Machine Learning in Astronomy: Possibilities and Pitfalls
Proceedings IAU Symposium No. 368, 2025
J. McIver, A. Mahabal & C. Fluke, eds.
doi:10.1017/S1743921322004872

Detecting complex sources in large surveys using an apparent complexity measure

David Parkinson[1] **and Gary Segal**[2,3]

[1]Centre for Theoretical Astrophysics, Korea Astronomy and Space Science Institute, 776 Daedeok-daero, Yuseong-gu, Daejeon 34055, Republic of Korea
email: DavidParkinson@kasi.re.kr

[2]School of Mathematics and Physics, University of Queensland, St Lucia, Brisbane, QLD 4072, Australia

[3]CSIRO Space and Astronomy, PO Box 76, Epping, 1710, NSW, Australia
email: gp.segal@gmail.com

Abstract. Large area astronomical surveys will almost certainly contain new objects of a type that have never been seen before. The detection of 'unknown unknowns' by an algorithm is a difficult problem to solve, as unusual things are often easier for a human to spot than a machine. We use the concept of apparent complexity, previously applied to detect multi-component radio sources, to scan the radio continuum Evolutionary Map of the Universe (EMU) Pilot Survey data for complex and interesting objects in a fully automated and blind manner. Here we describe how the complexity is defined and measured, how we applied it to the Pilot Survey data, and how we calibrated the completeness and purity of these interesting objects using a crowd-sourced 'zoo'. The results are also compared to unexpected and unusual sources already detected in the EMU Pilot Survey, including Odd Radio Circles, that were found by human inspection.

Keywords. Extragalactic radio sources; Astrostatistics; Astrostatistics tools

1. Introduction

We are entering an unprecedented era of astronomical data. In the radio wavebands the pathfinders to the Square Kilometer Array will conduct wide and deep surveys. The Australian Square Kilometer Array **Pathfinder** (ASKAP) is predicted to detect ~ 70 million radio galaxies with the Evolutionary map of the Universe (EMU) survey (Norris *et al.* 2011), in comparison to the ~ 2 million extragalactic radio sources currently detected. In the optical, the Legacy Survey of Space and Time (LSST), which will be conducted with the Vera Rubin Observatory (VRO) predicts a detection of 20 billion galaxies and a similar number of stars (Ivezić *et al.* 2019), in comparison with the 300 million galaxies and 80 million stars in the Dark Energy Survey DR1 (Abbott *et al.* 2018), and the 1.6 billion objects in GAIA DR2 (Gaia Collaboration *et al.* 2018).

Many of these objects detected by these new surveys will be of a type never before seen by astrophysicists. This work focuses on a measure of morphological complexity that can be used to identify new and interesting objects. The approach does not requiring learning the specific features of past interesting observations, and as such, can be calibrated on a small sample and then applied to a much larger sample. This method differentiates itself from many existing outlier methods as it does not require the computationally intensive task of computing the distance between specific observations and the rest of the data **in an abstract feature space.** Rather complexity can be computed for **a** single observation and compared to a pre-determined threshold.

© The Author(s), 2025. Published by Cambridge University Press on behalf of International Astronomical Union.

https://doi.org/10.1017/S1743921322004872 Published online by Cambridge University Press

2. Methodology

2.1. *Coarse-grained complexity*

We introduce the coarse-grained complexity in Segal *et al.* 2019 based on the notions of *effective complexity* (Gell-Mann 1994, Gell-Mann & Lloyd 1996) and *apparent complexity* (Aaronson, Carroll & Ouellette 2014). The apparent complexity is a measure of the entropy H of an object x computed after applying a smoothing function f, expressed as $H(f(x))$. The Shannon entropy of a probability distribution P can be defined as the expected number of random bits that are required to produce a sample from that distribution:

$$H(P) = -\sum_{x \in X} P(x) \log P(x). \tag{1}$$

As discussed in Aaronson, Carroll & Ouellette 2014, Segal *et al.* 2019, the Kolmogorov complexity $K(f(x))$ can be used as a proxy for the entropy of the smoothed function $H(f(x))$. The Kolmogorov complexity of x is the length of the shortest binary program $l(p)$, for the reference universal prefix Turing machine U, that outputs x; it is denoted as $K(x)$:

$$K(x) = \min_p \{l(p) \ : \ U(p) = x\}. \tag{2}$$

While the Kolmogorov complexity is uncomputable, its upper bound can be reasonably approximated by the compressed file size $C(f(x))$ using a standard compression program such as `gzip`.

Intuitively a complexity measure should provide low values for random data that does not contain structure or regularities. The coarse-grained complexity, like the apparent and effective complexity, can be defined as the compressed description length of regularities and structure after discarding all that is incidental. The coarse-grained complexity achieves this by applying a smoothing function f to the input x (which removes fine-grained noise while preserving the coarse-grained structure of the image) and calibrating this so that complexity values correctly partition data that has been expertly labelled (i.e. by human astronomers).

2.2. *Complexity Scan*

Source detection algorithms may make assumptions, and so miss interesting things. In- stead, we scan the region, computing the course-grained complexity within a sliding frame (square aperture, as shown in Fig. 1).

An important feature of the scan method is that frames are sampled without making any assumptions as to whether the frame contains a conventionally detected source or not (that is without using a source extraction tool or existing catalogue data). This helps reduce the risk of producing a sample that is biased towards preconceived notions of what is interesting, referred to as expectation bias by Norris 2017 and Robinson 1987. This process builds up a 'heat map' of complexity values intended to assist with the identification of the unexpected in new and large data with the goal of new scientific discoveries and surprise.

3. Data

3.1. *EMU Pilot Survey*

The Pilot Survey of the Evolutionary Map of the Universe (EMU-PS) was observed at 944MHz using the Australian Square Kilometre Array Pathfinder (ASKAP) telescope. The ASKAP telescope consists of 36 12-metre antennas spread over a region 6-km in diameter at the Murchison Radio Astronomy Observatory in Western Australia.

https://doi.org/10.1017/S1743921322004872 Published online by Cambridge University Press

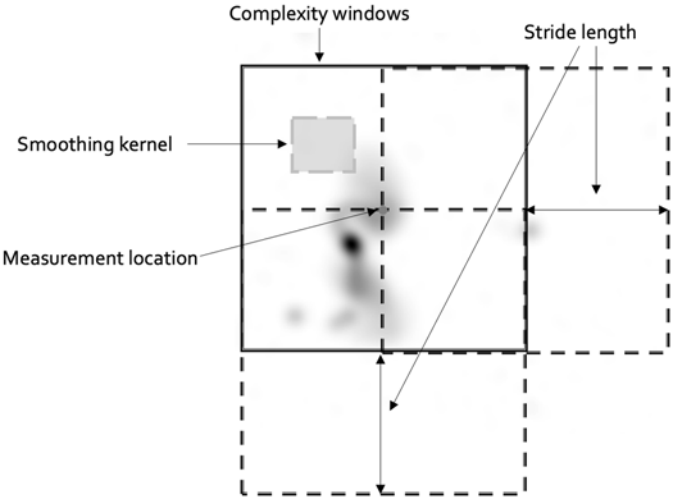


Figure 1. Illustration complexity scan parameters, including: the frame size (side length of solid black square), stride length (offset between position of solid black square and dashed black squares) and smoothing kernel size (grey shaded rectangle).

EMU- PS covers 270 square degrees of an area covered by the Dark Energy Survey at a spatial resolution of $\sim 11 - 13$ arcsec (Norris *et al.* 2021a).

While the primary goal of the EMU Pilot Survey is to test and refine observing parameters and the strategy for the main survey, the pilot in itself presents opportunity for new discoveries. Experience has shown (Norris 2017) that whenever we observe the sky to a significantly greater sensitivitywe make new discoveries. This goal has already been demonstrated through the successful identification of a new class of radio object, odd radio circles (ORCs, Norris *et al.* 2021b, Koribalski *et al.* 2021, Norris *et al.* 2022).

3.2. *Anomaly Zoo*

Truth labels are required to evaluate the effectiveness of alternative complexity thresholds for partitioning anomalous sources. To provide truth labels for the frames produced by the complexity scan we ran a project on the `Zooniverse.org` platform, titled *"Anomaly in the EMU Zoo"* (hereafter zoo), requesting expert astronomers to evaluate an unbiased sample of frames sub-sampled from the EMU-PS scan. Consensus from the zoo labels was then used to evaluate the Recall and Precision associated with prospective partition boundaries.

Expert volunteers were approached from within the Evolutionary Map of the Universe Survey Project and at the SPARCS 2021 conference. A sub-sample of 1627 frames from the EMU-PS scan ($n_{\rm total} = 365{,}000$) were presented to volunteers for classification through the Zooniverse project. 44 volunteers participated in the project, with 10 of these classifying more than 500 frames each.

The zoo asked the expert volunteers to evaluate frames sub-sampled from the EMU-PS Scan and to select an option that best describes the most interesting radio sources in each frame before moving on to the next. The four options presented for selection were:

- No sources/just noise
- One or more simple sources/unrelated simple sources
- At least one complex source/sources with multiple components
- Contains something unexpected/Anomaly

https://doi.org/10.1017/S1743921322004872 Published online by Cambridge University Press

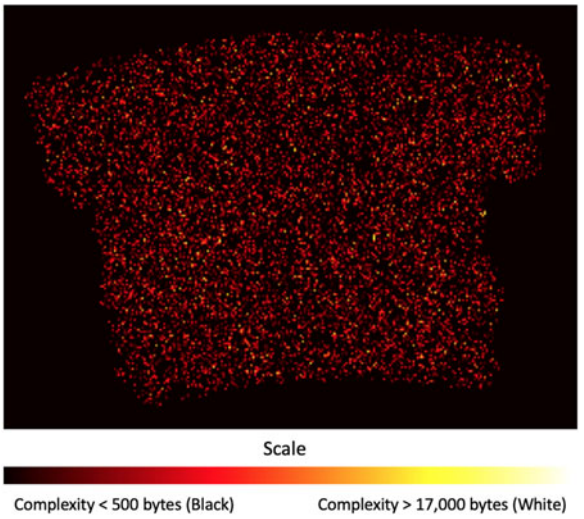


Figure 2. A complexity 'heat map' of the EMU-PS region, where the complexity of the frame is indicated by the brightness of the pixel. While many of the frames are red (grey) coloured, indicating the presence of a simple source, there are a few yellow (light grey) or white pixels, indicating more complex sources or low-surface brightness structures.

The zoo distinguished between complex and extended sources with multiple components, and sources that were deemed by the volunteers to be truly unexpected or anomalous. Only frames converging on a label through majority consensus were used to evaluate the effectiveness of the complexity to identify anomalous sources. Those frames from the zoo sample where majority consensus was not reached were excluded from the evaluation.

4. Results

4.1. *Complexity Heat Map*

We performed a complexity scan of the EMU-PS data using the method described. Com- plexity values are shown in Fig. 2 as a heat map overlaid on the EMU-PS field.

Visual inspection of a sub-sample of frames shows that the high-complexity value tail of the distribution comprises unusual, complex and extended objects with examples shown in Fig. 3. These include peculiar Tailed Radio Galaxies, **Fanaroff-Riley Class I (FRI) and Class II (FRII)** type sources with more inflated lobes than that of average such objects, Odd Radio Circles, and Giant Radio Galaxies including a previously unreported Giant Radio Galaxy **with projected linear length** of 1.94 Mpc.

4.2. *Anomaly Zoo Outcomes*

We use truth labels determined from the zoo to evaluate the effectiveness of alternative partitions for identifying anomalous objects. As an immediate benefit, an effective partition can be used to identify frames containing complex structures and unusual objects from the EMU-PS data and build an anomaly catalogue. An effective boundary can also be used when analysing new, and even larger surveys, including the subsequent full EMU survey which is anticipated to capture approximately 40 million sources.

Through sub-sampling of the EMU-PS Scan, the sample size for the zoo was limited to 1,627 frames. An enrichment sample was included to provide better representation of the type of observations found within frames beyond the 99.5th percentile. Classification

https://doi.org/10.1017/S1743921322004872 Published online by Cambridge University Press

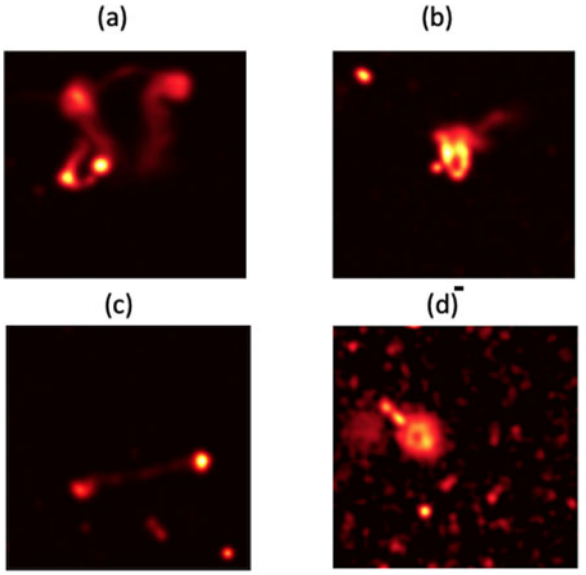


Figure 3. Examples of objects found within frames sampled above the 99th percentile complexity value from the EMU-PS Complexity Scan. These examples illustrate the breadth of objects found within frames in the complexity tail including, (a) the unusual source PKS 2130-538 (Otrupcek & Wright 1991) known as the Dancing Ghosts (Norris *et al.* 2021a), (b) a bright wide angle tail source on 2MASX J21291901-5053040 in cluster Abell 3771, (c) a previously unreported Giant Radio Galaxy **with projected linear length** of 1.94 Mpc (DES J221640.02-625241.6), (d) two odd radio circles, EMU PD J205842.8–573658 (ORC2) and EMU PD J205856.0-573655 (ORC3).

Table 1. Classification counts by consensus for each of the options provided in the EMU-PS Anomaly in the EMU Zoo. The enrichment sample supplements the zoo sub-sample with frames measuring complexity above the 99.5th percentile.

Classification	**Count: Non-bias sample (n = 1528)**	**Count: Enrichment (n = 99)**	**Count: Combined Sample (n = 1627)**
Anomalous	7	17	24
Complex	366	57	423
Simple	1059	12	1071
No source	31	0	31
No consensus	65	13	78

counts for each of the zoo classes are shown in **Tab**. 1. This table includes the anomaly counts both before and after the inclusion of the enrichment sample.

Only frames converging on a label through majority consensus were used to evaluate the effectiveness of the complexity to identify anomalous sources. Those frames from the zoo sample where majority consensus was not reached were excluded from the evaluation. 99.8% of the zoo sample frames received evaluations from 3 or more volunteers and 94% from 5 or more. We used the criterion that only frames receiving 5 or more evaluations were used to evaluate consensus, and be given a reliable 'truth' label. These restrictions were imposed to avoid the results being impacted by outlier evaluations that differed from the majority of expert volunteers.

We partition the data based on complexity and **signal-to-noise ratio (SNR)** values and use this to define a function based catalogue boundary (shown in Fig. 4).

https://doi.org/10.1017/S1743921322004872 Published online by Cambridge University Press

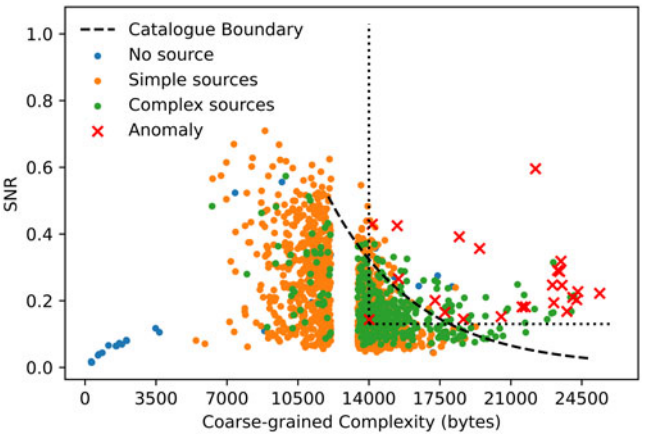


Figure 4. Scatter plot showing the coarse-grained complexity and the signal-to-noise ratio (SNR), and the relationship between them, along with the classification of frames derived through consensus from the zoo (Anomaly in the EMU Zoo).

5. Summary

The coarse-grained complexity can be used as a tool for identifying unusual and complex objects. We apply the method to new EMU-PS data (365,000 sampled frames containing at least the 220,000 Selavy catalogue sources) to identify and segment unusual sources (i.e. anomalies).

We used a Zooinverse project to produce crowd-sourced labels to evaluate the effectiveness of the approach and we identified an effective anomaly partition using the coarse-grained complexity and SNR values.

Results demonstrate the ability of the coarse-grained complexity to single out regions of the sky that contain complex and unusual sources, in a manner that can be computed at worst-case linear time complexity without reference to existing catalogue data. We propose the complexity measure as useful tool for identifying regions of interest in subsequent large and deep radio continuum surveys.

References

Aaronson S., Carroll S., Ouellette L., 2014, http://arxiv.org/abs/1405.6903v1
Abbott, T. M. C., Abdalla, F. B., Allam, S., Amara, A., Annis, J., *et al.* 2018, *ApJS* 239, 18
Gaia Collaboration *et al.* 2018, *A&A* 616, A1
Gell-Mann, M., 1994, *"The Quark and the Jaguar: Adventures in the Simple and the Complex,"* Henry Holt and Company
Gell-Mann, M. & Lloyd, S. 1996, *Complexity*, 2, 1
Ivezić, Ž, Kahn, S. M., Tyson, J. A., Abel, B., *et al.* 2019, *ApJ* 873, 2
Koribalski, B. S., Norris, R. P., Andernach, H., Rudnick, L., Shabala, S., Filipović, M., Lenc, E. 2021, *MNRAS Letters*, 505, 1
Norris, R. P., Hopkins, A. M., Afonso, J., Zinner, *et al.* 2011, *PASA*, 28, 3
Norris, R. P. 2017, *PASA*, 34, e007
Norris, R. P. *et al.* 2021a, *PASA*, 38, e046
Norris, R. P. *et al.* 2021b, *PASA*, 38, e0003
Norris, R. P. *et al.* 2022, *MNRAS*, 513, 1
Otrupcek, R. E. & Wright, A. E. 1991, *PASA*, 9, 1
Robinson, B. J. 1987, *PASA*, 7, 220
Segal, G., Parkinson, D., Norris, R. P., Swan, J., 2019, *PASP* 131, 1004

https://doi.org/10.1017/S1743921322004872 Published online by Cambridge University Press