**Abstract:**  
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 obtain 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 is able to identify targets in CA method more accurately than VSUP method. In addition, their speed of target identification was significantly faster in CA to VSUP method. But their preference between the two does not vary significantly.

**Chapter 7**

**Evaluation: Results and Numerical Analysis**

**7.1 Introduction**In the previous chapter, we have described the user study design including questionnaire presentation, data collection procedure, data structure, and the data storing mechanism. In this chapter, we discuss and analyze the study generated data with the help of statistical principles which are commonly used for user studies such t-test, and ANNOVA. The goal of the study was to evaluate user performance and user experience of our newly designed approach of uncertainty visualisation and generate quantitative and qualitative (user preferences, SUS, etc.) data. We will use that data in analysis, prepare results, eventually discuss the findings, and finally come to conclusions.

**7.2.1 Sample Population Demographics**

**Age and Gender**

The sample population of 32 participants had a distribution of 78.12% male (25/32), 21.88% female (7/32). Given that we did not have any plan to control for gender within the recruitment policy, we have recruited on a first come first join basis. All participants were in the age range of 22-35 years old.

**Education**

There were 25% CS grad students (8/32), 28.12% CS undergrad students (9/32), 34.37% ICT grad students (11/32), 3% Statistics undergrad students (1/32) and 9.37% telecom professionals (3/32).

**Prior experience in visualisation**

* All CS and ICT students had taken at least one course of visualisation/graphics design in their undergraduate/graduate level and 12 of them had conducted their undergraduate thesis related to visualization or graphics or image processing.
* Telecom professionals also came from a CS background, so they had taken Computer graphics course in their undergraduate level.
* All participants had played computer games many times.
* 15 participants have knowledge about animated movies.

**7.2 Study results**

We have obtained several kinds of data from the user study such as:

1. Quantitative Questionnaire Results.
2. Time utilization data for each component.
3. SUS data for CA and VSUP
4. NASA-TLX for CA and VSUP

We analyse all these data in various ways in the following sections which helps to reach conclusions from the study.

**7.2.1 Quantitative Questionnaire Results**

As we have four core components, we designed 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.

**7.2.1.1 One-way repeated measures ANOVA**The user performance results that we received from the study can be summarized as Table 7.1 and graphical box plot in Figure 7.2 and the complete raw data is attached in APPENDIX-L.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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 7.1: Data summary

The results of a one-way ANOVA can be considered reliable if the following assumptions are met:

1. the response variable (the dependent variable) is normally distributed.
2. the samples are independent.
3. the variances of populations are equal.

Since the sample are taken from independent interfaces of questionnaire, the requirement (2) fulfilled. Again, as per Keppel’s ratio rule of thumb [70], 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.

So, from Table 7.1, 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 shown in Table 7.2 which indicates the distributions of the components are in normal distribution for significance level of 0.005 and that satisfies the requirement (1) and we can conduct an ANOVA test. Additionally, we have also showed box-plot (Figure 7.1) and normal distribution graphs in Figure 7.2.

|  |  |  |  |
| --- | --- | --- | --- |
| Component | W | P | Status |
| Ca + Bubble | 0.915 | 0.015 | Normal |
| Ca + Grid | 0.932 | 0.045 | Normal |
| VSUP + Bubble | 0.911 | 0.012 | Normal |
| VSUP + Grid | 0.913 | 0.013 | Normal |

Table 7.2: Shapiro-Wilk Test of Normality

Chart, box and whisker chart

Description automatically generated

Figure 7.1: Box plot of user performance

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Ca+Bubble

Ca+Grid

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

VSUP+Bubble

VSUP+Grid

Figure 7.2: Normal Distributions for each component

We get the ANOVA summary as in Table 7.4.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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 7.4: ANOVA Summary

So, we briefly point out the findings from the ANOVA test as follows:

(1) Null and Alternative Hypotheses

The following null and alternative hypotheses need to be tested:

Ho:  *μ*1​ = *μ*2 ​= *μ*3 ​= *μ*4​ (Performances were equal for all components)

Ha: Not all means are equal (Performances were not equal for all components)

The above hypotheses will be tested using an F-ratio for a One-Way ANOVA.

(2) Rejection Region   
Based on the information provided, 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 = *R* = {*F*: *F* > 2.678}.

(3) Test Statistics  
The computed test statistic F equals 3.8499, which is not in the 95% region of acceptance:   
[-∞: 2.678].

(4) Decision about the null hypothesis   
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.*

(5) Conclusion  
It is concluded that the null hypothesis Ho *is rejected.* Therefore, there is not enough evidence to claim that all 4-population means are equal, 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.

The following Figure 7.3 summarizes the results of the One-Way ANOVA:

Line chart

Description automatically generated

Figure 7.3: ANOVA Results: F=3.85, p-value=0.0113, Ho rejected.

Finally, from Table 7.1 we see, CA+Bubble has significantly higher means compared other distributions and CA+Grid has closer mean with VSUP+Bubble, and VSUP+Grid has significantly lower mean among all. So, we can conclude CA has significantly better user experience compared to VSUP.

**7.2.1.2 Paired t-test**

We have generated the CA and VSUP data from the four components performance data by grouping and averaging the two pairs (CA+Bubble, CA+Grid and VSUP+Bubble, VSUP+Grid). Now the statistical summary of CA and VSUP data are shown in the following Table 7.5.

|  |  |  |
| --- | --- | --- |
| Group | CA | VSUP |
| Mean | 5.938 | 5.422 |
| SD | 1.105 | 1.078 |
| SEM | 0.195 | 0.191 |
| N | 32 | 32 |

Table 7.5: Summary of CA vs VSUP performance

We present test result of Shapiro-Wilk normality test for significance level of 0.005 in the following table 7.6 where we see both distributions do not differ significantly from normal distribution. We also show the normal distribution graphs in Figure 7.4.

|  |  |  |
| --- | --- | --- |
| Group | CA | VSUP |
| Skewness | -0.4622 | 0.07107 |
| Kurtosis | -0.8658 | -0.8737 |
| p-value | .017 | 0.017 |
| W | 0.916 | 0.956 |

Table 7.6: Shapiro-Wilk Normality Test

Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated

Figure 7.4: Normal Distribution CA (left), VSUP (right)

The following steps show the paired t-test results for the given data and draws conclusion from the test:

(1) Null and Alternative Hypotheses   
The following null and alternative hypotheses need to be tested using paired t-test:  
Ho: *μD*​​ = (*μ*1​ - *μ*2) >= 0 (performance of CA is higher or equal to performance of VSUP)  
Ha: *μD* = ​(*μ*1​ - *μ*2) < 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.

(2) Rejection Region   
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}

(3) Test Statistics   
The computed t-statistic = 3.61

(4) 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.

(5) Conclusion  
It is concluded that the null hypothesis Ho *is not rejected.*

Confidence Interval: The 95% confidence interval is 0.224 < *μD* ​< 0.807.

We can visualize the paired T-test scenario graphically as follows:

Chart

Description automatically generated  
Figure 7.5: Paired t-test sample with p-value=0.9995 for CA vs VSUP performance.

Finally, based on above statistical test results, analysis and hypothesize conclusion, we can essentially say that performance of CA quantitatively surpassed performance of VSUP.

**7.2.2 Time Utilization Results**

Our automated system tracked effective response time for every component separately. The statistical summary of the timing data is represented in the following table 7.7

|  |  |  |
| --- | --- | --- |
| Group | CA | VSUP |
| Mean | 8.675 | 9.647 |
| SD | 2.320 | 3.123 |
| SEM | 0.410 | 0.552 |
| N | 32 | 32 |

Table 7.7: Summary of CA vs VSUP timing

The Shapiro-Wilk tests on both distributions showed that they met the normality test with the following results:

For CA = W(32) = .959, p = .254

For VSUP = W(32) = .977, p = .716

The following steps show the paired t-test results for the given time data and draws conclusion from the test:

(1) Null and Alternative Hypotheses

The following null and alternative hypotheses need to be tested:

Ho: *μD*​​ = (*μ*1​ - *μ*2) <= 0 (CA response was equal or faster than VSUP response)  
Ha: *μD* = (*μ*1​ - *μ*2) > 0 (CA response was slower than VSUP response)

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

(2) Rejection Region

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}

(3) Test Statistics

The computed t-statistic is equal to -2.656

(4) Decision about the null hypothesis

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.

(5) Conclusion

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

Confidence Interval

The 95% confidence interval is −1.718 < *μD* ​< −0.226.

We can visualize the paired T-test scenario graphically as follows:

Chart

Description automatically generatedFigure 7.6: Paired t-test sample with p-value=0.9938 for CA vs VSUP timing.

Finally, 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.

**7.2.3 SUS Results**

The SUS provides a quick tool for measuring the usability of various kinds of systems based on user experience. It consists of a 10-item questionnaire with five scale response from participants starting from Strongly agree to Strongly disagree. It doesn’t have any right or wrong evaluation of any question and hence collectively its use is in classifying the ease of use of the system being tested. The best way to interpret the results is to normalize the scores to produce a percentile ranking. By convention of SUS scoring, based on Jeff Sauro [69], 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 given below:

|  |  |  |
| --- | --- | --- |
| Group | CA | VSUP |
| Mean | 60.078 | 61.094 |
| SD | 16.307 | 14.227 |
| SEM | 2.883 | 2.515 |
| N | 32 | 32 |

Table 7.7: SUS scores summary of CA vs VSUP

The Shapiro-Wilk tests on both distributions showed that they do not meet normality test with the following results:

For CA = W(32) = 0.913, p = 0.013

For VSUP = W(32) = 0.889, p = 0.003

The following steps show the Kruskal-Wallis Test results, 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.

(1) Null and Alternative Hypotheses

The following null and alternative hypotheses need to be tested:

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

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

The above hypotheses will be tested using the Kruskal-Wallis test.

(2) Rejection Region   
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}.   
  
(3) Test Statistics The computed H statistic is = 0.146   
  
(4) 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 the P-value approach: The p-value is *p* = 0.702, and since *p* = 0.702 ≥ 0.05, it is concluded that the null hypothesis is not rejected.

(5) ConclusionIt is concluded that the null hypothesis *Ho* is not rejected*.*

Chart, box and whisker chart

Description automatically generated

Figure 7.7: SUS rating plots for visualization methods

Finally, although the scores of the methods are slightly varying according to Figure 7.7, the differences (χ2 = 0.146, p = 0.702, df = 1) were not statistically significant as per Kruskal-Wallis test at *α* = 0.05.

**7.2.4 NASA-TLX Results**

The TLX stands for Task Load Index and is a measure of perceived workload. Just like SUS data, we have collected Nasa-Tlx load 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 20: (5-1) x 5.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Methods | NASA-TLX | Shapiro-Wilk Normality Test (α = 0.05) | | |
| Test Statistic (W) | p-value | Status |
| CA | Mental Demand | 0.906 | 0.009 | Not normal |
| Physical Demand | 0.914 | 0.014 | Not normal |
| Temporal Demand | 0.948 | 0.128 | Normal |
| Performance | 0.932 | 0.044 | Not normal |
| Effort | 0.942 | 0.085 | Not normal |
| Mental Frustration | 0.916 | 0.017 | Not normal |
| VSUP | Mental Demand | 0.863 | 0.001 | Not normal |
| Physical Demand | 0.903 | 0.007 | Not normal |
| Temporal Demand | 0.938 | 0.067 | Not normal |
| Performance | 0.887 | 0.003 | Not normal |
| Effort | 0.901 | 0.006 | Not normal |
| Mental Frustration | 0.877 | 0.002 | Not normal |

Table 7.8: Normality test results of NASA-TLX score

Since almost the datasets didn’t follow a normal distribution, 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. The following null and alternative hypotheses need to be tested with Kruskal-Wallis 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 shows the summary of such test results of Kruskal-Wallis test 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 7.9: Kruskal-Wallis test results of NASA-TLX

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.

**7.3 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*”.

**7.4 Summary of the results**

We obtained two types of results from the study i. Quantitative and ii. Qualitative. Based on a statistical analysis in the former sections, we can summarise as that the Quantitative results were better for CA than the VSUP approach whereas subjective results were not significantly different from each other. In other words, user performance and speed of target identification was significantly better in CA than VSUP although user preference was more or less similar.

**Chapter 8**

**Conclusions and Future Work**

In this thesis, we propose a novel approach of uncertainty visualisation in terms of Chromatic Aberration in web platform. There is an existing uncertainty visualisation system namely VSUP that presents a different approach of uncertainty visualisation. We conduct 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 statistically improved and faster compared to VSUP whereas in the subjective assessment do not vary statistically significantly.

Nevertheless, we admit that in real aberration the picture blurring happens very slowly from inner edge to outer edge but in our case, it just gives us a range of uncertainty for the prediction, so the whole 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. To mitigate the blurring effect additional research can be conducted such add adding additional color effects. In addition, further research could be conducted with more levels of uncertainties than were tested in both VSUP of Correll et al. [35] and the present work, for instance 8-levels instead of 4-levels.