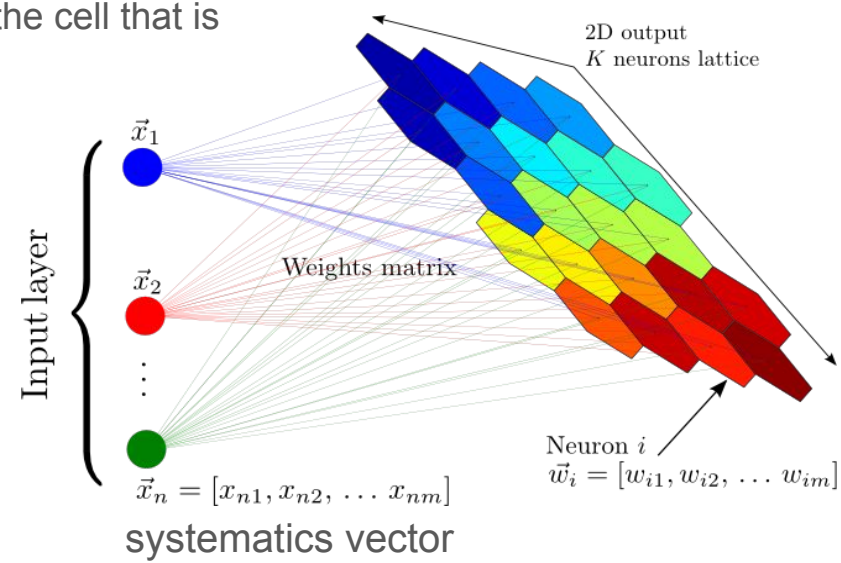


Self organising maps

Unsupervised artificial neural networks designed to project high-dimensional data onto a 2D map, preserving topological features of the space

After SOM training, each point (galaxy) is represented by the cell that is closest to it (the best-match unit, BMU)

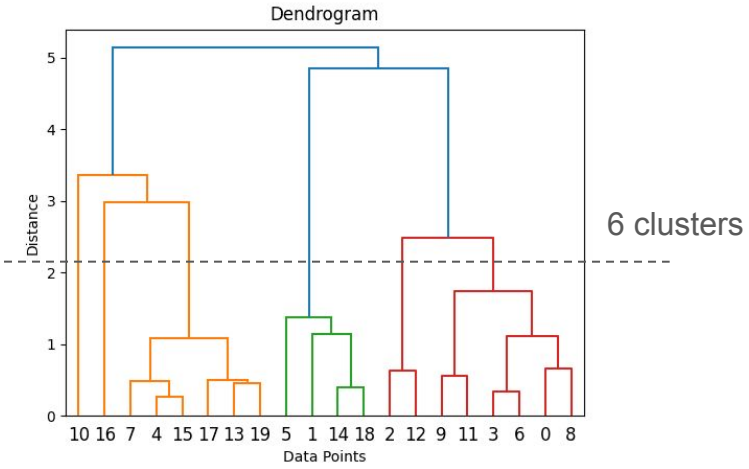


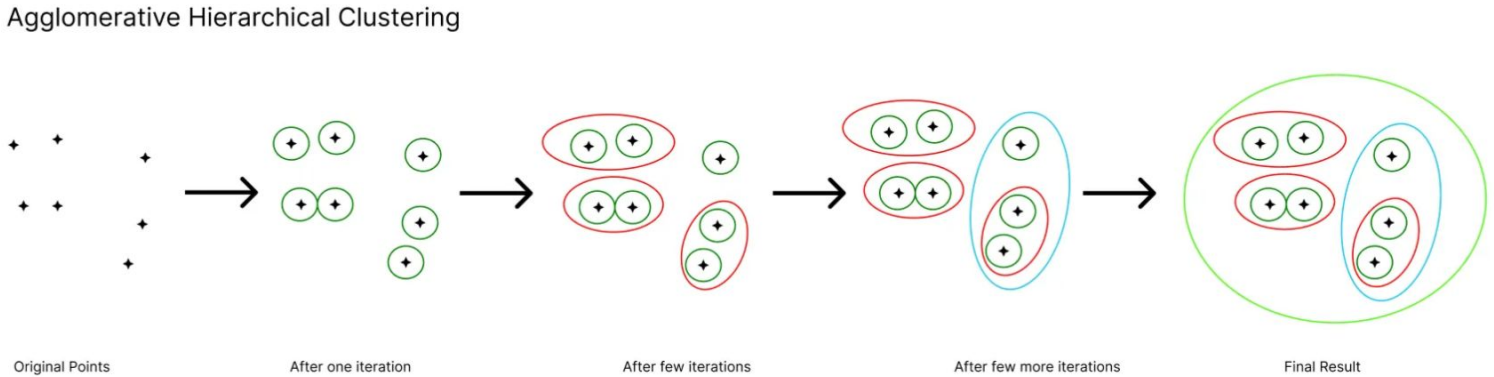


Agglomerative hierarchical clustering

Grouping the SOM cells into hierarchical clusters (HC) from bottom-up:

1) treat each SOM cell as its own cluster;
2) merge clusters according to some distance measure between clusters;
3) iteratively merge until there is only one cluster;
4) construct the dendrogram and cut it where there are number of clusters needed.





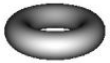
Recovering “Organized Randoms” with SOM+HC

Using SOM+HC to recover OR:

1) get the input: systematics vectors;

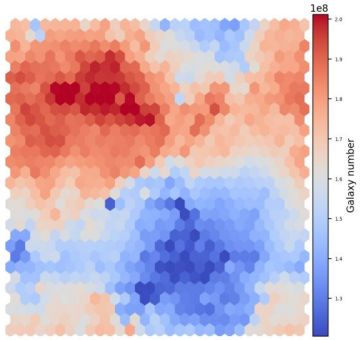
2) training the SOM from systematics vectors;

SOM grid type: hexagonal;

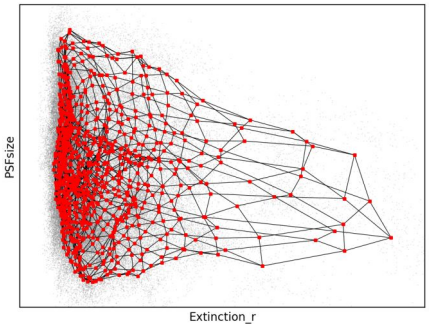
SOM topology: toroidal 

$\{s^\alpha\}$ 1. Input: systematics vectors

2. SOM training



galaxy number in each SOM cell



Position of SOM cells (red connected dots) and original systematics vectors (small gray dots) on the projected PSF size-extinction space

Recovering “Organized Randoms”

Using SOM+HC to recover OR:

1) get the input: systematics vectors;

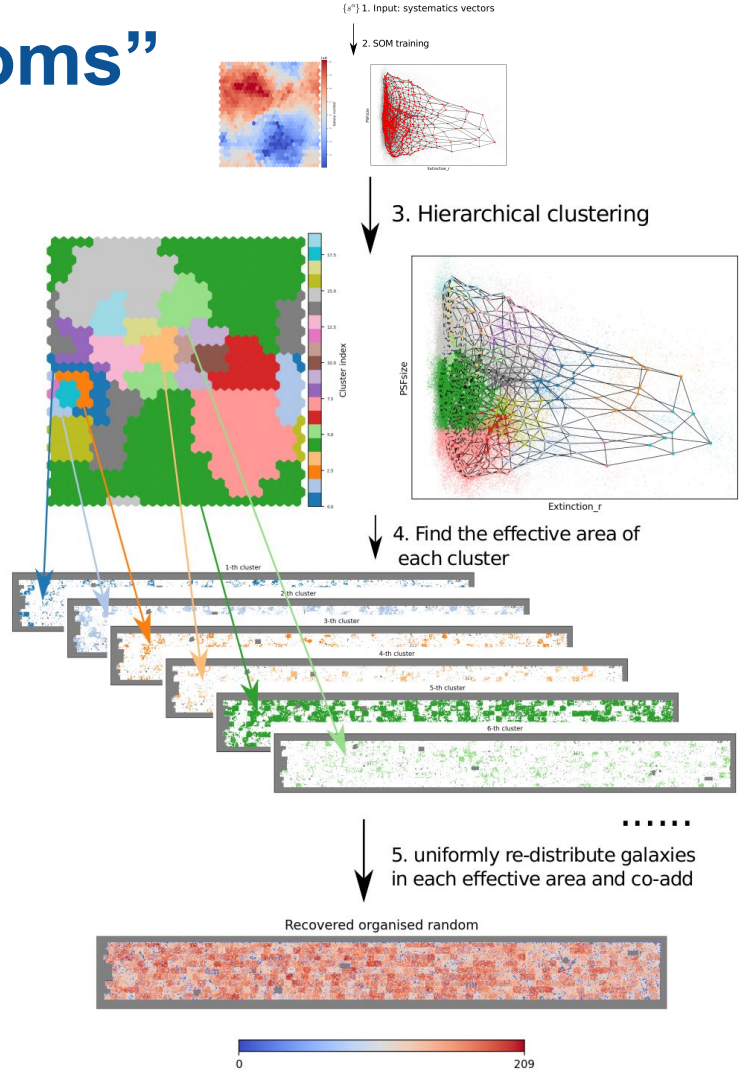
2) training the SOM from systematics vectors;

3) grouping SOM cells into hierarchical clusters (HC);

4) find the effective area of galaxies in each cluster and resample galaxies uniformly in each effective area;

5) combine these uniform randoms from all the clusters

Practical remark: the recovered OR is a pixelised sky map (“OR weight”). The pixel values are number of organised randoms in that pixel. Only pixels occupied by galaxies have the OR recovered



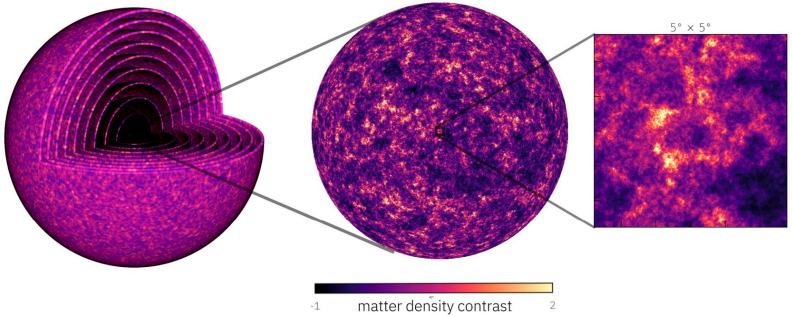
Validation of the method with mock galaxy catalogue

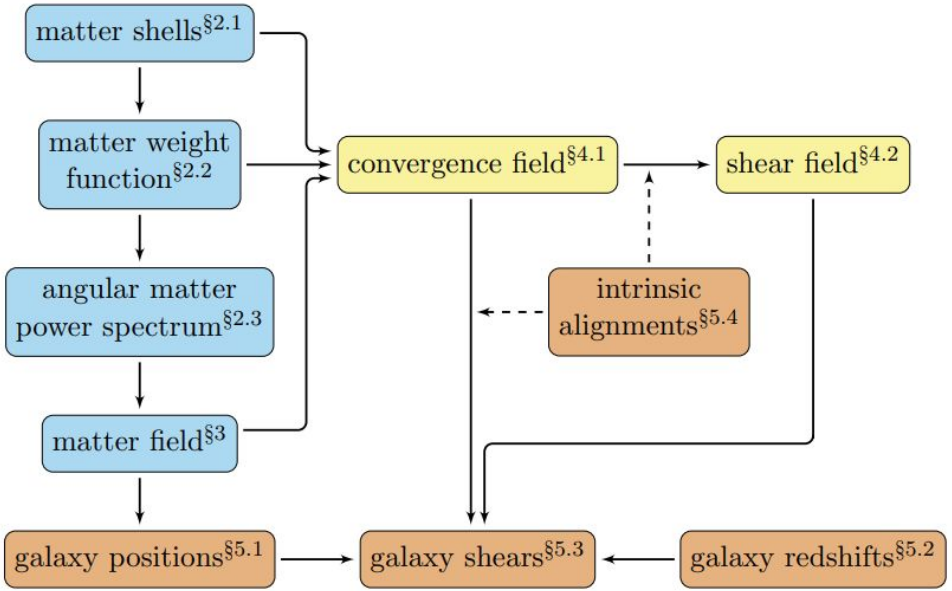
The Generator of Large-Scale Structure (GLASS)

GLASS (Tessore et al.,2023): a public code to generate mock data for LSS studies (https://glass.readthedocs.io/stable/);

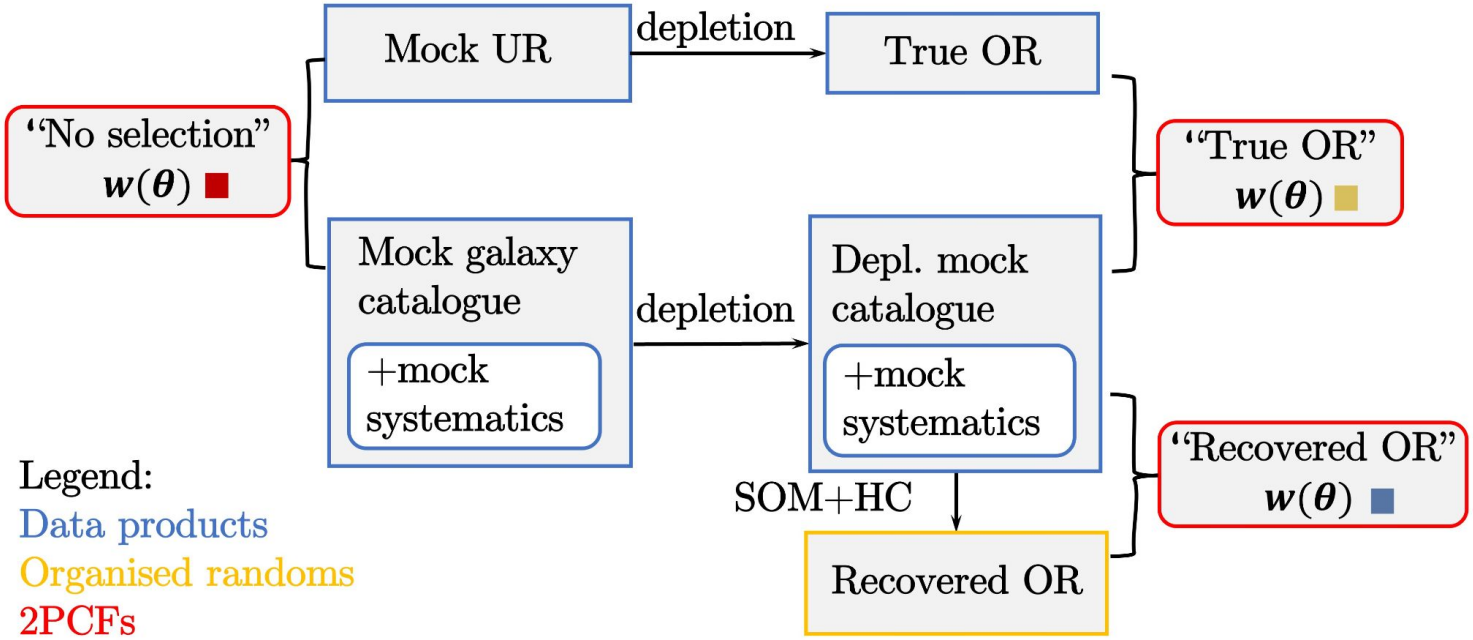
It takes an input matter power spectrum and generates a lognormal Gaussian fields on a sequence of shells throughout the light cone.

Galaxy positions and shears can then be generated accordingly given the redshift distributions.





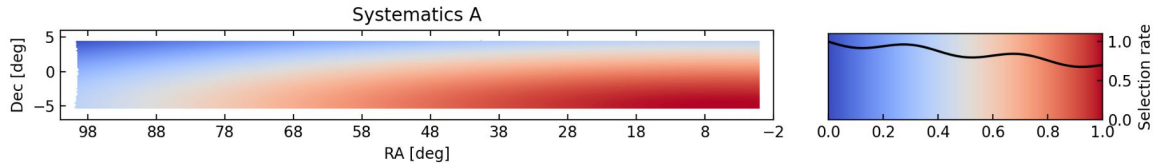
Validation I: toy models for systematics and selections



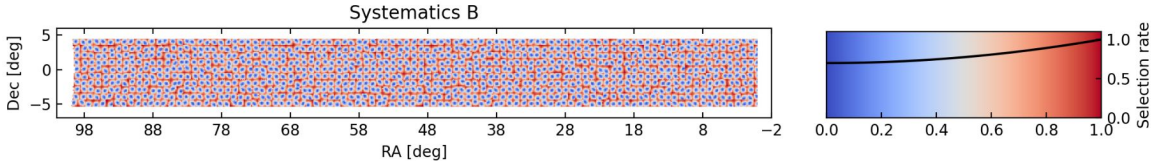
Validation I: toy models for systematics and selections

Generate galaxy samples from GLASS, then assign them with four systematics:

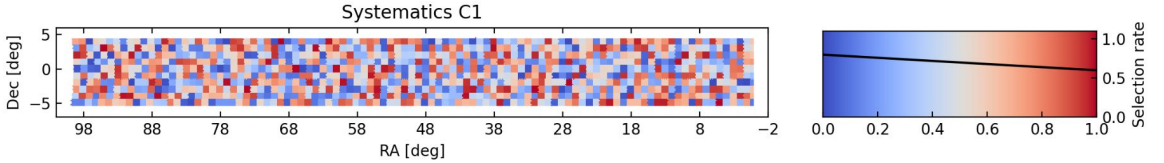
A: **large-scale variance**, mimicking galactic extinction;



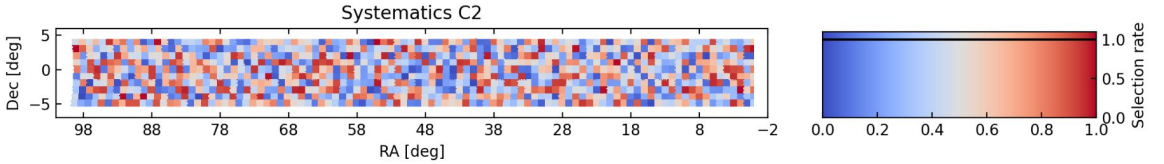
B: **intra-tile variance**, mimicking PSF shape across the focal plane;



C1: **inter-tile variance**, mimicking anisotropic seeing condition;

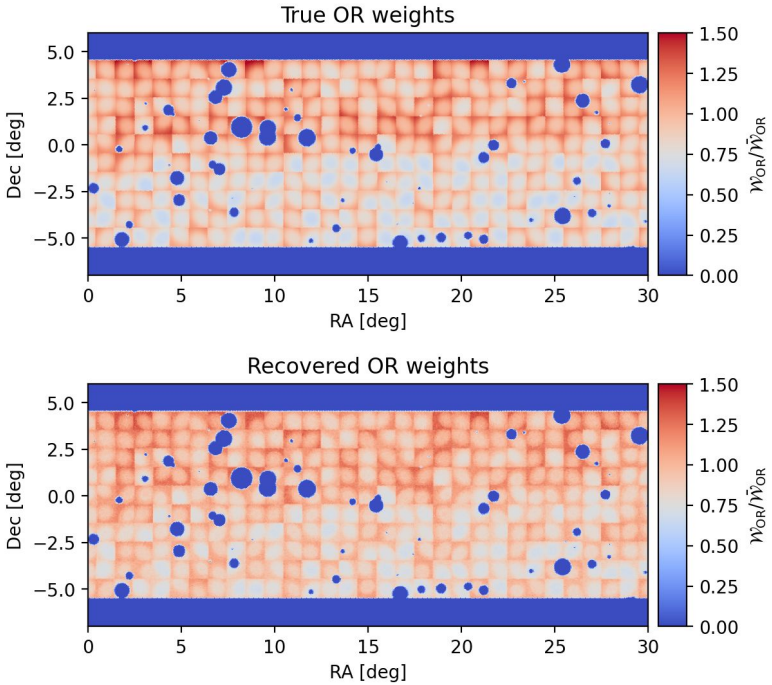


C2: **inter-tile variance**, with no selection (distractor)

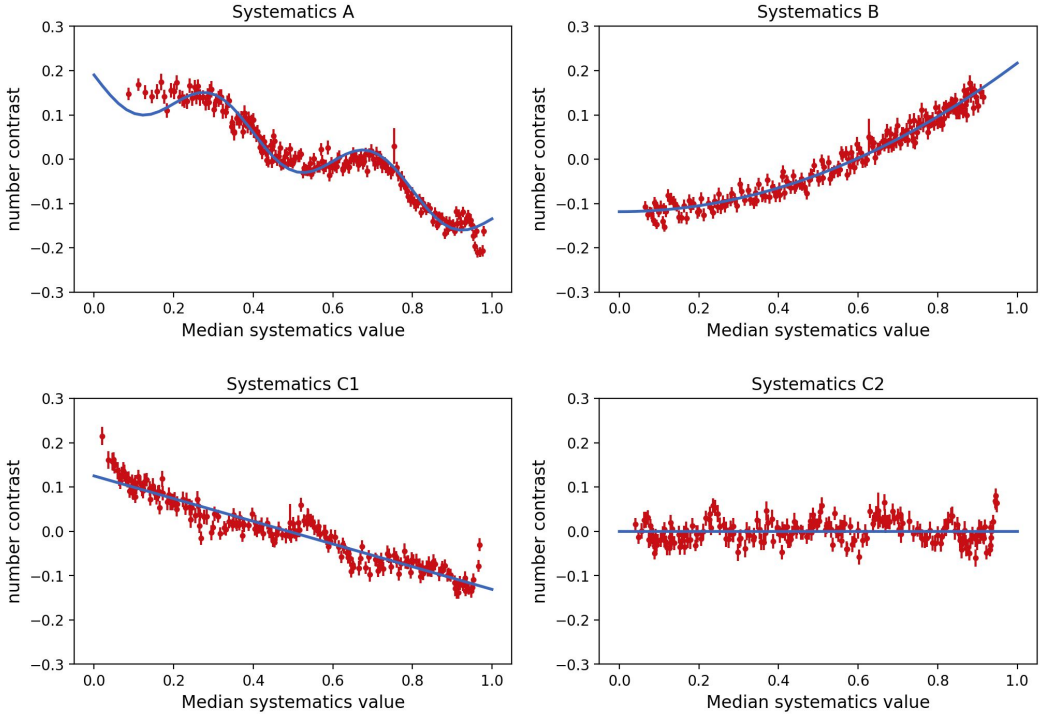


Validation I: toy models for systematics and selections

Recovered OR



Recovered selection functions



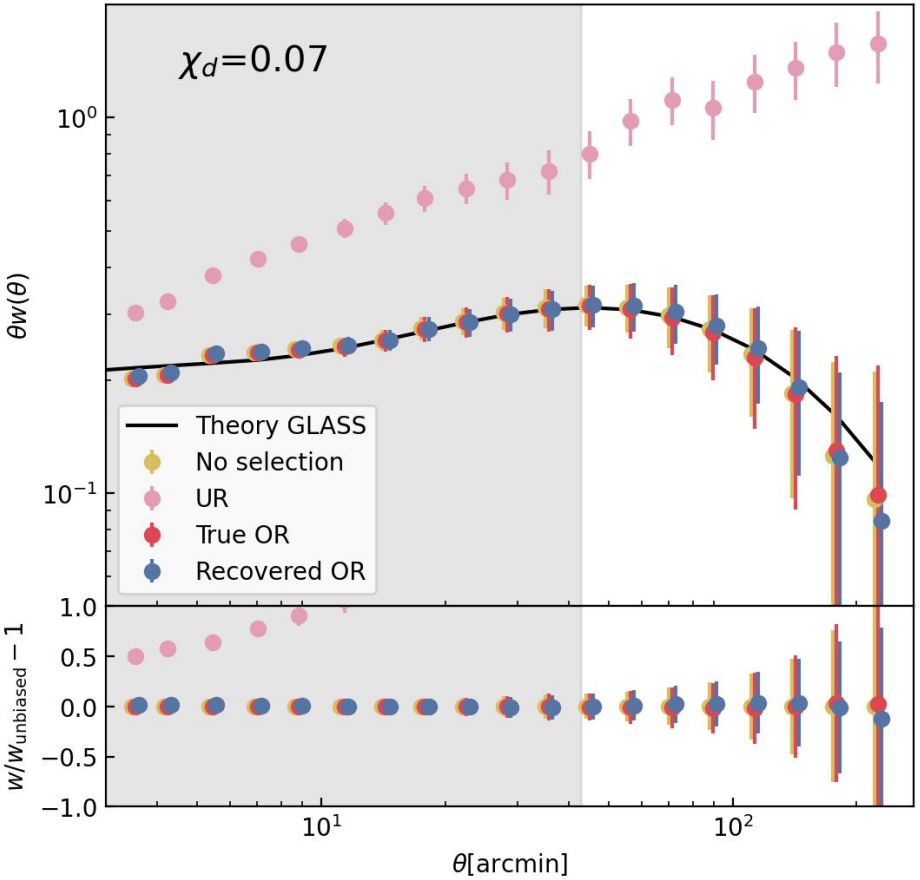
mean galaxy number contrast in each cluster
vs. median systematics in that cluster

Validation I: toy models for systematics and selections

The chi-square value between uniform random (UR) and “No selection” is **1016**

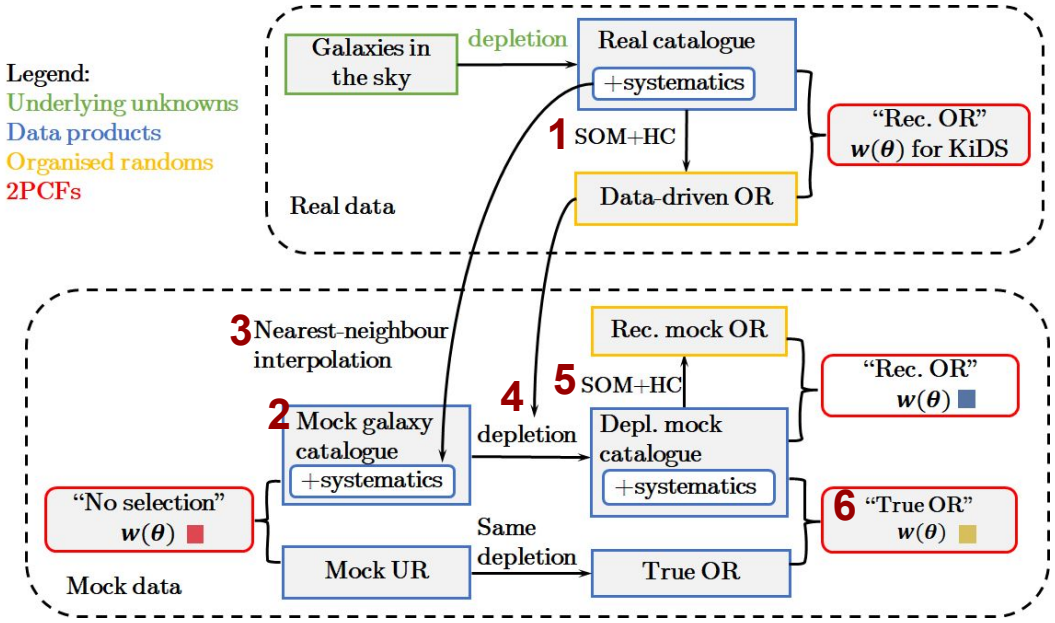
0 between No selection and True OR,

0.07 between “recovered OR” and No selection



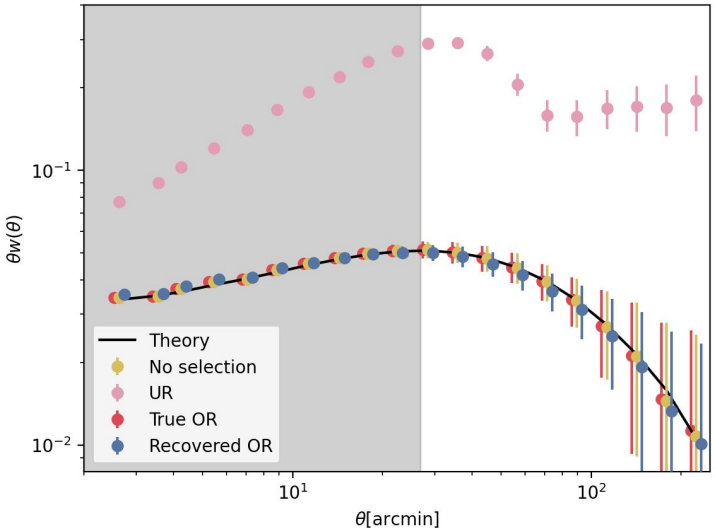
Validation II: Data-driven systematics

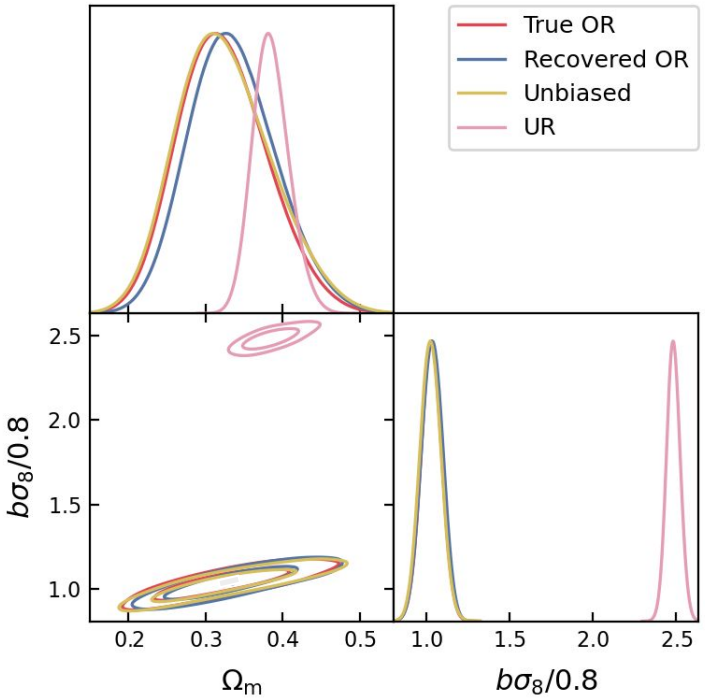
1. construct the “data-driven” OR with SOM+HC from the real data;
2. generate mock galaxies following the same footprint as the data
3. assign systematics to mock galaxies via nearest-neighbour interpolation from real systematics;
4. select the mock catalogue with the “data-driven” OR
5. recover the OR from the selected mock sample
6. measure the 2PCF



Validation II: Data-driven systematics

The chi-square value between UR and "No selection" is **8476**; recovered OR reduced it to **0.29**



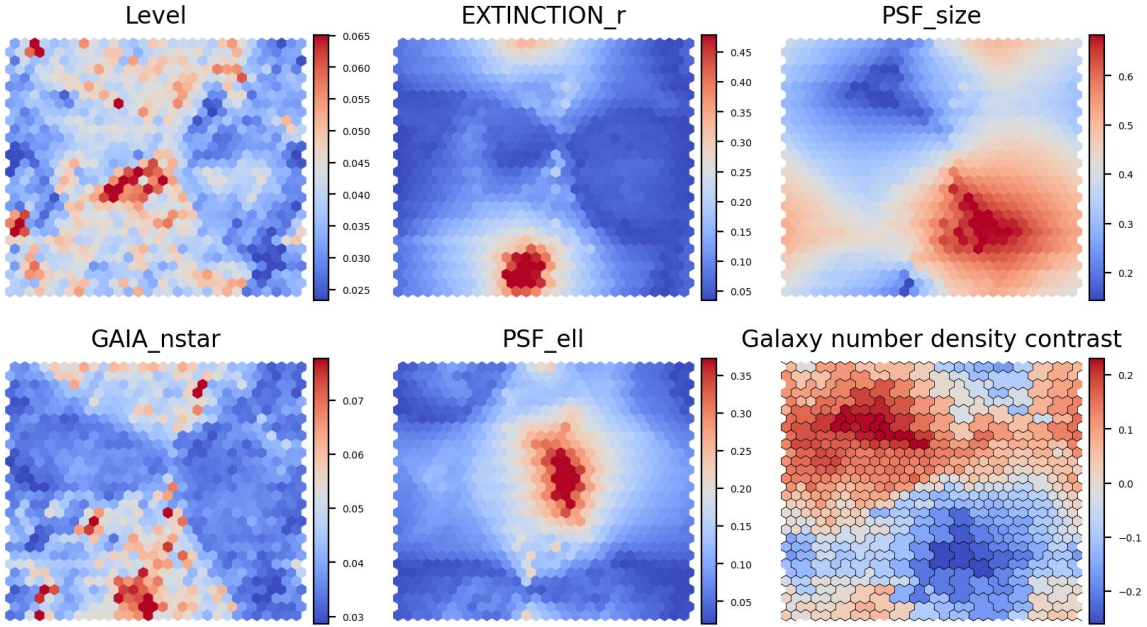


Since the mock catalogue is generated with arbitrary cosmology and n(z), we also conclude that the variable depth does not depend on cosmology and n(z)

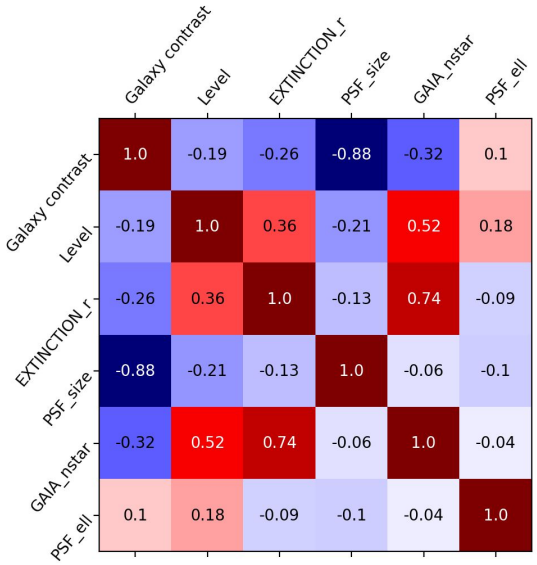
Application on KiDS-Legacy

KiDS-Legacy OR weight recovered from SOM+HC

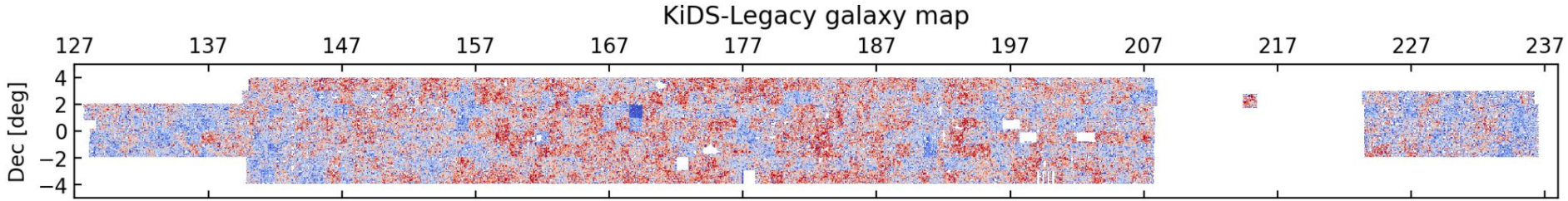
Mean systematics value and galaxy number in each SOM cell

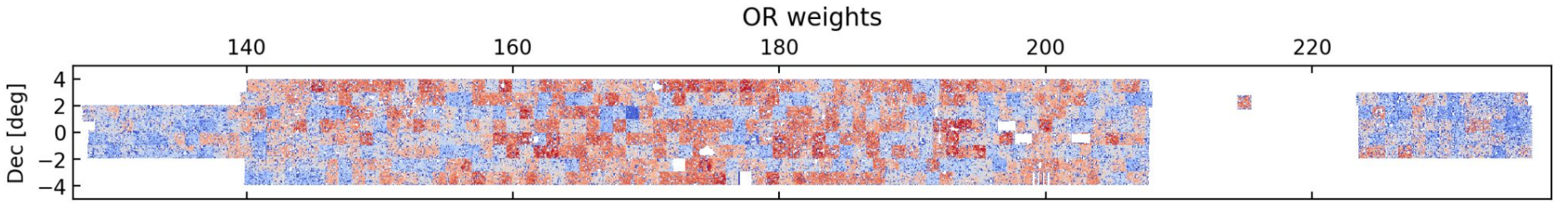


Correlation among systematics and galaxy numbers in all the clusters



KiDS-Legacy OR weight recovered from SOM+HC

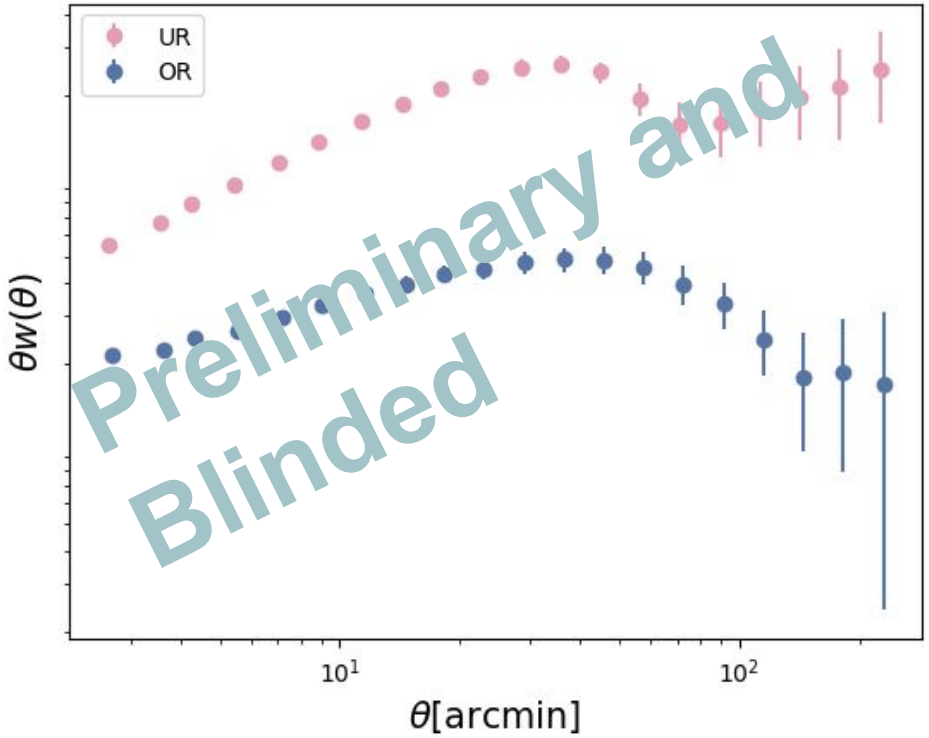




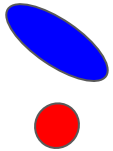
Only the northern field is shown here

KiDS-Legacy OR weight recovered from SOM+HC

blinded measurements (randomly shifted) with the whole legacy catalogue

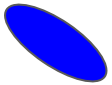


Future prospects: KiDS-Legacy 6x2pt

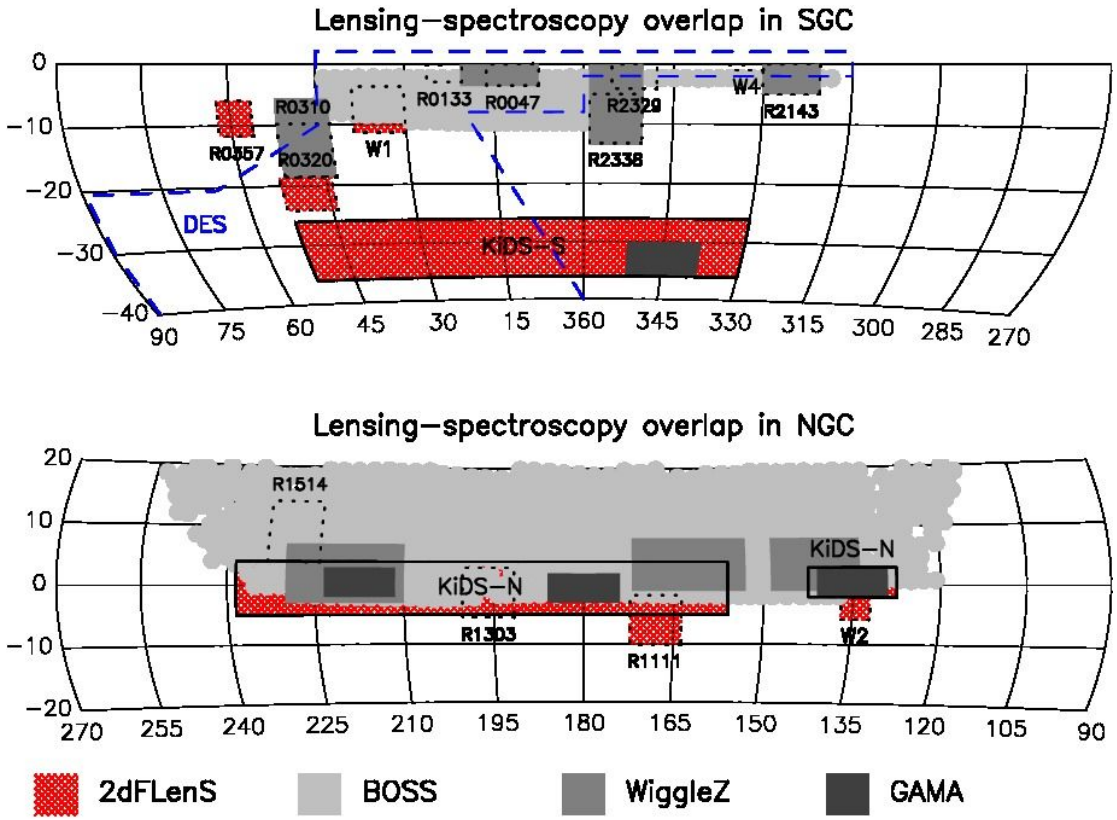


Photometric sample with shears, inaccurate redshift distribution (from KiDS-legacy)

Spectroscopic sample with known redshift distribution (from BOSS+2dFLenS)

			
	Spectroscopic (2dFLenS) galaxy lensing	**Photometric galaxy lensing**	Cosmic shear
	**Photo x Spec (BOSS) clustering**	**Photo-photo clustering**	
	Spec (BOSS)-clustering		

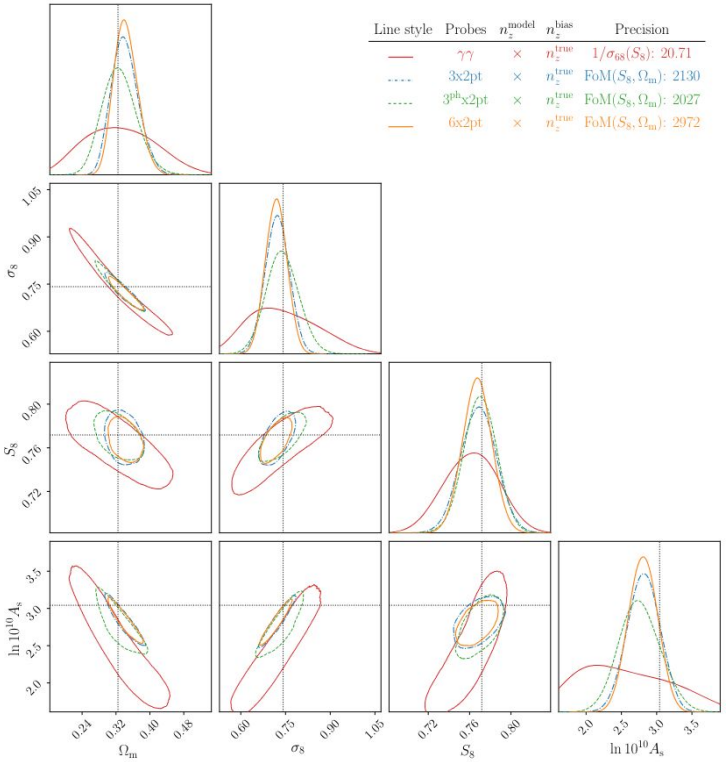
Future prospects: KiDS-Legacy 6x2pt



https://2dflens.swin.edu.au/survey.html

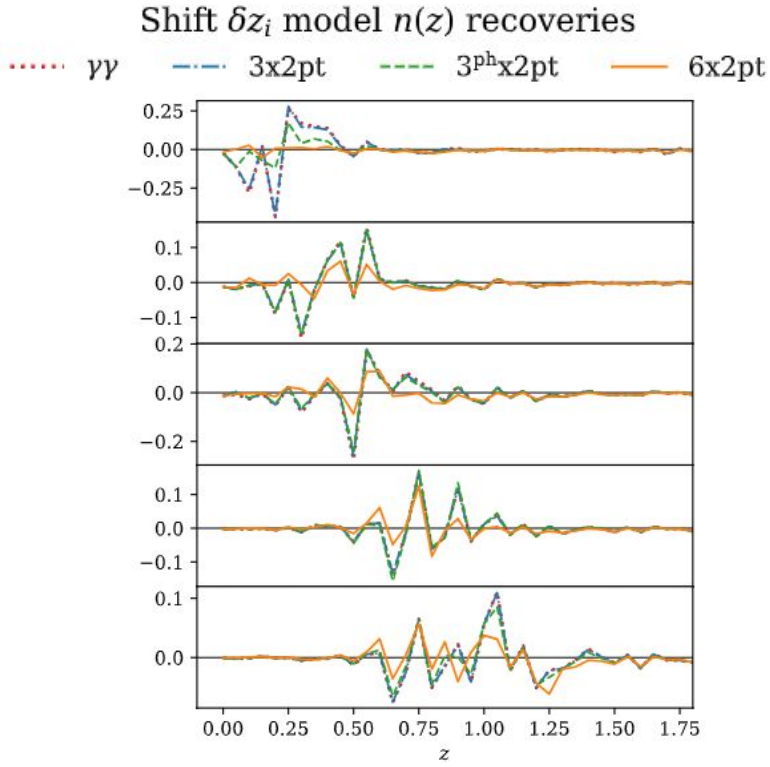
Future prospects: KiDS-Legacy 6x2pt

Forecast of cosmological parameters with different statistics



(Johnston et al., in prep)

Forecast of n(z) self-calibration



Summary

- Anisotropic systematics selection will **bias** the 2PCF if one uses uniform randoms;
- **“Organised randoms”** is used instead to correct the bias;
- We can recover the OR with the **SOM+HC method**. The method is validated with mock galaxy data
    - Advantages: this method is **model independent**, **flexible** enough to account for arbitrarily complicated selection functions;
    - Limitations: works only when the selections are cosmology-independent; bad performance for sparse sample (on the other hand, better performance for larger sample)

Future prospects

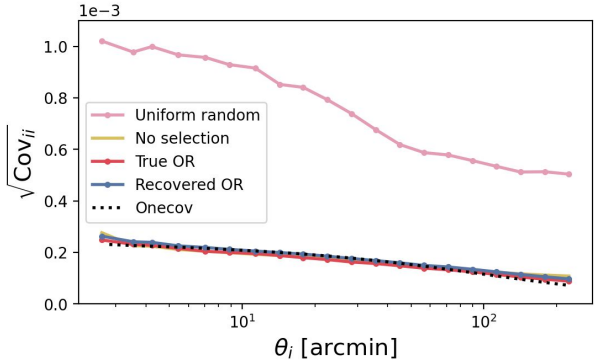
- apply it to KiDS legacy 6x2pt and 3x2pt analyses
- also useful for future surveys like Euclid and LSST

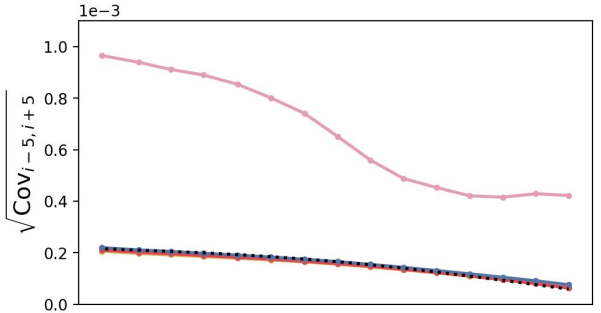
Backup slides

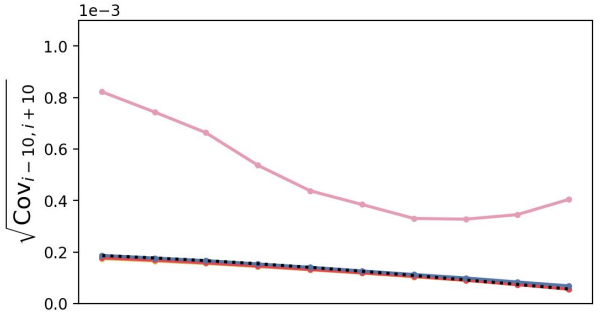
Covariance of 2PCF

In our test, the covariance matrices are calculated from 200 GLASS realisations.

We also calculate a theoretical covariance with the OneCovariance (Reischke et al., in prep) code (black dotted lines).







2PCF in harmonic space: pseudo-Cℓ

- We can also measure the 2PCF in harmonic space, i.e the angular power spectra.
- In practice, when the fields are partial/weighted-sky, we can measure the "pseudo-C$_\ell$" (PCL, Alonso et al., 2018)
- for galaxy field, we first construct the galaxy overdensity map by

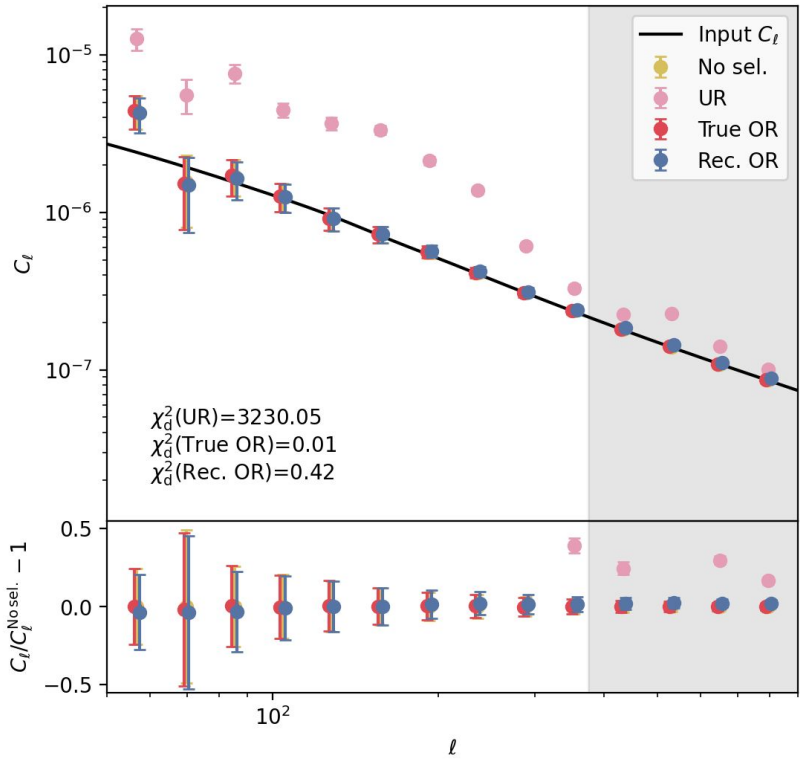
$$\delta(p) = \frac{N(p)}{w(p)\overline{N}} - 1,$$

- the "coupled PCL" is measured as

$$\tilde{C}_\ell^{ab} = \frac{1}{2\ell+1}\sum_m \tilde{a}^a_{\ell m}\tilde{a}^{b*}_{\ell m},$$

- it is connected to the underlying power spectra with the "mode-coupling matrix"

$$\tilde{C}_\ell^{ab} = \sum_{\ell'} M_{\ell\ell'}(w^a, w^b) C_\ell^{ab},$$



Find the optimal SOM+HC setup

The most important SOM+HC configuration is the number of cluster (NHC) and the angular resolution (described by Healpix Nside) of the OR weight map.

Too many HC: overfitting; sample noise in each cluster;
Too few HC: fail to capture variable selections;

Resolution too low: fail to capture variable depth;
Resolution too high: “unintentionally masked” regions on the OR weight causes over-correction

(Nside=4096 will see significant amount of holes on OR while Nside=2048 is still acceptable. We take 2048 as the optimal choice)

Find the optimal SOM+HC setup

Configuration NC_{KiDS}	Configuration NC_{rec}	χ^2_d	PTE_d	$\Delta\Omega_m[\sigma]$	$\Delta b[\sigma]$	$\chi^2_{\Delta\Omega_m,\Delta b}$
200	200	0.87	0.9997	0.14	0.4	0.24
400	400	0.31	1.0	0.18	-0.08	0.27
600	600	0.61	0.9999	0.23	-0.17	0.38
800	800	0.93	0.9996	0.18	-0.23	0.37
400(m=1.5)	400	0.43	1.0	0.32	0.27	0.1