Using Machine Learning to Predict the Correlation of Spectra Using SDSS Colour Magnitudes as an Improvement to the Locus Algorithm

Tom O'Flynn[a], Eugene Hickey[a,1], Kevin Nolan[a,2], Oisin Creaner[b,2]

[a] *Department of Applied Science Technological University Dublin D24FKT9 Ireland.*
[b] *Dublin Institute of Advanced Studies 10 Burlington Rd Dublin D04 C932 Ireland*

Abstract

The Locus Algorithm is a new technique to improve the quality of differential photometry by optimising the choices of reference stars. At the heart of this algorithm is a routine to assess how good each potential reference star is by comparing its sdss magnitude values to those of the target star. In this way, the difference in wavelength-dependent effects of the Earth's atmospheric scattering between target and reference can be minimised. This paper sets out a new way to calculate the quality of each reference star using machine learning. A random subset of stars from sdss with spectra was chosen. For each one, a suitable reference star, also with a spectrum, was chosen. The correlation between the two spectra was taken to be the gold-standard measure of how well they match up for differential photometry. The five sdss magnitude values for each of these stars were used as predictors. A number of supervised machine learning models were constructed on a training set of the stars and were each evaluated on a testing set. The model using Support Vector Regression had the best performance of these models. It was then tested of a final, hold-out, validation set of stars to get an unbiased measure of its performance. The dataset used, the model constructions, and performance evaluation are presented here.

1. Introduction

A wealth of astrophysics information is available through the study of the brightness of celestial objects as a function of time. For example, exoplanet detection by the transit method relies critically on measurements of intrinsic variability where such variability can be a small fraction of the total stellar brightness (Giltinan et al. (2011), Everett and Howell (2001)). Ground-based observations looking for such variability are complicated by the effects of the Earth's atmosphere which causes incoherent wavelength-dependent variations in the stellar flux detected. This can mask intrinsic variability and hamper the study of variable astrophysical phenomena (Smith et al. (2008)).

The technique of differential photometry has been developed in an attempt to mitigate the effects of the Earth's atmosphere on studies of stellar variability. Differential photometry uses references stars at small angular separations from the star of interest as comparators. Atmospheric effects should have similar effects on the measured flux from all of these stars causing them to vary in unison (Burdanov et al. (2014)). Because scattering in the Earth's atmosphere is wavelength dependent, the technique is especially successful if the target star and reference stars are spectrally similar (Milone and Pel (2011), Sterken et al. (2011)).

The Locus Algorithm (Creaner et al. (2022)) has been used to create catalogues of pointings suitable for differential photometry on astromonical targets based on a novel technique of choosing appropriate reference stars (Creaner et al. (2020) and Creaner EXO's). The algorithm no longer places the target in the centre of

[*] Corresponding author

Email addresses: tom.oflynn@tudublin.ie (Tom O'Flynn), eugene.hickey@tudublin.ie (Eugene Hickey), kevin.nolan@tudublin.ie (Kevin Nolan), creanero@cp.dias.ie (Oisin Creaner)

[1] Corresponding Author

[2] Equal contribution

the field of view but, in general, repositions it so as to include the best set of reference stars. Assessment of each reference star is performed by referring to the sdss catalogue and the colour band magnitudes therein. These magnitudes can be used to infer the overall shape of the star's spectrum. Stars that have similar spectra will be effected by scattering from the Earth's atmosphere to a more comparable degree that stars with dissimilar spectra. The original Locus Algorithm used a rational, but ad-hoc, method to estimate the correlation of stellar spectra based on differences between their g, r, and i sdss colour magnitudes (**?**). This was necessary for computational efficiency. The work presented here presents a more rigorous technique to estimate the correlation of stellar spectra based on machine learning. The subset of stars in sdss that have their spectra measured are used. These stars are paired off such that each pair has similar colour magnitude differences and are thus potentially a good match for differential photometry. This corresponds to the pre-filtering step employed by **?**. The correlation between each pair's spectra is calculated. This forms the basis of a goodness-of-fit between the two spectra. The sdss magnitudes (u, g, r, i, and z for both stars in the pair) are then used to train machine learning algorithms to predict this goodness-of-fit. The models produced are then applied to other pairs of stars, the test set, to evaluate their performance. The results show a significant improvement over the original ad-hoc Locus Algorithm routine, this model will be incorporated to future generations of the Locus Algorithm.

2. Data

This work uses 3556 stellar spectra from the SDSS SEGUE and BOSS observations and their physical parameters from the 13^{th} SDSS data release (Aguado et al. (2018)). The spectra are clipped to just the wavelengths contained in the sdss r band (between 550nm and 700nm). Stars are paired off based on their sdss colour magnitudes so that both stars in a pair are of similar colour. Specifically, both $(g_1 - r_1) - (g_2 - r_2)$ and $(r_1 - i_1) - (r_2 - i_2)$ will be between 0 and 0.3. 0.3 is thus called the *Colour Match Limit.* This ensures that these stars would be realistic matches for differential photometry. In addition, stars were chosen that had r colour magnitude values between 15 and 20. Care is taken to ensure that all 7112 stars in the 3556 pairs are unique. The SQL queries used to download physical parameters and the spectra are given in the supplementary materials for this paper. Correlations between spectra are calculated using the usual Pearson Correlation formula, equation (1).

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\ \sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}. \tag{1}$$

where x_i refer to the flux from the first star at a given wavelength, i, in units of $erg/cm^2/s$Å, y_i refer to the flux from the second star at the same wavelength. Figure 1 shows some pairs of stars along with their correlations. The first pair, A and B, are representative of the sample. The second pair, C and D, were chosen to have an unusually low correlation for this sample set.

Correlation is usually bounded by -1 and 1. And because these are spectra from stars and they have similar colour magnitudes, the correlations tend to be clustered near this higher end, see the histogram in figure 2A below. Machine learning algorithms work better with normally distributed values (need reference) and this is especially true when it comes to analysing model performance (another reference), so the correlation values were transformed. First of all by a logit transformation (2):

$$\operatorname{logit}(x) = \ln\left(\frac{1+x}{1-x}\right) \tag{2}$$

And then by scaling and normalising the values to have a mean of 0 and a standard deviation of 1. The resulting transformed values are shown in figure 2B.

The data is split into test and training sets, with 70% of the data (2668 samples) in the training set and the remainder (888 samples) in the test set. Each set has a representative sample of correlation values, to do this the original sample of 3556 pairs is split into five groups based on percentiles of the correlation and both testing and training sets get a commensurate proportion of each group. In additional, completely independent set of 526 pairs of stars are used for final validation of the chosen model.

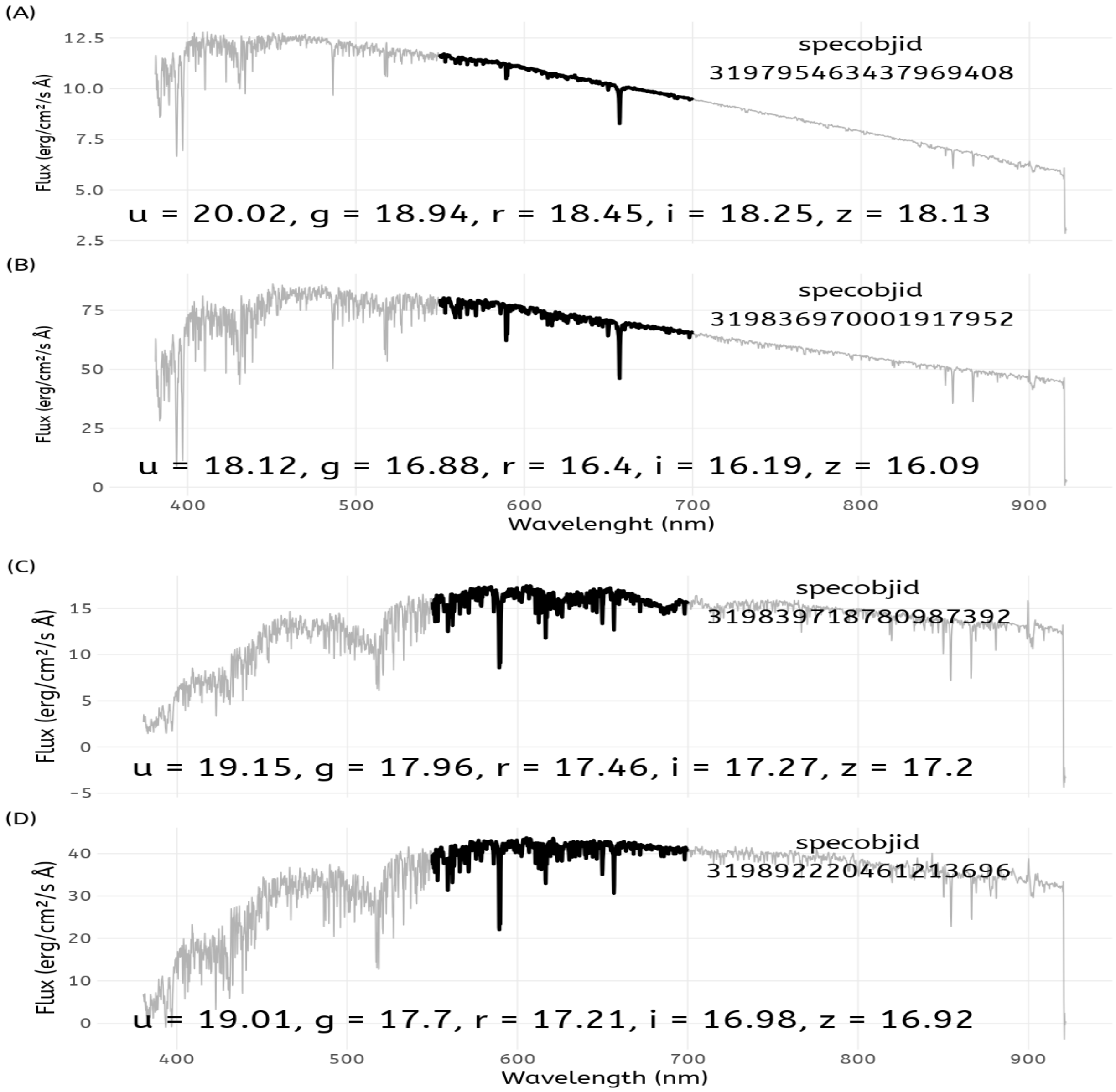


Figure 1: Two pairs of spectra downloaded from SDSS. The ugriz colour magnitudes for each star is given below its spectrum. The darkened area of the spectral line corresponds to the r-band wavelengths. The correlation between spectra A and C is 0.96. That between spectra B and D is 0.75.

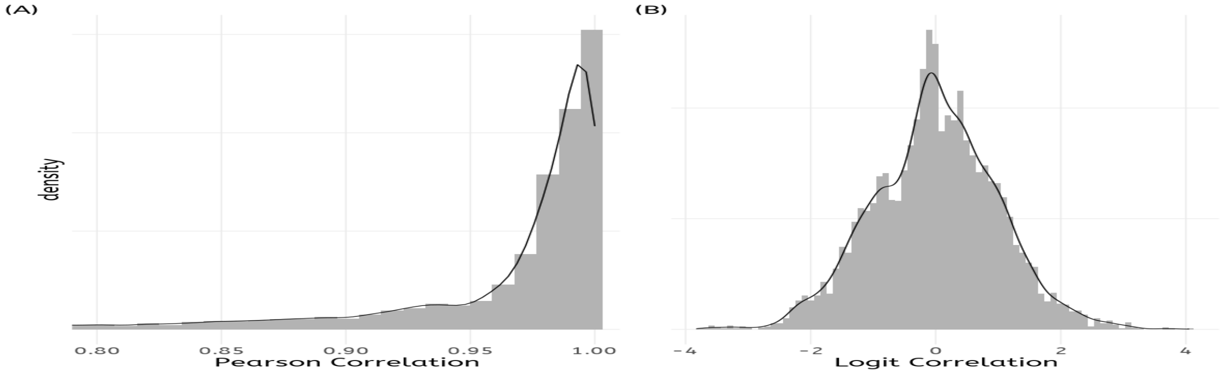


Figure 2: (A) Histogram of Pearson correlation values between r-band spectra between pairs of matched stars. (B) Values in (A) transformed by a logit function.

3. Models

Four types of machine learning algorithms were used: support vector regression (SVR), random forrest (RF), gradient boosting (GBM), and a linear model (LM). In each of the four cases, a model was built on the training set, using 10 predictors, namely the ugriz values for both stars in each pair. The target value was the logit(correlation) value. Cross validation, 20-fold repeated 10 times, was used to minimise model bias. Hyperparameters for each model were tuned. The results for RMSE, MAE, and R^2 for all models are given in table 1.The details are given below:

- *Support Vector Regression* This used a radial basis kernal function (Karatzoglou2004). A grid search on the hyperparameters Cost (C) and sigma (σ) was undertaken for values $8 < C < 25$ and $0.08 < \sigma < 0.20$. The model was optimised based on the RMSE of the training set. The best tune was obtained for $C = 20$ and $\sigma = 0.12$.
- *Stochastic Gradient Boosting* This used Friedman's gradient boosting algorithm (Boehmke and Greenwell (2019)). A grid search on the hyperparameters Number of Boosting Iterations (n.trees), Maximum Tree Depth (interaction.depth), Shrinkage (shrinkage), and Minimum Terminal Node Size (n.minobsinnode) was undertaken. The model was optimised based on the RMSE of the training set. The best tune was obtained for n.trees = 100, interaction.depth = 10, shrinkage = 0.1 and n.minobsinnode = 10.
- *Random Forrest* This used the *ranger* fast implementation of random forrests (Wright and Ziegler (2017)). The Minimum Node Size was set to be 5, the splitting rule was chosen to be *variance*. The number of variables to possibly split at each node (mtry) was tuned to be 6.
- *Linear Model* The final model was a linear model from the MASS R package (Venables and Ripley (2002)). There were no tunable parameters.

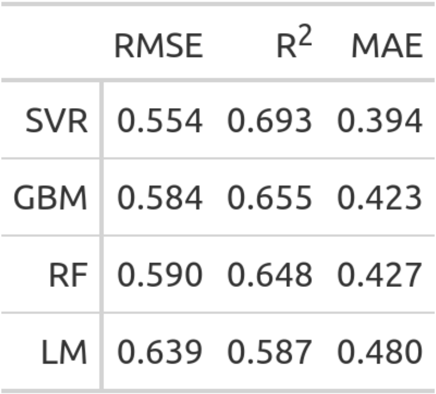
4. Model Evaluation

Table 1 gives the values for RMSE, MAE, and R^2 for each of the four values as they applied to the test set of 888 stellar pairs. To put these RMSE and MAE values in context, remember that the logit correlation values have been scaled to have a standard deviation of 1.

As can be seen from table 1, the best performing model was that produced by SVR. This is in line with expectations as our data set had few (10) predictors, they were all numeric as opposed to being a mix of numeric and categorical, and there were no missing values. These are all factors that favour SVR models.

Finally, the SVR model was applied to the validation set of 526 stellar pairs. The resulting predicted values has an RMSE of 0.564, an R^2 of 0.682, and an MAE of 0.399 when compared to the observed logit correlation values of the spectra.

Table 1: Results of applying the best tuned models on the test set. Root mean squared error (RMSE), R-squared value (R2), and mean average error (MAE) for each of the four model types. The models were built on the training set of 2668 stellar pairs. The values given below were obtained by applying each model to the test set of 888 stellar pairs



	RMSE	R^2	MAE
SVR	0.554	0.693	0.394
GBM	0.584	0.655	0.423
RF	0.590	0.648	0.427
LM	0.639	0.587	0.480

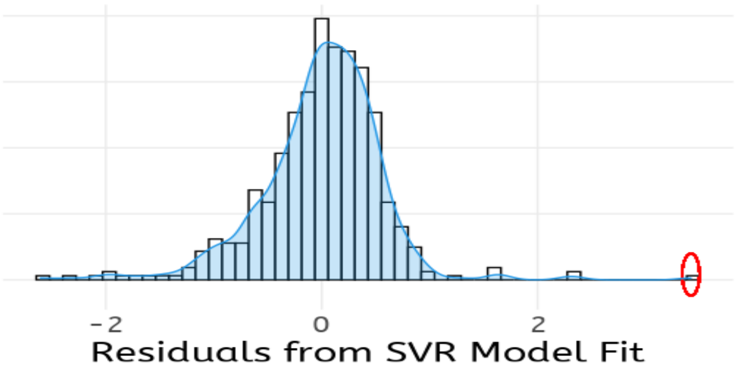


Figure 3: Residuals from the fit of the best performing SVR model on the validation dataset. Note one of the outlying values circled in red

The histogram of the residuals, the observations minus the predicted values, from this fit on the validation data is shown in figure 3. This histogram looks to be normally distributed with the addition of a significant number of outliers. Note the largest absolute residual with a value of 3.41 circled in red in figure 3. This particular value will be discusssed below. Figure **??** (A) shows the resulting values of observed logit correlation values against predicted logit correlation values. The points are coloured depending on the absolute value of the residual (how far they lie from the dashed line). The point at the bottom left corner is the one mentioned in the discussion of figure 3. Figure **??** (B) shows the resulting values of the residuals against predicted logit correlation values. Figure **??** (C) shows a quantile-quantile (QQ) plot of the residuals.

All of these plots point to the presence of outliers in the residuals. These were investigated. Looking at the histogram in figure 3. As stated earlier, there is a notable residual at 3.41, highlighted in red in the figure. The spectra of the pair of stars involved in this outlier were investigated and are shown in figure 5. Note that the top spectrum, from a star of spectral subclass CV located at $RA = 140.04°$ and $Dec = 0.7125°$, has emission lines rather than absorption lines more reminiscent of a galaxy rather than a star. This behaviour was noted for some, but not all, of the stellar pairs with high residuals. In all cases it involved a star with subclass CV.

Finally, the model presented here was compared with the model used in the original Locus Algorithm (**?**). That used a linear model of the differences in magnitudes in the g, r, and i SDSS bands between the two stars in each pair to calculate a rating for each pair. This is shown in equation 3.

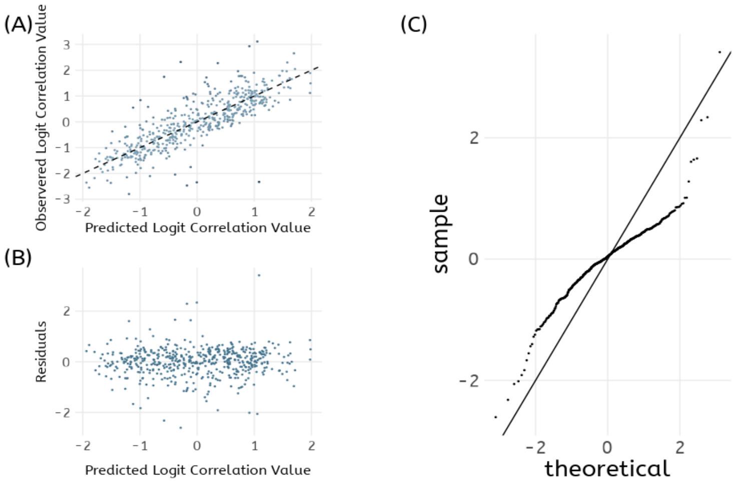


Figure 4: Observed versus predicted logit correlation values

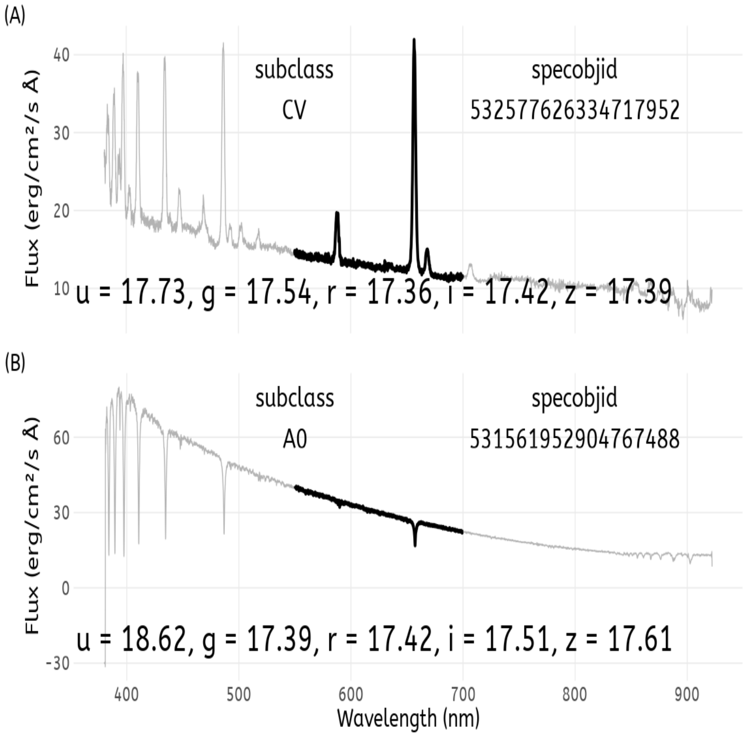


Figure 5: Spectra of the two stars involved in the outlier of the residuals histogram

$$rating = (1 - \left|\frac{\delta g - \delta r}{M}\right|) \times (1 - \left|\frac{\delta r - \delta i}{M}\right|) \tag{3}$$

Where M is the *Colour Match Limit* of 0.3 and δr stand for the difference between the r-band magnitudes for the two stars in each pair. Similarly for δg and δi.

When the ratings thus derived were compared with the logit correlation values derived from the spectra, an R^2 of 0.345 was obtained. This compares with the R^2 of 0.682 obtained by the SVR model here. There is a clear improvement in using the new machine learning approach to the Locus Algorithm.

5. Conclusions

The Locus Algorithm is a technique to refine differential photometry based on the wavelength-dependent nature of atmospheric effects. It relies on using reference stars with a good spectral match to the target star. The original Locus Algorithm used the linear difference between magnitudes from the SDSS astrophysics catalogue between pairs of stars to estimate their spectral match.

The work presented here examined machine learning techniques to estimate this spectral match. Stars were chosen from the SDSS catalogue where full spectra as well as photometric magnitudes were available. These stars were paired off with a colour-matched star from SDSS with also had a spectrum. The Pearson correlation between each pair of spectra was calculated. A dataset was produced which included this correlation along with the five SDSS magnitude values for each star. The correlations were then transformed by a logit transformation and scaled to have a standard mean and standard deviation. These pairs of stars were split into a training set, a testing set, and a validation set. The training set was used to produce a range of models using cross validation. Each model used ten predictors (the five SDSS magnitudes for both stars in the pair) to predict the spectral correlation values. The testing set was used to produce an unbiased estimate of model performance. And the validation set was used just once to get a final performance estimate for the best performing model.

The best performing machine learning model was based on Support Vector Regression using a radial kernel. The hyperparameters of this model were refined. The best performing configuration had an R^2 coefficient of 0.682 when applied to a validation set of 526 stars. This compares with the R^2 value of 0.345 from the original technique. This improvement is expected to lead to further improvements in the performance of the Locus Algorithm.

The machine learning model performed poorly on a subset of stellar pairs. In many cases, it was observed that one member of each pair had an unusual stellar spectrum with strong emission lines more typical of a galaxy rather than a star. These stars were invariably classified as CV type stars by SDSS.

References

D. S. Aguado, Romina Ahumada, Andres Almeida, Scott F. Anderson, Brett H. Andrews, Borja Anguiano, Erik Aquino Ortiz, Alfonso Aragon-Salamanca, Maria Argudo-Fernandez, Marie Aubert, Vladimir Avila-Reese, Carles Badenes, Sandro Barboza Rembold, Kat Barger, Jorge Barrera-Ballesteros, Dominic Bates, Julian Bautista, Rachael L. Beaton, Timothy C. Beers, Francesco Belfiore, Mariangela Bernardi, Matthew Bershady, Florian Beutler, Jonathan Bird, Dmitry Bizyaev, Guillermo A. Blanc, Michael R. Blanton, Michael Blomqvist, Adam S. Bolton, Mederic Boquien, Jura Borissova, Jo Bovy, William Nielsen Brandt, Jonathan Brinkmann, Joel R. Brownstein, Kevin Bundy, Adam Burgasser, Nell Byler, Mariana Cano Diaz, Michele Cappellari, Ricardo Carrera, Bernardo Cervantes Sodi, Yanping Chen, Brian Cherinka, Peter Doohyun Choi, Haeun Chung, Damien Coffey, Julia M. Comerford, Johan Comparat, Kevin Covey, Gabriele da Silva Ilha, Luiz da Costa, Yu Sophia Dai, Guillermo Damke, Jeremy Darling, Roger Davies, Kyle Dawson, Victoria de Sainte Agathe, Alice Deconto Machado, Agnese Del Moro, Nathan De Lee, Aleksandar M. Diamond-Stanic, Helena Dominguez Sanchez, John Donor, Niv Drory, Helion du Mas des Bourboux, Chris Duckworth, Tom Dwelly, Garrett Ebelke, Eric Emsellem, Stephanie Escoffier, Jose G. Fernandez-Trincado, Diane Feuillet, Johanna-Laina Fischer, Scott W. Fleming, Amelia Fraser-McKelvie, Gordon Freischlad, Peter M. Frinchaboy, Hai Fu, Lluis Galbany, Rafael Garcia-Dias, D. A. Garcia-Hernandez, Luis Alberto Garma Oehmichen, Marcio Antonio Geimba Maia, Hector Gil-Marin, Kathleen Grabowski, Meng Gu, Hong Guo, Jaewon Ha, Emily Harrington, Sten Hasselquist, Christian R. Hayes, Fred Hearty, Hector Hernandez Toledo, Harry Hicks, David W. Hogg, Kelly Holley-Bockelmann, Jon A. Holtzman, Bau-Ching Hsieh, Jason A. S. Hunt, Ho Seong Hwang, Hector J. Ibarra-Medel, Camilo Eduardo Jimenez Angel, Jennifer Johnson, Amy Jones, Henrik Jonsson, Karen Kinemuchi, Juna Kollmeier, Coleman Krawczyk, Kathryn Kreckel, Sandor Kruk, Ivan Lacerna, Ting-Wen Lan, Richard R. Lane, David R. Law, Young-Bae Lee, Cheng Li, Jianhui Lian, Lihwai Lin, Yen-Ting Lin, Chris Lintott, Dan Long, Penelope Longa-Pena, J. Ted Mackereth, Axel de la Macorra, Steven R. Majewski, Olena Malanushenko, Arturo Manchado, Claudia Maraston, Vivek Mariappan, Mariarosa Marinelli, Rui Marques-Chaves, Thomas Masseron, Karen L. Masters, Richard M. McDermid, Nicolas Medina Pena, Sofia Meneses-Goytia, Andrea Merloni, Michael Merrifield, Szabolcs Meszaros, Dante Minniti, Rebecca Minsley, Demitri Muna, Adam D. Myers, Preethi Nair, Janaina Correa do Nascimento, Jeffrey A. Newman, Christian Nitschelm, Matthew D Olmstead, Audrey Oravetz, Daniel Oravetz, Rene A. Ortega Minakata, Zach Pace, Nelson Padilla, Pedro A. Palicio, Kaike Pan, Hsi-An Pan, Taniya Parikh, James Parker, Sebastien Peirani, Samantha Penny, Will J. Percival, Ismael Perez-Fournon, Thomas Peterken, Marc Pinsonneault, Abhishek Prakash, Jordan Raddick, Anand Raichoor, Rogemar A. Riffel, Rogerio Riffel, Hans-Walter Rix, Annie C. Robin, Alexandre Roman-Lopes, Benjamin Rose, Ashley J. Ross, Graziano Rossi, Kate Rowlands, Kate H. R. Rubin, Sebastian F. Sanchez, Jose R. Sanchez-Gallego, Conor Sayres, Adam Schaefer, Ricardo P. Schiavon, Jaderson S. Schimoia, Edward Schlafly, David Schlegel, Donald Schneider, Mathias Schultheis, Hee-Jong Seo, Shoaib J. Shamsi, Zhengyi Shao, Shiyin Shen, Shravan Shetty, Gregory Simonian, Rebecca Smethurst, Jennifer Sobeck, Barbara J. Souter, Ashley Spindler, David V. Stark, Keivan G. Stassun, Matthias Steinmetz, Thaisa Storchi-Bergmann, Guy S. Stringfellow, Genaro Suarez, Jing Sun, Manuchehr Taghizadeh-Popp, Michael S. Talbot, Jamie Tayar, Aniruddha R. Thakar, Daniel Thomas, Patricia Tissera, Rita Tojeiro, Nicholas W. Troup, Eduardo Unda-Sanzana, Octavio Valenzuela, Mariana Vargas-Maga Na, Jose Antonio Vazquez Mata, David Wake, Benjamin Alan Weaver, Anne-Marie Weijmans, Kyle B. Westfall, Vivienne Wild, John Wilson, Emily Woods, Renbin Yan, Meng Yang, Olga Zamora, Gail Zasowski, Kai Zhang, Zheng Zheng, Zheng Zheng, Guangtun Zhu, Joel C. Zinn, and Hu Zou. The Fifteenth Data Release of the Sloan Digital Sky Surveys: First Release of MaNGA Derived Quantities, Data Visualization Tools and Stellar Library. dec 2018. doi: 10.3847/1538-4365/aaf651. URL http://arxiv.org/abs/1812.02759http://dx.doi.org/10.3847/1538-4365/aaf651.

Brad Boehmke and Brandon Greenwell. Hands-On Machine Learning with R. *Hands-On Machine Learning with R*, nov 2019. doi: 10.1201/9780367816377. URL https://www.taylorfrancis.com/books/mono/10.1201/9780367816377/hands-machine-learning-brad-boehmke-brandon-greenwell.

Artem Y Burdanov, Vadim V Krushinsky, and Alexander A Popov. Astrokit-an Efficient Program for High-Precision Differential CCD Photometry and Search for Variable Stars. *Translated from Astrofizicheskij Byulleten*, 69(3), 2014.

Oisín Creaner, Kevin Nolan, Niall Smith, David Grennan, and Eugene Hickey. A catalogue of Locus Algorithm pointings for optimal differential photometry for 23 779 quasars. *Monthly Notices of the Royal Astronomical Society*, 498(3):3720–3729, nov 2020. ISSN 13652966. doi: 10.1093/MNRAS/STAA2494.

Oisín Creaner, Kevin Nolan, Eugene Hickey, and Niall Smith. The Locus Algorithm: A novel technique for identifying optimised pointings for differential photometry. *Astronomy and Computing*, 38:100537, jan 2022. ISSN 2213-1337. doi: 10.1016/J.ASCOM.2021.100537. URL http://arxiv.org/abs/2003.04582.

Mark E. Everett and Steve B. Howell. A Technique for Ultrahigh-Precision CCD Photometry. *Publications of the Astronomical Society of the Pacific*, 113(789):1428–1435, nov 2001. ISSN 0004-6280. doi: 10.1086/323387/0.

Alan Giltinan, Dylan Loughnan, Adrian Collins, and Niall Smith. Using EMCCD's to improve the photometric precision of ground-based astronomical observations. *Journal of Physics: Conference Series*, 307(1), 2011. ISSN 17426596. doi: 10.1088/1742-6596/307/1/012010.

E. F. Milone and Jan Willem Pel. The High Road to Astronomical Photometric Precision: Differential Photometry. pages 33–68, 2011. doi: 10.1007/978-1-4419-8050-2_2. URL https://link.springer.com/chapter/10.1007/978-1-4419-8050-2_2.

Niall Smith, Alan Giltinan, Aidan O'Connor, Stephen O'Driscoll, Adrian Collins, Dylan Loughnan, Andreas Papageorgiou, Niall Smith, Alan Giltinan, Aidan O'Connor, Stephen O'Driscoll, Adrian Collins, Dylan Loughnan, and Andreas Papageorgiou. EMCCD Technology in High Precision Photometry on Short Timescales. *ASSL*, 351:257, oct 2008. doi: 10.1007/978-1-4020-6518-7_13. URL https://ui.adsabs.harvard.edu/abs/2008ASSL..351..257S/abstract.

Christiaan Sterken, E. F. Milone, and Andrew T. Young. Photometric Precision and Accuracy. pages 1–32, 2011. doi: 10.1007/978-1-4419-8050-2_1. URL https://link.springer.com/chapter/10.1007/978-1-4419-8050-2_1.

W. N. Venables and B. D. Ripley. Modern Applied Statistics with S. 2002. doi: 10.1007/978-0-387-21706-2. URL http://link.springer.com/10.1007/978-0-387-21706-2.

Marvin N. Wright and Andreas Ziegler. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77(1):1–17, mar 2017. ISSN 1548-7660. doi: 10.18637/JSS.V077.I01. URL https://www.jstatsoft.org/index.php/jss/article/view/v077i01.