USING MACHINE LEARNING TO UNDERSTAND STUDENTS' GAZE PATTERNS ON GRAPHING TASKS

Alex Lyford[1] and Lonneke Boels[2,3]
[1]Middlebury College, Vermont, USA
[2]Freudenthal Institute, Utrecht University, Utrecht, The Netherlands
[3]University of Applied Sciences, Utrecht, The Netherlands
alyford@middlebury.edu

Graphs are ubiquitous. Many graphs, including histograms, bar charts, and stacked dotplots, have proven tricky to interpret. Students' gaze data can indicate students' interpretation strategies on these graphs. We therefore explore the question: In what way can machine learning quantify differences in students' gaze data when interpreting two near-identical histograms with graph tasks in between? Our work provides evidence that using machine learning in conjunction with gaze data can provide insight into how students analyze and interpret graphs. This approach also sheds light on the ways in which students may better understand a graph after first being presented with other graph types, including dotplots. We conclude with a model that can accurately differentiate between the first and second time a student solved near-identical histogram tasks.

INTRODUCTION AND BACKGROUND

It is well-established that students of all ages struggle to interpret displays of univariate data, such as histograms (e.g., Bakker 2004; Cooper & Shore, 2008; Kaplan et al., 2014). Kaplan et al. (2014) revealed that after an introductory statistics course at a university, students were even more likely to confuse the horizontal and vertical axes in histograms. In a task on which students did perform better upon completion, the authors believed this was more likely due to the task rather than an actual improvement in student skills. These struggles, however, are not just limited to students. For example, Cooper and Shore (2010) found that nearly two thirds of primary and secondary instructors identified and interpreted a histogram as if it were a case-value plot. In a large review of the literature, several other misinterpretations were identified (Boels, Bakker, Van Dooren, & Drijvers, 2019). Histograms have therefore proven tricky to interpret—students often incorrectly believe that the bars in a histogram represent single data observations or that little variation in the heights of bars in a histogram indicates little variation in the underlying data (Boels, Bakker, & Drijvers, 2019; Lyford, 2017). It is therefore critical that researchers and instructors work to better understand exactly how students interpret graphs such as histograms.

In order to better understand how students interpret histograms, we provided students with a range of tasks related to histograms and used infrared sensors to track students' gazes throughout their attempts to solve the tasks (Boels, Bakker, & Drijvers, 2019; Boels, Bakker, Van Dooren, & Drijvers, 2022; Boels, Ebbes, et al., 2018). We are particularly interested in changes in students' interpretation strategies after solving several graph tasks, including a sequence of dotplot tasks. Using students' gaze data, we seek to answer the research question: *In what way can a machine learning algorithm quantify differences in students' gaze data when students interpret two near-identical histograms with several other graph tasks in between?*

From a previous qualitative study, we learned from students' cued recall interviews that students' gaze patterns on the graph area reveal their solution strategy for an item (Boels, Bakker, Van Dooren, & Drijvers, 2022). Figure 1 shows a visualization of a small subset of gaze data from one student on the *Before* task. Each circle represents a fixation point, and the numbers inside the circles represent the order of fixations. For this particular subset of data, the participant spent the majority of their time on the left side of the graph area, and the participant made several horizontal saccades (rapid eye movement between fixation points) between the bars on the left of the graph and the values on the vertical axis. More specifically, a mostly horizontal gaze pattern on a histogram typically indicates an incorrect strategy for finding the mean, as if the graph were a case-value plot. A mostly vertical gaze pattern typically indicates a strategy for finding the mean as if the graph indeed were a histogram.

Using gaze patterns on the graph area is in line with other research stating that such gaze patterns can reveal learning in more detail (e.g., Hyönä, 2010). We could not find a useful and meaningful way to separate the graph area into suitable areas of interest (AOI), the latter being a more common method

In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the Gap: Empowering & Educating Today's Learners in Statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11 2022), Rosario, Argentina.* International Association for Statistical Education. iase-web.org

for using spatial measures in eye-tracking research. Instead, in the qualitative study, a novel measure was used that better captured the gaze pattern. This measure is based on the graph area as a whole, as one AOI, and uses the direction and magnitude of saccades. Other AOIs (e.g., the axis title) and the order in those AOIs did not appear to be relevant.

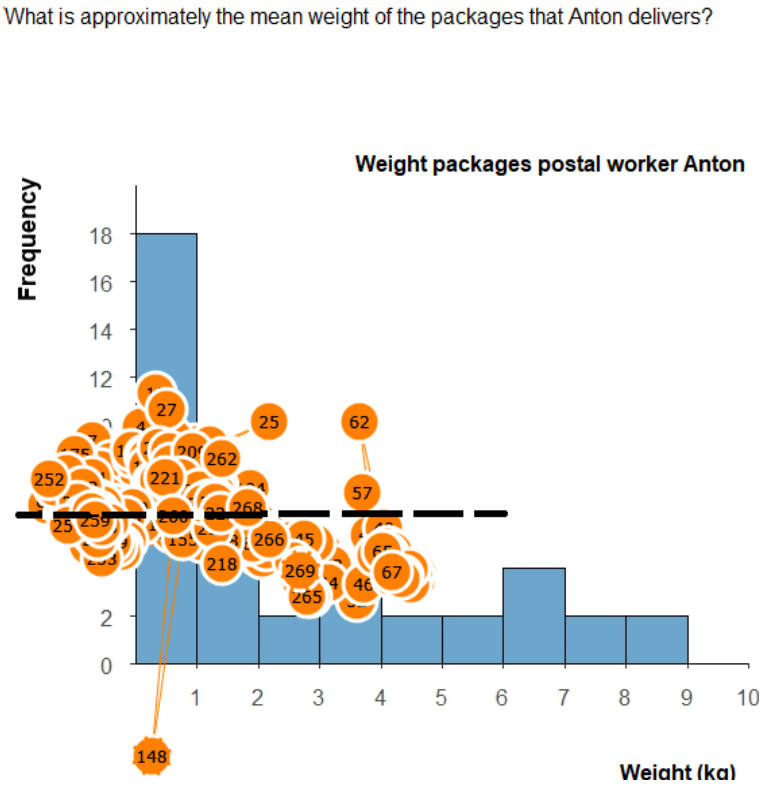

Figure 1. Visualization of *Before* task gaze data indicating an incorrect strategy

DATA COLLECTION AND RESEARCH QUESTIONS

We recruited 50 students from grades 10–12 at a public secondary school in The Netherlands. These students ranged in age from 15–19 years. Participation was voluntary, and participants received a small gift for their participation. Each of the 50 students answered 25 different tasks related to histograms, case-value plots, and dotplots (e.g., Boels, Bakker, Van Dooren, & Drijvers, 2022). This work focuses on two of the tasks, presented in Figure 2. Students were first shown the task on the left, henceforth referred to as the *Before* histogram task. Students then completed ten tasks, six with case-value plots and four with histograms, before they completed six tasks with dotplots. In these ten tasks, students were either asked to estimate the mean from the graph or to compare means. The six (non-stacked) dotplot tasks at the end of this sequence of tasks were intended to highlight and draw students' attention to specific features about a histogram that are often misunderstood, such as the heights of bars representing the number of observations in a given bin and not the value of a single observation. Because dotplots have only one axis, confusion about what the measured values represent is less likely. Finally, students were shown the task on the right, henceforth referred to as the *After* histogram task. In both tasks, students were asked: "What is approximately the mean weight of the packages that [Anton\Mo] delivers?" We note that these graphs are mirrored and otherwise essentially identical.

Students completed these tasks using a computer equipped with a Tobii XII-60 eye tracker. Students used a chin rest to stabilize their head movements, and gaze data were recorded throughout the experiment. The average accuracy of our eye-tracking data for the graph area is 13.4 pixels (0.27°), and the average precision is 0.58° without removing any of the participants' data. We then used these data to answer the following sub research questions:

- In what way do students' gaze patterns differ between the *Before* and *After* tasks?

- How can a machine learning algorithm predict whether gaze data come from a student's *Before* or *After* task?

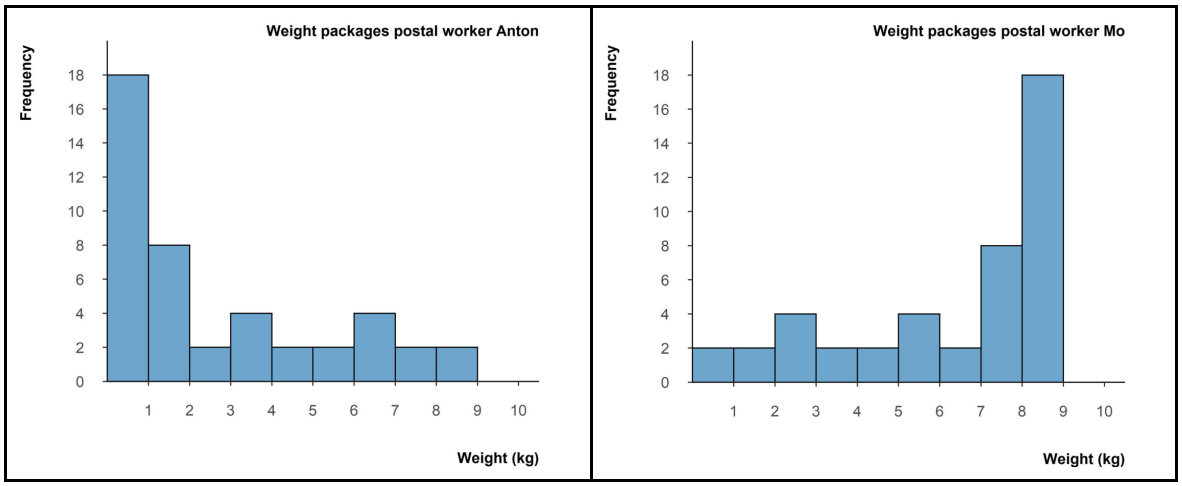

Figure 2. The *Before* and *After* histogram tasks

DATA PREPROCESSING AND MACHINE LEARNING

For our analysis, we only considered gaze data occurring in the graph area (i.e., the rectangular space formed by the positive horizontal and vertical axes). The horizontal and vertical coordinates of the gaze data were mirrored in the *After* task to match the corresponding location in the *Before* task to avoid that a machine learning algorithm would easily identify differences caused by the mirroring of the gaze pattern.

Our work uses random forests (see, for example, Breiman, 2001)—a type of machine learning algorithm—applied to the students' gaze data generated when completing these tasks. We create a set of features using fast transitions—known as saccades—between two positions (fixations) on the screen. In Figure 1, saccades are depicted as thin lines between the circles. An example of a feature is the magnitude of this saccade. The random forest we construct then uses these features to quantify differences in students' gaze patterns. In particular, we show students two near-identical histogram tasks with several other tasks with dotplots, case-value plots, and paired histograms in between.

A random forest combines several *decision* trees. A decision tree represents a series of binary decisions related to the presence or absence of certain features in a given set of gaze data. A fictive example of such a *decision* is: if at least ten percent of the saccades of a student have a magnitude between 100 and 200 pixels in length, inclusive, then the tree predicts that the eye movement belongs to the *After* task—otherwise it belongs to the *Before* task. This interval is notated as [100, 200], and the magnitude is an example of a feature we used as input into the random forest algorithm.

Figure 3 shows a superimposition of all saccades between 50 pixels and 600 pixels in length across all participants. We excluded saccades of less than 50 pixels in length given the accuracy of our eye tracking data and that these are most likely fixations. We also excluded saccades greater than 600 pixels in length because the graph area is only 600 pixels wide and 335 pixels high. We visualize these saccades in Figure 3 as if they are all centered at the origin.

Though we tried a number of different schemes, we ultimately classified saccades into discrete bins based on their direction and magnitude. Categorizing each of the saccades into the mutually exclusive bins shown in Table 1 allowed us to reduce the 'noise' in our data and enhanced the performance of our random forest. In other words, saccades of 100 pixels and 105 pixels are *essentially* the same length, and the random forest performed better when saccades of similar length and magnitude were placed into bins rather than treated continuously. The features used in our subsequent model are the proportion of saccades that fall into each of the given bins for a particular participant's gaze data.

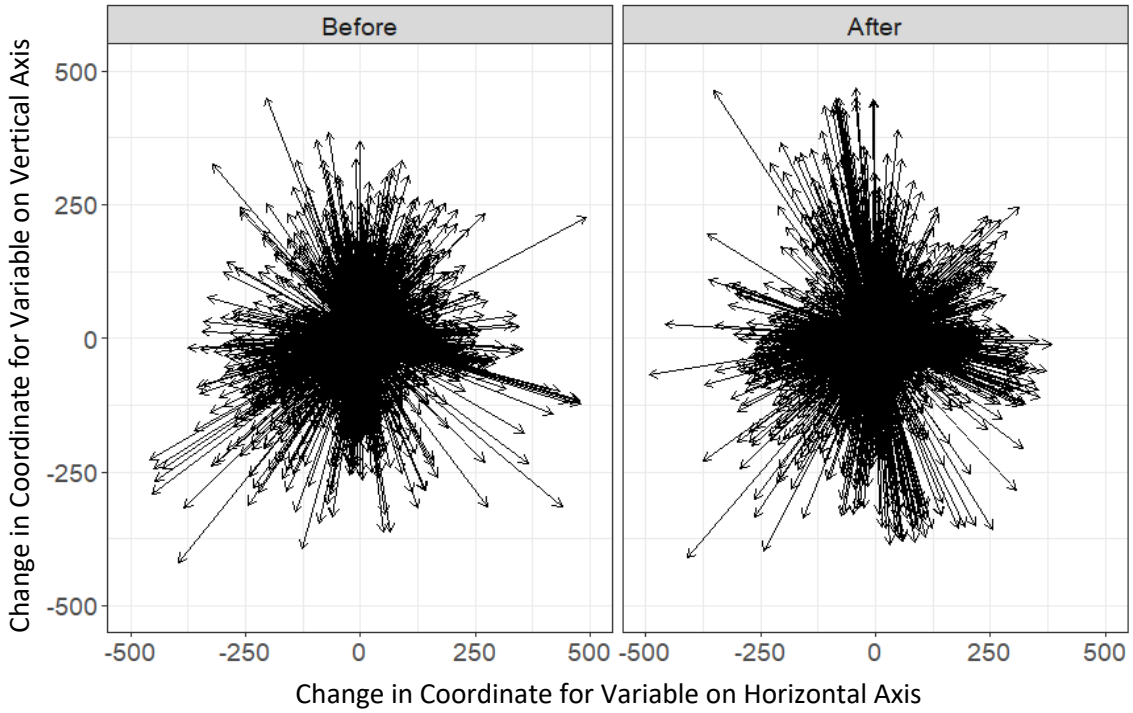

Figure 3. All saccades between 50 and 600 pixels in length centered at the origin

Table 1. Bins for feature creation

Direction	Magnitude
[0, π/4); [π/4, π/2); [π/2, 3π/4); [3π/4, π); [π, 5π/4); [5π/4, 3π/2); [3π/2, 7π/4); [7π/4, 2π)	[50, 100); [100, 200); [200, 400); [400, 600)

There are numerous statistical approaches that can predict and model a binary outcome, which in our case is whether the participant's gaze data came from the *Before* or *After* task. After trying several approaches, including linear discriminant analysis and neural networks, we decided to use random forests as our classifier. As explained below, random forests provide a reasonable mix of efficacy and interpretability. This allows us to both accurately classify gaze data as belonging to the *Before* or *After* task and also understand what features of the data allow the model to differentiate between the two tasks.

A random forest is a type of supervised learning algorithm—one that relies on training data to make predictions about future data. Though random forests can be used for quantitative responses, we use ours to predict whether the given gaze data come from a *Before* or *After* task. A random forest classifier is composed of many decision trees, each of which predicts whether the given gaze data come from a *Before* or *After* task. After each of the hundreds of trees makes a prediction, the random forest predicts the category with the majority of votes. Each individual tree is formed in the following manner.

A new, random set of data is generated by sampling from existing participants' gaze data with replacement. In other words, a participant's gaze data may appear multiple times in the new data set, or it may not appear at all. Then, a random subset of the features is chosen. Using a random selection of features makes the tree less accurate than using all of the features, but it allows for the creation of relatively uncorrelated trees, yielding a more accurate random forest. The given tree then iterates through each of the variables and a range of cutoffs, trying to find a cutoff that separates the data into mostly *Before* tasks and mostly *After* tasks. For example, suppose a large number of north-facing saccades of [100, 200) pixel magnitude (this roughly corresponds to students looking from the bottom of any given bar in the histogram to the top of the bar) were highly correlated with the first time a student looked a histogram—the *Before* task. The decision tree might decide to split responses into two groups based on whether or not a participant's gaze data contained at least ten percent of these medium-length, north-

facing saccades. This process of finding the best cutoff is repeated several times until certain stopping criteria are satisfied, yielding a tree of several binary decisions.

One important advantage of a random forest model is the ability to quantify how important each feature is to the model's overall performance. Because any given tree in the forest only uses a subset of the features, we can assess the importance of any feature by forcing the random forest to make predictions using only the trees where it did not have access to the given feature. Then, we can compare the accuracy of the entire forest (with all features) to the smaller forest only built using trees that did not use the given feature. The results of this analysis are provided in the following section.

Our random forest can reliably differentiate between students' gazes during the first and last histogram task, indicating that students may take different approaches to interpreting a histogram after being presented with dotplots. The importance of certain features in our model sheds light on some of these differences, helping us better understand what changed between a student's *before* and *after* histogram tasks.

RESULTS

Using leave-one-out cross validation, our random forest was 75% accurate, and treating *Before* as the positive case, our model had 72% sensitivity and 78% specificity. This is significantly better than random guessing (i.e., 50% accuracy, specificity, and sensitivity). Some participants' data were less reliable than others due to data loss, and this accuracy is likely improved with these participants' data removed. Some data loss in eye-tracking study is normal and occurs, for example, due to blinking. Other possible sources for data loss are wearing glasses or make-up, epicanthic eyes (skin fold on upper eyelid), or students looking above or below the screen while thinking.

Table 2 shows the quantification of feature importance for the five most important features in our random forest. The mean decrease in accuracy represents the decrease in the overall accuracy of the random forest compared to only using trees that did not have access to the given feature. Directions are provided both in radians and in cardinal directions, where SSE represents the south-by-southeast direction.

Table 2. Variable importance for random forest

Feature	Mean Decrease in Accuracy
Direction: [3π/2, 7π/4) (SSE), Magnitude: [100, 200)	10.5%
Direction: [3π/4, π) (WNW), Magnitude: [50, 100)	10.1%
Direction: [π/2, 3π/4) (NNW), Magnitude: [50, 100)	7.3%
Direction: [3π/4, π) (WNW), Magnitude: [100, 200)	6.8%
Direction: [3π/2, 7π/4) (SSE), Magnitude: [200, 400)	6.5%

The most important features were saccades of short-to-medium length, both in more horizontal directions (such as WNW) and in more vertical directions (such as NNW). At an aggregate level, this is confirmed by the saccades shown in Figure 2, where there are significantly more saccades in the SSE and WNW directions in the *After* task. In particular, Figure 2 also shows a substantially higher proportion of saccades of large magnitude in the vertical directions, most notably NNW and SSE, which the model confirms are the most important features for differentiating between *Before* and *After* tasks.

DISCUSSION AND IMPLICATIONS

What our results show is that the way in which students approached and looked at the *After* task was substantially and quantifiably different from the way they approached and looked at the *Before* task. However, we cannot say for certain that students learned how to interpret a histogram more correctly after completing tasks with dotplots.

We note that horizontal saccades may indicate that a participant is attempting to estimate the mean by finding the "balance point" in the *heights* of the bars (also called 'compensation,' Bakker, 2004), which is the correct way to estimate the mean of a case-value plot but the incorrect way to estimate the mean of a histogram. Because vertical saccades might indicate that a participant is trying to find the "balance point" of the histogram when estimating the mean, it is plausible that the increased proportion of vertical saccades is related to students figuring out—maybe through the support of dotplot items—how to correctly estimate the mean of a histogram.

Though our model is trained to classify data for these specific tasks, this machine learning approach is easily generalizable to gaze data for any tasks and is a step towards use in automatic identification of student strategies. As high-quality webcams become more ubiquitous and less expensive, it is conceivable that instructors could take advantage of our approach to predict whether a student is correctly or incorrectly interpreting a particular graph. If the algorithm detects—based on real-time eye-tracking data—that they are interpreting the graph incorrectly, the software could prompt the user with helpful feedback to nudge them in the correct direction. We also anticipate that this approach can help instructors and statistics education researchers better understand the misinterpretations that students have about graphs.

REFERENCES

Bakker, A. (2004). *Design research in statistics education: On symbolizing and computer tools* [Doctoral dissertation, Utrecht University]. CD-β Press. https://dspace.library.uu.nl/handle/1874/893

Boels, L., Bakker, A., & Drijvers, P. (2019b). Eye tracking secondary school students' strategies when interpreting statistical graphs. In M. Graven, H. Venkat, A.A. Essien, & P. Vale (Eds.). *Proceedings of the 43rd Conference of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 113–120). PME. https://www.igpme.org/publications/

Boels, L., Bakker, A., Van Dooren, W., & Drijvers, P. (2019a). Conceptual difficulties when interpreting histograms: A review. *Educational Research Review, 28*, Article 100291. https://doi.org/10.1016/j.edurev.2019.100291

Boels, L., Bakker, A., Van Dooren, W., & Drijvers, P. (2022*). Secondary school students' strategies when interpreting histograms and case-value plots: An eye-tracking study*. Manuscript submitted for publication.

Boels, L., Ebbes, R., Bakker, A., Van Dooren, W., & Drijvers, P. (2018). Revealing conceptual difficulties when interpreting histograms: An eye-tracking study. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July 2018) Kyoto, Japan.* ISI/IASE https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_8E2.pdf

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cooper, L. L., & Shore, F. S. (2008). Students' misconceptions in interpreting center and variability of data represented via histograms and stem-and-leaf plots. *Journal of Statistics Education, 16*(2). https://doi.org/10.1080/10691898.2008.11889559

Cooper, L. L., & Shore, F. S. (2010). The effects of data and graph type on concepts and visualizations of variability. *Journal of Statistics Education, 18*(2). https://doi.org/10.1080/10691898.2010.11889487

Hyönä, J. (2010). The use of eye movements in the study of multimedia learning. *Learning and Instruction, 20*(2), 172–176. https://doi.org/10.1016/j.learninstruc.2009.02.013

Kaplan, J., Gabrosek, J., Curtiss, P., & Malone, C. (2014). Investigating student understanding of histograms. *Journal of Statistics Education, 22*(2). https://doi.org/10.1080/10691898.2014.11889701

Lyford, A. (2017). *Investigating undergraduate student understanding of graphical displays of quantitative data through machine learning algorithms* [Doctoral dissertation, University of Georgia]. Ex Libris. https://getd.libs.uga.edu/pdfs/lyford_alexander_j_201705_phd.pdf