Supplementary Material – Exploring Time Series Segmentations Using Uncertainty and Focus+Context Techniques

Christian Bors, Christian Eichner, Christian Tominski,
Silvia Miksch, Heidrun Schumann, Theresia Gschwandtner

March 30th, 2019

The following supplementary material file contains the questions participants received during our evaluation, followed by the overall scores and completion times for each of the groups. The study was designed as a between-subject study, meaning every participant answered every question for one of the visualization designs (participants' questions were randomized to mitigate learning effects). A total of 111 persons participated in the study. The participants were undergraduate computer science students attending a lecture on information design and visualization, so they had basic experience with information visualization. Participants first received a short introduction, to familiarize them with the data at hand, and how it could be interpreted appropriately.

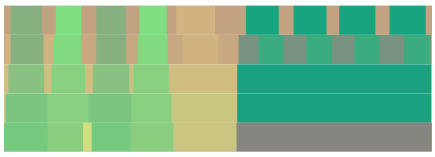
The study results were tested against the hypotheses (see Section 2.1) using Friedman Tests to test for statistical significance of Hypothesis H_2 (see Section 2.2) and a post-hoc Nemenyi Test to determine the significant pairs, if significance is found. Non-equivalence tests were conducted to test hypotheses H_0 (Section 2.3), H_1 (Section 2.4), and H_3 (Section 2.6). TODO: add p-value

Since non-significance was found for H_2, we also tested this hypothesis for non-inferiority (Section 2.5).

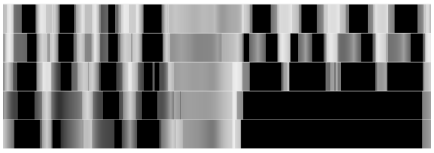
Sections 3 show the test results for all hypotheses, and Section 4 gives general implications that can be drawn from the evaluation results.

1 Visualization Designs

For the study we developed four different uncertainty visualization designs (see Figure 1).

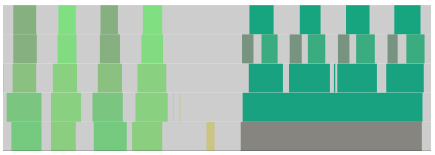


(a) The regular segmentation result design encodes labels as colors and visualizes the segments over time.



(b) The Uncertainty Heatmap design only encodes uncertainty as saturation.

(c) The Gradient Uncertainty design extends the regular segmentation result label view by encoding uncertainty as saturation.



(d) The Threshold Uncertainty design only shows segment labels for timestamps with result uncertainty below the specified threshold (in this case 40%).

Figure 1: Visualization designs showing result uncertainty for uncertainty-aware segmentation result overview.

2 Questions

Questions 1 to 6 are used for testing hypotheses H_0, H_1, and H_2. Questions 7 to 9 are used for testing hypothesis H_3. The questions 1 to 6 are exemplified with the composite visualization, showing the computed segments of a result over time (top), alongside the associated uncertainties as line charts (bottom).

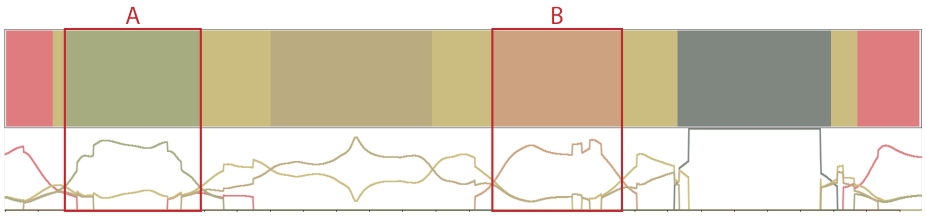


Figure 2: Question 1: Out of the highlighted areas (red frames), which is the most certain?

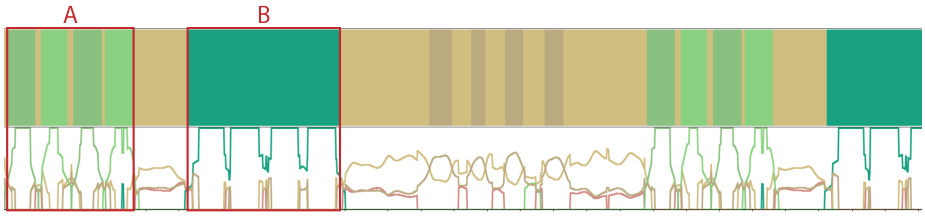


Figure 3: Question 2: Out of the highlighted segments (red frames), which is the most certain?

Figure 4: Question 3: Out of the highlighted areas (red frames), which is the most certain?

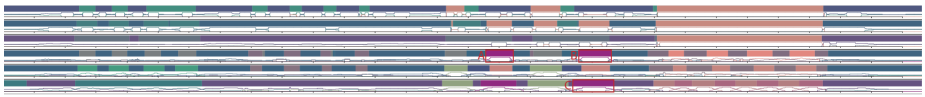


Figure 5: Question 4: Out of the highlighted segments (red frames), which is the most certain?

3 User Study Results - Uncertainty in Time Series Segmentation Results

3.1 Hypotheses

H_0 The *Gradient Uncertainty Plot* does not perform significantly worse than a composite view of the regular visualization of segmentation results

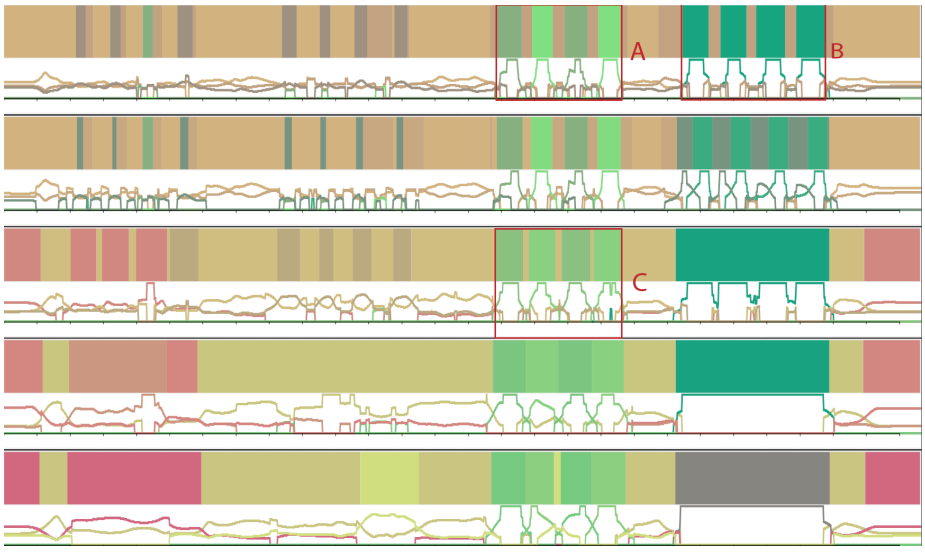


Figure 6: Question 5: Out of the highlighted areas (red frames), which is the most certain?

Figure 7: Question 6: Please sort the following highlighted Segments from Most Certain to Least Certain.

as colored bars plus an additional line plot showing result uncertainty.
H_1 The *Gradient Uncertainty Plot* does not perform worse than the *Uncertainty Heatmap* plot showing result uncertainty.

Figure 8: Question 7: Out of the highlighted areas (red frames), which has less uncertainty (Area Chart Variant)?

Figure 9: Question 8: Out of the highlighted areas (red frames), which has less uncertainty (Area Chart Variant)?

Figure 10: Question 9: Out of the highlighted areas (red frames), which area has the least overall uncertainty (Area Chart Variant)?

H_2 The *Gradient Uncertainty Plot* is more effective than an interactive *Threshold Uncertainty Plot* for assessing result uncertainties of a large number of segmentation results, H_2a especially with limited vertical space available.
H_3 The *Heatband Uncertainty Plot* is not inferior to the *Area Uncertainty Plot* for showing value uncertainty.

3.2 Hypothesis Testing

H_2 will be tested using a Friedman test to calculate statistical significance, and a post-hoc Nemenyi test determining if the design pair in question, i.e., gradient - threshold, are significantly different, *followed by a superiority test.*

H_0, H_1, and H_3 will be tested using a non-inferiority test, evaluating if one used method is not significantly inferior to another. Using an equivalence test and only observing the *lower bound* will yield the test for non-inferiority (`https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3019319/`).

The bounds are calculated based on the statistical power of 0.95, the number of study participants $n = 111$, and the Significance level $\alpha = 0.05$, yielding the upper and lower bounds, of which only the lower bound will be of interest.

3.3 Significance Tests

Tests for significant differences between designs. Here we try to find significance particularly between the pair Gradient and Threshold plots, which would confirm H_2 with a significant pair Gradient Uncertainty plot - Threshold plot.

3.3.1 Friedman Test - Error and Completion Time over all questions

Questions 1 to 6 error and Completion Time, including post-hoc Nemenyi test:

##

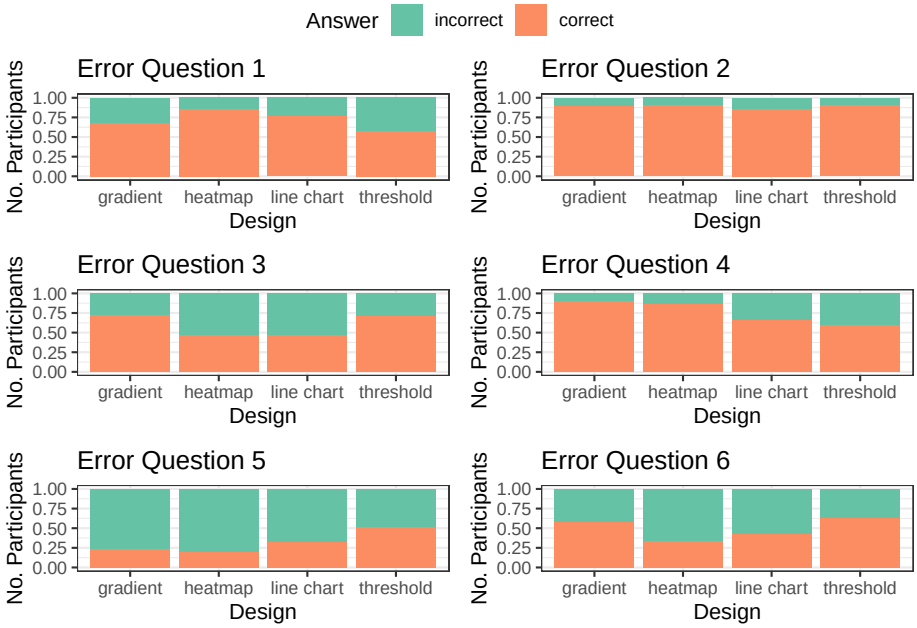


Figure 11: Results – Error Rates per question.

```
##  Friedman rank sum test
##
## data:  u_scores_combined$question , u_scores_combined$design
## and u_scores_combined$id
## Friedman chi-squared = 19.341, df = 3, p-value = 0.0002324

##
##  Friedman rank sum test
##
## data:  u_scores_combined$time , u_scores_combined$design
##        and u_scores_combined$id
## Friedman chi-squared = 286.03, df = 3, p-value < 2.2e-16

##
##  Pairwise comparisons using Nemenyi multiple comparison test
##              with q approximation for unreplicated blocked data
##
## data:  question and design.f and id
##
##            gradient heatmap line chart
## heatmap    0.224    -       -
## line chart 0.082    0.966   -
## threshold  0.974    0.446   0.206
```

Figure 12: Results { Completion times per question.

```
##
## P value adjustment method: none

##
##  Pairwise comparisons using Nemenyi multiple comparison test
##                    with q approximation for unreplicated blocked data
##
## data:  time and design.f and id
##
##                gradient heatmap line chart
## heatmap       1.9e-12  -          -
## line chart 0.04        3.4e-14  -
## threshold   2.9e-14   < 2e-16  2.8e-09
##
## P value adjustment method: none
```

3.3.2 Plots for Error and Completion Time over All Questions

3.3.3 Result

No signi cant pairs for scores were found, however, the di erence in Completion Time is signi cant.

3.3.4 Friedman Test - Error and Completion Time for Questions 4 and 5

An error rate that is signi cantly lower (especially for questions 4 and 5) would con rm that Gradient Uncertainty plots performs better than Threshold plots for use cases where vertical space is limited.

```
##
##  Friedman rank sum test
##
## data:  u_scores_q45$question , u_scores_q45$design
##          and u_scores_q45$id
## Friedman chi-squared = 5.0174, df = 3, p-value = 0.1705

##
##  Friedman rank sum test
##
## data:  u_scores_q45$time , u_scores_q45$design
##          and u_scores_q45$id
## Friedman chi-squared = 160.9, df = 3, p-value < 2.2e-16

##
##  Pairwise comparisons using Nemenyi multiple comparison test
##                    with q approximation for unreplicated blocked data
##
## data:  time and design.f and id
##
##                gradient heatmap line chart
## heatmap       2.6e-07   -          -
## line chart 0.0085     3.5e-14 -
## threshold   2.8e-10   < 2e-16 0.0035
##
## P value adjustment method: none
```

3.3.5 Error

Error Rate: No Signi cance.
Completion Time: Signi cant di erences between all designs. Order: 1.Uncertainty Heatmap , 2. Gradient Uncertainty plot , 3. composite line chart , 4. Threshold plot .

3.3.6 Friedman Test - Error and Completion Time for Questions 3 - 6 (Vertical Comparison)

An error rate that is signi cantly di erent especially for questions 3 - 6 would con rm that Gradient Uncertainty plots performs better than Threshold plots for use cases where vertical space is limited.

```
##
##  Friedman rank sum test
##
## data:  u_scores_q3456$question , u_scores_q3456$design
##          and u_scores_q3456$id
## Friedman chi-squared = 49.709, df = 3, p-value = 9.214e-11

##
##  Friedman rank sum test
##
## data:  u_scores_q3456$time , u_scores_q3456$design
##          and u_scores_q3456$id
## Friedman chi-squared = 243.87, df = 3, p-value < 2.2e-16

##
##  Pairwise comparisons using Nemenyi multiple comparison test
##                  with q approximation for unreplicated blocked data
##
## data:  question and design.f and id
##
##                gradient heatmap line chart
## heatmap       0.0041    -          -
## line chart 0.0069    0.9986   -
## threshold   0.9999    0.0034   0.0058
##
## P value adjustment method: none

##
##  Pairwise comparisons using Nemenyi multiple comparison test
##                  with q approximation for unreplicated blocked data
##
## data:  time and design.f and id
##
##                gradient heatmap line chart
## heatmap       1.2e-10   -          -
## line chart 0.009      3.9e-14  -
## threshold   4.1e-14   < 2e-16  9.1e-07
##
## P value adjustment method: none
```

3.3.7 Plots for Error and Completion Time over Questions 4-5 and 3-6

3.3.8 Results

Error Rate - Signi cance between pairs:

- Gradient Uncertainty plot and Uncertainty Heatmap (0.0041)
  - { Gradient Uncertainty plot performed signi cantly better
- Gradient Uncertainty plot and line plot (0.0069)
  - { Gradient Uncertainty plot performed signi cantly better
- Threshold plot and Uncertainty Heatmap (0.0034)
  - { Threshold Uncertainty plot performed signi cantly better
- Threshold plot and line plot (0.0058)
  - { Threshold Uncertainty plot performed signi cantly better

Completion Time: Signi cant di erences between all designs. Order: 1.Uncertainty Heatmap , 2. Gradient Uncertainty plot , 3. composite line chart , 4. Threshold Uncertainty plot .

3.4 Non-Equivalence Test of Gradient Uncertainty Plot vs Composite Uncertainty and Segmentation Result Plot (H_0)

Testing for non-inferiority (error is lower) of Error (q1 - q6) and completion times (t_q1 - t_q6) between Gradient Uncertainty plot - line plot (H_0).

```
##
##   TOST INDEPENDENT SAMPLES T-TEST
##
##   TOST Results
##   -------------------------------------------------------
##                                        t          df       p
##   -------------------------------------------------------
##     question      t-test           3.192      1330      0.001
##                   TOST Upper      -0.413      1330      0.340
##                   TOST Lower       6.80       1330      < .001
##
##     time          t-test           0.228      1330      0.819
##                   TOST Upper      -3.376      1330      < .001
##                   TOST Lower       3.83       1330      < .001
##   -------------------------------------------------------
##
##
## Equivalence Bounds
```

```
## ------------------------------------------------------------------
##                                    Low              High           Lower          Upper
## ------------------------------------------------------------------
## question       Cohen's d         -0.198            0.198
##                  Raw             -0.0950           0.0950         0.0407         0.127
##
## time           Cohen's d         -0.198            0.198
##                  Raw            -11.0433          11.0433        -4.3428         5.742
## ------------------------------------------------------------------
```

3.5 Non-Equivalence Test of Gradient Uncertainty Plot vs Uncertainty Heatmap (H_1)

Testing for non-inferiority (error is lower) of Error (q1 - q6) and completion times (t_q1 - t_q6) between Gradient Uncertainty plot - Uncertainty Heatmap (H_1).

```
##
##   TOST INDEPENDENT SAMPLES T-TEST
##
##   TOST Results
##   ---------------------------------------------------
##                                        t          df          p
##   ---------------------------------------------------
##     question      t-test             2.57       1330        0.010
##                    TOST Upper       -1.03       1330        0.151
```

```
##                      TOST  Lower       6.18      1330      < .001
##
##      time            t-test             2.06      1330       0.040
##                      TOST  Upper      -1.55      1330       0.061
##                      TOST  Lower       5.66      1330      < .001
##   ------------------------------------------------------
##
##
## Equivalence Bounds
## ---------------------------------------------------------------------
##                                    Low             High          Lower        Upper
## ---------------------------------------------------------------------
## question       Cohen's d         -0.198          0.198
##                  Raw               -0.0946         0.0946       0.0244       0.111
##
## time           Cohen's d         -0.198          0.198
##                  Raw              -13.1132        13.1132       1.5003       13.476
## ---------------------------------------------------------------------
```

3.6 Non-Equivalence Test of Gradient Uncertainty Plot vs Threshold Uncertainty Plot (H_2)

Testing for non-inferiority (error is lower) of Error (q1 - q6) and completion times (t_q1 - t_q6) between Gradient Uncertainty plot - threshold (H_2)

```
##
```

```
##   TOST INDEPENDENT SAMPLES T-TEST
##
##   TOST Results
##   -------------------------------------------------------
##                                          t          df       p
##   -------------------------------------------------------
##      question      t-test           0.287       442      0.774
##                    TOST Upper      -3.32        442      < .001
##                    TOST Lower       3.89        442      < .001
##
##      time          t-test          -2.355       442      0.019
##                    TOST Upper      -5.96        442      < .001
##                    TOST Lower       1.25        442       0.106
##   -------------------------------------------------------
##
##
## Equivalence Bounds
## ---------------------------------------------------------------------
##                                   Low          High        Lower          Upper
## ---------------------------------------------------------------------
## question      Cohen's d       -0.342       0.342
##                 Raw             -0.170       0.170       -0.0641        0.0911
##
## time          Cohen's d       -0.342       0.342
##                 Raw             -22.510      22.510      -24.9997       -4.4147
## ---------------------------------------------------------------------
```

3.7 Non-Equivalence Test of Area Plot vs. Heat Bands (H_3)

Testing for non-inferiority (error is lower) of Error (q1 - q3) and completion times (t_q1 - t_q3) between area plot - heat bands (H_3).

```
##
##   TOST INDEPENDENT SAMPLES T-TEST
##
##   TOST Results
##   -------------------------------------------------
##                                        t          df        p
##   -------------------------------------------------
##      question      t-test            1.46       664       0.145
##                    TOST Upper       -2.15       664       0.016
##                    TOST Lower        5.06       664      < .001
##
##      time          t-test           -1.29       664       0.197
##                    TOST Upper       -4.90       664      < .001
##                    TOST Lower        2.31       664       0.010
##   -------------------------------------------------
##
##
## Equivalence Bounds
## ---------------------------------------------------------------
##                                  Low          High        Lower          Upper
## ---------------------------------------------------------------
## question      Cohen's d        -0.279        0.279
##                 Raw             -0.119        0.119       -0.00625       0.102
##
## time          Cohen's d        -0.279        0.279
##                 Raw             -21.581       21.581      -17.58762      2.134
## ---------------------------------------------------------------
```

Error Rate

Non-inferiority confirmed in q_1, q_2, and q_3.
Equality confirmed in q_2 and q_3.
Area plot is superior in q_1.

Completion Time

Equality (and subsequently non-inferiority) confirmed in q_1, q_2, and q_3.

4 Hypotheses Tested

H_0 Gradient Uncertainty Plot vs. Composite Uncertainty Visualization

Error Rate: Gradient Plot is superior to Composite Uncertainty Visualization
Completion Time: Equality confirmed.

H_0 non-inferiority confirmed, even superiority of gradient plot for errors.

H_1 Gradient Uncertainty Plot vs. Uncertainty Heatmap

Errors: Gradient Plot is superior to Uncertainty Heatmap

Completion Time: Heatmap is superior to Gradient Plot.

H_1 non-inferiority confirmed.

H_2 Gradient Uncertainty Plot vs. Threshold Uncertainty Plot

Errors: Gradient Plot is not significantly better than Threshold Uncertainty Plot, pairs not significant according to post-hoc Nemenyi test (p=0.974).

Completion Time: Gradient Plot is significantly better than Threshold Uncertainty Plot.

H_2 can only be confirmed for completion times.

H_{2a} - Limited Vertical Space

Errors: Friedman Test non-significant

Completion Time: Gradient Plot is significantly better than Threshold Uncertainty Plot.

H_2a is not confirmed for errors, but can again be confirmed for completion times.

H_3 Difference between Heatband and Area Charts Uncertainty

Errors: Equivalence confirmed.
Completion Time: Equivalence confirmed.
H_3 can be con rmed with equivalence.

5 Implications

For Question 1 and 2 comparisons had to be made between segments from one result, meaning that horizontally comparisons could be made well using line charts or heatmaps. However, in Questions 3 to 6, comparison had to be made across segmentation results visualized as rows, which seems to be more difficult when using the Composite Visualization: There were noticable differences in results for Question 3, 4, and 6 where the Gradient Uncertainty Plot outperformed the Composite Visualization (H_0), while times employed using the Gradient Uncertainty Plot were not significantly longer.

Question 4 was aimed to test the effectiveness of uncertainty visualization designs for limited vertical space, in which the Gradient Uncertainty Plot had significantly higher error than the Composite (H_0) and Threshold Uncertainty Visualization (H_2) and Completion Time not inferior to other designs, except for the Uncertainty Heatmap (H_1).

Question 5 had the overall worst error rate, which we infer was due to the difficulty of the question being two very similar segment uncertainties. In this case, the Threshold Uncertainty Plot significantly outperformed the Gradient Uncertainty Plot (H_2) and Uncertainty Heatmap. However, the completion time was still significantly worse than both of these designs. Error were also low for the Gradient Uncertainty Plot, which was out of line with other questions with multiple segmentation results visualized (Question 3-6). Overall, completion times were highest for the Threshold Uncertainty Plot (median completion time: 26s), with the Gradient Uncertainty Plot showing lower completion times (median completion time: 19s).

Two questions in the test were more difficult to answer (Q1, Q5): differences between uncertainty in the segments and areas were smaller than in other questions. Participants took longer to answer these questions, and had worse error rates compared to similar questions:

Question 1 and 2 are similar, horizontal intervals must be compared:

- Mean Error Q1: 0.277027, Q2: 0.1036036
- Median Completion Time Q1: 29, Q2: 12

Question 4 and 5 are similar, horizontal and vertical comparison with vertical space available:

- Mean Error Q4: 0.2387387, Q5: 0.6779279
- Median Completion Time Q4: 18, Q5: 23

Question 5 even had error rates above 50%, except for the Uncertainty Threshold Plot. This implies that the aggregated uncertainty of an interval is hard to judge mentally and without visual support. We suggest employing an explicit aggregated uncertainty visualization.