Evaluation of Noise Annotation Lines: Using Noise to Represent Thematic Uncertainty in Maps

Christoph Kinkeldey*, Jennifer Mason**, Alexander Klippel**, Jochen Schiewe*

** Lab for Geoinformatics and Geovisualization (g2lab), HafenCity University Hamburg, Germany*

*** GeoVISTA Center, Department of Geography, The Pennsylvania State University, USA*

Noise annotation lines are a promising technique to visualize thematic uncertainty in maps. However, their potential has not yet been evaluated in user studies. In two experiments we assessed the usability of this technique with respect to a different number of uncertainty levels as well as the influence of two design aspects of noise annotation lines: the grain and the width of the noise grid. We conducted a web-based study utilizing a qualitative comparison of two areas in 150 different maps. We recruited participants from Amazon Mechanical Turk with the majority being non-experts with respect to the use of maps.

Our findings suggest that for qualitative comparisons of 'constant uncertainty' (i.e., constant uncertainty per area) in thematic maps, noise annotation lines can be used for up to 6 uncertainty levels. During comparison of 4, 6, and 8 levels, the different designs of the technique yielded significantly different accuracies. We propose to use the 'large noise width, coarse grain' design that was most successful. For 'mixed uncertainty' (i.e., uncertainty is not constant per area) we observed a significant decrease in accuracy, but for 4 levels the technique can still be recommended.

This article is a follow-up to our conference paper reporting on preliminary results of the first of the two experiments (Kinkeldey et al. 2013).

Keywords: uncertainty, geovisualization, user evaluation, usability, AMT

1. Introduction

Uncertainty is inherent in all geospatial data arising from various sources such as measurement errors and inaccuracies, model ambiguity, and vagueness, or loss of quality during processing of the data (Atkinson and Foody 2002; Heuvelink and Brown 2008; Shi 2010; Zhang and Goodchild 2002). With many applications, ignoring uncertainty can result in misleading or unusable results: 'Error-laden data, used without consideration of their intrinsic uncertainty, are highly likely to lead to information of dubious value' (Zhang and Goodchild 2002, p. 3). Past research suggested that communicating uncertainty through visualization can support analyses and decision-making (Deitrick and Edsall 2006; Hope and Hunter 2007) and can increase analysts' trust in their results (Fisher et al. 2012). A variety of methods for visualizing uncertainty exist, especially in the area of GIScience and scientific visualization (MacEachren et al. 2005; Pang 2008; Brodlie et al. 2012). Typically, uncertainty is categorized by type, i.e., thematic (also: attribute), geometric (positional), or temporal uncertainty. To display different types of uncertainty, different approaches can be combined, e.g., integrated views or adjacent views, static or dynamic approaches, the use of interaction, etc. This results in a high number of possible approaches and it can be difficult to choose a suitable technique for a specific application. Different typologies for uncertainty visualization exist (Thomson et al. 2005; Senaratne et al. 2012) but mainly focus on data characteristics (dimensionality, type etc.) and do not account for other aspects such as the tasks involved. Thus, they can only offer limited support for the selection of suitable techniques. Besides typologies, other categorizations of the techniques can also be helpful, for instance the distinction between intrinsic and extrinsic approaches (Gershon 1998). Intrinsic approaches utilize visual variables from existing objects in the visualization to represent uncertainty, mostly including visual variables from cartography. In addition to the seven visual variables described by Bertin (1983), variables including symbol focus and clarity are used for intrinsic displays (MacEachren 1992; MacGranaghan 1993). Extrinsic approaches, on the other hand, incorporate additional graphical objects to represent uncertainty, e.g., glyphs (Pang 2001) or other objects such as bars or dials that are added to the display. Unlike most intrinsic variables, they can be visually separated from the other content.

In this research we evaluate an extrinsic method we term *noise annotation lines,* a method first described as procedural annotations by Cedilnik and Rheingans (2000). The technique is a promising way to display thematic uncertainty in maps that involve heterogeneous geometries, e.g., land cover maps (Figure 1). This work contributes to the evaluation of extrinsic uncertainty visualization methods by testing usability aspects of noise annotation lines, with a focus on the impact of design factors on the usability of the method. After discussing related work (section 2) we report the results from two web-based experiments (section 3) and discuss the implications of the experiments in section 4. In the last section we conclude our findings and the limitations of the experiment and provide suggestions for future evaluation in this field.

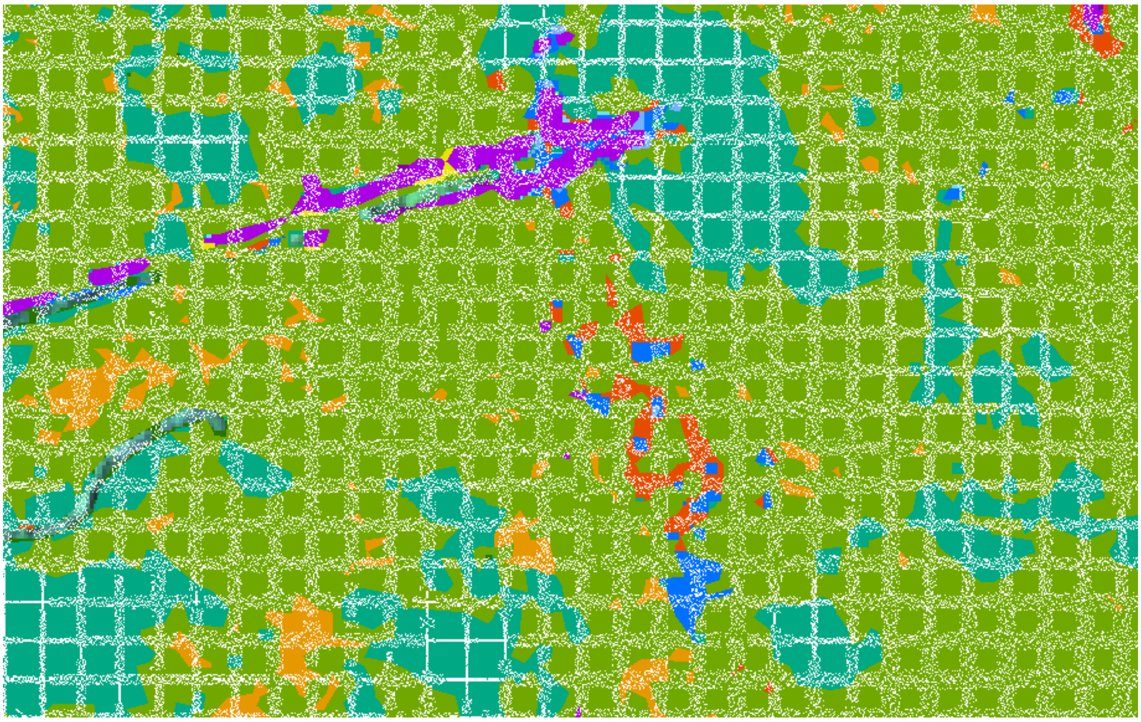

Figure 1. Noise annotation lines representing classification uncertainty of a vegetation land cover map. The local width of the noise grid indicates the degree of uncertainty (larger noise width represents a higher degree of uncertainty).

2. Related Work

So far there has been very little research to investigate the usability of the noise annotation lines technique. An exception is a qualitative assessment by Zuk and Carpendale (2006) who evaluated noise annotations along with the three other types suggested by Cedilnik and Rheingans (2000): width, sharpness, and amplitude. This was done from a theoretical standpoint using heuristics from theories introduced by Bertin, Tufte, and Ware. Zuk and Carpendale point out that the data-ink ratio of a noise grid is relatively small, compared with other annotation types (e.g., the amplitude grid). Thus, they hypothesize that the noise grid may not be able to represent as many uncertainty levels as the other types but they do not provide further evidence for this assumption. They remark that more formal testing of procedural annotations, especially concerning perceptual aspects, is needed.

Similar to the work we present here, a number of studies by Kardos and colleagues evaluate a technique called ‘trustree’ that depicts uncertainty in census maps by varying the level of detail locally (Kardos et al. 2007; Kardos et al. 2008). They found that the visual metaphor of ‘detail’, i.e., a coarser grid in uncertain areas is more usable than a metaphor of ‘clutter’ that represents uncertain areas with a finer grid.

Despite the variety of methods for representing uncertainty, a systematic evaluation of their usability is still needed (MacEachren et al. 2005). Studies on extrinsic approaches deal with various data and display types, including multivariate vector glyphs (Wittenbrink et al. 1996) or 3D displays (Newman and Lee 2004). Thus, the results do not always help in choosing a suitable technique for 2D maps since the requirements for displaying uncertainty differ. For instance, in 3D environments the question of clutter through occlusion in different perspectives is important but does not exist with 2D maps.

3. Evaluation of Noise Annotation Lines

Many thematic maps such as land cover maps contain objects of high geometric variability, that is, objects that differ considerably in size and shape. Representing uncertainty integrated into such maps is challenging compared to maps with more homogeneous objects such as choropleth maps. For maps with geometrically diverse areas, extrinsic methods, especially those based on uniform grids, seem promising because they are independent of the underlying geometry. *Noise annotation lines* (Kinkeldey et al. 2013) are a grid-based extrinsic method. A regular grid is placed onto the map and is altered locally to represent the degree of uncertainty (Figure 2). Cedilnik and Rheingans (2000) proposed four different versions of annotations: variation of width, sharpness, noise, and amplitude. We focus here on the noise grid because we expected noise to be a particularly suitable metaphor for uncertainty. This was substantiated in a qualitative pre-test where people found the noise display particularly intuitive (Kinkeldey and Schiewe 2012). Additionally, recent research suggests 'noise was seen as a promising graphic variable, worth further investigation' (Vullings et al. 2013). The noise grid is varied in size locally to represent the level of uncertainty: the more uncertain the underlying content, the more scattered the noise line. Regardless of the level of uncertainty, the number of noise particles and their size (grain) remain constant. An important characteristic of this approach is that it only represents the values beneath the lines, where the values in the cells of the grid are not depicted thus showing a generalization of the uncertainty data. However, since the size of the grid cells can be varied according to the scale of the map, a compromise can be made between maximum coverage of uncertainty data and minimum occlusion of the underlying content. Related research suggested that '[f]rom a design standpoint, selection of an appropriate number of levels should be guided first by task demands, such as the level of detail necessary for people to differentiate among potential actions. Information should not be displayed at a greater level of detail than is required by the task.' (Bisantz et al. 2009, p. 78). This could be one of the potential advantages of noise annotation lines and makes this method promising for use in maps.

We plan to use the noise annotation lines approach for exploratory analyses of change in land cover maps. Understanding general changes of land cover does not require precise values of uncertainty, rather a qualitative estimation to compare general differences. Thus, we tested the qualitative comparison of uncertainty between different areas, i.e., no specific uncertainty values had to be retrieved from the display and no legend was provided.

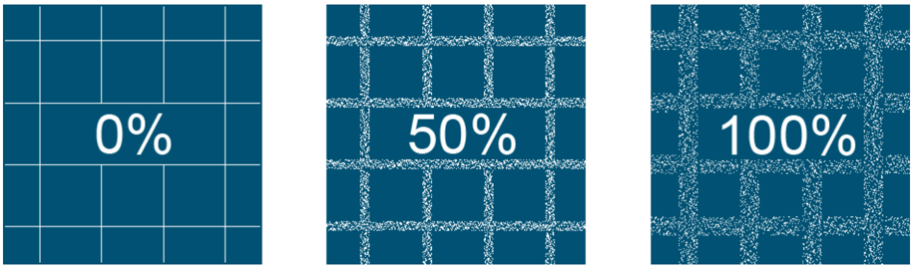

Figure 2. Variation of noise width to represent uncertainty: the higher the uncertainty the larger the with of the noise grid.

Experiment 1

Research questions

The goal of this experiment was to make a first estimation about the usability of noise annotation lines to represent thematic uncertainty in maps. As there are different ways to display the noise grid, the first question that emerged was whether changes in the design affect user performance when areas are compared in terms of their uncertainty. The second question was how many levels of uncertainty could be easily distinguished, and third, if and how user performance changes when uncertainty does not remain constant over the areas in the map (mixed uncertainty). These broad questions led to the following specific research questions:

1. How do different design parameters impact the usability of noise annotation lines as a representation of thematic uncertainty in a land cover map?

2. How does the number of uncertainty levels affect user performance?

3. Can users accurately compare the overall degree of uncertainty between two defined areas when the values vary within the areas (mixed uncertainty)?

Variables

The appearance of noise annotation lines can be changed by altering different design parameters. Our hypothesis was that these variations will impact the effectiveness and efficiency of the uncertainty display. The following three major design parameters were chosen for evaluation: the number of uncertainty levels, the width of the noise grid, and its grain (Table 1).

Table 1. Factors used in experiment 1.

Factor	**Number of levels**	**Levels**
Noise width	2	Small (40%), Large (50%)
Noise grain	2	Fine (1x1), Coarse (2x2)
Uncertainty levels	3	4, 5, 6 levels

The width of the noise grid was defined with respect to the size of the grid cells (Figure 3). With a smaller noise width the grid covers less area, and there is less space to represent different levels of uncertainty. Consequently, the choice of this parameter was a compromise between the visual interference of the grid with the underlying content and the number of levels that are discernible. If the grid width is too large, the grid occludes more of the underlying content and the structure of the noise grid is not preserved. This effect already occurs with a noise width of 60% of the grid cell size. On the other hand, if the width is too small, it limits the number of uncertainty levels to be discerned. Thus, we chose 40% and 50% of the grid cell size as levels to assess for this factor.

The grain of the noise particles was the second design parameter we manipulated (Figure 4). A finer grid consists of a higher number of small particles, while a coarse grid contains fewer, but larger particles. Since we kept a constant pixel resolution across all maps in this study we implemented 1x1 pixels and 2x2 pixels for the factor ‘grain’.

For the third factor, we varied the number of uncertainty levels. We chose a minimum of 4 levels because a pre-test revealed that three levels (0%, 50%, and 100% uncertainty) are straightforward to discern in contrast to 4 levels that already led to errors in uncertainty value retrieval. We hypothesized that a variation up to 6 levels would be appropriate to determine the limit of levels that people are able to compare. This resulted in 4, 5, and 6 uncertainty levels so subjects had to discern intervals of 33%, 25%, and 20% uncertainty (Table 2).

As dependent variables, response accuracy and the time the subjects needed for each question were measured during the experiment.

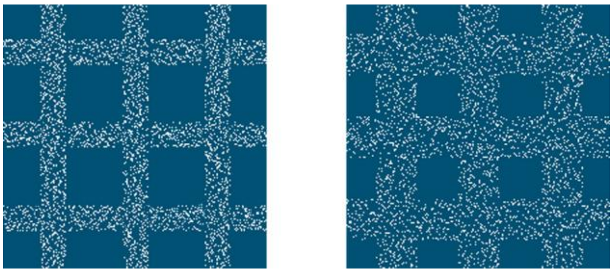

Figure 3. Design parameter ‘noise width’. Both grids represent the same degree of uncertainty (100%), but with different widths: 40% (left) and 50% (right) of the grid cell size.

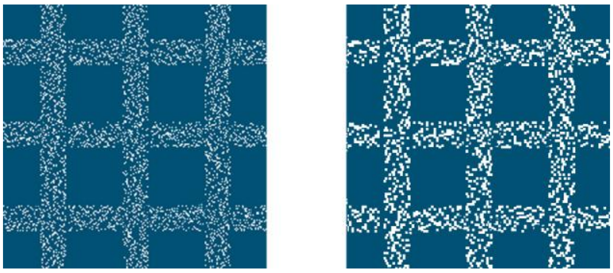

Figure 4. Design parameter ‘noise grain’. Both grids represent the same degree of uncertainty (100%), but with different grain sizes: ‘fine’ (left) and ‘coarse’ (right).

Table 2. Factor ‘uncertainty levels’ in experiment 1.

Number of Levels	**Interval**	**Levels**
4 levels	33%	0%, 33%, 66%, 100%
5 levels	25%	0%, 25%, 50%, 75%, 100%
6 levels	20%	0%, 20%, 40%, 60%, 80%, 100%

Task

For the main part of the study we chose a uniform task for all maps: a comparison of uncertainty between two areas. This was done in a qualitative way (no specific values had to be retrieved) because during real world analyses it is rarely the case that the user needs exact uncertainty values. Instead, with most of the tasks it is more important to know how uncertainty of one area relates to uncertainty in a different area. For example, if an area is classified as water body, high uncertainty can be evidence for a misclassification. But this is just the case if uncertainty is relatively higher than with other areas of this type - the specific amount is not of interest. Thus, we asked participants to compare uncertainty between two marked areas: ‘A’ and ‘B’. The question and possible answers remained the same for all maps: ‘Which area is more uncertain?’ Potential answers included ‘A is more uncertain’, ‘B is more uncertain’, ‘A and B are equal’ and ‘I can’t tell’. Since the questions were mandatory the latter answer was included to minimize nonsense answers when participants could not read the map or had little confidence in their answers.

Stimuli

We created ten different maps per factor combination to establish ten repetitions. All maps were taken from the same vegetation land cover map representing equally sized areas (100 m x 100 m) at the same scale. Furthermore, the size of the noise grid cells in all maps was the same (4 m). We varied the background colors according to a qualitative color scheme recommended by ColorBrewer (Brewer et al. 2003). The utilized color scheme (‘Paired’) is indicated to be colorblind-safe and laptop-/LCD-friendly. In each map, two square areas in the size of 3 x 3 noise grid cells were drawn on areas of the same color and labeled ‘A’ and ‘B’ (Figure 5). The value within the areas was either equal or differed by one level, e.g., 66% vs. 100% with 4 levels or 60% vs. 40% with 6 levels. We placed the squares on areas with the same background color, either light blue or light green. These two colors have a very similar contrast distance from the white color of the grid. Hence, we varied the color but not the contrast between grid lines and the background. Additionally, the low contrast between grid and background assures we evaluate the more critical point to find the limits of this technique.

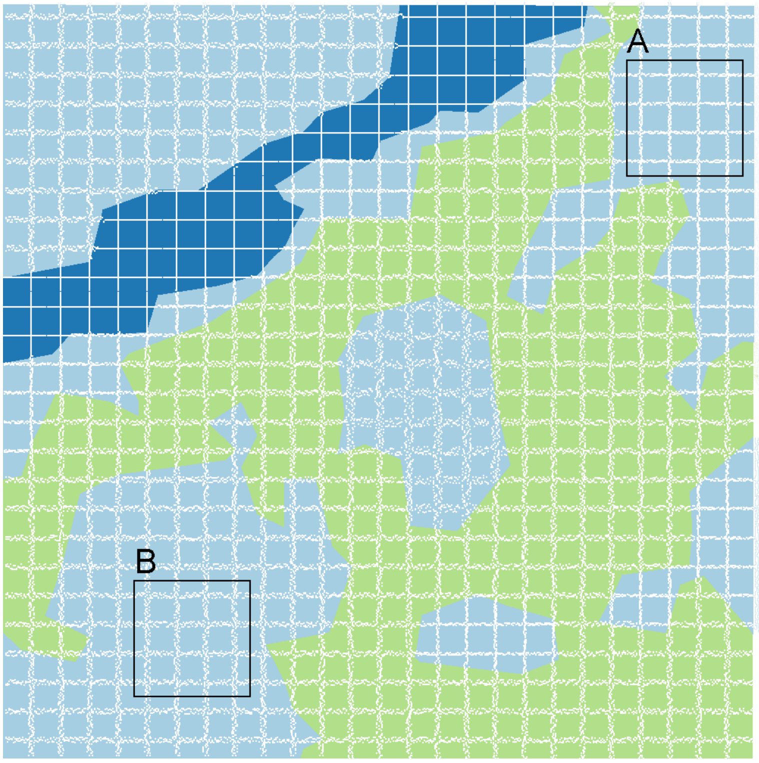

Figure 5. Example map representing constant uncertainty (uncertainty remains constant per map region).

In order to answer the third research question (constant versus mixed uncertainty), we included maps showing a mixed representation of uncertainty. This means, in contrast to the constant case, uncertainty values do not remain constant within each area (Figure 6). Thus, uncertainty is not constant in the marked regions A and B either. For the mixed uncertainty case, we did not involve all combinations as with the constant case. In order to keep the number of combinations low we only varied the number of uncertainty levels and not noise width and grain. The three levels (4, 5, and 6 uncertainty levels) were repeated ten times, resulting in 30 maps. Overall, each participant answered 150 (120 constant + 30 mixed uncertainty) questions in the main portion. In the survey, the maps with constant uncertainty and those with mixed uncertainty were shown in a randomized order.

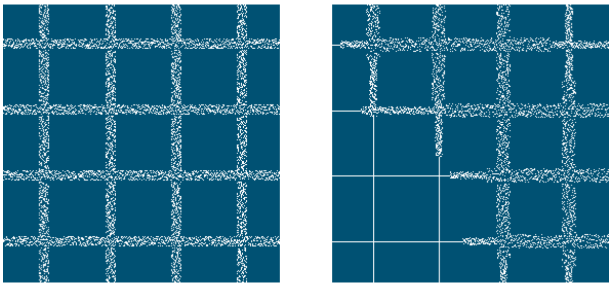

Figure 6. Constant (left) vs. mixed uncertainty distribution (right).

We conducted the experiment as a web-based survey, making it possible to recruit participants via Amazon Mechanical Turk. Web-based experiments have proven to be suitable for evaluating usability aspects of uncertainty visualization (Aerts et al. 2003). There have been substantial review articles by Rand (2012) and Crump et al. (2013) showing that behavioral experiments via AMT yield essentially the same results as those in controlled environments. Crump et al. (2013) were able to replicate findings of popular reaction times tasks requiring participant attention in AMT to that of traditional laboratory settings. In an online cognitive and perceptual experiment by Germine et al. (2012), they found that their data from challenging timed tasks were similar in quality to performance data that can be gathered in a laboratory setting despite being anonymous, uncompensated, and unsupervised. Additionally, guidelines exist that describe how experiments based on Amazon Mechanical Turk can be designed to yield valid results (Mason and Suri 2012). In our case we were aware we could not control the display type, color calibration, or distractions that potentially influence the participant; that is why our experiments are similar to field experiments rather than controlled lab experiments. But this fact could even make the results more valid for the use in real world applications.

Survey

The survey comprised of the following parts:

1) Introduction: The participants were provided with an introductory explanation of

uncertainty and noise annotation lines. Three figures of noise annotation lines were shown (no, medium and high uncertainty) to clarify the method. We also included a note to not use a smartphone or similar device and when using a tablet, to not zoom in and out to make sure that all subjects see each map in its entirety when answering the questions.

2) Personal information: We asked for gender, age and a self-assessment in terms of experience with uncertainty visualization in maps.

3) Maps: The main section of the study showed the 150 maps in succession including its uncertainty. In each map, the two areas A and B were compared. In order to avoid bias and learning effects, we randomized the order of the questions.

4) Comments: An opportunity to provide feedback on the survey.

All questions except the comments at the end were mandatory. We used LimeSurvey (http://www.limesurvey.org), survey software freely available under an open source license. We decided to use version 1.92+ since we noticed problems with the randomization that occurred in the latest version (2.05).

Participants

We recruited participants using the online crowdsourcing service Amazon Mechanical Turk (AMT, http://mturk.com). The reasons for utilizing this service are threefold: First, it is efficient to recruit subjects, second, we aimed to obtain participants with different backgrounds and expertise (not only from our domain) and third, paid participants were likely to be motivated to finish the survey even though it took 20 to 30 minutes. Participants were reimbursed with $0.50 for their participation. Among the 32 participants, 17 declared themselves as female and 15 as male. Regarding age, most of the participants classified themselves between 20 and 29 years old (16/32), followed by 50 to 59 years (8/32) and 30 to

39 (5/32). Very young people and the group 40 to 49 years were barely represented. People from 60 years and older did not participate at all.

Concerning the subjects' experience with uncertainty maps, we asked three questions: If they had known about the concept of uncertainty before, how often they used maps, and if they had seen a map including uncertainty information before. From the three answers we determined a level of experience per participant (little, average, extensive experience). More than half of the participants (18 out of 32) had little experience while roughly one-quarter had average experience (7/32) or extensive experience (7/32) with uncertainty maps. This is not surprising as one can expect that participants acquired via Amazon Mechanical Turk will be primarily lay people.

Results

We obtained the results through rounds of ten participants and checked the integrity and validity of the data after each step. In the end the turnout was 32 because of incomplete replies that we ignored. Since there were ten maps for each combination in our 3x2x2 factorial design we had 150 answers from each subject, totaling 4800 single answers in the main section. In case a participant responded that he or she was not able to provide an answer, we treated this as a missing value. Since there were only 96 missing values out of 4800 (2%) we made the assumption that sufficient responses were collected for each participant and map. We computed accuracy for each participant and factor combination as the percentage of correct answers.

Figure 7 shows the mean accuracy and the standard error for the maps with constant uncertainty. The charts are grouped by combination of the factors 'noise width' and 'noise grain' and each chart shows the accuracies for 4, 5, and 6 uncertainty levels. Generally, for all factor combinations, the mean values are higher than 76%. The 'small noise width' conditions (two charts on the left) show a stronger trend of decreasing accuracy with an increase of uncertainty levels from 4 to 5. In the ‘large noise width’ conditions (two charts on the right) the accuracy values are, compared to the other conditions, lower for 4 levels and roughly equal for 5 levels, but increasing for 6 levels again, especially for the coarse grain grid (increase from 77% for 5 levels to 83% for 6 levels).

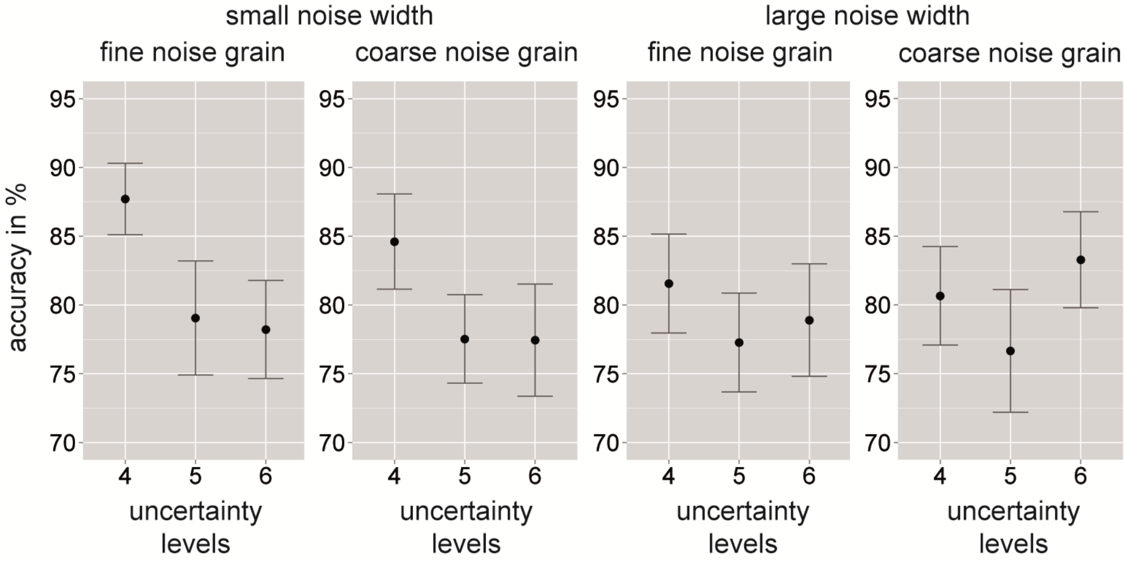

Figure 7. Experiment 1: Accuracy for constant uncertainty (mean and standard error).

Repeated measures ANOVA with the three factors revealed the following: Mauchly’s test of sphericity showed that the assumption of sphericity is violated for the factor ‘uncertainty levels’ ($\chi^2(2) = 7.23$, p = .27) and the interaction of ‘uncertainty levels’ and ‘noise grain’ ($\chi^2(2) = 8.22$, p = .16). Hence, we used the Greenhouse-Geisser correction as suggested by Tabachnick and Fidell (2007). Of the three main effects, only ‘uncertainty levels’ is statistically significant (F(1.647,51.065)=7.024, p=.004, $\eta^2 = .185$): the accuracy decreases with a higher number of levels. The factors ‘noise width’ and ‘noise grain’ do not significantly change user performance, however, there is a statistically significant interaction effect of ‘uncertainty levels’ and ‘noise width’ (F(1.960,60.747)=7.295, p=.002, $\eta^2 = .19$).

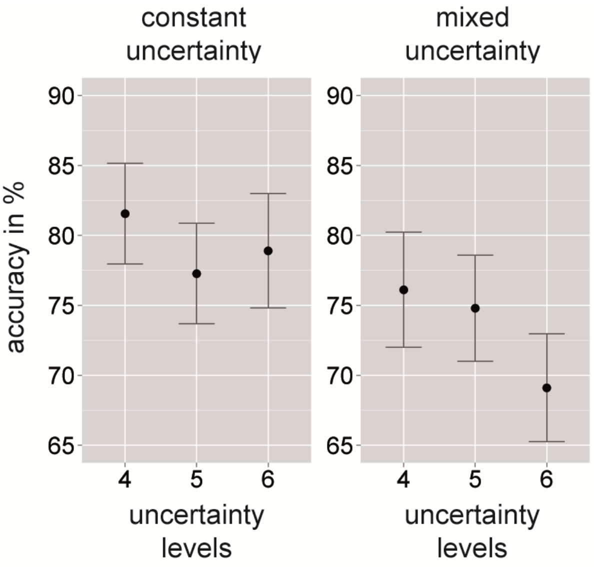

Figure 8. Experiment 1: Accuracy for constant vs. mixed uncertainty (mean and standard error) for the factor combination ‘large noise width’, ‘fine noise grain’.

Our third research question addressed the change from constant to mixed uncertainty data (Figure 6) and its effect on user performance. We did not vary noise grain and width for this comparison and selected a salient combination (‘fine grain, large width’) visualizing 4, 5, or 6 uncertainty levels. A graphical comparison of accuracy between constant and mixed case can be found in Figure 8. The most obvious difference is that with mixed data, accuracy is generally lower than in the constant case. In contrast to constant uncertainty the results for the mixed case show a consistent decrease of accuracy with growing number of uncertainty levels.

Repeated measures ANOVA provided further insight. Mauchly’s test indicated that only the main effect of uncertainty levels is statistically significant ($\chi^2(2) = 8.56$, p = .014). Therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity. The main effect ‘constant vs. mixed’ is statistically significant (F(1,31)=10.59, p=.003, η^2 = .255). Likewise, the main factor ‘uncertainty levels’ is statistically significant (F(1.60,49.68)=3.912, p=.035, η^2 = .112) but there is no significant interaction between ‘constant vs. mixed’ and ‘uncertainty levels’ (F(2,62)=1.74, p=.183). Hence, the results for the number of uncertainty levels (the more uncertainty levels the lower the accuracy) is present in this comparison, too, and making judgments in which multiple levels of uncertainty (mixed case) have to be taken into consideration as further decreasing the accuracy of user responses.

Experiment 2

In the first experiment the use of 4, 5, and 6 uncertainty levels yielded relatively high accuracy, even with 6 levels (>76% mean accuracy with a maximum standard error of 4.5%). In order to determine the limitations of noise annotation lines with respect to the number of uncertainty levels, we repeated the experiment with up to 8 levels. The setup was the same as with experiment 1, that is, the structure and the task were not changed. However, we removed the maps with 5 levels and added maps using 8 uncertainty levels. Thus, the uncertainty levels changed to 4, 6 and 8 (Table 3). The other two factors 'noise width' and 'noise grain' remained the same as in experiment 1.

Table 3. Levels for factor 'uncertainty levels' in experiment 2.

Uncertainty levels	Step	Levels
4 levels	33%	0%, 33%, 66%, 100%
6 levels	20%	0%, 20%, 40%, 60%, 80%, 100%
8 levels	14%	0%, 14%, 28%, 43%, 57%, 71%, 86%, 100%

Participants

We recruited participants in the same way as in experiment 1, i.e., via Amazon Mechanical Turk offering the same amount for a complete set of answers. We had the same number of responses as in experiment 1 (32 full responses) after sorting out incomplete datasets.

There were more female participants (18) than male (14). The distribution of age was similar to experiment 1; most of the participants classified themselves to be between 20 and 29 years old (14/32), followed by 30 to 39 (8/32) and 40 to 49 years (6/32). Very young (below 20) and older respondents (over 60) were not represented. Self-assessment of experience resulted in 14 subjects with little, 11 with average, and 7 with extensive experience interacting with uncertainty in maps. Hence, the overall experience was higher than with experiment 1 where half of the subjects had little experience and fewer subjects in the average experience group.

Results

Figure 9 shows the mean accuracies for all factor combinations. Generally, accuracy is lower than in experiment 1 (between 4% and 12% lower for 4 levels and 2.5% and 9.5% for 6 levels). While most results were similar to experiment 1 there were a number of participants in experiment 2 with constantly low accuracy. Since we could not determine the reason for their low performance we did not treat them as outliers.

As expected, the decrease in accuracy with more uncertainty levels than 6 is higher: for 8 levels, accuracies are generally lower than with the same design using 4 or 6 levels. For example, in the ‘small noise width’, ‘fine noise grain’ condition, accuracy decreases from 75.6% (4 levels) to 60.7% (8 levels). Again, the magnitude of change depends on the design of the grid: the ‘coarse grain, large width’ condition results in higher accuracy values (76.9% for 4 levels to 70.9% for 8 levels).

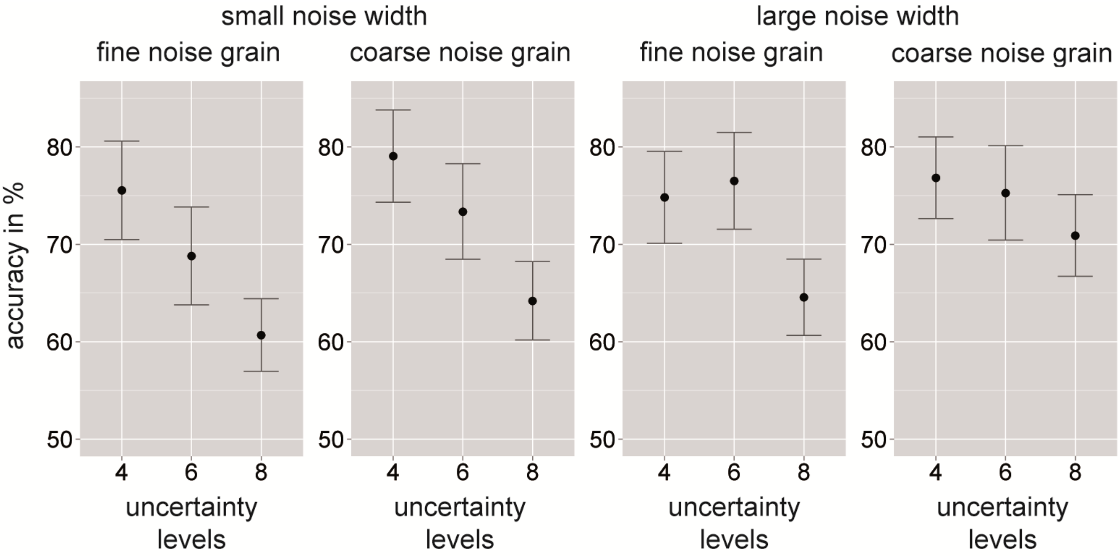

Figure 9. Experiment 2: Accuracy for constant uncertainty (mean and standard error).

Repeated measures ANOVA showed that Mauchly's test of sphericity is not significant for experiment 2, so we assumed sphericity. In contrast to experiment 1, results suggested that all three factors have a statistically significant impact on accuracy: 'uncertainty levels' (F(1.993,61.777)=15.337, p<0.001, η^2 = .331), 'noise width' (F(1,31)=5.958, p=.021, η^2 = .161), and 'noise grain' (F(1,31)=6.039, p=.02, η^2 = .163). Additionally, we found a statistically significant interaction effect between 'uncertainty levels' and 'noise width' (F(2,62)=3.875, p=.026, η^2 = .111). Contrast revealed that this interaction effect is specific to the change from 4 to 6 uncertainty levels but did not apply to 8 uncertainty levels. These results largely reflect experiment 1 showing that more uncertainty levels lead to lower accuracy. The expected additional challenge using 4, 6, and 8 uncertainty levels is nicely reflected in the increased η^2 for the main effect 'uncertainty levels'. It becomes obvious that in more challenging situations, design characteristics become more important and can support reasoning with uncertainty: salient visual characteristics can off-set the negative effect of higher numbers of uncertainty levels.

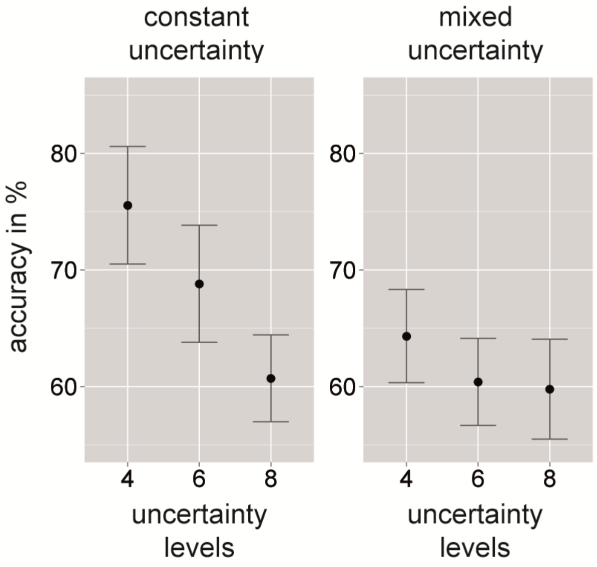

Figure 10. Experiment 2: Accuracy of constant vs. mixed uncertainty (mean and standard error) for the factor combination 'large noise width', 'fine noise grain'.

Comparing constant and mixed uncertainty representations, accuracy values for the mixed case were lower than for the maps that depicted constant uncertainty (Figure 10). Mauchly's test was not significant, thus we assumed sphericity. ANOVA revealed that uncertainty levels were again significant (F(2,62)=7.444, p=.001, η^2 = .194): more levels result in lower accuracy. Additionally, there is a significant difference between constant and mixed uncertainty (F(1,31)=15.255, p<.001, η^2 = .330), indicating the difficulties participants have when uncertainty is not constant per area (mixed). There are no significant interaction effects.

4. Discussion

The results from the two experiments offer insights into various aspects of uncertainty visualization. The first thing to note is that, as expected, the *number of levels* has an influence on people's abilities to make judgments about uncertainty. Simply put, the fewer levels participants have to distinguish, the better their performance was in terms of correct answers. This result can be explained by numerous studies in both perception and cognition literature that indicate that the more information that has to be distinguished and kept in working memory, the more difficult it is to reason with this information (Lloyd and Bunch 2005).

The downward trend of user performance can be halted using appropriate visualizations, which is an important finding for research on visualizing uncertainty. Experiment 2 revealed that when comparing 4, 6, and 8 uncertainty levels, the two design parameters of the grid had a significant impact on the decrease of user performance. The most visually salient design of the grid (large noise width and coarse grain) was able to leverage the negative effect of an increased number of uncertainty levels (as revealed by the significant interaction effects and the graph in Figure 9). In this combination, accuracies in comparing the two areas were higher than expected: even for 8 uncertainty levels we observed a mean accuracy of 71% compared to 61% with the small width, fine grain design. The more levels to distinguish, the more important it is to use a salient design for the grid. For 6 levels we observed a mean accuracy of more than 75%. Taking into account that we evaluated the case of very low contrast to the background, we see this as recommendation for the use of noise annotation lines with up to 6 uncertainty levels.

The comparison of constant and mixed visualizations of uncertainty shows that accuracy with mixed uncertainty is generally lower. This confirms the assumption that the more complex the information is that is offered to participants, the more errors they make. We only compared one factor combination (fine grain and large width) for constant and mixed uncertainty visualization and expected that the increase in the number of levels would lead to lower user performance. Regarding the dependency on the number of uncertainty levels the two experiments show ambiguous results. In experiment 1 (4, 5, 6 levels) only the accuracy values for mixed uncertainty decreased with more uncertainty levels. In experiment 2 (4, 6, 8 levels) accuracy decreased in both cases (constant and mixed) but the extent was higher with constant uncertainty. Hence, mixed uncertainty generally leads to lower accuracy but with up to 8 levels the effect of decreasing accuracy is weaker. This can be explained by the implicit difference between the way of comparing constant and mixed uncertainty: In the constant case the comparison task will be successful if the noise grid in the two areas can be visually distinguished. With mixed data subjects have to estimate the overall uncertainty in both areas first and subsequently compare them. The estimation part of the task could be less influenced by the number of uncertainty levels than the direct comparison lessening the effect of decreased accuracy with more uncertainty levels.

Given the difficulty of controlling time in web-based studies (compared to lab experiments) it does not come as a surprise that no significant effects related to response times could be observed. For instance we recognized that the loading times of the maps varied a lot depending on the Internet connection. We tried to minimize these effects by compressing the map images but there were still differences.

5. Conclusions and Outlook

We have presented a study to evaluate noise annotation lines for visualizing thematic uncertainty. In a web-based survey, subjects compared the uncertainty of two equally-sized areas A and B. This was done for constant and mixed uncertainty representations (see section 3). From the experiments, the following results can be concluded:

- the number of uncertainty levels has a significant influence on the participants' judgment (generally, with more levels, user performance decreases),
- the variation of the design parameters 'noise width' and 'noise grain' has a significant impact on user performance with a higher number of (up to 8) levels, meaning that
- the decrease of user performance when more uncertainty levels need to be distinguished can be counterbalanced with changes in the design of the noise grid,
- more complex uncertainty information ('mixed' uncertainty) leads to a significant decrease in user performance, and,
- there are no significant effects with respect to response time.

All in all, we could show the potential of noise annotation lines for the representation of thematic uncertainty in qualitative comparison tasks. We can recommend their use in maps that are geometrically diverse or already make extensive use of color so that intrinsic techniques are hard to apply. Our findings show that the technique can successfully be used in qualitative analyses of up to 6 uncertainty levels when more salient grid designs are used.

There are several *limitations* worth mentioning with respect to the two experiments. First, we did not account for the impact of noise annotation lines in the readability of the map, i.e., we did not evaluate if the background could still be read despite the noise grid. Second, we did not evaluate the influence of background colors with different contrast levels. Third, the fact that response times did not show any effect could be due to our setup. There was no time pressure or incentive for quick replies.

Based on the limitations we recommend the following aspects as *future work*. The first one is the evaluation of response time. As already discussed above, our experiments did not show any effects regarding time. For applications in which quick comparisons are of importance this could be of high interest, for instance in dynamic decision making (Kobus et al. 2001). We are convinced that there are alternative experimental designs that could help reveal time effects, e.g., the use of incentives for quick answers combined with a reward for the most accurate responses. A second aspect worth investigating is the intuitiveness of noise annotation lines. So far we experienced that when people see noise annotation lines for the first time they seem to understand the concept very quickly. This could be a potential strength of the technique and has not been evaluated systematically so far. A third aspect worth looking at is the comparison with other types of annotations. Zuk and Carpendale (2006) assumed that with a different type of annotation, a grid representing uncertainty by amplitude of a sine curve, more levels of uncertainty may be discerned than with a noise grid (because of the higher data-to-ink-ratio). It would be interesting to evaluate if this assumption is true and how much difference there is. Finally, making the grid size interactively adaptable by the user is a promising approach for the use of noise annotation lines in exploratory analyses. When the data is unknown the user can reveal patterns in the data by altering the grid width. This may open this approach for other usages than the representation of uncertainty – for explorative analysis of other data, e.g., pollution or population density, it seems promising as well.

References

Aerts, J.C., Clarke, K.C., and A.D. Keuper. 2003. Testing Popular Visualization Techniques for Representing Model Uncertainty. *Cartography and Geographic Information Science*, 30 (3), 249–261.

Atkinson, P.M., and G.M. Foody. 2002. “Uncertainty in Remote Sensing and GIS: Fundamentals.” In *Uncertainty in remote sensing and GIS*, edited by G.M. Foody and P.M. Atkinson. Hoboken, NJ: J. Wiley.

Bertin, J. 1983. *Semiology of graphics*. University of Wisconsin Press.

Bisantz, A.M., Stone, R. T., Pfautz, J., Fouse, A., Farry, M., Roth, E., Nagy, A. L., and Gina Thomas. 2009. “Visual Representations of Meta-Information.” *Journal of Cognitive Engineering and Decision Making,* 3 (1), 67–91.

Brewer, C.A., Hatchard, G.W., and M.A. Harrower. 2003. “ColorBrewer in Print: A Catalog of Color Schemes for Maps.” *Cartography and Geographic Information Science,* 30 (1), 5–32.

Brodlie, K., Allendes Osorio, R., and A. Lopes. 2012. “A Review of Uncertainty in Data Visualization.” In *Expanding the Frontiers of Visual Analytics and Visualization,* edited by J. Dill, R. Earnshaw, D. Kasik, J. Vince, and P. C. Wong. Springer London, 81–109.

Cedilnik, A., and P. Rheingans. 2000. “Procedural Annotation of Uncertain Information.” In *Proceedings of IEEE Visualization 2000,* 77–84.

Crump, M. J. C., McDonnell, J. V., Gureckis, T. M., and S. Gilbert. 2013. “Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research.” *PLoS ONE, 8*(3), e57410.

Deitrick, S., and R. Edsall. 2006. “The Influence of Uncertainty Visualization on Decision Making: An Empirical Evaluation.” In *Progress in Spatial Data Handling:* Springer Berlin Heidelberg, 719–738.

Fisher, D., Popov, I. O., Drucker, S. M., and m. c. schraefel. 2012. “Trust me, i’m partially right: incremental visualization lets analysts explore large datasets faster.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing System,* 2012, 1673–1682.

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., and J. B. Wilmer. 2012. “Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments.” *Psychonomic Bulletin & Review* 19 (5), S. 847-857.

Gershon, N. 1998. “Visualization of an Imperfect World.” *IEEE Computer Graphics and Applications,* 18 (4), 43-45.

Heuvelink, Gerard B. M., and J. D. Brown. 2008. “Uncertain Environmental Variables in GIS.” In *Encyclopedia of GIS,* edited by S. Shekhar, and H. Xiong. Springer US, 1184-1189.

Hope, S., and G. J. Hunter. 2007. “Testing the Effects of Positional Uncertainty on Spatial Decision-Making.” *International Journal of Geographical Information Science,* 21 (6), 645-665.

Kardos, J., Benwell, G., and A. Moore. 2007. “Assessing different approaches to visualise spatial and attribute uncertainty in socioeconomic data using the hexagonal or rhombus (HoR) trustree.” *Computers, Environment and Urban Systems,* 31 (1), 91–106.

Kardos, J., Moore, A., and G. Benwell. 2008. “Exploring Tessellation Metaphors in the Display of Geographical Uncertainty.” In *Geospatial Vision,* edited by A. Moore, and I. Drecki, Springer Berlin Heidelberg, 113-140.

Kinkeldey, C., Mason, J., Klippel, A., and J. Schiewe. 2013. “Assessing the Impact of Design Decisions on the Usability of Uncertainty Visualization: Noise Annotation Lines for the Visual Representation of Attribute Uncertainty.” In *Proceedings of 26th International Cartographic Conference ICC 2013 Dresden (Germany),* 25-30 August 2013.

Kinkeldey, C., and J. Schiewe. 2012. „Visualisierung thematischer Unsicherheiten mit Noise Annotation Lines.“ *Kartographische Nachrichten,* 62 (5), 241–249.

Kobus, D. A., Proctor, S., and S. Holste. 2001. “Effects of experience and uncertainty during dynamic decision making.” *International Journal of Industrial Ergonomics,* 28 (5), 275–290.

Lloyd, R., and R. Bunch, 2005. “Individual differences in map reading spatial abilities using perceptual and memory processes.” *Cartography and Geographic Information Science,* 32 (1), 33–46.

MacEachren, A. M. 1992. “Visualizing Uncertain Information.” *Cartographic Perspectives,* 13, 10–19.

MacEachren, A. M., Robinson, A., Hopper, S., Gardner, S., Murray, R., Gahegan, M., and E. Hetzler. 2005. “Visualizing Geospatial Information Uncertainty: What We Know and What We Need to Know.” Cartography and Geographic Information Science, 32 (3), 139-160.

MacGranaghan, M. 1993. “A Cartographic View of Spatial Data Quality.” *Cartographica: The International Journal for Geographic Information and Geovisualization,* 30 (2), 8–19.

Mason, W., and S. Suri. 2012. “Conducting behavioral research on Amazon’s Mechanical Turk.” *Behavior Research Methods,* 44 (1), 1-23.

Newman, T. S., and W. Lee. 2004. “On visualizing uncertainty in volumetric data: techniques and their evaluation.” *Journal of Visual Languages & Computing,* 15 (6), 463–491.

Pang, A. 2001. “Visualizing Uncertainty in Geo-spatial Data.” In *Proceedings of the Workshop on the Intersections between Geospatial Information and Information Technology,* 2001.

Pang, A. 2008. Visualizing Uncertainty in Natural Hazards. Risk Assessment, Modeling and Decision Support. In *Risk Assessment, Modeling and Decision Support,* edited by A. Bostrom, S. French, and S. Gottlieb, Springer Berlin Heidelberg, 261–294.

Rand, D. G. 2012. “The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments: Evolution of Cooperation.” *Journal of Theoretical Biology*, *299*(0), 172–179.

Senaratne, H., Gerharz, L., Pebesma, E., and A. Schwering. 2012. “Usability of Spatio-Temporal Uncertainty Visualisation Methods.” In *Bridging the Geographic Information Sciences. 15th AGILE International Conference,* edited by J. Gensel, D. Josselin, and D. Vandenbroucke, Springer Berlin Heidelberg, 3-23.

Shi, W. 2010. *Principles of modeling uncertainties in spatial data and spatial analyses*. Boca Raton: CRC Press/Taylor & Francis.

Tabachnick, B. G., and L. S. Fidell. 2013. *Using multivariate statistics*. 6th ed. Boston: Pearson Education.

Thomson, J., Hetzler, E., MacEachren, A. M., Gahegan, M., and M. Pavel. 2005. “A typology for visualizing uncertainty”. In *Proceedings SPIE 5669, Visualization and Data Analysis 2005,* edited by R. F. Erbacher, J. C. Roberts, M. T. Gröhn, and K. Börner, 146-157.

Vullings, L. A. E., Blok, C. A., Wessels, C. G. A. M., and J. D. Bulens. 2013. “Dealing with the Uncertainty of Having Incomplete Sources of Geo-Information in Spatial Planning.” *Applied Spatial Analysis and Policy,* 6 (1), 25-45.

Wittenbrink, C. M., Pang, A. T., and S. K. Lodha. 1996. “Glyphs for visualizing uncertainty in vector fields.” *Visualization and Computer Graphics, IEEE Transactions on,* 2 (3), 266–279.

Zhang, J., and M. F. Goodchild. 2002. *Uncertainty in geographical information*. London: Taylor & Francis.

Zuk, T., and S. Carpendale, 2006. “Theoretical Analysis of Uncertainty Visualizations”. In *Proceedings SPIE 5669, Visualization and Data Analysis 2006*, edited by R. F. Erbacher, J. C. Roberts, M. T. Gröhn, and K. Börner, 146-157.