Clustering Statistics for Cosmology

Author: Syed Faisal ur Rahman

1- Lahore University of Management Sciences (LUMS)
2- Blockchain Laboratories LLC
3- National Center for Big Data and Cloud Computing- NEDUET

Basic Definitions

Understanding Large-Scale Structures: Clustering Statistics

- Clustering statistics are tools used in cosmology to study the spatial distribution of galaxies and matter in the universe.
- They help us explore how galaxies group together, forming cosmic structures like clusters, filaments, and voids.
- Clustering statistics provide insights into the underlying cosmological principles that govern the evolution of the universe.

Two-Point Spatial Correlation Function

The two-point spatial correlation function, denoted by ξ(r), quantifies the excess probability of finding a pair of galaxies at a given separation r compared to a random distribution. It is defined as:

$\xi(r) = (DD(r) - DR(r)) / DR(r)$

where:

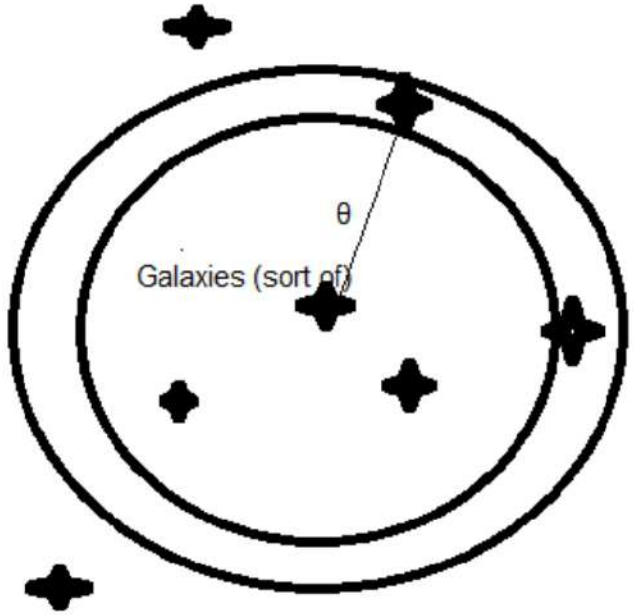
DD(r) is the number of galaxy pairs with a separation between r and r + dr

DR(r) is the expected number of galaxy pairs with a separation between r and r + dr assuming a random distribution

The two-point spatial correlation function is typically plotted as a function of separation r.

Read for more: https://ned.ipac.caltech.edu/level5/March12/Coil/Coil2.html

Two-Point Angular Correlation Function



The two-point angular correlation function, denoted by w(θ), quantifies the excess probability of finding a pair of galaxies at a given angular separation θ compared to a random distribution. Computin It is defined as:

$w(\theta) = (DD(\theta) - DR(\theta)) / DR(\theta)$

where:

DD(θ) is the number of galaxy pairs with an angular separation between θ and θ + dθ

DR(θ) is the expected number of galaxy pairs with an angular separation between θ and θ + dθ assuming a random distribution

The two-point angular correlation function is typically plotted as a function of angular separation θ.

Read for more: https://ned.ipac.caltech.edu/level5/March12/Coil/Coil3.html

Landy-Szalay Estimator

The Landy-Szalay estimator is a popular method for estimating the two-point spatial correlation function. It is based on the idea that the number of galaxy pairs with a separation between r and r + dr is proportional to the area of a concentric shell with radii r and r + dr. The estimator is defined as:

$$\xi_{LS}(r) = \frac{DD(r) - 2DR(r) + RR(r)}{RR(r)},$$

where:

RR(r) is the number of random pairs with a separation between r and r + dr

The Landy-Szalay estimator is a simple and efficient method for estimating the two-point spatial correlation function, but it can be biased for small separations.

AstroML Two-Point Correlation Function Methods

AstroML provides a variety of methods for estimating the two-point spatial and angular correlation functions. These methods include the Landy-Szalay estimator and the Pearson estimator. AstroML also provides methods for estimating the covariance matrix of the two-point correlation function, which is important for error estimation.

AstroML uses the bootstrap resampling technique, which involves randomly drawing multiple sub-samples from the data with replacement. Similar to the jackknife method, the two-point correlation function is calculated for each sub-sample, and the covariance matrix is estimated from the variations in the measurements across the different sub-samples.

AstroML also provides methods for estimating the covariance matrix directly from the data, without the need for resampling. These methods typically involve fitting a model to the two-point correlation function and assuming that the model residuals are normally distributed. The covariance matrix can then be estimated from the variance of the model residuals.

Read for more: https://www.astroml.org/user_guide/correlation_functions.html

TreeCorr Two-Point Correlation Function Methods

TreeCorr is a Python package that provides a variety of efficient methods for estimating the two-point spatial correlation function. These methods are based on the idea of using a tree data structure to efficiently compute the pair counts needed to estimate the two-point correlation function. TreeCorr also provides methods for estimating the covariance matrix of the two-point correlation function.

This covariance matrix provides essential information about the uncertainties associated with the estimated 2PCF values, enabling researchers to conduct reliable statistical analyses and make informed inferences.

TreeCorr specifically provides implementations for the following methods:

Jackknife resampling: This method involves omitting one data point or patch at a time and computing the 2PCF for each sub-sample. The covariance matrix is then estimated from the variations in the 2PCF measurements across the different sub-samples.

Bootstrap resampling: This method involves randomly sampling data points or patches with replacement and computing the 2PCF for each sub-sample. The covariance matrix is then estimated from the variations in the 2PCF measurements across the different sub-samples.

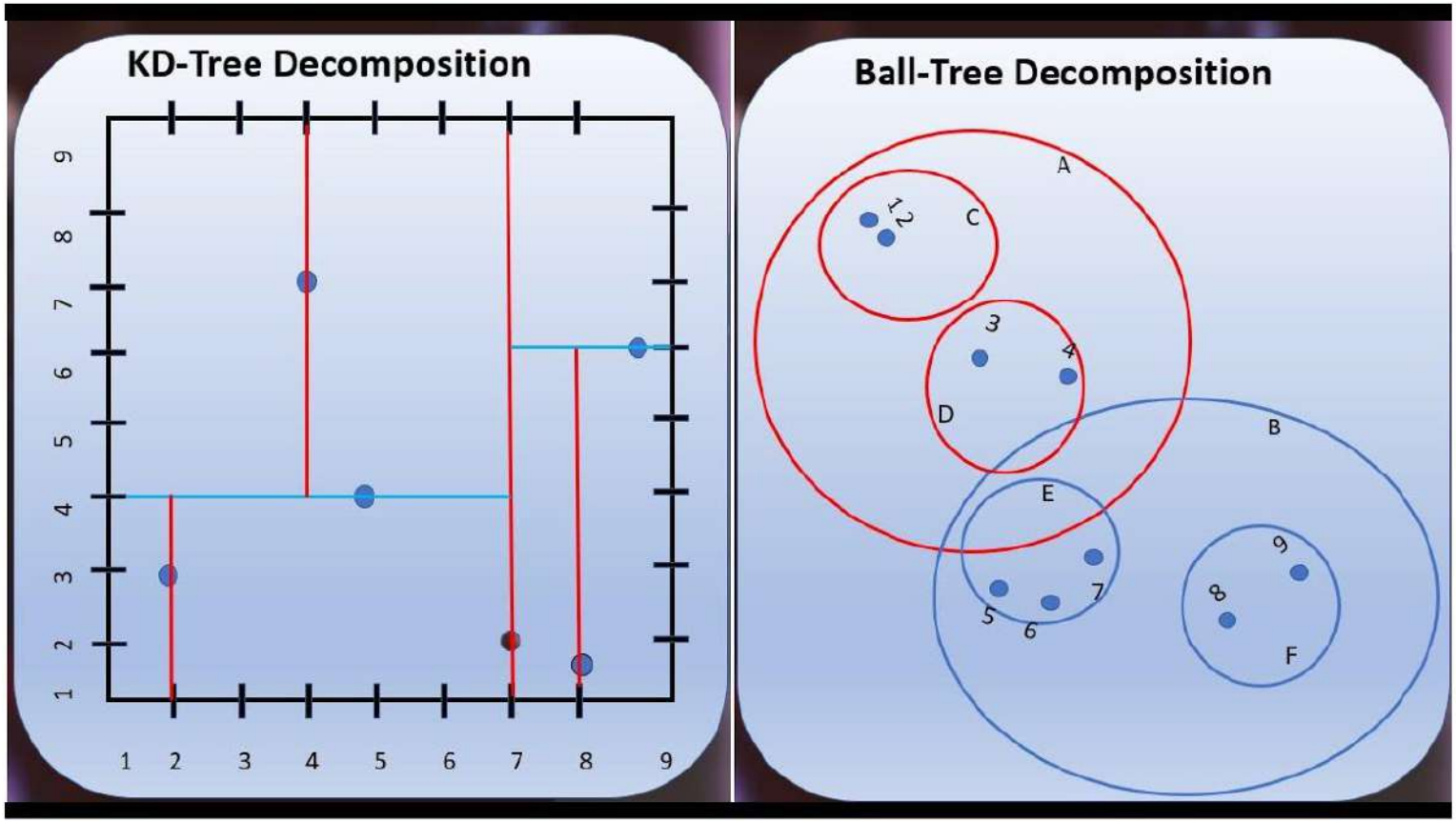
Shot noise covariance: This method estimates the covariance matrix based on the Poisson shot noise associated with the data. This is a simple and efficient approach that is well-suited for datasets with low noise levels.

Patch-based covariance: This method considers the spatial distribution of data points or patches and estimates the covariance matrix based on the variations in the 2PCF measurements across different patches. This approach is particularly useful for datasets with non-uniform data distributions.

Read for more: https://rmjarvis.github.io/TreeCorr/_build/html/index.html

k-d Tree and Ball Tree

A k-d tree and a ball tree are both types of tree data structures that can be used to efficiently compute the pair counts needed to estimate the two-point correlation function. A k-d tree recursively partitions the data space into a hierarchy of hypercubes, while a ball tree recursively partitions the data space into a hierarchy of spheres. Both k-d trees and ball trees can be used to efficiently find all pairs of points within a given distance of each other.



For More: https://towardsdatascience.com/tree-algorithms-explained-ball-tree-algorithm-vs-kd-tree-vs-brute-force-9746debcd940

https://www.youtube.com/watch?app=desktop&v=fy40y3UFkNE

Concepts in action!

AUTOCORRELATION FUNCTIONS, COSMOLOGY AND INVESTIGATING THE CMB COLD SPOT WITH EMU-ASKAP RADIO CONTINUUM SURVEY

Ref: https://arxiv.org/abs/2109.10734

Introduction

Galaxy angular-power spectrum provides information about the distribution of matter by using galaxy counts as a proxy.
In this study, we are going to estimate autocorrelation angular power spectrum and angular autocorrelation function (ACF) for EMU 5 sigma sources and then compare them with results from NVSS and SUMSS.

Angular Power Spectrum and Autocorrelation Function

We can write the angular power spectrum, Cl_{gg} as [1]:

$$Cl_{gg} = 4\pi \int_{kmin}^{kmax} \frac{dk}{k} \Delta^2(k) \{Wl_g(k)\}^2 \qquad \text{Equation (1)}$$

Here $\Delta^2(k)$ is the logarithmic matter power spectrum, which can be calculated as:

$\Delta^2(k) = \frac{k^3}{2\pi^2} P(k)$

In this, P(k) is the matter power spectrum. $Wl_g(k)$ is galaxy window function.
For this analysis we adopted Limber's approximation as discussed in [8][9]. This will give an expression for Cl_{gg} as:

$$Cl_{gg} = 4\pi \int_0^{z*} \frac{dz}{c} \frac{H(z)}{X^2(z)} \{Wl_g(z)\}^2 \; P(k) \qquad \text{Equation (2)}$$

With k being approximated as [10]:

$$k = \frac{l + \frac{1}{2}}{\chi^2(z)}$$

With the window function defined as [9]:

$$Wl_g(z) = b(z)\frac{dN}{dz} + \frac{3\Omega_m}{2c}\frac{H_0^2}{H(z)}(1+z)\mathrm{X}(z)\int_z^{z_*} dz'\left(\frac{\mathrm{X}(z') - \mathrm{X}(z)}{\mathrm{X}(z')}\right)(\alpha(z') - 1)\frac{dN}{dz'}$$

Equation (3)

The second part after '+' sign represents the contribution from the magnification bias.

It is dependent on $\alpha(z)$ which is the slope of the integral count ($N(>S) = CS^{-\alpha}$).
In surveys like NVSS or SUMSS, where α is close to 1, this part doesn't play a significant role.

Magnification bias contribution also depends on the shape of $\frac{dN}{dz}$.
We checked for EMU 5-sigma using SKADS data and found it to play a negligible part in the overall galaxy autocorrelation power spectrum.

However, with real data, it will be good to measure it again after carefully measuring α and $\frac{dN}{dz}$.

In figure (1), we used SciPy's curve_fit [13] for the model, $N(z) = \frac{dN}{dz} = \left(\left(\frac{z}{a}\right)^{b}\right) exp\left(\frac{-cz}{a}\right)$ with a=0.40086143 b=0.60527401 c=0.54263201. Blue '+' represent SKADS data points, green solid line represents our model and others are fit from numpy's polyfit [14] function which are used for comparison here.

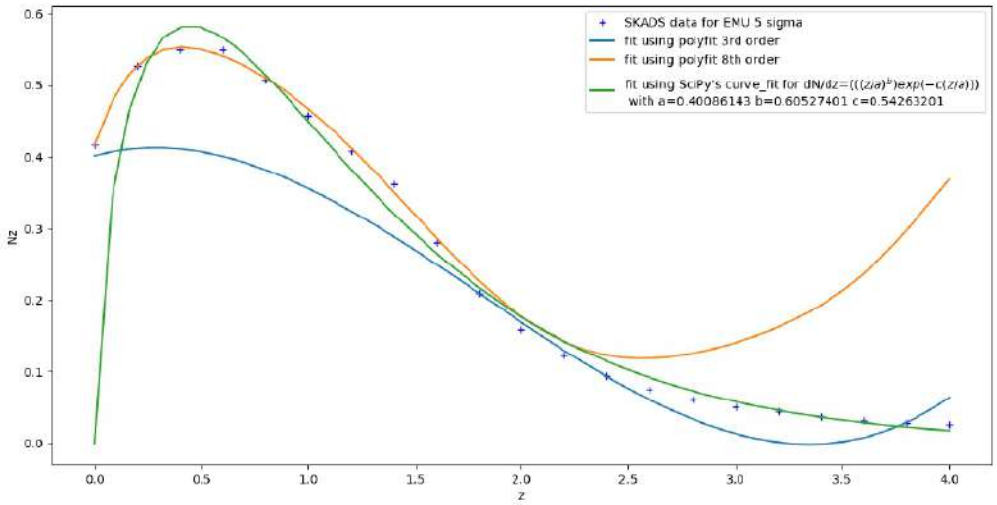


Figure 1:$\frac{dN}{dz}$ from SKADS for EMU 5 sigma.

For EMU 5 sigma, galaxy bias b(z) is used as a weighted average of bias of different source types in each redshift bin and after z=3, we use a constant bias (see figure 2).

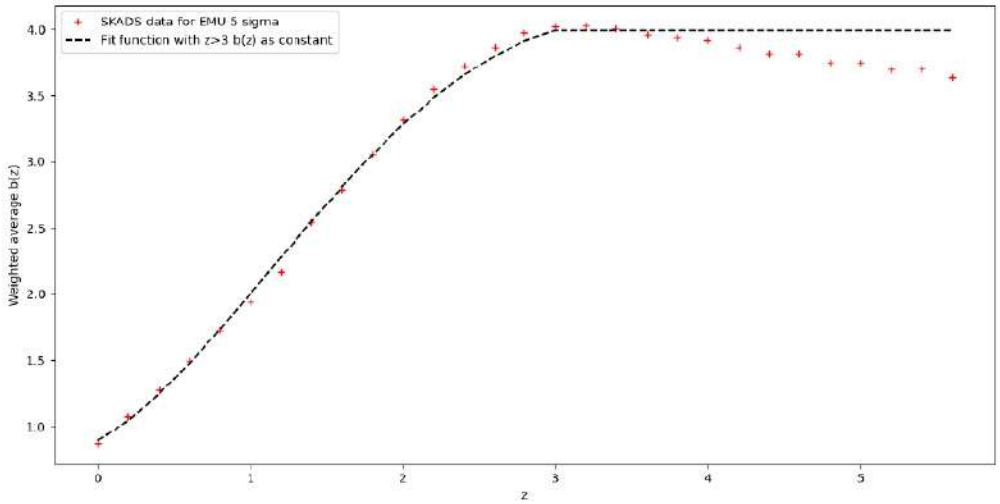


Figure 2: linear weighted average bias b(z) for EMU 5 sigma sources from skads and the best fit function. In the fit function, we use a constant value after z=3. Here, weight is based on the proportion of source types present in each redshift bin.

We use redshift distribution for NVSS as provided by [15] [16] and galaxy bias as discussed in [17]. For NVSS, we use a redshift dependent bias function [17]:

$$b(z) = 0.90[1 + 0.54(1 + z)^2]$$

From 'Cl' values, we can get the angular correlation function (ACF) as [1]:

$$w_{gg}(\theta) = \sum_{l=l_i}^{lmax} \frac{2l+1}{4\pi} Cl_{gg} Pl(cos\theta)$$ Equation (4)

For our analysis we used l_i=2 and l_{max}=1000. This method of converting 'Cl' values into ACF values is suitable for both theory and large scale observations with large sky coverage (fsky).

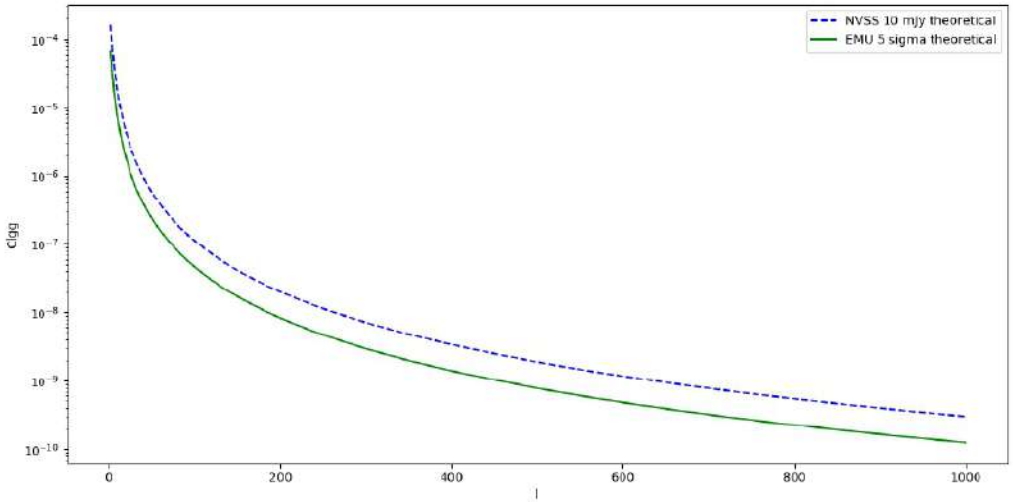


Figure 3: Comparison of autocorrelation angular power spectrum between theoretical estimates under Λ-CDM assumption for EMU 5 sigma and NVSS.

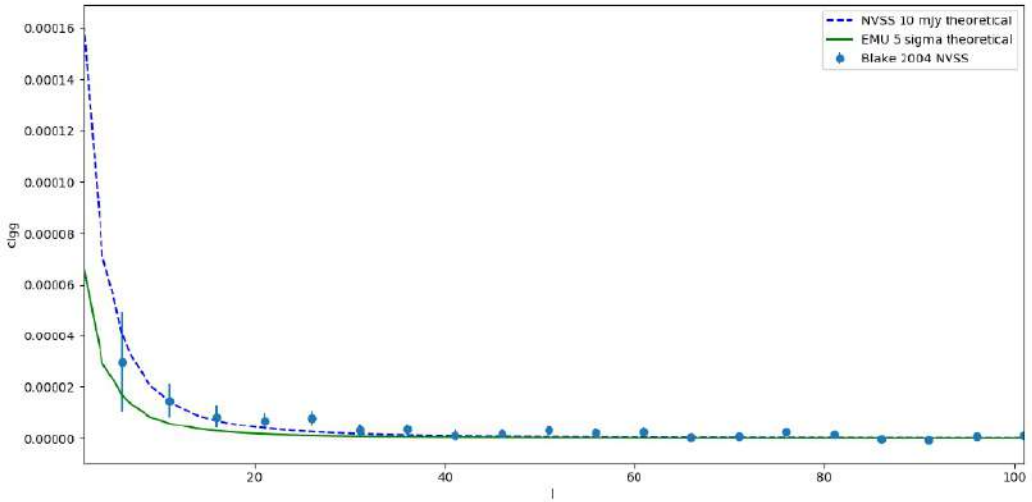


Figure 4: Comparison between theoretical estimates for EMU 5 Sigma and NVSS with Blake 2004 values.

For small sky patches, a suitable estimator is Landy-Szalay estimator [18], which can be used as:

$$w_{gg}(\theta) = \frac{DD(\theta) - DR(\theta) + RR(\theta)}{RR(\theta)}$$ Equation (5)

Here,

DD=Two point count-count correlation from galaxy catalog.

DR=Two point cross correlation of counts from galaxy and random catalogs.

RR=Two point count-count correlation from random catalog.

Random catalog is developed in such a way that number of sources in random catalog are greater than or equal to ten times the number of sources in the galaxy catalog.

We use 100 bootstrap iterations to compute mean and variance for ACF estimates. For SUMSS, sources, we use $-65° < declination < -50°$ and $290° < right\ ascension < 340°$ with flux>20 mJy. For NVSS's flux > 10 mJy configuration, we use $5° < declination < 58°$ and $12° < right\ ascension < 34°$.

We also removed sources > 1Jy from our analysis. We use WMAP 9 years cosmological parameter results to perform theoretical calculations [19]. We can see angular power spectrum results in figures (3 and 4) where we can see comparison between EMU 5 sigma estimates with theoretical values of NVSS and Blake 2004 results

Our ACF results can be seen in figures (5 and 6) where EMU 5 sigma results are compared with NVSS flux > 10 mJy, SUMSS > 20 mJy and NVSS ACF power law fit results from [20].

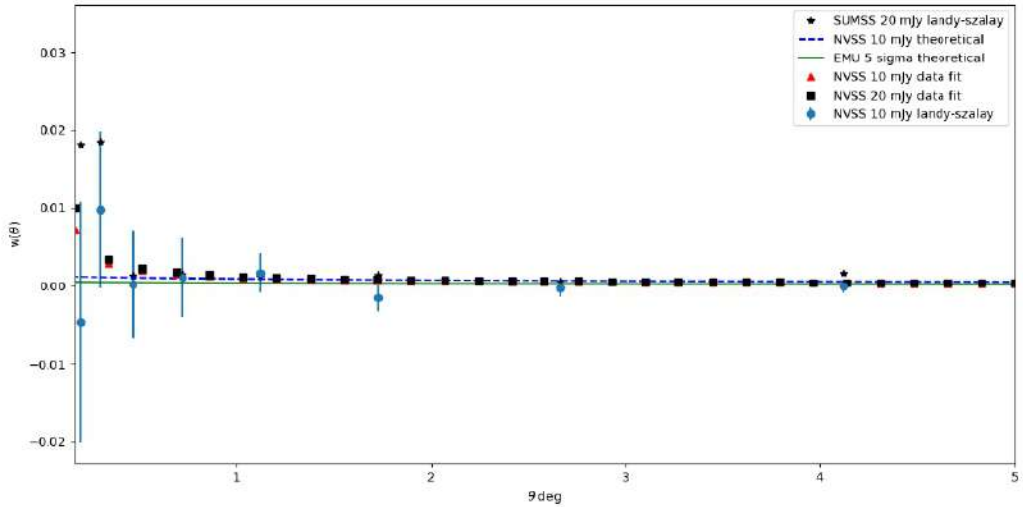


Figure 5: Comparison of autocorrelation function (ACF)'s from EMU 5 sigma, NVSS theory, NVSS power law fits from Blake 2002, and NVSS and SUMSS ACF's measured using AstroML's Landy-Szalay estimator.

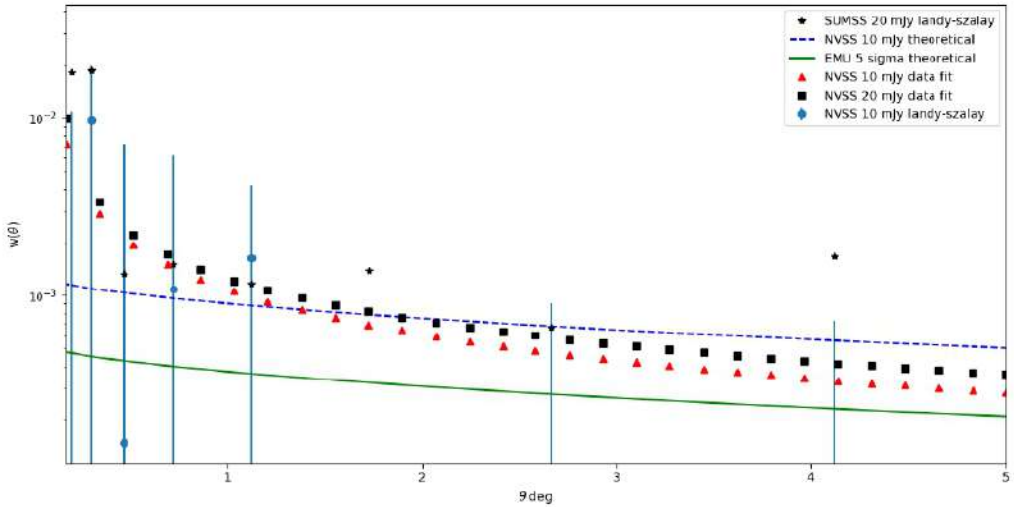


Figure 6: Comparison of autocorrelation function (ACF)'s from EMU 5 sigma, NVSS theory, NVSS power law fits from Blake 2002, and NVSS and SUMSS ACF's measured using AstroML's Landy-Szalay estimator. Here, the results are scaled to focus on lower theta contribution.

Differential and integral source count.

In order to understand the distribution of sources and estimate confusion, magnification bias and other useful quantities, differential and integral source counts provide useful means (Condon 2007, Rahman & Iqbal 2019). Theoretically, we can calculate integral source counts from differential count power law distributions as [21] [5]:

$N(S>S_{min}) = \int n(S)\,dS$ Equation (6)

Here n(S) is the differential source count power law probability distribution in $Jy^{-1}\ Sr^{-1}$ as [21] [5]:

$$n(S) = \frac{dN}{dS} = kS^{-\gamma}$$ Equation (7)

For EMU-5 sigma sources using SKADS, we get an estimate of k ≈ 57.24 and γ ≈ 2.18 [4] [5].

Smin in equation (6) is can be the rms or any suitable the lower bound value based on the survey science goals and technical specifications.

Integral source counts from observational data are usually are used to fit a power law curve in the form of:

$$N(S) = CS^{-\alpha}$$ Equation (8)

This gives us 'α' which is not only useful in measuring the magnification bias part of equation (3).

Using SKADS database, we get α ≈ 1.18 an C ≈ 48.5 [4] [5].

In order to perform statistical error analysis for galaxy continuum surveys, the shot-noise measurements play an important part.

Shot noise estimates or measurements are required to calculate the signal to noise ratios, measure error bars, and obtain correct covariance matrices, especially in relation to the theoretical or observed 'Cl' values obtained during cross or autocorrelation studies.

We can define shot-noise as:

$$Shot\ Noise = \frac{\Delta\Omega}{N}$$

Where ΔΩ =observed area of the survey in steradian and N=number of sources observed in the total survey area. Shot-noise can be calculated from the number count per steradian (Ns), by using the simple relation:

$$Shot\ Noise = 1/Ns$$

Where Ns is the number of sources per steradian. Ns can be calculated using equation (8) by using the differential source count power law fit from equation (7).

Reference:

1 C. Blake, P. G.Ferreira, J. Borrill, 2004, Mon. Not. R. Astron. Soc.,351,923

2 M. Loverde, L, Hui, E. Gaztañaga,"Lensing corrections to features in the angular two-point correlation function and power spectrum", 2008, Physical Review D, vol. 77, Issue 2, id. 023512

3 R. Norris et al. 2011 https://ui.adsabs.harvard.edu/abs/2011PASA...28..215N/abstract (EMU description paper)

4 S.F. Rahman ,2015, Theoretical estimates of integrated Sachs–Wolfe effect detection through the Australian Square Kilometre Array Pathfinder's Evolutionary Map of the Universe (ASKAP- EMU) survey, with confusion, position uncertainty, shot noise, and signal-to-noise ratio analysis, CJP, Vol 93. No. 4, pp. 384-394

5 S. F. Rahman, M.J. Iqbal , 2019, Astronomy Reports 63, pages515–526 arXiv:1612.08226 [astro-ph.CO] https://link.springer.com/article/10.1134%2FS1063772919070072

6 A. Raccanelli et al., 2012, Cosmological Measurements with Forthcoming Radio Continuum Surveys, Mon. Not. R. Astron. Soc.,424, 801

7 Condon J.J. et al.,1998, Astrophys. J., 115,1693

8 D.N. Limber , 1953, Astropys. J.,117,134

9 Bianchini, F. et al., 2015, ApJ 802 64

10 A. Hojjati, L. Pogosiana , Gong-Bo Zhao, Journal of Cosmology and Astroparticle Physics, 2011, 08,005,w011, doi:10.1088/1475-7516/2011/08/005

11 C. Blake, T. Mauch, E. M. Sadler,2004, Angular clustering in the Sydney University Molonglo Sky Survey, Monthly Notices of the Royal Astronomical Society,

12 R.J. Wilman, et al., 2008, Mon. Not. R. Astron. Soc., 388, 1335

13 E. Jones, E. Oliphant, P. Peterson, 2001, http://www.scipy.org/.

14 C.R. Harris, et al., 2020, Array programming with NumPy. Nature 585, 357–362 . https://doi.org/10.1038/s41586-020-2649-2

15 B. Stölzner, A. Cuoco, J. Lesgourgues, M. Bilicki, 2018, .Phys. Rev. D 97, 063506

16 T. Brinckmann, J. Lesgourgues,2019, MontePython 3: boosted MCMC sampler and other features, Phys. Dark Univ.24, 100260, arXiv:1804.07261

17 Planck Collaboration, 2016b, A&A. 594, A21

18 S. Landy, A.S. Szalay, 1993, Astrophysical Journal v.412, p.64

19 G. Hinshaw , et al., 2013, Astrophys. J.Suppl. Ser., 208,19

Hands-on Examples:

1. Two point angular correlation function (ACF) using AstroML
2. Two point angular correlation function (ACF) using TreeCorr (better for the datasets requiring masking of regions)