Strategies for Detecting Difference in Map Line-up Tasks

Johanna Doppler Haider[1], Margit Pohl[1], Roger Beecham[2], and Jason Dykes[3]

[1] University of Technology, Vienna, johanna.haider@igw.tuwien.ac.at
[2] University of Leeds, Leeds, r.j.beecham@leeds.ac.uk
[3] City University London, j.d.dykes@city.ac.uk

Abstract. The line-up task hides a plot of real data amongst a line-up of decoys built around some plausible null hypothesis. It has been proposed as a mechanism for lending greater reliability and confidence to statistical inferences made from data graphics. The proposition is a seductive one, but whether or not line-ups guarantee consistent interpretation of statistical structure is an open question, especially when applied to representations of geo-spatial data. We build on empirical work around the extent to which statistical structure can be reliably judged in map line-ups, paying particular attention to the strategies employed when making line-up judgements. We conducted in-depth experiments with 19 graduate students equipped with a moderate background in geovisualization. The experiments consisted of a series of map line-up tasks with two map designs: choropleth maps and a centroid-dot alternative. We chose challenging tasks in the hope of exposing participants' sensemaking activities. Through structured qualitative analysis of think-aloud protocols, we identify six sensemaking strategies and evaluate their effects in making judgements from map line-ups. We find five sensemaking strategies applicable to most visualization types, but one that seems particular to map line-up designs. We could not identify one single successful strategy, but users adopt a mix of different strategies, depending on the circumstances. We also found that choropleth maps were easier to use than centroid-dot maps.

Keywords: Graphical inference · cognitive strategies · spatial autocorrelation · geovisualization · visual perception · sensemaking· thinking-aloud.

1 Introduction

If statistical graphics are to be used in data analysis and reporting, there needs to be reassurance that the statistical effect implied by a graphic can be reliably perceived. The possibility of a mismatch between statistical effect and its visual perception is especially relevant to geovisualization. Whilst maps convey information around the location and extent of phenomena that may be difficult to imagine using non-visual techniques, they may also lead to artefacts that are incidental to the statistical structure under investigation and that may even induce interpretation of false structure.

The graphical line-up test [24] is a practical means of effecting more reliable interpretation. Graphical line-ups, as depicted in Figure 1, can be considered as visual equivalents of test statistics. Line-up tests were developed in analogy to line-ups in the criminal justice system. The accused (the real data set) is hidden among several innocents (decoys). The innocents are data sets that conform to the null hypothesis. The null hypothesis assumes that there are no significant differences among the data sets. Significant differences between data sets can be tested statistically but also visually by human observers. If an impartial observer, an individual who has not previously seen the plot, is able to correctly identify the real from the decoys, then this lends confidence to the claim that a statistical effect exists – or rather, following null hypothesis significance testing, that the observed data are not consistent with the specified null hypothesis.

Graphical line-up tests are straight-forward to implement and conceptually appealing. They offer much potential to geo-spatial analysis [1]. However, they do not fully negate concerns around reliability of perception. Recent empirical studies have demonstrated that perception of statistical effects varies systematically with the intensity of effect [20], with visualization design [7] and in the case of geovisualization, with the geometric properties of the regions being studied [1]. Whilst there is evidence to suggest that these variations in perception are sufficiently systematic to be modelled (e.g. [7]), the evidence is less compelling for representations of geo-spatial data in choropleth maps (e.g. [1]).

Beecham et al. [1] speculated around the various explanations for why this is the case – why it is that, after modelling for variation due to intensity of statistical effect and geometric irregularity, there is much variation in perception of statistical structure encoded in maps. Elsewhere, Hofmann [8] and later VanderPlas & Hofmann [23] investigated whether or not ability to make correct judgements in (non geo-spatial) line-up tests varies as a function of individuals' perceptual capability and reasoning or some other demographic characteristics. Whilst both studies found variation in individual ability to interpret line-up tests, this variation was not consistent with demographics or visual abilities.

This study attempts to address the problem from a different perspective. Through structured qualitative analysis of think-aloud protocols, we attempt to expose the *sensemaking processes* through which judgements are made during line-ups displaying geo-spatial data. Specifically, we wanted to find out whether participants adopt different sensemaking strategies to solve the tasks. If this is the case, some variation might be explained by the use of different strategies. Using the materials made available by Beecham et al. [1], we developed a series of map line-up tasks with line-ups consisting of nine data

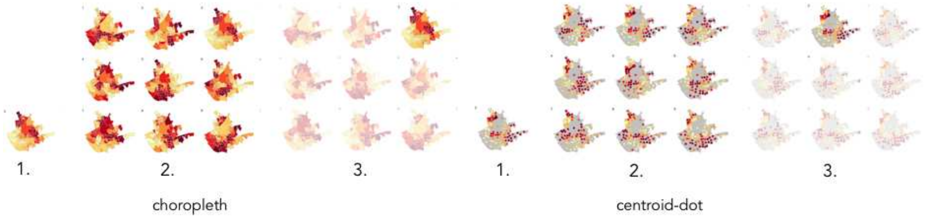

Fig. 1. A line-up is a visual equivalent to a test statistic. 1. Analyst observes neighbourhood-level pattern of crime rates. 2. Informal observer asked to pick the real data from a group of decoys constructed under CSR. 3. If the real is correctly selected from the decoys, we reject the null that crime distributes independently of location.

graphics: one plot of *real* data hidden amongst a set of eight decoy plots. We conducted a randomised controlled study (N=19) where the following conditions were varied:

- Geovisualization design: choropleth map | centroid-dot map
- Geometric irregularity: artificial grid | real geography, regularly shaped | real geography, irregularly shaped
- Graphic size: small | large
- Statistical intensity: low | high

We analysed how ability to perform line-up tasks – that is, to correctly identify the real from the decoys – varies by these different conditions. We also paid attention to participants' perceived confidence in making line-up judgements under different conditions and their preferences amongst the different conditions. We conducted a qualitative study, therefore the sample size is fairly small due to the extensive analysis process of the thinking aloud protocols.

The three main contributions of our investigation are:

- An exposition of the cognitive strategies users adopt when performing map line-up tasks. We identify six cognitive strategies; most are strategies generalisable across visualization types, but one is specific to geo-spatial data.
- Findings around the factors influencing performance of map line-up tasks. These factors may result from differences in the stimulus (map size, low/high statistical intensity) or from the strategies the participants adopted (cognitive strategies, time spent on task). The most important factor influencing performance is time spent on task.
- Insight into the role of geovisualization design in influencing task performance. We compared choropleth maps and centroid-dot maps to find out which of the two supported more reliable judgements of statistical structure in maps. We conjectured that centroid-dot maps would be associated with higher success rates, especially in the more irregular geographies, as they overcome the problems associated with different sizes and shapes of spatial units, visual artefacts that are inherent to chropleth maps. Our investigation indicates that this assumption is inaccurate. Participants performance is better when using choropleth maps.

2 Related work

The process of making inferences from graphics can be regarded as one of sensemaking. Individuals tend to apply a range of different strategies or heuristics, often in combination. Newell and Simon [17] describe how heuristics can be used to cut down the large problem space to manageable dimensions. They especially describe two heuristics – hill-climbing and means-end analysis. Gigerenzer [5] argues that reasoning processes in everyday situations are often based on a specific heuristic – gut feeling. Based on empirical research he shows that this heuristic can at times be very efficient. Fast inferences made from visualization can also be described as resulting from this heuristic. Lemaire and Fabre [15] discuss cognitive strategies from a conceptual point of view. They distinguish between general and domain specific strategies and argue that many reasoning processes

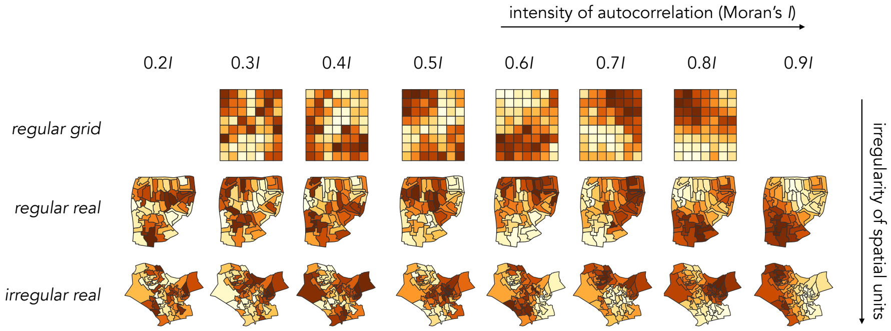

Fig. 2. We base our geographic stimuli – spatial regions – on conditions used by Beecham et al. [1]. We chose to test performance with two background levels of autocorrelation (*Moran's I*): low (0.3) and high (0.8).

are a combination of the two. In Information Visualization and Visual Analytics such issues have been discussed within the framework of sensemaking theories. In this context, Klein's approach has been especially influential [9–11]. Klein distinguishes between five different processes that enable people to gain insight: making connections, finding coincidences, emerging curiosities, spotting contradictions, and being in a state of creative desperation [9]. Sedig et al. [22] point out that complex cognitive activities can be described at different levels of abstraction. Lee et al [14]. analysed sensemaking processes and developed a model that is rather similar to Klein's approach.

Doppler Haider et al. [3] developed a model of sensemaking strategies, which is partly based on Klein's research [9]. They identify the following general sensemaking strategies: comparing (finding connections), laddering, storytelling, summarising, eliminating, verifying. They also found task specific sensemaking strategies which will not be discussed here. Pohl and Doppler Haider [19] provide a general overview of the literature on heuristics and sensemaking strategies. Our research is especially influenced by the last two papers. In the study presented in this paper we focus on sensemaking strategies used when comparing levels of autocorrelation in maps. We especially want to compare the strategies found for geo-spatial vizualisations with the strategies found for other visualization types and analyse whether there are strategies specific for detecting autocorrelation. In addition, we also want to analyse whether there are strategies that are more successful than others.

3 Study

3.1 Analytic background and study aim

When presenting data in maps, analysts are often concerned with the role of space, or spatial association, in phenomena: to what extent are high crime rates concentrated in certain neighbourhoods of a city and low crime rates concentrated in others? *Spatial autocorrelation* is a concept used widely for describing this tendency [18] and *Moran's I* a formal statistic for quantifying the amount or *intensity* of autocorrelation that exists. A test statistic for spatial autocorrelation is typically derived by comparing an observed intensity of *Moran's I* against a distribution that would be expected under *complete spatial randomness* (CSR) or some sensible prior knowledge [1].

Beecham et al. [1] measured the precision with which differences in spatial autocorrelation can be perceived in choropleth maps through a large crowd-sourced experiment. They found that precision varies within different stimuli (geometric irregularity and intensity of statistical effect, cf. Figure 2). As the intensity of autocorrelation structure increases, the difference in statistical effect necessary to correctly discriminate that structure decreases. Further, as geometric irregularity increases, so too does the difference in statistical effect necessary to correctly discriminate that structure. They also found much variation in the ability to discriminate structure that could not be explained through the experiment conditions. This variation may relate to physical artefacts introduced in choropleth maps which was not systematically controlled for. Or it may relate to differences in qualitative heuristics – *strategies* – applied by participants when making judgements.

3.2 Study conditions

This research aims to expose and characterise the *strategies* used when making judgements in *map* line-up tests and the study conditions tested were generated using resources published through Beecham et al. [1]. A summary of the conditions tested is displayed in Table 1. We vary geometric irregularity in the same way as Beecham et al.: we use the same geographic regions, cf. Figure 3, but add a further set of three regions with approximately the same levels of geometric irregularity and twice the number of spatial units (from ≈ 50 to ≈ 100 units).

One hypothesis for the large variation in perception identified in Beecham et al. [1] relates to geovisualization design. In a choropleth map, the entirety of each spatial units is given a single colour, which can result in salience bias in favour of larger regions and other artefacts that are incidental to statistical effect. We therefore also test a centroid-dot alternative. Introducing the centroid-dot maps brings some additional challenges: with a white background, dark dots gain greater saliency, whilst the contrary is true of a black background. A light grey background appeared to minimise these artefacts. Additionally, we design line-up tests with two intensities of baseline statistical effect: *Moran's I* of 0.3 (low) and 0.8 (high).

3	*geometric irregularity* (grid, regular, irregular)	×
2	*map size* (small ≈ 50 units, large ≈ 100 units)	×
2	*statistical effect* (low 0.3, high 0.8 – *Moran's I*)	×
2	*geovis types* (choropleth, centroid-dot)	×
19	participants	=
456	tests overall	
24	unique test conditions	

Table 1. Study conditions.

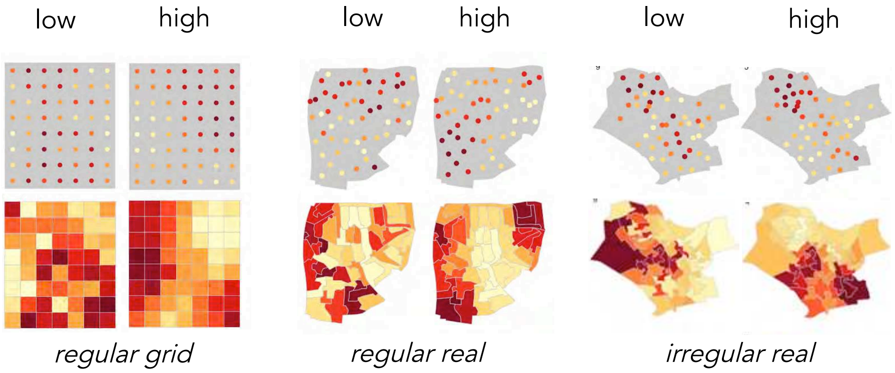

Fig. 3. The difficulty of line-up tests is assumed to increase with increasing irregularity of spatial units, i.e., grids and regular geometries are easier to assess than more complex, irregular geometries. The lower autocorrelation condition is assumed to be easier than the high condition for each geometry.

Since we wish to expose the various strategies employed when making line-up judgements we wished to generate line-up tasks that were relatively challenging. We therefore selected ”difference” values for decoy plots based on the thresholds published in [1]. These thresholds loosely represent the minimum difference necessary to correctly judge between two choropleth maps 75% of the time – a quantity referred to as *just noticeable difference* (JND) – and take into account the modelled influence of geometric irregularity and intensity of statistical effect. By choosing difference values in this way, we hope to control for the influence of geometry and intensity of statistical effect – that is we *expect* no difference in performance due to these factors.

3.3 Design and tasks

We used a within-subject ($N = 19$), counter-balanced design, where every participant performed 24 line-up tasks in random order. Each line-up was composed of nine images: eight *decoy plots* and one correct *target*, randomly positioned in a 3×3 array. The *target* was made different from the decoys by increasing its autocorrelation value in line with the thresholds published in Beecham et al. [1] – one JND higher than the decoys given the geometric irregularity (grid $|$ regular $|$ irregular) and baseline intensity of *Moran's I* (low $|$ high). All eight decoys contain approximately the same autocorrelation level. Note that each decoy is unique – even if it contains the same spatial autocorrelation value. The decoys were generated using a permutation approach published in [1].

For each map line-up we asked the following three questions: 1) Which is the plot with the highest spatial autocorrelation? 2) How confident are you in your choice? 3) Are there possible alternatives? We specifically decided to develop challenging tasks that forced participants to reflect explicitly about the problem and possible solutions. In this way, we were better able to study the strategies used by the participants than with simple tasks that can be solved at a glance. In easier scenarios, participants are less able to verbalise how they reached a solution because the reasoning process is fast and unconscious.

3.4 Data set

We arrived at threshold values for our conditions based on the experience of two pretests. Important considerations here were the time taken to complete the experiment (we wished to keep this to within 60 minutes) and generating stimuli of sufficient levels of difficulty to trigger slow sensemaking processes, and therefore expose *strategies*, rather than testing for pre-attentive perceptive abilities.

Difference in the decoys of the line-ups Beecham et al's [1] JND thresholds were generated under a very different setting to ours. Rather than a full map line-up, participants had to compare two images at a time in an established staircase procedure where the difficulty level changed due to participant performance. The aim was to encourage learning and improve performance to the extent that the JND level represents the minimum perceptible difference between the two stimuli. Since only two stimuli are used, the intuitive explanation of JND – the difference necessary to correctly discriminate 75% of the time – cannot be easily transferred to our study since the 75% figure must also include some chance-guessing.

We started by constructing line-up tests using the minimum JND threshold values published by Beecham et al [1] and completed two pre-tests with two and three participants in each. We hypothesise that a 50:50 success rate would suggest tests that are sufficiently challenging to expose user strategies, provide sufficient number of correct and incorrect tests in order to analyse the circumstances under which correct and erroneous judgements are made, as well as maintain the motivation of the participants.

In a second round we increased the difference to the median JND thresholds used by Beecham et al [1] and found that the target was too easy to identify with the effect that almost all answers were correct. Based on this observation we chose the value midway between minimum and mean JND's per geometry $(mean(JND) - (mean(JND) - min(JND))/2)$ and Moran's I which yielded the anticipated 50:50 performance (compare results).

3.5 Participants and procedure

We conducted a study with 19 computer science students with a Bachelor's degree or higher, from which eleven were male and eight female. Participants were between 23 and 31 years old with fair knowledge of geovisualization (average 3.15 on a 5-point Likert scale). We asked *"How familiar are you with map visualizations?"* with ranges from extremely, moderately, somewhat, slightly to not at all familiar. One participant reported a mild red-green colour perception deficiency, who afterwards stated that she did not feel challenged with the task. We chose a colour-blind safe colour palette from colorbrewer2.org for the visualisations. Participants were trained on both map types with different data than in the experiment.

Participants trained for 10-15 minutes prior to the experiment. They were furnished with an explanation of spatial autocorrelation and six examples, one per map type and geography (2×3). In the trial we first collected demographic information, followed by the line-up tasks and finally preferences on the map types. We asked participants to complete line-up tasks as depicted in Figure 1 with the exception that we specifically asked participants to identify the plot with the *greater* level of spatial autocorrelation. Participants were given feedback on these answers, but in the experiment proper, no feedback on participant performance was given. Additionally, we deliberately provided no context around the phenomena and spatial processes under investigation. Special attention was paid to instances where participants provided explanations behind judgements that included storytelling. Experiment sessions lasted between 50 and 60 minutes.

4 Results

4.1 Participant performance

Overall around 50% of line-ups were correctly answered: 213 correctly and 219 incorrectly identified the target from the decoys. For each test condition we consider the number of participants that performed the line-up correctly. Figure 4 displays this information as well as a frequency plot of participants' self-reported confidence in their answers for that condition on a 5-point Likert scale (1=not at all confident; 5=very confident).

The test condition with the highest success rate was the small centroid-dot map with a regular geometry and low level of baseline autocorrelation. Comparing success levels between geovisualization type, we found, counter to expectation, that the choropleth maps were associated with higher success rates than were centroid-dot maps (Cohen's d. effect size 0.64). An even larger effect was observed between the high and low baseline autocorrelation cases ($d.1.64$), with the low autocorrelation conditions associated with higher success rates than the lower cases. There is no obvious difference in success rates between map size (small and large) ($d = .02$).

On average, participants needed 42 minutes to complete the line-ups. There is a small correlation between used time and performance ($\rho\ (T, P)$ = .27). The greater the time spent studying the line-ups, the better the performance. We have eight participants with a good success rate of more than 50%. The remaining 10 answered less than half of the test cases correctly. The individual performances differ significantly. The best performance is 18 out of 24 line-up tasks correctly answered (75%), the worst performance is 5 out 24 (20,8%). We can neither observe a clear increase nor decrease in performance over time: whilst participant performance did not improve over time, neither did it deteriorate towards the end of the experiment, although verbal protocols include statements regarding fatigue in the last third of the experiment (compare Figure 5).

4.2 Participant preference

Participants expressed a strong preference for choropleth maps over the centroid-dot maps. This was true of all geometries (grid 29:9, regular 32:6, irregular 32:6). Only a

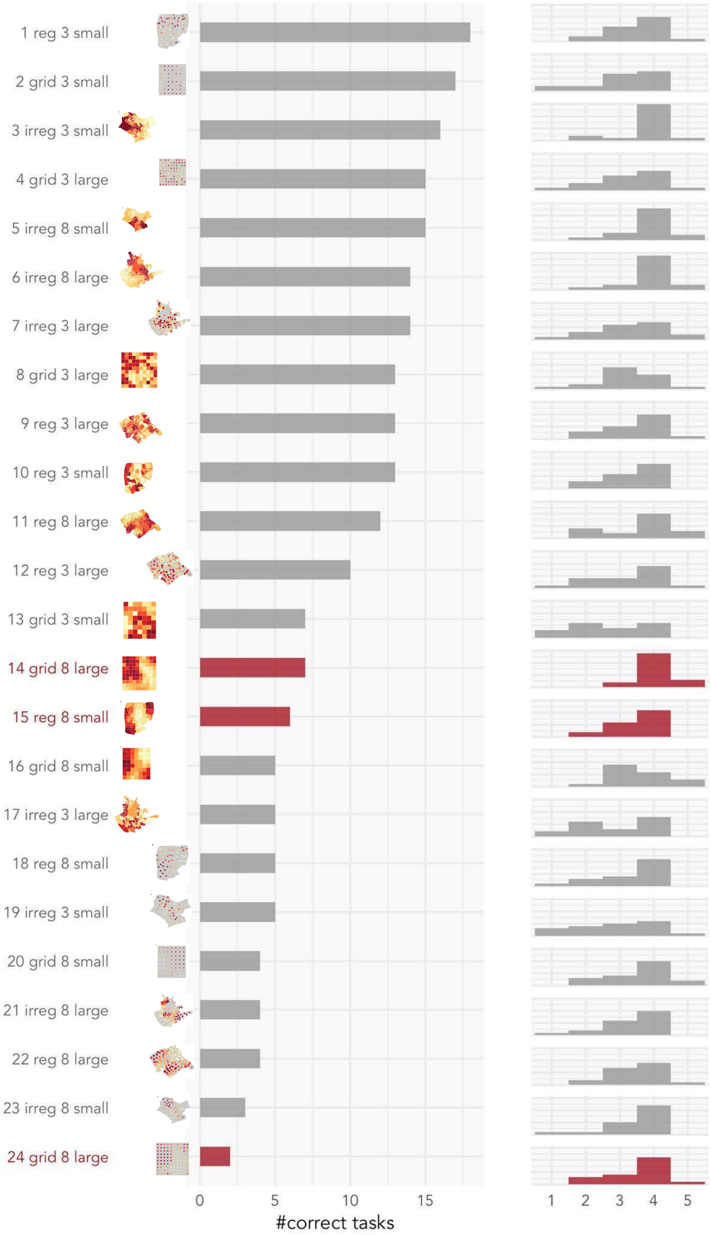

Fig. 4. Success and confidence per condition shows that participants were overconfident in the incorrect cases. Red lines show conditions that get discussed in the strategies examples. The map line-up stimuli with strategies used to form a judgement are shown in Figure 6.

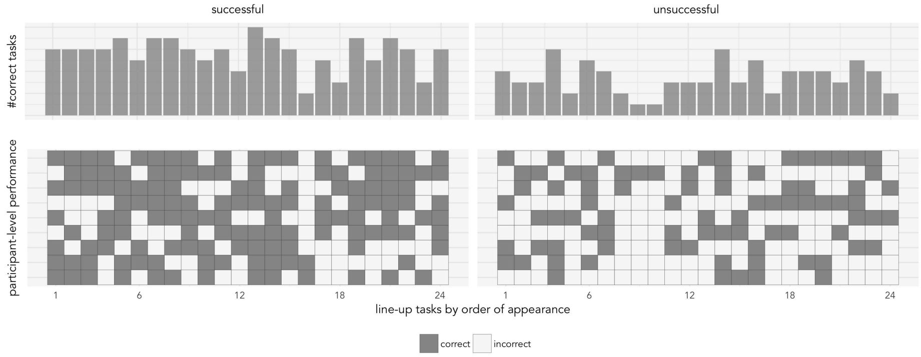

Fig. 5. Participant performance over time. We split *successful* participants (those with a success rate of $> 50\%$) from *unsuccessful* participants (with a success rate $< 50\%$). The left-most column represents the first test that participants performed; the right-most column the final (24th) test they performed. The bottom graphic displays individual, participant-level performance – each row represents a participant and each column identifies their n^{th} test. The top graphic displays counts of these columns – the number of correct identifications for the n^{th} test that participants performed. Note that the conditions shown to participants were randomised – therefore we do not expect patterns of performance over time to be a function of relative difficulty.

small number of participants preferred the centroid-dot maps and in a small number of specific cases: in the grid geometry five participants expressed a preference for the centroid-dot maps.

Overall the centroid-dots irritated the participants due to overlaps and differently sized areas between the dots. Nevertheless, when reporting this frustration, participants reflected on the importance of inter-zone distances. In one dot case a participant pitied a ”lonesome” dot, as it ”stands all alone by itself”, reporting that it would especially catch their attention. Our design seemingly did not overcome the saliency problem with large areas and may have exacerbated it.

4.3 Participant confidence

Asked directly about their confidence, participants were more confident with the choropleth maps (mean 3.49) than with the centroid-dot maps (mean 3.33). The top four mean confidence values were reported in choropleth map conditions. The success rate and reported confidence of each condition is shown in Figure 4. A chi-square test of independence showed no significant linear association between the confidence and success rate, $\chi^2(1) = 6.334, p = .175, (V = .118)$. This is confirmed by visually scanning Figure 4. An interesting observation, hidden by the summary statistic here is that high success rates are generally associated with high levels of confidence, but also that the lowest levels of confidence do not appear at the bottom of the graphic – where the success rate is lowest.

Participants were asked to optionally provide a ’second choice’ for the target. On average only 6 of the 19 participants named a possible alternative for the line-up test.

Code	*Frequency*	*Agreement rate*
Cluster	499	99.6%
Transition	355	98.7%
Outlier	354	99.1%
Colour	162	97.6%
Figure	70	85.1%
Gut feeling	58	93.9%

Table 2. The sensemaking codes show a high inter-coder agreement (Cohen's Kappa $\kappa = .976$, N= 19).

There is a small negative correlation between number of alternatives and mean confidence ($\rho\ (A, C)$ = $-.190$) – the higher the confidence the fewer the number of alternatives. This association is confirmed by the fact that the condition with the fewest alternatives offered was associated with a high success rate ($> 75\%$), whereas that with the most alternatives offered (12 participants in total volunteered an alternative target) had a success rate of only 50%.

4.4 Strategies

After transcription of the think-aloud protocols one researcher suggested possible codes for sensemaking strategies, which were consequently discussed with a second researcher. For the interpretation of the thinking aloud data we used qualitative content analysis [16, 21]. This is an empirical methodology for the systematic investigation of textual data that preserves some of the advantages of quantitative content analysis but still yields rich qualitative information about textual communication. Qualitative content analysis can either adopt a bottom-up or a top-down approach. The top-down approach requires investigators to develop a frame of categories before the study, the bottom-up approach consists of a repeated processing of the material with the goal to structure the material in a way to derive the categories from the material itself. We adopted a bottom-up approach but were inspired by categories identified in previous research. The final six codes were used by both researchers and their high inter-coder agreement (Cohen's Kappa $\kappa = .976$, compare Table 2) shows that the codes are efficient and easily distinguishable. The strategies will be described in detail in the following.

To assess the performance of strategies proved to be difficult but we found a dependency of performance and strategy change. Better performing participants switched less frequently, i.e., applied fewer strategies than participants with a success rate of less than 50%. They employed clusters followed by outliers and transitions as most successful strategies, as did the participant with the overall best performance. The worse performing group could not employ the transition strategy as successfully, and was rather successful with analysing clusters and outliers. A clear pattern, however, does not emerge.

The individual performances differ to a high degree, from the best performance with 18 out of 24 line-up correct cases (75%) to the worst performance of 5 out of 24 (20,8%). The best participant took more time (55 min) than the average (42 min) for the tasks and looked at every map in close detail. She reported that the centroid-dot condition was harder as was the higher autocorrelated condition than the lower, a truly good assessment as 5 out of the 6 wrong answers showed low autocorrelations. The best participant used 81 strategies (the average ≈ 83). Most often *Clusters* was used and secondly, *Outliers*. The participant with the least correct answers reported a red-green colour-vision impairment but afterwards reported that she had no problems discerning the colours. She noted that lighter colours seemed to stick out less than the dark ones. Although she completed the line-ups in average time, far more strategies (118) than the average (≈ 83) were used. We observed quick switching between strategies and insecurity about whether a strategy was useful for the task. Interestingly, three out of the five correct line-ups were the supposedly more difficult ones, the low autocorrelated centroid-dot regular real and the high autocorrelated irregular real small and large choropleth map. Most often *Clusters* were used to come to a decision and secondly, *Outliers*, which were the only successfully used strategies that led to correct answers.

Searching for clusters The grouping of units into clusters of the same colour has a big influence on the decision about autocorrelation. *Examples: "There is one big cluster in the middle", "Here we have two clusters on the sides", "I take the one with the fewer clusters", "I will choose the one with the big centered cluster above this one which also looks nice, with the left to right separation".*

Identifying clusters was the most dominantly used strategy, including mentioning the number of clusters, the size of clusters and the position of the cluster. If the form of a cluster was explicitly mentioned the statements were coded as both, cluster and figure strategy. We can summarise the following observations.

- Bigger and fewer clusters were favoured. The size and consequently the number of cluster had an effect on the decision making. If there are fewer clusters in the decoys it can happen that they look more homogeneous than the higher autocorrelated plot with, e.g., two clusters having a smooth transition, and, therefore, a higher Moran's I value. This happened for example in the small, regular, choropleth map line-up (compare *Cluster* line-up in Figure 6 plot nr.6: one yellow and one red cluster).
- Centred clusters were favoured. The position of the cluster influenced the decision in some cases and centred clusters had a greater effect than those on the side. The cluster in the middle got over-emphasised, e.g., in the case of small, grid, high autocorrelated centroid-dot maps (compare *Position* line-up in Figure 6 plot nr.5).

These strategies seem intuitive and can lead to good decisions, however, they are not reliable for autocorrelation judgements. The clusters in the high autocorrelated regular small choropleth map condition, for example, were often wrongly interpreted and decoy plots seemed to fit better to these strategies than the correct real (compare low success rate in Figure 7 and *Cluster* example in Figure 6).

Analysing transitions This strategy summarises all statements on transitions, where participants looked for smoother changes within each plot. *Examples: "There is a nice transition to the center", "This evolves beautifully from light to dark".*

Transition was remarkably well reported in the condition of high autocorrelated regular large choropleth line-up, which led to a good success rate compared to the centroid-dot (compare Figure 7) and fairly confident decisions (compare *11 reg 8 large* in Figure 4).

Elimination due to single outliers Participants used single outliers in a plot as a reason to exclude the plot from the possible range of answers. Hence, this elimination strategy was sometimes heavily applied when no positive example stood out. The maximum of 8 times per line-up, however, was rarely employed, but instead, a switch to a different strategy happened with the reduced set of plots. *Examples: "I will exclude this because of these high contrasts here", "Here are also these high contrasts", "There is a very light one right next to a very dark one".*

Emphasis on colour Dark hues had a greater impact than light ones. Darker colour hues were more often explored than the yellow, light hues when looking for clusters and transitions in both the choropleth map and the centroid-dot map. Some participants reported on looking at orange *"mid-level"* colours, others again tried to look at both and change their focus when they were stuck in an impasse, but those were the exception. *Examples: "I mostly look at dark areas", "Maybe I should look more on light colours.. I will try that now.".*

Regarding outliers light and dark hues were equally distracting, e.g., *"One light in the middle of this dark area"*, as well as *"..but then there is this one dark dot here".*

Recognising shapes (storytelling) Storytelling as a possibility of designing visualizations has been discussed in the visualization community [12, 13]. It has been argued that the development of a storyline supports users to form connections between disparate facts to make them memorable. Similar issues have been discussed in cognitive psychology for some time (see eg. [26]). In cognitive psychology, the emphasis is on the activities of the study participants and how they make sense of the material that is being presented to them. Participants try to construct coherent models based on this material. We called a strategy storytelling when participants identified meaningful shapes in the maps that helped them to solve the tasks. Figure-like cluster arrangements have a big impact. *Examples: "Here are dark clouds", "A blob or an island", "..like a mountain range", "This looks somewhat like Vienna".*

If a random plot resembles some kind of figure, a recognisable pattern with some kind of meaning, it has a strong effect on the user's decision, although it is unrelated to the task of judging autocorrelation, e.g., *"This looks somehow like a man with a stick"* (compare decoy nr.8 of the *Figure* line-up in Figure 6). This can, of course, also help if the figure is seen in the correct real, as it was the case in the low autocorrelated large choropleth grid condition where the highest autocorrelated plot was described as the shape of a *pincer* (compare real of condition *8 grid 3 large* in Figure 4). Confidence, however, was not high in this case (compare confidence in Figure 4).

Trusting a "gut feeling" This strategy summarised statements about first impressions and *gut feelings*, where no rationale could be verbalised. It is hard to discern aesthetics and other statements from this category, but we wanted to grasp how often participants were stuck in an impasse and had to give up like this, or just liked to trust their guts. We think it is due to the difficulty of the line-up tasks that this strategy was not employed very often. *Examples: "I don't know, it is just a feeling", "This was my first impression"*.

Although participants mostly could not make a decision at first glance, they often had first choices as a starting point. However, this depends very much on the participant and the verbal reporting (participants usually do not verbalise in the same amount of detail). For example, participant 17 mentioned a first idea in 16× of the 24 line-ups, where she then stuck with this choice only in five of the cases. The explanation then was, e.g., *"Here I will stick to my gut feeling"*. More cautious people, on the other hand, stated their gut feeling only when they were more or less sure about it, e.g., participant 19 verbalised 8× a first idea and then stuck with it in seven of the cases.

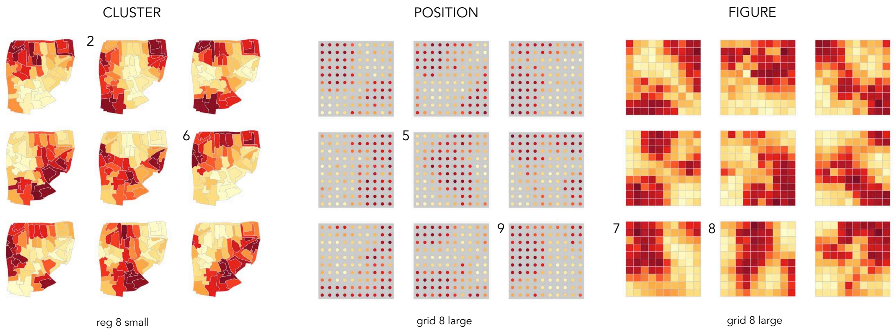

Fig. 6. Participants favoured plot 6 over 2 in the left choropleth line-up because of the smaller number of clusters; In the centroid-dot map (center) most participants chose plot 5 because of the centered cluster instead of the smoother, higher autocorrelated plot 9 leading to the worst success rate. Most participants chose plot 8 in the right choropleth line-up because of a perceived figure although plot 7 shows a higher autocorrelation value.

5 Discussion

5.1 Strategies

We conducted an investigation to study the processes underlying the perception and interpretation of autocorrelation in maps and how persons with knowledge in computer science and visualization try to solve such tasks. One of the main goals was to identify the most important cognitive strategies people use in such a context. We identified six such strategies: searching for clusters/identifying connections, analysing transitions, identifying outliers, emphasising colour, seeing meaningful figures (storytelling) and gut feeling.

The most popular trategies are *searching for clusters, analysing transitions, identifying outliers* and *emphasising colour*. The other two strategies are less important. Contrary to our expectation, we could not find any relationship between the strategy

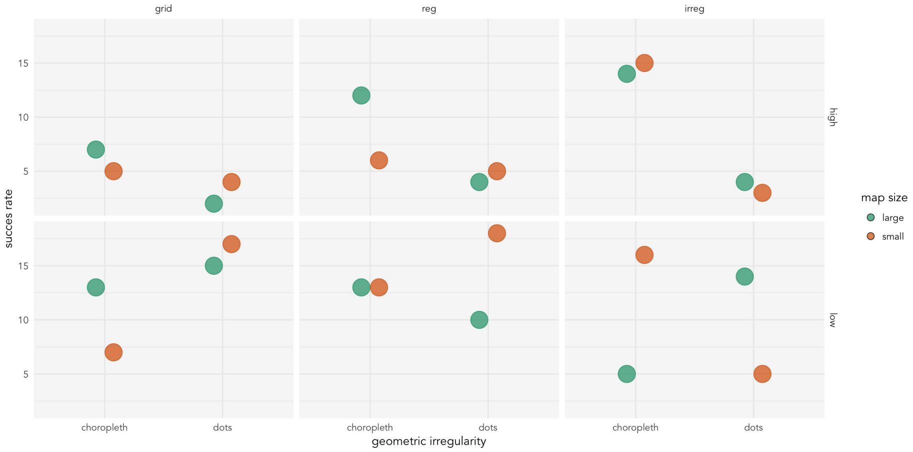

Fig. 7. Each dot represents a unique line-up – a line-up representing each experiment condition. The vertical position of dots varies according to the *success rate* for that line-up – the number of participants that correctly identified the real from the decoys.

used and the performance of the participants. We specifically assumed that the strategy analysing transition would be more successful than the others because it is tightly related to the task of finding autocorrelation. This was not the case. We also had assumed that participants would use this strategy most often because it seems to be most appropriate for the task of finding autocorrelation but they relied most often on the strategy of finding clusters.

Some of the strategies can be applied to other tasks and are not specific to the spatial and spatial-autocorrelation or map-lineup context: finding clusters, identifying outliers (elimination), storytelling and gut feeling [3, 6]. Especially finding clusters seems to be a very general strategy. It has some similarities with finding connections in Klein's model [9] where it is an important element. Klein argues that finding connections often occurs in the context of problem solving activities. In our experiment participants described clusters as elements having some connection with each other. From this they concluded that autocorrelation exists. Doppler Haider et al. [3] also found that finding connections is a strategy that is often adopted. Finding connections seems to be a kind of default strategy people adopt when they are not sure how to solve a problem. This might be the reason why participants applied it very often in our study although it is probably not the most obvious strategy. Elimination is often used when people work with visualizations. Newell and Simon [17] describe the frequent behaviour of excluding elements from consideration as strategy to make the problem space more manageable.

We would like to point out that the tasks participants had to solve were deliberately challenging. In many cases it was impossible to detect the correct solution at a glance, but participants had to study the map line-ups repeatedly. In this process they applied several different strategies in succession. This conforms to research from everyday thinking indicating that people do not adhere to fixed algorithms when solving problems, but use strategies pragmatically and apply them depending on the context [25]. Using different strategies seems to be an advisable approach. Nevertheless, using too many different strategies rather is an indication of confusion and should be avoided. Future work should try to clarify the issue whether combining different strategies is helpful in the context of map line-ups or not. Our experiment indicates that challenging tasks in map line-ups are difficult to solve. Drawing statistical inferences under such conditions is not a straight-forward task.

5.2 Performance and confidence

The quantitative analysis also provides interesting results. The performance of the participants was influenced by level of autocorrelation (high vs. low autocorrelation), but not by map size. This is surprising as we have deliberately designed our line-ups to control for this effect using empirically-informed prior knowledge [1]. We assume that the 0.3 and 0.8 cases are equally difficult. That we find a lower success rate associated with 0.8 may suggest that Beecham et al.'s [1] modelling might have overrepresented the effect of statistical intensity, that the staircase procedure used in Beecham et al. [1] disproportionately improves ability to detect differences in the high autocorrelation cases and/or that line-up designs 'work harder' where the level of statistical effect is low.

There is one approach that is helpful to solve map line-up tasks. This is analysing the material in detail and spending more time on the tasks. This is probably an indication of increased motivation. This also shows that challenging map line-up tasks require the users to study the material repeatedly for a considerable amount of time.

There was no relationship between confidence about the correctness of the results and the performance of the participants. The participants apparently had difficulties to assess the correctness of their solutions, especially when the tasks were very difficult. This probably indicates that the participants need more training in detecting autocorrelation.

5.3 Map design and spatial autocorrelation

We also compared choropleth and centroid-dot maps. We had assumed that centroid-dot maps would be associated with higher success rates, especially in the more irregular geographies, as they negate certain artefacts inherent to choropleth maps. Contrary to our expectation, participants performed significantly better with choropleth maps. This might be due to the fact that people are better acquainted with choropleth maps and feel more comfortable with them. Subjective ratings of the participants indicate that they prefer choropleth maps to dot maps. This could be addressed by adapting the centroid-dot design, e.g., varying size of dots, prominence of region borders or different colour palettes.

From our investigation, we can derive a few tentative ideas about factors influencing perception of autorcorrelation in maps. There are some factors that might distract viewers from detecting the real data. We saw that bigger and fewer clusters as well as clusters in the center were favoured. This indicates that other autocorrelations might be overlooked. Shapes conveying some imaginary meaning resulting in storytelling may also distract users.

The different autocorrelation levels (low and high) were received with mixed feelings. On the one hand participants liked the higher autocorrelation plots, calling them *"more pleasing"* just to realise that it made the task not easier because *"all look very good now"* (highest autocorrelated plot was not such an obvious visual outlier). In the high autocorrelated conditions participants would quickly find one or more possible candidates and while looking at each plot participants would add more and more to this subset ending up with a long list of possible answers.

We observed higher discomfort when confronted with the lower autocorrelated line-ups. Participants did not like them at first sight because they found them chaotic and could not make out any clusters or good transitions, i.e., the two most frequently applied strategies could not be applied.

5.4 Methodological challenges

Thinking-aloud is a good method to analyse thought processes while a task is solved or a specific tool is used. It can show up what participants have in mind while they are working [2,4]. This information cannot be complete because humans are not thinking about everything they do and look at, and only verbalise certain information at a time, hence, this method has its limits.

The analysis of our verbal protocols indicate differently employed reading directions of line-ups. The starting position was deliberated by some participants - if the top most left plot or the centre position was more investigated than other. One participant was concerned, e.g., that she was always coming back to the centre. The participant stated *'I don't really know why I decided for plot 5, maybe because it's in the middle'*. From the protocols we could observe starting from the top left to the right line-by-line, chaotically jumping between plots as well as pairwise comparison between likely candidates for the answer. For a complete analysis, however, eye-tracking is necessary to see in which order the plots get fixated and commented on. This would provide more detail on the specifics of the strategies used to solve line-ups and help establish whether there are differences in performance. It would also provide detail on whether the position of the target plot makes a difference when judging spatial autocorrelation in map line-up tasks.

6 Limitations and future work

This work aimed to identify sensemaking strategies used in solving line-up tasks in a qualitative manner. We could identify sensemaking strategies that were used in the line-up tasks, but the relationship between usage of strategies and performance is still not entirely clear. More work is necessary to clarify this issue.

We observed that participants adopted several different strategies for the same task, but we do not know whether or not such combinations are beneficial. Analysing the order of strategies seems like an important next step - left to right, or top to bottom or scan then check - are very different. We might find that performance is better if the outlier is checked first as a baseline. Capturing eye movements may provide detailed information that could inform this work.

Similarly to Beecham et al [1] we found great variability among participants concerning performance and strategies used. The reason for this is not clear. It would also be interesting to investigate whether people use different strategies for different levels of difficulty of the tasks or different levels of autocorrelation. We also noticed that the confidence about the solution deviated from the performance. Participants were sometimes very confident about incorrect solutions.

The sample we investigated was fairly small and consisted of computer science students. When conducting qualitative research, it is not possible to analyse larger samples because qualitative methods are very time-consuming. As far as the composition of the sample is concerned, we think that computer science students are representative of the kind of persons who might work with interactive choropleth maps and who might be interested in this topic. Nevertheless, the results may not apply to a wider population.

7 Conclusion

We conducted a study of graphical line-up tests to investigate how statistical inferences are drawn from data graphics. In this context, we are especially interested in the users' sensemaking strategies. We conducted an in-depth experiment with 19 graduate students with choropleth and centroid-dot maps. We could identify six strategies users adopt to solve challenging line-up tasks. Some of the strategies are comparable to strategies used when interacting with other visualizations (clustering/finding connections, elimination, storytelling). There is one strategy that is specific for the interpretation of map line-up tasks (transition).

Our research indicates that a successful detection of autocorrelation depends on a combination of various different strategies. Further research should clarify on which conditions specific strategies can successfully be applied. It is, for example, possible that strategies like storytelling are less successful and might be distracting. We found that using too many different strategies is an indication of confusion and should be avoided. Spending more time on the task, in contrast to that, is a condition for a successful solution of the tasks. Detecting autocorrelation sometimes apparently requires extensive exploration. We also compared choropleth maps and centroid-dot maps. In general, choropleth maps out performed centroid-dot maps. We were also able to identify some adverse conditions that may distract users (large clusters in the centre, dark clusters). We think that the results of such research can help us begin to understand the reliability of map lineup tests, how people conduct them and how we may train map readers to interpret autocorrelation and perform these tests more reliably.

8 Acknowledgments

The research reported in this paper has received funding from the European Union 7th Framework Programme FP7/2007-2013, through the VALCRI project under grant agreement no. FP7-IP-608142, awarded to B.L. William Wong, Middlesex University London, and Partners.

References

1. Beecham, R., Dykes, J., Meulemans, W., Slingsby, A., Turkay, C., Wood, J.: Map LineUps: Effects of spatial structure on graphical inference. IEEE Transactions on Visualization and Computer Graphics **23**(1), 391–400 (Jan 2017)
2. Boren, T., Ramey, J.: Thinking aloud: Reconciling theory and practice. IEEE Transactions on Professional Communication **43**(3), 261–278 (Sep 2000)
3. Doppler Haider, J., Seidler, P., Pohl, M., Kodagoda, N., Adderley, R., Wong, B.L.W.: How Analysts Think: Sense-making Strategies in the Analysis of Temporal Evolution and Criminal Network Structures and Activities. In: Proceedings of the Human Factors and Ergonomics Society 61st Annual Meeting - 2017. Austin (2017)
4. Ericsson, K.A., Simon, H.A.: Protocol Analysis. MIT press Cambridge, MA (1993)
5. Gigerenzer, G.: Gut Feelings: The Intelligence of the Unconscious. Penguin (2007)
6. Haider, J., Pohl, M., Hillemann, E.C., Nussbaumer, A., Attfield, S., Passmore, P., Wong, B.L.W.: Exploring the Challenges of Implementing Guidelines for the Design of Visual Analytics Systems. Proceedings of the Human Factors and Ergonomics Society Annual Meeting **59**(1), 259–263 (Sep 2015)
7. Harrison, L., Yang, F., Franconeri, S., Chang, R.: Ranking visualizations of correlation using Weber's Law. IEEE Conference on Information Visualization (InfoVis) **20**, 1943–1952 (2014)
8. Hofmann, H., Follett, L., Majumder, M., Cook, D.: Graphical Tests for Power Comparison of Competing Designs. IEEE Transactions on Visualization and Computer Graphics **18**(12), 2441–2448 (Dec 2012)
9. Klein, G.: Seeing What Others Don't: The Remarkable Ways We Gain Insights. PublicAffairs, a Member of the Perseus Book Group, New York, USA (2013)
10. Klein, G., Moon, B., Hoffman, R.R.: Making sense of sensemaking 1: Alternative perspectives. IEEE intelligent systems **21**(4), 70–73 (2006)
11. Klein, G., Moon, B., Hoffman, R.R.: Making sense of sensemaking 2: A macrocognitive model. IEEE Intelligent systems **21**(5), 88–92 (2006)
12. Kosara, R., Mackinlay, J.: Storytelling: The next step for visualization. Computer **46**(5), 44–50 (2013)
13. Lee, B., Riche, N.H., Isenberg, P., Carpendale, S.: More than telling a story: Transforming data into visually shared stories. IEEE computer graphics and applications **35**(5), 84–90 (2015)
14. Lee, S., Kim, S., Hung, Y., Lam, H., Kang, Y., Yi, J.S.: How do people make sense of unfamiliar visualizations?: A grounded model of novice's information visualization sensemaking. IEEE Transactions on Visualization and Computer Graphics **22**(1), 499–508 (2016)
15. Lemaire, P., Fabre, L.: Strategic aspects of human cognition: Implications for understanding human reasoning. Methods of though: Individual differences in reasoning strategies pp. 11–56 (2005)
16. Mayring, P.: Qualitative Inhaltsanalyse: Grundlagen und Techniken Beltz Verlag Weinheim (2003)
17. Newell, A., Simon, H.A.: Human Problem Solving, vol. 104. Prentice-Hall Englewood Cliffs, NJ (1972)
18. O'Sullivan, D., Unwin, D.: Geographic Information Analysis. John Wiley & Sons, New Jersey, USA, 2 edn. (2010)
19. Pohl, M., Doppler Haider, J.: Sense-making strategies for the interpretation of visualizations—bridging the gap between theory and empirical research. Multimodal Technologies and Interaction **1**(3) (2017)
20. Rensink, R., Baldridge, G.: The perception of correlation in scatterplots. Computer Graphics Forum **29**, 1203–1210 (2010)

21. Schreier, M.: Qualitative Content Analysis in Practice. Sage Publications (2012)
22. Sedig, K., Parsons, P., Liang, H.N., Morey, J.: Supporting Sensemaking of Complex Objects with Visualizations: Visibility and Complementarity of Interactions. In: Informatics. vol. 3, p. 20. Multidisciplinary Digital Publishing Institute (2016)
23. VanderPlas, S., Hofmann, H.: Spatial Reasoning and Data Displays. IEEE Transactions on Visualization and Computer Graphics **22**(1), 459–468 (Jan 2016)
24. Wickham, H., Cook, D., Hofmann, H., Buja, A.: Graphical Inference for Infovis. IEEE Transactions on Visualization and Computer Graphics **16**(6), 973–979 (2010)
25. Woll, S.: Everyday Thinking: Memory, Reasoning, and Judgment in the Real World. L. Erlbaum Associates, Mahwah, N.J (2001)
26. Zwaan, R.A., Magliano, J.P., Graesser, A.C.: Dimensions of situation model construction in narrative comprehension. Journal of experimental psychology: Learning, memory, and cognition **21**(2), 386 (1995)