Evaluating the Impact of Binning 2D Scalar Fields

Lace Padilla, P. Samuel Quinan, Miriah Meyer, and Sarah H. Creem-Regehr

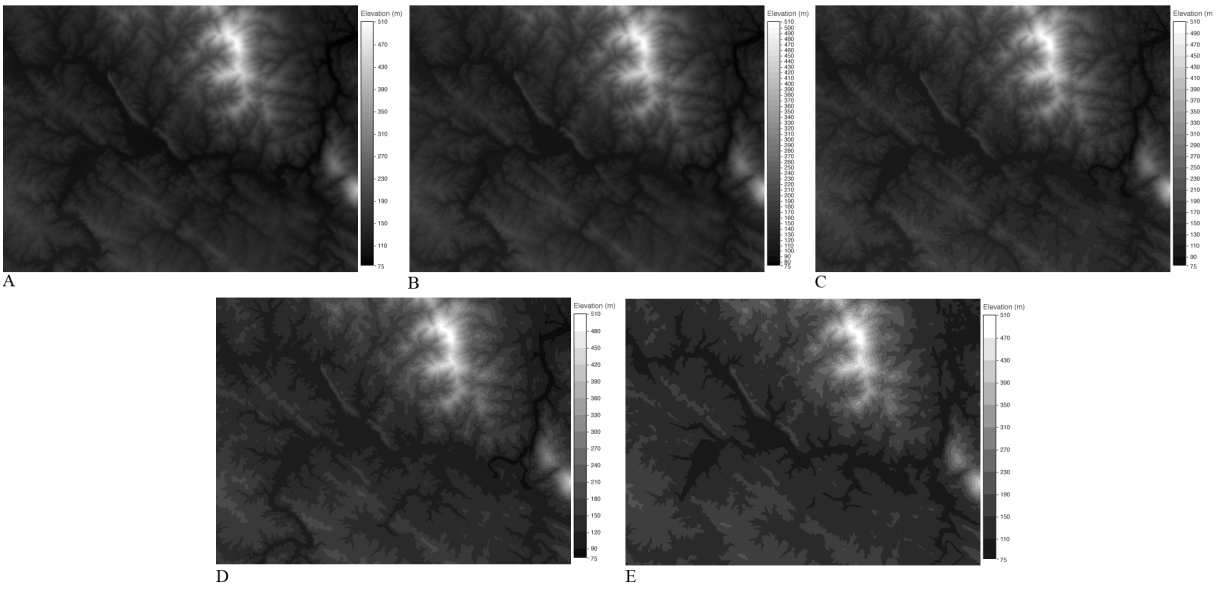


Fig. 1: Experimental stimuli for five binning conditions: A. Continuous, B. 10m binning, C. 20m binning, D. 30m binning, E. 40m binning

Abstract— The *expressiveness principle* for visualization design asserts that a visualization should encode all of the available data, and only the available data, implying that continuous data types should be visualized with a continuous encoding channel. And yet, in many domains binning continuous data is not only pervasive, but it is accepted as standard practice. Prior work provides no clear guidance for when encoding continuous data continuously is preferable to employing binning techniques or how this choice affects data interpretation and decision making. In this paper, we present a study aimed at better understanding the conditions in which the expressiveness principle can or should be violated for visualizing continuous data. We provided participants with visualizations employing either continuous or binned greyscale encodings of geospatial elevation data and compared participants' ability to complete a wide variety of tasks. For various tasks, the results indicate significant differences in decision making, confidence in responses, and task completion time between continuous and binned encodings of the data. In general, participants with continuous encodings were faster to complete many of the tasks, but never outperformed those with binned encodings, while performance accuracy with binned encodings was superior to continuous encodings in some tasks. These findings suggest that strict adherence to the expressiveness principle is not always advisable. We discuss both the implications and limitations of our results and outline various avenues for potential work needed to further improve guidelines for using continuous versus binned encodings for continuous data types.

Index Terms—Geographic/Geospatial Visualization, Qualitative Evaluation, Color Perception, Perceptual Cognition

1 Introduction

A foundational design principle in visualization is the *expressiveness principle*, which states that a visual encoding should express all of the relationships in the data, and only the relationships in the data [24, 35]. For a continuous data type, this implies that a continuous encoding channel is a good choice. In practice, however, domains such as cartography [43] and meteorology [36] have strong conventions that visualize continuous data with a discrete encoding. These domains rely on visual channels, such as color and saturation to encode a continuous function defined over two-dimensional space, known as a *2D scalar field*. They commonly do so by employing discrete colormaps or contour lines, also called isarithmic maps [43].

Existing literature provides little guidance about encoding continuous, 2D scalar fields with binned colormaps, or how this design decision affects data interpretation and decision making. Research into properties of colormaps for encoding continuous data types has largely focused on continuous colormaps [2, 28, 38, 48]. This line of research provides guidance on how to capture properties of the data, such as divergence around a center point [48] or emphasis on one end of the data range [2]. These papers go so far as proposing corresponding binned colormaps, but do not make claims, or even discuss, their efficacy for continuous data. Work on transfer function design has also proposed methods for binning colors, but with a focus on volumetric scalar fields, with the underlying goal of classifying materials or features [12], as opposed to directly understanding the continuous nature

- *L. Padilla is with the University of Utah Department of Psychology. E-mail: lace.m.k.padilla@gmail.com*
- *S. Quinan and M. Meyer are with the University of Utah School of Computing. E-mail: psq,miriah@cs.utah.edu.*
- *S. Creem-Regehr is with the University of Utah Department of Psychology. E-mail: sarah.creem@psych.utah.edu.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx/

of a scalar field itself.

An exception is work in cartography that explores the efficacy of using binned colormaps for encoding continuous values in choropleth maps. This work suggests that bins increase speed and accuracy [16]. However, the tasks that a choropleth supports are different from those of a 2D scalar field, making it unclear whether the results hold more generally across continuous data types. Additional work has focused on contouring 2D scalar fields, such as digital elevation models (DEMs), which involves connecting points of equal value with isolines. While pragmatic advice is offered for determining distances between isolines [51], no empirical studies have evaluated the impact of contouring on decision making.

The aim of this work was to test whether the decision to forgo the expressiveness principle when visualizing 2D scalar fields influences decision making across a broad spectrum of tasks. To this end, we conducted a user study in which participants were provided with continuous or binned greyscale encodings of the same topographical 2D scalar field, as shown in Figure 1. Then we tested participants' abilities to complete a range of increasingly complex tasks, in a geospatial context. Our results demonstrate that decisions made with binned encodings, that did not conform to the expressiveness principle, were either as accurate or better than those that followed the principle. These findings suggest that strict adherence to the expressiveness principle is not always advisable for visualizing 2D scalar fields.

In the rest of this paper, we lay out an overview of existing work on continuous versus binned colormaps, a description of our study and results, and a discussion of the findings and their implications for visualization. Section 2 includes discussions from cartography regarding the decision to use continuous or segmented mappings, as well as visualization work on encoding data using color and binning data effectively. This is followed by a detailed description of the experimental setup of our user study, along with results, in Section 3. We then conclude in Section 4 with a discussion of the implications of the findings and recommendations for future work.

2 Related Work

Work in cartography and visualization offers inconsistent views on binning continuous data. Modern approaches in cartography advocate for binning the data, while visualization suggests that it should be encoded with a continuous channel. Given the disagreement in the literature, we detail research from both disciplines as motivation for the present work.

2.1 Cartography

A prevalent assumption in cartography regarding choropleth maps is that binned colormap encodings of quantitative data are more comprehensible than continuous encodings [16, 51]. This perspective started a lively discussion in 1973 when Waldo Tobler developed a method of creating continuous greyscales using a line plotter [47]. Employing Tobler's method, cartographers had the option between continuous greyscales or discrete tone encodings, and thus a heated debate ensued [43]. Jean-Claude Muller was a key advocate for the new continuous method, conducting various studies demonstrating that on printed maps generated using a facsimile method [34], users could discern similar map features with continuous encodings that are discernable by binned encodings [32, 33]. Muller's and Tobler's work was contested notably by Michael Dobson in a series of commentaries [10, 9] where he detailed a laundry list of grievances with continuous encodings of choropleth maps, including but not limited to: increased time to perform tasks, reduced ability to control the amount of information presented, decreased pattern discrimination, and increased mental workload [9]. While these specific findings may not necessarily generalize directly to either 2D scalar field visualizations or modern display technologies, the perspectives exemplify relevant considerations for the trade-offs between the richness of presented data and usability.

Many of the tasks performed with choropleth maps are fundamentally different from those performed with 2D scalar field visualizations, as the latter are often used to make judgments about continuous features in the data. Thus, it is unclear to what extent standard practices used for binning choropleth maps extend to 2D scalar fields. Other work in cartography and GIS, however, has focused on specifications for contouring 2D scalar fields such as digital elevation models [51]. This work suggests that contouring aids in interpreting continuous features [20, 15]. Contouring involves defining spans between isolines and, in the case of contour maps, encoding the span between isolines with a discrete encoding such as color or greyscale. As the resulting visualizations can be similar between contouring and discrete binning, there is likely a relationship between tasks that both methods support. Defining the appropriate relative distance between isolines is domain specific and commonly a judgment call by the designer. Watson recommends that, *"By looking at the range of heights in your dataset, you will get an idea of the levels that suit your data"* [51]. More advanced techniques in GIS include methods for generating contours for a given task based on classifications of data features such as peaks or valleys (For a review see [15]). Various techniques for representing surface topology as graphs have also been proposed, notably the *Reeb graph* [41] and the *Morse-Smale complex* [8]. These approaches represent the relationship of critical topological features as nodes in a network visualization [23]. However to our knowledge, no user studies in GIS have been conducted to assess the influence of discretization techniques on decision making.

2.2 Visualization

An extensive body of research within the visualization literature considers the role of color for encoding data (For a review see [54]). Here, we summarize work concerning decisions and issues with encoding continuous data with color.

Previous work in visualization concerning color encoding channels primarily recommends a binned colormap for categorical and ordinal data and a continuous one for quantitative data [2, 28, 48]. Much of the work on discrete colormaps centers around Brewer's perceptually optimized colormaps [7, 16, 6]. These colormaps are widely adopted [54] and have inspired guidelines [53] and generative systems [52]. Alternatively, Healey outlines a systematic method for generating effective isoluminant color schemes with up to seven colors by controlling for color distance, linear separation, and color category [18].

For continuous colormaps, the literature largely focuses on principles of colormap design, such as smart defaults [30] or domain-specific recommendations [36]. However, it is generally accepted that there is no one best color scale, and a variety factors dictate effective color usage [38]. For example, Ware offers experimental evidence that hue variation is better for extracting absolute values, while luminance change enables enhanced form comprehension in continuous 2D scalar fields [50]. Bergman et al. provide continuous colormap recommendations based on data class, task, and spatial frequency [2]. Rheingans outlines various design considerations for creating successful continuous colormaps: goals, the nature of the data, the intended audience, the visualization as a whole, and cultural connotations [38]. Additionally, Tominski et al. [48] provide a formalized set of task-driven continuous colormap recommendations based on characteristics of the data, tasks and goals, the user, and the display.

Where principles for continuous colormap design have been generalized to corresponding binned colormaps, the discussion is limited to applying those colormaps to nominal or ordinal data [28, 48]. Bergman et al. [2] provide one of the only explicit discussions about encoding continuous data discretely. They argue that using a binned colormap on continuous data is useful for *segmentation* tasks, where the goal is to explicitly show multiple features. There is, however, no discussion of what effect a binned encoding might have on *isomorphic* tasks, where the goal of the representation is to reflect the structure in the data. Additionally, as Mittelstadt et al. point out, many of the existing guidelines for creating effective colormaps are geared toward a single elementary task, while real-world scenarios typically require more complex, combined tasks [28]. The goal of our work is to look at the influence of binned colormaps across a wide range of both segmentation and isomorphic tasks for 2D scalar fields.

Work on visualizing continuous functions in three-dimensional space using color largely focuses on techniques for volume visualiza-

tion. Transfer function (TF) design explores methods and techniques to specify optical properties, including both color and opacity, for different portions of volumetric datasets. TFs are predominately used to classify materials based upon features in a scalar field [12], which are time-consuming to generate manually. As such, there have been a variety of attempts to automate or optimize the data classification step based on the underlying features of the data [22, 13, 49]. This classification role closely resembles Bergman et al.'s notion of segmentation tasks discussed above, where the goal is to elucidate features, as opposed to isomorphic tasks, which we are also interested in testing.

3 Methods and Aims

Even though the expressiveness principle suggests that continuous data should be encoded with a continuous channel, it is likely that some visualization practitioners have discovered, through experience, that binning continuous data types is beneficial in many situations. We argue that violations of the expressiveness principle should be used to calibrate efficacy. Visualization efficacy is domain specific, but can generally be gauged by speed, accuracy, and confidence in task performance. Focusing on the geospatial domain, we predict that forgoing the expressiveness principle for the 2D scalar field data used in the current study will likely increase accuracy and confidence in task performance, which is consistent with work suggesting that binned choropleth maps are preferable. However, binning may reduce speed, showing a classic speed-accuracy trade-off.

To test our hypothesis, we employed a between-subjects design, with five different groups of participants each performing tasks with one of five different binning conditions. Each of the binning groups was comprised of participants from Amazon's Mechanical Turk with participation criteria set to master-class workers in the US. The average screen size used by the population in this study was 1600 x 900 cm, the median was 1280 x 1024 cm, and the mode was 1366 x 768 cm. The five binning conditions, shown in Figure 1, include one perceptually linear continuous greyscale encoding, along with four different approximately perceptually linear binned greyscale encodings. These were based on binning intervals ranging from fine to coarse. Type of task was a within-subject variable, with each participant completing 12 randomly ordered tasks ranging in complexity.

3.1 Stimuli Generation

All of the experimental stimuli were created from a DEM projected into the continuous range [0, 100]. This range allowed us to start with a dataset where the data range explicitly matched the range of the lightness channel, L^*, in the commonly used CIELAB color space. CIELAB was designed to be perceptually uniform when evaluating small color differences [44]. These differences are defined based on a Euclidean distance metric, where one unit of Euclidean distance ($\Delta\, E^*_{ab} = 1$) was designed to equal one just-noticeable color difference (JND). Other work, suggests that $\Delta\, E^*_{ab} = 2.3$ is a more accurate benchmark for just-noticeable color differences in CIELAB space [26]. While more accurate color difference spaces and color appearance models exist [31], the simplicity of CIELAB has made it the most practical choice for designers [45].

We generated the perceptually linear continuous greyscale encoding by mapping the elevation data values to CIELAB lightness, L^*, while leaving $a^* = b^* = 0$. We utilized the Mahy et al. [26] benchmark of $\Delta\, E^*_{ab} = 2.3$ to create four binning conditions. The smallest bin was *nearly continuous* ($\Delta\, E^*_{ab} = 2.3$), which resulted in a set of bins approximately 1 JND apart, starting at the highest elevation. The three coarser binnings were then generated by doubling ($\Delta\, E^*_{ab} = 4.6$), tripling ($\Delta\, E^*_{ab} = 6.9$), and quadrupling ($\Delta\, E^*_{ab} = 9.2$) the initial bin size. The resulting images are shown in Figure 1. As none of these quantities neatly divide into the data range, the lowest elevation bins were smaller and variable in size. Therefore, we specifically avoided asking questions regarding the lower elevations as this inconsistency could have influenced participants' responses.

Greyscale is generally considered to have a limited perceived dynamic range and approximations for the number of JNDs varies [38]. The data range of our continuous encoding matches the theoretical maximum of 100 JNDs [38], and we base our finest binning on the most conservative estimate of 43 JNDs [26].

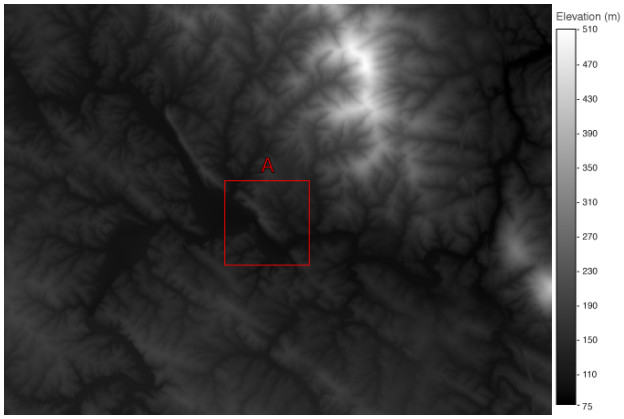


Fig. 2: An example of the continuous encoding for the Highest Point in Region task.

It should be noted that we did not account for perceptual effects such as *simultaneous contrast* that have been well documented for greyscale [50]. Recent work by Mittelstädt et al. has proposed and subsequently improved upon post-processing methods that counter contrast effects [29]. However, prioritizations based on task are required. As such, we could not appropriately optimize a single image for all tasks using this technique.

The CIELAB pixel values were converted to sRGB to create the images used in the study. The sRGB color space is non-linear standardized RGB specification, which is widely accepted as the default RGB specification in web-imaging [44]. Given that we ran the study online, sRGB was a logical choice. [1]

For those stimuli requiring defined regions, the regions were layered on the sRGB images via post-processing (e.g., Figure 2). Additionally, to ensure that our visual stimuli more accurately matched the nature of our questions, we linearly scaled the legend values such that $\Delta\, E^*_{ab} = 2.3$ equates to a 10-meter difference in elevation. The five binning techniques or test conditions are named based on this scale (10m n = 100, 20m n = 99, 30m n = 99, 40m n = 100, and Continuous n = 101). From the participants' point of view, our questions regarded topographical data spanning a range of [75, 510] meters. We collected responses using this range and for appropriate tasks transformed the response values back to the original [0, 100] range for analysis.

3.2 Tasks

This study examines how visualization techniques that forgo (i.e. binned encodings) and conform (i.e. continuous encoding) to the expressiveness principle influence the general public's interpretations of data. Brehmer and Munzner [5] provide a thorough typology of visualization tasks that we selected from to test this aim. We focused on discovery-based tasks, which involve generating predictions and testing theories in the data. This categorization of tasks incorporates *search* and *query*. Search tasks can be categorized as *lookup* (when both the target and location are known), *browse* (when the target is unknown, but the location is known), *locate* (when the target is known, but the location is unknown) and *explore* (when neither the target or the location are known). When the target or set of targets has been located, then a query step is performed, where the participants *identify*, *compare* or *summarize* the targets. We did not address the *how* and *what* sections of Brehmer and Munzner's [5] typology since these are most relevant to interactivity, which falls outside of the scope of the current work. Further, we were interested in assessing performance on both elementary (on single values) and synoptic (on sets of values) tasks [48, 1]. As such, we utilized comparisons, rankings, and

[1] Alternative methods include Szafir et al.'s parametric color difference model [45], which was designed to incorporate the real-world perceptions of target populations like Amazon Mechanical Turk.

summarizations of regions to this aim.

In an extensive initial pilot study, we tested six tasks that were chosen to represent a broad spectrum of possible visualization tasks ranging in complexity. We were particularly interested in including tasks that included the integration of multiple features, since we had observed users attempting this in applied settings in our previous work [37]. In the simplest task, participants were instructed to click on the highest location on a map, which was intended to test the participants' ability to locate relevant values with the different binning techniques. Prior work has found that participants are quite accurate at mouse cursor location precision with errors ranging from 2%-3% [3], making mouse clicks a reasonable choice of response type. In the most complex task, participants were asked to rank locations according to flood risk, which involved integrating both elevation information and distance from a river. This task was both complex and modeled after real-world scenarios. While our pilot study was informative, the problem was that glyphs were placed on the visualizations to indicate specific points, which occluded relevant values. The present study addressed this issue by asking participants to make relative judgments about regions rather than occluded points. Additionally, the present work uses mouse-click responses and comparisons because they were the most successful task types in the pilot study. We further expanded the spectrum of tasks to include previously untested variants of the aforementioned tasks along with asking participants to summarize fields of data.

For all of the tasks, participants were shown one topographical map with a single binning technique. Time to complete each task was recorded along with a subjective measure of confidence in task performance. After each task, participants were presented with a 1-7 Likert scale to measure their confidence (1 = not at all confident, 7 = very confident). Additionally, we asked participants questions about monitor display capacities, included black level, white saturation, and gradient banding. Screenshots of each task and the display questions are available in the supplementary materials.

3.2.1 Locate or Explore and Identify Tasks

Depending on a participant's approach, the following tasks could be completed by either a *locate* or *explore* strategy followed by an identification step, according to Brehmer and Munzner [5]. For example, if a participant identified a relevant value on the legend (e.g., the highest elevation) and then looked for that value, this would be a locate task — seeking for a known target in an unknown location. If instead the participant searched for characteristics in the data which are unknown targets (e.g., relative height or slope), this would be an exploration task [5]. Participants then identified the target via mouse click.

Highest Point. Participants were asked to "Click on the highest point in the image."

Highest Point in Region. Participants were asked to "Click on the highest point inside box A." This task was intended to be a more complex version of the former task, by prohibiting participants from distinguishing the highest elevation on the legend and then identifying that location on the map. Instead, participants had to locate the highest point within a region, which was outlined in red and overlaid on the map (See Figure 2). All of the tasks that required responses in regions used the same type of red outlined overlay. While participants could have looked up values in the legend, this task did not require this strategy and likely primarily depended on participants' ability to make relative comparisons between elevations within the region.

Locate 275m. Participants were asked to "Click on any point in the image that has an elevation of 275 m." This task was intended to require participants to use the legend then search in the visualization for the relevant value.

3.2.2 Locate or Explore and Compare Tasks

For these tasks, participants could use either a locate or explore strategy, followed by a comparison of targets.

Compare Heights. Participants were asked to "Find the highest point in each box. Comparing these locations, which box contains the higher location?" This question was a comparison task that could be implemented without occluding the relevant locations on the map (which would have occurred if specific points were overlaid on the map). Although allowing participants to interactively remove and display a glyph would have been one solution to the occlusion issue, we felt that it was important to limit the scope of these tasks to static images. Thus, we created a task requiring participants first to identify the highest location in two regions and then to compare those values.

Rank Heights. Participants were asked to "Find the highest elevation in each box. Then rank each box according to the elevation of its highest location." This task was intended to increase complexity from the Compare Heights task. Rather than comparing the highest locations in two regions, the participants were instructed to rank the highest locations in three regions.

3.2.3 Explore and Identify Task

Exclusively explore tasks required participants to make judgments using characteristics of the data rather than known targets.

Steepest. These are one of the most difficult and rarely employed tasks with the general public. We tested whether participants could identify the relationship between points by using the colloquial understanding of slope. Participants were provided the following description of slope, "Terrain is steeper where there is a larger change in elevation between adjacent points." In this task, participants were asked to "Click on the steepest point in the map."

3.2.4 Lookup and Identify Task

Lookup. Lookup tasks are characterized by both the target and location being known. Participants were asked to identify the height value of the steepest point they selected in the previous question. This question was presented on the same page as the Steepest task, and asked, "What are the highest and lowest elevations adjacent to the point you clicked on?"

3.2.5 Explore and Compare Task

Steepness Compare. This question was intended to assess participants' ability to compare relationships in the data through exploration, using the "steepest point" task described above. Participants were provided with the instructions, "In each box, find the steepest point. Comparing these locations, which box contains the steepest point?"

Steepness Rank. This was one of the most difficult tasks that we tested. Participants were given the instructions, "Inside each box, find the steepest point. Rank each box according to their steepest points."

3.2.6 Browse and Summarize

Browse tasks are characterized by the location being known and the target being unknown. The following browse tasks used specific regions with in which the participants were asked to determine characteristics about the region.

Average Height. This task was intended to test participants' ability to integrate a large set of relationships in the data. We asked participants to provide a summary statistic from a specified region. Participants were asked, "What is the average elevation of the area in box A?" All participants were screened at the start of the experiment for a basic understanding of how to calculate an average from a string of numbers. While calculating numeric averages is fundamentally different than visual spatial aggregation, we wanted to ensure that participants were familiar with the term average and related concepts.

3.2.7 Browse and Compare

Continuous. In this task, participants summarized two regions and then compared the regions. Participants were provided with the instructions, "Find the average elevation in each box. Comparing these locations, which box contains the higher average elevation?"

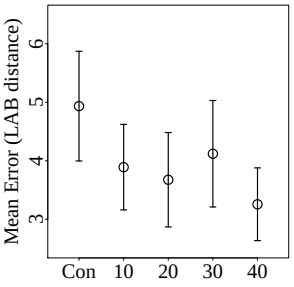


Fig. 3: Highest Point Task. Error bars (95% CI)

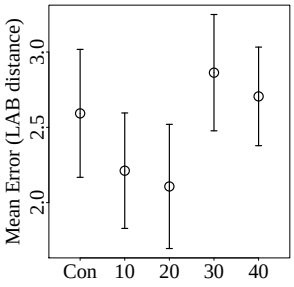


Fig. 4: Highest Point in Region Task. Error bars (95% CI)

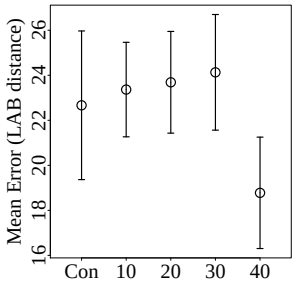


Fig. 5: Locate 275m Task. Error bars (95% CI)

3.2.8 Combination Task

The final combination task involved a multistep iterative process that could include all or a subset of the previously defined processes in any order, depending on a participant's strategy.

Steepest at 355. This question was intended to increase complexity and relevance by combining a locate task and a browse task. Many tasks that are performed in real-world scenarios require combinations of tasks. Participants were asked to "Click on the steepest point in the map at an elevation of 355 m."

3.3 Analysis

3.3.1 Overview of Results

As an overview, there were no cases where there was a clear advantage in terms of accuracy with the continuous encoding.The results described in more detail below show that for some tasks there was no difference in accuracy between binned and continuous encoding (e.g., Compare Heights and Rank Heights tasks), but in other tasks, binned encodings showed higher accuracy than continuous (e.g., Highest Point and Lookup tasks). Interestingly, we found that, generally, participants were faster to complete the tasks with the continuous encoding than with some of the binned encodings. These findings suggest that while participants with continuous encodings appear to complete tasks more quickly, continuous encodings did not improve accuracy. The following sections will detail the analyses and results for each task.

For all of the following ANOVA analyses, post hoc Tukey HSD comparisons were used to test significant differences between binning techniques for each task, correcting for multiple comparisons. In addition, all analyses controlled for effects of display parameters. (See supplementary material for full output.) The task accuracy analyses and results are presented for each task separately, since the nature of the task determined the analyses used (Sections 3.3.2 - 3.3.9). Analyses and results for task completion time and confidence ratings, summarized across tasks, are reported in Sections 3.3.10 and 3.3.11.

3.3.2 Locate or Explore and Identify Tasks

1. Highest Point. Participants were more accurate when using the 40m binning display compared to the continuous display. This result was revealed by a one-way between-subjects ANOVA testing the effect of binning technique on accuracy of click location. Click accuracy was calculated by subtracting the L value at the clicked location from the L value of the highest location (100), creating an error score in CIELAB space distance. Outliers that were two standard deviations (SD) above the mean were removed (6% of trials). The analysis revealed there was a significant effect of binning technique on click location accuracy, $F(4,463) = 2.634, p = .033, \eta_p^2 = .021$. The mean error score for continuous $(M = 4.93, SD = 4.55)$ was significantly higher than 40m $(M = 3.25, SD = 3.06)$ at a $p < .05$ criterion (See Figure 3).

2. Highest Point in Region. Participants were more accurate using the 20m binning technique, specifically, when compared to the 30m binning. This result was revealed by a one-way between-subjects ANOVA testing the effect of binning technique on accuracy of click location within a region. Click accuracy was calculated by subtracting the L value at the clicked location from the L value of the highest location in the region (22.40), creating an error score in CIELAB space distance. Outliers $2SD >$ mean were removed (3.8 % of trials). The analysis revealed there was a significant effect of binning technique on click location accuracy, $F(4,472) = 2.744, p = .028, \eta_p^2 = .023$. The mean error score for 20m $(M = 2.10, SD = 2.04)$ was significantly lower than 30m $(M = 2.86, SD = 1.89)$ binning, $p < .05$ (See Figure 4).

3. Locate 275m. Participants were more accurate using the 40m binning technique, specifically when compared to the 30m binning. This result was revealed by a one-way between-subjects ANOVA testing the effect of binning technique on accuracy of clicking a location at an elevation of 275 m. Click accuracy was calculated by subtracting the L value at the clicked location from the L value corresponding to an elevation of 275 m (45.95), creating an error score in CIELAB space distance. The analysis revealed that there was a significant effect of binning technique on click location accuracy, $F(4,491) = 2.763, p = .027, \eta_p^2 = .021$. Mean error score for 40m $(M = 18.77, SD = 12.46)$ was significantly lower than 30m $(M = 24.12, SD = 12.87)$ binning, $p < .05$ (See Figure 5).

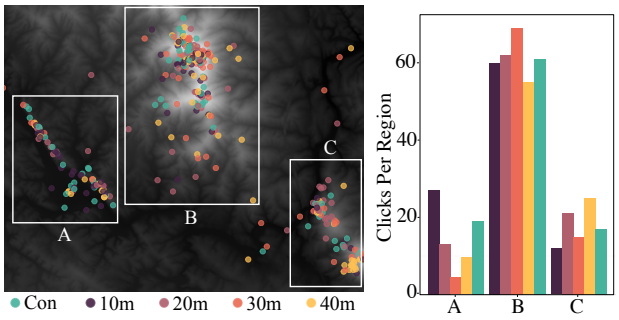
3.3.3 Locate or Explore and Compare Task

4. Compare Heights. A binomial logistic regression found that there was no significant effect of binning technique on a height comparison, $\chi^2(df = 7) = 7.189, p = .40$. While participants were overall more likely to correctly choose the higher elevation (73% responded correctly), this did not vary across binning condition.

5. Rank Heights. An ordinal logistic regression was used to test the effect of binning technique on rankings of regions by height. Participants responded to this question by entering rankings of 1-3 (three indicating the highest region and one the lowest) for regions A, B, and C. The ordinal logistic regression equation did not significantly predict rankings when using binning technique and region as predictors, $\chi^2(df = 9) = 10.63, p = 0.30$. In general, participants performed poorly on this task: 35% correctly ranked B as the highest, 52% correctly ranked A second, and 35% correctly ranked C as lowest, which makes it difficult to compare the binning conditions.

3.3.4 Explore and Identify Task

6. Steepest. Preliminary diagnostic checks of the data revealed that participants predominantly clicked in three separate areas (See Figure 6a). We assessed steepness by looking at the DEM's gradient, which we approximated using a Sobel operator. The point with the greatest gradient magnitude (44.33), was located in Area A. However, this point was located proximal to a low elevation depression in the map. The vast majority of clicks were in Area B, which was proximal to a peak feature. This pattern of responses suggests that participants' click decisions may have been influenced by an assumption about the relationship between steepness and peaks. Because of this initial pattern, we decided to provide a qualitative description of the number of clicks per binning condition in each area separately. As shown in Figure 6b, there are different proportions of total clicks in regions A and C across



(a) Click locations overlaid on the continuous encoding, showing three main regions.

(b) Number of clicks per region by binning technique.

Fig. 6: Steepest Task

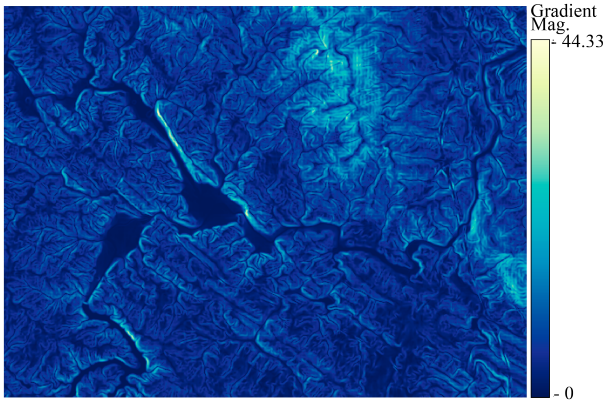


Fig. 7: A visualization of the spatial frequency of the DEM used in this study.

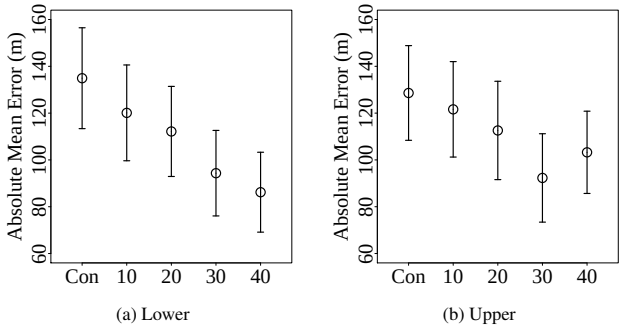
binning conditions. We can speculate that different binning techniques influenced some participants' incorrect assumptions relating elevation and steepness; however, future studies are needed to fully understand these effects.

An additional complicating factor when asking participants to make judgments of steepness comes from the high spatial frequency of the gradient magnitude (See Figure 7). In the areas with the greatest gradient magnitude, relatively close points (i.e., only a few pixels apart) could have vastly different gradient magnitude values. While beyond the scope of this study, future work should investigate strategies to account for this.

3.3.5 Lookup and Identify Task

7. Lookup. This task followed the Steepest Point task and asked participants to report the lowest and highest values adjacent to their click; thus, accuracy of lowest and highest points were analyzed separately. For the lowest value, participants were the least accurate using the continuous encoding, specifically when compared to the 30m and 40m binnings. Accuracy was calculated by subtracting the reported lower adjacent elevation from the actual lower adjacent elevation, creating an error score in CIELAB space distance. Outliers $> 2SD$ above the mean were removed (7.8% of trials). A one-way between-subjects ANOVA, (controlling for reported higher adjacent elevation), showed there was a significant effect of binning technique, $F(4,451) = 4.418, p = .001, \eta_p^2 = .057$. The mean error score for the continuous encoding ($M = 134.91, SD = 105.20$) was significantly less accurate than the 30m binning ($M = 94.33, SD = 87.21$) and 40m binning ($M = 86.20, SD = 82.88$), $p < .05$ (See Figure 8a).

A similar analysis was performed on the reported upper adjacent elevation. Outliers $> 2SD$ above the mean were removed (2% of trials). There was a significant effect of binning technique on reporting the upper adjacent elevation, $F(4,479) = 2.602, p = .03, \eta_p^2 = .044$. However, post hoc Tukey HSD comparisons did not reveal significant differences between the conditions at the $p < .05$ level when accounting for multiple comparisons. To understand the main effect of binning, planned contrast codes were generated which compared the continuous encoding to the binned encodings. Similar to the lower elevation analysis above, we found that the continuous encoding ($M = 128.60, SD = 100.90$) was less accurate than binned encoding ($M = 107.36, SD = 97.16$), $F(1,490) = 7.79, p = 0.005$ (See Figure 8b).



(a) Lower

(b) Upper

Fig. 8: Lookup Task. Error bars (95% CI)

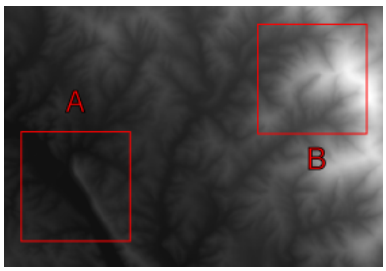


Fig. 9: An example portion of the continuous encoding from the Steepness Compare task, showing Area A and Area B.

3.3.6 Explore and Compare Task

8. Steepness Compare. A binomial logistic regression found that there was no significant effect of binning technique on a steepness comparison, $\chi^2(df = 7) = 1.35, p = .98$. Area A contained the steepest point with a magnitude gradient of 44.33, and Area B contained the second steepest point with a magnitude gradient of 41.65 (See Figure 9). 62% of participants incorrectly selected Area B as containing the steeper point. Similar to the Steepest task, these findings suggest that participants' prior understanding of topography and an assumption about a connection between steepness and peaks could have biased incorrect responses. Additionally, these findings may be influenced by the issues related to the high spatial frequency of the gradient magnitude noted in Section 3.3.4.

9. Steepness Rank. An ordinal logistic regression was used to test the effect of binning technique on rankings of the greatest gradient magnitude between regions. Participants responded to this question by entering rankings of 1-3 (three indicating the greatest gradient magnitude region and one the least) for regions A, B, and C. Each of these regions were selected because they contained the 3rd, 4th, and 5th steepest points (See Figure 10). The ordinal logistic regression equation did significantly predict rankings when using binning technique and regions as predictors $\chi^2(df = 9) = 137.79, p < .00$, but binning technique did not effect gradient magnitude rankings. Both the regions and the order of rankings were significant predictors of rankings.

This task and the prior tasks relating to steepness judgments suggest that a number of different factors likely influenced a reduced effect of binning, such as prior assumptions about how elevation peaks and slopes relate and variable gradient magnitudes.

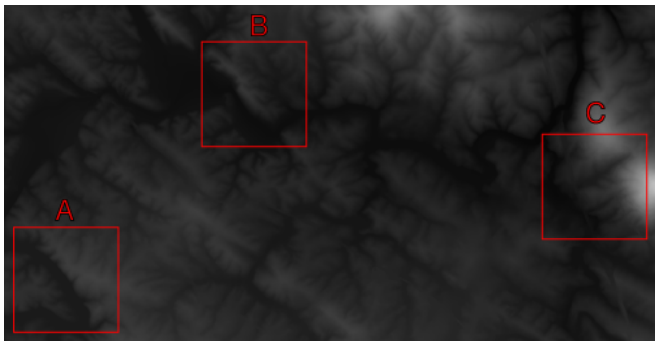


Fig. 10: An example of the continuous encoding in the Steepness Rank Task.

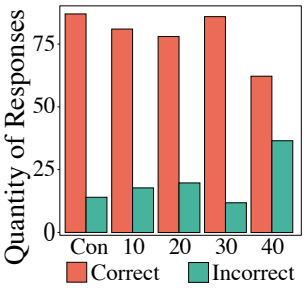


Fig. 11: Proportion of correct and incorrect responses in the Average Height Compare Task for each encoding condition.

3.3.7 Browse and Summarize

10. Average Height. A one-way between-subjects ANOVA, testing the effect of binning technique on reported average height in a region, while controlling for the display parameters, found no effect of binning on response accuracy $F(4,474) = 1.241, p = .293$, after removing outliers $> 2SD$ above the mean error score (3.4% of trials). The finding may be due to the large magnitude of errors in all of the binning conditions $(M = 101.105, SD = 47.44)$.

3.3.8 Browse and Compare

11. Average Height Comparison. *Participants were least accurate using the 40m binning technique*. This result was revealed by a logistic mixed model, testing the effect of binning technique on a comparison of steepness. Comparison accuracy was calculated by coding correct responses as 1 and incorrect rankings as 0. The analysis revealed that the 40m binning displayed a significantly lower proportion of accurate responses (63%) compared to the 10m (82%), 20m (79%), 30m (87%) binnings and the continuous encoding (86%). The model accounted for a significant proportion of variance compared to a baseline model with no predictors, $\chi^2(df = 7) = 22.80, p = .001$ (See Figure 11).

3.3.9 Combination task

12. Steepest at 355. Preliminary diagnostic checks of the data revealed that the majority of participants were clicking at elevations much lower than 355m as instructed (See Figure 12). We defined regions based on relative proximity in elevation to 355m, making three groups (below 255, 255–455, above 455) and qualitatively assessed the frequency of clicks in each region by binning condition. It is clear that many participants did not correctly identify the 355 elevation, as the greatest number of clicks fell in the "low" category. Overall, the wide distribution of click responses makes it difficult to assess differences between the binning conditions and suggests that the task was difficult for users.

3.3.10 Task Completion Time

Eight of the 12 tasks showed significant differences in task completion time between binning techniques, with the continuous encoding displaying consistently faster performance than some of the binnings.

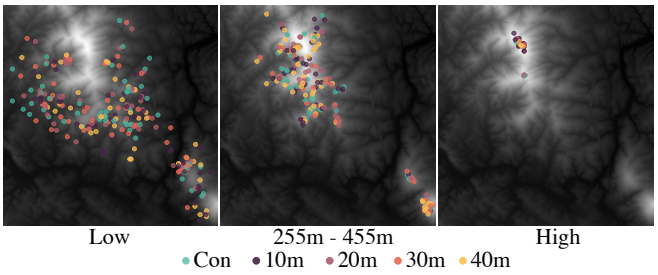


Fig. 12: Distribution of click locations for low, middle, and high elevation groups, for the Steepest a 355m Task.

These results were revealed by one-way between-subjects ANOVAs computed for each task, examining the effect of binning condition on task completion time. Table 1 includes mean, standard deviation statistics, and Tukey HSD post hoc tests of significant differences accounting for multiple comparisons for each binning technique in all 12 tasks.[2] Figure 13 shows mean time for each task that demonstrated significant differences between binning techniques.

3.3.11 Task Confidence

Four tasks showed significant differences in task confidence between binning techniques. However, no consistent relationship between binning and confidence was observed (See Table 2 and Figure 1 in the supplementary material). Further, overall confidence ratings were relatively consistent between tasks. These results were revealed by one-way between-subjects ANOVAs computed for each task, examining the effect of binning condition on task confidence ranking. These results suggest that confidence ratings were not a very sensitive measure of differences in encoding conditions and future work might use other subjective ratings such as preference or ease of task.

3.3.12 Discussion of Results

Significant differences in accuracy of decision making, response speed, and response confidence were found between continuous and binned encodings. The results were task-specific, but overall, our findings suggest that the continuous encoding increased speed but rarely accuracy in many tasks. The largest differences among binning conditions were found in the Highest Point and Lookup tasks, suggesting that the continuous encoding technique was the least effective in those situations. Our findings suggest that the choice of binning versus continuous encoding of 2D scalar fields is important to consider as an influence on user task performance.

More specifically, we found that for the Locate or Explore and Identify Tasks (Section 3.3.2), the binning technique influenced task performance. For example, those with the continuous encoding had worse performance than those with the 40m binning in the Highest Point task, which may have been due to a pop-out effect that the 40m binning produced (See Figure 14 for comparisons). The other Locate or Explore and Identify tasks varied based on binning interval.

Responses for all tasks that utilized a steepness judgment (e.g., Explore and Identify task (Section 3.3.4), Explore and Compare tasks (Section 3.3.6) and the Combination task (Section 3.3.9)) appear to be highly influenced by prior assumptions about the relationship between elevation peaks and slopes and variable gradient magnitudes as described in Section 3.3.4.

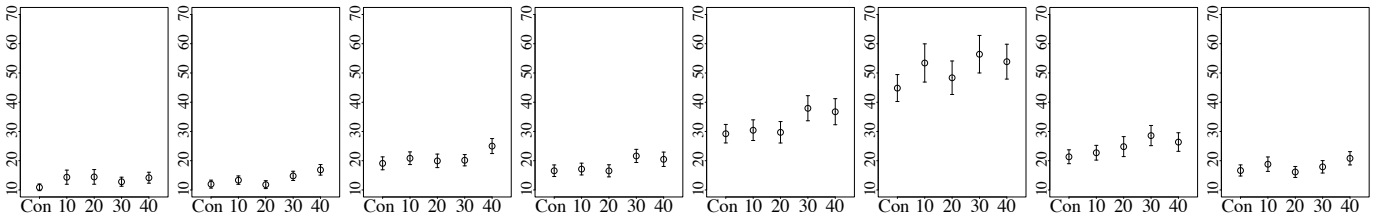
In the Lookup and Identify Task (Section 3.3.5), those with the continuous encoding also performed the worst. This effect may have been due to the difficulty of locating values on the continuous legend. Work by Brewer [7, 17, 6] suggests that identifying relevant values

[2]For the first task, binning technique significantly affected time. However, post hoc comparisons did not reveal significant differences between the conditions at the $p < .05$ level when accounting for multiple comparisons. To understand the main effect of binning, contrast codes were generated which compared the continuous encoding to the binned encodings, which is shown in the first row of the table.

Table 1: Task completion time, with ** = $p < .000$, * = $p < .05$. Outliers refers to the percent of outliers two standard deviations beyond the mean that were removed.

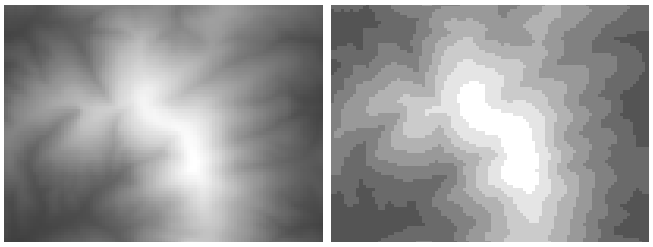
Task	Continuous		Binned Encodings				F	p	outliers	figure
	μ	σ	μ	σ						
1	**10.90***	5.75	13.95*	10.54			7.79	**.005****	1.4	a

Task	Continuous		10m		20m		30m		40m		F	p	outliers	figure
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ				
2	**11.99***	6.83	**13.35***	7.14	**11.81***	6.71	**15.74***	10.20	**16.87***	9.0	7.679	**.000****	2	b
3	**19.10***	10.96	20.84	10.76	**19.95***	11.30	**20.16***	9.48	**25.00***	12.50	7.679	**.000****	4.4	c
4	**16.56***	9.88	**17.14***	10.00	**16.54***	10.25	**21.63***	11.17	20.49	12.24	4.88	**.000****	2.4	d
5	**29.24***	15.85	30.43	17.34	**29.71***	22.10	**37.94***	21.16	36.75	22.10	4.652	**.001****	3.2	e
6	**44.83***	23.32	53.42	32.57	48.37	28.41	**56.41***	31.31	53.87	29.72	2.522	**.04***	2.4	f
8	20.35	15.25	19.16	11.78	20.17	12.43	23.17	16.00	22.82	15.52	1.462	.213	2.6	
9	31.71	14.81	32.01	16.86	29.56	16.02	36.32	18.64	33.86	18.72	1.5	.201	2.8	
10	**21.34***	11.81	**22.72***	12.63	24.81	16.86	**28.60***	17.04	26.38	15.33	3.629	**.006****	3.6	g
11	**16.66***	9.57	18.81	12.18	16.15	9.36	17.90	10.34	**20.82***	11.09	2.977	**.019***	3.4	h
12	23.72	15.44	27.93	17.60	25.20	15.59	26.87	16.60	30.03	19.07	2.011	.09	3.9	



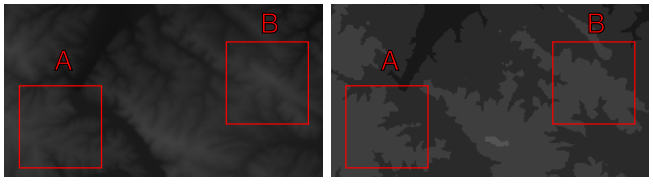
(a) Highest Point (b) Highest in Region (c) Locate 275m (d) Compare Heights (e) Rank Heights (f) Steepest (g) Avg. Height (h) Avg. Compare

Fig. 13: Mean Time (s). Error bars (95% CI)



(a) Continuous (b) 40m

Fig. 14: An example of the highest portions of the continuous and 40m encodings.



(a) Continuous (b) 40 m

Fig. 15: An example of portions of the continuous and 40m encodings for the Average Height Comparison Task.

on a choropleth map legend with many bins (over 12) is difficult. Indiscernibility of specific values on the continuous legend may have contributed to the inefficacy of this encoding.

For the Browse tasks (Sections 3.3.7–3.3.8), task difficulty likely led to large errors when participants were asked to report the average elevation of a region. However, participants were reasonably effective at comparing the average height of two regions and those with the 40m binning were the least effective at this comparison. This may have been due to the 40m binning's lack of relevant encodings for making average height comparisons (See Figure 15 for comparisons).

The response time data was the most consistent across tasks. Overall, these results show that the continuous encoding was significantly faster than some of the binning techniques for many tasks. This finding is not consistent with work that suggests that binning data facilitates task completion speed [16, 51, 10, 9].

While significant differences in task confidence were observed, confidence ratings did not consistently vary across binning conditions.

4 Discussion and Future Work

While the expressiveness principle implies that encoding continuous data continuously is advantageous, some domains have strong conventions that advocate for a discretized encoding of continuous data. The existing literature lacks systematic evaluation of the effects of binning 2D scalar fields on human performance. Our goal in this work was to directly compare performance using several different binning intervals and a continuous encoding, across a range of tasks for 2D scalar fields. Consistent with our initial hypothesis, we found that accuracy in performance was superior using binning in some tasks. Although in other tasks we did not show differences between binning and continuous representations, there were no tasks that showed a clear advantage for continuous encoding. This result stands in contrast to one of the dominant principles in visualization and is a finding that should be pursued further in both controlled laboratory studies and real-world contexts.

Our findings on task completion time are also notable. In contrast to assertions in cartography, the current study reveals that a binned encoding does not facilitate faster task completion times, compared to a continuous encoding. Rather, we found that performance on the continuous encoding condition was the fastest, across a majority of the tasks, even when accuracy was worse. One possible explanation for this speed-accuracy trade-off is that tasks which require more effort can activate a slower and more deliberative cognitive process [19]. Judgments made using fast cognitive processes are more error prone and subject to cognitive biases than judgments made using slower more contemplative cognitive processes [21]. The continuous encoding may have produced quick responses that did not allow for careful inspection of the visualization or thoughtful consideration of the task.

Additionally, one of the aims of this study was to create a methodology for testing increasingly complex and ecologically valid tasks, which are traditionally difficult to assess and as such rarely utilized.

Table 2: Task confidence, with ** = $p < .000$, * = $p < .05$

Task	Continuous		10m		20m		30m		40m		F	p	figure
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ			
1	5.68	1.35	5.95	1.28	5.76	1.31	5.87	1.21	6.12	1.36	1.66	.15	
2	5.01	1.34	4.87	1.45	4.92	1.49	5.17	1.27	4.90	1.52	0.73	.57	
3	4.12	1.55	**3.78***	1.42	4.06	1.49	**4.62***	1.32	**4.59***	1.63	5.91	**.000****	a
4	**4.99***	1.45	**4.89***	1.43	**4.85***	1.49	4.57	1.67	**3.98***	1.63	7.05	**.000****	b
5	4.52	1.49	4.42	1.47	4.25	1.58	4.50	1.47	4.04	1.68	1.72	.14	
6	4.36	1.73	4.49	1.74	4.67	1.52	5.01	1.48	4.55	1.71	2.22	.06	
7	3.89	1.57	3.93	1.49	4.00	1.47	4.11	1.56	4.02	1.72	0.29	.88	
8	**4.85***	1.47	5.26	1.58	**5.39***	1.57	**5.53***	1.41	5.33	1.67	2.77	**.02***	c
9	4.61	1.46	4.53	1.52	4.66	1.53	4.86	1.34	4.97	1.58	1.5	.20	
10	3.83	1.37	3.59	1.27	3.61	1.43	3.98	1.28	3.93	1.44	1.75	.13	
11	4.66	1.54	4.78	1.26	4.67	1.49	**5.17***	1.39	**4.40***	1.63	3.59	**.006****	d
12	3.85	1.47	3.81	1.68	3.67	1.56	4.11	1.57	4.23	1.71	2.00	.09	

We chose to increase task complexity by having participants make judgments concerning features and summary statistics in addition to points and values, increasing the number of comparisons, and requiring multistep judgments. Multistep visual comparison tasks are highly dependent on spatial working memory (SWM), and differences in SWM may account for the inconsistent performance observed. In particular, we suspect that ranking tasks may not be appropriate for spatial decision making as this type of task may be confounded by limits on participants' SWM. Further work is needed to disambiguate the demands on SWM from the increased task complexity. Another issue is found in the multistep judgment tasks (i.e., Steepest at 355m task), where it is unclear how to disambiguate participants' strategies from the influence of the binning technique. The inconsistent findings of the Steepest at 355m task may have resulted from participants prioritizing clicking on the correct elevation or on a steepest location rather than a combination of the two. Each strategy would produce vastly different responses, which may have influenced the uninterpretable variability in the current findings.

As discussed briefly in Sections 3.3.4 and 3.3.6, interpretations of the questions that involved steepness were difficult. We believe that this is, in part, because the image gradient we used as the basis for steepness had high spatial frequency information. Gradient approximations are generally known to be sensitive to high frequency information such as noise [12, 14]. While there are a variety of gradient estimation methods we could have pursued, we chose to use the Sobel operator, which includes a smoothing component [12, 14]. Smoothing obfuscates more of the data and the influence of smoothing on decision making requires further study. Given that our results also left open the possibility of other factors influencing the results, such as the conflation of steepness with height, it was not clear what amount of smoothing would have been appropriate. While our findings are specific to 2D scalar fields embedded in $\mathbb{R}^2$, we suspect that our results are also relevant for 2D scalar fields embedded in $\mathbb{R}^3$, such as those in bioengineering [25]. This study provides motivation for future work aimed at assessing the influence of individual components of spatial data, such as spatial frequency, and determining appropriate smoothing for specific conditions.

As we were interested in performance across a spectrum of tasks, we did not attempt to compare an optimized binning technique. There are a variety of data-informed optimization strategies for binning within the visualization literature [46, 40, 11]. Additionally, standardized binning approaches for choropleth maps from cartography [42] might also be generalizable. However, it is not clear how these various optimizations interact with tasks. Given that existing research suggests that binning strategies that are highly effective for a particular task are likely inadequate for a larger set of tasks [28], optimizing binning for a variety of tasks remains an open problem. We believe a particularly fruitful area of inquiry would be investigating the effects of interactive interfaces for helping a user customize bins according to evolving task demands. In such a system, to what extent does the *act* of binning, itself, become a vehicle for exploration of the data? What sorts of default settings are desirable? A number of potential avenues for future work exist in this space.

The primary implications of this work are for topographical analysis, geography, cartography and meteorology, suggesting that designers should consider selectively employing the expressiveness principle to promote accuracy or speed. If accuracy is the main objective of the visualization, then binning may be more effective. If speed is of interest, then following the expressiveness principle using continuous encoding may be desirable. In many cases, designers will want to strike a balance between speed and accuracy, requiring a fine tuning of the binning intervals based on task demands and data. Further, the tasks developed in this study should be applicable to any domain using 2D scalar fields. While not tested here, designers should also consider the influence of both legend labels [27] and even coarser bins (e.g., three or five) that are commonly used with choropleth maps [16]. Previously studied labeling effects may interact with very coarse binnings, adding to cognitive load. The possible influences of varying spatial frequencies and topological characteristics in the underlying data also warrant future study.

A secondary implication for this work involves the use of rainbow colormaps. Rainbow colormaps have been marked as a poor choice for encoding data in part because they are not perceptually continuous, resulting in perceptual bands that hide details [4, 39, 30] (cf.[6]). Even though large portions of the visualization community reject this technique, rainbow colormaps are still widely used, even as defaults. Our results call into question the extent to which the natural discretization produced by rainbow colormaps should truly be considered a deficiency. Possible future directions for work include comparing continuous to binned rainbow colormaps to understand if the appeal of this approach comes from a general preference for discretization.

In sum, contrary to the expressiveness principle, no cases were found in which a continuous encoding of 2D scalar field data was advantageous for task accuracy, and for some tasks, specific binned encodings facilitated accuracy. Furthermore, we found support for the counterintuitive finding that decisions with binned encoding were slower than those made with continuous encoding. Given this apparent trade-off between accuracy and speed of performance, it is important to consider specific task-goals when deciding whether to adhere to the expressiveness principle. In addition, we developed and tested several different tasks that ranged in complexity and relevance to real-world problems and offer insights on types of tasks and data characteristics to avoid. Our approach and findings contribute to a greater understanding of the use of binning for 2D scalar data and lead to many possibilities for future research.

Acknowledgments

This work was funded by NSF grant IIS-1212806.

References

[1] N. Andrienko and G. Andrienko. *Exploratory analysis of spatial and temporal data: a systematic approach*. Springer Science & Business Media, 2006.

[2] L. D. Bergman, B. E. Rogowitz, and L. A. Treinish. A rule-based tool for assisting colormap selection. In *Visualization, 1995. Visualization '95. Proceedings., IEEE Conference on*, pages 118–125, 444, Oct 1995.

[3] B. Bied Sperling and T. S. Tullis. Are you a better mouser or trackballer?: A comparison of cursor-positioning performance. *ACM SIGCHI Bulletin*, 19(3):77–81, 1988.

[4] D. Borland and R. M. Taylor II. Rainbow color map (still) considered harmful. *IEEE computer graphics and applications*, (2):14–17, 2007.

[5] M. Brehmer and T. Munzner. A multi-level typology of abstract visualization tasks. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2376–2385, 2013.

[6] C. A. Brewer, A. M. MacEachren, L. W. Pickle, and D. Herrmann. Mapping mortality: Evaluating color schemes for choropleth maps. *Annals of the Association of American Geographers*, 87(3):411–438, 1997.

[7] C. A. Brewer and E. Press. *Designing better maps: A guide for GIS users*. Taylor & Francis, 2006.

[8] E. Danovaro, L. De Floriani, L. Papaleo, and M. Vitali. A multi-resolution representation for terrain morphology. In *International Conference on Geographic Information Science*, pages 33–46. Springer, 2006.

[9] M. W. Dobson. Choropleth maps without class intervals?: a comment. *Geographical Analysis*, 5(4):358–360, 1973.

[10] M. W. Dobson. Perception of continuously shaded maps. *Annals of the Association of American Geographers*, 70(1):106–107, 1980.

[11] M. Eisemann, G. Albuquerque, and M. Magnor. Data driven color mapping. In *Proc. EuroVA: International Workshop on Visual Analytics 2011*, pages 1–4, Bergen, Norway, May 2011.

[12] K. Engel, M. Hadwiger, J. M. Kniss, C. Rezk-Salama, and D. Weiskopf. *Real-time Volume Graphics*. A. K. Peters, Ltd., Natick, MA, USA, 2006.

[13] I. Fujishiro, T. Azuma, and Y. Takeshima. Automating transfer function design for comprehensible volume rendering based on 3d field topology analysis. In *Visualization '99. Proceedings*, pages 467–563, Oct 1999.

[14] R. C. Gonzalez and R. E. Woods. *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2006.

[15] E. Guilbert. Multi-level representation of terrain features on a contour map. *Geoinformatica*, 17(2):301–324, 2013.

[16] M. Harrower and C. A. Brewer. ColorBrewer. org: an online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003.

[17] M. Harrower and C. A. Brewer. Colorbrewer. org: an online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003.

[18] C. G. Healey. Choosing effective colours for data visualization. In *Proceedings of the 7th Conference on Visualization '96*, VIS '96, pages 263–ff., Los Alamitos, CA, USA, 1996. IEEE Computer Society Press.

[19] J. Hullman, E. Adar, and P. Shah. Benefitting infovis with visual difficulties. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2213–2222, 2011.

[20] E. Imhof. *Cartographic relief presentation*. ESRI, Inc., 2007.

[21] D. Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[22] G. Kindlmann and J. W. Durkin. Semi-automatic generation of transfer functions for direct volume rendering. In *Proceedings of the 1998 IEEE Symposium on Volume Visualization*, VVS '98, pages 79–86, New York, NY, USA, 1998. ACM.

[23] S. Liu, D. Maljovec, B. Wang, P.-T. Bremer, and V. Pascucci. Visualizing high-dimensional data: Advances in the past decade. In *Proc. Eurographics Conf. Visualization*, pages 20151115–127, 2015.

[24] J. Mackinlay. Automating the design of graphical presentations of relational information. *Acm Transactions On Graphics (Tog)*, 5(2):110–141, 1986.

[25] R. MacLeod, C. Johnson, and M. Matheson. Visualizing bioelectric fields. 13(4):10–12, 1993.

[26] M. Mahy, L. Van Eycken, and A. Oosterlinck. Evaluation of uniform color spaces developed after the adoption of cielab and cieluv. *Color Research & Application Application*, 19(2):105–121, 1994.

[27] J. Matejka, M. Glueck, T. Grossman, and G. Fitzmaurice. The effect of visual appearance on the performance of continuous sliders and visual analogue scales. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5421–5432. ACM, 2016.

[28] S. Mittelstädt, D. Jäckle, F. Stoffel, and D. A. Keim. ColorCAT: Guided Design of Colormaps for Combined Analysis Tasks. In E. Bertini, J. Kennedy, and E. Puppo, editors, *Eurographics Conference on Visualization (EuroVis) - Short Papers*. The Eurographics Association, 2015.

[29] S. Mittelstädt and D. A. Keim. Efficient contrast effect compensation with personalized perception models. *Computer Graphics Forum*, 34(3):211–220, 2015.

[30] K. Moreland. *Advances in Visual Computing: 5th International Symposium, ISVC 2009, Las Vegas, NV, USA, November 30-December 2, 2009. Proceedings, Part II*, chapter Diverging Color Maps for Scientific Visualization, pages 92–103. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[31] N. Moroney, M. D. Fairchild, R. W. G. Hunt, C. Li, M. R. Luo, and T. Newman. The ciecam02 color appearance model. In *IS&T/SID 10th Color Imaging Conference*, pages 23–27, 2002.

[32] J.-C. Muller. Perception of continuously shaded maps. *Annals of the Association of American Geographers*, 69(2):240–249, 1979.

[33] J.-C. Muller. Visual comparison of continuously shaded maps. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 17(1):40–52, 1980.

[34] J.-C. Muller and J. L. Honsaker. Choropleth map production by facsimile. *The Cartographic Journal*, 15(1):14–19, 1978.

[35] T. Munzner. *Visualization Analysis and Design*. CRC Press, 2014.

[36] P. S. Quinan and M. Meyer. Visually comparing weather features in forecasts. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):389–398, Jan 2016.

[37] P. S. Quinan and M. Meyer. Visually comparing weather features in forecasts. *Visualization and Computer Graphics, IEEE Transactions on*, 22(1):389–398, 2016.

[38] P. L. Rheingans. Task-based color scale design. *Proc. SPIE*, 3905:35–43, 2000.

[39] B. E. Rogowitz and L. A. Treinish. Data visualization: the end of the rainbow. *IEEE Spectrum*, 35(12):52–59, Dec 1998.

[40] P. Schulze-Wollgast, C. Tominski, and H. Schumann. Enhancing visual exploration by appropriate color coding. In *Proceedings of the International Conference in Central Europe on Graphics, Visualization and Computer Vision (WSCG'05)*, pages 203–10, Bergen, Norway, 2005.

[41] Y. Shinagawa and T. L. Kunii. Constructing a reeb graph automatically from cross sections. *IEEE Computer Graphics and Applications*, 11(6):44–51, 1991.

[42] T.-K. Shyy, I. Azeezullah, I. Azeezullah, R. J. Stimson, and A. T. Murray. 14 classification for visualizing data: integrating multiple attributes and space for choropleth display. *Handbook of Research Methods and Applications in Spatially Integrated Social Science*, page 265, 2014.

[43] T. A. Slocum. *Thematic cartography and geovisualization*. Prentice hall, 2009.

[44] M. Stone. *Field Guide to Digital Color*. A K Peters / CRC Press, 7 2003.

[45] D. A. Szafir, M. Stone, and M. Gleicher. Adapting color difference for design. In *Color and Imaging Conference*, pages 228–233. Society for Imaging Science and Technology, 2014.

[46] D. Thompson, J. Bennett, C. Seshadhri, and A. Pinar. A provably-robust sampling method for generating colormaps of large data. In *Large-Scale Data Analysis and Visualization (LDAV), 2013 IEEE Symposium on*, pages 77–84, Oct 2013.

[47] W. R. Tobler. Choropleth maps without class intervals? *Geographical analysis*, 5(3):262–265, 1973.

[48] C. Tominski, G. Fuchs, and H. Schumann. Task-driven color coding. In *Information Visualisation, 2008. IV '08. 12th International Conference*, pages 373–380, July 2008.

[49] L. Wang, X. Chen, S. Li, and X. Cai. *Neural Information Processing: 13th International Conference, ICONIP 2006, Hong Kong, China, October 3-6, 2006. Proceedings, Part II*, chapter General Adaptive Transfer Functions Design for Volume Rendering by Using Neural Networks, pages 661–670. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[50] C. Ware. Color sequences for univariate maps: theory, experiments and principles. *IEEE Computer Graphics and Applications*, 8(5):41–49, Sept 1988.

[51] D. Watson. *Contouring: a guide to the analysis and display of spatial data*. Elsevier, 2013.

[52] M. Wijffelaars, R. Vliegen, J. J. Van Wijk, and E.-J. Van Der Linden. Generating color palettes using intuitive parameters. *Computer Graphics Forum*, 27(3):743–750, 2008.

[53] A. Zeileis, K. Hornik, and P. Murrell. Escaping rgbland: Selecting colors for statistical graphics. *Computational Statistics & Data Analysis*, 53(9):3259 – 3270, 2009.

[54] L. Zhou and C. Hansen. A survey of colormaps in visualization. *IEEE Transactions on Visualization and Computer Graphics*, PP(99):1–1, 2015.