Identifying Complex Sources in Large Astronomical Data Sets Using a Coarse-grained Complexity Measure

Gary Segal[1,2], David Parkinson[1,3] , Ray P Norris[2,4], and Jesse Swan[2,5]

[1] School of Mathematics and Physics, University of Queensland, St Lucia, Brisbane, QLD 4072, Australia; g.segal@uq.edu.au
[2] CSIRO Astronomy and Space Science, PO Box 76, Epping, 1710, NSW, Australia
[3] Korea Astronomy and Space Science Institute, Daejeon 34055, Republic of Korea; davidparkinson@kasi.re.kr
[4] Western Sydney University, Locked Bag 1797, Penrith South, 1797, NSW, Australia; Ray.Norris@csiro.au
[5] School of Natural Sciences, University of Tasmania, Private Bag 37, Hobart 7001, Australia

Abstract

The volume of data that will be produced by the next generation of astrophysical instruments represents a significant opportunity for making unplanned and unexpected discoveries. Conversely, finding unexpected objects or phenomena within such large volumes of data presents a challenge that may best be solved using computational and statistical approaches. We present the application of a coarse-grained complexity measure for identifying interesting observations in large astronomical data sets. This measure, which has been termed apparent complexity, has been shown to model human intuition and perceptions of complexity. Apparent complexity is computationally efficient to derive and can be used to segment and identify interesting observations in very large data sets based on their morphological complexity. We show, using data from the Australia Telescope Large Area Survey, that apparent complexity can be combined with clustering methods to provide an automated process for distinguishing between images of galaxies which have been classified as having simple and complex morphologies. The approach generalizes well when applied to new data after being calibrated on a smaller data set, where it performs better than tested classification methods using pixel data. This generalizability positions apparent complexity as a suitable machine-learning feature for identifying complex observations with unanticipated features.

Key words: methods: statistical – techniques: image processing – radio continuum: galaxies

Online material: color figures

1. Introduction

The universe is very large, and we have only just begun to scratch the surface in terms of identifying the different objects and events that it contains. Each generation of instrumentation and infrastructure has expanded our knowledge and sample size significantly, often resulting in unexpected scientific results. For example, of the 10 greatest discoveries by the *Hubble Space Telescope*, only one was listed in its key science goals (Norris 2017a). The next generation of astrophysical instruments will be collecting petabytes of data per day. This represents a significant opportunity for making unplanned discoveries due to the large volume of observational data that will be made available, but is beyond the limit for the amount of information that can be examined directly by the human astronomical community on any reasonable timescale.

One good example of this expansion in surveying and collection capability is in the area of extragalactic radio astrophysics. We currently know of about 2.5 million extragalactic radio sources, but future surveys will increase this number by several orders of magnitude. The Australian Square Kilometre Array Pathfinder (ASKAP) and the Evolutionary Map of the Universe survey are predicted to increase this number to about 70 million (Norris 2017b) in the continuum. The Square Kilometre Array (SKA) may increase this number into the billions (Jarvis et al. 2015).

Progress that has been made with supervised machine-learning approaches such as convolution neural networks has demonstrated the potential effectiveness of these approaches for identifying and classifying observations in astronomical surveys based on their features (Aniyan & Thorat 2017; Karpenka et al. 2013; Kim & Bailer-Jones 2016; Kessler et al. 2010; Dieleman et al. 2015; Huertas-Company et al. 2015; Charnock & Moss 2017; Alger et al. 2018). Supervised approaches may become highly effective at identifying observations that have been previously considered interesting, utilizing features associated with the interesting observations used for training and testing, but overlooking new observations whose features have little in common with past interesting observations. Such an approach can be said to suffer from an expectation bias, discussed by Norris (2017a) and Robinson (1987). Unsupervised learning methods, or the use of selected features that have been demonstrated to generalize well, may

therefore be considered preferable candidates for segmenting new observations and detecting the unexpected.

A good example of an unsupervised learning approach for finding outliers inside the ensemble was the use of random forests by Baron & Poznanski (2017). Here they utilized the correlation structure in feature space to identify interesting observations using learned features based on the correlation structure of galaxy spectra. They used random forests to identify interesting features in galaxy spectra by learning the difference between real and synthetic observations, where the correlation structure was removed from the feature space in the synthetic case. Outliers were then identified based on a measure of similarity between objects, where the authors counted how often every pair of real objects was classified as real in the same leaf of a given tree. This approach demonstrates the use of learned interesting features, but learning not based on previous observations, to perform subsequent outlier detection. A potential drawback of this approach, as with many other unsupervised outlier detection methods, is the computational burden of having to compute pairwise comparisons. The use of selected features that generalize well presents a computationally efficient option when applying machine-learning approaches to segment the large volumes of data being processed in the early stages of the image analysis pipeline.

This paper presents the application of a coarse-grained complexity measure for identifying observations with complex and interesting morphologies in large astronomical data sets. Aaronson et al. (2014) have demonstrated the effectiveness of a coarse-grained complexity measure, termed apparent complexity, at capturing human intuition and perceptions of complexity. This measure provides a quantitative description of a notion of complexity informally proposed by Gell-Mann (1994) as a phenomenon that first increases and then decreases with the rising entropy of a closed system. A potential drawback of apparent complexity as a formal description of complexity is that it relies on assumptions regarding human perceptions. Conversely, this very connection to human perceptions suggests that this measure should be effective at identifying complex observations that are likely to be of interest to a human observer. This paper shows that, by learning an appropriate smoothing function, the apparent complexity can be used to partition a sample based on the morphological complexity and how interesting the observations are.

We present the application of apparent complexity as a machine-learning feature, which can be used by itself or together with the signal-to-noise ratio (S/N) to effectively identify radio galaxies with complex morphologies in the presence of noise. We show that the measure generalizes well when applied to new and larger data sets with a more expansive range of varied morphologies. We also compare approaches using the measure to more traditional classification approaches, logistic regression, and support vector machines (SVMs) applied to pixel data, and show that the use of the complexity measure generalizes better when applied to new and larger data after being calibrated on a smaller data set.

Apparent complexity is fast and computationally efficient to compute, as the measure only requires applying a smoothing function and compression algorithm, both of which can be implemented at worst-case linear time complexity:

$$T(n) = O(n). \tag{1}$$

Implementations of constant time $O(1)$ and linear time $O(n)$ median filters are detailed in Perreault & Hebert (2007). Being able to leverage fast and efficient tools that generalize well is likely to be desirable when identifying interesting and unexpected observations in very large scientific data sets such as those that will be produced by ASKAP and the SKA.

We envisage a potential application for the apparent complexity measure in the early stages of the image analysis pipeline to segment very large data sets and identify smaller samples of complex radio sources that are appropriate for additional analysis requiring more computationally intensive methods or manual inspection. This is likely to include cross-matching with observations at other frequencies to determine whether and how the host galaxy is coincident with radio components. It may also serve in the early identification of new and unexpected observations.

Using data from the Australia Telescope Large Area Survey (ATLAS; Norris et al. 2006) we show that apparent complexity can be used to distinguish between images of galaxies which have been classified as having simple and complex morphologies. We also show that the approach generalizes well when applied to new data.

The paper is structured as follows: Section 2 frames the theory in terms of Kolmogorov complexity and discusses the theoretical merits of apparent complexity as an attractive candidate for identifying interesting astronomical observations in large data sets. Section 3 outlines empirical methods and results that show apparent complexity can be used to distinguish between images of galaxies with simple and complex morphologies. It will also show that a smoothing function calibrated on a small labelled sample with few interesting observations is able to generalize well when applied to a much larger sample containing a larger collection of complex morphologies. Testing on data containing a larger sample of simple and complex observations than that used for training, we show that clustering using the complexity measure generalizes better than traditional classification methods using pixel data. In Section 4 we summarize our conclusions.

2. Theory

In this paper we measure the complexity C of some observation x, subject to a function f that extracts only the non-incidental information from measurements. As such, the measured complexity will depend on the function f, which can

be calibrated to align with the interests and perceptions of the scientific observer, and so should be considered as $C(f(x))$. It is envisioned that the association between the computed measure $C(f(x))$ and the morphological complexity of observations will generalize well.

2.1. Apparent Complexity

Apparent complexity has been defined by Aaronson et al. (2014) as the entropy H of an object x after applying a smoothing function f, $H(f(x))$. 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:

$$H(P) = -\sum_{x \in X} P(x) \log P(x). \quad (2)$$

By Shannon's noiseless coding theorem the minimum average description length L of a sample is close to the Shannon entropy:

$$H(P) \leqslant L \leqslant H(P) + 1. \quad (3)$$

The Kolmogorov complexity $K(f(x))$ can be used as a proxy for the entropy of the smoothed function $H(f(x))$, as proposed by Aaronson et al. (2014). The seeming analogy between the concept of entropy and program size has been previously recognised (Chaitin 1975). The Kolmogorov complexity, or prefix 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 $K(x)$:

$$K(x) = \min_p\{l(p): U(p) = x\}. \quad (4)$$

A thorough treatment is provided by Li & Vitanyi (2008). The Kolmogorov complexity has the advantage of being well-defined for a particular description of a system such as an image of a galaxy. This is not the case for the Shannon entropy which is defined in terms of the possible states of the system. 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 (Aaronson et al. 2014), such as gzip.

The issue with using the approximated Kolmogorov complexity directly as measure of complexity is that it is maximized by random information. Intuitively a complexity measure should provide low values for random data that do not contain structure that is of interest to the observer (Zenil et al. 2012). Aaronson et al. (2014) have shown that the apparent complexity measure is able to achieve this by applying a smoothing function f to the input x.

While the Kolmogorov complexity of a random sequence is large, the apparent complexity of the same sequence becomes small with smoothing, as fluctuations are removed where the average or median information content becomes homogeneous at the coarse-grained resolution. Accordingly, we define the apparent complexity as the compressed description of regularities and structure after discarding all that is incidental. The apparent complexity will be small for both simple and random sequences.

The objective of applying the smoothing function f when deriving $C(f(x))$ is to remove incidental or random information, such as instrumental noise, which is incomprehensible to the observer even though it may have a physical basis. Comprehensibility here is defined with respect to the observer of information, in this case scientists with specific interests. Comprehensible information has a meaningful structure within feature space, which in the case of images refers to the spatial distribution of bits of information across available channels. The apparent complexity measure can be calibrated to align with expert distinctions between meaningful structure and noise by adjusting the measurement resolution through a smoothing function so that complexity values correctly partition expertly labelled data.

The apparent complexity measure of an image does not rely on the presence of any particular structures or structural elements and is invariant to rigid motions of the plane. The measure makes only explicit assumptions regarding the choice of coarse-graining level and the scale of the image. Previous data are therefore used only to calibrate the coarse-graining level (i.e., the appropriate measurement resolution).

Apparent complexity runs into obstacles as a well-defined measure of complexity. First, the uncomputability of the Kolmogorov complexity prohibits the concept from being defined in terms of an optimal compression. It has been proven by Chaitin et al. (1995) that there can be no procedure for finding all theorems that would allow for further compression. Furthermore, the problem of distinguishing between meaningful structure and incidental information, especially in finite data, may fail to be well-defined. Different smoothing functions and different coarse-graining levels will retain different distinct regularities in the data.

These theoretical challenges in objectively defining the apparent complexity can be circumvented when the approach is applied to the segmentation of observations by complexity. Here the apparent complexity can be calibrated to coincide with notions of complexity adopted by the observer.

3. Analysis of Radio Continuum Data

Segmentation based on apparent complexity can be used to identify complex images or complex regions within an image. We demonstrate this approach using radio continuum images from the ATLAS survey, to distinguish between simple and complex radio sources. Here we define "simple" sources as single unresolved components, and "complex" sources as anything else, including bent-tail galaxies and extended radio sources (e.g., Fanaroff-Riley I, Fanaroff-Riley II) containing bright radio components in combination with diffuse plume-like
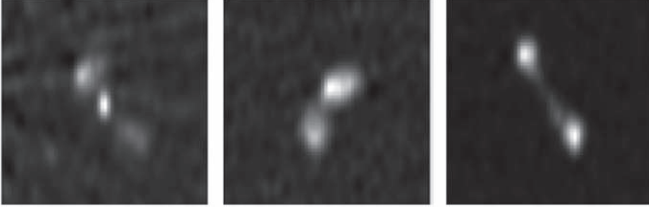
Figure 1. Complex radio sources with multiple components.
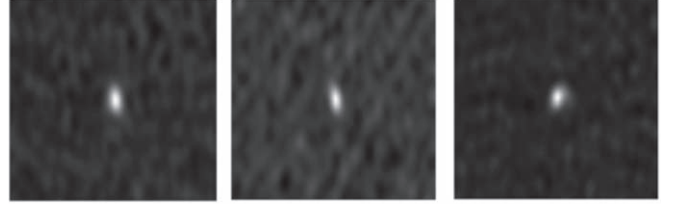
Figure 2. Simple unresolved radio sources.

jets. Figure 1 provides examples of complex radio sources and Figure 2 provides examples of simple sources.

It is expected that two simple radio (unresolved) sources, representing spatially separated galaxies that are randomly associated, will contain a significant amount of shared information in their morphologies, due to the similarity of the basic components. Conversely, a true complex source is likely to contain components of a differing nature, such as radio lobes of differing luminosity and jets and plumes with luminosity gradients. By the unique decompression property, distinct components will require additional bits of information in the combined compressed description (Cilibrasi & Vitanyi 2005). Accordingly the apparent complexity of a complex radio source should be larger than that of two simple sources in proximity.

The measure of apparent complexity can be used as a proxy for how interesting a radio image of a galaxy is, on the basis that radio sources with a larger apparent complexity contain a larger number of distinct and meaningful components that are likely to be of interest to an observer. We first demonstrate this intuition by distinguishing between artificially generated doubles and true complex sources, before using the measure to segment the ATLAS data with clustering methods.

3.1. Method

We approximate the apparent complexity measure by applying a median filter f with a window size calibrated to 10 pixels to an image x, and then calculating the gzip (Levine 2012) file size $C(f(x))$ as an upper bound on the Kolmogorov complexity $K(f(x))$. We apply this approach to the 256 by 256 pixel radio continuum images as follows.

1. Load a centred image as a 256 by 256 matrix of 8 bit channel pixel intensity values.
2. Crop from the centre of the image to create a 64 by 64 matrix.
3. Filter the matrix using a percentile based threshold (P90) for pixel intensity values.
4. Apply a median filter using a learned window size ($h = 10$) to produce a smoothed 64 by 64 matrix (to remove random information and retain structural information).
5. Compress the smoothed array using gzip.

6. Measure the compressed image size to estimate an upper bound of the Kolmogorov complexity.

We adopt a median filter with window size calibrated at $h = 10$ pixels, using a square structuring element with data extended at image borders through reflection. The choice of window size is the only free parameter in this method, and is learned from some small training set, as described in Section 3.3.

The median filter was selected because it completely removes noise and incidental values in regions predominately without flux measurement and retains the strength of signals in regions dominated by actual flux measurements. It is also better at preserving edges (than, for example, a Gaussian filter) given the expected salt-and-pepper noise.

Images were filtered based on a percentile-based threshold for pixel intensity values set at the 90th percentile (P90). This threshold was chosen through visual inspection of the images based on the removal of noise and imaging artifacts concentrated at lower flux values. Appendix B shows the the difference in average apparent complexity between simple and complex ATLAS DR1 images across a range of measurement resolutions at different percentile based thresholds, namely, P40, P50, P60, P70, P80, and P90. At thresholds lower than P60 the images become noise dominated and effective separation between simple and complex images is no longer achieved through smoothing. These results support the appropriateness of a large threshold, such as P90, which provides both reasonable estimation and clearer separation of the average apparent complexity for simple and complex sources. This parameter has not been finely tuned, however, and alternative threshold values my provide superior performance. Furthermore, this value was held fixed during calibration and testing so that the measurement resolution was the only free parameter.

Images were cropped to produce a 64 by 64 matrix. The matrix size was selected to retain all information (measurements) associated with the source of interest while reducing the image size to facilitate faster and more efficient processing.

The gzip compression software was selected to provide lossless compression post-smoothing. This allows information loss to be calibrated through the smoothing process. The gzip software was used due to its speed, broad usage, and its adoption in prior work by Aaronson et al. (2014).

Table 1
ATLAS DR1 and DR3 Samples

DR3 (n = 4825)	Complex Obs	Simple Obs
ELAIS	72	1892
CDFS	97	2764
DR1 (n = 708)	Complex Obs	Simple Obs
CDFS	34	674

3.2. Survey Sample

ATLAS data consist of deep radio continuum imaging of the the *Chandra* Deep Field South (CDFS) and the European Large Area *ISO* Survey (ELAIS). The data are described in Norris et al. (2006) and Middelberg et al. (2008) (DR1), and Franzen et al. (2015) (DR3).

Table 1 shows the number of identified radio sources within each field and data release. The table also provides a breakdown between sources that have been classified by human inspection from Norris et al. (2006) and R. P. Norris et al. (direct communication) as having simple and complex morphology.

The sources were provided as 256 by 256 pixel portable network graphics files. The images provided were pre-processed as detailed in Norris et al. (2006) and Franzen et al. (2015).

Labels were provided for ATLAS DR1 files identifying which sources had been classified as simple and complex. The data were then used to learn the smoothing function window size using images of galaxies that were manually labeled as having complex and simple morphologies. The window size of the smoothing function was chosen so as to maximize the difference between the average apparent complexity of observations labelled complex and simple in the ATLAS DR1 sample.

As shown in Table 1, the ATLAS DR3 data provide a much larger sample containing more complex sources. The analysis for ATLAS DR3 was conducted "blind", where labels were not provided with the source files. The success of the approach could therefore be judged independently using the DR3 data.

3.3. Smoothing Function Calibration

The appropriate window size for a smoothing function can be learned from a training set by maximizing the difference between the apparent complexity of observations expertly labelled as complex and simple. As suggested by Aaronson et al. (2014) there may also be natural choices for selecting the smoothing function suggested by our physical ability to actually observe systems and our knowledge of the systems properties.
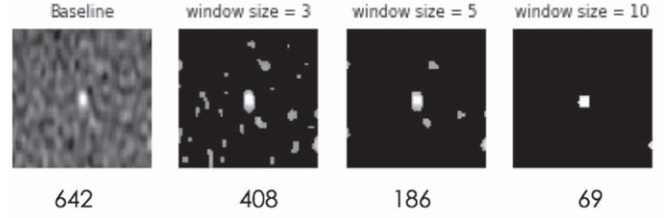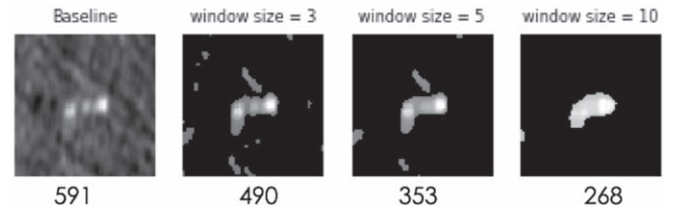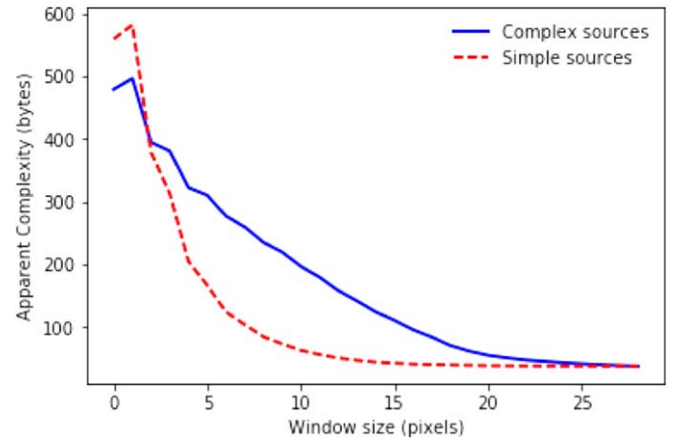

Figure 3. ATLAS DR1 sources with progressive smoothing function window sizes and complexity scores in bytes.

Figure 4. Comparison of the average apparent complexity between simple and complex ATLAS DR1 sources across changes in the size of the smoothing function window.

(A color version of this figure is available in the online journal.)

Applying an appropriate smoothing function to the radio images appears to remove random information and retain information contained within a structure comprised of distinct regions with different mean or median pixel values. The isolation of structure in a complex source is shown through the progressive application of coarser median filters in Figure 3. Images containing more random information show an initial rapid reduction in apparent complexity.

Figure 4 compares the average apparent complexity of complex and simple ATLAS DR1 images across changes in the smoothing function window size. As this increases, the apparent complexity of the simple and complex sources decreases, since information is being removed but, importantly,

they decrease at different rates. The near-exponential shape of the curve representing simple sources suggests that the measured complexity of these images consists of random information at the baseline pre-processing level. Conversely, the changing apparent complexity of the images labelled as complex, which reverts to a closer to linear rate of decrease, suggests a greater content of coarser, more comprehensible information. Where the apparent complexity curves of both complex and simple sources become flatter and converge, the smoothing function window size is large enough to remove both random and comprehensible content.

This analysis shows that there is an appropriate smoothing function window size that provides a clear separation, on average, between images that have been expertly classified as complex and simple sources in the ATLAS DR1 data. The separation between the average apparent complexity values of simple and complex sources becomes clear between a window size of 5 to 25 pixels, with the largest separation achieved toward the centre of this range. We fixed our choice to a diameter of 10 pixels to use when applying this function to the ATLAS DR3 data. The separation disappears at larger coarse-graining levels where the smoothing filter blacks out the image, removing all content. At null and low smoothing levels the images represent the baseline pre-processing level of the ATLAS images.

It is interesting to note that, at the baseline pre-processing level, the average apparent complexity of images representing simple sources exceeds the average apparent complexity of images representing complex sources. The rapid reduction in complexity of these images with increased smoothing suggests that this higher initial apparent complexity is likely to be attributable to a higher content of random information. This is in contrast to the truly complex images that retain a greater percentage of the measured complexity at equivalent smoothing levels.

3.4. Experimental Results

We computed the apparent complexity of the ATLAS DR3 images, using the method described in Section 3.1. The resulting distribution of values is shown in Figure 5, showing that the apparent complexity values for expertly classified simple and complex sources slightly overlap but with complex sources concentrated in the heavier right tails of the distributions.

Without reference to the expertly classified labels for the images, we adopted a threshold of the approximate P90 of the apparent complexity distribution (of the combined sample) segmenting the heavy tails that we interpreted as being influenced by a second overlapping distribution assumed to represent complex sources. In this way we assume a segmentation boundary at an apparent complexity threshold of 300 bytes, equivalent to selecting approximately the top 10% of complexity values.
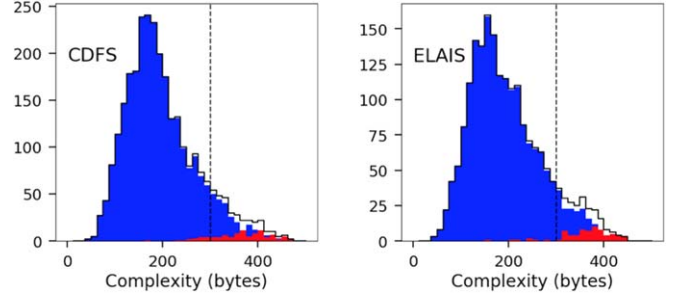
Figure 5. Distributions of the apparent complexity of the total sample (black solid line), subdivided into simple (blue) and complex (red) radio sources (human classified) in the CDFS (left) and ELAIS (right) fields from ATLAS DR3. The black dashed vertical line gives the 300 byte boundary we assume in order to partition the two populations by apparent complexity.
(A color version of this figure is available in the online journal.)

Table 2
Confusion Matrix (P90 Complexity Cut)

CDFS (n = 2861)	Complex Obs	Simple Obs
Prediction: complex	81	244
Prediction: simple	16	2520
ELAIS (n = 1964)	Complex Obs	Simple Obs
Prediction: complex	65	167
Prediction: simple	7	1725

The performance of this binary segmentation with respect to the true nature (expert classification) is shown in Table 2. These results show that the apparent complexity measure, at the selected partition boundary, can be used to correctly identify 86% (i.e., a recall of 0.86) of the interesting observations from the combined DR3 samples with a 91% reduction in the non-interesting data volume following classification.

Applying clustering methods to the complexity measurements allows us to implement an automated process for detecting interesting and unexpected observations. We apply clustering methods to the complexity measurements to avoid reliance on a large training set that would be needed to represent the feature space and class boundaries. To demonstrate the generalizability of the apparent complexity measure we calibrate the smoothing function on a small sample (ATLAS DR1, $n = 708$) and test it using a much larger sample (ATLAS DR3, $n = 4825$).

These considerations reflect the intended use of the approach to identify complex observations, with often unexpected features, in new data that are extensively larger and more varied than any previous data sets, such as the new data that will be produced by next-generation instrumentation such as ASKAP and the SKA.

To demonstrate an automated partitioning process we performed cluster analysis by fitting a two-component Gaussian mixture model (GMM) using the expectation maximisation
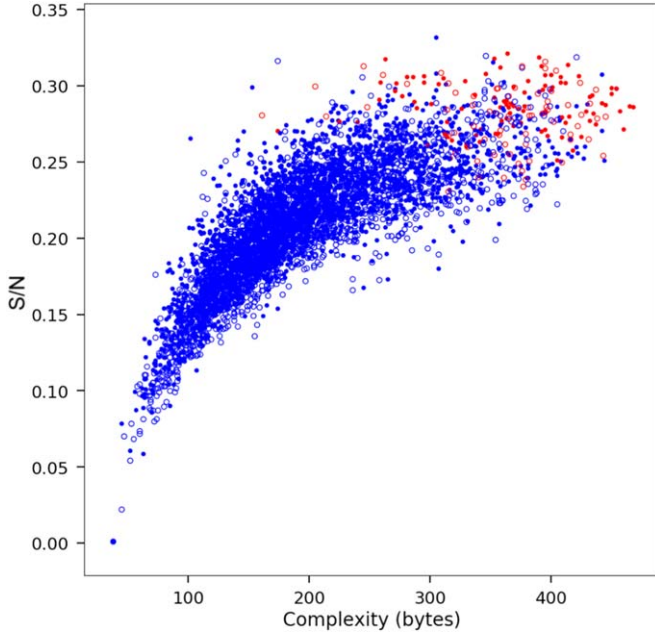
Figure 6. Scatter plot demonstrating the effectiveness of apparent complexity and S/N to partition simple (blue) and complex (red) radio sources in the CDFS (filled circle) and ELAIS (empty circle) fields from ATLAS DR3. (A color version of this figure is available in the online journal.)

Table 3
Confusion Matrix (Complexity Partition Based on GMM)

CDFS (n = 2861)	Complex Obs	Simple Obs
Prediction: complex	73	141
Prediction: simple	24	2623
ELAIS (n = 1964)	Complex Obs	Simple Obs
Prediction: complex	65	149
Prediction: simple	7	1743

Table 4
Confusion Matrix (Complexity and S/N GMM Partition)

CDFS (n = 2861)	Complex Obs	Simple Obs
Prediction: complex	83	140
Prediction: simple	14	2624
ELAIS (n = 1964)	Complex Obs	Simple Obs
Prediction: complex	65	126
Prediction: simple	7	1766

algorithm. Binary segmentation using this approach is shown in Table 3. Results based on this approach correctly identify 82% (i.e., a recall of 82) of the interesting observations from the combined DR3 samples with a 94% reduction in the non-interesting data volume following classification.

Type II errors, representing the incorrect classification of complex sources as simple, may be due to the removal of meaningful information by the smoothing function or potentially by the sparse representation of complex features, discernible to a human observer, but having little impact on the information content of the image. Classification errors may also be attributable to the allocation of the partition boundary.

Type I errors, representing the incorrect identification of simple sources as complex, may be due to the presence of non-random information deemed by a human observer to be incidental and not contributing to the complexity of the source itself. An example could be a telescope imaging artifact containing structure, such as a point-spread function originating from a brighter source. Examples of telescope imaging artifacts include the diagonal lines shown at baseline in Figure 3.

Alternatively, type I errors may be explained by the retention of random information not removed by the smoothing function. Figure 4 shows that there is a large amount of random information in the simple sources at baseline, and this suggests there is a risk that in some images random information will produce incidental structure that may not be removed through smoothing. Where random information is retained after smoothing, segmentation is likely to be improved by incorporating thresholds in both the

apparent complexity and the S/N, as random information is likely to be distributed more uniformly across the available channel intensity values. The S/N can be calculated as the reciprocal of the coefficient of variation for the channel intensity values.

Figure 6 shows both the apparent complexity and the S/N for the the ATLAS DR3 samples and the relationship between them, where the samples have been partitioned into simple and complex by human inspection. In this figure the true complex radio sources are clustered at larger apparent complexity and S/N values, towards the top right edge of the scatter plot. This figure shows the extra information provided by the S/N in segmenting simple and complex sources.

We automated the segmentation process by fitting Gaussian mixture models to segment and remove observations with low complexity and S/N values (i.e., the tail shown in Figure 6) and segment the remaining complexity and S/N values to identify the most interesting observations (i.e., simple versus complex). The results are shown in Table 4.

In assessing the effectiveness of the approach, the most important consideration is the reduction of the type II error rate, measuring the effectiveness of the approach at identifying as many of the interesting observations as possible. Minimising the type I error rate is of importance in providing a significant reduction in the volume of non-interesting data flagged for further investigation. Given a significant reduction in the total data volume and a very low type II error rate, the contamination of the segregated sample as measured by precision was not deemed to be of primary concern. For this reason the informedness measure, as described in Appendix A, was chosen to assess performance. The informedness measure incorporates both Type I errors (false positives) and Type II errors (false negatives) and describes the

Table 5
Summary of Experimental Results, Including Recall, False Positive Rate, and Informedness Measure

CDFS ($n = 2861$)			
Method	Recall	False Positive Rate	Informedness
Percentile cut: complexity without smoothing	0.04	0.12	−0.07
Percentile cut: coarse-grained complexity	0.84	0.09	0.75
GMM: coarse-grained complexity	0.75	0.05	0.70
GMM: coarse-grained complexity and S/N	0.86	0.05	0.81
ELAIS ($n = 1964$)			
Method	Recall	False Positive Rate	Informedness
Percentile cut: complexity without smoothing	0.03	0.12	−0.09
Percentile cut: coarse-grained complexity	0.90	0.09	0.81
GMM: coarse-grained complexity	0.90	0.08	0.82
GMM: coarse-grained complexity and S/N	0.90	0.07	0.84

improved performance of the approach with respect to chance, costing true positives and false positives in a way analogous to how a bookmaker fairly prices the odds (Powers 2011). A detailed description of these metrics is provided in Appendix A.

Based on the results shown in Table 4 the combined sample (CDFS and ELAIS) produces a recall of 0.88 (88% true positives) and a false positive rate of 6%. This represents an informedness of 0.82. The CDFS sample provides a recall of 0.86 and informedness of 0.81 while the ELAIS sample provides a recall of 0.90 and informedness of 0.84. These results show that approximately 90% of expertly classified complex sources were contained within the largest 10% of complexity values.

Results show an improvement in performance when incorporating the S/N, with an improved recall within the CDFS sample and a reduced false positive rate within the ELAIS sample as shown in Table 5. By reducing the likelihood thresholds used for classification the recall can be further improved at the expense of the false positive rate.

To demonstrate the importance of applying an appropriate smoothing function, an experiment was run to partition the data without smoothing. The images were classified by selecting approximately the top 10% of observations based on apparent complexity score, as was also done using the smoothed images. The drastic reduction in performance is shown by comparing the results in Tables 2 and 6, and comparing the summary measures for the ranked complexity, with and without smoothing, as shown in Table 5. These results demonstrate the importance of the smoothing function when partitioning interesting observations. Results suggest that the learned smoothing function is successfully able to isolate comprehensible content associated with the meaningful structural information used by astronomers to manually classify the radio sources.

The results shown in Table 5 also demonstrate that the smoothing function window size learned from the smaller CDFS DR1 sample is able to generalize well when applied to the larger CDFS and ELAIS DR3 samples containing a larger collection of complex morphologies.

3.5. Validation Using Synthetic Data

Synthetic radio images were constructed by placing simple radio sources in close proximity, comparable to the distances between components associated with real complex sources with two components (i.e., real doubles). In accordance with the unique decompression property the distinct components of true complex sources should produce larger apparent complexity values. The apparent complexity was calculated for images containing only simple radio sources, synthetic sources combining two simple sources, and real complex radio sources as shown in Figure 7. The apparent complexity values distinguish between the true complex sources and synthetic sources consisting of two simple sources in close proximity, with true complex sources producing larger complexity values.

3.6. Method Comparisons

The proposed approach presents two key attributes. First, apparent complexity is fast and computationally efficient to compute. The measure requires only applying a smoothing function and compression algorithm, both of which can be implemented to scale at worst-case linear time complexity, making the measure appropriate for analysing large data sets. Second, the approach generalizes well when using only a very small training set to calibrate the smoothing function. Combined, these attributes make the approach a good candidate for detecting the unexpected in large data produced from new astronomical surveys. These surveys will produce data where the volume, and likely scope, of observations will greatly exceed any existing data available for training classifiers.

However, the question remains as to whether this approach performs as reliably as traditional binary classification approaches applied to pixel data. We tested our method,

Table 6
Confusion Matrix (P90 Complexity Cut without Smoothing)

CDFS (n = 2861)	Complex Obs	Simple Obs
Prediction: complex	4	321
Prediction: simple	93	2443

ELAIS (n = 1964)	Complex Obs	Simple Obs
Prediction: complex	2	230
Prediction: simple	70	1662

comparing the proposed approach to more traditional classification methods, namely logistic regression and SVMs.

Feature preparation for logistic regression and SVM training and testing involved flattening the images and sorting the resulting one-dimensional arrays by pixel intensity value. The ranked or distributional representation of pixel values are more suitable for classification, as they are invariant to rigid motions of the plane. Logistic regression was implemented with the loss function optimized using gradient descent.[6] The SVM was implemented using a two-degree polynomial kernel specified with a coefficient 1.

Training and testing on ATLAS DR1 data were performed using cross-validation, splitting the DR1 set in two separate training and testing sets each time, with each set of 354 images containing 17 complex images. Thus, the training data had a comparable range of complex observations to the data used for testing. These experiments employed the same ATLAS DR1 data used in other studies (L. F. Park et al. 2019, in preparation; Alger et al. 2018) to allow comparison to additional approaches including those that incorporate feature learning such as convolution neural networks. The approaches tested here achieved comparable performance to those tested by L. F. Park et al. (2019, in preparation) in the binary classification of simple and complex morphology in the ATLAS DR1 data.

Table 7 shows results from experiments using logistic regression and SVM models implemented using the ranked pixel data. In terms of the informedness and recall metrics, these binary classification approaches achieve results comparable to the method using the coarse-grained complexity measure.

Having trained on ATLAS DR1 (n = 708), we then tested using ATLAS DR3 (n = 4825). The results are also shown in Table 7. We found that logistic regression applied to the ranked pixel data provided an informedness of 0.64 (across both CDFS and ELAIS samples) and the SVM applied to the ranked pixel data provided an informedness of 0.48, both of which were considerably smaller than the informedness of 0.82 produced when using the apparent complexity measure. Thus the

---

[6] Here we implement gradient descent using the `Tensorflow` package (Abadi et al. 2015) having achieved superior performance to the implementation of the Library for Large Linear Classification using `scikit-learn` (Pedregosa et al. 2011).

complexity measure was better able to generalize when using a smaller training set. This shows that the smoothing scale can be effectively calibrated using only a small number of complex images to distinguish between simple images and a broad range of complex images. In contrast, the classification methods applied to the pixel data were unable to replicate this performance, being "under-trained" for such an expanded data set.

These tests demonstrate that the proposed approach provides comparable performance to traditional classification methods applied to pixel data when the training data are representative of the range of complex observations in the testing data. They also demonstrate, however, that the use of the proposed complexity measure provides superior performance to the traditional classification methods when the training data set is much smaller than the testing data set and the training data contain only a small sample of the range of complex observations found in the testing data. This latter scenario is representative of the challenge that comes with larger astronomical surveys.

4. Conclusions

We introduce apparent complexity as a measure to identify interesting or complex observations in large astronomical data sets. The apparent complexity is defined as the entropy of an image after applying a smoothing function, and is approximated by the file size of a corresponding image after smoothing and compression. The measure can be computed at worst-case linear time complexity, making it well suited for analysing very large data sets to identify smaller and relevant samples for additional analysis requiring more computationally intensive methods or manual inspection.

Using imaging data from the ATLAS radio continuum survey, we have shown that apparent complexity, combined with the S/N, can be used to partition images of galaxies into those with simple and complex morphologies. Partitioning the images by fitting Gaussian mixture models using the expectation maximization algorithm, we are able to implement an automated process to identify complex sources with a recall of 0.88 and informedness of 0.82 across both CDFS and ELAIS samples. Results also show that approximately 90% of expertly classified complex sources were contained within the largest 10% of complexity values.

We demonstrate that the association between apparent complexity and the morphological complexity of observations generalizes well, positioning the measure as a candidate machine learning feature for identifying complex observations with unanticipated morphological features (i.e., unknown unknowns) in the presence of noise. We have shown that the apparent complexity with a smoothing function window size learned from the smaller ATLAS DR1 sample (n = 720) is able to generalize well when applied to the much larger ATLAS DR3 sample (n = 4825) containing a larger collection of complex morphologies. The use of apparent complexity to segment images and identify interesting observations should generalize well across

Real single component sources
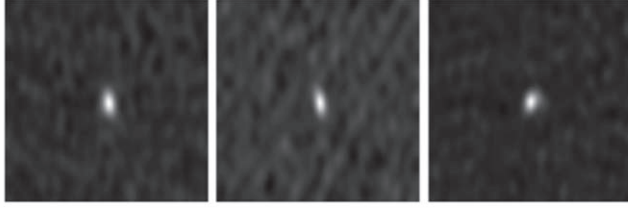
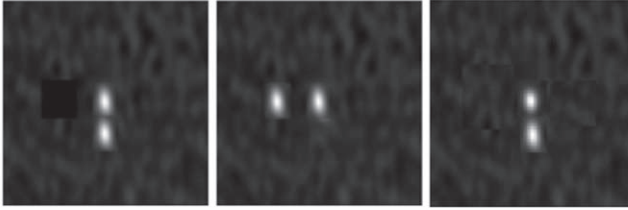125*   93*   111*

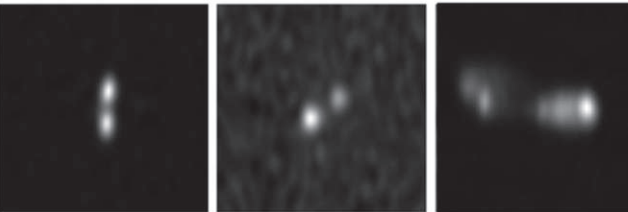## Synthetic doubles (actually two singles)

127*   158*   152*

Real doubles

204*   233*   403*

* Apparent complexity values

Figure 7. Apparent complexity (in bytes) calculated for images containing only simple radio sources, synthetic sources combining two simple sources, and real complex radio sources. These results show that the apparent complexity values can be used to distinguish between the true complex sources and simple sources in close proximity.

Table 7
Classification Performance of GMM Using Calibrated Coarse-grained Complexity Compared to Alternative Classification Approaches

DR1 CDFS ($n = 708$ cross val)			
Method	Recall	False Positive Rate	Informedness
Logistic regression	1.00	0.19	0.81
SVM	0.88	0.02	0.86
GMM: coarse-grained complexity	0.85	0.06	0.79
DR3 CDFS ($n = 2861$)			
Method	Recall	False Positive Rate	Informedness
Logistic regression	0.99	0.36	0.63
SVM	0.61	0.03	0.58
GMM: coarse-grained complexity	0.75	0.05	0.70
GMM: coarse-grained complexity and S/N	0.86	0.05	0.81
DR3 ELAIS ($n = 1964$)			
Method	Recall	False Positive Rate	Informedness
Logistic regression	0.94	0.30	0.65
SVM	0.36	0.02	0.34
GMM: coarse-grained complexity	0.90	0.08	0.82
GMM: coarse-grained complexity and S/N	0.90	0.07	0.84

testing data. Training on ATLAS DR1 ($n = 708$) and testing on ATLAS DR3 ($n = 4825$), logistic regression applied to pixel data provided an informedness of 0.64 and the SVM applied to pixel data provided an informedness of 0.48, both being smaller than the informedness of 0.82 produced when using the apparent complexity measure.

We envisage a potential application for the apparent complexity measure at the early stages of the analysis pipeline for radio images. The early identification of complex radio sources allows these observations to be segmented for additional analysis. This is likely to include cross-matching with observations at other frequencies to determine whether the host galaxy is coincident with any of the radio components. It may also serve in the early identification of new and unexpected observations that may arise as part of next-generation surveys, for example the SKA (Jarvis et al. 2015; Wilkinson 2015).

While a large number of the most interesting observations are likely to be more complex than less interesting observations, we do acknowledge that there will be interesting outliers in scientific data sets that are relatively simple. Further work to understand the intersection of the sets of outlying and complex

other samples, including different types of observational data outside of radio frequencies. Further testing with large radio samples and new types of data will be needed to test this hypothesis.

Experiments were conducted comparing the proposed approach using the apparent complexity measure against more traditional classification methods, namely logistic regression and SVMs, applied to pixel data. These experiments demonstrated that the use of the coarse-grained complexity measure provides comparatively superior classification performance when the training data contain only a small sample of the complex observations found in the

observations will clarify the potential limitations and best applications of a complexity-based approach.

Appendix A
Evaluation Measures

For a binary classification problem a 2 by 2 contingency table can be constructed to represent counts of false positives (FP), true positives (TP), false negatives (FN) and true negatives (TN) as depicted by Table 8. Such a table is referred to as a confusion matrix and depicts both the counts of Type I errors (FP) and Type II errors (FN).

Quantities can be derived using the information contained within a binary confusion matrix to measure the performance of a classifier. Two such quantities are precision

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (5)$$

and recall

$$\text{Recall} = \frac{TP}{TP + FN}. \quad (6)$$

Precision determines the number of correct positive classifications as a fraction of positive classifications, while recall determines the number of correct positive classifications as a fraction of the total number of actual positives (so the fraction of positive objects that have been missed would be $1 - \text{recall}$, in the binary classification case).

An alternative framework for measuring performance involves the use of receiver operating characteristic (ROC) curves. The use of ROC curves to construct a comparative framework has been adopted in the machine-learning literature (Fürnkranz & Flach 2005). These approaches account for chance level performance and can also be used to account for the cost weightings of negative and positive cases. ROC analysis examines the FP rate (FP/RN) versus the TP rate (TP/RP) and presents equivalent information to ratios calculated in the vertical direction of the confusion matrix presented in Table 8.

The maximum distance of the ROC curve from the 45 degree chance line is equivalent to Youden's J statistic (Youden 1950) and the informedness measure (Powers 2011). The informedness measure is equivalent to the subtraction of the FP rate (FPR) from the TP rate (TPR) as follows:

$$\textbf{Informedness} = \frac{TP}{TP + FN} - \frac{FP}{TN + FP} = TPR - FPR. \quad (7)$$

Table 8
Confusion Matrix for a Binary Classification Problem

	Class +	Class −	Total
Prediction +	TP	FP (Type I error)	Predicted positives (PP)
Prediction −	FN (Type II error)	TN	Predicted negatives (PN)
Total	Real positives (RP)	Real negatives (RN)	

This measure is also equivalent to a chance adjusted version of recall:

$$\textbf{Informedness} = \text{recall} - FPR. \quad (8)$$

Powers (2011) shows that informedness is an unbiased estimator of above chance performance. The measure incorporates both FPs and FNs and describes the improved performance of the measured classifier with respect to chance, costing TPs and FPs in a way analogous to how a bookmaker fairly prices the odds (Powers 2011). For this reason the measure is also referred to as bookmaker informedness. The informedness measure is defined on a $(-1, 1)$ interval and gives equal weighting to the TP and FP rate.

Informedness appears appropriate for evaluating the effectiveness of alternative approaches at detecting and classifying interesting observations in large astronomical data. The informedness measure relates to the following objectives of classification.

1. Maximize TP rate (i.e., minimize the type II error rate) providing assurance that actual interesting observations are available for analysis.
2. Minimize FP rate (i.e., minimize the type I error rate) to minimize the data burden and place minimal unnecessary burden on data transmission and storage infrastructure.

Removing FPs reduces storage and data handling requirements and the associated costs. Retaining a smaller subset of observations that are likely to be interesting allows these to be directed to low-latency storage options where they can be easily retrieved.

Due to the likely small number of actual interesting observations compared to normal observations the metric is likely to be more sensitive to small changes in the TP count and less sensitive to small changes in the FP count.

Appendix B
Sensitivity Analysis: Pixel Intensity Threshold

We consider a range of possible values for the threshold pixel intensity, which removes random information from the
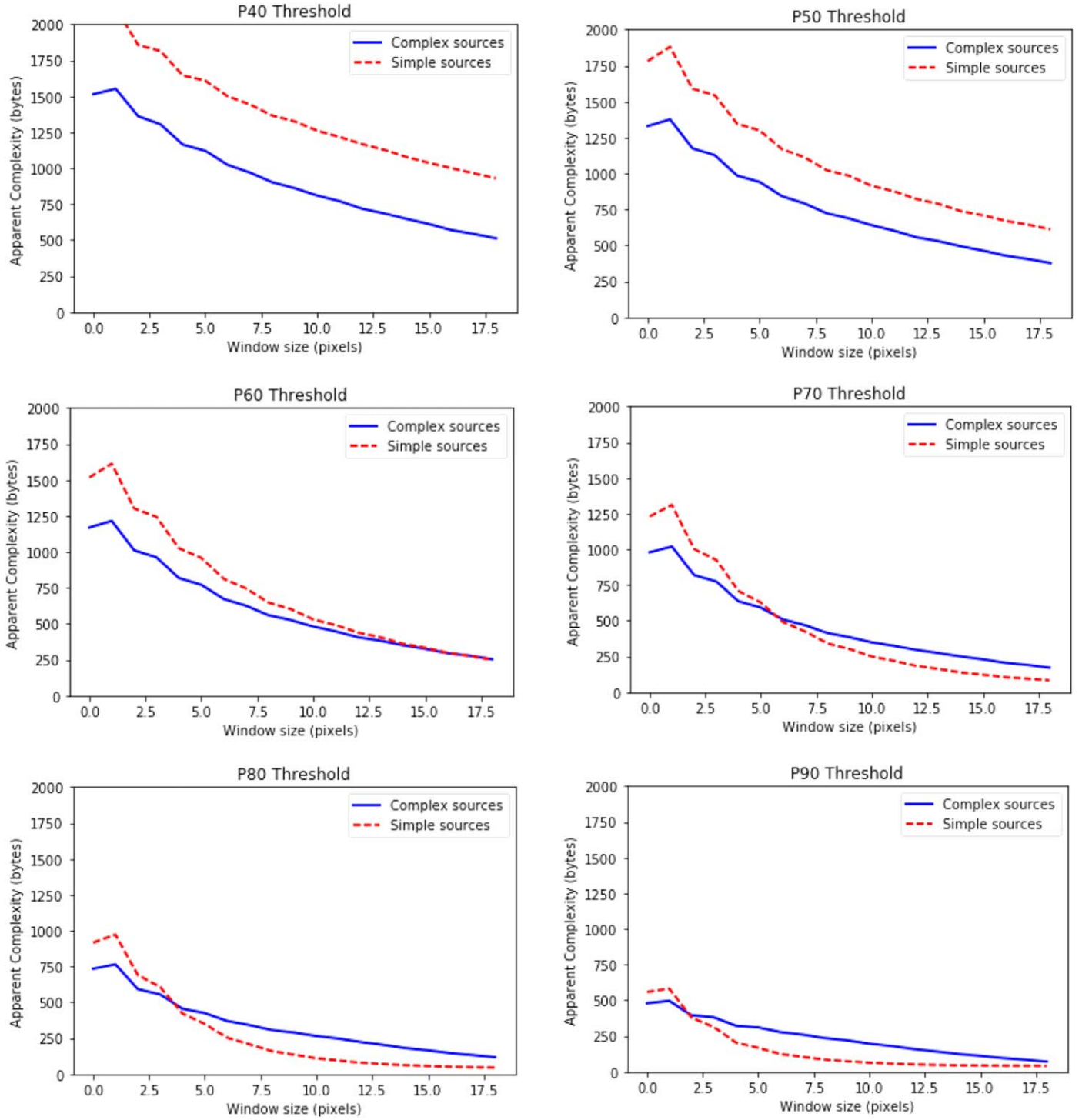
Figure 8. Plots showing the differences in average apparent complexity between simple and complex ATLAS DR1 images as a function of smoothing scale at different percentile-based pixel intensity thresholds.
(A color version of this figure is available in the online journal.)

image, and preserves the non-random or "true" image information particular to the object. Figure 8 shows the differences in average apparent complexity between simple and complex ATLAS DR1 images at different percentile-based pixel intensity thresholds, namely, P40, P50, P60, P70, P80, and P90, as a function of smoothing scale. At thresholds lower than P60 the images become noise-dominated and effective separation between simple and complex images is no longer possible through smoothing. The simple images produce larger complexity values attributable to noise. At all possible threshold values, the effective separation between simple and complex images disappears as the smoothing scale greatly exceeds the size of source features. These results support the appropriateness of a large threshold, such as P90, which provides clear separation between simple and complex sources and larger average apparent complexity for complex sources in the presence of noise. This parameter has not been finely tuned, however, and alternative threshold values may provide superior performance. Furthermore, this value was held fixed during calibration and testing so that the smoothing resolution was the only free parameter.

ORCID iDs

David Parkinson ⬤ https://orcid.org/0000-0002-7464-2351

References

Aaronson, S., Carroll, S., & Ouellette, L. 2014, arXiv:1405.6903v1
Abadi, M., Agarwal, A., Barham, P., et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, https://www.tensorflow.org/
Alger, M. J., Banfield, J. K., Ong, C. S., et al. 2018, MNRAS, 478, 5547
Aniyan, A. K., & Thorat, K. 2017, ApJS, 230, 20
Baron, D., & Poznanski, D. 2017, MNRAS, 465, 4530
Chaitin, G. 1975, J. Assoc. Comput. Mach., 22, 329
Chaitin, G., Arslanov, A., & Calude, C. 1995, CDMTCS Research Report Series CDMTCS-008
Charnock, T., & Moss, A. 2017, ApJL, 837, L28
Cilibrasi, R., & Vitanyi, P. 2005, ITIT, 51
Dieleman, S., Willett, K., & Dambre, J. 2015, MNRAS, 450, 1441
Franzen, T. M. O., Banfield, J. K., Hales, C. A., et al. 2015, MNRAS, 453, 4020
Fürnkranz, J., & Flach, P. 2005, Mach Learn, 58, 39
Gell-Mann, M. 1994, The Quark and the Jaguar: Adventures in the Simple and the Complex (New York: Henry Holt and Company)
Huertas-Company, M., Gravet, R., Cabrera-Vives, G., et al. 2015, ApJS, 221, 8
Jarvis, M., Bacon, D., Blake, C., et al. 2015, arXiv:1501.03825
Karpenka, N. V., Feroz, F., & Hobson, M. P. 2013, MNRAS, 429, 1278
Kessler, R., Bassett, B., Belov, P., et al. 2010, PASP, 122, 1415
Kim, D.-W., & Bailer-Jones, C. A. L. 2016, A&A, 587, A18
Levine, J. 2012, Media Types, Tech. Rep. RFC, 6713, The "application/zlib" and "application/gzip" (doi:10.17487/RFC6713)
Li, M., & Vitanyi, P. 2008, An Introduction to Kolmogorov Complexity and Its Applications (3rd ed.; New York: Springer)
Middelberg, E., Norris, R. P., Cornwell, T. J., et al. 2008, AJ, 135, 1276
Norris, R. P. 2017a, PASA, 34, e007
Norris, R. P. 2017b, NatAs, 1, 671
Norris, R. P., Afonso, J., Appleton, P. N., et al. 2006, AJ, 132, 2409
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825
Perreault, S., & Hebert, P. 2007, ITIP, 16, 2389
Powers, D. M. 2011, Journal of Machine Learning Technologies, 2, 37 (http://hdl.handle.net/2328/27165)
Robinson, B. J. 1987, PASA, 7, 220
Wilkinson, P. 2015, in Proc. of Advancing Astrophysics with the Square Kilometre Array (AASKA14) (Giardini Naxos, Italy), 65
Youden, W. J. 1950, Cancer, 3, 32
Zenil, H., Delahaye, J.-P., & Gaucherel, C. 2012, Cmplx, 17, 26