ENHANCING DATA SCIENCE LEARNING THROUGH THE USE OF IMAGES

Karsten Lübke[1], Gero Szepannek[2] and Martin Vogt[3]
[1]FOM University of Applied Sciences, Germany
[2]Hochschule Stralsund – University of Applied Sciences, Germany
[3]Trier University of Applied Sciences, Germany
karsten.luebke@fom.de

Image analysis represents a crucial and dynamic field within the realm of data science and machine learning, enabling automated interpretation and analysis of images through statistical methods. Beyond its practical applications, such as image recognition, images offer a valuable pedagogical tool for teaching various multivariate analysis techniques, including cluster analysis, principal component analysis, and k nearest neighbors. By employing straightforward R code, images can be transformed into tidy data formats, primed for multivariate analysis. The resultant analysis outcomes can then be translated back into images, affording students the opportunity to visually comprehend the impact of techniques like cluster analysis, principal component analysis or k nearest neighbors. This approach effectively bridges the gap between abstract multivariate analysis concepts and concrete understanding, as students can visually perceive how cluster centroids or principal components reduce the complexity of the original (image) data.

INTRODUCTION

The current discourse surrounding artificial intelligence predominantly centers on generative models, such as ChatGPT and Midjourney. Unlike conventional statistical and data science tasks that focus on data modeling e.g. for classification purposes, generative models also exhibit the ability to generate novel text and images. While tabular data inherently embodies structured information, text and images, on the other hand, lack such inherent organization and are referred to as unstructured data. However, prior to statistical analysis, both text and images undergo a mathematical transformation that enables their conversion into a structured mathematical representation. Although instructional resources for textual data exist (e. g., Boehm and Hanlon, 2021), their counterparts for image analysis remain relatively scarce. As argued by Ridgway (2016) educators should harness the ongoing data revolution for pedagogical purposes. This paper aims to present a diverse range of practical applications and provides open educational resources for effectively incorporating images in the classroom. These materials are particularly suitable for introductory courses in multivariate analysis and machine learning.

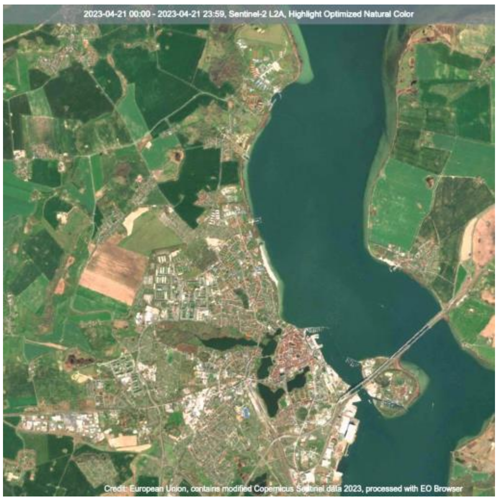
IMAGES AS DATA

There appears to be a consensus that data visualizations, i.e., images, are essential and beneficial (Schwab-McCoy et al., 2021 or Hsu et al., 2022) in statistical and data science education. However, it is noteworthy that images can also serve as valuable tools for elucidating concepts related to data structures and data preprocessing, which are integral topics in various data science curricula (Schwab-McCoy et al., 2021).

Undoubtedly, the realm of images encompasses a multitude of forms, ranging from photographs to paintings, with an extensive repository readily available on students' smartphones or the internet. Depending on the subject matter of the course, satellite images, exemplified in Figure 1, present an added opportunity to foster civic engagement. For instance, students can engage in comparative analyses of different cities, such as determining which city offers more green areas, or examine the temporal evolution of land use patterns, such as changes in farmland colors across seasons. Satellite images can be for example obtained from the Sentinel Hub (https://www.sentinel-hub.com/). The data (i.e. images) exported from Sentinel Hub EO Browser are licensed under the Attribution 4.0 International license (CC BY 4.0, https://creativecommons.org/licenses/by/4.0/). Figure 1 showcases the Stralsund area in Germany on April 21, 2023, employing the Sentinel-2 L2A data, with a highlighted optimized natural color. The downloaded image was subsequently optimized and cropped. The image offers distinct features, including the city of Stralsund located in the southwest corner, the island of Rügen to the east, the Strelasund between Stralsund and Rügen, the

In: EM Jones (Ed.), *Fostering Learning of Statistics and Data Science*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
July 2023, Toronto, Canada.

bridge connecting the two regions, and the presence of the small island of Dänholm in between both. The northwest and northeast sections exhibit diverse farmland landscapes.



2023-04-21 00:00 - 2023-04-21 23:59, Sentinel-2 L2A, Highlight Optimized Natural Color

Credit: European Union, contains modified Copernicus Sentinel data 2023, processed with EO Browser

Figure 1. Satellite (Sentinel-2 L2A) data of Stralsund (Germany)

The acquired data is in JPG format, which can be imported into the R programming language utilizing the "jpeg" package (Urbanek, 2022) and the "readJPEG()" function. The resulting R object corresponds to an array with dimensions defined by the pixel height, width, and the number of channels. In the case of RGB images, three channels are involved, representing different colors, each with a value ranging from 0 to 1. For Figure 1, we are dealing with a 1280x1280x3 array, ascertained by employing the "dim()" function in R. This presents an excellent opportunity to delve into the principles of tidy data, which advocates for tabular data structure with variables as columns and observations as rows (Wickham, 2014). It is essential to highlight that numerous statistical methods presuppose the utilization of structured data. In the context of images, each pixel can be regarded as an observation, while the red, green, and blue components serve as variables. Following the conversion of the array into a data frame, the image depicted in Figure 1 is encoded as 1280x1280=1,638,400 observations with five variables (three colors plus two (x, y) coordinates). Table 1 prints the first six observations.

Table 1. Pixel coordinates and colors

x-coordinate	y-coordinate	Red	Green	Blue
1	1280	0.5568627	0.5647059	0.5137255
1	1279	0.5568627	0.5647059	0.5137255
1	1278	0.5568627	0.5647059	0.5215686
1	1277	0.5490196	0.5529412	0.5215686
1	1276	0.5411765	0.5450980	0.5215686
1	1285	0.5333333	0.5372549	0.5137255

CLUSTER ANALYSIS

Identifying subgroups within a dataset is a prevalent task in unsupervised learning (e.g., Witten et al., 2021), often covered in multivariate statistics, data science, or machine learning courses. Numerous clustering algorithms exist, with k-means being one of the most commonly employed approaches. In essence, it serves as a means of simplifying complexity by focusing on k centroids rather than analyzing each individual observation. In the case of Figure 1, which comprises 1,638,400 pixels or observations, there are a total of 118,058 unique RGB color values. This complexity can be effectively reduced to a smaller number, such as k=8 clusters. To achieve this, a k-means clustering procedure is performed specifically on the three color variables. Figure 2 displays the colors of the resulting 8 centroids, along with their corresponding cluster sizes.

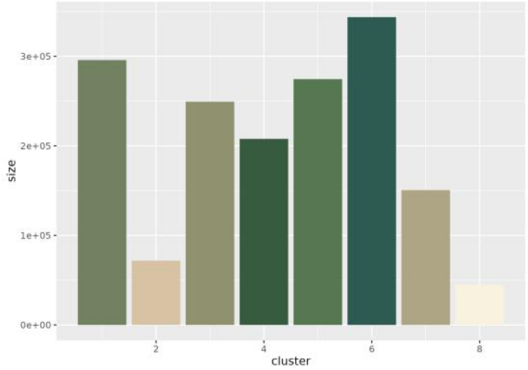


Figure 2. Cluster centroids and sizes

Instead of relying solely on the color of each cluster centroid, employing more contrasting colors to visualize the results can offer enhanced clarity. By utilizing a colorblind-friendly palette, this approach presents an opportunity to foster inclusivity within the classroom (Dogucu et al., 2023), as exemplified in Figure 3.

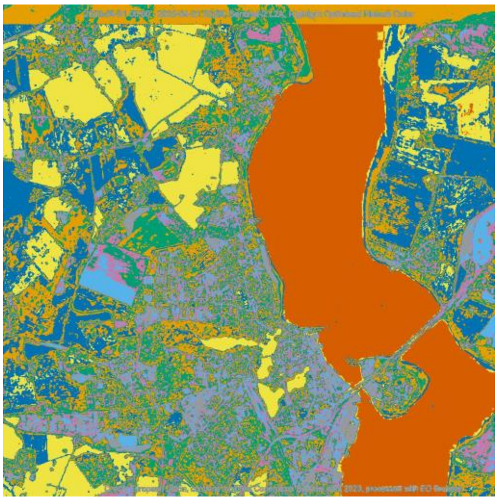


Figure 3. Satellite image based on 8 cluster in Okabe and Ito palette

Additionally, replacing the color of individual observations with the color of their corresponding cluster centroid, as depicted in Figure 4, allows students to grasp the overarching patterns exhibited in Figure 1. In an instructional setting, students can experiment with different values for k, catering to varying course levels. This exploration can encompass the introduction of selection criteria, as well as automatic methods, with subsequent comparisons of the obtained results.

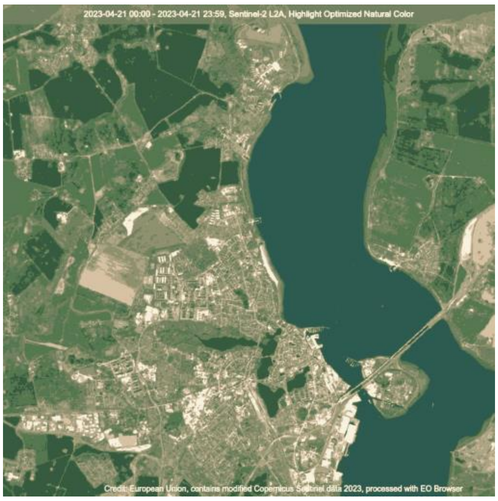

2023-04-21 00:00 - 2023-04-21 23:59, Sentinel-2 L2A, Highlight Optimized Natural Color
Credit: European Union, contains modified Copernicus Sentinel data 2023, processed with EO Browser

Figure 4. Satellite image based on 8 color cluster center in centroid colors

PRINCIPAL COMPONENT ANALYSIS

Principal component analysis (PCA) represents another widely used unsupervised learning technique (e.g., Witten et al., 2021). While cluster analysis is primarily employed to reduce dimensionality at the observational level, it is important to note that clustering of variables is also a possibility. In contrast, principal component analysis operates by reducing dimensionality within the variable space itself. In our specific application, this entails transforming the original three RGB colors to a single grayscale value. In the case of Figure 1, the first principal component captures nearly 95% of the total variation. In order to visualize the outcome, it is necessary to scale the scores to values between 0 and 1. Therefore, in addition to the utilization of multivariate analysis methods, this particular application presents an opportunity to discuss various scaling techniques for variables. The resulting image is displayed in Figure 5.

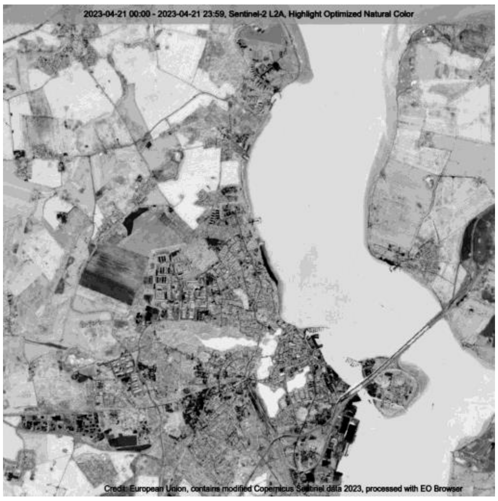

2023-04-21 00:00 - 2023-04-21 23:59, Sentinel-2 L2A, Highlight Optimized Natural Color
Credit: European Union, contains modified Copernicus Sentinel data 2023, processed with EO Browser

Figure 5. Satellite image based on first PCA score

Naturally, the second principal component can also be visualized accordingly, representing an additional 4% of the total variation.

SUPERVISED LEARNING

Cluster analysis and principal component analysis are both unsupervised learning methods that can be directly applied to a wide range of image data. However, when additional variables, such as land-cover type (e.g., water, forest), are available, supervised learning methods like classification trees or multinomial logistic regression can also be employed. Similarly, linear regression can be utilized, for instance, when modeling population size based on satellite imagery.

The metaphor of the "undiscovered island" (Martin, 2003) in linear regression can be effectively supported by visual representations using images. We are presented with a fragmented and noisy image, as depicted in Figure 6, and our objective is to reconstruct the original Figure 1.

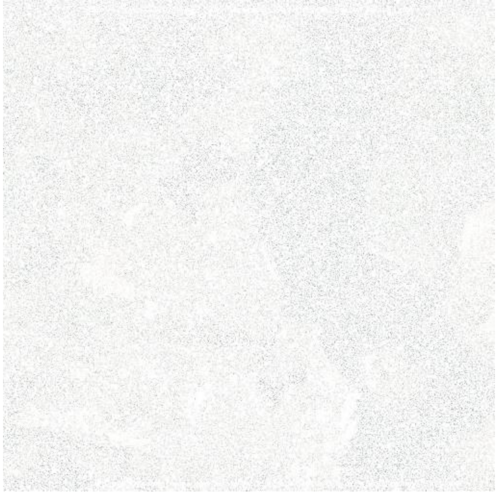


Figure 6. Noisy and fragmented satellite image

K-nearest neighbors, a classical supervised learning algorithm, offers a comprehensive platform to discuss various aspects of machine learning (Mike & Hazzan, 2022). A common task in data science, covered extensively in numerous courses, involves predicting outcomes for unseen data. By applying this algorithm to the pixel coordinates and with the three color values as target variables based on the data of Figure 6, the predicted full image data is illustrated in Figure 7.

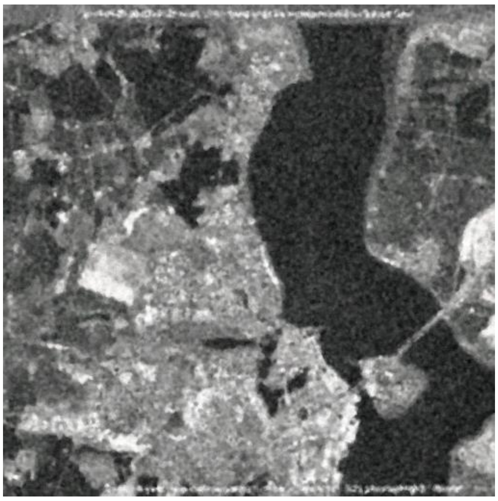


Figure 7. Predicted full image by k-nearest neighbors

EVALUATION

Based on our experience, we have observed that the inclusion of images tends to enhance the motivation of students with a specific interest in practical applications. We provide illustrative examples where images, particularly satellite data, can be leveraged for marketing purposes, such as identifying areas with a high concentration of private swimming pools. Moreover, the utilization of images facilitates the monitoring of economic and ecological activities.

As of now, our findings are primarily based on anecdotal evidence. However, it is worth noting that our observations suggest a high level of student engagement when exposed to such analyses. Students are able to witness firsthand how statistical methods aid in reducing the complexity of real-world multivariate data, enabling the capture of crucial aspects within the dataset.

CONCLUSION

In the rapidly evolving landscape of data, it is crucial to cultivate student learning in the fields of statistics and data science. It is equally important for students to recognize that data can manifest in diverse modalities. One effective approach that educators can adopt is the integration of image analysis in their teaching practices. By carefully selecting visually appealing images and facilitating discussions around the underlying conceptual ideas and algorithms, students can become actively engaged with the material. Moreover, the inclusion of tangible visual outputs, can further enhance the learning experience. It is important to acknowledge that, at present, our evidence is primarily based on anecdotal observations. Therefore, there is a pressing need for formal evaluation and further research to validate these outcomes.

It is worth noting that our study exclusively focused on the application of classical multivariate analysis methods to image data. We did not delve into the exploration of specialized techniques designed explicitly for image analysis.

SUPPLEMENTAL MATERIAL

The R Code for the analysis is available from https://github.com/luebby/IASE_2023-Imageanalysis.

ACKNOWLEDGMENT

This article contains modified Copernicus Sentinel data 2023, processed by Sentinel Hub. Lastly, it is important to disclose that the authors utilized ChatGPT and DeepL for the sole purpose of editing/lectorate this manuscript.

REFERENCES

Boehm, F. J., & Bret M. Hanlon, B. M. (2021).What Is Happening on Twitter? A Framework for Student Research Projects With Tweets, *Journal of Statistics and Data Science Education, 29*, S95-S102.

Dogucu, M., Johnson, A. A., & Ott, M. (2023). Framework for Accessible and Inclusive Teaching Materials for Statistics and Data Science Courses, *Journal of Statistics and Data Science Education*.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning* (2ed). New York: Springer.

Hsu, J. L., Jones, A., Lin, J.H., & Chen, Y.R. (2022). Data visualization in introductory business statistics to strengthen students' practical skills, *Teaching Statistics, 44*, 21-28.

Martin, M. A. (2003) “It's Like… You Know”: The Use of Analogies and Heuristics in Teaching Introductory Statistical Methods, *Journal of Statistics Education, 11*.

Mike, K., & Hazzan, O. (2022). Machine learning for non-majors: a white box approach. *Statistics Education Research Journal, 21*.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ridgway, J. (2016). Implications of the Data Revolution for Statistics Education. I*nternational Statistical Review, 84*, 528-549.

Schwab-McCoy, A., Baker, C. M., & Gasper, R. E. (2021). Data science in 2020: Computing, curricula, and challenges for the next 10 years. *Journal of Statistics and Data Science Education, 29,* S40-S50.

Urbanek S (2022). *jpeg: Read and write JPEG images*. R package version 0.1-10.

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software, 59,* 1-23.