Spatialization Design: Comparing Points and Landscapes

Melanie Tory, David W. Sprague, Fuqu Wu, Wing Yan So, and Tamara Munzner, *Member, IEEE*

Abstract— Spatializations represent non-spatial data using a spatial layout similar to a map. We present an experiment comparing different visual representations of spatialized data, to determine which representations are best for a non-trivial search and point estimation task. Primarily, we compare point-based displays to 2D and 3D information landscapes. We also compare a colour (hue) scale to a grey (lightness) scale. For the task we studied, point-based spatializations were far superior to landscapes, and 2D landscapes were superior to 3D landscapes. Little or no benefit was found for redundantly encoding data using colour or greyscale combined with landscape height. 3D landscapes with no colour scale (height-only) were particularly slow and inaccurate. A colour scale was found to be better than a greyscale for all display types, but a greyscale was helpful compared to height-only. These results suggest that point-based spatializations should be chosen over landscape representations, at least for tasks involving only point data itself rather than derived information about the data space.

Index Terms— Spatialization, Information Landscape, User Study, Numerosity, 3D, 2D, Colour, Greyscale, Surface, Points

1 Introduction

People are familiar with spatial concepts such as distance and height as part of their everyday life. Spatialization [8] takes advantage of this knowledge by using a spatial metaphor to display abstract, non-spatial data. Spatializations have been used for many applications, but most commonly for visualizing document collections [4][25]. For example, news articles from an archival database can be arranged in a map-like layout to illustrate themes [25]. The spatial arrangement of items is typically created through a dimensionality reduction technique such as multi-dimensional scaling (MDS) or principle component analysis. Such layouts allow the user to infer the similarity of items by observing their spatial distance on the display (the distance-similarity metaphor). Research suggests that spatializations promote understanding of high dimensional relationships, by enabling users to easily see similarities, clusters, and outliers [1][2][11].

We address the challenge of how to visually represent spatializations once the two-dimensional (2D) layout of points has been determined. Many different visual representations of the points are possible. Each point can be represented as a dot, with the dots coloured by some property of the data. For example, the price or fuel economy of a specific model in a car database could be represented using point colour. Alternately, a surface can be fitted to the 2D arrangement of points to produce a visualization resembling a landscape. A classic example of a three-dimensional (3D) information landscape is Themescape [25], where coloured 'hills' represent common themes in a collection of documents. In Themescape, the height of each 'hill' represents the density of points at that location. However, height could also be used to represent a variable in the data itself, such as the time a document was created. Colour can be added to 3D landscapes; colour and height can represent two different variables or can redundantly encode the same variable. Keeping the points on a plane and then fitting a surface to the points can be used to create 2D visualizations resembling topographic maps.

We categorize spatializations into two groups based on the graphical mark used to represent data:

- **Points**: Spatializations that show only points.
- **Information Landscapes**: Spatializations where a surface has been fitted to the set of underlying points. Points may be shown on the surface. We refer to these simply as landscapes.

We also identify the following characteristics of spatializations that may affect their usability:

- **Dimensionality (2D or 3D)**: In 3D landscapes, the points are first arranged in 2D space using dimensionality reduction, and then the landscape height is used to encode some component of the data. In 3D point visualizations, the height dimension may be used to encode a component of the data similar to 3D landscapes, or the dimensionality may be directly reduced to 3D instead of 2D.
- **Colouring method**: Points and landscapes may be coloured using greyscale or colour scale mappings, or may have no colour mapping applied.

It is currently unclear which of these many options is best suited to different visual analysis tasks.

1.1 Empirical Knowledge about Graphical Encoding and Numerosity

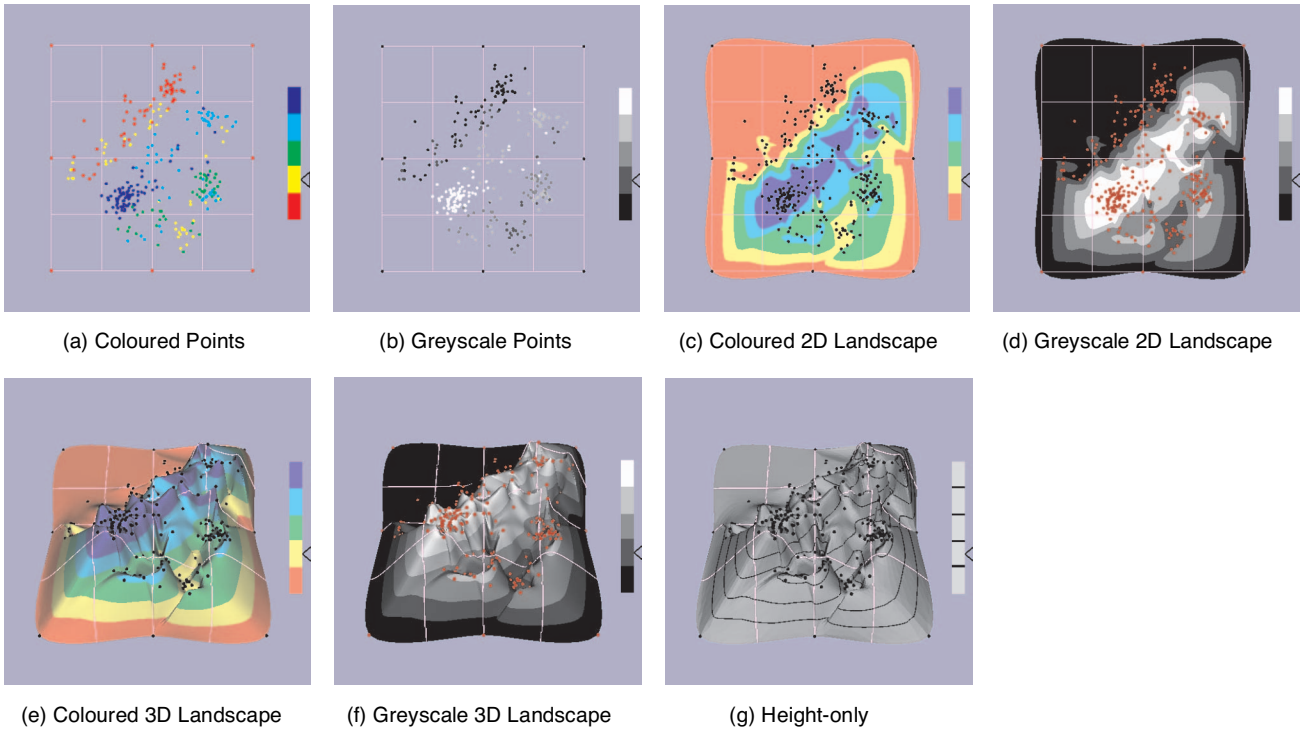
Redundantly encoding data using two or more retinal variables has been shown to improve perceptual salience and task performance for some retinal variables and tasks [3][18]. However, it remains unclear whether 3D landscapes that redundantly encode data using height with either greyscale or colour will have similar benefits. 3D displays often suffer from occlusion and clutter, and can be difficult to interact with. For example, Cockburn *et al.* [6] found that 3D worlds were more difficult to perceive and analyze. For these reasons, redundantly encoding information using height plus colour or greyscale may actually be detrimental.

We also investigate the use of colouring methods. We focus on data analysis tasks involving continuous data, such as a car's price in an automobile dataset. Intuitively, one might choose a lightness or saturation scale for this data. However, only approximately four lightness levels can be easily distinguished as compared with five to ten colours [22]. For tasks where users need to focus on particular ranges of values, having easily distinguishable colours is important [19]. Hence, colour may be a better choice for this type of task. Furthermore, in 3D landscapes the surface must be shaded to enable people to accurately judge surface shape and height [22]. Using the lightness channel to additionally represent data may therefore be confusing. By contrast, hue-based encodings may be correctly perceived irrespective of shading and lighting, due to colour consistency. These results suggest that a hue scale may be better than a lightness scale for 3D landscapes.

We focus on a task where users need to estimate the number of points (numerosity) of a specified colour within a spatial area.

- *Melanie Tory, David W. Sprague, Fuqu Wu, and Wing Yan So are with the University of Victoria, E-Mail: {mtory, dsprague, fuquwu}@cs.uvic.ca and vivian@uvic.ca.*
- *Tamara Munzner is with the University of British Columbia, E-Mail: tmm@cs.ubc.ca.*

Manuscript received 31 March 2007; accepted 1 August 2007; posted online 27 October 2007. Published 14 September 2007.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.



(a) Coloured Points (b) Greyscale Points (c) Coloured 2D Landscape (d) Greyscale 2D Landscape

(e) Coloured 3D Landscape (f) Greyscale 3D Landscape (g) Height-only

Fig. 1 Point-based displays and information landscapes used in our experiment. All displays show the same data.

Experimental evidence [7] suggests that clustering, density, and dot regularity will all affect perceived numerosity. In addition, Healey *et al.* [12] have shown experimentally that visually estimating the proportion of points of a particular colour is a pre-attentive process. However, Healey *et al.*'s task required users to compare the number of points of different colours, whereas in our experiment the other colours are distractors. In addition, their work considered only point-colour, not surface colour. Hence, although surface colour may enable users to rapidly identify the region of interest, it remains unclear whether estimating the numerosity in a coloured region can be done in parallel with identifying the target region.

1.2 Common Visualization Assertions

Despite the considerable theoretical foundations establishing the efficacy of points and the challenges associated with 3D displays, such as navigation and occlusion, many researchers have proposed using landscapes (e.g., [1][4][8][13][25]). Proponents of landscapes suggest several reasons why landscapes may be beneficial. Using a landscape metaphor may facilitate pattern recognition and spatial reasoning [21]. Information landscapes may also avoid problems with some dimensions obfuscating others, may simplify the amount of data to be presented, and may display data in a way that is optimal for information processing [4]. Several authors suggest that the landscape metaphor is easily understood by most users and facilitates hierarchical clustering of data items [1][11][13][14]. For 3D spatializations, the landscape surface may also provide a constant reference to reduce disorientation when navigating in 3D space [4] and to aid depth perception [17]. Most of these assertions about the benefits of information landscapes have not been tested empirically.

Another common assertion is that the rainbow colour scale is a poor choice for representing quantitative data. Although categorical data is commonly represented using hue, it is generally believed that ordinal and continuous data should be represented using lightness or saturation scales. However, as described in the previous section, there are reasons to believe that a rainbow colour map may be a very good choice for some tasks and for 3D landscapes.

Similarly, redundantly encoding data using two or more retinal variables is generally considered better than encoding data using a single retinal variable. Although there is experimental evidence supporting this claim for some retinal variables, it may not be true for all representations. It is particularly unclear whether redundantly encoding data using height and colour or height and greyscale will provide any benefit compared to a single encoding.

1.3 Our Experiment

We designed an experimental study to empirically test the assertions described above. We compared a subset of spatializations that are most likely to be effective data display methods. Our primary objective was to compare 2D points to both 2D and 3D landscapes. We also compared a greyscale to a colour scale, and considered a 3D landscape with no colour mapping. We do not consider points with no colour or greyscale mapping because they are unable encode data outside of horizontal spatial location. We also do not consider 3D points and 2D landscapes without colour or greyscale because they have previously been shown to be ineffective [8]. Fig. 1 illustrates the seven types of displays compared in our study. We designed our study to answer the following questions:

1. Is performance better when data is represented using landscapes or points?
2. For representing data using landscapes, which single retinal variable supports the best performance: colour, greyscale, or height?
3. In landscapes, what is the benefit or drawback of redundantly encoding data using colour and height, or greyscale and height?
4. Does a colour scale or greyscale support better performance for tasks where users need to focus on a specific range of quantitative data?

As our benchmark task, we chose a non-trivial search task involving spatial areas and estimation. To design this task, we first used a spatialization to explore experimental data from one of our prior user studies, and examined the low-level mental operations we performed. A frequently occurring low-level operation was to

identify areas of the display containing many points in a given value range. For example, we asked ourselves questions such as, "Where in the display are the trials with the fastest response times?", "Does this area also contain low error levels?", and, "Does this area largely correspond to one experimental condition?". Our benchmark task represents this style of data analysis. Participants estimated which area of the display contained the most points in a given value range, as described in section 3.2. Our task was designed to satisfy the following criteria:

- Non-trivial task complexity. The task involved multiple points, and was more complex than a simple lookup, comparison, or outlier search task.
- Response time of less than 1 minute. To test many factors (several display types and data complexities), the task needed to be simple enough to repeat many times.
- Tasks involving both spatial layouts (multidimensional similarity) and dimension specific values. We wanted to include all the information that is typically available at any given time when analyzing information landscape data.

We should note that spatial areas in our task were explicitly defined by a grid; this characteristic was artificially imposed to enable our software to easily collect and evaluate answers. Users would define spatial areas more dynamically and imprecisely in real analysis.

2 Related Work

Several experiments have considered point spatializations. Montello *et al.* [16] demonstrated that people naturally equate distance with similarity, but visual illusions and emergent features (e.g. belonging to a common cluster) can override the distance-similarity metaphor. Hornbæk and Frøkjær [14] report that information retrieval using a 2D point display was not more effective than using text summaries, but subjects preferred the spatialization. Chalmers [4] reported that a 3D point spatialization had usability problems due to occlusion and scene complexity. Similarly, Newby [17] reports that although people were able to use a 3D point spatialization, they had difficulty judging distances between items and became disoriented when navigating through 3D space. Westerman and Cribbin [23] directly compared 2D and 3D point spatializations, where dimensionality was reduced to two and three dimensions respectively, for a simple search task. They showed that participants performed better with 2D points than with 3D points and that the *goodness-of-fit* of the layout algorithm also affected performance.

Experiments with landscape spatializations have been more limited. In a series of studies, Fabrikant [8] experimentally compared spatializations similar to those used in our experiment. She demonstrated that people could intuitively understand the distance-similarity metaphor, landscape representations of non-spatial data, and the relationship between 3D landscapes and the underlying data points. She found that 2D landscapes were usually faster than 3D landscapes for simple distance and density judgement tasks, but that accuracy of some tasks was higher with 3D than 2D. For point-based displays, she found that only 2D points were effective; 3D point-based displays were very difficult to understand. Fabrikant's experiments provide a useful starting point for investigating spatialization design. However, her studies were limited to landscapes where height and colour represented point density rather than some other variable. They did not consider different colouring methods and they examined only simple tasks. Our experiment builds on Fabrikant's thesis work by addressing these limitations.

Other variations of spatialized displays have also been considered empirically. Fabrikant *et al.* have investigated perceptual issues in spatializations where there are graphical links between objects [10] and in spatializations that have discrete regions, analogous to choropleth maps [9]. Cribbin and Chen [5] demonstrated that visually connecting the most similar nodes in a 2D point spatialization could improve performance at some tasks. These types of spatializations are not considered in our study.

3 Method

3.1 Experimental Design

We used a 7 display × 5 target level × 3 data complexity within-subjects design. Order of all conditions was randomized. Measures were response time, correctness, and questionnaire responses about various aspects of the task.

3.2 Task

Participants were asked to identify which spatial area contained the most points of a specified target value range. These value ranges were usually specified by a unique colour or greyscale value and will hereafter be referred to as *target levels*. Fig. 2 gives two examples. For each trial, the target level was identified by the position of a black unfilled triangle next to the legend bar at the right. The target level and data values of points in the display changed for each trial. To specify an answer, participants selected one of 16 screen areas by placing the mouse cursor over the area and pressing the space bar on the keyboard with their other hand. Possible answers (areas) were delineated using a pink grid as shown in Fig. 2.

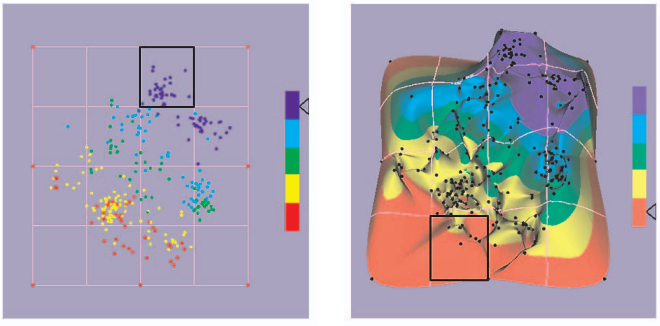


Fig. 2 Example trials from our experiment. Target levels are 5 (blue) in the left example and 1 (red) in the right example. Correct answers are highlighted with black outlines.

3.3 Participants

Eighteen participants (6 male and 12 female) were recruited from a university student population. All participants reported normal or corrected-to-normal vision and passed the Ishihara Test [15] for common colour vision deficiencies. The average age was 23. Most participants had experience with 3D video games and computer-based maps. Two had done substantial computer graphics programming.

3.4 Apparatus

3.4.1 Computer and Software

Experimental software was written in Java, using the Visualization Toolkit [20] to create visual stimuli. Experiments were run using an AMD Athlon 64 bit dual core PC running at 1.99 GHz, with 1GB of RAM, a Radeon x1600 series video card, and WindowsXP. The display was a 17" LCD at 1024 × 768 resolution. Participants interacted with the software using a standard mouse and keyboard. Only the spacebar key on the keyboard was used.

The graphic display could be rotated, translated, and zoomed by depressing the left, middle, and right mouse buttons respectively and then dragging. Buttons beneath the graphic display enabled the user to set pre-defined viewpoints: a bird's-eye view (as in Fig. 2 left) and an oblique view (as in Fig. 2 right). These buttons were labelled with images showing the different viewpoints and the labels changed on each trial to match the current display type.

3.4.2 Stimuli

Each participant completed 105 experimental trials, representing all combinations of 7 displays × 5 target levels × 3 data complexities. 105 questions were created, each with a particular target level, set of data values, and one best answer. Target levels and data for each question were selected at random. Each participant answered all 105 questions, but different displays were randomly assigned to each question for each person. Trials were presented in random order.

Stimuli were created from a multidimensional environmental dataset containing 5000 rows and 290 columns. We used real rather than synthetic data to ensure the spatializations were realistic. Participants were not told anything about the nature of the data. The values in the table were normalized to range from 0 to 1, to ensure consistent heights in the 3D landscapes. The entire dataset was laid out in 2D space using the MDSteer system for multidimensional scaling [24] and then the 300 lowest stress points were selected for use in the experiment. The number of points was chosen to make the task challenging but still manageable in a short time.

Positions of the points in 2D space remained constant throughout all trials. For each trial, one of the 290 columns in the dataset was selected, and the value for each point using this dimension was displayed graphically using:

- Point colour or greyscale (for point conditions),
- Surface colour or greyscale (for all landscape conditions except height-only), and / or
- Surface height (for 3D landscape conditions).

Low values appeared as valleys in the 3D displays and as colours at the bottom of the legend. Each of the 290 columns in the data set was assigned a complexity value of low, medium, or high according to the visual complexity of the display when that data column was used. Complexity was judged by visual inspection. Examples of the three complexity levels are shown in Fig. 3.

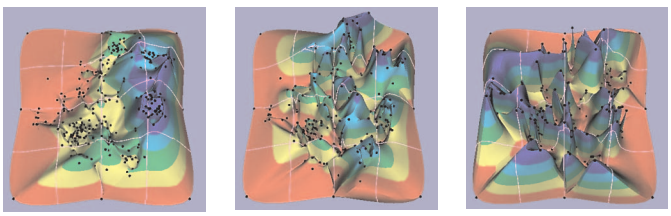


Fig. 3 Examples of the three data complexity levels. Occlusion increases with higher complexity levels for 3D displays.

For point-based displays, points were directly displayed and were assigned a colour or greyscale level based on their data value. For the other displays, a graphic surface was created. For 3D displays, points were first moved to a height representing their data value. Points were then triangulated to create a surface and the surface was smoothed to make it appear like continuous terrain. Contours were extracted at the interpolated boundaries between target value ranges. Contours were shown using colour bands, greyscale bands, or contour lines depending on landscape type. Points were coloured red in greyscale landscapes and black in the other landscapes. Point colour was chosen to maximize visibility.

Greyscale levels ranged from black to white, with three intermediate shades of grey. Colour scales used a five level hue scale ranging from red to blue. For coloured points, fully saturated colours were used to maximize point visibility. However, for coloured 2D and 3D landscapes the saturation was reduced to 0.75 because large patches of highly saturated colour were hard on the eyes. The saturation difference can be seen in the colour legends of Fig. 2.

Point-based displays and 2D landscapes were initially rendered from a bird's-eye viewpoint and 3D landscapes were initially rendered from an oblique viewpoint. These standard viewpoints are shown in Fig. 2. Participants could change the viewpoint interactively through rotation / translation / zoom or by pressing the standard viewpoint buttons.

3.5 Procedure

Participants were first screened for colour blindness using the Ishihara test [15]. They then filled in a questionnaire about demographic information and computer experience, and reviewed a printed tutorial that described the task, procedure, and instructions on how to use the test system. The tutorial intentionally did not describe the data represented or how to interpret the displays. Participants were shown all seven types of displays one by one, and at the same time, they were asked to experiment with the corresponding display in our test system. The experimenter ensured participants understood and tried out all operations in the test system, repeating examples if necessary. Participants then started the test. Each test session consisted of one practice session with 14 trials followed by three real sessions with 35 trials each. Trial order, including assignment to sessions, was random. Participants could ask questions during the practice session, but not during the real sessions. Five-minute breaks were enforced between real sessions. After the testing was completed, participants rated the different displays on three criteria and were encouraged to make comments about the task and displays.

For each trial, participants were asked to be reasonably confident in their answer, but to estimate the answer rather than count the points. Participants were told that their responses and response times were being recorded by the test system. Each trial began when the participant pressed an onscreen *Ready* button using the mouse, and ended when the participant pressed the spacebar (with the mouse cursor placed over a screen area to indicate the answer).

3.6 Hypotheses

Our primary hypotheses, based on our four major research questions, were as follows:

H1: Landscapes would be slower than points.

H2: When data was encoded by a single variable in a landscape, colour would be the best, followed by greyscale. Height would be the worst.

H3: Redundantly encoding data using height plus either colour or greyscale would be less effective than single encoding using colour or greyscale, respectively. In particular, redundant encoding would increase response time with no increase in accuracy.

H4: Colour would be faster and more accurate than greyscale for all display types.

We also had the following secondary hypotheses:

H5a: Low and high target levels would be faster and more accurate than middle target levels.

H5b: Differences between target levels would be greatest for greyscale conditions and 3D conditions.

H6a: Higher data complexity would increase errors and response time.

H6b: Differences between data complexity levels would be greatest for the 3D conditions.

4 Results

Results were analyzed statistically using repeated measures analysis of variance (ANOVA) followed by Bonferroni-corrected pairwise comparisons. In approximately 1% of trials, landscape colouring did not render correctly; data for these trials were replaced with the mean values for the series. One participant did not completely fill in the questionnaire; this participant's rating data was discarded. Data was transformed using a natural log or square root when this improved the fit to a normal curve. When Mauchly's Test of Sphericity indicated it was necessary, we used the Huynh-Feldt correction. Factors in the analysis were display (7 display types), target level (5 levels), and landscape complexity (3 levels). Target level 1 was the bottom and level 5 was the top. We used paired-samples t-tests to

compare 3D to 2D (for coloured and greyscale landscapes), and to compare the two colour scales (greyscale and colour). Below we describe the most interesting results. ANOVA results are given in tables with F, p, degrees-of-freedom (DOF), and η^2_p values. Post-hoc tests are indicated by p-values only. Many post-hoc test results have been aggregated with only the largest p-value given.

4.1 Response Time

Mean response time for each display / target level combination is shown in Fig. 4. Overall statistical results are shown in Table 1.

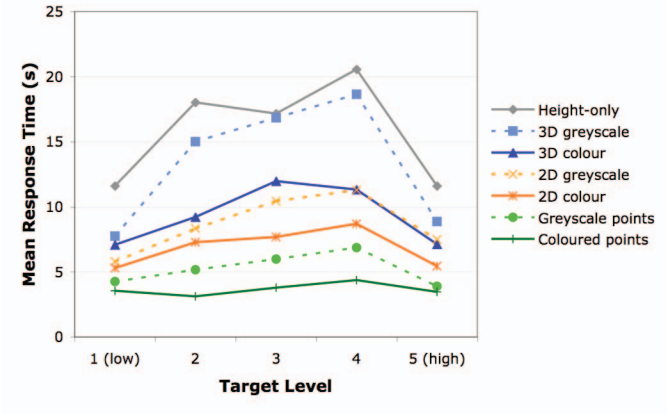


Fig. 4 Mean reponse time for displays with each target level.

Table 1. Statistical Results for Response Time

Effect	*F*	*DOF*	*p*	η^2_p
Display	123.6	6, 102	< 0.001	0.88
Target Level	77.8	4, 68	< 0.001	0.82
Complexity	108.7	2, 34	< 0.001	0.87
Display * Target Level	2.7	19.6, 332.5	< 0.001	0.14
Complexity * Target Level	10.2	8, 136	< 0.001	0.38
Complexity * Display * Target Level	1.5	28.5, 485.5	0.043	0.08

All displays were significantly different from all others ($p \leq 0.04$), except that 3D colour was not significantly different from 2D greyscale. Coloured points were the fastest, followed by greyscale points and 2D coloured landscapes. Height-only was the slowest. Colour was significantly faster than greyscale (t=7.2, df=17, p<0.001). When considering only the coloured and greyscale landscapes (blue and orange lines in Fig. 4), 2D landscapes were faster than 3D (t=7.5, df=17, p<0.001).

Response time was faster for the highest and lowest value ranges. All target levels were significantly different from all others ($p \leq 0.05$), except 1 and 5 (top and bottom levels) and 3 and 4 (middle levels). However, this trend was not consistent across all display types. As shown in Fig. 4, the differences between target levels were largest for 3D displays, smaller for 2D displays and greyscale points, and very small for coloured points. Response time also increased with increasing data complexity, with all complexity levels significantly different from each other overall ($p < 0.001$, data not shown). This trend did not vary significantly with display type, but we note that complexity appeared to have a smaller effect on coloured points than on the other displays.

4.2 Errors

We considered the following measures of trial error (where AN = number of points of the target level in the participant's answer, BN = number of points of the target level in the best answer, and TN = total number of points of the target level in the whole display):

- Binary: whether the best answer was selected or not
- Percent from best = 100% × |(AN – BN) | / BN
- Difference over total = 100% × |(AN – BN)| / TN

All measures resulted in the same overall trends, providing confidence in their validity. We therefore conducted a statistical analysis only on *percent from best*. Intuitively, this measures the percentage difference (in the number of target points) between the given answer and the correct answer. Mean error (percent from best) for each display / target level combination is shown in Fig. 5, and mean error for each display / data complexity combination is shown in Fig. 6. Overall statistical results are shown in Table 2.

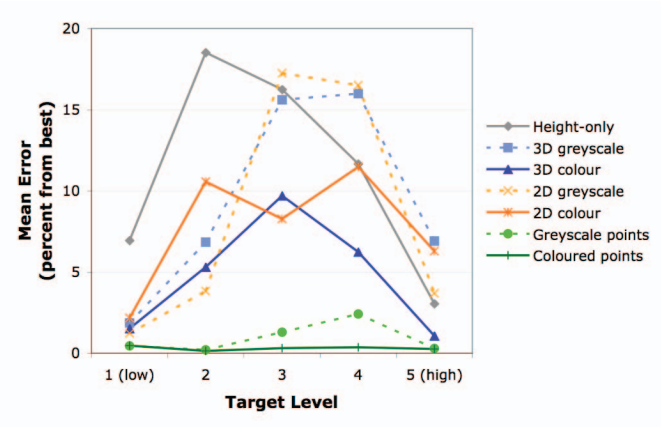


Fig. 5 Mean error level (percent from best) of the displays with each target level.

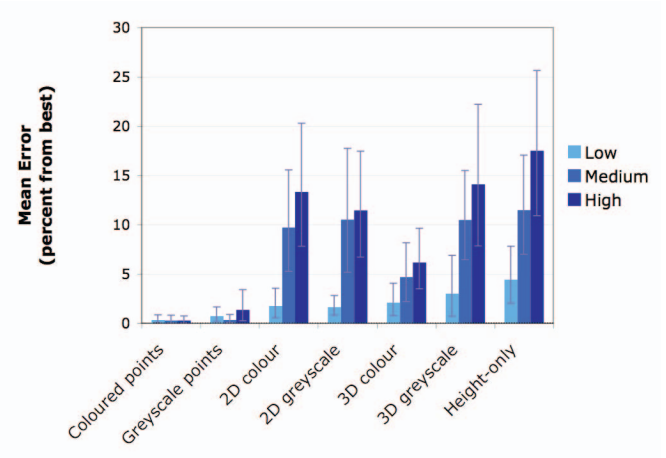


Fig. 6 Mean error level (percent from best) for each display type with low, medium, and high data complexities. Error bars show 95% confidence intervals.

Coloured points and greyscale points had significantly lower error levels overall than the other display conditions ($p \leq 0.004$). Colour had significantly lower error than greyscale (t=2.3, df=17, p=0.035). Error levels were higher for target levels in the middle of the scale (middle value ranges), with all target levels significantly different from each other ($p \leq 0.043$), except that level 1 was not significantly different from level 5, and level 3 was not significantly different from level 4. This trend occurred only for the landscapes (not for points), as shown in Fig. 5. Overall, the error level increased with increasing landscape complexity, as shown by the different shades of blue in Fig. 6. All complexity levels were significantly different ($p \leq 0.022$). Significant differences between complexity levels did not occur for points, but occurred for all landscape-style displays ($p \leq 0.038$). Of the landscapes, complexity had the smallest

effect on 3D colour. Only the lowest and highest complexity levels were significantly different for the 3D colour display.

Table 2. Statistical Results for Error

Effect	*F*	*DOF*	*p*	η^2_p
Display	21.2	4.6, 79	< 0.001	0.56
Target Level	31.3	4, 68	< 0.001	0.65
Complexity	44.7	2, 34	< 0.001	0.72
Display * Complexity	3.8	12, 204	< 0.001	0.18
Display * Target Level	2.5	17.0, 289.1	0.001	0.13
Complexity * Target Level	9.6	7.6, 129	< 0.001	0.36
Complexity * Display * Target Level	1.5	35.1, 596.6	0.027	0.08

4.3 Ratings

Participants were asked to rate the seven display types on 3 criteria using a 5-point Likert scale with 1 being disagree and 5 being agree:
[Distracting] The interface was distracting to complete the task.
[Find Dots] It was easy to find dots of the target level.
[Overall] The interface made it easy to do the task overall.

Average results are shown in Fig. 7. Overall statistical results are shown in Table 3. Responses on the three questions generally agreed. Note that a good display received low values on [Distracting] and high values on [Find Dots] and [Overall].

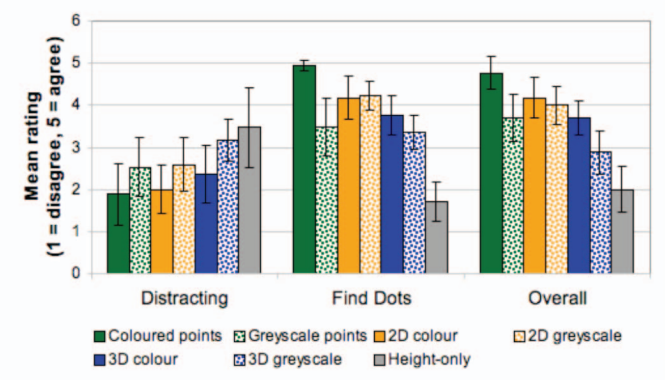


Fig. 7: Mean ratings of the three criteria. Error bars show 95% confidence intervals. Good displays received low values on [Distracting] and high values on [Find Dots] and [Overall].

Table 3. Statistical Results for Ratings

Criterion	*F*	*DOF*	*p*	η^2_p
Distracting	3.5	3.0, 47.5	0.022	0.18
Find Dots	23.0	4.5, 71.4	< 0.001	0.59
Overall	16.9	4.3, 68.8	< 0.001	0.51

The height-only display was least preferred, and was rated significantly worse than all other displays on [Find Dots] ($p \leq 0.011$) and worse than all other displays except 3D greyscale on [Overall] ($p \leq 0.005$). Coloured points was most preferred and was rated significantly better than all other displays except 2D colour on both [Find Dots] and [Overall] ($p \leq 0.032$). The remaining display types were not rated significantly different.

5 Discussion

Our results supported our primary hypotheses, but not some of our secondary hypotheses, as shown in Table 4. Coloured points supported the best performance. 2D coloured landscapes performed second best and may therefore be suitable to some applications. 3D landscapes that redundantly encoded the data using colour and height were slower than 2D landscapes using colour alone, with no difference in accuracy. Height-only was least effective. A detailed discussion of specific research questions is given below.

Table 4. Summary of Our Hypotheses and Whether Each Was Supported (Y = Yes, N = No)

H1: Landscapes would be slower than points.	Y
H2: When data was encoded by a single variable in a landscape, colour would be the best, followed by greyscale. Height would be the worst.	Y
H3: Redundantly encoding data using height plus either colour or greyscale would be less effective than single encoding using colour or greyscale, respectively. In particular, redundant encoding would increase response time with no increase in accuracy.	Y
H4: Colour would be faster and more accurate than greyscale for all display types.	Y
H5a: Low and high target levels would be faster and more accurate than middle target levels.	Y
H5b: Differences between target levels would be greatest for greyscale conditions and 3D conditions.	N
H6a: Higher data complexity would increase errors and response time.	Y
H6b: Differences between data complexity levels would be greatest for the 3D conditions.	N

5.1 Discussion of the Research Questions

Is performance better when data is represented using landscapes or points? Coloured points were the fastest, most accurate, and most highly rated display, as predicted by **H1**. In addition, they showed minimal degradation in performance with increasing landscape complexity and had the most consistent performance with different target levels. Coloured points were the most effective and greyscale points performed second best. Task performance with landscapes was consistently less accurate (by 4-10%) and required more time (from 1.9× – 4.2× as long) compared to points. The magnitude of this difference depended on the landscape's dimensionality and colouring, as discussed below. In summary, points were substantially better than landscapes for our numerosity task.

For representing data using landscapes, which single retinal variable supports the best performance: colour, greyscale, or height? Representing data values using landscape height alone was very ineffective, supporting **H2**. Height alone was by far the slowest, most error-prone, and least well-liked display. When compared with coloured points (the best performing display), height-only required on average 4.2 times as long and had an average error level of 10.4% (as compared to 0.3%). Participants' comments suggested that height-only landscapes did not provide enough helpful information for users to determine which levels the dots were on, or which level a particular contour represented. This was particularly true since many 'peaks' in the landscape did not reach the highest level and many 'valleys' did not reach the lowest level.

2D landscapes were also less effective than points, but this difference was far less substantial. 2D landscapes required on average 2.1 times as long as coloured points, had an average error of 7.1%, and were rated second best (with coloured points being first). 2D coloured landscapes were significantly faster than 2D greyscale landscapes, supporting **H2**.

In landscapes, what is the benefit or drawback of redundantly encoding data using colour and height, or greyscale and height? **H3** was supported: 2D landscapes were significantly faster than 3D landscapes with no significant difference in accuracy. Thus, redundantly encoding data using height plus greyscale or height plus colour was detrimental compared to greyscale or colour alone.

Cognitive complexity of 3D displays may have contributed to the difference [23], but the most likely cause is occlusion. In 3D landscapes, the user needed to rotate the display to alleviate occlusion. This need to rotate was partially caused by the default oblique viewpoint. However, using a bird's eye viewpoint would have made the default 3D landscape appear very similar to the 2D landscape and seemed like an unnatural initial view. These problems appear to outweigh any potential benefit of redundant encoding.

Five participants developed an unexpected strategy for dealing with occlusion in the 3D landscapes. These participants sometimes rotated the landscape approximately 180 degrees around the horizontal axis, so they could view the landscape upside down. Turning the 'hills' into 'valleys' in this way sometimes made it easier to see details, particularly in complex landscapes. This example illustrates the awkward workarounds that participants find for dealing with non-optimal display conditions.

Does a colour scale or greyscale support better performance for tasks where users need to focus on a specific range of quantitative data? A hue scale was better overall for our numerosity task, supporting **H4**. Colour scale displays were significantly faster and more accurate than greyscale. Our task required users to focus on a particular data range. This is easiest when points in the range are pre-attentively distinct from other points, which is most easily done using colour. These results emphasize that choosing an optimal colour scale depends more on the user's task than on the data type.

Preference differences between colour and greyscale were greatest for point-based displays, moderate for 3D landscapes, and small for 2D landscapes. Participants' comments suggest that the greyscale levels were harder to distinguish in small points than in the large grey areas of the 2D landscapes. For example, one participant commented, "It was hard to distinguish dark grey [dots] from black [dots]". Greyscales were also problematic in 3D landscapes because grey levels were confounded with surface shading. Surface shading could not simply be eliminated because prior research has indicated it is important for understanding 3D landscapes [8].

How do target level and data complexity affect performance with different spatializations? Overall, performance was higher with low and high target levels compared to middle levels, supporting **H5a**. We expect this occurred because the user could simply look for peaks or valleys. Performance also degraded with increasing data complexity, supporting **H6a**. However, our data did not support hypotheses **H5b** and **H6b**. **H5b** predicted that differences between target levels would be greatest for greyscale conditions (where distinguishing middle greyscale values may be more difficult than black or white) and 3D conditions (where peaks appear as physical hills). Instead, target level affected all landscapes, not just 3D and greyscale ones. Similarly, **H6b** predicted that differences between data complexity levels would be greatest for 3D conditions because of occlusion. Performance with 3D landscapes did vary with this factor, as expected, but performance with 2D landscapes also degraded significantly. Consequently our predictions that 2D would be more tolerant than 3D to changes in target level and data complexity were not supported. Notably, only point displays seemed unaffected by changes in the target level and data complexity. High and low target levels do not have emergent features (i.e. peaks and valleys) for point displays. The tolerance of point displays to changes in data complexity suggests that points could be processed in parallel when estimating numerosity within spatial regions.

5.2 Colour Design

Our result that points were very effective agrees with Healey *et al.*'s study [12], which demonstrated that estimating numerosity of coloured points is pre-attentive. Our experiment further suggests that estimating numerosity may not be pre-attentive with background colouring. Several possible explanations could account for this difference. We initially thought that large patches of colour might be easier to perceive than small points of colour. Although large patches may attract the eye, this may be counter-intuitive for the task since people might naturally search for large patches, when the size provides no information about the number of points. In addition, identifying points of the target range and estimating their number may require two separate steps when the surface is coloured, but may be done in parallel when the points are coloured. Additionally, data values were ambiguous in landscapes when the points were on or near a colour boundary, whereas point colour was unambiguous.

Although points supported superior performance, we found a qualitative advantage for landscapes when a greyscale was used. In particular, participants found it easier to judge grey levels of large patches on the landscape surface than on small points.

We considered many colour scale options for this study. Since the data covers a continuous range, we initially planned to compare a continuous, isoluminant colour scale to a continuous greyscale. However, our pilot analysis revealed that most queries involved focusing on points within a range of values, a task that is easier with a segmented, non-isoluminant colour map [22]. We might have also segmented the height variable, but felt that choppy "terraced" landscapes would look unnatural. As a result, we essentially showed categorical data using colour but continuous data using height. This mismatch did not appear to confuse participants, but may be an issue to explore in future studies, along with continuous colour maps. We also note that in real practice, we would expect users to set up custom segmented colour maps with thresholds appropriate for their task, rather than our five equally-spaced bins, perhaps by alternating between continuous and categorical colouring.

5.3 Choice of Benchmark Task

Choice of benchmark tasks is often a difficult issue affecting the generality of the results. We modelled our task on a mental operation where people identify areas containing many points within a value range. Among other criteria, we chose this task because it occurred frequently during a pilot analysis of how people use spatialization tools. This frequency, and the fact that the task must be performed mentally and often several times in quick succession, suggested that this task would substantially affect many higher-level activities.

However, we recognize that tasks of a different nature may be better suited to landscapes. Specifically, landscapes can show a continuous distribution of values across the data space, not just the values at discrete data points. For this reason, landscapes may be beneficial for tasks involving this type of derived information. Examples might include understanding the 'shape' of a data space, such as its 'peakiness' or slope, remembering this shape, and identifying points in areas with different shapes. In addition, surface colour could be useful for judging the size of a spatial area containing points with similar values, or the number of large areas with a given value range. Landscapes may also provide better spatial context, such as whether points are in a peak or valley.

Our benchmark task may be construed as having a bias toward point spatializations. We contend that the mental operations used in our task would be performed frequently in many types of high-level analysis. For this type of operation, point spatializations have the inherent advantage of representing data values directly. Some landscapes had a small performance cost, and might be worth considering when the high-level task involves analyzing the shape of the data space. Further research is needed to validate whether landscapes provide better performance when analyzing derived information, and to determine when these benefits are worthwhile, considering the performance cost for some basic mental operations.

5.4 Additional Limitations

We did not tell participants what the data represented and did not explain how to interpret the displays, to assess whether the displays were understood intuitively. Consequently, results could change for complicated tasks or specific user populations, particularly users experienced with visual data analysis. Our experiment was also

limited to a relatively small number of points. Landscapes may be worth considering with more points, particularly when the number exceeds the display resolution so that points overlap. Future experiments should examine these factors. In addition, our experiment only considered landscapes derived from data values themselves. Landscapes derived from the density of points, such as Themescape [25], may be interpreted differently. This is particularly true since point density would be redundantly encoded by the landscape and the point positions, whereas in our experiment data values were sometimes singly encoded by the landscape.

Information landscapes also have the potential to display at least two variables simultaneously by encoding one using height and one using colour or greyscale, a condition that was not tested in our study. Our results suggest that encoding two variables in this way should be done with great caution since error levels were high when data was encoded using landscape surface properties, especially height. Representing two variables in a landscape display – one using height and one using surface colour – can be expected to be even more perceptually complex, and may allow only approximate judgements about each variable. It is also not clear whether unmatched height and surface colour values would be confusing.

6 Conclusion and Future Work

We presented an experiment comparing different methods of visually representing spatializations, to determine which representations are better for a non-trivial search task. Our results demonstrated that point-based displays were substantially more effective than landscapes. Landscape-style representations varied in their effectiveness. Performance was better with 2D than 3D landscapes, so little or no benefit was found for redundantly encoding data using colour and height or greyscale and height. Height-only 3D landscapes were by far the least effective. A colour scale was found to be better than a greyscale for all display types, but a greyscale was helpful compared to no colour scale at all (height-only).

We plan to conduct further studies that compare different spatialization designs for additional abstract tasks as well as higher-level tasks. We also plan to explore methods of representing multiple variables in a spatialization. The current experiment provides a useful starting point for these future investigations, and should provide valuable information to inform spatialization design.

Acknowledgements

This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] U. Bischoff, N. Diakopoulos, F. Lösch, and Y. Zhou. ThemeExplorer: A Tool for Understanding the History of the Field of Information Visualization. *InfoVisFun 2004*, October 2004.

[2] U. Brandes and T. Willhalm. Visualization of bibliographic networks with a reshaped landscape metaphor. In *Proceedings of the Symposium on Data Visualization (VisSym '02)*, pages 159-164, 2002.

[3] C.M. Carswell and C.D. Wickens. Information integration and the object display: and interaction of task demands and display superiority. *Ergonomics*, 30(3): 511-527, 1987.

[4] M. Chalmers. Using a landscape metaphor to represent a corpus of documents. In *Proceedings of the European Conference on Spatial Information Theory (COSIT '93)*, volume 716 of Lecture Notes in Computer Science, pages 377-390, September 1993.

[5] T. Cribbin and C. Chen. Visual-spatial exploration of thematic spaces: a comparative study of three visualisation models. In *Electronic Imaging 2001: Visual Data Exploration and Analysis VIII*, pages 199--209, January 2001.

[6] A. Cockburn and B. McKenzie. Evaluating the effectiveness of spatial memory in 2D and 3D physical and virtual environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '02),* pages 203-210, April 2002.

[7] F.H. Durgin (1995). Texture density adaptation and the perceived numerosity and distribution of texture. In *Journal of Experimental Psychology: Human Perception and Performance*, volume 21, pages 149-169, 1995.

[8] S. I. Fabrikant. *Spatial metaphors for browsing large data archives*. Ph.D. thesis. University of Colorado-Boulder, Boulder, CO, 2000.

[9] S.I. Fabrikant, D.R. Montello, and D.M. Mark. The distance-similarity metaphor in region-display spatializations. *IEEE Computer Graphics and Applications*, 26(4): 34-44, July/August 2006.

[10] S.I. Fabrikant, D.R. Montello, M. Ruocco, and R.S. Middleton. The distance-similarity metaphor in network-display spatializations. *Cartography and Geographic Information Science*, 31(4): 237-252, 2004.

[11] M. Granitzer, W. Kienreich, V. Sabol, K. Andrews, and W. Klieber. Evaluating a system for interactive exploration of large, hierarchically structured document repositories. In *Proceedings of the IEEE Symposium on information Visualization (InfoVis'04)*, pages 127-134, October 2004.

[12] C.G. Healey, K.S. Booth, and J.T. Enns. High-speed visual estimation using preattentive processing. ACM Transactions on Computer-Human Interaction, 3(2): 107-135, 1996.

[13] E. Hetzler and A. Turner. Analysis experiences using information visualization. *IEEE Computer Graphics and Applications,* 24(5): 22-26, September 2004.

[14] K. Hornbæk and E. Frøkjær. Do thematic maps improve information retrieval? In the *Seventh IFIP Conference on Human-Computer Interaction (INTERACT'99)*, pages 179-186, August 1999.

[15] S. Ishihara. Ishihara's test for colour blindness. Found at *http://www.toledo-bend.com/colorblind/Ishihara.html*, 2007.

[16] D.R. Montello, S.I. Fabrikant, M. Ruocco, and R.S. Middleton. Testing the First Law of Cognitive Geography on Point-Display Spatializations. In *Proceedings of the Conference on Spatial Information Theory (COSIT '03)*, pages 316-331, September 2003.

[17] G. B. Newby. Empirical study of a 3D visualization for information retrieval tasks. *Journal of Intelligent Information Systems,* 18(1):31-53, January 2002.

[18] L. Nowell, R. Schulman, and D. Hix. Graphical Encoding for Information Visualization: An Empirical Study. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis'02)*, pages 43-50, October 2002.

[19] B.E. Rogowitz and L.A. Treinish. How not to lie with visualization. *Computers in Physics*, 10(3): 268-274, May/June 1996.

[20] W. Schroeder, K. Martin, and W. Lorensen. *The Visualization Toolkit*, Prentice Hall PTR, second edition, 1998.

[21] R. Spence. *Information Visualization*, Addison Wesley Publishers, first edition, 2000.

[22] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers, second edition, 2004.

[23] S.J. Westerman and T. Cribbin. Mapping semantic information in virtual space: dimensions, variance and individual differences. *International Journal of Human Computer Studies,* 53(5): 765-787, November 2000.

[24] M. Williams and T. Munzner. Steerable, progressive, multidimensional scaling. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis'04)*, pages 57-64, October 2004.

[25] J. A. Wise, J.J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis*'95), pages 51-58, October 1995.