Report on the Ph.D. thesis entitled "Environmental dependence of galaxy properties using marked statistics" by Unnikrishnan Potty Sureshkumar (M. Sc.)

The doctoral dissertation presented by the candidate, M.Sc. Unnikrishnan Potty Sureshkumar, deals with the well known *"nature vs. nurture"* problem which emerge in many aspects of the study of the large scale structure of our Universe. More specifically, in his thesis the candidate tries to address and scrutinize in a physically well-based and statistically-strong quantitative way a more clear connection between the "large-scale" clustering properties of gravitational structures (in this case, galaxies) and the "local" environmental and physical properties of the same structures. As it is reported in the dissertation, such (cor)relation is well known and studied in literature, but this work has some new features which make it important for the development of the field:

- its results are based on the deepest and most complete (both in volume and magnitude) galaxy survey data up to date, reaching a redshift of ~ 0.4 and a limiting magnitude $r < 19.8$ mag;
- his work makes an extensive use of a relatively new tool for the analysis of the galaxy clustering properties, the marked correlation function, which is not yet widespread in the scientific community, although its positive properties are shown here quite clearly. Indeed, it allows to literally weight the dependence of the clustering properties on different and separated physical "marks", giving a more clear insight into the problem;
- assessing the validity of the marked correlation function is quite important also in light of the huge amount of data which we will get in the very next future from the most detailed surveys which we have ever run. This thesis clearly shows that it should be used preferentially in any present and future analysis to gain some real new deep insight into the problem.

The thesis is 134 pages long and consists of: a short Introduction, in which most of the concepts and methodologies which will be exposed in the thesis are generically introduced; Chapter 2, describing the data sets which have been used for the analysis; Chapter 3, which is a very nice and useful technical and fully-comprehensive compendium of the main core elements which are employed by the candidate in his analysis; four main Chapters $(4-7)$ displaying the main results of his research activity; a Conclusion section; and an Appendix which complements some information given in the text. The bibliography is very rich and contains ~ 230 entries.

The results from Chapters 4, 5 and 6 are based on two journal papers (one of them already published and another one still under peer review, at the date in which this report is signed) and two conference proceedings, of which the candidate was co-author together with his Ph.D. supervisor, prof. dr. hab. Agnieszka Pollo, his auxiliary supervisor, dr. Anna Durkalec, and with other researchers which I understand are mostly connected to the projects/surveys whose data have been used in this works (the GAMA survey, more specifically). The papers are relatively new and for that reason not yet cited as they might deserve, so that I will not comment on this aspect. In Chapter 7, instead, new results are reported which have not yet been submitted to any scientific journal.

In the following, I will explicitly discuss, organizing them by chapters, some main criticisms and comments which I would like to be clarified by the candidate. To start with, I have a general comment which I guess fits well at this point: I have found in many places in this work 1) short but really net and presumably "self-explanatory" statements and 2) long list of references, both given without any detailed explanation or a clear link to a broader context. I understand this is a doctoral dissertation, so it is a quite specialistic work, but it would have been probably good to add more information to guide the reader.

I will just give a couple of examples here to state clearly what I mean. Example 1: on pag. 7, the author writes: *"These observations are consistent with the framework of the ΛCDM cosmology"*. Why? How? Maybe a couple of paragraphs would have been helpful. I understand that for someone specialized on the topic such a conclusion might sound trivial, but I would have appreciated more explanations. Example 2: last paragraph of pag. 9: there is a long list of references given, stating that they *"explored"* an equally long list of topics, but the author does not really give any detail about their final results, and how they connect with his work. I think that giving more (of course, selected) details would have been more helpful also to understand more clearly the weight, the novelty and the role of the works produced by him and which constitute the main core of this thesis.

1 Comments

1.1 Chapter 1

I am aware that cosmology is not the main goal and focus of this thesis, but in this first chapter there are some inconsistencies and mistakes which should be taken into consideration, to avoid any misleading by anyone reading this thesis.

- There is no homogeneity in the use of the terms "Universe" and "universe", which are both present. Given the topic and the standard cosmological context which is assumed here, I guess that the right version should be "Universe".

- Eq. (1.1) are more commonly defined as Einstein field equations (i.e. with plural), given its tensorial nature. Probably the singular was a typo.

- *"Applying Einstein field equations to FLRW metric..."*: actually, it works on the other way, i.e. you need to provide a metric to the field equations in order to solve them.

- The density ρ in Eq. (1.4) should be generally defined as the density of the stress-energy tensor. Thus, it refers not only to matter, but to radiation too (and to dark energy, although here this component is explicitly given as a cosmological constant).

- There is something not clear with Eq. (1.5). By dimensionless density parameters we generally refer to constants defined as:

$$\Omega_{i,0} = \frac{\rho_{i,0}}{\rho_{c,0}}$$

  where: the suffix 0 means they are evaluated at present time, or equivalently at redshift $z = 0$; i refers to which component is considered (matter, radiation, etc...); and the

critical density of the Universe today, $\rho_{c,0}$, is defined in Eq. (1.6). The function $\Omega_i(z)$, which is presumably reported in Eq. (1.5), instead, is correctly defined as:

$$\Omega_i(z) = \frac{\rho_i(z)}{\rho_c(z)} = \frac{H_0^2}{H^2(z)} \frac{\rho_i(z)}{\rho_{c,0}} ,$$

where

$$\rho_c(z) = \frac{3H^2(z)}{8\pi G} .$$

- As a consequence, there is misleading in what are the Ω_i appearing in Eq. (1.7). Are they meant to be the $\Omega_{i,0}$ or the $\Omega_i(z)$? Of course, Eq. (1.7) holds in both cases; my point is that a clear nomenclature and definition should be used.

- On pag. 6: instead of *"inflation theory"* I think it would be more correct to speak about an "inflationary paradigm" or "scenario". Many people in the cosmology community question the status of theory for cosmological inflation.

- On pag. 8: the author refers to simulations like UNIVERSEMACHINE and SHARK. Why to discard or not even mention other simulations? Many of them are available nowadays, with some probably even more detailed than those referred in the text. Is there any specific reason for that choice? If yes, which one?

- As a cosmologist, I have a question which is probably trivial for astronomers, but not for people who may want to use data for cosmological analysis purposes. Within this thesis, a fiducial cosmology is fixed. As cosmologists, we know that while at the background level differences among cosmological models might be smeared out (up to some level of accuracy which is given by uncertainties from data), at the perturbation level the things are more tricky, and similar models might behave very differently. And here we are dealing with clustering of gravitational structures, which is exactly connected to this latter point. I am thus wondering if the candidate has any idea of how much would weight the choice of another cosmological background, different from assuming a cosmological constant for example, on (some of) the results provided here.

1.2 Chapter 2

- GAMA II is introduced on pag. 13 without any previous description or specification in the text. Does it possibly refer to a second release? In Chapter 3.2 both GAMA II and the third GAMA data release are cited at the same time, so the candidate could clarify this point.

- On pag. 13 there is another case of a long list of references (penultimate paragraph, starting with *"There have been many studies in GAMA using 2pCF in the past..."*), about which the author does not specify the main outcomes, which would have been helpful to stress the main differences with the present work.

- I understand that the Outer Rim run, from which the cosmoDC2 synthetic galaxy catalogue is taken, is based on a different cosmological background than that set by the

candidate in this thesis. Although they both are ΛCDM, they have different values for most, if not all, their corresponding cosmological parameters. Can this lead to any bias in the performed analysis? Should it be taken into account, or not?

1.3 Chapter 3

This chapter is particularly appreciated because it provides a short but very clear and fully-exhaustive introduction to all the steps which the author (or anyone working on the topic) has to take in order to proceed with such type of analysis.

- At the beginning of Chapter 3.3 it is written: *"To have a meaningful comparison, it is essential that this random sample reflects the same sky distribution and redshift distribution of the real galaxy sample."* While the sky distribution is (very likely) largely cosmological model-independent, the redshift distribution is crucial for calculating distances among galaxies. But distances are cosmological model-dependent. And the fiducial cosmological model in Farrow et al. 2015 is different from the one used here. Has this point be considered somewhere in the analysis? How does it affect the results?

- The Landy & Szalay estimator is said to be *"often preferred"*. Although one reason to prefer it with respect at least some of the other estimators which have been defined so far is provided in the Appendix, I am wondering how much general are the results based on its use, and which differences would be expected from other estimators (if meaningful).

- Looking at Fig. 3.4, I am wondering why the author has chosen to consider $N_r = 5 \times N_d$, when it is clear that for such a value there is still some variation mostly at small scales, while for $N_r/N_d > 10$ there is practically no difference when changing the values. Is the reason just due to computational time? Do you expect to have negligible differences if a different choice would have been made? Did you make a check of that?

- On pag. 29 it is written that *"the projected 2pCF can be used to recover the real space 2pCF devoid of RSD"*. I guess it is just a matter of rephrasing/terms, but I would like to have a clarification about the meaning of such a statement. I understand that what is done is:

  - to calculate/measure the real space 2pCF (with RSD), $\xi(r_p, \pi)$;
  - to integrate it (numerically) to get the projected 2pCF, $\omega_p(r_p)$;
  - to assume that the real space 2pCF without RSD is given by the power law Eq. (3.3), so that $\omega_p(r_p)$ can be analytically given by Eq. (3.10);
  - to use $\omega_p(r_p)$ to constrain the parameters r_0 and γ.

  Thus, the last point is what is meant for "recovering the real space 2pCF". Am I correct?

- I have a question about the power law Eq. (3.3): is this the best (after comparison with data) model emerging from literature? Are there alternatives, pointing to departures from it, maybe at small scales?

- From pag. 33 it is not clear how the χ^2 is minimized. Is the author using a grid? In such a case, how much fine is the grid and how do you estimate the errors on the parameters? Is he using a Monte Carlo method? In such a case, how he tested that the statistics is reliable?

- On pag. 35: I think that the correct definitions are *eigenvalues* and *eigenmodes*, not with separated words;

- In Fig. 3.13 and 3.14, some contours plots are cut in r_0. Is this just a "visual" cut, or it is due to a limited range in the grid which is used to minimize the χ^2? In the latter case: is it taken into account when the errors on r_0 are estimated?

- On pag. 47, Eq. (3.23): I guess that under the square root at the denominator the variances σ^2 should appear.

1.4 Chapter 4

- On pag. 52 it is stated that the J and K-band magnitudes are extrapolated. Given that extrapolation procedure is always tricky (leading to "diverging" estimated values, exploding errors, etc...), I would like to ask how much reliable are such extrapolations in this specific case.

- On pag. 54, when the real and the random catalogues are compared, also referring to Fig. 4.2, it is stated that *"the agreement between the redshift and sky distributions of the real and random samples* are confirmed. Looking at Fig. 4.2 one might claim that the agreement is not so perfect. May I ask how, technically, this agreement is quantified?

- On pag. 56, from the caption of Fig. 4.4, I understand that the parameters (r_0, γ) are derived from a fit of the projected 2pCF function, $\omega_p(r_d)$. Could they be derived from M_p which I understand should also have much smaller (relative) errors?

- On pag. 58 and 61: why the J-band is not mentioned as a good proxy? It seems to me that is it as much good as K-band.

- On pag. 60 and in the conclusions of Chapter 4, the stellar mass is considered as the galaxy property with the strongest environmental dependence. But I have understood that both values of MCFs > 1 and < 1 are good tracers for environmental dependence, while ~ 1 is not. And it seems that SFR is as good as stellar mass (1.4 vs 0.5 in MCFs). So, both of them should be considered as the best ones. Is that right?

1.5 Chapter 5

- There is no uniformity in referring to the limiting magnitude of the GAMA sample: although being always the same quantity (Petrosian), in Chapter 2 it is used simply r, while in Chapter 5 it is given as r_{petro}.

- What are the specific criteria for choosing the sub-samples $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}, \mathcal{M}, \mathcal{N}, \mathcal{P}, \mathcal{Q}, \mathcal{R}, \mathcal{S}$? I guess that answering this question will also answer the following: is it not possible to

define the same sub-samples (considering that there is already some overlapping) to be used once and for all through the full analysis?

- I have probably missed it, but I have not seen any discussion about the maximum anti-correlation shown in Fig. 5.6 at $\sim 0.1 h^{-1}$ scales, when the SFR is used as proxy, as well as the maximum correlation in M_{WISE} and $M_{PROSPECT}$, at the same scales, from Fig. 5.7. Are not they important?

1.6 Chapter 6

I have no specific comments on this chapter.

1.7 Chapter 7

- On pag. 104: it is said that 12 jackknife samples are used to estimate uncertainties. Is 12 a reasonable number? Should not it be larger?

- On pag. 105: what is meant by *"random scrambling method"*?

- The final main conclusion from this chapter is that the virtual cosmoDC2 catalogue does not reproduce the real clustering properties which can be derived from GAMA. They are actually very different. What should then be concluded about cosmoCD2? Should it be disregarded? I understand that more studies must be performed, but is there any clue about the reasons for such a discrepancy?

2 Conclusions

The comments and the questions which I raised in the previous sections are not meant to undervalue this doctoral thesis, are just formal requirements. For what concerns what matters, i.e. the content, in my opinion the thesis fulfills all the necessary requirements to be presented for the doctoral degree so that I recommend the admittance of M.Sc. Unnikrishnan Potty Sureshkumar for the defence.

Szczecin, 07.07.2022
dr hab. Vincenzo Salzano, prof U.S.
Institute of Physics, University of Szczecin