Visualizing Uncertainty with Chromatic Aberration

In recent years an increasing array of research are being conducted by researchers in the field of uncertainty visualization that attempt to determine the impact of representations on users’ perception and evaluate its effectiveness in decision making. Uncertainties are often an integral part of data and by nature model predictions also contain significant amounts of uncertain information. A prominent example of uncertainty, COVID-19 is a respiratory infectious disease caused by novel coronavirus. Due to its unprecedented challenges over time and frequent changes of strains, scientists and researchers are investigating the available data to discover the patterns in different demographic areas and examine the effect of vaccinations against different variants. In this study, we explore a novel idea for a visualization to present predictive model uncertainties using Chromatic Aberration (CA). We first utilized existing machine learning models to generate predictive results using Covid-19 pandemic data and calculated the corresponding model uncertainties for the most impacted countries with respect to number of new-cases, new-deaths, and new-vaccination for different countries. We then visualized the data itself and its associated uncertainties with an artificially spatially separated channels of red, green, and blue color components. This chromatic aberration representation has been evaluated in a comparative user study. From quantitative analysis it is observed that user can identify targets in CA method more accurately than the state-of-the-art VSUP approach. In addition, their speed of target identification was significantly faster in CA as compared to the VSUP method. But their preference between the two does not vary significantly.

**1 Introduction:**Uncertainty visualization is an ongoing area of research but a topic that many practitioners avoid due to the additional complexity that it experiences. There are various studies conducted for uncertainty representations, for example: textual representation such as captions or tooltips [24], graphical representations such as glyphs [16, 25], custom color palettes such as VSUP [19], bivariate choropleth maps [22] and texture patterns [17]. But as far we know, no uncertainty representation has made use of Chromatic Aberration (CA). To accomplish the purpose, we had to go through the steps such as collect relevant data from reputable sources, generate uncertainty data from model predictions which is accomplished by feeding collected data into machine learning models by doing some required preprocessing. Uncertainties are then calculated from the resultant forecasts [4], visualize the uncertainty along with data using CA, as well as competing existing methods in comparable manner, conduct a controlled human-computer interaction experiment to evaluate the effectiveness of the new visual representation, and explain experimental results with numerical analysis and draw conclusions.

In modern world, data is an essential part of human life, and all data has certain amount of uncertainty either in known or unknown form. Since the uncertainties in data can be originated from different phases or sources, it’s important to analyze and measure the amount uncertainty in the data. That’s why we intend to generate uncertainty data from an authentic way. Considering that fact in mind, the study procedure is subdivided into three major phases i. Generate time series forecasted data from COVID-19 data using four machine learning predictive models, ii. Calculate corresponding uncertainties for different countries and visualize uncertainties in terms of Chromatic Aberration (CA) in a graphical presentation, and iii. Conduct user studies to evaluate user perceptions and applicability with commonly used visualizations.

CA is a color distortion or alteration that is sometimes seen on high contrast edges of objects in photographs. Since different colors of light refract to different angles upon traveling through materials with refractive indices [6], the resulting images may appear to be distorted [7]. Since CA is an image quality problem, most of the research focusing CA are conducted to fix the problem and improve image quality thereby. On the other hand, uncertainty is the problem of data quality and relevant research are conducted mostly regarding reducing it to improve data certainty. But existing research has been conducted to visualize uncertainty with traditional approaches such as glyphs [16] or VSUPs [19]. Since our goal is neither to improve image quality nor data quality, we borrowed the term CA for our research to represent uncertainty as a novel approach in the field of visualization.

**2. Related Work**Uncertainty is an unavoidable part of data and due to complexity people usually avoid it to represent in visualization work. The term uncertainty can synonymously refer to some other integral property of data such as data quality, error in data, accuracy in prediction, etc. Prediction data generation is becoming way popular day by day by using different machine learning predictive models such as neural network model for instance MLP, LSTM and GRU in [43] for performance evaluation, ARIMA, PROPHET [1, 2, 4] for time series analysis, XGBoost machine learning algorithm for cholera epidemics predictions linked with weather variable [3]. A decision-supporting tool [5] for medical centers and health-care services has been proposed for influenza prediction for Belgium and liver disease predictive analysis in [6]. All these works have been conducted without concerning uncertainty whereas Botchen et al. [17] focuses on uncertainty that occurs during data acquisition and demonstrates texture-based techniques to visualize uncertainty in time-dependent 2D flow fields. Being error in data inherent, improper or eliminated presentations in visualizations can mislead decision making for data analysts, hence Kamal et al. [18] discusses state-of-the-art approaches such as the quantiﬁcation method to uncertainty visualization, along with the concept of uncertainty and its sources. Bonneau et al. [12] explores uncertainty in the visualization domain by comparing different results (such as a weather forecast) to detect similarities or differences with comparative visualization technique e.g., comparison of border areas than individual pixels.

A statement on the position of uncertainty visualization today is explained in Griethe et al. [13] that defines the basic concept of uncertainty and discusses sources and necessary measures. Through a human-subjects experiment Deitrick et al. [14] integrates visualization with the study of uncertainty, Lundstrom et al. [15] proposes animation methods to convey uncertainty in the rendering and

Pang et al. [16] introduces a wide variety of new uncertainty visualization methods like adding glyphs, adding geometry, modifying attributes, modifying geometry, animation and Wittenbrink et al. [25] designs uncertainty vector glyphs, Finger et al. [23] introduces blended icons, Kay et al. [24] presents a novel mobile interface for visualizing uncertainty for transit predictions. Being uncertainty a multi-faceted concept, the researchers of [20, 22] investigate how data uncertainty visualized in maps might influence the process of spatial decision-making and determine the accuracy of estimates. The authors Greis et al. [21] published a web-based game on Facebook and compared four representations that communicate different amounts of uncertainty information to the user and compared the results to show how uncertainty regularly influences decision making in our daily life.

From a vision perspective, chromatic aberration leads to various forms of color imperfections or tempering in the image. Koh et. al. [7] presented a user study to observe the effect on users’ judgment with Lateral Chromatic Aberration (LCA) for Chart Reading in Information Visualization on display devices and suggested some guidelines (e.g., using eyeglasses) for designers to avoid such issues and [9, 10] proposed image warping technique to resolve such problems. Colour is widely used in information visualisation to deliver different types of information such as extreme values, patterns and attribute values and color size illusion is one of the common problems in this domain. Yoo et. al. [8] aim to identify appropriate interventions and propose design guidelines of image warping to reduce the illusion effect when size judgement is critical. Real cameras have an aperture through which light falls on an image plane to register the image, but diffraction is an issue on this process, so Lee et al. [11] present a novel rendering system for defocus blur and lens effects by approximating optical aberrations.

**2.1 VSUP**

Both uncertainty visualization and understanding uncertainty are complex and critical tasks. One of the most common approaches of uncertainty visualisation is to encode data values and uncertainty values independently, using two visual variables in a bivariate map. These resulting bivariate maps can be difficult to interpret, and the discriminability of marks can be reduced due to the interference between visual channels. To address this issue, Correl et al. [19] introduces Value-Suppressing Uncertainty Palettes (VSUPs) and we highlight this study as it is the comparator approach of our user study. We see that VSUPs allocates smaller ranges of the visual channel to data when uncertainty is high and larger ranges when uncertainty is low. This allocation of visual variables promotes patterns of decision-making that make efficient use of uncertainty information, discouraging comparison of values in unreliable regions of the data, and promoting comparison in regions of high certainty. In traditional bivariate maps Figure 2.1(left), outputs for each combination of value and uncertainty might be represented as a 2D square whereas VSUP approaches it as arcs mapping larger number of outputs for smaller and smaller sets of outputs for higher uncertainty.

But the main limitation of that research is they filter out higher uncertainty values by grouping them altogether which suppresses the values for decision making when uncertainties are high. Due to this higher uncertainty elimination aspect the designers need to carefully consider if this representation is suitable and desirable for certain systems. Another limitation is, since both uncertainty and value are represented by a single color, the perceptual non-separability of color channels are well-known, and which requires the concept of a limited “budget” of distinguishable marks. To achieve the limited budget criteria, it necessitates one to quantize the data. Due to the data quantization, uncertainty visualisation for continuous (or all discrete) values are not possible with limited color budgets.

**3. Uncertainty Data Generation**Good quality data is an important part in data visualization research. Without having an authentic dataset research cannot be conducted properly and cannot succeed in the long run. That’s why we chosen COVID dataset from WHO authorized data repository and applied data preparation strategy such as cleaning, validating, and consolidating raw data before feeding into machine learning models. Model are built by exploring and tuning hyperparameters [4, 6] to get better prediction results. We used MLP, CNN, LSTM and ARIMA algorithms to create models and since the model training is time consuming task, we picked top 100 infected countries based on number of new cases to train models. Uncertainties are calculated from the ranges of predicted values for every time step (day) during the specified 200 days of forecasting period. That means we have a lower bound, mean, and upper bound of the predictions for each time step. So, the difference between upper and lower limit is the grey area of model prediction. As the number of new cases are larger and highly varying for different countries and different time, the predictions results were varying by similar ranges. To compensate the larger values and higher variations and to accommodate in display devices, we needed to normalize the data so that the uncertainty data can belong to a single digit.

**3.1 Uncertainty Data Scaling**

ALGORITHM 1: Data scaling

1. Repeat step 2 and 3 for all countries and store them in an array named: all\_countries\_avg\_uncertainties
2. Calculate total uncertainty (county\_total\_uncertainties) of a country.
3. Calculate average uncertainties of a country as follows:

country\_avg\_uncertainty = county\_total\_uncertainties/number\_of\_days

1. Find maximum average uncertainty from all countries:  
   max\_uncertainty = find\_max\_uncertainty(all\_countries\_avg\_uncertainties)
2. country\_uncertainty\_normal = country\_avg\_uncertainty / max\_uncertainty
3. scaling\_factor = 9
4. country\_uncertainty = country\_uncertainty\_normal \* scaling\_factor

We have shown top-level algorithms in the previous sections to generate uncertainty data from the model predictions. Since the uncertainty values are independent of the display, we need to scale the values. The following pseudo code is used to scale the uncertainty data.

From the algorithm, we see all steps are self-descriptive. Up to step 5, we have calculated the normalized form of uncertainty for every country, that means uncertainties are below or equal to 1 for all countries. So, we have set *scaling\_factor = 9* and multiplied it with the country’s normal uncertainty to display those smaller values in display in terms of pixels. For example: the countries that have higher uncertainties will be in normal form such as 1, 0.9, 0.8 and after multiplying with scaling\_factor those will be 9, 8.1, 6.4 and so on. So, in this way, we could allocate 1 pixel per unit of uncertainty and that helped to visualize the default view in a human recognizable manner and easily understandable the higher uncertainty countries.

The following tables 1 and 2 show snapshot of top 5 uncertainty countries and comparison of obtained uncertainties from all four engaged models. We notice that the uncertainties are independent on infection count and which is realistic result based on actual occurrences of infections in different countries. We also observe that MLP gives us better result and hence we use the MLP results as reference all through the discussion of this paper.

|  |  |  |  |
| --- | --- | --- | --- |
| Country | Actual Count | Predicted Count | Uncertainty |
| United States | 14,851,118 | 15,652,300 | 7.00 |
| India | 15,693,425 | 7,409,636 | 4.28 |
| Brazil | 7,219,982 | 7,409,636 | 3.64 |
| Kazakhstan | 667,009 | 651,009 | 2.43 |
| France | 2,088,610 | 2,307,005 | 2.15 |

Table 1: Top 5 uncertainty countries using MLP model

Table 2: Uncertainty comparisons of models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Country | MLP | CNN | LSTM | ARIMA |
| United States | 7.00 | 7.00 | 3.44 | 7 |
| India | 4.28 | 0.61 | 7.00 | 3.52 |
| Brazil | 3.64 | 0.51 | 3.24 | 1.27 |
| Kazakhstan | 2.43 | 0.42 | 0.35 | 0.17 |
| France | 2.15 | 0.31 | 0.81 | 0.56 |

**4 Background Architecture of CA**

We have seen an example of lateral chromatic aberration in Figure 1.2 (Chapter 1) where all lights with different wavelengths do not focus to the same convergent point because lights having shorter wavelength refract more than the lights with longer wavelength. Inspired by that phenomenon, we can consider a circle that represents the predicted number of new cases for a country in a specific day. But since there is associated uncertainty of the prediction, a single circle will not be sufficient to represent bivariate (number of cases and uncertainty) distribution. So, instead of single circle if we use three different circles with separated RGB color channels, we can then apply lateral shifting from the center of the circle by the amount of uncertainty and blend them together and the resultant outcome would be an approximate representation of CA. The following Figure 1 shows such a geometric arrangement on a unit radius circle.

Diagram, engineering drawing

Description automatically generated

Figure 1: Underlying Geometry of CA

To draw a circle representing aberration as per the above explanation if we draw 3 circles, let’s call them 3 chromatic circles, then we can render the technique with the following simple algorithm:

ALGORITHM 2: CA Construction Formula

1. Let’s consider the center of the target circle at (x, y).
2. Radius (radial offset) of the circle is ‘r’ represents uncertainty.
3. Draw the first chromatic circle with color (R, 255, 255) with a shifted location of (x, y + r) where ‘r’ denotes red color channel.
4. Draw the second chromatic circle with color (255, G, 255) with a shifted location of (, ) where ‘G’ denotes green color channel.
5. Draw the third chromatic circle with color (255, 255, B) with a shifted location of (, ) where ‘B’ denotes blue color channel.
6. Set the standalone css ‘mix-blend-mode’ to ‘darken’ to blend all three circles to get the resultant CA appearance.

**4.1 Examples of CA in Shapes**By using the above formula explained in section 3.9, a resultant aberration is presented with the uncertainty for the country India (IND) in Figure-9 below. The center dark-grey area represents the predicted number of new cases, and the color separated edges represent the amount of uncertainty in that prediction. So, each of the items in the figure represent different amount of uncertainties as indicated side by side.

|  |  |
| --- | --- |
| **Shape, circle  Description automatically generated**  10% | **Diagram, venn diagram  Description automatically generated**  90% |
| **Shape, rectangle, square  Description automatically generated**  7% | **Chart  Description automatically generated**  40% |

Figure 2:CA representation on Bubbles (top) and Rectangles (bottom)

**5. User Study Design**Uncertainty visualisation is one of the complex challenges in the visualisation domain and designing a valid user study is also important. The study design usually prepares a particular set of questions that depends on the nature of the research, goal of the research, and the availability of resources, etc. There are various types of user studies such as experimental/interventional studies, descriptive studies, observational studies, within/between subject studies, and so on. We have conducted a within-subject comparative study with the measures of i. Task time ii. Error Rate and iii. Subjective assessments (NASA-TLX, SUS).

According to lam et al. [27] user performance is predominantly measured in terms of objectively measurable metrics such as time and error rate, yet it is also possible to measure subjective performance such as work quality. The commonly used metrics are task completion time and task accuracy. On the other hand, the goal of evaluating user experience is to understand to what extent the visualization supports the intended tasks as seen from the participants’ eyes to provide subjective feedback and to probe for requirements and needs.

**5.1 Study Materials**We have developed a dynamic webpage with the content of study materials to conduct the study session entirely remotely online including color blindness/vision test to ensure participants are capable to decern color and detect objects properly. To maintain similarity with Correll et al. [19], we presented a set of Ishihara plates [26] in the page for blindness test and participants were asked to detect and answer. The webpage was developed with HTML, CSS, JavaScript, and D3.js for frontend and PHP in backend, deployed in the webspace of Dalhousie University, NS, Canada. We used CSS color blending to represent Chromatic Aberration which does not work properly in Google Chrome/Safari but Firefox/Microsoft Edge. Component Questions and Post Session Questionnaire (PSQ) are two categories of the questions. As said PSQ includes subjective assessments (NASA-TLX and SUS) whereas Component Questions includes: i. CA + Bubble ii. VSUP + Bubble iii. CA + Grid and iv. VSUP + Grid. The whole questionnaire is attached in Appendix B.

Chart, bubble chart

Description automatically generated  
Figure 3: CA+Bubble Questionnaire Interface

Chart

Description automatically generated  
Figure 4: VSUP+Bubble Questionnaire Interface

Figure 3 and Figure 4 shows PSQ for CA+Bubble and VSUP+Bubble module questionnaire including the introductory markups for better understanding the reader of the paper. In the study session, the markups were not shown because main researcher clarified the underlying mechanism to the participants and/or answered any question the participants had.

Chart, diagram

Description automatically generated with medium confidence  
Figure 5: CA+Grid Questionnaire Interface

Chart

Description automatically generated

Figure 6: VSUP+Grid Questionnaire Interface

Figure 5 and Figure 6 show PSQ for CA+Grid and VSUP+Grid modules including the introductory markups for better understanding the reader of the paper. In the study session, the markups were not shown because main researcher clarified the underlying mechanism to the participants and/or answered any question the participants had.

Graphical user interface, application

Description automatically generated

Figure 7: PSQ Interface for SUS

Graphical user interface, text, application, email

Description automatically generated

Figure 8: PSQ Interface for NASA-TLX

Figures 7 and 8 show PSQ interfaces for NASA-TLX Work-Load test related questions and System Usability Test (SUS) questions. We separated both UIs in top-down section where CA and VSUP are at top and bottom respectively as caption indicated. Since the underlying mechanism is same for both CA+Bubble and CA+Grid, they are grouped together and placed at the top of the UI in the CA section. Similarly, VSUP+Bubble and VSUP+Grid are grouped together for the same reason and placed at the bottom in VSUP section of the UI. For both CA and VSUP, we have shown the same question. For SUS we have questions to answer in the scale range of 1 to 5 where is for NASA-TLX it is 1 to 21 scales.

**5.2 Recruitment**Since the participants play a central role in any user study, it’s important to find the suitable participants for the study based on the attributed research domain. As we have four components in our study and each component has eight random questions, we decided to hire (4 x 8 = 32) participants to give equal emphasis to every component and questions. We got the participants mostly (97%) from undergraduate/graduate students from Computer Science/ICT background. We conducted color blindness test to ensure they do not suffer with vision problem. Since the study conducted fully online participants also agreed to have good internet connection and own a computer during session.

5.3 **Questionnaire Setup**   
The existing evaluation of uncertainty representation named VSUP uses grid-chart method with a custom color set. We will be comparing VSUP with Chromatic Aberration (CA) using both a grid-chart and bubble-chart. So, the questionnaire arrangement is made with the sections as: A: CA + Bubble, B: CA + Grid, C: VSUP + Bubble and D: VSUP + Grid. To make the comparison fair, we have grouped our uncertainties to 4 levels since VSUP also uses four levels of uncertainties. In our case, we have quantized our CA data and made four equidistant values of uncertainties such as 33, 52, 71, 90 to draw the aberration in both circles and rectangles. In addition, to fill the circles and rectangles of CA, we have used the eight standard VSUP colors to make the evaluation consistent.

**5.3 Counter Balancing**

Each component consists of eight questions. The order of the questions is selected randomly which means no two participants would get the questions in same order and the components themselves were presented to the participant in “Balanced Latin Squares” method of counter balancing mechanism proposed in [28] to give equal emphasis to each component throughout the study and balance the learning effect. Many empirical evaluations of input devices or interaction techniques are comparative. A new device or technique is often compared against alternative devices or techniques. There are two common designs for such experiments *within-subjects design* and *between-subjects design*. We have used the former because we were able to test every component of the system by every participant. If we consider the four components as A, B, C, and D and based on counter balancing method there will be four different component approaching flows(orders) and in this way, 5th and 9th participant will get 1st participant’s order whereas 6th and 10th participants will get 2nd participant’s order and so on. This approach ensures, no two consecutive participants will get the same order of components and ultimately 8 participants among 32 will get each component first.

**5.4 Data Collection and Storing**

We have developed the webpage in such a way that the system can automatically track the status of every answer whether correct or wrong, which options they selected, and which available options were there to be correct answers. It also tracks amount of time it requires for every module individually. SUS and NASA-TLX responses are stored in the same object but with different name so that it can be processed easily as required. That means it keeps a record of every question from starting to end in a JSON object and on completion it stores the json in a file on the server.

**6. Results and Numerical Analysis**

We have obtained several kinds of data from the user study such as: i. Quantitative Questionnaire Results, ii. Time utilization data for each component, iii. SUS data for CA and VSUP, iv. NASA-TLX for CA and VSUP. We analyse all these data (attached in Appendix A) in various ways in the following sections which helps to reach conclusions from the study.

**6.1 Quantitative Questionnaire Results**

As we have four core components, we designed the study content for each component individually and collected the log data for each component separately. As we already stated, there were 8 questions for each component and every question carried 1 point. For answering correctly, the participant gains one point and do no lose any points for wrong answers. So, a participant can gain minimum 0 point and maximum 8 points for a component. That point achievement is considered as the user performance of the study and we are going to analyse the user performance on the basis of ANOVA for four components and t-test for two grouped (CA and VSUP) components.

* + 1. **One-way repeated measures ANOVA**

The result of a one-way ANOVA is considered reliable when some assumptions hold such as: the response variable (the dependent variable) is normally distributed, the samples are independent, the variances of populations are equal.   
  
Since the sample are taken from independent interfaces of the questionnaire, requirement 2 fulfilled. Again, as per Keppel’s ratio rule of thumb [30], if the ratio of the larger variance to the smaller variance is less than 1.5, then we can assume the variances are approximately equal.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Groups** | **N** | **Mean** | **Std. Dev.** | **Variance** | **Std. Error.** |
| CA + Bubble | 32 | 6.2813 | 1.301 | 1.692 | 0.23 |
| CA + Grid | 32 | 5.5938 | 1.2916 | 1.668 | 0.2283 |
| VSUP + Bubble | 32 | 5.6563 | 1.4053 | 1.975 | 0.2127 |
| VSUP + Grid | 32 | 5.1875 | 1.2032 | 1.456 | 0.2127 |

Table 3: ANOVA Data summary

So, from Table 3, we see that variances are equal which conforms condition (3). Since conditions 2 and 3 are met, we need to ensure data is normally distributed. On this purpose, we conducted Shapiro-Wilk Normality Test and obtained results: CA+Bubble (W=0.915, P=0.015), CA+Grid (W=0.932, P=0.045), VSUP+Bubble (W=0.011, P=0.012) and VSUP+Grid (W=0.913, P=0.013) which indicates the distributions of the components are approximately in normal distribution which satisfies requirement (1) and we can conduct an ANOVA test. Additionally, we have also showed boxplot (Figure 4 - left).

To discuss and conclude on results, we need to first define Null (Ho) and Alternative Hypotheses (Ha) as follows (hypotheses will be tested using an F-ratio for a One-Way ANOVA):  
 *Ho:  μ1​ = μ2 ​= μ3 ​= μ4​ (Performances were equal for all components)  
Ha: Not all means are equal (Performances were not equal for all components)*

**Description**: Based on the information provided, we get the test result shown in Table 4 for the significance level is *α*=0.05, and the degrees of freedom are *df*1​=3 and *df*2​=3, therefore, the rejection region for this F-test is  *R* = {*F*: *F* > 2.678}. The computed test statistic F equals 3.8499, which is not in the 95% region of acceptance: [-∞: 2.678]. Decision about the null hypothesis are: p-value equals 0.0113, [p ( x ≤ F ) = 0.988735 ]. It means that the chance of type1 error (rejecting a correct H0) is small: 0.01126 (1.13%). The smaller the p-value the stronger it supports H1. Again, from the sample information we get that F = 3.85 > *Fc*​=2.678, it is then concluded that *the null hypothesis is rejected.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **Degrees of Freedom**  **DF** | **Sum of Squares SS** | **Mean Square MS** | **F-Stat** | **P-Value** |
| Between Groups | 3 | 19.5875 | 6.5292 | 3.8499 | 0.0113 |
| Within Groups | 124 | 210.2851 | 1.6958 |  |  |
| Total | 127 | 229.8726 |  |  |  |

Table 4: ANOVA Test Results Summary

Chart, box and whisker chart

Description automatically generated Line chart

Description automatically generated

Figure 4: Box plot of user performance (left), ANOVA Results: F=3.85, p-value=0.0113 (right)

**Conclusion**: Finally, it is concluded that the null hypothesis Ho *is rejected* at the *α=0.05* significance level. In other words, the difference between the averages of some groups is big enough to be statistically significant. Figure 4 (right) shows the results of the One-Way ANOVA. And from normality test results we see, CA+Bubble has significantly higher means compared other distributions. CA+Grid has closer mean with VSUP+Bubble, and VSUP+Grid has significantly lower mean among all. So, we can conclude CA group has significantly better user performance compared to VSUP.

**6.1.2 Paired t-test**

We have generated the CA and VSUP data from the four components performance data by grouping the two pairs CA = (CA+Bubble and CA+Grid) and VSUP = (VSUP+Bubble and VSUP+Grid). The statistical summary of performance data is CA (Mean=5.938, SD=1.105, SEM=0.195, N=32), VSUP (Mean=5.422, SD=1.078, SEM=0.191, N=32). By using Shapiro-Wilk normality test for significance level of 0.005, we obtain results CA (p-value=0.017, W=0.916, S=-0.4622, K=-0.8658), VSUP (p-value=0.017, W=0.956, S=0.07107, K=-0.8737).

To discuss and conclude on results, we need to first define Null (Ho) and Alternative Hypotheses (Ha) as follows (hypotheses need to be tested using paired t-test):   
 *Ho: μD​​ = (μ1​ - μ2) >= 0 (performance of CA is higher or equal to performance of VSUP)   
Ha: μD = ​(μ1​ - μ2) < 0(performance of CA is less than performance of VSUP)*

This corresponds to a left-tailed test, for which a t-test for two paired samples be used.

**Discussion**: Based on the information provided, the significance level is *α*=0.05, and the critical value for a left-tailed test is *tc*​ = −1.696. The rejection region for this left-tailed test is *R*={*t* : *t* < −1.696}. The computed test-statistic = 3.61. Decision about the null hypothesis: since it is observed that *t* = 3.61 ≥ *tc* ​= −1.696, it is then concluded that *the null hypothesis is not rejected.* Using the P-value approach: The p-value is *p* = 0.9995, and since *p* = 0.9995 ≥ 0.05, it is concluded that the null hypothesis is not rejected.

**Conclusion:** It is concluded that the null hypothesis *Ho is not rejected.* Confidence Interval: The 95% confidence interval is 0.224 < *μD* ​< 0.807. So, based on statistical test results, analysis and hypothesize conclusion, we can say that performance of CA quantitatively surpassed the performance of VSUP.

**6.1.3 Time Utilization Results**

Our automated system tracked effective response time for every component separately. The statistical summary of the timing data is CA (Mean=8.675, SD=2.320, SEM=0.410, N=32) and VSUP (Mean=9.647, SD=3.123, SEM=0.552, N=32). The Shapiro-Wilk tests on both distributions showed that they met the normality test with the results CA (W(32) = 0.959, p = 0.254) and VSUP (W(32) = 0.977, p = 0.716).

To discuss and reach in conclusion from results, we need to first define Null (Ho) and Alternative Hypotheses (Ha) as follows (hypotheses need to be tested using paired t-test):   
 *Ho: μD​​ = (μ1​ - μ2) <= 0 (CA response was equal or faster than VSUP response)   
Ha: μD = (μ1​ - μ2) > 0 (CA response was slower than VSUP response)*   
  
This corresponds to a right-tailed test, for which a t-test for two paired samples are used.

**Discussion**: Based on the information provided, the significance level is *α* = 0.05, and the critical value for a right-tailed test is *tc*​ = 1.696. The rejection region for this right-tailed test is *R*={*t* : *t* > 1.696}. The computed test-statistic is equal to -2.656. Since it is observed that *t* = −2.656 ≤ *tc* ​= 1.696, it is then concluded that *the null hypothesis is not rejected.*

Using the P-value approach: The p-value is *p* = 0.9938, and since *p* = 0.9938 ≥ 0.05, it is concluded that the null hypothesis is not rejected.

**Conclusion:** It is concluded that the null hypothesis *Ho is not rejected.* The 95% confidence interval is −1.718 < *μD* ​< −0.226.So, based on above statistical test results, analysis and hypothesize conclusion, we can essentially say that user performance in CA method was faster than VSUP method.

**6.2 SUS Results**

The SUS provides a quick tool for measuring the usability of various kinds of systems based on user experience [44]. It consists of a 10-item questionnaire with five scale response from participants starting from Strongly agree to Strongly disagree. Collectively its use is in classifying the ease of use of the system being tested. We interpret the results by normalizing the scores to produce a percentile ranking. By convention of SUS scoring, based on Sauro [29], we converted SUS results to SUS scores by the following rules:

* For odd items: subtract one from the user response.
* For even-numbered items: subtract the user responses from 5
* This scales all values from 0 to 4 (with four being the most positive response).
* Add up the converted responses for each user and multiply that total by 2.5. This converts the range of possible values to a range from 0 to 100 instead of 0 to 40.

The statistical overview of the scores is CA (Mean=60.078, SD=16.307, SEM=2.883, N=32) and VSUP (Mean=61.094, SD=14.227, SEM=2.515, N=32). The Shapiro-Wilk tests on both distributions showed that they do not meet normality test for CA (W(32) = 0.913, p = 0.013) and VSUP (W(32) = 0.889, p = 0.003). We needed run Kruskal-Wallis Test on the data, which is non-parametric alternative to the paired t-test since the distributions are not normal. The purpose of the test is to assess whether or not the samples come from populations with the same population median.

To reach out in conclusion, we need to define Null (Ho) and alternative hypothesis (Ha) as follows:

*Ho*: The samples come from populations with equal medians.

*Ha*: The samples come from populations with medians that are not all equal.

**Discussion**: Based on the information provided, the significance level is *α*=0.05, and the number of degrees of freedom is *df* = 2 – 1 = 1. Therefore, the rejection region for this Chi-Square test is *R* = {*χ*2: *χ*2 > 3.841}. The computed test (H) statistic is = 0.146. Decision about the null hypothesis: since it is observed that *χ2=0.146 ≤ χc2​ = 3.841*, it is then concluded that the null hypothesis is not rejected.

Using p-value approach: *p* = 0.702, and since *p* = 0.702 ≥ 0.05, it is concluded that the null hypothesis is not rejected.

**Conclusion**: It is concluded that the null hypothesis *Ho* is not rejected*.*

Finally, although the scores of the methods are slightly varying in naked eyes, the differences (χ2 = 0.146, p = 0.702, df = 1) were not statistically significant as per Kruskal-Wallis test at *α* = 0.05.

**6.3 NASA-TLX Results**

TLX stands for Task Load Index and is a measure of perceived workload [39]. Just like SUS data, we have collected Nasa-TLX test data from our online system. A TLX method increments of high, medium, and low estimates for each point result in 21 gradations on the scales. To score, we subtract 1 from the given rating in the range of 1-21, and multiply by 5. For example, if user gives a rating 5, the score would be (5-1) x 5 = 20. Applying Shapiro-Wilk Normality Test for α = 0.05, we find the test results as:

CA = Mental Demand (W=0.906, p-value=0.009), Physical Demand (W=0.914, p-value=0.014), Temporal Demand (W=0.948, p-value=0.128), Performance (W=0.948, p-value=0.014), Effort (W=0.942, p-value=0.085), Mental Frustration (W=0.916, p-value=0.017)

VSUP = Mental Demand (W=0.863, p-value=0.001), Physical Demand (W=0.903, p-value=0.007), Temporal Demand (W=0.938, p-value=0.067), Performance (W=0.887, p-value=0.003), Effort (W=0.901, p-value=0.006), Mental Frustration (W=0.877, p-value=0.002).

Other than temporal demand, none of the group found to be normal distribution. Hence, we used the Kruskal-Wallis non-parametric test to evaluate the differences across the two methods of uncertainty representations (CA and VSUP) on NASA-TLX ratings provided by the participants.

We define null (*Ho*) and alternative hypotheses (*Ha)* as the following way for our test.

*Ho: The samples come from populations with equal medians*

*Ha: The samples come from populations with medians that are not all equal*

The following table 5 shows the summary of such test results at the α = 0.05 significance level:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **NASA-TLX** | **X2** | **P** | **df** | **H** | **Conclusion** |
| Mental Demand | 0.19 | 0.6626 | 1 | 0.19 | *Not Rejected* |
| Physical Demand | 0.062 | 0.8038 | 1 | 0.062 | *Not Rejected* |
| Temporal Demand | 0.018 | 0.8932 | 1 | 0.018 | *Not Rejected* |
| Performance | 3.61 | 0.0574 | 1 | 3.61 | *Not Rejected* |
| Effort | 0.062 | 0.8038 | 1 | 0.062 | *Not Rejected* |
| Mental Frustration | 0.173 | 0.6772 | 1 | 0.173 | *Not Rejected* |

Table 5: Kruskal-Wallis test results of NASA-TLX

**Conclusion**: No statistically significant differences were found between the learning conditions on: mental demand (χ2 = 0.19, p = 0.6626, df = 1), physical demand (χ2 = 0.62, p = 0.8038, df = 1), temporal demand (χ2 = 0.018, p = 0.8932, df = 1), performance (χ2 = 3.61, p = 0.0574, df = 1), effort (χ2 = 0.62, p = 0.8038, df = 1), and mental frustration (χ2 = 0.61, p = 0.6772, df = 1) for the significance level α = 0.05.

**7 User Comments:**

Although participants did not offer many informative comments, we note a few comments that were made during the experiment. Participants (4, 21) commented that “*CA representation is deterministically difficult*” but we also noted that in these cases the comment was the opposite of their performance given that they performed better in CA than VSUP. It is interesting, nonetheless. Some other participants (19, 24) made a more nuanced comment, stating that “*CA representation is complex but gives more confidence to find target*”. Another comment that was commonly expressed by participants (14, 25, 31) is that “*Colors are very close in VSUP which made them puzzled to select target*”.

**8 Conclusions and Future Work**

In this paper, we propose a novel approach for uncertainty visualisation, namely Chromatic Aberration. We conducted a within subject comparative user study with VSUP and our system to assess user performance accuracy/error rate, task completion time, and subjective assessment with NASA-TLX and SUS. From numerical analysis and evaluation of the results, we see user performance and perception is both statistically improved and faster compared to VSUP whereas in the subjective assessment do not vary significantly.

Nevertheless, we note that in real chromatic aberration the chromatic blurring appears continuously from inner edge to outer edge. But in our case, it just gives us a range of uncertainty for the prediction, so the edges are with the same bright color. However, our simplified implementation allows us to reduce the aberration to both double and/or single parameter, which facilitates chromatic aberration tuning with regards to the amount of represented uncertainty. It also allows one to implement the approach relatively easily using standard d3 and SVG operations. However, additional research could be conducted that examine more sophisticated effects. In addition, further research could be conducted with more levels of uncertainties than were tested in both in Correll et al. [19] and the present work, for instance 8-levels instead of 4-levels. The role of CA might also be explored in animated visualizations. And finally, other future work may refine and expand upon some of our other experimental designs such as the starfish streamgraph layout briefly discussed.

**REFERENCES**

1. Song, Xin; Xiao, Jun PhD; Deng, Jiang PhD; Kang, et al. Time series analysis of. influenza incidence in Chinese provinces from 2004 to 2011. Received March 1, 2016, Accepted May 20, 2016, Medicine: June 2016 - Volume 95 - Issue 26 - p e3929
2. Christophorus Beneditto, Aditya Satrio et al. Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET, https://doi.org/10.1016/j.procs.2021.01.036
3. Leo J, Luhanga E, Michael K. Machine Learning Model for Imbalanced Cholera Dataset in Tanzania. The Scientific World Journal. 2019 Jul; 2019: p. 1–12.
4. Emrah Gecili, Assem Ziady, Rhonda D. Szczesniak. Forecasting COVID-19 confirmed cases, deaths and recoveries: Revisiting established time series modeling through novel applications for the USA and Italy. **Received:** June 30, 2020; **Accepted:** December 5, 2020; **Published:** January 7, 2021.
5. Miranda GHB, Baetens JM, Bossuyt N, Bruno OM, Baets BD. Real-time prediction of influenza outbreaks in Belgium. Epidemics. 2019 Sep; 28: p. 100341.
6. Muktevi Srivenkatesh, Performance Evolution of Different Machine Learning Algorithms for Prediction of Liver Disease. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-2, December 2019.
7. K. Koh, B. Kim & J. Seo. 2014. Effect of lateral chromatic aberration for chart reading in information visualization on display devices. Advanced Visual Interfaces. Como, Italy, 289-292.
8. H. S. Yoo. 2007. Color illusions on liquid crystal displays and design guidelines for information visualization. Master of Science, Virginia Tech.
9. T. Boult & W. Wolberg. 1992. Correcting chromatic aberrations using image warping. CVPR, Champaign, IL, 684–87.
10. M. K. Johnson & H. Farid. 2006. Exposing digital forgeries through chromatic aberration. Multimedia and security, Geneva, Switzerland, 48-55.
11. S. Lee, E. Eisemann & H.P. Seidel. 2010. Real-time lens blur effects and focus control. ACM SIGGRAPH, Los Angeles, 1-7.
12. Bonneau et al. Overview and State-of-the-Art of Uncertainty Visualization, The University of Grenoble, France e-mail: Georges-Pierre.Bonneau@ujf-grenoble. fr. ISBN: 978-1-4471-6496-8.
13. Henning Griethe et al. The Visualization of Uncertain Data: Methods and Problems. Computer Graphics, 18051 Rostock, Germany. January 2006, 2,988 reads, 36 publications, 262 citations.
14. Deitrick, S., Edsall, R.: The influence of uncertainty visualization on decision making: An empirical evaluation. In: Progress in Spatial Data Handling, pp. 719–738. Springer Berlin Heidelberg (2006).
15. Lundstr¨om, C., Ljung, P., Persson, A., Ynnerman, A.: Uncertainty visualization in medical volume rendering using probabilistic animation. IEEE Transactions on Visualization and Computer Graphics 13(6), 1648–1655 (2007).
16. Pang, A., Wittenbrink, C., Lodha., S.: Approaches to uncertainty visualization. The Visual Computer 13(8), 370–390 (1997).
17. R.P. Botchen; D. Weiskopf; T. Ertl. Texture-based visualization of uncertainty in flow fields. VIS 05. IEEE Visualization, Minneapolis, MN, USA. 23-28 Oct. 2005. **ISBN:** 0-7803-9462-3.
18. Aasim Kamal · Parashar Dhakal, et al. Recent advances and challenges in uncertainty visualization: a survey. May 2021, Journal of Visualization 24(5):1-30.
19. Michael Correll, Dominik Moritz, Jeffrey Heer. Value-Suppressing Uncertainty Palettes.‬‬‬‬‬‬‬‬‬ Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. April 2018. Paper No.: 642 Pages 1–11.
20. Michelle Korporaal, Ian T. Ruginski, and Sara Irina Fabrikant. Effects of Uncertainty Visualization on Map-Based Decision Making Under Time Pressure. Human-Media Interaction, a section of the journal Frontiers in Computer Science. Received: 22 May 2020. doi: 10.3389/fcomp.2020.00032.
21. Miriam Greis, Passant El Agroudy, et al. 2016. Decision-Making under Uncertainty: How the Amount of Presented Uncertainty Influences User Behavior. In Proceedings of the 9th Nordic Conference on Human-Computer Interaction. ACM, 52. DOI: http://dx.doi.org/10.1145/2971485.2971535.
22. Lydia R Lucchesi and Christopher K Wikle. 2017. Visualizing uncertainty in areal data with bivariate choropleth maps, map pixelation and glyph rotation. Stat (2017). DOI:http://dx.doi.org/10.1002/sta4.150.
23. R. Finger and A. M. Bisantz. Utilizing graphical formats to convey uncertainty in a decision-making task. Theoretical Issues in Ergonomics Science, 3(1):1–25, 2002.
24. M. Kay, T. Kola, J. R. Hullman, and S. A. Munson. When (ish) is my bus?: User-centered visualizations of uncertainty in everyday, mobile predictive systems. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pages 5092–5103. ACM, 2016.
25. C. M. Wittenbrink, A. T. Pang, and S. K. Lodha. Glyphs for visualizing uncertainty in vector fields. IEEE transactions on Visualization and Computer Graphics, 2(3):266–279, 1996.
26. LeGrand H Hardy, Gertrude Rand, and M Catherine Rittler. 1945. Tests for the detection and analysis of color-blindness. I. The Ishihara test: an evaluation. JOSA 35, 4 (1945), 268–275.
27. Heidi Lam, Enrico Bertini, et al. Empirical Studies in Information Visualization: Seven Scenarios. IEEE Electronic Library (IEL) Journals. 06 December 2011. Page(s):1520 – 1536. ISSN: 1077-2626.  IEEE Electronic Library (IEL) Journals.
28. I. Scott MacKenzie, Within-subjects vs. Between-subjects Designs: Which to Use? https://www.yorku.ca/mack/RN-Counterbalancing.html.
29. Jeff Sauro, PhD. Measuring Usability with the System Usability Scale (SUS). https://measuringu.com/sus. February 3, 2011.
30. Keppel, Geoffrey: Design and Analysis, 3rd Edition. Prentice-Hall.
31. Jason Brownlee. Deep Learning Models for Univariate Time Series Forecasting. https://machinelearningmastery.com/how-to-develop-deep-learning-models-for-univariate-time-series-forecasting.
32. Aayush Agrawal, Building Neural Network from scratch. https://towardsdatascience. com/building-neural-network-from-scratch-9c88535bf8e9.
33. Max Schneider, Michelle McDowell et al. Effective uncertainty visualization for aftershock forecast maps. Natural Hazards and Earth System Sciences. Discussion started: 3 September 2021. https://doi.org/10.5194/nhess-2021-237..
34. Maria Riveiro. 2007. Evaluation of uncertainty visualization techniques for information fusion. In 10th International Conference on Information Fusion. IEEE, 1–8. DOI: http://dx.doi.org/10.1109/ICIF.2007.4408049.
35. J. Hullman, X. Qiao, M. Correll, A. Kale, and M. Kay. In pursuit of error: A survey of uncertainty visualization evaluation. IEEE transactions on visualization and computer graphics, 25(1):903–913, 2019..
36. Shapiro, S. S.; Wilk, M. B. (1965). "An analysis of variance test for normality (complete samples)". Biometrika. 52 (3–4): 591–611.
37. S. Greenberg and B. Buxton, “Usability Evaluation Considered Harmful (Some of the Time),” Proc. Conf. Human Factors in Computing Systems (CHI), pp. 217-224, 2008.
38. Shapiro, S. S.; Wilk, M. B. (1965). "An analysis of variance test for normality (complete samples)". Biometrika. 52 (3–4): 591–611..
39. NASA (1986). Nasa Task Load Index (TLX) v. 1.0 Manual.
40. Brooke, J. (1986). SUS: a "quick and dirty" usability scale. In P. W. Jordan; B. Thomas; B. A. Weerdmeester; A. L. McClelland (eds.). Usability Evaluation in Industry. London: Taylor and Francis.
41. Stéfan van der Walt and Nathaniel Smith. 2015. Mpl colormaps. https://bids.github.io/colormap/, (2015).
42. Michael Correll and Michael Gleicher. 2014. Error bars considered harmful: Exploring alternate encodings for mean and error. IEEE Transactions on Visualization and Computer Graphics 20, 12 (2014), 2142–2151. DOI: • <http://dx.doi.org/10.1109/TVCG.2014.2346298>
43. Sathler C, Luciano J. Predictive modeling of dengue fever epidemics: A Neural Network Approach. 2017. Data Science for Drug Discovery, Health and Translational Medicine. December 10, 2017. I590.
44. Ken Brodlie, Rodolfo Allendes Osorio, and Adriano Lopes. 2012. A review of uncertainty in data visualization. In Expanding the frontiers of visual analytics and visualization. Springer, 81–109. DOI: <http://dx.doi.org/10.1007/978-1-4471-2804-5_6>