The thesis of Mr Szymona Nakonecznego focuses on building good-quality quasar catalogues from the Kilo-Degree Survey (KiDS) imaging survey using various machine learning techniques. Overall, I find the main results presented in the thesis reliable and highly valuable for the comminity. The writing is concise and to the point. The work presented in the thesis resulted in two first-author publications and a final science chaper which could be submitted to a science journal after moderate revisions. Below I have some detailed comments and suggestions on the individual chapters in the thesis.

In Chapter one, an introduction to the thesis is given, starting from the importance of quasars as a large-scale structure (LSS) tracer, how to build large catalogues of quasars and estimate their key properties, to the various datasets and methodologies used in this thesis. In my opinion, this introduction chapter could be made much more extensive. For example, a brief review on previous key studies of quasar clustering (including dependence on redshift and physical properties such as luminosity) should be included. Similarly, adding a brief review on the connection between quasars (as a special phase in the evolutionary history of galaxies) and other galaxy populations would be very beneficial. Furthermore, a review focused on previous studies of building quasar catalogues from photometric data and their pros and cons would be highly relevant.

In Chapter two, an overview of the datasets and methodology used in this thesis is given. The data part is mainly divided into KiDS DR3 and DR4. The construction of the inference sets from DR3 and DR4 is explained in some detail. The mismatch in the imaging depth between KiDS and the training dataset from SDSS is highlighted. It would be good to add some basic information such as areal coverage of KiDS DR3 and DR4, and comparisons with other major ground-based optical imaging surveys (such as DES and HSC) in terms of depths and spatial resolution.

The training set comes from the SDSS DR14 spectroscopic sample, therefore the selection biases inherent in SDSS DR14 would be transferred to the quasar catalogues constructed from the KiDS data. I think it would be helpful to discuss the types of quasars likely to be missed in this approach, which should have been studied (at least partially) already in previous publications.

Here it is mentioned that the WISE data could be very helpful in classifying quasars and estimating photo-zs. Due to the relatively small fraction of KiDS sources matched with WISE, the WISE data is eventually not used. In future work, it would be interesting to see quantitatively how the addition of the WISE data can help with the classification and photo-z estimation. In future work, it could also be interesting to explore combining images (or morphological and structural parameters derived from optical imaging, in addition to the stellarity index already used in the current approach) with magnitude-based features in the classification of quasars.

The methodology part of Chapter two can be divided into two components, one on machine learning and the other on correlation analysis. The machine learning part introduces the specific models used in the thesis, feature engineering and validation procedures, for both DR3 and

DR4. The use of unsupervised ML methods such as the t-SNE method is very effective and helpful in visualising limitations in the training data and the motivation of emplying various selection cuts. The correlation analysis part is very brief.

Chapter three presents the construction of a quasar catalogue using the KiDS DR3. The feature importance study is interesting and gives insights into the most relevant features in identifying quasars. The tests from the three ML methods give very comparable results. The investigation into potential reasons for misclassification reveals interesting trends with redshifts. Potentially including redshifts will help reduce misclassifications. I understand photo-z information is however not included in the analysis as the KiDS photo-z estimates are not optimised for quasars as explained in the thesis. In the future, it might be interesting to see if the photo-z estimation step could be improved to be opmised for both galaxies and quasars (for example by using better SED templates) and then use the photo-z info to improve the quasar classification. This chapter also includes various validation tests for better understanding of the reliability and completeness of the quasar catalogue. A recommendation on the probability threshold is given based on these findings. I find these tests very helpful in convincing the reader of the quality of the catalogue. I have a suggestion for an additional test which is to use radio data (some of the KiDS fields have LOFAR radio imaging for example), because some quasars would have radio emission but radio stars are very rare.

Chapter four presents extended results using the KiDS DR4 and combining optical photometry with near-infrared photometry. The design of a faint extrapolation test for overfitting is very useful. The three ML methods also give fairly comparable performances for the classification task but ANN is shown to be clearly the best method for redshift estimation. Two ANNs are constructed to perform quasar classification and redshift estimation separately. I think this is an interesting approach giving the best results in both tasks. The investigations into purity and completenes of the resulting quasar catalogue (as well as redshift error) and dependence on r-band magnitude give credible and very useful results.

The last science chapter Chapter five presents a first look into the correlation functions and tentative bias constraints using the quasar catalogues constructed from the KIDS DR4. Auto-correlations of the quasars and cross-correlations between quasars and CMB lensing maps are measured. I think this analysis is very interesting as a quick first look, but clearly more work is needed to refine the analysis and turn this into a publication eventually. For example, in measuring the auto-correlation functions, selection effects which can result in variations of the source density across the sky are very important to be taken into account properly. In comparison, the angular selection effects in measuring the cross-correlation functions are not so important as long as these effects from the quasar catalogue and the CMB lensing maps are not correlated. In the next step, the PhD candidate should discuss in detail the angular selection effects and make sure the measured correlation functions are not affected by this. This is probably the most important step in making sure the correlation functions presented are convincing. In addition, it would also be interesting to show the correlation functions in bins of redshift.

At the end of this chapter, analysis of the bias factor for the quasars is presented very briefly. At the moment, it is difficult to fully understand this part as many details are missing. For example, how is the bias model as a function of redshift actually fitted? Do you assign a bias factor to each quasar depending on its redshift? What is the reduced chi2? The model fit, by eye, is not a very good description of the data. Do you have any explanations for that? The comparison with Sherwin et al. (2012) has minimal information. What sample was used in Sherwin et al.? How does their analysis method differ from this analysis? Is the comparison fair? Clustering of quasars has been studied extensively in the literature and so a much more extensive comparison with previous results is sorely needed. These are all important points to be addressed in future refinement of the work presented in the last science chapter of this thesis..

In summary, my view is that the thesis meets all the legal and customary requirements for PhD dissertations and I propose that Mr. Nakoneczny be admitted to further stages of his doctoral procedure.

Lingyu Wang

05/09/2022