© 2019 IEEE. This is the author's version of the article that has been published in IEEE Transactions on Visualization and Computer Graphics. The final version of this record is available at: 0.1109/TVCG.2019.2934807

A Comparison of Visualizations for Identifying Correlation over Space and Time

Vanessa Peña-Araya, Emmanuel Pietriga, Anastasia Bezerianos

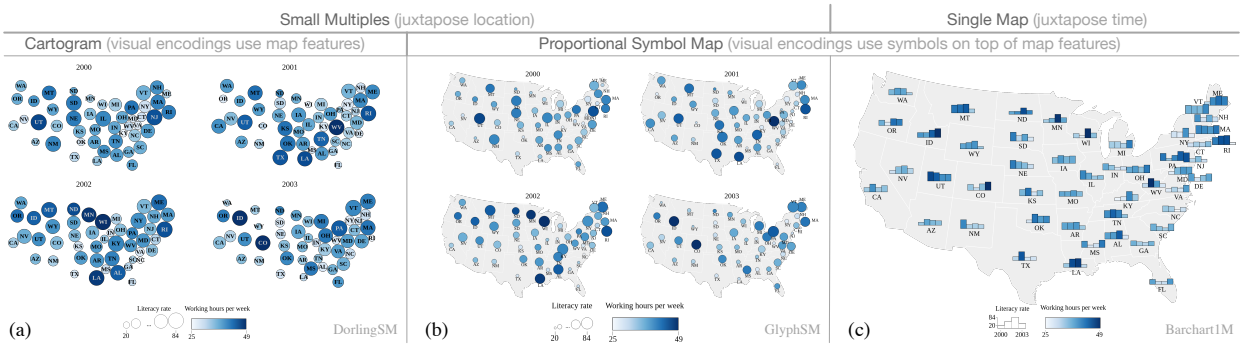


Fig. 1. The three visualizations compared in our study. (a) Dorling cartograms as small multiples, (b) proportional symbols (circles) on maps as small multiples, and (c) proportional symbols (bar charts) on a single map. In this example, each map shows the values of two artificially-created variables over four years. In each case, both variables have an overall positive correlation (Pearson correlation coefficient ≥ 0.75) and no monotonic evolution.

Abstract— Observing the relationship between two or more variables over space and time is essential in many domains. For instance, looking, for different countries, at the evolution of both the life expectancy at birth and the fertility rate will give an overview of their demographics. The choice of visual representation for such multivariate data is key to enabling analysts to extract patterns and trends. Prior work has compared geo-temporal visualization techniques for a single thematic variable that evolves over space and time, or for two variables at a specific point in time. But how effective visualization techniques are at communicating correlation between two variables that evolve over space and time remains to be investigated. We report on a study comparing three techniques that are representative of different strategies to visualize geo-temporal multivariate data: either juxtaposing all locations for a given time step, or juxtaposing all time steps for a given location; and encoding thematic attributes either using symbols overlaid on top of map features, or using visual channels of the map features themselves. Participants performed a series of tasks that required them to identify if two variables were correlated over time and if there was a pattern in their evolution. Tasks varied in granularity for both dimensions: time (all time steps, a subrange of steps, one step only) and space (all locations, locations in a subregion, one location only). Our results show that a visualization's effectiveness depends strongly on the task to be carried out. Based on these findings we present a set of design guidelines about geo-temporal visualization techniques for communicating correlation.

Index Terms—geo-temporal data, bivariate maps, correlation, controlled study, bar chart, Dorling cartogram, small multiples

1 Introduction

Understanding phenomena often requires looking at multiple variables, their inter-relationships, and how these evolve over time. Take Hans Rosling's visualization of the demographics of countries in his seminal 2006 TED talk [55]. Looking at the life expectancy and the fertility rate together is key to understanding the phenomenon at hand. Watching their co-evolution provides many of the insights unveiled by the speaker.

In many cases, the data will also feature a spatial dimension. Rosling refers to individual countries, but also different groups of countries multiple times. The spatial dimension plays an important role in his story, even if it is only indirectly represented in the scatterplot. Again, understanding the interplay between the considered variables, and the spatial arrangement of the entities they describe, can yield key insights.

- *Vanessa Pena-Araya is with Univ. Paris-Sud, CNRS, INRIA, Université Paris-Saclay. E-mail: vanessa.pena-araya@inria.fr.*
- *Emmanuel Pietriga is with Univ. Paris-Sud, CNRS, INRIA, Université Paris-Saclay. E-mail: emmanuel.pietriga@inria.fr.*
- *Anastasia Bezerianos is with Univ. Paris-Sud, CNRS, INRIA, Université Paris-Saclay. E-mail: anastasia.bezerianos@lri.fr.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

This famous example illustrates the potential of multivariate geo-temporal data visualization as a storytelling device. The speaker communicates insights about two variables that are related thematically, and that describe a phenomenon that is situated both spatially and temporally [2]. Beyond data storytelling, geo-temporal visualization can also support the analysis of such phenomena. The context, however, is different. While animation can illustrate temporal evolution when telling a story, it will often not be as effective for analysis purposes [65]. Moreover, depending on the application domain considered, information about group membership (*e.g.*, a country belonging to a particular continent) might not be sufficient to understand what role the spatial dimension plays in the phenomenon. Thus more detailed information about the topological relationship between entities might be necessary.

The problem of designing an effective visual representation in this context is challenging, as multiple data of different nature must be combined, each having specific characteristics: the thematic variables that describe the first-class entities in the dataset (life expectancy, fertility rate), the spatial properties of those entities (countries, continents), and the evolution of the thematic variables over time (years). Design choices will influence how well the representation can enable analysts to detect correlations between variables over space and time. It is thus important to identify guidelines to inform such designs.

Prior studies have compared geo-temporal visualization techniques for a single variable that evolves over space and time [21, 39, 40, 58].

Others have looked at two variables on a map (bivariate maps), but at a specific point in time [15, 18, 45]; or at how to visualize the correlation between two variables [31, 52, 53, 69], including visualizations that can be used to depict temporal evolution [26], but not in a geospatial context. To our knowledge, how effective visualization techniques are at communicating correlation between two thematic variables, that evolve over both space and time, remains to be studied.

We identify the different strategies used to combine thematic, spatial and temporal data into a visualization. The first design choice to be made concerns the combination of thematic variables in the representation: is the representation **juxtaposing all locations for a given time step**; or **juxtaposing all time steps for a given location**. The second choice concerns the visual encoding of thematic variables: either **overlaying symbols on top of map features**; or **using visual channels of the map features themselves**.

We discuss design variations for each strategy and identify three candidate techniques (see Fig. 1). Our study is designed to evaluate participants' ability to identify whether two variables are correlated over time or not, and if they are, if there is a pattern to their evolution. As we expect the techniques to fare differently depending on the number of time steps and the number of geographical entities to consider, we test them on tasks that vary both in temporal and in geographical granularity. Our results confirm this intuition, leading to a set of design guidelines about visualization choices for effectively communicating correlations in thematic geo-temporal data.

2 Related work

We first review some of the available visualizations categorized by how they combine space and time, and then how thematic variables are encoded to create multivariate maps. We finally discuss research related to perception studies of visualizations of correlated geo-temporal data.

2.1 Visualizing Combined Dimensions of Space and Time

Maps are the most direct visual representation of geo-temporal data. When combining both dimensions of space and time, thematic variables can be displayed by either juxtaposing locations (*e.g.*, small multiples of compact map representations); or juxtaposing time (*e.g.*, glyphs that represent multiple time steps overlaid on locations on a single map).

Juxtaposing location. From this category, small multiples are the most popular technique. For example, Johnson *et al.* [29] use small multiples to observe the correlation of Internet adoption with GDP and with population over the years. Animation can also be considered as a technique that juxtaposes location on maps that are presented in a sequence. Animation has been used to smooth the transition between views [9], or combined with symbols to depict change [32].

Juxtaposing time. The most common approach is to use glyphs in 2D (*e.g.*, [3, 17, 33, 47, 61]) or 3D (*e.g.*, [64]) on top of a single map. Additionally, the 3rd dimension has been used to juxtapose time over a map. For example, Space Time-Cubes [35] arrange time steps on the z-axis, effectively piling up the maps that correspond to each one of them. They have been used in several applications, *e.g.*, [19, 43, 50].

2.2 Visually Encoding Thematic Variables

Visually encoding data on a map can be done using two main strategies: mapping thematic attributes to visual properties of the map features; or overlaying symbols (*e.g.*, basic shapes such as circles, or glyphs such as pie charts and bar charts) on top of a base map, which remains untouched. As stated by Elmer [15], the number of possibilities to create bivariate or multivariate maps can range from dozens to hundreds (the declarative model of Jo *et al.* [28] for multiclass density maps shows numerous examples). Thus in this section we focus on those representations that are most commonly used or studied.

Encodings that use visual channels of the map features. Choropleth maps are among the most popular in this category [22, 41, 62]. They visually encode thematic attribute values using the map features' fill color. A bivariate type of choropleth, called value-by-alpha maps, allows for two variables to be displayed at the same time by combining color hue and transparency for each map feature [18].

	Juxtapose location	Juxtapose time	No time
Visual encodings use symbols on top of map features	[6] [16] [58]*	[33]† [39, 40]	[15]* [18]* [30]* [63]† [71] [38] [70] [4]
Visual encodings use map features	[21] [44] [45]*	[39, 40] [45]*	[15]* [18]* [30]* [63]† [23]* [24] [42] [62]

Table 1. Categorization of studies comparing geo-spatial visualizations. The first two columns represent the juxtaposition strategy. The third groups studies which compare visualizations that do not include time. The two rows represent the categories of visual encodings (symbols or map features). (*) indicates studies that consider more than one quantitative variable, and (†) studies that consider one quantitative and one qualitative variable. Note that some references are included in more than one cell as they make comparisons across categories.

Cartograms, use size as a visual encoding channel, and deform geographical shapes proportionally to the variable of interest [46]. There are four major types of cartograms: contiguous, non-contiguous, Dorling and rectangular. Contiguous cartograms distort regions to make their size reflect the thematic variable's value, preserving topology, and in particular adjacency, at the cost of statistical accuracy. Non-contiguous cartograms rescale each region of the map independently. They yield better statistical accuracy but fail to preserve topology (geographical regions are no longer contiguous). Dorling cartograms [12] yield more abstract representations of the geographical entities, replacing each region with a circle (Fig. 1-a). The circle's area can be mapped to a thematic variable. The position of circles is computed so as to preserve the overall topology, putting each circle as close to its original location as possible, adjusting their actual position to avoid circles overlapping one another. Finally, rectangular cartograms are similar to Dorling cartograms, but use rectangles to represent each region, yielding even more abstract representations of the geographical entities. Bivariate cartograms [66] use color or shade to encode a second variable in addition to that mapped to size. A recent variation on bivariate cartograms was presented by Nusrat *et al.* [45], in which two variables are visually encoded with size.

Encodings that use visual channels of symbols on top of map features. Overlaying thematic glyphs on top of a base map (*"symbols on maps"* [25]) gives more flexibility compared to mapping data to the attributes of the map features themselves. A wide variety of glyphs can be used to encode multivariate data. They are typically placed on top of geographical regions, on an independent layer. Proportional circles are the most frequently-used shape, but other basic shapes like squares, triangles or any other symbol can also be used [66]. Beyond simple shapes, more elaborate glyphs have been proposed; from generic glyph designs such as star glyphs or Chernoff faces [7] to domain-specific ones such as those used in meteorology [68].

2.3 Perception Studies on Correlated Geo-Temporal Data

We now summarize the studies we consider most relevant to geospatial visualization. From the extensive literature, we selected a subset using keyword searches involving *maps*, *geographical*, *geo-temporal*, *empirical study*, *evaluation*. We filtered out papers that were more than 20 years old, ones that consider numerical metrics but not visual perception (*e.g.*, [1, 41]), or that evaluated a new proposed technique in isolation (*e.g.*, [14, 37]). The final set of articles can be seen in Table 1.

We observe that most work on evaluating map-based visualizations does not focus on temporal evolution. From the results of those that do, we conclude that choosing the best-suited technique will depend on the task. For example, for analyzing statistical data over time and space, the results of Boyandin *et al.* [6] indicate that users get more insights with small multiples than with animation. This is confirmed by Robertson *et al.* [54] for the analysis of trends using non geo-spatial visualizations. For identifying moving patterns, Griffin *et al.* [21] show that animation leads to better results than small multiples. Other studies that consider temporal change focus on comparing only two points in time (*e.g.*, [44, 45]). They do not provide insights about the compared techniques' performance for identifying trends over space and time.

© 2019 IEEE. This is the author's version of the article that has been published in IEEE Transactions on Visualization and Computer Graphics. The final version of this record is available at: 0.1109/TVCG.2019.2934807

Regarding visual encoding, we observe that most studies do not focus on more than one quantitative variable at the same time. Particularly regarding correlation, two of them study user performance for tasks that require analyzing the relationship between two variables. The first, from Gao *et al.* [18], compares value-by-alpha maps with non-contiguous cartograms and proportional symbol maps. The latter displayed better overall performance. The second is from Elmer [15], who evaluated eight different visual encodings for bivariate maps. He focused on studying the effectiveness of different combinations of visual variables for the analysis of patterns. His results indicate that the eight combinations were consistent in accuracy, showing the utility of bivariate maps. Time was not considered in these studies.

Other research studies the perception of spatial autocorrelation [4,34] (how much a phenomena is dependent on spatial location). Yet other studies investigate the perception of correlation in visualizations that do not involve maps [26, 31, 52, 53, 69]. While such studies relate to our work, none of them considers all dimensions (correlation of *two* variables, over both space *and* time) simultaneously.

3 Study Rationale and Hypothesis

The literature describes many visualization techniques capable of encoding two thematic variables in a geo-temporal context. As it would be impractical to test them all, we discard general strategies that are ill-suited to the context of visual analysis, and identify representative techniques based on the strategies briefly introduced earlier. We then motivate our tasks, formulate our hypotheses, and explain how we have generated the synthetic datasets used in the study.

3.1 Selection of Visualization Techniques

Our first decision is to discard techniques that use animation to convey the temporal evolution of thematic variables. There has been much discussion about the role of animations [8] and their effectiveness [65], with sometimes-contradictory findings. But there seems to be relatively broad consensus that they are ineffective for detailed analyses of multiple variables over sequences of many time steps: showing only a single step at a time, they require users to remember previously-seen steps, thereby increasing cognitive load [27].

We also discard techniques that use 3D representations. These can provide more opportunities for mapping data attributes to visual variables (see, *e.g.*, [64]), which can be useful when visualizing multivariate data. But they typically force users to interact more with the representation, and require more elaborate means of navigation because of the higher number of degrees of freedom, among other pitfalls [60].

To make our study tractable, we make one final choice: to focus on visualizations based on how they represent the information, independently of any interaction technique. This means that we consider only static visualizations, in which elements can neither be filtered nor highlighted. As we discuss later in Sec. 7.1, follow-up studies should investigate how adding interaction impacts performance, but as this is the first empirical study to investigate the perception of correlation over space and time, there are already many factors to include before considering interaction techniques.

Based on these choices, we identify strategies used to combine thematic, spatial and temporal data into one visual representation. 1) We first categorize visualizations according to how they organize thematic variables. They can juxtapose values for all locations at a given time step, yielding **small-multiples** maps. Or they can juxtapose values for all time steps at a given location, yielding a **single map**. 2) We then categorize visualizations according to how thematic variables are visually encoded [15]. They can be mapped to the visual properties of symbols overlaid on top of the corresponding map features, eventually forming a **proportional symbol map** [18]. Or they can be mapped to the visual properties of the map features themselves. Both choropleth maps and cartograms fall in this category, but we only consider **cartograms** here. Indeed, encoding two thematic variables on choropleth maps is mostly limited to fill color hue, saturation and brightness, but these often interfere in terms of visual perception. Variations on the original design exist, such as, *e.g.*, Banded Choropleth Maps [14], but have not proven effective so far.

Combinations of these different strategies each yield multiple design variations. To avoid having to handle an unmanageable number of conditions, we choose at most one design per combination of strategies, and limit ourselves to designs that are actually used in practice. Those choices are rationalized below, taking into account the fact that our two thematic variables are quantitative in nature.

Proportional Symbol Map + Small Multiples: these techniques juxtapose values for all locations at a given time step. They consist of multiple identical base maps, one for each time step, with symbols superimposed on top of map features. The symbols' visual channels encode the thematic variables, showing individual values for the corresponding time step. We select circles, as they are the most frequently used shape [66], mapping the thematic variables to their radius and fill color brightness, respectively. This technique, which we refer to as `GlyphSM` in the study, is illustrated in Fig. 1-b.

Proportional Symbol Map + Single Map: these techniques juxtapose values for all time steps at a given location. They consist of a single base map. Because all values for all time steps are juxtaposed, we can create miniature bar charts [28], encoding one of the thematic variables using bar length instead of circle radius. Length is considered a more effective encoding channel than area, and this also makes for a more compact glyph than juxtaposed circles would. The second variable is mapped to each bar's fill color brightness. This technique, which we refer to as `Barchart1M`, is illustrated in Fig. 1-c.

Cartogram + Small Multiples: these techniques juxtapose values for all locations at a given time step and encode thematic attributes directly on the map features, without using symbols. They consist of multiple cartograms, one for each time step in the dataset. Among all variations on cartograms (discussed in Sec. 2.2), prior studies have shown that contiguous cartograms and Dorling cartograms perform best overall [44]. We chose Dorling cartograms over contiguous cartograms as results of previous studies indicate they yield higher statistical accuracy and are better suited to *summarize* tasks, therefore better aligned with the analysis of correlations. This technique, which we refer to as `DorlingSM`, is illustrated in Fig. 1-a.

Cartogram + Single Map: while instances of this combination do exist, all the ones we identified are somewhat contrived. Indeed, it is difficult to have a single small glyph meet all requirements: (i) show two thematic variables; (ii) show individual values for each of them, (iii) for each time step; and (iv) preserve the global topology of map features. One possibility would be to take the above Dorling cartogram, slice the circles radially into as many time steps (transforming them into pie charts), and map the thematic variables to each slice's radius and fill color, effectively creating a rose chart. Such a design, however, makes it difficult to compare values across entities. Other possibilities exist, involving, *e.g.*, augmented donut charts or treemaps, but none of these is in reasonably widespread use and none stands out as a promising technique. We thus did not include this combination in the study.

3.2 Task Motivation

Our goal was to compare the effectiveness of visualization techniques, when it comes to identifying the correlation between two variables and its evolution over time. We had no hypothesis about which part is more difficult: detecting different types of correlations (positive / negative / non-existent), or characterizing their evolution (following a trend or not). We thus treat them as a single integrated task, that requires viewers to identify both potential correlations and their trends. We varied the combinations of these factors in our tasks to cover their range, but without exhaustively testing all combinations (Sec. 3.4) and without making any assumption about their difficulty. Such integrated tasks fall under *"characterize the relationship among multiple map features"* in Roth's task taxonomy [56].

To construct our tasks, we used the geo-temporal framework proposed by Peuquet [51], that describes the linked triad of "what", "where" and "when". Each task corresponds to a question of the type *when* + *where* $\rightarrow$ *what*, where *what* is the participant's characterization of the correlation and its evolution.

We varied the *when* and *where* in a way similar to other research (*e.g.*, [20, 59]), using three granularity levels. In particular, the classifi-

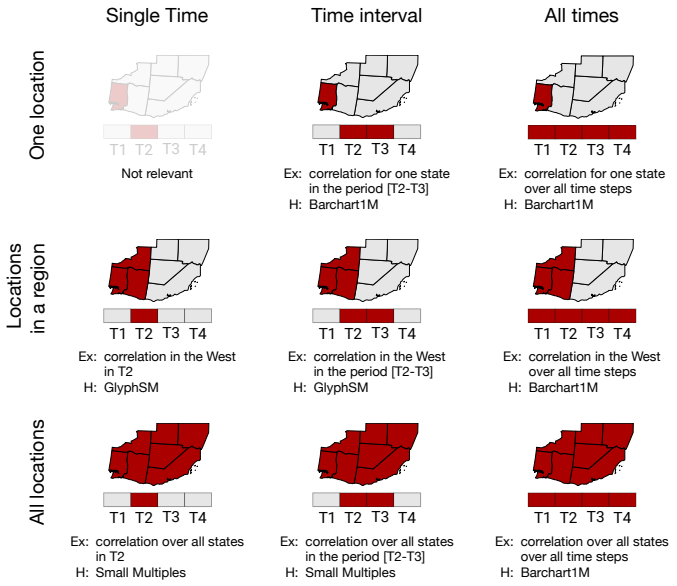


Fig. 2. Summary of tasks based on spatio-temporal granularity. In each cell, the image illustrates the task, together with an example (Ex), and our hypothesis (H) about which visualization will perform best overall.

cation of granularity levels for time (*when*) is divided in (i) one time, (ii) a time interval, and (iii) all times. Space (*where*) is categorized as (i) one location, (ii) locations in a region, and (iii) all locations. Crossing the spatial and temporal dimensions results in a matrix of nine possible tasks illustrated in Fig. 2, together with a concrete example. Correlation at one location in one point in time (top-left cell) is not meaningful and was discarded as a task. We thus ended up with eight possible spatio-temporal tasks.

We hypothesized that the best-performing visualization would depend on the task considered. Specific hypotheses are detailed in Sec. 3.3, and the techniques hypothesized to perform best for each task are also indicated in Fig. 2.

3.3 Hypotheses

The following hypotheses capture our expectations and were formulated before data was collected:

H1: We expect small multiples (`GlyphSM`, `DorlingSM`) to result in better performance (less time and fewer errors) than single maps (`Barchart1M`) for tasks that require analysis at *one point in time* only. The search for the desired point in time is done only once across small multiples, and then the focus is on the spatial information that is grouped closely together. Whereas for a single map the specific point in time needs to be identified repeatedly across map features (bar charts).

H2: For *time intervals* in a single location, we expect better performance (time or errors) for a single map, as all information is colocated (one bar chart) (**H2.1**). When it comes to locations in a region, or to all locations, small multiples (`GlyphSM`, `DorlingSM`) will fare better than single maps (`Barchart1M`) (**H2.2**). We expect that repeatedly identifying the right time interval across multiple locations in a single map will make this visualization slower and lead to more errors.

H3: We expect single maps, that juxtapose time (`Barchart1M`), to result in better performance (time or errors) than small multiples (`GlyphSM`, `DorlingSM`) for tasks that require analysis over all time steps. Indeed, small multiples require users to continuously change their focus between many maps to see trends for locations and make comparisons. This is not the case for single maps as they allow getting an overview of the behavior at each location quickly and identify trends.

H4: We expect that among small multiple techniques, `GlyphSM`, which overlays symbols on a base map, will feature better performance across all tasks. Cartograms (`DorlingSM`) adjust the layout of features in each map independently, thus making it hard to identify and match them across small multiples.

3.4 Dataset and Task Construction

For the setup of our experiment we use the map of the United States (*i.e.*, map features are US states) over nine years of temporal evolution (*i.e.*, a point in time is a year).

The geography of the US states provides good diversity in terms of size of individual features (*e.g.*, Texas compared to South Carolina) and density of those features (*e.g.*, west coast compared to east coast). In trials involving a single location, we varied the size of target features (smaller & larger states) and density of the surrounding geographic area. We grouped locations in contiguous regions using the four cardinal points: north, south, west and east. These regions were selected as they represent common geographic division of countries or other administrative levels. Regions were determined by drawing an imaginary line that divided the country into two equally-sized areas, vertically for east and west, horizontally for north and south. This resulted in areas of varying density across trials. To avoid participants fixating their gaze over discontinuous areas, especially for tasks involving a subset of locations in a region (Fig. 2, second row), we removed Alaska and Hawaii from the map. This resulted in a total of 48 locations, a fair amount of locations to analyze.

Regarding time granularity, we define all time spans to be nine years long (a number that utilizes the space of a small multiple setup). Time intervals were made of four consecutive steps, selected in the middle of the range so as not to favor single maps – identifying the first or last part of the small bar charts is much easier. Four years represent almost half of the total time steps, which allows us to balance the amount of patterns (correlations to identify) and noise (additional data-points to make the task realistic).

The two variables were presented to participants as *literacy rate* and *working hours per week*. Nevertheless, to control the displayed correlation and trends within the different spatio-temporal constraints, we used artificially-created datasets. We initially created variables that followed normal distributions, as other perception studies about the visualization of correlation do [26,53]. With this type of distributions, it is common that points do not follow strict patterns of both increasing at the same time (in case of positive correlation), or one increasing as the other decreases (in case of negative correlation). This is not a problem with scatterplots, as the overall distribution of many points helps convey the overall relationship. However, in our case, the number of points in time was small, minimum 4 for time intervals and maximum 9 for all time steps. Thus, even if one point did not follow the pattern, it would suggest that there was no correlation. We instead generated pairs of points using a random linear regression model with added Gaussian-centered noise,[1] as the difference between values could be evaluated more clearly. The obtained points were checked to ensure that they follow the pattern for the desired time range. To make the generated distributions closer to actual literacy rate and working hours per week, we scaled our generated data between values extracted from Rosling's GapMinder example. For instance, for the variable assigned to literacy rate, we scaled between a minimum within $[20,30]$ and a maximum within $[75,85]$. For the variable assigned to working hours per week, the minimum varied within $[25,35]$ and the maximum within $[40,50]$.

For each task, we created a dataset that followed particular spatio-temporal patterns. The possible correlation patterns were: positive correlation ($r \geq 0.75$) with and without monotonic evolution; negative correlation ($r \leq -0.75$) with and without monotonic evolution; and no correlation ($|r| \leq 0.2$). These patterns were enforced for the space and time granularities considered in each task (*e.g.*, a time range or all times).[2] We added distractors for the locations and time points that were not the focus of the task by including 1/3 of data points that did not follow the assigned pattern.

To increase reliability, our design included three repetitions per task, that were aggregated in our analysis. To avoid learning for each repetition, we varied the selected location, region, point in time and time interval. We generated one dataset per task repetition that, for

[1] Data was created with Scikit-learn [48], using `make_regression`.

[2] We note that for tasks that require analysis in one point in time, it was not relevant to create two variables with monotonic evolution.

© 2019 IEEE. This is the author's version of the article that has been published in IEEE Transactions on Visualization and Computer Graphics. The final version of this record is available at: 0.1109/TVCG.2019.2934807

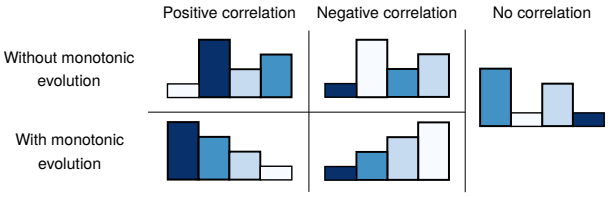


Fig. 3. Schematic illustration of all possible answers for tasks in Fig. 2. The three trial repetitions included combinations, such that each correlation type (positive, negative, no-correlation) appeared once. When temporal evolution was applicable, one of the positive/negative correlations was coupled with monotonic evolution, while the other was not.

the spatial and temporal constraints required by the question (time and space granularity), followed a different correlation pattern. For the three repetitions, there was always: one trial with no correlation, one with correlation (positive or negative) but no monotonic evolution, and one with correlation (positive or negative) and monotonic evolution. Fig. 3 illustrates the different configurations.

To avoid participants remembering answers across visualizations, from each generated dataset we derived two additional variations by shuffling data over states, over years, or both. Thus, for each task repetition displayed in each visualization, the participants would observe a dataset with the same structure but with different layouts. In total, this resulted in 80 datasets: 8 tasks $\times$ 3 repetitions $\times$ 3 datasets (1 original + 2 shuffled variations) $=$ 72 for main trials $+$ 8 for training.

4 Study Design

The study was designed to evaluate, for each of the tasks, the three visualizations introduced earlier. Supplemental material containing dataset generation code, experiment data, analysis scripts and detailed results are available at `http://ilda.saclay.inria.fr/spacetimecorr`.

4.1 Experimental Design

We used a within-subjects design where all participants were exposed to all three visualization techniques. For each technique, a participant had to perform 8 training trials, and 8 measured tasks $\times$ 3 repetitions $=$ 24 main trials. Repetitions considered one of each possible correlation types: Positive, Negative or No-Correlation. For tasks that involved analysis over time, answers also included monotonic choices (Fig. 3).

Technique presentation order and dataset variations were counterbalanced across participants using Latin squares. Tasks were grouped by time granularity (one point in time, time interval, all times) and their order of presentation was counterbalanced as well. For each time granularity, the order of geographical granularity was randomized. Within each group of space and time granularity, the three task repetitions were also randomized. In total, the experiment consisted of 18 participants $\times$ 3 visualizations $\times$ 8 tasks $\times$ 3 repetitions $=$ 1296 trials.

4.2 Apparatus and Participants

We used a 27" Apple Thunderbolt Display set to its default resolution (2560$\times$1440 pixels). The web user interface was implemented in Django and visualizations were generated with D3 [5] and Vega [57]. We made sure that all visualizations were of similar size by keeping their width consistent (adjusting height to keep the original aspect ratio). All visualizations fit comfortably on the screen and did not require scrolling. More specifically, the dimensions were 1350$\times$996 pixels for `GlyphSM` and `DorlingSM`, and 1350$\times$849 pixels for `Barchart1M`.

We recruited 18 participants before starting the experiment, a number that allowed us to counterbalance technique presentation order. We continuously recruited participants until we arrived at this pre-defined number. Our participant exclusion criteria included: not completing all conditions, or failing any of the 3 training trials. Given the complexity of the task, we assumed task learning would transfer across techniques. Thus, an excluded participant would have to be replaced with another participant with an equivalent configuration of technique, dataset and task presentation ordering. We had to replace a single participant who declared during the second session that she had misunderstood how to perform the tasks in the first session.

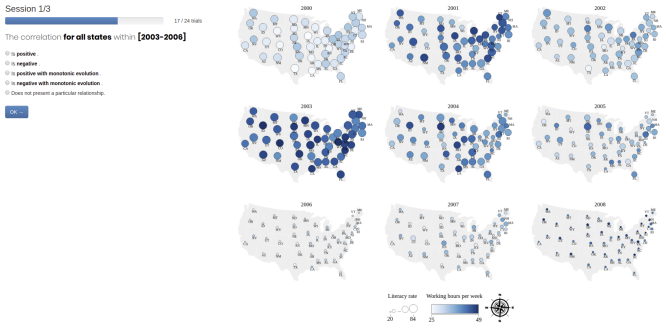


Fig. 4. Web interface used to conduct the experiment. Visualization = `GlyphSM`; task performed on a Time Interval, for All Locations.

From the final 18 participants (10 female and 8 male), none reported any color deficiency. All had normal or corrected-to-normal vision. Age ranged from 23 to 40 ($M = 27.6$, $SD = 4.9$) and most of them were students (13/18) from either a PhD or a Masters' program. Their backgrounds were mainly HCI, Computer Science and Visualization. They were all volunteers, and did not receive any monetary compensation.

4.3 Procedure

First pilots of our study showed that conducting the tasks was mentally demanding. We thus divided the study in three sessions, one per visualization, performed on three different days (that could be consecutive and at most 9 days apart). Each session consisted of three parts: introduction, training, and main trials. In the first session, participants signed a consent form, were told that they could withdraw at any time, and filled out a demographic questionnaire.

1) Introduction and training. The experimenter explained the visualization to be used in the session, along with examples of how correlation and monotonic evolution looked on it. Further training was conducted, that consisted in answering eight trials, one per task (described next). After finishing each trial, the system would indicate if the answer was correct or not. If participants made no error and declared that they had no further question, they would start the main trials. Otherwise, the experimenter would add further training trials.

2) Completion of main trials. Fig. 4 shows a trial screenshot. On the left are the overall progress, the question asked and possible answers. On the right is the generated visualization for that condition. Before each trial a map was shown, highlighting the location(s) that the trial would be about. Our aim was to reduce potential bias due to prior knowledge of the United States' geography, and to ensure there was no ambiguity about geographical features to consider such as, *e.g.*, which states constitute a region.

Participants completed 24 main trials per session (visualization). In this phase, they did not get any feedback about the correctness of their answers. They were instead asked to report the level of confidence in the answer they had just given (low, medium, high).

Once trials were completed for a visualization, participants filled out a post-hoc questionnaire about the strategies used to complete the eight tasks, and how easy it was to complete each one of those tasks. After finishing the third session, participants filled out a final questionnaire, in which they were asked to rank the visualizations. A representative image of each visualization was displayed in the form to help participants remember them. The entire experiment (3 sessions) took approximately one hour and a half.

4.4 Measures

For each task, we defined three metrics, two objective, one subjective:

- `Task completion time:` measured from the moment participants saw the trial screen until they submitted an answer. We computed the average over the 3 repetitions.
- `Error rate:` computed as the number of incorrect answers per task multiplied by the total number of repetitions.

- `Self-reported confidence`: measured on a 3-point Likert scale (high, medium, low).

For each technique, we recorded:

- `Strategies to complete the trials`: described as free text.
- `Self-reported difficulty to complete each type of task`: measured on a 5-point Likert scale from very easy (5) to very difficult (1).

5 Results

We analyze, report, and interpret all our inferential statistics using graphically-reported point estimates and interval estimates [11, 13].

We report sample means for **Completion Time** and **Error Rate** and 95% confidence intervals (CIs), indicating the range of plausible values for the population mean. For our inferential analysis we use means of differences and their 95% confidence intervals (CIs).[3] We use BCa bootstrapping to construct all confidence intervals (10,000 iterations). Since in our *per-task* analysis we test specific predictions rather than a universal null hypothesis, no correction for multiple comparisons was performed [10, 49]. A p-value approach of our technique can be obtained following the recommendations from Krzywinski and Altman [36]. Finally, we also report percentages for self-reported **Confidence** results.

We analyzed a total of 1296 trials (18 participants $\times$ 72 trials). All reported analyses were planned before the experiment started.

We first provide an overview across tasks.[4] Since our hypotheses are task dependent, we then perform a detailed per-task analysis.

5.1 Overall results across tasks

Completion Time: Fig. 5 shows completion times of all tasks collectively. Mean times per technique are on the left, mean differences on the right. Mean times are shorter for `GlyphSM` (23.7sec) followed by `DorlingSM` (26sec) and `Barchart1M` (30.7sec). There is strong evidence that `Barchart1M` is slower than `GlyphSM` (by 7.0sec on average) and evidence that it is also slower than `DorlingSM`, although the difference is smaller (4.5sec on average).

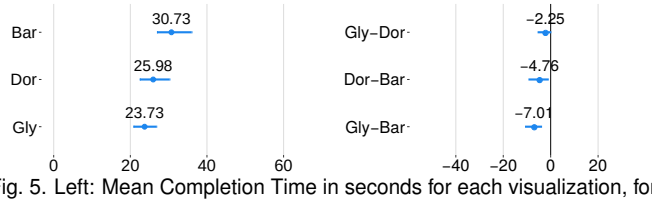


Fig. 5. Left: Mean Completion Time in seconds for each visualization, for all tasks. Right: Pairwise comparisons for each visualization. Error bars represent 95% Bootstrap confidence intervals.

Error Rate: Fig. 6 shows error rates for all tasks collectively, with mean error rates per technique on the left and mean differences on the right. Mean error rates are lower for `GlyphSM` (7.4%) followed by `DorlingSM` (8.1%) and `Barchart1M` (8.6%). There is no evidence that error rates were different across techniques. Thus the main differentiation we can make across techniques comes from completion time.

Confidence: Fig. 7 shows the self-reported confidence for each visualization, for all tasks. Confidence is high for all three visualizations in more than half the trials, although more so for `GlyphSM` (64% of trials) than for `DorlingSM` (57%) and `Barchart1M` (53%).

5.2 Results per task

Next we report results per task, grouped by *temporal granularity* for legibility purposes (analyses were performed per task). The values

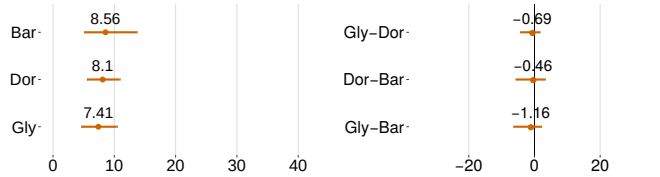


Fig. 6. Left: Mean Error Rate in % for each visualization, for all tasks. Right: Pairwise comparisons for each visualization. Error bars represent 95% Bootstrap confidence intervals.

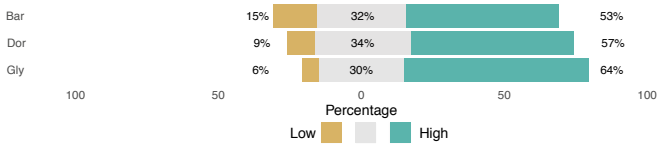


Fig. 7. Self-reported confidence across all tasks per visualization.

and CIs for means and differences of means for both Completion Time and Error Rate can be seen separately for each task in Fig. 8, with the direction of our hypothesis indicated by a gray background. Self-reported Confidence per task can be seen collectively in Fig. 9.

5.2.1 SingleTime correlation tasks

In tasks involving a single time step we expect small multiples techniques to fare better (**H1**). Completion times and error rates (means and CIs) for these tasks are found in the leftmost column of Fig. 8.

Completion Time: is faster with small multiples (`GlyphSM` , `DorlingSM`) and slower for `Barchart1M` for both geographic granularity tasks. Looking at mean differences, there is strong evidence that `Barchart1M` is slower than both small multiples techniques, by more than 27sec for Region, and by more than 32sec for AllLocations tasks. Results are inconclusive for the difference between `GlyphSM` and `DorlingSM` in both tasks.

Error Rate: Similar to the results for completion time, for both Region and AllLocations tasks, `GlyphSM` had the best performance, followed by `DorlingSM` and `Barchart1M` with the highest error rate. Looking at mean differences, there is strong evidence that `Barchart1M` is more error prone than `GlyphSM` for both types of geographic granularities. There is weak evidence that `Barchart1M` is also more error-prone than `DorlingSM` for Region (but no evidence of a difference for AllLocations). Finally, `DorlingSM` appears more error-prone than `GlyphSM` for both tasks (strong evidence of this difference for Region, and weak for AllLocations).

Confidence: (self-reported by participants) corroborates these findings. For both tasks that considered Single Time, confidence is high for small multiples techniques (`GlyphSM` and `DorlingSM`) but low for `Barchart1M` (see top of Fig. 9).

Summary for SingleTime: Overall, the tendencies for the two tasks that focus on correlations for a Single Time are similar, irrespective of whether we consider a geographical region or all locations. We have evidence that using the small multiples visualizations (`GlyphSM`, `DorlingSM`) takes less time (less than 20sec) and causes less errors than `Barchart1M`, supporting **H1**. There is also evidence of differences between `GlyphSM` and `DorlingSM` when it comes to errors, with `DorlingSM` being more error prone, supporting **H4**.

5.2.2 Time Interval correlation tasks

In time interval tasks, we expect different performance across geographic granularities (**H2**), with a single map (`Barchart1M`) faring better for tasks involving one location (**H2.1**), and small multiples faring better for tasks involving a region or all locations (**H2.2**). Completion times and error rates (means and CIs) are found in the middle column of Fig. 8.

Completion Time: When considering OneLocation, we observe that completion time is indeed on average lower for `Barchart1M` (22.2sec), followed by `GlyphSM` (25.8sec) and then `DorlingSM` (29.3sec). Looking at the mean differences, there is evidence that `Barchart1M` is faster than `DorlingSM` (by 7sec on average). It may

[3] A CI of *differences* that does not cross 0 provides evidence of differences - the further away from 0 and the smaller the CI the stronger the evidence.

[4] We counterbalanced visualization order across participants to mitigate learning (Sec. 4.1). An unplanned analysis indicates that although participants improved over sessions (performed best in the 3rd visualization presented than in the 1st), there was indeed no evidence of asymmetric learning across `Barchart1M` and `GlyphSM`, thus counterbalancing worked for them (there is some Time improvement for `DorlingSM`). Analyses/charts are available as supplementary material.

© 2019 IEEE. This is the author's version of the article that has been published in IEEE Transactions on Visualization and Computer Graphics. The final version of this record is available at: 10.1109/TVCG.2019.2934807

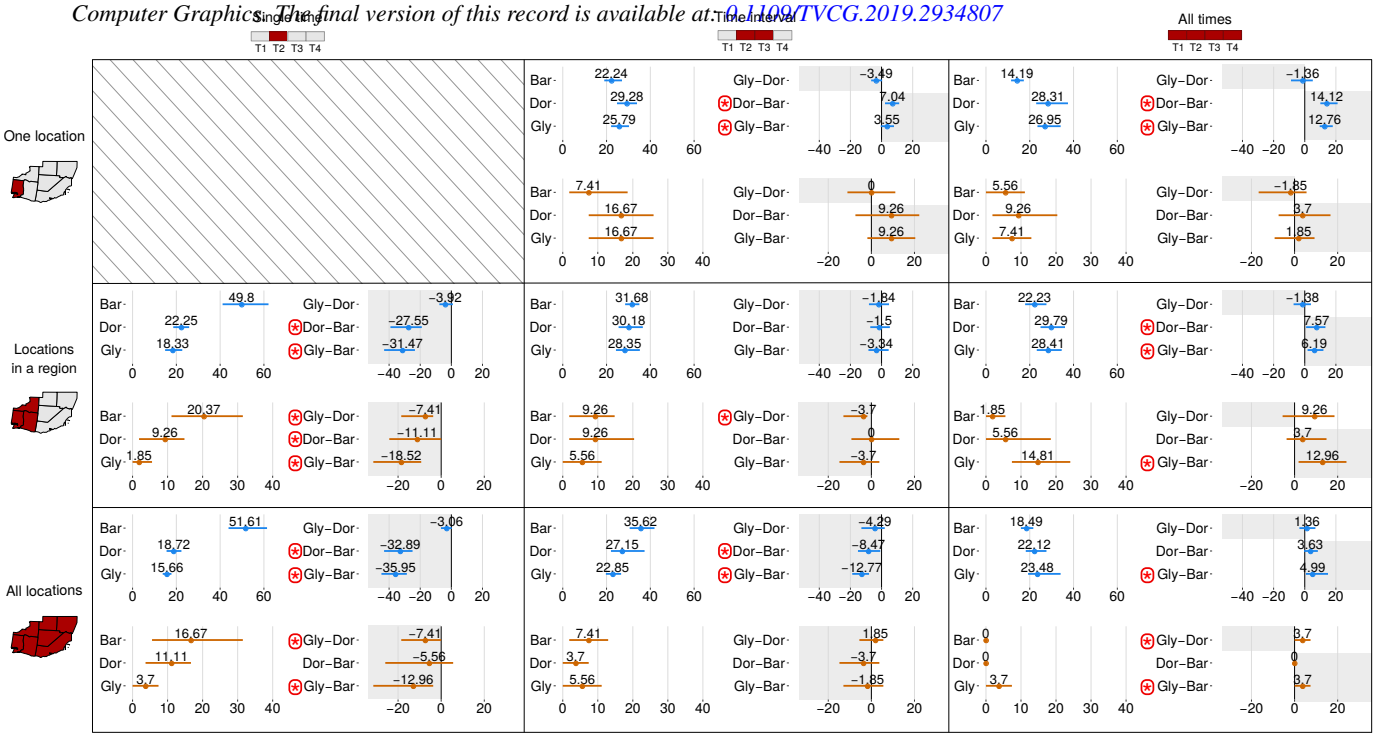


Fig. 8. Results for Completion Time (sec) and Error Rate (in %) for each task in Fig. 2. In each cell (task), Mean values per visualization are seen on the left and means of pairwise differences on the right. Error bars represent 95% Bootstrap confidence intervals. Gray rectangles indicate the direction of our hypotheses. Evidence of differences are marked with a ⊛ (the further away from 0 and the tighter the CI, the stronger the evidence).

be the case that `Barchart1M` is also faster than `GlyphSM` and that `GlyphSM` is faster than `DorlingSM`, but evidence is not conclusive.

The completion time for REGION is close for all three techniques (`GlyphSM` 28.4sec, `DorlingSM` 30.2sec, and `Barchart1M` 31.7sec) and we do not have evidence of differences looking at the mean differences. The same pattern is found in ALLLOCATIONS as `GlyphSM` (22.8sec) is faster than the other techniques, followed by `DorlingSM` (27.1sec) and `Barchart1M` (35.6sec). Looking at the mean differences, we have evidence that `Barchart1M` is slower than both `GlyphSM` and `DorlingSM` (by 12.7sec and 8sec on average). We do not have evidence of a difference between `GlyphSM` and `DorlingSM`.

Error Rate: For these tasks, we observe that the lowest error rate depends on the geographical granularity considered. `Barchart1M` is better for ONELOCATION (7.4%), `GlyphSM` for REGION (5.6%) and `DorlingSM` for ALLLOCATIONS (3.7%). Looking at mean differences for ALLLOCATIONS there is indeed evidence that `DorlingSM` is more error prone than `GlyphSM` (by 3.7% on average) for REGION, but no evidence of other differences.

Confidence: The second row of Fig. 9 shows the self-reported confidence for TIME INTERVAL. We observe that confidence for ONELOCATION is high in more than half of the trials for `Barchart1M` and `GlyphSM` (over 60%), but lower for `DorlingSM` (45%). For tasks in REGION and ALLLOCATIONS, we observe that it is higher for both `GlyphSM` and `DorlingSM` (over 60%) and lower for `Barchart1M` (54% and 50% respectively).

Summary for TIMEINTERVAL: The tendencies for the three tasks that focus on correlations for a time interval change significantly depending on the spatial granularity. For a single location, `Barchart1M` is faster than the small multiple techniques (`GlyphSM`,`DorlingSM`), supporting **H2.1**. This behavior is reversed when considering all locations on the map. `Barchart1M` becomes the slowest visualization, supporting the part of **H2.2** related to all locations. In both tasks, we found no evidence of difference in error rates. The situation is less clear when multiple locations in a region have to be considered. We found no evidence of differences for any of the measures, contrary to the prediction of **H2.2** related to geographical regions. We observe no difference between `GlyphSM` and `DorlingSM`. **H4** is thus not supported.

5.2.3 ALL TIME

In tasks involving all time steps we expect a single map (*i.e.*, `Barchart1M`) to fare better (**H3**). Completion times and error rates (means and CIs) for these tasks are in the rightmost column of Fig. 8.

Completion Time: is lower with `Barchart1M` than with both small-multiples visualizations. Looking at the mean differences, there is strong evidence that `Barchart1M` is faster than `GlyphSM` and `DorlingSM` for both ONELOCATION and REGION tasks. For ALLLOCATIONS task, there is also strong evidence that `Barchart1M` is faster than `GlyphSM` (by 4.9sec on average) but evidence is not conclusive regarding `Barchart1M` being faster than `DorlingSM`. There is no evidence of a difference between `GlyphSM` and `DorlingSM` for any geographical granularity.

Error Rate: is lowest in `Barchart1M` for ONELOCATION and REGION tasks. For ALLLOCATIONS, the error rate is 0% for both `Barchart1M` and `DorlingSM` (and thus, no CI is computed). There is evidence that `Barchart1M` is less prone to errors than `GlyphSM` for REGION, but this evidence is weak for ALLLOCATIONS (and we see no evidence of a difference for ONELOCATION). There is also weak evidence that `DorlingSM` is also less error prone than `GlyphSM` (by 3.7%) for ALLLOCATIONS.

Confidence: is high in over 60% of trials for most visualizations and geographic granularities, with high-confidence trials for `DorlingSM` being a bit lower (around 50% of trials) for the ONELOCATION and REGION.

Summary for ALLTIME: The tendencies for the three tasks that focus on correlations over all time steps are fairly similar, with `Barchart1M` being generally faster than small multiples (`GlyphSM`, `DorlingSM`), thus supporting **H3**. Again, we do not find evidence of a difference between `GlyphSM` and `DorlingSM`. **H4** is not supported.

6 PER-TASK DISCUSSION AND DESIGN RECOMMENDATIONS

We observed that, overall, small multiples were faster across tasks, but their error rates were not different from those of a map with bar charts. Nevertheless, as hypothesized, looking at the individual tasks we see that the performance changes depending on the task at hand. Next,

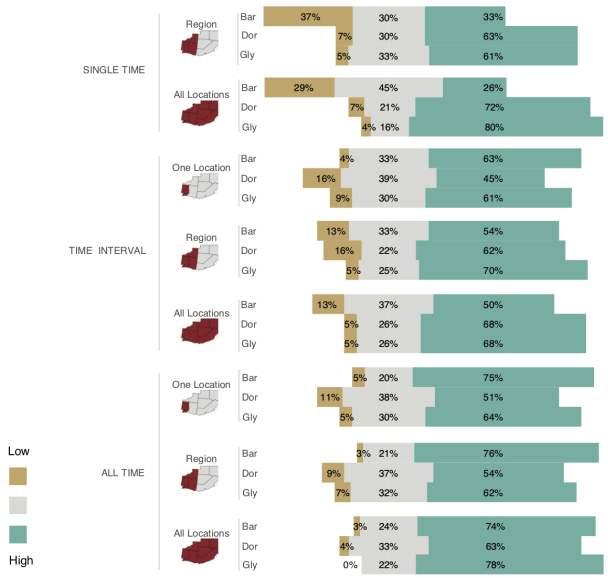


Fig. 9. Reported self-confidence per task (in %).

we summarize and discuss our findings, and distill them into design recommendations (summarized in Fig. 10).

SINGLETIME: The correlation of thematic variables on geographic maps has been studied before for a single point in time [15, 18]. We add to these findings, by identifying that the tendencies for correlation tasks on a single point in time are similar, irrespective of whether we consider a geographical region or all locations. Using the small multiple visualizations (`GlyphSM`, `DorlingSM`), participants were almost twice as fast as when using a single map with bar charts (`Barchart1M`), as they only needed to focus on a single cell of the small multiples, since that cell juxtaposes all spatial information for that point in time. The tasks are slower with a single map with bar charts (`Barchart1M`), since participants needed to visually search for the specific time step across multiple bar charts and synthesize their findings. Error rates for these tasks follow similar trends. Our findings thus confirm **H1**.

When it comes to small multiples, there is a tendency for the proportional symbol map (`GlyphSM`) to be less error prone than the Dorling cartogram (`DorlingSM`), supporting **H4**. This is likely the case because the position of symbols shifts between multiples in the cartogram case, making it hard to re-identify them. This tendency was also observed when comparing proportional symbol maps with non-contiguous cartograms in the work of Gao *et al.* [18]. However, in their case, it was for the overall performance over multiple tasks, not just for correlation identification, and the differences observed were not significant.

These tendencies were consistent with the self-perceived difficulty of conducting these tasks in the exit questionnaire. It was stated often that it is hard to make analyses for one time step with bar charts.

R1: *For identifying correlations at a specific point in time, small-multiples visualizations are better.*

TIMEINTERVAL: When participants have to identify correlation tendencies and evolution over a time interval, the situation is less clear. The tendencies change significantly depending on the geographic granularity (consistent with **H2**). When considering a single location, a single map with bar charts (`Barchart1M`) is faster than the small-multiples techniques (`GlyphSM`, `DorlingSM`), as all temporal information is grouped closely together and participants just needed to identify the temporal interval on a single bar chart. Whereas for small multiples, after identifying the relevant time cells, participants needed to then identify, in each cell, the specific location and collate their findings. This is consistent with **H2.1**.

The findings are reversed when considering all locations, consistent with **H2.2**. Here, a map with bar charts is slower, because it is the visualization where information is scattered and needs to be collated from across different areas. Participants first had to go through all (or almost all) bar charts to identify the specific interval, and collate the information to identify tendencies. Whereas for small multiples, they only needed to focus on a few time steps and look for overall patterns.

One of the most interesting findings from this study is the inconclusive evidence for tasks where a geographic region has to be considered across a time interval (this part of **H2.2** is not confirmed). The lack of observed differences may be due to low statistical power. But we believe it is more likely due to this task being more balanced in the amount of information that needs to be collated across different areas for the different techniques. Here, for a single map with bar charts, participants still had to identify the specific bars across multiple bar charts – but not all of them. When using small multiples, they could focus on a few time steps, but still had to identify the desired geographic area in each one of them. There is likely a tradeoff when it comes to tasks that involve spatial regions and time intervals. When considering subsets of time, it looks like the less spatial locations have to be considered, the better a single map is. Inversely, the more spatial locations, the better small multiples become. More generally, it is likely that a single map with bar charts likely works best for simple geography and complex temporal patterns, and small multiples when geography is complex but the temporal variability is simple. Future work needs to determine exactly when to transition between visualizations. We are not aware of any previous work that has considered correlation tasks that require gathering information across subsets of space and time.

R2: *For identifying correlations and temporal evolution over a subset of time steps and a subset of locations, there is no clear winner. If there are only a few locations, consider using a single map with bar charts. If there are many locations, prefer small multiples.*

ALL TIME: The tendencies for the three tasks that focus on correlations for all time steps are again consistent, with `Barchart1M` being faster, in accordance with **H3**. Even though participants had to collate both spatial and temporal information, a single map with bar charts was faster. This representation makes it easy to see trends over time (correlation and monotonic evolution) that are juxtaposed in the individual bar charts. Collating this information seems to be fast irrespective of how many geographic regions are taken into account. Small multiples seem slower, likely because determining temporal trends necessitates comparing several locations across cells before identifying a trend.

The self-perceived difficulty to conduct the task for all time steps was also consistent with objective measures. A single map with bar charts was perceived, overall, as easier to use than both small-multiples visualizations, and several participants commented that it was easy to observe evolution over time on the single map with bar charts.

R3: *For identifying correlations and temporal evolution over all time steps, irrespective of the number of locations, a single map with bar charts is better.*

Small multiples: We found evidence that the two small-multiple techniques (`GlyphSM`, `DorlingSM`) were different mainly when considering a single point in time (partially confirming **H4**), with `DorlingSM` being slower and more error prone. Participants' comments indicate that they had difficulty matching a location, or sets of locations, across small multiples with `DorlingSM`, since positions of locations shifted. Nevertheless, this cost is not seen in tasks considering more than one time step. This may be due to low statistical power, or because this cost is small when it comes to more challenging tasks (time intervals or all time steps) that require collating information across small multiples.

R4: *For small multiples, there is some evidence that proportional symbol maps are better than Dorling, especially for a single time point.*

7 General Discussion

Our findings generally followed our original hypothesis. We thus believe that our reasoning, that difficulty in each technique depends on whether the information to be collated is juxtaposed or distant, is sound; and that our results reflect true tendencies.

The one exception relates to correlation tasks on subranges in time and space. We had originally thought that small-multiple variations would prevail in this situation, but we were unable to detect a trend. We believe that our setup of this task may reflect a similar difficulty in collating temporal information (for small multiples) and spatial

© 2019 IEEE. This is the author's version of the article that has been published in IEEE Transactions on Visualization and Computer Graphics. The final version of this record is available at: 0.1109/TVCG.2019.2934807

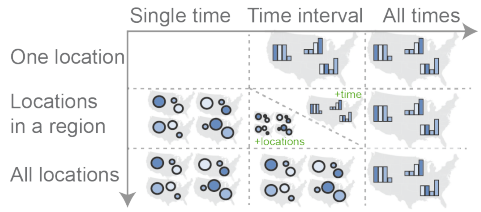


Fig. 10. Summary of our recommendations for the different tasks. For tasks on subranges of space and time (middle cell) there is no clear winner, but the table structure suggests small multiples work best for increased spatial complexity, and a single map with bar charts for increased temporal complexity (although the transition point is not known).

information (for the single map with bar charts). Looking at Fig. 10, we observe that this middle cell is a transition cell between tasks better suited to small multiples and tasks better suited to single maps. For example, if we look at the middle column (Time interval), it seems that one (and likely a few) geographic locations are best seen on single maps, but as more locations are added small multiples become more competitive. Or if we look at the middle row (Locations in a region) it seems that one (and likely a few) time steps may be better viewed with small multiples, but as time increases a single map with bar charts becomes better. It is interesting to consider what are the tipping points of these shifts (number of time steps, number of geographic locations), in order to determine when to transition between visualizations.

For all visualizations, collating information across different areas (bars from different bar charts for the single map, and locations across cells for the small multiples) is challenging. In our study, we focus on static visualizations, but the addition of highlighting would likely reduce the differences we found, by making it easier to collate information (*e.g.*, highlight Washington in all small multiples, or 2011 in all bar charts). Nevertheless, we believe the high-level effects would still hold (to a lesser extent) as they are due to the fact that information is dispersed across the visualization. If filtering is considered, we believe behavior will likely revert to the results at the corners of Fig. 10. For example, filtering on time interval 2009-2011 would either remove or fade other years out, making this a task similar to ones over all time steps. Similarly, if the East US is the focus, the system would remove or fade other locations out, making this a task similar to those involving all locations. More importantly, the actions performed to select or filter time steps or geographical locations could themselves be used as an indication of what is the user's focus, and used to transition to the best visualization for the task.

7.1 Limitations and Future Work

Interaction was deliberately not considered in this first study, as we primarily aimed at evaluating the specific influence of space and time at different granularities on users' ability to identify correlations with different visualizations. Thus, we wanted to avoid adding further factors to an already complex experimental design. Our discussion section above provides initial thoughts about how interaction could affect our results, but further work is needed to verify them, and to consider the use of interaction as a means to transition between visualizations.

The number of steps used to detect correlation in our tasks is limited (nine time steps per location), which required us to use datasets with a strong relationship between two variables. Data extracted from measures of real world phenomena is unlikely to present such strong patterns, making it harder to detect potential correlations. This could alter our results, although we believe the general trends would persist.

Another limitation is that we only used a single map of the US, which necessarily captures only a subset of geographical configurations. It is possible that countries with more diverse shapes (*e.g.* Chile, Italy or United Kingdom) would lead to different results, as the identification of individual locations or regions might be different. We attempted to mitigate any bias in identifying the locations of interest by displaying, prior to each trial, the geographic region of interest. Nevertheless, further experimentation is needed. Moreover, diversity of irregularity of locations can impact spatial autocorrelation in geospatial visualizations that use irregular geometries to represent thematic variables, such as choropleths [4, 42, 67]. While it is possible that effects might differ somewhat in other types of maps, we feel that the general trends should hold: our techniques use regular shapes to represent thematic variables, and thus the size and number of items compared likely weigh more in the complexity of the task (*e.g.*, occlusion or clutter of elements might impact the interpretation of patterns). To this end, in our trials we varied the size of locations and their density. Finally, although the analysis of data using a map of a known country could have led to bias given preconceptions about the geographical distribution, we believe this to be unlikely given the extensive training, and the number of map features and time steps involved.[5] In summary, while we believe that overall trends would persist across different maps, future work needs to consider more diverse geographic maps.

For the small multiples tested, we expected that Dorling cartograms (*i.e.*, visualizations that use visual channels of the maps features themselves to encode thematic variables) would perform worst than proportional symbol maps, as was the case in previous work [18, 30]. In our context this was observed mainly when considering tasks at a single point in time. It is very likely this effect will be more pronounced in other spatial tasks that involve more continuous geographic changes and correlations that vary spatially (*e.g.*, identifying transmission patterns).

We recruited users who were already knowledgeable about visualization, and gave them additional training. Opportunities for such training may not be available to the general public. While we believe general trends will still apply, it is possible that non-trained users would have lower accuracy rates or would not dedicate as much time as our participants to perform the tasks. Additionally, they might be more familiar with one of the three tested techniques, which would bias results in its favor. Future work should investigate the learning curve of each technique and analyze how well they fare when used by novices with a more diverse background and lack of training. A next step in that direction would be to conduct a crowdsourced study.

Finally, we decided to combine two different association tasks in one (*i.e.*, the type of correlation and its evolution), as we felt they were tightly coupled when performed in the context of geo-temporal analysis. Due to this combination, our analysis does not provide finer details on the difficulty of each subtask. Future work could study each one separately to gain more insights about how correlation and trends are detected individually. For example, we expect that complex temporal tasks, such as detecting and characterizing monotonic evolution, is easier on single maps with bar charts (as each one directly encodes this evolution); whereas complex geographical tasks, such as detecting transmission patterns, may be easier with small multiples.

8 Conclusions

We presented a study on identifying correlation in spatio-temporal visualizations. We considered eight tasks that associate two variables over different granularity levels for both time and space. The compared visualizations combine different strategies to represent thematic variables: juxtaposing either time or space (a single map with bar charts *vs.* small-multiple maps); encoding variables either using symbols overlaid on top of map features, or using visual channels of the map features themselves (proportional symbol maps *vs.* cartograms). We provide a set of design recommendations depending on the task at hand. In our context, the technique using the map features' own visual channels to encode thematic variables (cartograms) performed worst only when a single point in time was considered. Our results further indicate that for tasks that consider the evolution over all time steps, a visualization that represents data on a single map (juxtaposing time) is more effective and easier to interpret than small multiples. Small multiples (juxtaposing space) are better suited for tasks that require the comparison of variables for one point in time over several geographical locations. When dealing with time intervals and spatial regions, our results suggest that there is a continuum of performance between visual representations (juxtapose time *vs.* space), raising questions for future research.

[5] We did not find any warning signs of such pitfalls (*e.g.*, participants taking very little time to finish the tasks and making numerous errors).

REFERENCES

[1] M. J. Alam, S. G. Kobourov, and S. Veeramoni. Quantitative measures for cartogram generation techniques. *Comput. Graph. Forum*, 34(3):351–360, June 2015. doi: 10.1111/cgf.12647

[2] L. An, M.-H. Tsou, S. E. Crook, Y. Chun, B. Spitzberg, J. M. Gawron, and D. K. Gupta. Space–time analysis: Concepts, quantitative methods, and future directions. *Annals of the Association of American Geographers*, 105(5):891–914, 2015. doi: 10.1080/00045608.2015.1064510

[3] N. Andrienko and G. Andrienko. Interactive visual tools to explore spatio-temporal variation. In *Proc. of the working conference on Advanced visual interfaces (AVI)*, pp. 417–420. ACM, 2004. doi: 10.1145/989863.989940

[4] R. Beecham, J. Dykes, W. Meulemans, A. Slingsby, C. Turkay, and J. Wood. Map LineUps: Effects of spatial structure on graphical inference. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):391–400, Jan. 2017. doi: 10.1109/TVCG.2016.2598862

[5] M. Bostock, V. Ogievetsky, and J. Heer. D3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011. doi: 10.1109/TVCG.2011.185

[6] I. Boyandin, E. Bertini, and D. Lalanne. A qualitative study on the exploration of temporal changes in flow maps with animation and Small-Multiples. *Comput. Graph. Forum*, 31(3pt2):1005–1014, June 2012. doi: 10.1111/j.1467-8659.2012.03093.x

[7] H. Chernoff. The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, 68(342):361–368, 1973. doi: 10.1080/01621459.1973.10482434

[8] F. Chevalier, N. H. Riche, C. Plaisant, A. Chalbi, and C. Hurter. Animations 25 years later: New roles and opportunities. In *Proc. of the International Working Conference on Advanced Visual Interfaces*, AVI '16, pp. 280–287. ACM, 2016. doi: 10.1145/2909132.2909255

[9] P. Craig, N. R. Seïler, and A. D. O. Cervantes. Animated geo-temporal clusters for exploratory search in event data document collections. In *18th International Conference on Information Visualisation (IV)*, pp. 157–163. IEEE, 2014. doi: 10.1109/IV.2014.69

[10] G. Cumming. The new statistics: Why and how. *Psychological Science*, 25(1):7–29, 2014. doi: 10.1177/0956797613504966

[11] G. Cumming and S. Finch. Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, pp. 170–180, 2005. doi: 10.1037/0003-066X.60.2.170

[12] D. Dorling. *Area Cartograms: Their Use and Creation*, chap. 3.7, pp. 252–260. John Wiley & Sons, Ltd, 2011. doi: 10.1002/9780470979587.ch33

[13] P. Dragicevic. Fair statistical communication in HCI. In J. Robertson and M. Kaptein, eds., *Modern Statistical Methods for HCI*, chap. 13, pp. 291–330. Springer International Publishing, 2016. doi: 10.1007/978-3-319-26633-6_13

[14] Y. Du, L. Ren, Y. Zhou, J. Li, F. Tian, and G. Dai. Banded choropleth map. *Pers. Ubiquit. Comput.*, 22(3):503–510, June 2018. doi: 10.1007/s00779-018-1120-y

[15] M. E. Elmer. Symbol considerations for bivariate thematic maps. *In Proceedings of the 26th International Cartography Conference (ICC)*, 2013.

[16] C. Fish, K. P. Goldsberry, and S. Battersby. Change blindness in animated choropleth maps: An empirical study. *Cartogr. Geogr. Inf. Sci.*, 38(4):350–362, Jan. 2011. doi: 10.1559/15230406384350

[17] G. Fuchs and H. Schumann. Visualizing abstract data on maps. In *Proceedings of the 8th International Conference on Information Visualisation (IV)*, pp. 139–144. IEEE, 2004. doi: 10.1109/IV.2004.1320136

[18] P. Gao, Z. Li, and Z. Qin. Usability of value-by-alpha maps compared to area cartograms and proportional symbol maps. *Journal of Spatial Science*, pp. 1–17, Mar. 2018. doi: 10.1080/14498596.2018.1440649

[19] P. Gatalsky, N. V. Andrienko, and G. L. Andrienko. Interactive analysis of event data using space-time cube. *Proceedings of the 8th International Conference on Information Visualisation (IV)*, pp. 145–152, 2004. doi: 10.1109/IV.2004.1320137

[20] S. Goodwin, J. Dykes, A. Slingsby, and C. Turkay. Visualizing multiple variables across scale and geography. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):599–608, Jan. 2016. doi: 10.1109/TVCG.2015.2467199

[21] A. L. Griffin, A. M. MacEachren, F. Hardisty, E. Steiner, and B. Li. A comparison of animated maps with static small-multiple maps for visually identifying space-time clusters. *Annals Assoc. of American Geographers*, 96(4):740–753, 2006. doi: 10.1111/j.1467-8306.2006.00514.x

[22] D. Guo, J. Chen, A. M. MacEachren, and K. Liao. A visualization system for space-time and multivariate patterns (vis-stamp). *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1461–1474, 2006. doi: 10.1109/TVCG.2006.84

[23] H. Hagh-Shenas, S. Kim, V. Interrante, and C. Healey. Weaving versus blending: a quantitative assessment of the information carrying capacities of two alternative methods for conveying multivariate data with color. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1270–1277, Nov. 2007. doi: 10.1145/1179622.1179848

[24] R. Han, Z. Li, P. Ti, and Z. Xu. Experimental evaluation of the usability of cartogram for representation of GlobeLand30 data. *ISPRS International Journal of Geo-Information*, 6(6):180, 2017. doi: 10.3390/ijgi6060180

[25] R. L. Harris. *Information Graphics: A Comprehensive Illustrated Reference*. Oxford University Press, 2000.

[26] L. Harrison, F. Yang, S. Franconeri, and R. Chang. Ranking Visualizations of Correlation Using Weber's Law. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1943–1952, Dec. 2014. doi: 10.1109/TVCG.2014.2346979

[27] M. Harrower. The cognitive limits of animated maps. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 42(4):349–357, 2007. doi: 10.3138/carto.42.4.349

[28] J. Jo, F. Vernier, P. Dragicevic, and J. Fekete. A declarative rendering model for multiclass density maps. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):470–480, Jan 2019. doi: 10.1109/TVCG.2018.2865141

[29] T. Johnson, C. Acedo, S. Kobourov, and S. Nusrat. Analyzing the evolution of the internet. In *17th IEEE Eurographics Conference on Visualization (EuroVis)*, vol. 17, 2015. doi: 10.2312/eurovisshort.20151123

[30] S. Kaspar, S. I. Fabrikant, and P. Freckmann. Empirical study of cartograms. In *25th International Cartographic Conference (ICC)*, vol. 3, p. 5, 2011.

[31] M. Kay and J. Heer. Beyond weber's law: A second look at ranking visualizations of correlation. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):469–478, Jan. 2016. doi: 10.1109/TVCG.2015.2467671

[32] S. Kim, S. Jeong, I. Woo, Y. Jang, R. Maciejewski, and D. Ebert. Data flow analysis and visualization for spatiotemporal statistical data without trajectory information. *IEEE Transactions on Visualization and Computer Graphics*, PP(99):1–1, 2017. doi: 10.1109/TVCG.2017.2666146

[33] S. Kim, R. Maciejewski, A. Malik, Y. Jang, D. S. Ebert, and T. Isenberg. Bristle maps: a multivariate abstraction technique for geovisualization. *IEEE Transactions on Visualization and Computer Graphics*, 19(9):1438–1454, Sept. 2013. doi: 10.1109/TVCG.2013.66

[34] A. Klippel, F. Hardisty, and R. Li. Interpreting spatial patterns: An inquiry into formal and cognitive aspects of Tobler's first law of geography. *Annals of the Association of American Geographers*, 101(5):1011–1031, 2011. doi: 10.1080/00045608.2011.577364

[35] M.-J. Kraak. The space-time cube revisited from a geovisualization perspective. In *Proc. 21st International Cartographic Conference (ICC)*, pp. 1988–1996, 2003.

[36] M. Krzywinski and N. Altman. Points of significance: Error bars. *Nature Methods*, 10(10):921–922, Oct. 2013. doi: 10.1038/nmeth.2659

[37] J. Li, Z.-P. Meng, M.-L. Huang, and K. Zhang. An interactive radial visualization of geoscience observation data. In *Proceedings of the 8th International Symposium on Visual Information Communication and Interaction*, VINCI '15, pp. 93–102. ACM, 2015. doi: 10.1145/2801040.2801061

[38] Y.-N. Li, D.-J. Li, and K. Zhang. Metaphoric transfer effect in information visualization using glyphs. In *Proceedings of the 8th International Symposium on Visual Information Communication and Interaction*, VINCI '15, pp. 121–130. ACM, 2015. doi: 10.1145/2801040.2801062

[39] M. A. Livingston and J. W. Decker. Evaluation of trend localization with multi-variate visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2053–2062, Dec. 2011.

[40] M. A. Livingston and J. W. Decker. Evaluation of multivariate visualizations: a case study of refinements and user experience. In *Visualization and Data Analysis 2012*, vol. 8294, p. 82940G. SPIE - International Society for Optics and Photonics, Jan. 2012. doi: 10.1117/12.912192

[41] L. McNabb, R. S. Laramee, and R. Fry. Dynamic choropleth maps – using amalgamation to increase area perceivability. In *Proceedings of the 22nd International Conference Information Visualisation (IV)*, pp. 284–293, July 2018. doi: 10.1109/IV.2018.00056

[42] L. McNabb, R. S. Laramee, and M. Wilson. When size matters: towards

© 2019 IEEE. This is the author's version of the article that has been published in IEEE Transactions on Visualization and Computer Graphics. The final version of this record is available at: 0.1109/TVCG.2019.2934807

evaluating perceivability of choropleths. In *Proceedings of the Conference on Computer Graphics & Visual Computing*, pp. 163–171. Eurographics Association, 2018. doi: 10.2312/cgvc.20181221

[43] T. Nakaya and K. Yano. Visualising crime clusters in a space-time cube: An exploratory data-analysis approach using space-time kernel density estimation and scan statistics. *Transactions in GIS*, 14(3):223–239, 2010. doi: 10.1111/j.1467-9671.2010.01194.x

[44] S. Nusrat, M. J. Alam, and S. Kobourov. Evaluating cartogram effectiveness. *IEEE Transactions on Visualization and Computer Graphics*, 24(2):1077–1090, Feb. 2018. doi: 10.1109/TVCG.2016.2642109

[45] S. Nusrat, M. J. Alam, C. Scheidegger, and S. Kobourov. Cartogram visualization for bivariate Geo-Statistical data. *IEEE Transactions on Visualization and Computer Graphics*, 24(10):2675–2688, Oct. 2018. doi: 10.1109/TVCG.2017.2765330

[46] S. Nusrat and S. Kobourov. The state of the art in cartograms. *Comput. Graph. Forum*, 35(3):619–642, June 2016. doi: 10.1111/cgf.12932

[47] J. H. Park, S. Nadeem, and A. Kaufman. GeoBrick: exploration of spatiotemporal data. *The Visual Computer*, 35(2):191–204, Feb 2019. doi: 10.1007/s00371-017-1461-y

[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[49] T. V. Perneger. What's wrong with bonferroni adjustments. *BMJ*, 316(7139):1236–1238, 1998. doi: 10.1136/bmj.316.7139.1236

[50] S. Peters and L. Meng. Visual analysis for nowcasting of multidimensional lightning data. *ISPRS International Journal of Geo-Information*, 2(3):817–836, 2013. doi: 10.3390/ijgi2030817

[51] D. J. Peuquet. It's about time: A conceptual framework for the representation of temporal dynamics in geographic information systems. *Ann. Assoc. Am. Geogr.*, 84(3):441–461, Sept. 1994. doi: 10.1111/j.1467-8306.1994.tb01869.x

[52] R. A. Rensink. The nature of correlation perception in scatterplots. *Psychon. Bull. Rev.*, 24(3):776–797, June 2017. doi: 10.3758/s13423-016-1174-7

[53] R. A. Rensink and G. Baldridge. The perception of correlation in scatterplots. *Comput. Graph. Forum*, 29(3):1203–1210, Aug. 2010. doi: 10.1111/j.1467-8659.2009.01694.x

[54] G. Robertson, R. Fernandez, D. Fisher, B. Lee, and J. Stasko. Effectiveness of animation in trend visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1325–1332, Nov 2008. doi: 10.1109/TVCG.2008.125

[55] H. Rosling. The best stats you've ever seen. `https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen`, 2006. TED, Last accessed: 2019-03-27.

[56] R. E. Roth. An empirically-derived taxonomy of interaction primitives for interactive cartography and geovisualization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2356–2365, Dec. 2013. doi: 10.1109/TVCG.2013.130

[57] A. Satyanarayan, R. Russell, J. Hoffswell, and J. Heer. Reactive vega: A streaming dataflow architecture for declarative interactive visualization. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):659–668, Jan 2016. doi: 10.1109/TVCG.2015.2467091

[58] R. Scheepens, H. v. d. Wetering, and J. J. v. Wijk. Non-overlapping aggregated multivariate glyphs for moving objects. In *IEEE Pacific Visualization Symposium*, pp. 17–24, Mar. 2014. doi: 10.1109/PacificVis.2014.13

[59] J. Schiewe. Task-Oriented visualization approaches for landscape and urban change analysis. *ISPRS International Journal of Geo-Information*, 7(8):288, July 2018. doi: 10.3390/ijgi7080288

[60] B. Shneiderman. Why not make interfaces better than 3d reality? *IEEE Comput. Graph. Appl.*, 23(6):12–15, Nov. 2003. doi: 10.1109/MCG.2003.1242376

[61] A. Slingsby. Tilemaps for summarising multivariate geographical variation. In *IEEE VIS Poster*, 2018.

[62] J. Stewart and P. J. Kennelly. Illuminated choropleth maps. *Annals of the Association of american Geographers*, 100(3):513–534, 2010. doi: 10.1080/00045608.2010.485449

[63] H. Sun and Z. Li. Effectiveness of cartogram for the representation of spatial data. *The Cartographic Journal*, 47(1):12–21, 2010. doi: 10.1179/000870409X12525737905169

[64] C. Tominski, P. Schulze-Wollgast, and H. Schumann. 3D information visualization for time dependent data on maps. In *Proceedings of the 9th International Conference on Information Visualisation (IV)*, pp. 175–181. IEEE, 2005. doi: 10.1109/IV.2005.3

[65] B. Tversky, J. B. Morrison, and M. Betrancourt. Animation: Can it facilitate? *Int. J. Hum.-Comput. Stud.*, 57(4):247–262, Oct. 2002. doi: 10.1006/ijhc.2002.1017

[66] J. A. Tyner. *Principles of map design*. Guilford Publications, 2014.

[67] M. O. Ward, G. Grinstein, and D. Keim. *Interactive data visualization: foundations, techniques, and applications*. AK Peters/CRC Press, 2015.

[68] C. M. Wittenbrink, A. T. Pang, and S. K. Lodha. Glyphs for visualizing uncertainty in vector fields. *IEEE Transactions on Visualization and Computer Graphics*, 2(3):266–279, Sept. 1996. doi: 10.1109/2945.537309

[69] F. Yang, L. T. Harrison, R. A. Rensink, S. L. Franconeri, and R. Chang. Correlation judgment and visualization features: A comparative study. *IEEE Transactions on Visualization and Computer Graphics*, 25(3):1474–1488, Mar. 2019. doi: 10.1109/TVCG.2018.2810918

[70] B. Yost and C. North. Single complex glyphs versus multiple simple glyphs. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '05, pp. 1889–1892. ACM, 2005. doi: 10.1145/1056808.1057048

[71] J. Zhang, A. Malik, B. Ahlbrand, N. Elmqvist, R. Maciejewski, and D. S. Ebert. TopoGroups: Context-Preserving visual illustration of Multi-Scale spatial aggregates. In *Proceedings of the Conference on Human Factors in Computing Systems*, CHI '17, pp. 2940–2951. ACM, 2017. doi: 10.1145/3025453.3025801