Journal of Education and Social Development, 2020, 4–1
July. 2020, pages 18-24
doi: 10.5281/zenodo.3961030
http://www.ibii-us.org/Journals/JESD/
ISBN 2572-9829 (Online), 2572-9810 (Print)

Analyze and visualize cathodoluminescence data obtained from images of a photovoltaic cell using the R language

Mienie Roberts[1, *], Taylor Harvey[1], James Sullivan[1], Aida Torabi[1]

[1]Department of Science and Mathematics, Texas A&M University-Central Texas, 1001 Leadership Pl., Killeen, TX 76549

*Correspondence: Mienie Roberts, Department of Science and Mathematics, Texas A&M University-Central Texas, 1001 Leadership Pl., Killeen, TX 76549, Email: dekock@tamuct.edu,

Received on 06/05/2020; revised on 07/25/2020; published on 07/26/2020

Abstract
We use the R-software to analyze cathodoluminescence data obtained from images of a photovoltaic cell. We build an application to visualize the characterization of the luminescence of solar cells by using the Shiny package. The Graphical User Interface can import data, create a heatmap to indicate high quality crystal structure, and output several variables of interest.

Keywords: Data analysis, Data visualization, R, RStudio, Shiny, Data Science, Statistics education

1 Introduction

"During the last several years, we have observed an exponential increase in the demand for data scientists in the job market" (Belloum et al., 2019). We will adopt Wing's definition for Data Science, since there are many different definitions in the literature to describe this term. "Data science is the study of extracting value from data" (Wing, 2019). An undergraduate data science education curriculum typically consists of a combination of computer science and statistics courses. "With increasing demand for training in data science, extracurricular or 'ad hoc' education efforts have emerged to help individuals acquire relevant skills and expertise" (Demasi et al., 2020). In this paper we introduce a student project for designing and developing an application to analyze and visualize a large data set. Hicks and Irizarry in *A Guide to Teaching Data Science* advocate for an update to the statistics curriculum that would bring applications to the forefront rather than a theoretical focus (Hicks & Irizarry, 2018). We will discuss an "ad hoc" approach of "High investment, long-term cohesion" (HILT) education as an effort to better prepare students for a career in data science. "High investment, long-term cohesion (HILT) efforts require multiple investments (e.g., time, resources, cost) to persist over months or years. To do so, some efforts required hierarchies of training for researcher or software development mentors (e.g., "train the trainer" models)" (Demasi et al, 2020). In our case the mentors were RStudio certified. Prototypical HILT efforts include a focus on hands-on research projects or software development through close mentoring relationships for an extended period of time (in our case one academic year). "Data science, like computer science, requires a mix of theory and practice. Similar to how we now run software projects as part of most computer science curriculum, we should add practical projects to data science curricula" (Berthold, 2019). Through our "ad hoc" project, we address the need for teaching statistical *thinking* rather than statistical *recipes.* We promote the use of case-based, inquiry-led learning and teaching activities (Fawcett, 2018) by serving as mentors on a project where the student is required to build an application to visualize real-life data. The project requires the student to translate data collected from a Scanning Electron Microscope (SEM) obtained from images of a photovoltaic cell into an interactive online tool to visualize the characterization of the luminescence of solar cells. We collaborated with the Engineering Technology department to create an interactive application to upload a dataset and facilitate the data visualization process. The application is interactive, visual, dynamic, and assists research related to the development of efficient photovoltaic cells.

2 Methods

2.1 User Interface and distribution

"R language is a powerful tool used in a wide array of research disciplines and owes a large amount of its success to its open source and adaptable nature" (Carson and Basiliko, 2016). R relies heavily on packages and is widely used in the STEM (Science, Technology, Engineering, and Mathematics) disciplines. "It is also an 'all-in-one' environment that streamlines the data analysis workflow from data management and analysis to graphical data presentation" (Carson and Basiliko, 2016). For our purposes the R language encapsulates the cleaning, transfer, analysis and visualization of the data using several packages including Shiny, an R package that facilitates the creation of java-based web applications (hereafter "Shiny"; (Chang et al., 2015)). "A well-designed Shiny app can be extremely immersive. With its easy-to-use graphical user interface it can effectively 'bring to life' existing R code, allowing users to interact with functions and control statements using sliders, radio buttons, and text entry boxes" (Fawcett, 2018). "The development of graphical user interface (GUI) applications hosted on web-based servers such as Shiny can make complex workflows accessible across operating systems and internet browsers to those without programming knowledge" (Reyes et al., 2019). The cathodoluminescence application in Shiny combines the analytical capacities of the R environment with the user-friendliness of the Shiny user interface (Wsola et al., 2017). The application serves as a tool to improve the speed of the analysis and visualization of a large and messy set of data. The application eliminates routine tasks, allowing more time for higher order thinking (Rowell, 2014). As a result, the researchers are able to contextualize the data and draw conclusions from the heat maps and outputs. The strength of the Shiny applications' framework is its reactive programming, which links input and output data such that changes to the input results in updates to the output area without having to refresh the program, allowing users to seamlessly explore data. The Overall Workflow Diagram (Fig. 1) illustrates how the analytical procedure runs in the background to render graphical outputs without reloading the page, creating a seamless data exploration experience (Wszola et al., 2017).

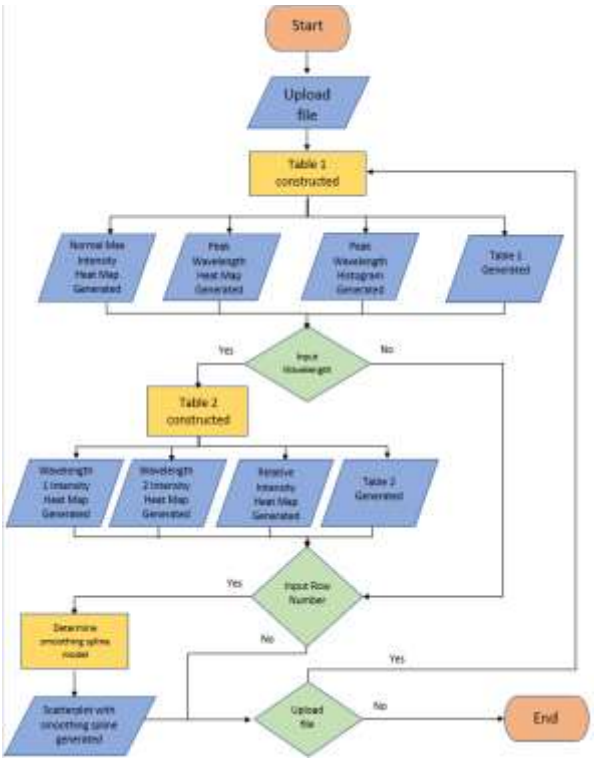


Fig. 1. Overall workflow diagram of the application.

First, the user uploads a Comma Separated Values (CSV) file containing raw data from the Scanning Electron Microscope. The algorithm cleans and organizes the columns from the input file and creates a data frame (Table) containing only four variables of interest: The "Coordinates" (first two columns), the "Wavelengths," and the "Luminescence values." Then the algorithm uses the values in this table to generate and display the following: The "Normal Maximum Intensity Heat map", the "Peak Wavelength Heat map", and the "Peak Wavelength Histogram." The table is also displayed. The heat maps are interactive in the sense that when the user clicks on a tile of interest, the reactive heat map displays the corresponding row number, coordinates, and peak wavelength. Next, the user has the option to input the value of the wavelength or row of interest. If the user selects the wavelength option, a second table/data frame is generated and a "Relative Intensity Heat Map" is displayed with the table. The "Row number" option generates a "Smoothing spline model" with a corresponding "Scatterplot." Finally, the "Download" button which accompanies every plot or table enables the user to download any of the outputs. The plots are downloaded as Scalable Vector Graphics (SVG) files and the tables can be downloaded as Comma Separated Values (CSV) files.

The application and code presented here were developed based on open-source tools (packages: data.table (Wickham, 2019), shiny (Wickham, 2020), dplyr (Wickham, 2016), ggplot2 (Wickham, 2016), plotly (Wickham, 2020), readr (Wickham, 2018), and hrbrthemes (Wickham, 2020)) and can thus run on most systems with sufficient memory to run basic java-based applications (Wszola et al, 2017).

2.2 Link to application

Below is a link the Shiny application:

https://sulley.shinyapps.io/HeatMapApp/?_ga=2.148470772.27375924.1591131362-1912021327.1580833746

Data files and code can be found at:
https://github.com/JSulley/Solar-Cell-Research

3. Results

The Cathodoluminescence Shiny application display is a web page organized by a navigation bar with the following options: "Welcome," "Upload File," "Wavelength Input," "Smooth spline Interpolation Display Input," "Dataset Plots," "Relative/Wavelength Intensity Plots," "Tables," "Peak Wavelength Histogram," "Export all Spectra," and "About." When a user opens the application, they are directed to the "Welcome" tab, which contains a static explanation of how to navigate the application (Fig. 2) with the following tabs: "File format for upload," "Downloading Plots/Tables," and "New Features."

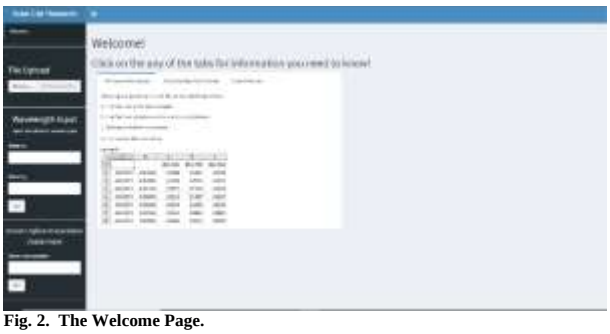


Fig. 2. The Welcome Page.

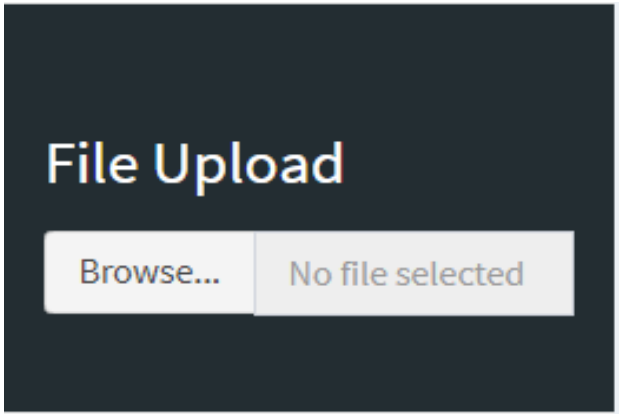


Fig. 3. CSV File Upload Button.

The "Upload File" button prompts the user to upload a CSV file (Fig. 3). This file should contain the cathodoluminescence data obtained from the scanning electron microscope. The dataset includes the following variables: "Wavelengths," "Coordinates," and "Luminescence" (or intensity values).

The "Wavelength Input," button navigates to a page which prompts the user to enter the values of two wavelengths. This is of interest to a user who wants to investigate the behavior of the intensity values between two different wavelengths (individually or combined). The application will then create corresponding heat maps (Fig. 4).

Fig. 4. Peak wavelength heat map flow chart.

The "Smooth Spline Interpolation Display Input" button navigates to a page that displays the "Smooth Spline Interpolation Plot." Since the "Normal Max Intensity Heat Map" does not allow a user to examine the relationship between the normalized intensities and the wavelengths for a specific point directly.

The "Dataset Plots" button navigates to a page which displays the corresponding plots.

"Relative/Wavelength Intensity Plots" button navigates to a page containing the resulting plots which can be downloaded.

The "Tables" button navigates to a page displaying the overall data and comparing intensities between the corresponding wavelengths displayed as a table.

"Peak Wavelength Histogram" button navigates to an interactive histogram to visualize the distribution of wavelengths and a scatterplot with the smoothing spline curve for a certain coordinate to visualize the relationship between the luminescence and wavelengths (Fig. 5).

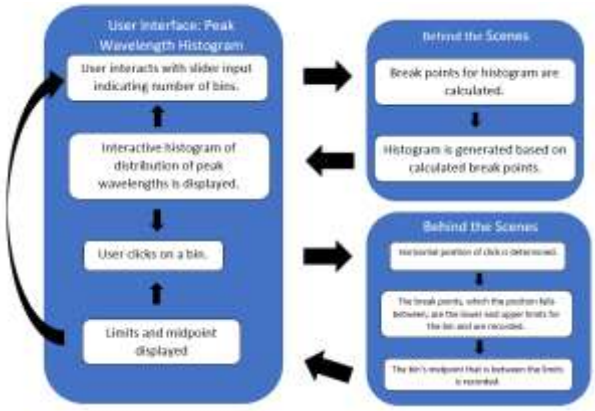


Fig. 5. Peak Wavelength Histogram flow chart.

The "Export All Spectra" button navigates to a page which provides the user with the option to export the plots as Scalable Vector Graphic (SVG) files.

The "About" button navigates to a page which includes information about the application, developers and cooperating agencies.

In summary, the combination of these features enables the user to examine the data as a whole or view specific areas.

3.1 Normal Maximum Intensity Heat Map

The "Normal Maximum Intensity Heat Map" represents the normalized distribution of the luminescence for the cathodoluminescence data. We explain this with an example dataset. The dataset is obtained from a CdMgTeSe thin film used in a solar cell. The dataset contains the variables (columns): "Coordinates" and the "Luminescence values." The rows represent the "Wavelengths."

Every coordinate has a series of luminescence values that correspond to a particular wavelength. Note that we extract the absolute maximum value for every coordinate. However, a coordinate can have two distinct wavelengths that correspond to the same absolute maximum value. To account for this, we perform a smoothing spline interpolation to ensure a unique maximum value for every coordinate.

As a result, the following process is executed to form the heat map for the dataset:

(1) For a coordinate, the program determines the smoothing spline model that best fits the data. "Wavelength" is treated as the independent variable; "Luminescence" is treated as the dependent variable.
(2) The algorithm then uses the model to predict the luminescence values of each wavelength.
(3) From the calculated luminescence values, the program determines the highest value and records it along with the coordinates and the wavelength. This information is stored in a row of a data frame.
(4) The data frame consists of four columns: the first two columns contain the coordinate values, the third column contains the wavelength value corresponding to the absolute maximum, and the last column contains the absolute maximum value.
(5) The algorithm continues to the next coordinates and repeats steps 1-3 until the last absolute maximum value is recorded.
(6) After obtaining all the absolute maximum values, the program determines which absolute maximum is the greatest and divides every absolute maximum by that value as a normalizing procedure.
(7) We use the newly constructed data frame to record the values by coordinate and we colorize based on a scale where red indicates higher values and black indicated lower values (Fig. 6).

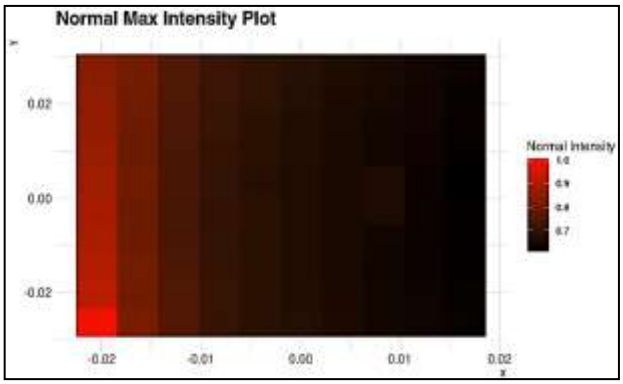
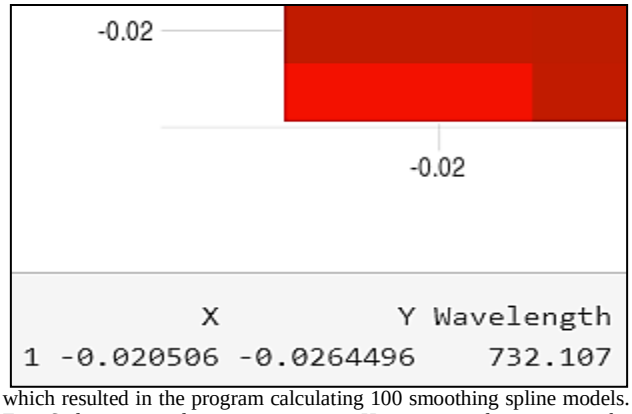


Fig. 6. Normal Maximum Intensity Heat Map. The tiles are shaded based on the value of the normal intensity.

Fig. 7. Output of the Normal Maximum Intensity Heat map.

Note: The relationship between the luminescence and the wavelengths is distinct for every coordinate. The dataset consists of 100 coordinates,



which resulted in the program calculating 100 smoothing spline models. Fig. 6 does not indicate composition. However, it does portray the electronic properties of the thin film through the range of normal intensity values. The colors are used to signal homogeneity.

A key feature of this plot is the interactivity. With a click of a mouse on any given tile, the user is presented with coordinates along with the normal max intensity. The value to the left represents the row number corresponding to the tile (Fig. 7).

3.2 Peak Wavelength Heat Map

Another value of interest is the wavelength corresponding to the absolute maximums. This value is obtained from the "Peak Wavelength Heat Map." The heat map is generated from the wavelength column in the existing data frame. Since the wavelengths correspond to the absolute maximum values, these result in the peak wavelengths. The peak wavelengths are represented as tiles and are color coded. Once again, red indicates higher values and black indicates lower values. Again, the heat map is interactive, therefore the user is presented with the row number corresponding to the selected tile, the coordinates, and peak wavelength (Fig. 8).

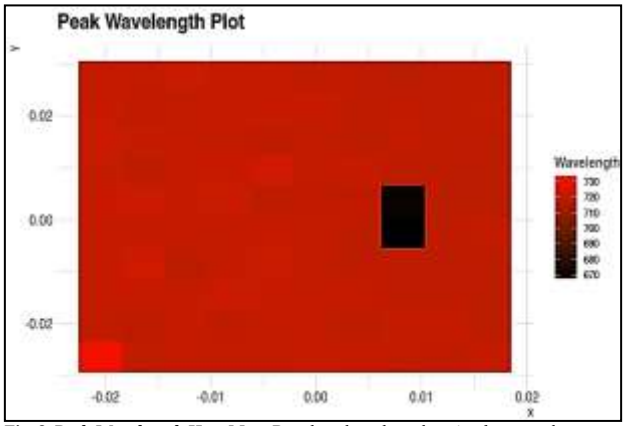


Fig. 8. Peak Wavelength Heat Map. Based on the colors, there is a large gap between the peak wavelength values.

3.3 Intensity Behavior Between Two Wavelengths

Since there is a wide range of values on the "Peak Wavelength Heat Map," we compare the intensities between the two wavelengths. The next three plots are major components for observing any trends that may occur either individually or combined. Note that the relative intensity refers to the ratio of Wavelength 1 and Wavelength 2.

3.3.1 Wavelength 1 and Wavelength 2 Heat Maps

Starting individually, the user is required to enter two distinct wavelength values and press the "Go" button when prompted (Fig. 9).

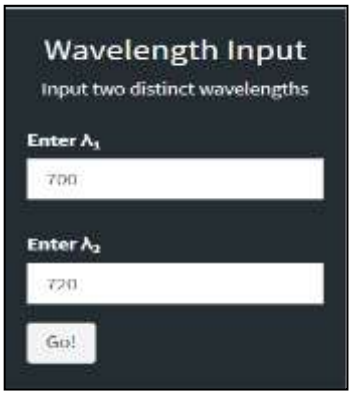


Fig. 9. Wavelength Input. The user will use this to generate all three heat maps. The entry can also take decimal values.

The program then initiates the same process as before: forming a smoothing spline model on the coordinate, using the given wavelengths and corresponding luminescence values. This time, however, instead of predicting the intensity values for all the wavelengths from the data set, the program only predicts the intensity values for the two given wavelengths. The algorithm records this value and continues to the next coordinate. The data recorded is not modified. After the program iterated through every coordinate, two heat maps for the respective wavelengths are displayed representing the individual intensities (Fig. 10 & 11). When the user clicks on a tile for one of the heat maps, the row number, coordinates, predicted intensity (not normalized), and relative intensity will be displayed (Fig. 12).

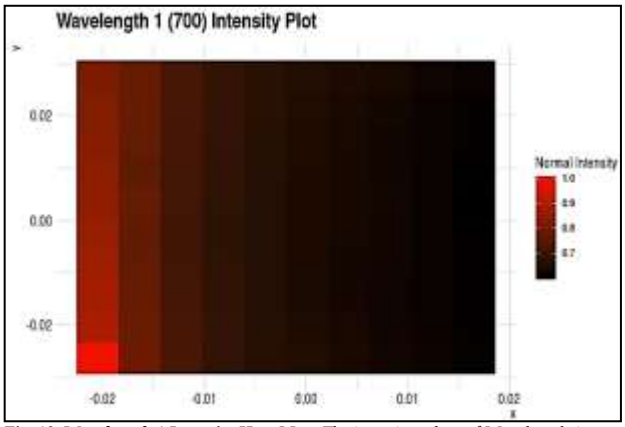


Fig. 10. Wavelength 1 Intensity Heat Map. The intensity values of Wavelength 1

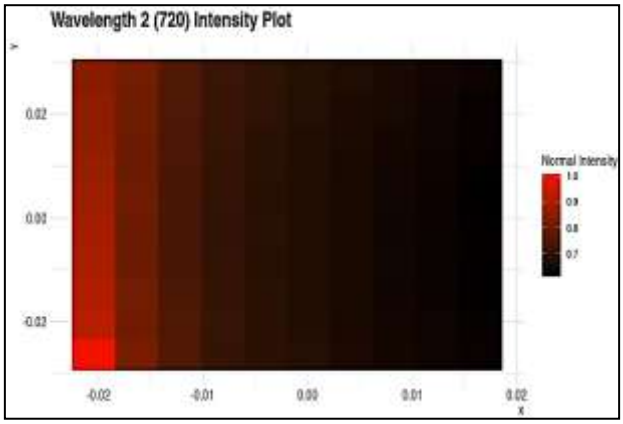


Fig. 11. Wavelength 2 Intensity Heat Map. The intensity values of Wavelength 2.

3.3.2 Relative Intensity Heat Map

The Relative Intensity heat map (Fig. 12) is generated by taking the ratio of the intensity values based on coordinate. Again, red indicates higher values and black indicates lower values.

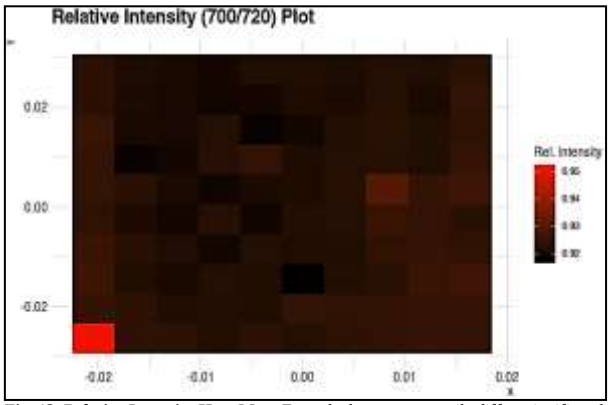


Fig. 12. Relative Intensity Heat Map. From the heat map, one tile differs significantly from all the other tiles.

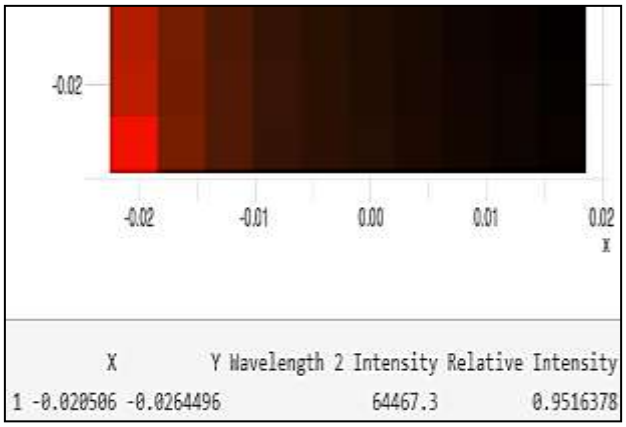


Fig. 13. Printed Information. Output for Wavelength 2.

After clicking on a tile, the user is presented with the row number, coordinates, predicted intensities for both wavelengths, and the relative intensity at the coordinate (Fig. 13). We obtain a similar image corresponding to wavelength 1 (Fig 14).

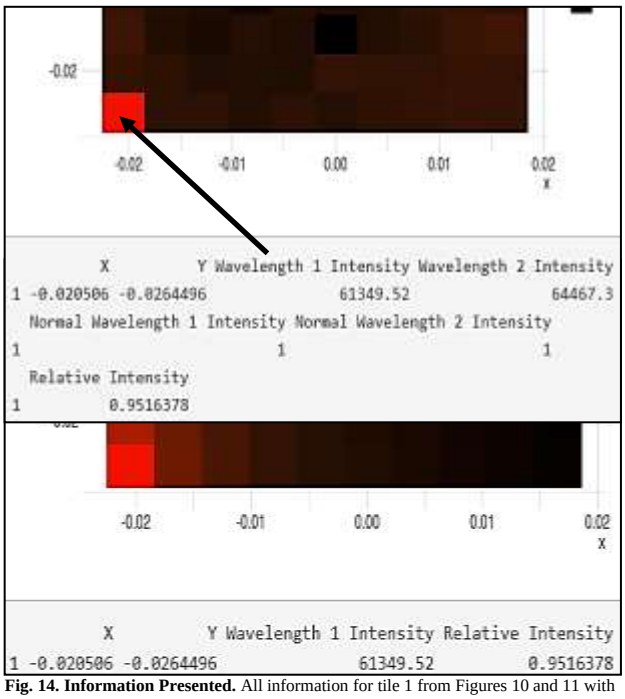


Fig. 14. Information Presented. All information for tile 1 from Figures 10 and 11 with the normalized numbers.

3.4 Tables

In addition to heat maps the application can also display the results in table format.

3.4.1 Table 1: Cathodoluminescence Data

The table describing the overall dataset displays the coordinates, peak wavelengths, absolute maximum intensity (not normalized) and the normalized intensity. This information stems from both the "Normal Maximum Intensity Heat Map" and the "Peak Wavelength Heat Map" making it easily available for the user (Fig. 16).

Table 1

Download Table 1

X	Y	λ_{peak}	Intensity	Normal Intensity
-0.02	-0.03	732.11	64486.35	1.00
-0.02	-0.02	720.39	59166.15	0.92
-0.02	-0.01	721.44	58367.96	0.91
-0.02	-0.01	721.18	57650.03	0.89
-0.02	-0.00	721.18	57103.08	0.89
-0.02	0.00	722.22	56398.65	0.87

Fig. 16. Table 1. Information shown for the first 6 coordinates.

3.4.1 Table 2: Wavelengths

Table 2 displays the results for the wavelengths. The entries correspond to the predicted luminescence value for the entered wavelengths including the normal intensities for both wavelengths and the relative intensity (Fig. 17).

Table 2

Download Table 2

λ_1 Intensity	λ_2 Intensity	Normal λ_1 Intensity	Normal λ_2 Intensity	Relative Intensity
61349.52	64467.30	1.00	1.00	0.95
54706.18	59164.51	0.89	0.92	0.92
54022.93	58344.29	0.88	0.91	0.93
53348.20	57644.32	0.87	0.89	0.93
52702.65	57067.84	0.86	0.89	0.92
52084.88	56335.72	0.85	0.87	0.92

Fig. 17. Table 2. Information regarding wavelengths individually and combined.

3.5 Histograms

In order to investigate the distribution of the peak wavelengths, the application provides the user with an option to generate an interactive histogram representing the distribution of the peak wavelengths (Fig 18). A slider input provides the user with options on the number of bins. The "Peak Wavelengths Histogram" has an additional interactive feature that allows a user to click on a bin causing the midpoint, lower and higher limits to appear (Fig. 19).

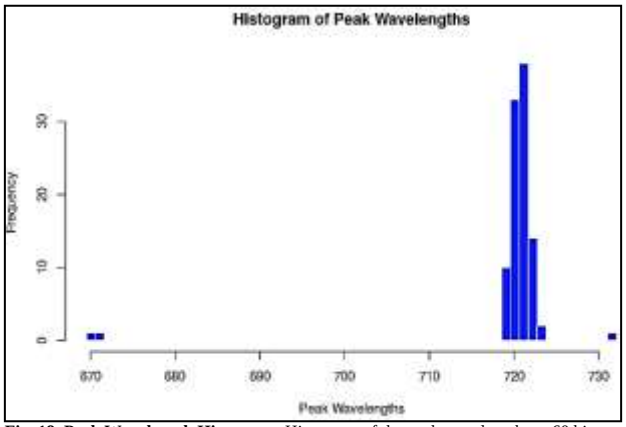


Fig. 18. Peak Wavelength Histogram. Histogram of the peak wavelengths at 60 bins. The slider input is above the plot.

3.6 Smoothing Spline Interpolation Plot

The smoothing spline interpolation provides additional insights into the data. The above-mentioned results represent general aspects of the cathodoluminescence data, such as the wavelengths and luminescence values. The algorithm involved in generating the "Normal Max Intensity Heat Map" does not account for an important point of interest: What is the relationship between the normalized intensities and the wavelengths for a specific point? In the previous example, the algorithm created 100 smoothing spline models since the relationship between the value of the luminescence and the value of the wavelength differs at every coordinate. The application attends to this question with an additional feature: the scatterplot of the normalized intensities and wavelengths with the smoothing spline curve for a specific coordinate, by normalizing original luminescence values before determining the smoothing spline model for a chosen coordinate. The reason is because the original intensity values are not scaled, (normalized in this case), whereas the predicted values by the smoothing spline model are. As a consequence, the transformation of the dependent variable explains the relationship between the two variables and show the smoothing spline curve through the coordinates.

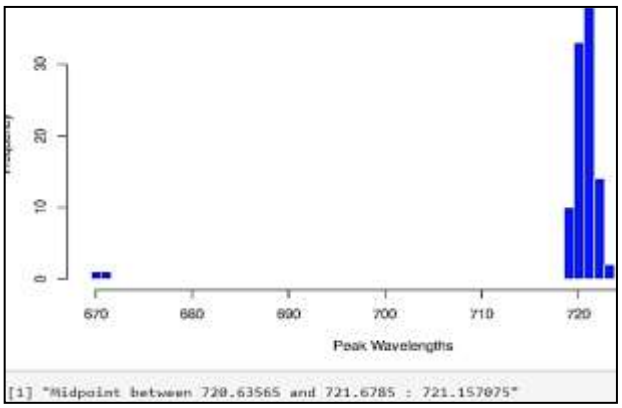


Fig. 19. Output midpoint and limits. Information for the class with the highest frequency.

By clicking on the tile of interest the algorithm can determine the row number. This value can also be entered manually into the "Smooth Spline Interpolation Display Input" (Fig. 20)to generate a "Smooth Spline Interpolation plot" (Fig. 21).

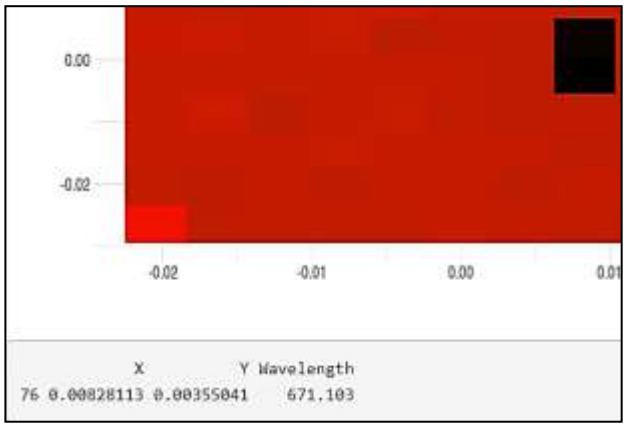


Fig. 20. Smoothing Spline Interpolation Input.

3.7 Downloading Plots/Tables

The program also allows the user to download any of the outputs. This can be done by pressing the download button that accompanies the plot/table. For the plots, they are downloaded as SVG (Scalable Vector Graphics) format, while the tables are downloaded as CSV format.

3.8 Conclusion and future work

In the above sections we introduced designing and developing of a Shiny application as an 'ad hoc' approach to teaching data science topics. In this project, we explored several features of the application, such as the

cleaning of data obtained from SEM, the transfer of variables of interest, the analysis and visualization of the data in terms of interactive heat maps, scatterplots, and histograms, and the option to export the displays

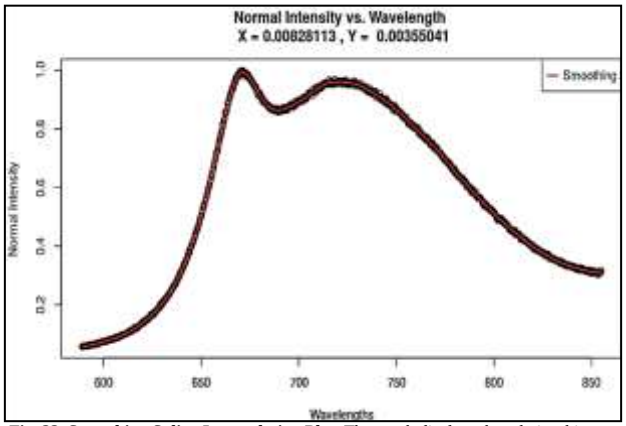


Fig. 22. Smoothing Spline Interpolation Plot. The graph displays the relationship between both the normal intensity and wavelengths along with the coordinates smoothing spline model.

in table format. These features allow solar cell researchers to investigate certain luminescence characteristics of photovoltaic cells. The informal 'ad hoc' effort of a project-based approach to teach sophisticated data skills to a mathematics major with a HILT approach delivered positive results and several challenges. A cohesive collaboration between the mentee and engineers resulted in an effective data product. Limited resources in terms of time posed a challenge. Learning platforms (e.g. Datacamp) supplemented with contact hours from mentors proved to be effective. Future projects will include more students and structured use of online learning platforms. We also want to improve the speed and efficiency of the current application to optimize for maximum functionality. One feature that we would like to add is the ability to apply machine learning techniques (neural networks) for image classification of the crystals on the photovoltaic cells.

Funding

This work has been supported by the National Science Foundation's Research Experiences for Undergraduates grant.

Conflict of Interest: none declared.

References

Belloum ASZ, Koulouzis S, Wiktorski T, Manieri A. *Bridging the demand and the offer in data science.* Concurrency Computat Pract Exper. 2019;31:e5200. https://doi.org/10.1002/cpe.5200

Berthold, M. R. (2019). *What does it take to be a successful data scientist?* Harvard Data Science Review. https://doi.org/10.1162/99608f92.e0eaabfc

Carson M. and Basiliko N. (2016) *Approaches to R education in Canadian universities. F1000Research.* ssional, Boston. 2016.

Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J. Package "Shiny." 2015; Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.677.6503&rep=rep1&type=pdf.

DeMasi O, Paxton A, Koy K (2020) *Ad hoc efforts for advancing data science education.* PLoS Comput Biol 16(5): e1007695. https://doi.org/ 10.1371/journal.pcbi.1007695.

Fawcett, L. (2018) *Using Interactive Shiny Applications to Facilitate Research-Informed Teaching and Learning,* Journal of Statistics Education, **26:1**, DOI: 10.1080/10691898.2018.1436999

Irizarry, R.A. (2020). *The Role of Academia in Data Science Education.* Harvard Data Science Review, 2(1). https://doi.org/10.1162/99608f92.dd363929

Reyes, A.L.P., Silva, T.C., Coetzee, S.G. *et al.* GENAVi: *a shiny web application for gene expression normalization, analysis and visualization. BMC Genomics* 20, 745 (2019). https://doi.org/10.1186/s12864-019-6073-7

Rowell, G.H. (2004), *"Assessment of Using Technology for Teaching Statistics,"* Paper presented at the ARTIST Roundtable Conference on Assessment in Statistics held at Lawrence University, August 2004.

Wickham H, RStudio. Package "data.table". 2019. Available from: https://cran.r-project.org/web/packages/data.table/data.table.pdf

Wickham H, Francois R. Package "dplyr." 2016. Available from: https://cran.r-project.org/web/packages/dplyr/dplyr.pdf.

Wickham H, Chang W, RStudio. Package "ggplot2". 2016. Available from: https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf.

Wickham H, RStudio. Package "hrbrthemes". 2020. Available from: https://cran.r-project.org/web/packages/hrbrthemes/hrbrthemes.pdf

Wickham H, RStudio. Package "plotly". 2020. Available from: https://cran.r-project.org/web/packages/plotly/plotly.pdf

Wickham H, RStudio. Package "readr". 2018. Available from: https://cran.r-project.org/web/packages/readr/readr.pdf

Wickham H, RStudio. Package "shiny". 2020. Available from: https://cran.r-project.org/web/packages/shiny/shiny.pdf

Wickham H, RStudio. Package "shinydashboard". 2018. Available from: https://cran.rproject.org/web/packages/shinydashboard/shinydashboard.pdf

Wing, J. M. (2019). *The Data Life Cycle.* Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.e26845b4.

Wszola LS, Simonsen VL, Stuber EF, Gillespie CR, Messinger LN, Decker KL, et al. (2017*) Translating statistical species-habitat models to interactive decision support tools.* PLoS ONE 12(12): e0188244. https://doi.org/10.1371/journal.pone.0188244