MORFOMETRYKA—A NEW WAY OF ESTABLISHING MORPHOLOGICAL CLASSIFICATION OF GALAXIES

F. FERRARI1, R. R. DE CARVALHO2, AND M. TREVISAN2,3
1IMEF—FURG, Rio Grande, RS, Brazil; fabricio@ferrari.pro.br
2INPE/MCT, Divisão de Astrofísica, S. J. dos Campos, Brazil
3Institut d'Astrophysique de Paris, F-75014, Paris, France

Received 2015 August 21; accepted 2015 September 21; published 2015 November 17

ABSTRACT

We present an extended morphometric system to automatically classify galaxies from astronomical images. The new system includes the original and modified versions of the CASGM coefficients (Concentration $C_1$, Asymmetry $A_3$, and Smoothness $S_3$), and the new parameters entropy, $H$, and sparityality $\sigma_v$. The new parameters $A_3$, $S_3$, and $H$ are better to discriminate galaxy classes than $A_1$, $S_1$, and $G$, respectively. The new parameter $\sigma_v$ captures the amount of non-radial pattern on the image and is almost linearly dependent on $T$-type. Using a sample of spiral and elliptical galaxies from the Galaxy Zoo project as a training set, we employed the Linear Discriminant Analysis (LDA) technique to classify EFIGI (Baillard et al. 4458 galaxies), Nair & Abraham (14,123 galaxies), and SDSS Legacy (779,235 galaxies) samples. The cross-validation test shows that we can achieve an accuracy of more than 90% with our classification scheme. Therefore, we are able to define a plane in the morphometric parameter space that separates the elliptical and spiral classes with a mismatch between classes smaller than 10%. We use the distance to this plane as a morphometric index ($M_i$) and we show that it follows the human based $T$-type index very closely. We calculate morphometric index $M_i$ for $\sim$780k galaxies from SDSS Legacy Survey—DR7. We discuss how $M_i$ correlates with stellar population parameters obtained using the spectra available from SDSS—DR7.

Key words: galaxies: fundamental parameters – galaxies: general – galaxies: photometry – galaxies: statistics – techniques: image processing

1. INTRODUCTION

The study of the formation and evolution of galaxies in general requires their systematic observations over a large redshift domain. Data sets for local (today’s systems) and distant galaxies (their actual progenitors) must be consistently gathered to avoid biases, a procedure that requires knowledge of the very evolution we are seeking to understand. From an astrophysical perspective, mechanisms regulating star formation, e.g., ram-pressure (Gunn & Gott 1972), harassment (Moore et al. 1996), starvation (Larson et al. 1980), depend on the distance from the center of the potential well of a cluster; their effects on the stellar population of a galaxy depend on how efficiently the local environment is capable of removing the interstellar gas and affecting the star formation history. Thus, morphology, in a general sense, is just a snapshot reflecting all these processes imprinted in the galaxy image at a given time.

Traditionally, galaxy morphology has been addressed visually: an expert examines the images of galaxies and identifies features (or the absence of them, in the case of early-type galaxies) which distinguish the object as belonging to a specific class, as done in Hubble (1926), de Vaucouleurs (1959), Sandage (1975), van den Bergh (1976), Lintott et al. (2008, 2011), among many others. This classification paradigm is strongly subjective, prone to errors, and cannot be applied to the number of galaxies present in modern surveys. For instance, Fasano et al. (2012) compare their classification with that of de Vaucouleurs et al. (1991, RC3 catalog, see their Figure 2). As it is clearly seen there is an uncertainty of approximately 1 in $T$-type within Fasano et al. (2012) authors and about 2.5 between RC3 and Fasano et al. (2012). Comparison between EFIGI and NA2010, for 1438 galaxies in common (Baillard et al. 2011, Figure 32), exhibits a similar sort of inconsistency, with an uncertainty between 2 and 3 in $T$-type. This tell us that, for example, visual classification does not agree when distinguishing an S0 from an Sa or an E from an S0. Thus, it is imperative to quantify the morphology of a galaxy as a measurable quantity—morphometry—that can be coded in an algorithm. However, in spite of its uncertainties, visual classification is still important, because automated techniques would have difficulty doing classifications like (R,R,0) SAB(r, nr0)/a, which is much more valuable than just knowing whether a galaxy is S, S0, or E. Also, regarding RC3, it should be noted that its classifications were made based on photographic plates or sky survey charts. Modern digital images allow greater consistency between multiple classifiers and have the potential to greatly improve on RC3.

Two approaches for galaxy morphology have been widely explored recently: parametric—those which model the light distribution as a bulge plus disc plus other less important components representing a few percent of the total light (e.g., Peng et al. 2002; Simard et al. 2002); non-parametric—those which use the measured properties of the light distribution, like concentration, asymmetry (e.g., Abraham et al. 1996). Each approach has its virtues and vices, as discussed, for example, in Andrae et al. (2011).

One relatively successful non-parametric system is the concentration, asymmetry, smoothness, Gini and M20 (CASGM) system, presented in Abraham et al. (1994, 1996), Conselice et al. (2000), and Lotz et al. (2004). This basic set has been enlarged with other quantities such as the Sérsic model parameters (Sérsic 1968), and the Petrosian radius (Petrosian 1976), among others. These quantities may not work properly in the high redshift regime and this has been studied in recent papers (e.g., Freeman et al. 2013). These authors use Multi-mode, Intensity, and Deviation statistics, MID, to detect disturbances in the galaxy light distribution and show that it is very effective at $z \sim 2$.
The previously mentioned way of establishing galaxy morphology answers two immediate needs. First, it is possible to reproduce human classification by positioning the galaxies in the space of these parameters. In such a supervised classification, a set of visually classified galaxies is used to train a discriminant function that will assign to each new galaxy a probability of belonging to each class. The second reason for establishing a galaxy morphometry system is that we can seek structures, in the quantitative morphology parameter space, that may yield clues for the physical reasons for their formation and evolution that are not visible in the currently human-based mode. Furthermore, a system such as the Hubble tuning fork classification does not account for all the details that we can currently measure in galaxy images, and it does not hold as we go deep even at a moderate redshift \( z = 0.25 \). To evaluate this, a new quantitative classification procedure is needed, both to handle the large amount of data becoming available with the new surveys, and also to help us find the physical processes driving galaxy evolution.

The paper is organized as follows: in Section 2 we discuss similar works; in Section 3 we describe the data sets used; in Section 4 we define new non-parametric methods to quantify galaxy morphology; in Section 5 we present the MORFOMETRYKA algorithm. We apply the MORFOMETRYKA code to galaxy samples described in Section 6, where we also test the robustness of the measured parameters and explore the ability of them to classify galaxy morphologies. In Section 7 we propose a new morphometric index \( M_f \). In Section 8 we compare \( M_f \) with other physical parameters and a summary is presented in Section 9.

2. RELATED WORK

There have been several attempts to classify galaxies automatically, beginning with Abraham et al. (1994, 1996), among others. Here we briefly mention recent works based on machine learning and on morphometric parameters. The list is not meant to be exhaustive but rather to present different approaches followed in the last few years.

Huertas-Company et al. (2011) used a system based on colors, shapes and concentration to train a support vector machine to classify \( \sim 700 \)k galaxies from the SDSS DR7 spectroscopic sample. For each galaxy, they estimate the probabilities of being E, S0, Sab, or Scd. It is not a pure morphometric classification, since it includes colors.

Scarlata et al. (2007) analyzed 56,000 COSMOS galaxies with the ZEST algorithm, using five non-parametric diagnostics \((A, C_1, G, M_{20}, q)\) and Sérsic index \( n \). They perform principal component analysis (PCA) and classify galaxies with three principal components. They find contamination between galaxy classes in parameter space (see their Figure 10), although they do not state clearly the success rate of the classification.

Andrae et al. (2011) present a detailed analysis of several critical issues when dealing with galaxy morphology and classification. Several morphological features are intertwined and cannot be estimated independently. They show the dependence between \( C \) and \( n \), which is also presented here in a different form in Appendix C. The authors claim that parameter based approaches are better for classification, and state that a system such as CASGM has serious problems. However, they do not show it in practice.

Dieleman et al. (2015) present a Neural Network machine to reproduce Galaxy Zoo classification. They work directly in pixel space, using a rotation invariant convolution that minimizes sensitiveness to changes in scale, rotation, translation and sampling of the image. The algorithm obtains an accuracy of 99% relative to the Galaxy Zoo human classification; however, since the human classification is also error prone, as discussed in Section 1, their algorithm also reproduces the errors in the human classification.

Freeman et al. (2013) introduced MID (multimode, intensity and deviation) statistics designed to detect disturbed morphologies, and then classified 1639 galaxies observed with the Hubble Space Telescope WFPC3 with a random forest. It is one of the few works that state the detailed classifier performance, in terms of the confusion matrix coefficients.

3. DATA AND SAMPLE SELECTION

We use several databases derived from SDSS DR7 (Abazajian et al. 2009), for which we analyze \( r \) band images. They are: the Baillard et al. (2011) database (hereafter EFI); the Nair & Abraham (2010) database (hereafter NA); and the SDSS DR7 complete Legacy database and a volume limited subsample, hereafter referred as LEGACY and LEGACY–\( z_r \), respectively. We also use the Galaxy Zoo collaborative project visual classification (Lintott et al. 2008, 2011). The number of galaxies in the original databases, those successfully processed by MORFOMETRYKA, and those that have a Galaxy Zoo classification are listed in Table 3.

The databases are used with different purposes, namely, training, validation, and classification. In training phase, the galaxies in the databases that have a Galaxy Zoo classification are used to train a classifier machine (Section 6.2). For validation, we use a cross validation scheme to attest how well our classifier performed compared to Galaxy Zoo human classification. In the classification stage, we use the trained classifier to linearly separate LEGACY galaxies in two classes (elliptical—E or spiral—S) in the morphometric parameters space (Section 6.2). The galaxy distance to the separating hyperplane is then proposed as a morphometric index \( M_f \) (Section 7). The databases EFI and NA, for which we have \( T \)-type values, are further used to support out argument that \( M_f \) based on the classifier discriminant function, can reflect the galaxy morphological type.

The classification scheme from the Galaxy Zoo project (Lintott et al. 2008, 2011) was used to train our supervised morphometric classifier. The Galaxy Zoo project provides simple morphological classifications of nearly 900,000 galaxies drawn from the SDSS–DR6.

Below we discuss each sample in detail.

3.1. The EFI Catalog

The EFI catalog was specifically designed to sample all Hubble morphological types. It provides detailed
morphological information of galaxies selected from standard surveys and catalogs (Principal Galaxy Catalogue, Sloan Digital Sky Survey, Value-Added Galaxy Catalogue, HyperLeda, and the NASA Extragalactic Database). The sample is essentially limited in apparent diameter, and offers a detailed view of the whole Hubble sequence. The final EFIGI sample comprises 4458 galaxies for which there is imaging in all the \textit{ugriz} bands in the SDSS-DR4 database. For these galaxies, the EFIGI reference dataset provides visually estimated morphological information as well as re-sampled SDSS imaging data. The photometric catalog is more than 80% complete for galaxies with $10 < m_{\text{petro},g} < 14$, where $m_{\text{petro},g}$ is the Petrosian magnitude in the $g$ band.

3.2. The NA Sample

Nair & Abraham (2010) provide detailed visual classifications for 14,034 galaxies selected from the SDSS spectroscopic main sample described in Strauss et al. (2002). They used the SDSS-DR4 photometry catalogs to select all spectroscopically targeted galaxies in the redshift range $0.01 < z < 0.1$ down to an apparent extinction-corrected magnitude limit of $g < 16$ mag. Objects mistakenly classified as galaxies have been removed, leading to the final sample of 14,034 galaxies. Their final catalog provides $T$-types, the existence of bars, rings, lenses, tails, warps, dust lanes, arm flocculence, and multiplicity for all galaxies.

3.3. The SDSS Legacy and Legacy-zr Samples

Our target sample of galaxies was retrieved from SDSS-DR7 (Abazajian et al. 2009) by selecting all objects spectroscopically classified as galaxies (see Appendix A.2 for a full query). SDSS Frames and psFields were obtained and stamps and PSF (point-spread function) were generated from them (see Appendix A for details). Our final catalog comprises 804,974 objects.

The subsample LEGACY–zr is volume limited at redshift $z < 0.1$ and $m_{\text{petro},r} < 17.78$, where $m_{\text{petro},r}$ is the extinction corrected Petrosian magnitude in the $r$ band. This magnitude limit roughly corresponds to the magnitude at which the SDSS spectroscopy is complete (Strauss et al. 2002). The redshift limit of $z < 0.1$ provides a complete sample for $M_{\text{petro},r}$, where $M_{\text{petro},r}$ is the SDSS Petrosian absolute magnitude in the $r$ band.

For 570,685 galaxies, those for which zWarning $= 0$ in the SDSS database, we derived ages, metallicities, stellar masses, and velocity dispersions using the spectral fitting code STARLIGHT (Cid Fernandes et al. 2005). Before running the code, the observed spectra are corrected for foreground extinction and de-redshifted, and the single stellar population (SSP) models are degraded to match the wavelength-dependent resolution of the SDSS spectra, as described in la Barbera et al. (2010). We adopted the Cardelli et al. (1989) extinction law, assuming $R_V = 3.1$.

We used SSP models based on the Medium resolution INT Library of Empirical Spectra (Sánchez-Blázquez et al. 2006), using the code presented in Vazdekis et al. (2010), using version 9.1 (Falcón-Barroso et al. 2011). They have a spectral resolution of $\sim 2.5$ Å, almost constant with wavelength. We selected models computed with a Kroupa (2001) universal IMF with slope $= 1.30$, and isochrones by Girardi et al. (2000). The basis grids cover ages in the range of 0.07–14.2 Gyr, with constant log(Age) steps of 0.2. We selected SSPs with metallicities $[\text{M}/\text{H}] = \{-1.71, -0.71, -0.38, 0.00, +0.20\}$.

4. Quantitative Galaxy Morphology

The basic morphometric measurements of the CAGM system are fully described by Abraham et al. (1994, 1996), Bershady et al. (2000), Conselice et al. (2000), and Lotz et al. (2004), among others. Relevant modifications of these parameters and the new parameters introduced in this work are discussed in this section.

We define the region with the same axis ratio and position angle as the galaxy (see Section A.2) and with major axis equal $N_R R_p$, where $R_p$ is the Petrosian Radius and $N_R = 2$, as the Petrosian region. Most measurements are made with pixels in this region, except if otherwise stated. A central region of the size of the PSF FWHM is masked before calculating $A$, $S$, and $\sigma_v$.

4.1. Concentration $C_1$ and $C_2$

The concentration index $C$ is the ratio of the circular radii containing two fractions of the total flux of the galaxy (Kent 1985), where these percentages are chosen to maximize the distinction between systems and minimize seeing effects. The concentration depends on the determination of the radius that contain some fraction of some measure of the total luminosity of the galaxy. In this work, we have adopted the Petrosian luminosity as the total luminosity $L_T$, which is the maximum value of $L(R)$ inside the Petrosian region. The measured $L(R)$ is spline interpolated and then the point where it attains some fraction $f$ of $L_T$ is found by evaluating the spline at the point. In this way we obtain $R_{20}$, $R_{50}$, $R_{80}$, and $R_{90}$ and finally

\[ C_1 = \log_{10} \left( \frac{R_{80}}{R_{20}} \right) \]  

and \[ C_2 = \log_{10} \left( \frac{R_{90}}{R_{50}} \right). \]

Note that we dropped the factor 5 that is usually in the definition of $C$, so that all morphometric measurements used will fall approximately in the range $[0, 1]$, and thus statistic standardization would have little effect and may be optional. The concentration $C_1$ is more sensitive to the seeing effect that is more pronounced in the central regions and thus on $R_{20}$; $C_2$ is more sensitive to the noise that is more important in the outer regions and thus on the measure of $R_{90}$.

4.2. Asymmetry $A_1$, $A_2$, $A_3$

The asymmetry coefficient $A$ is determined comparing a source image with its rotated counterpart. We measure the asymmetry $A_1$, as defined by Abraham et al. (1996), with the exception that we do not subtract the background asymmetry, for we find that this procedure makes the asymmetry estimation unstable and sensitive to the selected sky (and hence to the stamp) size. Instead, we consider only the galaxy portion inside the Petrosian region. Even so, $A_1$ depends heavily on the noise and on the image sampling. To address this problems we used two new asymmetry measurements defined as

\[ A_2 = 1 - r(I, I_x) \]  \hspace{2cm} (1) 

and \[ A_3 = 1 - s(I, I_x), \]  \hspace{2cm} (2)
where $r()$ and $s()$ are the Pearson and Spearman correlation coefficients (Press et al. 2002), respectively. $I$ is the image and $I_r$ is its $\pi$-rotated version. The rationale behind this formulation is that those pixels made up mostly of noise will not contribute to $A_2$ and $A_3$ since the correlation between them will tend toward zero. Furthermore, correlation coefficients are more immune to convolution and thus less affected by seeing effects. The Pearson coefficient tends to accumulate close to unity, so $A_3$ has proven not so useful as $A_1$ and $A_3$. The center of rotation is chosen to minimize the asymmetry measurements.

### 4.3. Gini Coefficient

The Gini coefficient $G$ measures the flux distribution among the pixels of a galaxy image. The Gini coefficient for the image pixels in the Petrosian region is calculated exactly as shown in Lotz et al. (2004), i.e., for $n$ pixels with values $I_i$ in increasing order we have

$$G = \frac{1}{n(n-1)} \sum_{i=1}^{n} (2i - n - 1)I_i,$$

where $\bar{I}$ is the average value.

### 4.4. Smoothness

The smoothness coefficient $S$ (a.k.a. clumpiness) in general measures the small scale structures in the galaxy image. Here we consider three different measures of smoothness. $S_1$ is calculated as shown in Lotz et al. (2004), except that the filter used is a Hamming window (Hamming 1998) with size $[R_0/4]$. Following the same reasoning as for asymmetry, we define the modified smoothness $S_2$ and $S_3$ as

$$S_2 = 1 - r(I, I^F)$$

and

$$S_3 = 1 - s(I, I^F)$$

where $I^F$ is the filtered image. As with asymmetry, $S_3$ has proven to be more useful than $S_2$.

### 4.5. Entropy

The entropy of information $H$ (Shannon entropy, e.g., Bishop (2007)) is used here to quantify the distribution of pixel values in the image. For a random variable $I$, the entropy $H(I)$ is the expected value of the information $\log[p(I)]$.

$$H(I) = -\sum_k p(I_k) \log[p(I_k)],$$

where $p(I_k)$ is the probability of the occurrence of the value $I_k$, $k$ refers to a specific value, and $K$ is the number of bins considered. For discrete variables, $H$ reaches the maximum value for a uniform distribution, when $p(I_k) = 1/K$ for all $k$ and hence $H_{\text{max}} = \log K$. The minimum entropy is that of a delta function, for which $H = 0$. We then have the normalized entropy

$$\widehat{H}(I) = \frac{H(I)}{H_{\text{max}}} \quad 0 \leq \widehat{H}(I) \leq 1.$$  

Smooth galaxies will have low $H$, while clumpy galaxies will have high $H$.

### 4.6. Spirality $\sigma_\psi$

None of the CASGM parameters take into account the spiral arms, rings, and bars in galaxies, albeit they are a major and important emphasis of human-based classification. We devise a parameter to take it into account. As done in Shamir (2011), we first transform the standardized galaxy image to polar coordinates $(r, \theta)$. In $(r, \theta)$ space a bulge appears as a band in the lower region of the diagram, a bar as two vertical lines, and spiral arms as inclined bands. See Figure 1. We then calculate the gradient magnitude $|\nabla I|$ and direction $\psi$ fields of the polar image. Most points in this direction field $\psi$ for an elliptical galaxy will point to the bottom, while for a spiral galaxy there will be many orientations corresponding to arms, rings, and bars. The standard deviation $\sigma_\psi$ for the field direction values will be smaller for an elliptical compared to that of a spiral, and hence can be used to estimate the amount of characteristic structures. To avoid regions of noise, we make the measurements in regions where the gradient magnitude is greater than the median of the magnitude field.

Figure 2 shows a density plot of $\sigma_\psi$ versus $T$-type for the EFIGI database (a subsample of galaxies were selected such that there are 45 objects in each $T$-type), where a clear linear relationship is seen. So, $\sigma_\psi$ is a good diagnostic for $T$-type, provided there is enough spatial resolution to distinguish spiral arms, rings, and bars. This is the case for the EFIGI database, marginally for NA and not for LEGACY in general, as inferred from the discussion in Section 6.1 and Figure 4.

### 5. THE MORFOMETRYKA ALGORITHM

We developed a standalone application to automatically perform all the structural and morphometric measurements over a galaxy image, called MORFOMETRYKA\(^4\) (MFMTK). MFMTK reads the input stamp image and related PSF for a given galaxy and performs various measurements explained in detail in Appendix A. MFMTK is currently implemented in an object-oriented fashion in Python 2.7,\(^5\) with the aid of scientific libraries SciPy and Numpy (Oliphant 2007), Matplotlib (Hunter 2007) and PyFits.\(^6\) Figure 3 is an example output of MFMTK run for EFIGI data of galaxy PGC 9445.

The MORFOMETRYKA basic output is: sky background value and standard deviation; image centers $(x_0, y_0)_{\text{CoL}}, (x_0, y_0)_{\text{peak}}$; Sérsic parameters for 1D surface brightness profile fitting $(I_{s1D}, R_{s1D}; n_{1D})$ and for 2D image fitting $(I_{s2D}, R_{s2D}, n_{2D}, q_{2D}, PA_{2D}, (x_0, y_0)_{2D})$, Petrosian Radius $R_{p}$; radii $R_2$, $R_{50}$, $R_{80}$, $R_{90}$, and concentrations $C_1$ and $C_2$; asymmetries $A_1$, $A_2$, $A_3$, and fitted center for $A_1$ and $A_3$; smoothness $S_1$ and $S_3$; Gini coefficient $G$; second moment $M_{20}$; gradient field direction value $\psi$ and standard deviation $\sigma_\psi$; quality flags QF. Optionally, all maps (star masks, segmentation map, polar image and so on) are saved. MORFOMETRYKA takes about 12 seconds to process a $256 \times 256$ galaxy and $45 \times 45$ PSF image on a 2.5 Ghz processor. The version used in this work was 5.0.

### 6. SUPERVISED CLASSIFICATION

Our morphological classification is based on the linear discriminant method that separates galaxies in two main classes

\(^4\) [http://morfometryka.ferrari.pro.br]

\(^5\) Python Software Foundation. Python Language Reference, version 2.7. Available at [http://www.python.org]

\(^6\) PyFits is a product of the Space Telescope Science Institute, which is operated by AURA for NASA.
in morphometric parameter space. We train the classifier using the classification from the Galaxy Zoo (Lintott et al. 2008, 2011). The process was done independently for EFIGI, NA, LEGACY, and LEGACY–zr data sets.

Our main goal is not only to classify galaxies in a way that reproduces the human classification but also to establish a basis for a morphometric space where galaxy classes are separated, allowing further studies where the human classifier cannot be used. Thus, we use a linear discriminant and also we seek the smallest set of independent parameters that may yield a reliable classification that is physically meaningful.

### 6.1. Feature Selection

Given that we have so many measured quantities for each galaxy, some of them may be redundant or irrelevant and we need to select those which are more relevant to the classification algorithm. Many feature selection algorithms tend to diminish the importance of quantities that correlate with each other. A criterion that avoids that is the maximum information content (MIC, Albanese et al. 2013). MIC is based on the mutual information and the information entropy: it compares, given the parameters and the known class, which one possesses the greater mutual information with the class variable, i.e., which one will have greater impact in the classification. The normalized values for MIC are shown in Figure 4.

We may have more information on how efficiently each feature helps to separate the classes by examining the feature histogram separated by classes, as shown in Appendix D, in Figure 10 (EFIGI), Figure 11 (NA), Figure 12 (LEGACY) and Figure 13 (LEGACY–zr). We must be aware that we are seeing marginal probability distribution functions (PDF) on each variable and this is not equivalent to analyze the multivariate PDF of all parameters together.

First, we note that the features with the highest discriminant power are those related to the light concentration (Sérsic n, C1 and C2). Since they are equivalent (Appendix C), and also equivalent to the Petrosian Radii (Appendix B), we retain only C1 and C2, which are not parametric and more robust.

Comparing the asymmetry measures in the histograms in Appendix D, we see that A3 is able to discriminate classes better than A1, which is confirmed by the MIC values. The Gini coefficient is very poor at separating E from S, as it is M20. Compared to Gini, entropy H works better in separating classes, and the introduced S3 is better than the original S1. The axis ratio q is good for E but indifferent for S, so it is not used. The spirality $\sigma_\psi$ is also good to discriminate classes but it is crucially dependent on angular resolution—its importance decreases from EFIGI to NA, to LEGACY, in the same sense as the mean angular resolution decreases.

Finally, based on the MIC analysis, we choose this set of parameters

$$x = \{ C_1, A_3, S_3, H, \sigma_\psi \},$$

for they constitute a minimal set of independent parameters that yield a reliable classification. Four of the chosen parameters are new, used here for the first time. A3 and S3 are enhanced

---

**Figure 1.** Illustration about how spirality is measured. EFIGI spiral SB PGC 2182 is above and lenticular PGC 2076 is below. From left to right: original image, standardized image ($q = 1$, PA = $0^\circ$), image in polar coordinates, gradient field of polar image.

**Figure 2.** Density plot for $\sigma_\psi$ vs. $T$-type for the EFIGI sample. For this plot, a subsample of galaxies were selected so that there are 45 galaxies in each $T$-type.
versions of standard parameters, $H$ is first applied in morphometric studies, and $\sigma_{\psi}$ is completely new.

### 6.2. Linear Discriminant Analysis

A simple linear classifier may be represented by a discriminant function, which for a given input vector $\mathbf{x}$ that contains $d$ morphometric measurements (Duda et al. 2000; Bishop 2007) gives

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0,$$

where $\mathbf{w}$ is the weight vector and $w_0$ is the threshold. The input vector is assigned to the class $\mathcal{C}_1$ if $f(\mathbf{x}) > 0$ and to $\mathcal{C}_2$ otherwise. The decision boundary or surface is a hyperplane defined by $f(\mathbf{x}) = 0$, for which $\mathbf{w}$ is a normal vector and $-w_0/||\mathbf{w}||$ is its normal distance to the origin. The decision function corresponds to the perpendicular distance from $\mathbf{x}$ to the decision surface.

When using the Bayes Decision Theory the expressions for $\mathbf{w}$ and $w_0$ are assigned as follows: an object belongs to class $\mathcal{C}_1$ if

$$P(\mathcal{C}_1, \mathbf{x}) > P(\mathcal{C}_2, \mathbf{x}) \quad \text{(for class $\mathcal{C}_1$)}$$

and to $\mathcal{C}_2$ otherwise. Since the evidence $p(\mathbf{x})$ is the same for both classes, the Bayes rule in Equation (8) is equivalent to

$$p(\mathbf{x}, \mathcal{C}_1) P(\mathcal{C}_1) > p(\mathbf{x}, \mathcal{C}_2) P(\mathcal{C}_2),$$

where $p(\mathbf{x}, \mathcal{C}_i)$ is the class conditional probability density function (CCPDF) and $P(\mathcal{C}_i)$ the prior. We assume that the CCPDF is multivariate normal density

$$p(\mathbf{x}, \mathcal{C}_i) = \frac{1}{(2\pi)^{d/2} |\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i) \right],$$

where $\mu_i$ are the mean and $\Sigma_i$ is the covariance matrix of $\mathbf{x}$ for class $\mathcal{C}_i$. The decision rule Equation (9), or equivalently its
logarithm, is then
\[- \frac{1}{2} (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) + \ln P(C_1) >
\]
\[- \frac{1}{2} (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) + \ln P(C_2). \]

The terms involving \(x'\Sigma^{-1}x'\) are general quadratic forms and if we expand them in Equation (11) we have a quadratic classifier. But instead, we consider identical covariance matrices \(\Sigma_1 = \Sigma_2 = \Sigma\), which yield a linear classifier. Expanding the terms in Equation (11), and ignoring those that are identical for both classes, we have
\[
\Sigma^{-1}(\mu_1 - \mu_2) + \frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1
\]
\[
+ \frac{1}{2} \mu_2^T \Sigma^{-1} \mu_2 + \ln \frac{P(C_1)}{P(C_2)} > 0. \]

If we refer to Equations (7) and (12), we then have
\[
w = \Sigma^{-1}(\mu_1 - \mu_2) \]
\[
w_0 = \frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2} \mu_2^T \Sigma^{-1} \mu_2 + \ln \frac{P(C_1)}{P(C_2)}, \]

which completes our linear classifier.

### 6.3. Classifier Performance

We may estimate the classifier performance by means of a confusion matrix or contingency table, which is a comparison of the actual class with the predicted class for each object. The performance is evaluated calculating several scores based on true positives (TP, hits), true negatives (TN, correct rejections), false positives (FP, false alarms) and false negatives (FN, misses). See, for example, Hackeling (2014).

The accuracy \(A\) is the fraction of hits relative to the total number of classifications
\[
A = \frac{TP + TN}{TP + TN + FP + FN};
\]
Precision \(P\) is the fraction of positive predictions that are correct
\[
P = \frac{TP}{TP + FP};
\]
Sensitivity \(R\) is the fraction of the truly positive instances that the classifier recognizes
\[
R = \frac{TP}{TP + FN};
\]
and \(F_1\) score is the harmonic mean between sensitivity and precision
\[
F_1 = \frac{2TP}{2TP + FP + FN}.
\]

We test the performance of the classifier by a tenfold cross validation: for each database we selected those samples with known Galaxy Zoo classifications and partitioned them into 10 parts; in each of 10 runs 1 of the parts was used as a validation sample and the other 9 parts as training samples. In each run, the scores \(A, P, R, F1\) were calculated; their final averages are shown in Table 2. For all databases the classifier usually performs better than 90%, namely 90% of the time the automated classifier agrees with the visual classification. If we consider that the performance in the human classification is also of that order, and that those classifications were used to train the classifier, then this performance can be considered very good and it is the best figure that we could expect without using a classifier that would incorporate the errors in it.

### 7. MORPHOMETRIC INDEX

As stated in Section 6.2, the discriminant function \(f(x)\) is the distance of \(x\) to the plane that separates classes, here ellipticals and spirals. Based on that, we propose to use \(f(x)\) to represent the galaxy type, which we call the morphometric index \(M_i\). Figure 5 shows the comparison of \(M_i\) with the \(T\) type from EFIGI and NA samples. There is a clear linear relationship between \(M_i\) and \(T\) type, justifying the use of \(M_i\) as a morphometric index. In Section 8 we extend this argument by comparing \(M_i\) with other galaxy physical characteristics. By construction, \(M_i\) is negative for early-type and positive for late-type galaxies.

A linear regression between \(T\) and \(M_i\) could be used to calibrate \(M_i\) as an inferred \(T\)-type, but since \(T\) is a subjective parameter we prefer to maintain \(M_i\) in its own scale, and as a pure morphometric measure, estimated solely based on the values of \(x\). For a binary classification, the magnitude of the direction vector \(w\) has no importance, and in fact, \(\|w\|\) could be different depending on the details of the linear discriminant analysis (LDA). But since we want to use this distance as a physical measure, we prefer to normalize \(w\) and the distance to the plane \(w_0\) in the morphometric index \(M_i\), so that
\[
M_i = \hat{w}^T x + \hat{w}_0,
\]
where
\[
\hat{w} = \frac{w}{\|w\|} \quad \text{and} \quad \hat{w}_0 = \frac{w_0}{\|w\|}.
\]

The final values for the LEGACY database are
\[
\hat{w} = \{-0.832, 0.249, 0.451, 0.190, -0.079\}
\]
\[
\hat{w}_0 = 0.018
\]

As we see in Equation (14), \(w_0\) depends on the class priors \(P(C_i)\). Usually, the relative frequency for each class \(N_i/N\) (\(N\), the number of objects in class \(C_i\), \(N\) total number of objects) is used as the prior, since it gives the probability that a new object belongs to class \(C_i\) if we know nothing about it. But in the EFIGI and NA databases, the relative frequency is biased, as it was designed to contemplate each morphological \(T\) type with approximately the same number of objects. So, the priors and hence \(w_0\) for EFIGI and NA could not be applied to other databases with different relative frequencies. LEGACY and LEGACY–\(z\) databases, on the contrary, have priors that may reflect real distribution of classes of galaxies, since no morphology was used to select the objects. Briefly, the intercept term in the linear relationships in Figure 5 may be
biased by the selection effects in the databases; the linear character, however, is unaffected.

The linear regression between $T$-type and $M_i$, shown in Figure 5, is a least square solution using a robust Theil–Sen estimator, which computes the median slope among all pairs of points in a set, implemented in Scikit-Learn library (Pedregosa et al. 2011). Note that $T \approx 20 M_i$.

In order to have $M_i$ as a trustable morphological indicator we need to establish how accurate it is, which in principle can be done by propagating the error from each of the parameters $C_1, A_3, S_3, H$, and $\sigma_{\psi}$. However, we find it more realistic to compute the signal-to-noise ratio of the galaxy as $S/N = I_{2D}/skybgstd$ (see Appendix A) and see how $M_i$ varies with it. As we can see from Figure 6, there seems to be no trend between them and the blue area indicates that most galaxies have $S/N$ around 10 and average $M_i$ slightly around 0.1, which is slightly above 0.0, where we would expect.

8. COMPARISON WITH OTHER PHYSICAL PARAMETERS

We present in this paper a new approach for galaxy morphological classification that is not focused on recovering visual classification, although this is done remarkably well (see Section 6.3). In this work, the parameters defining the morphology of a galaxy are physically motivated and to confirm how successful we were in reaching this goal we compare $M_i$ with quantities measured from the spectrum of the galaxies we measured.

Figure 7 exhibits how $\text{Age}_L$ (age weighted by luminosity), $\text{Age}_M$ (age weighted by mass), eClass (a single parameter classifier based on PCA analysis, retrieved from the SDSS database), and central velocity dispersion $\sigma$ correlate with $M_i$.

Notice that the SDSS spectra reflect properties of the central region of galaxies. In Figure 7(a) we see that, overall, negative $M_i$ corresponds to older systems. $\text{Age}_L$ reflects more recent episodes of star formation and in this case, as $M_i$ goes to very negative, the systems do not present any recent star formation, namely these are very old galaxies. There is a ridge of old systems extending from $M_i = -0.5$ up to 0.1 and then a significant drop in age as $M_i$ tends to 0.2. For $0.1 < M_i < 0.4$ we see that $\text{Age}_L$ is around 1.5 Gyr. These are the late type spirals which exhibit a...
considerable amount of star formation. This figure clearly shows that morphology, in this case manifested by the parameter $M_i$, varies continuously from an old to a young stellar population, which is an important aspect of any morphological quantifier—to reflect stellar population properties of galaxies.

Figure 7(b) is similar to Figure 7(a) but plots $\text{Age}_M$ instead, which, contrary to $\text{Age}_L$, reflects the whole star formation history of a galaxy. The same trend is seen here, however, since the SDSS spectra samples only the central region of the galaxies, there is a population that dominates the figure for $\text{Age}_M$ around 10.0 Gyr, from early (negative $M_i$) to late types (positive $M_i$).

Figure 7(c) exhibits how $M_i$ is related to eClass, a parameter designed to express differences in stellar populations among different galaxies and then serve as a discriminant between early and late type systems. We find that the relation between these two quantities is not linear, which is what we would expect if both reflected morphology in a one-to-one relationship. What we see is that for $-0.5 < M_i < 0.2$ eClass is concentrated around $-0.15$ (early type systems), with a scatter that increases as $M_i$ increases. Then for $M_i > 0.2$ eClass increases steadily, reaching an eclass of 0.5 for $M_i > 0.4$. Both eClass and $M_i$ are associated with morphology, although eClass is primarily associated with stellar population and $M_i$ is derived solely based on image morphometry. $M_i$ is more sensitive to morphology, particularly in the early-type systems domain ($M_i < 0$).

Finally, in Figure 7(d) we present the relation with the central velocity dispersion $\sigma$ (corrected for an aperture of $R_e/8$, where $R_e$ is the effective radius of the galaxy) even though, with a large scatter, a clear relation exists between $M_i$ and $\sigma$, which is remarkable considering that $M_i$ is solely photometric. In summary, these comparisons show that $M_i$ is reliable in separating the different morphological types according to their

Figure 7. Relationship between $M_i$ and $\text{Age}_L$ (age weighted by luminosity), $\text{Age}_M$ (age weighted by mass), eClass, and central velocity dispersion $\sigma$. 

The Astrophysical Journal, 814:55 (15pp), 2015 November 20
Ferrari, de Carvalho, & Trevisan
stellar population properties, a performance not seen in other previously proposed morphological quantifiers.

9. SUMMARY

We present a new method to establish morphological classification of galaxies that is physically motivated although it matches what is done visually for the very nearby universe equally well. In the following, we summarize the main aspects of the classification system proposed here and the verification analysis.

1. We developed a pipeline that automatically estimates morphometric parameters from galaxy images. Measured parameters include concentration, $C_1$, asymmetry, $A_2$, and smoothness, $S_1$, which were slightly modified with respect to the conventional ones. We also make use of two new extra parameters: entropy $H$ and sparsity $\sigma_s$.

2. MORFOMETRYKA measures several quantities per galaxy, which brings the question of which ones are more adequate for establishing the morphological type of the system. We use a method called MIC to select the relevant features avoiding redundancy. The new introduced morphometric parameters have a better discriminant power than previously used ones. MIC analysis resulted in the minimum number of independent parameters listed in item 1. The relationship between concentration, Petrosian radius, and Sérsic index $n$ is derived in Appendices B and C.

3. Our supervised classification is based on Galaxy Zoo and tested with different data sets: EFIGI, NA, LEGACY, and LEGACY–$\tau r$. The LDA method is used to determine the decision surface that separates early from late type systems and the distance from this surface will indicate how early or late the system is. It is exactly this distance that we propose as a morphological index, $M_i$.

4. Classification performance was evaluated using the confusion matrix, from which we measured accuracy, precision and sensitivity scores, with a tenfold cross validation scheme. We obtain final scores better than 90%.

5. Another independent validation comes from comparing $M_i$ with stellar population quantities and velocity dispersions that were established using the spectra available in DR7 together with the spectral fitting code STARBURST99. We note that $M_i$ correlates with eClass and it shows that classifying early-type galaxies solely as eClass < 0 can significantly contaminate the sample with late-type systems that have $M_i > 0.2$.

We thank the referee Ronald Buta for many suggestions that helped improve the manuscript. We also thank Karina Machado and Diana Adamatti (C3-FURG) for lending us their processing infrastructure; Juliano Marangoni (IMEF-FURG) for discussions on statistics; and Valérie de Lapparent (IAP-Paris) for help with EFIGI data. R.R.d.C. would like to thank Francesco La Barbera for insightful discussions on the topic over the years. R.R.d.C. acknowledges financial support from FAPESP through a grant # 2014/11156-4. M.T. acknowledges financial support from FAPESP (process #2012/05142-5) and CNPq (process #204870/2014-3).

APPENDIX A

MORFOMETRYKA ALGORITHM DETAILS

Here, we provide a detailed description of the various measurements in MORFOMETRYKA. MFMTK is logically divided into four main blocks (classes in programming parlance): STAMP—basic data reading and low level, low complexity geometrical measurements; PHOTOMETRY—luminosity distribution, star masking and Petrosian radius estimation; SÈRÈSIC—1D and 2D luminosity distribution fitting; MORPHOMETRY—measurements of the morphometric parameters used later on for establishing the galaxy’s morphology. The package also includes auxiliary applications MAKEMYSDSS for retrieving SDSS frames and cutting stamps and LDA CLASSIFY to perform the LDA. In the following, logical units are written in SMALL CAPS, and algorithm code is in typewriter.

A.1. Cutting Stamps

The list of all objects, containing ObjIDs, RA, DEC, run, rerun, camcol, field, and petroRad, is generated with the following SQL query on SDSS CasJobs.

```
SELECT p.objID, p.ra, p.dec, p.run, p.rerun, p.camcol, p.field, p.petroRad_r
FROM DR7.SpecObj as s JOIN DR7.PhotoObj AS p ON s.
    bestObjID = p.objID
WHERE s.specclass = 2.
```

From this list, we build a set of unique combinations of (run, rerun, camcol, field) and the required SDSS Frames and psFields are downloaded. We do it exactly as for DR7 Frames and psFields but we download the DR10 files since they refer to the same region of the sky, i.e., the raw data are the same, but the image processing algorithms were improved from DR7 to DR10. Also, DR10 frames are calibrated in nanomaggies7 (Lupton et al. 1999). For each object, the relative Frame is loaded and a square region of size 10 petroRad_r centered in the object’s R.A. and Decl. is cut. The PSF for the same position is generated with the SDSS read_PSF application from the psField file. The stamp FITS file header is updated with the astrometry and relevant frame keywords. If the object is in the frame border, i.e., if it has less than 90% of pixels in the frame, a FITS header keyword FLAGINC and a header comment “MFMTK: incomplete stamp” are written.

A.2. Basic Image Processing

The process starts with the target image gal0 and the associated PSF, which is measured from the second moment collapsed in the y direction. The sky background $skybg$ is estimated from the median of all pixels from the four corners of the image (squares of typical width of 10 pixels, $skyboxsz$). The accuracy of the sky background estimate $skybgstd$ is set by the standard deviation of the aforementioned set of pixels.

The segmentation is done on the gal0itr image, which is the gal0 image median filtered with a window of size segS (typically 5 pixels); high frequencies are filtered from the image to avoid sharp edges in the segmented regions. Regions are then selected by histogram thresholding: those pixels whose intensity are greater than the threshold $\text{median}(gal0itr)+\text{segK}$ mad

---

7 https://www.sdss3.org/dr10/algorithms/magnitudes.php
8 The median estimate is less affected than the mean by outliers, and such sky background estimate has proven to be accurate enough even if a star occupies several pixels in one of the corners.
which is similar to sigma-clipping consists of one or more labeled regions. The pixels receive the same label. At this stage the segmentation connected. The spatial information is taken into account by performing a connected-component labeling, where 4-connected pixels receive the same label. On the other hand, pixels outside the segmentation region are nil. Geometric measurements in this section are done in the gal0seg image.

The galaxy image center is estimated in two distinct ways. First, the peak center \((x_0, y_0)_{\text{peak}}\), referring to the locus where the intensity is maximum, is estimated from the CoL (first moment) of the \(5 \times 5\) matrix around the pixel with the highest intensity, attaining sub-pixel precision.

For an image \(I(x, y)\) the standard image moments are defined as

\[
m_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p y^q I(x, y) \, dx \, dy,
\]

and the CoL \((x_0, y_0)_{\text{CoL}}\) are given by \(x_0 = m_{01}/m_{00}\) and \(y_0 = m_{01}/m_{00}\). The translational invariant moments \(m_{pq}\) are determined by replacing \(x\) and \(y\) by \((x - x_0)\) and \((y - y_0)\), respectively, in \(m_{pq}\). The axis lengths are given by

\[
\lambda_1 = \sqrt{\frac{\mu_{20} + \mu_{02} + \Lambda}{m_{00}}} \quad \lambda_2 = \sqrt{\frac{\mu_{20} + \mu_{02} - \Lambda}{m_{00}}}
\]

where

\[
\Lambda = \sqrt{\left(\mu_{20} - \mu_{02}\right)^2 + 4\mu_{11}^2}
\]

from which we define

\[
a = \max(\lambda_1, \lambda_2) \quad b = \min(\lambda_1, \lambda_2).
\]

Furthermore, we can calculate the position angle of the main axis by

\[
PA = \frac{1}{2} \arctan(2\mu_{11}/\mu_{20} - \mu_{02}).
\]

The 1D Sérsic parameters are measured by fitting the Sérsic law (Sérsic 1968)

\[
I(R) = I_n \exp \left[ -b_n \left( \frac{R}{R_n} \right)^{\nu} - 1 \right]
\]

with \(b_n = 2n - \frac{1}{3}\), to the 1D surface brightness profile \(I(R)\). The minimizations are done with a Levenberg–Marquardt algorithm, in a least-squared sense. The fits are bounded by adding a square penalty function for parameters outside the specified range. The boundaries are: \(\min[I(R)] < I_{n,1D} < \max[I(R)]\), \(1 < R_{e,1D} < \max[R]\), \(\frac{1}{2} < n_{1D} < 50\). The output parameters are \(I_{n,1D}, R_{e,1D}, n_{1D}\).

The 2D fitting applies Equation (20), convolved with the PSF and with \(R\) replaced by \(R = \sqrt{x'^2 + y'^2/q_{2D}^2}\), where

\[
x' = (x - x_{0,2D}^0) \cos(\text{PA}_{2D}) - (y - y_{0,2D}^0) \sin(\text{PA}_{2D})
\]

\[
y' = (x - x_{0,2D}^0) \sin(\text{PA}_{2D}) + (y - y_{0,2D}^0) \cos(\text{PA}_{2D}).
\]

Coordinates \(x, y\) refer to positions in the galaxy image. The two-dimensional Sérsic function is fitted directly to the galaxy image, except that pixels outside the galaxy, as defined by the Petrosian Region, flagged stars and central circular region of 1 PSF FWHM are masked. The algorithm is the same as for 1D fitting, with the following boundaries: the center \((x_0, y_0)_{2D}\) cannot vary more than 15% compared to \((x_0, y_0)_{\text{peak}}\); \(I_{n,2D}\) must be within image pixel values range; \(R_{e,2D}\) cannot be greater than the image half-diagonal; \(\frac{1}{2} < n_{2D} < 20\) and \(q_{2D} = b/a < 1\). This setup has been proven in simulation and in real galaxies to be the most stable, converging for most galaxies in the samples. The fit free parameters are \(\{x_0, y_0, \text{PA}, q, I_n, R_n, n\}_{2D}\).

Based on simulations of synthetic Sérsic galaxies, we found that the 2D fitting is better than 1D at recovering “true” parameters from images. Since the 2D is more unstable to initial parameters, we use the 1D results as the initial guess for the 2D fit.
A.5. Quality Flags

For reference, a series of conditions is evaluated and informative Quality Flags (QF) are saved. They are not conclusive but may indicate situations when the condition occurs. For example, if for a given object the $R_{n,2D}$ is of the order of the PSF, $n_{2D} \sim 0.5$ (a Gaussian) and $b/a \sim 1$, the object is probably a star, a MORFOMETRYKA target selection error. Other QFs indicate that the fitting routine did not converge in situations of a crowded field. For detecting crowded fields, we define the asymmetry $A_4$ as the distance between $r_{\text{Col}} = (x_0, y_0)_{\text{Col}}$ and $r_{\text{peak}} = (x_0, y_0)_{\text{peak}}$ in units of $R_p$, in percentage,

$$A_4 = 100 \frac{r_{\text{Col}} - r_{\text{peak}}}{R_p},$$

which attains values greater than $\sim 10$ in crowded fields, and is used to turn the QF64 on.

The QFs are summarized in Table 3.

A.6. Petrosian Quantities

Petrosian (1976) defined a function $\eta(R)$ which is the ratio of the mean intensity inside $R$ to the intensity at the isophote $R$

$$\eta(R) = \frac{\langle I \rangle (R)}{I(R)}. \quad (23)$$

The Petrosian radius is the distance from the galaxy center where the fraction in Equation (23) has some constant value

$$\eta(R_p) = \eta_0.$$

Here we use $\eta_0 = 5$. The virtue of $\eta$ is that both the numerator and denominator have the same dependence with the distance, hence $\eta$ is distance independent. The Petrosian radius is used as an implicit scale length for each galaxy.

APPENDIX B
PETROSIAN RADIUS AND SÉRSIC INDEX EQUIVALENCE

The mean intensity within radius $R$ for a Sérsic model is the integrated luminosity up to $R$ by the region area $\langle I \rangle = L(R)/AA = \pi R^2$, so for $x = b_0(R/R_0)^{1/n}$, we have (see for example Ciotti & Bertin 1999; Graham & Driver 2005)

$$\langle I \rangle (R) = 2n I_n e^{b_0} \frac{\gamma(2n, x)}{x^{2n}}. \quad (24)$$

We then have for the Petrosian function Equation (23)

$$\eta(R) = \frac{2n\gamma(2n, x)}{x^{2n} e^{-x}}. \quad (25)$$

We have to solve

$$\frac{2n\gamma(2n, x_p)}{x_p^{2n} e^{-x_p}} = \eta_0 \quad \text{with} \quad x_p = x(R_p), \quad (26)$$

to obtain $R_p$ as a function of $n$. This equation is transcendental and can only be solved for $R_p$ numerically. However, for practical purposes, we can write an empirical Petrosian radius function

$$R_p(n) = R_n R_p^{n-n_0} \exp\left[\left(-\frac{n-n_0}{a}\right)^{\alpha}\right] \quad (27)$$

whose parameters $R_{\text{max}} = 5.8$, $n_0 = -1.11$, $a = 2.04$ and $\alpha = 0.8$ provide a fit better than 1% over the range $0.3 < n < 15$, as shown in Figure 8.
CONCENTRATION AND SÉRSIC INDEX EQUIVALENCE

In the case of the Sérsic law, the integrated luminosity within radius $R$ (Ciotti & Bertin 1999) is

$$L(R) = 2\pi n I_n R^2 \frac{e^b}{b^{2n}} \gamma(2n, x), \quad (28)$$

with $x \equiv b_n(R/R_n)^{1/n}$. Hence the total luminosity $L_T = L(R \to \infty)$ is

$$L_T = 2\pi n I_n R^2 \frac{e^b}{b^{2n}} \Gamma(2n). \quad (29)$$

From Equations (28) and (29) we have the equation for the $R_f$, which attains some fraction $f$ of the total luminosity

$$\gamma(2n, x_f) = f \Gamma(2n) \quad \text{with} \quad x_f = x(R = R_f) \quad (30)$$

Equation (31) cannot be solved analytically (except for $n = 1/2$) and the solution must be found numerically. Figure 9 shows the numerical solution for $1/2 < n < 15$.

Again, we can write an empirical function

$$C(n) = C'(\frac{n}{n'})^{\beta}, \quad (32)$$

which approximates the solution in the specified range with an error smaller than 2% for $C_1$ in the range $1 < n < 15$, with $C' = 2.91$, $n' = 32.44$ and $\beta = 0.48$. 

---

**Figure 10.** Distribution of feature values among morphometric classes for the EFIGI database.

**Figure 11.** Distribution of feature values among morphometric classes for the NA database.
APPENDIX D
HISTOGRAM OF MORPHOMETRIC PARAMETERS FOR DATABASES

Figures 11, 12 and 13 present the histogram for the parameters $A_1, A_3, C_1, C_2, \sigma_\psi, S_1, S_3, G, H$ and $M_{20}$ measured for the database EFIGI, NA, LEGACY, and LEGACY–zr, respectively. Red lines refer to elliptical galaxies and blue lines refer to spiral galaxies, as classified by Galaxy Zoo.

REFERENCES
Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2009, ApJS, 182, 543
Abraham, R. G., Valdes, F., Yee, H. K. C., & van den Bergh, S. 1994, ApJ, 432, 75
Abraham, R. G., van den Bergh, S., Glazebrook, K., et al. 1996, ApJS, 107, 1
Albanese, M. F., Visintainer, R., Riccadonna, S., et al. 2013, Bioinformatics, 29, 407
Andrae, R., Jahnke, K., & Melchior, P. 2011, MNRAS, 411, 385
Baillard, A., Berin, E., de Lapparent, V., et al. 2011, A&A, 532, A74
Bershady, M. A., Jangren, A., & Conselice, C. J. 2000, AJ, 119, 2645
Bishop, C. M. 2007, Pattern Recognition and Machine Learning (Berlin: Springer)
Cardelli, J. A., Clayton, G. C., & Mathis, J. S. 1989, ApJ, 345, 245
Cid Fernandes, R., Mateus, A., Sodré, L., Stasińska, G., & Gomes, J. M. 2005, MNRAS, 358, 363
Ciotti, L., & Berin, G. 1999, A&A, 352, 447
Conselice, C. J., Bershady, M. A., & Jangren, A. 2000, ApJ, 529, 886
de Vaucouleurs, G. 1959, HDP, 53, 275
de Vaucouleurs, G., de Vaucouleurs, A., Corwin, H. G., Jr, et al. 1991, Third Reference Catalogue of Bright Galaxies (New York, NY: Springer)
Dieleman, S., Willett, K. W., & Dambre, J. 2015, MNRAS, 450, 1441
Duda, R., Hart, P. E., & Stork, D. G. 2000, Pattern Classification (2nd ed.; New York: Wiley-Interscience)
Falcón-Barroso, J., Sánchez-Blázquez, P., Vazdekis, A., et al. 2011, A&A, 532, A95

Figure 12. Distribution of feature values among morphometric classes for the LEGACY complete sample.

Figure 13. Distribution of feature values among morphometric classes for the LEGACY–zr sample.
