RESEARCH-ARTICLE

Text-Color Hybrid Labeling for Multiclass Map Visualization: A Comparative Evaluation of Four Annotation Strategies

XINYAO CHEN, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, China

XINYUAN ZHANG, Chinese University of Hong Kong, Hong Kong, Hong Kong

TENG MA, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, China

LINGYUN YU, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, China

YU LIU, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, China

Published: 01 December 2025

Citation in BibTeX format

VINCI 2025: Proceedings of the 18th International Symposium on Visual Information Communication and Interaction
December 1 - 3, 2025
Linz, Austria

Text-Color Hybrid Labeling for Multiclass Map Visualization: A Comparative Evaluation of Four Annotation Strategies

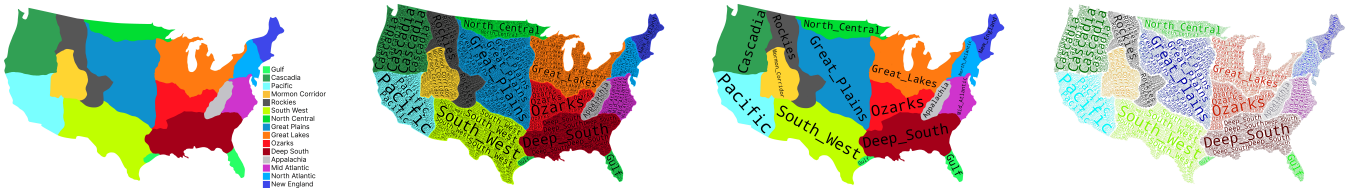
Xinyao Chen
Xi' an Jiaotong-liverpool University
Suzhou, Jiangsu, China
Xinyao.Chen21@student.xjtlu.edu.cn

Xinyuan Zhang
The Chinese University of Hong Kong, Shenzhen
Suzhou, Jiangsu, China
xinyuanzhang2@link.cuhk.edu.cn

Teng Ma
Department of Computing
Xi' an Jiaotong-liverpool University
Suzhou, Jiangsu, China
teng.ma@xjtlu.edu.cn

Lingyun Yu
Department of Computing
Xi' an Jiaotong-liverpool University
Suzhou, China
Lingyun.Yu@xjtlu.edu.cn

Yu Liu*
Department of Computing
Xi'an Jiaotong-Liverpool University
Suzhou, Jiangsu, China
Yu.Liu02@xjtlu.edu.cn



(a) Legend-Aside Map (b) Label-Fill Map (c) Label-Fit Map (d) Colored Label-Fill Map

Figure 1: Four text–color hybrid annotation strategies for multiclass map visualization. (a) Legend-Aside Map. (b) Label-Fill Map. (c) Label-Fit Map. (d) Colored Label-Fill Map.

Abstract

Prior work has identified the shortcomings of color-only encodings for maps with many categories, yet systematic comparisons of hybrid text–color strategies remain scarce. We therefore ran an 80-participant crowdsourced study on choropleth maps with 8–13 categories—approaching the 10-hue perceptual limit—to compare four annotation designs (Legend-Aside, Label-Fill, Label-Fit, Colored Label-Fill) across Count, Identify, Compare, and Rank tasks. Results show that the Label-Fit Map—with a single, large in-situ label—yields the highest accuracy and speed and ranks first in readability; Legend-Aside excels in simple counting and side-by-side comparisons. These findings deliver clear, task-specific guidelines for enhancing multiclass map readability and efficiency, informing the design of more effective map visualizations. All supplementary materials are available at our GitHub repository.

CCS Concepts

• **Human-centered computing** → **Visualization techniques**.

*Corresponding author

VINCI 2025, Linz, Austria

ACM ISBN 979-8-4007-1845-8/25/12
https://doi.org/10.1145/3769534.3769610

Keywords

Map Visualization, Crowd-source Experiment, Geographical Data

ACM Reference Format:
Xinyao Chen, Xinyuan Zhang, Teng Ma, Lingyun Yu, and Yu Liu. 2025. Text-Color Hybrid Labeling for Multiclass Map Visualization: A Comparative Evaluation of Four Annotation Strategies. In *Proceedings of the 18th International Symposium on Visual Information Communication and Interaction (VINCI 2025), December 01–03, 2025, Linz, Austria.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3769534.3769610

1 Introduction

Map data are ubiquitous across domains like urban planning [30], socio-economic analysis [31], and environmental monitoring [31]. These scenarios demand effective visualization techniques to convey both spatial distribution and inter-category relationships. Among these approaches, choropleth maps remain the most prevalent: each region is colored by category to reveal geographic patterns.

However, pure color encoding suffers severe limitations as category counts rise. Colin Ware [32] recommends using no more than ten distinct hues for reliable symbol identification against varied backgrounds; similar constraints are validated in studies on color-category perception (e.g., [6], [11]). In practice—whether mapping land-cover classifications, linguistic distributions, or political affiliations—datasets often include dozens of classes, far exceeding color's capacity. Consequently, users may face increased cognitive load, higher error rates, and constant legend consultation, all of which degrade speed and accuracy.

To mitigate these challenges, dual-encoding schemes that combine text and color have been proposed and applied [35]. Yet the design space—and their relative effectiveness—remains underexplored. In this work, we define and evaluate four hybrid text-color designs (**Legend-Aside Map**, **Label-Fill Map**, **Label-Fit Map**, and **Colored Label-Fill Map**) across four representative map-reading tasks: count, identification, comparison, and ranking. We measure task completion time, accuracy, and self-reported readability, comprehension, aesthetic appeal, and complexity.

Our contributions are twofold:

- A first systematic empirical comparison of four text-color hybrid map designs on real-world, multi-category data.
- Actionable, evidence-based design guidelines for categorical map visualization.

Our findings can inform GIS software defaults, cartography curricula, and standards for interactive dashboards.

2 Related Work

2.1 Map Data Visualization

In cartography, map data refers to spatial datasets that have been processed, generalized, and symbolized for effective thematic display [14, 26]. Such data marry geometric primitives (points, lines, and polygons) with rich attribute tables, then undergo projection, simplification, and styling to optimize legibility and emphasize the intended message [1]. Well-formatted map data serve as the backbone of decision-making across a wide spectrum of fields. In public health, disease-incidence maps guide intervention strategies [36]; in ecology, ecosystem-service inventories reveal supply–demand balances for sustainable management [27]. Urban planners leverage them for infrastructure siting [7]. Emergency responders depend on rapid, accurate map updates during disasters [2].

Various visualization techniques have been developed to render these structured datasets [13, 35]. The most prevalent forms include:

- Choropleth Maps: Entire areal units are shaded according to categorical or aggregated quantitative values, capitalizing on our ability to perceive hue differences to reveal spatial patterns and clusters.
- Proportional-Symbol Maps: Point symbols (e.g., circles, squares) are sized in accordance with underlying numeric values, allowing precise comparison of magnitudes at discrete locations.
- Dot-Density Maps: Uniform symbols ("dots") are randomly or regularly placed within regions to represent counts, conveying density through point frequency.
- Isarithmic (Isopleth) Maps: Contour lines or gradient bands interpolate continuous phenomena (such as elevation or temperature) across space.
- Cartograms: Geographic shapes are deliberately distorted so that their areas become proportional to a chosen variable, highlighting relative significance over geographic fidelity.
- Flow Maps: Arrows or lines of varying thickness depict movement volumes or directions between origins and destinations.

Among these, choropleth mapping remains the prevalent choice for visualizing both categorical and aggregated quantitative map data, due to its intuitive mapping of region → color → value (or class). Its widespread adoption underscores the importance of addressing the perceptual and cognitive limits inherent in color-based encoding—a challenge this work seeks to overcome through hybrid text–color annotation strategies.

2.2 Text in Visualization

In visualization design, text fulfills two principal roles: *annotation*—providing labels, captions, and explanatory notes, and *encoding*—conveying data values, categories, or textual information directly through typographic and spatial properties.

Annotation. . Text annotations, such as axis labels, legends, and tooltips, enrich visualizations by contextualizing graphical elements without altering their encoding. Direct labeling techniques—placing names on map regions or data points—minimize the need for external legends and reduce visual search time, thereby improving task performance in identification and comparison. Interactive text annotations further support drill-down analysis, allowing users to access details on demand while preserving a clean overall view.

Encoding. . When text operates as an encoding channel, it directly represents information through typography, positioning, or semantic content. This encompasses two common scenarios:

- **Category and Quantitative Encoding:** Typography (font size, weight, color) can map to numeric values or discrete classes, offsetting limitations of color and shape palettes in complex datasets. For example, hybrid map designs overlay labels within regions to disambiguate categories beyond palette constraints [3, 22].
- **Text Visualization:** In scenarios where the data itself is textual—such as document term frequencies or keyword importance—word clouds (tag clouds) use font size and spatial arrangement to encode frequency or relevance. Word cloud algorithms (e.g., Wordle [29], SparkClouds [17]) exemplify text-first encoding, revealing semantic patterns and term prominence through typographic variation and layout [10].

2.3 Text in Map Visualization

Text are integral to map visualizations, serving both as contextual annotations and as direct encoding channels tailored to geographic data. While annotation aligns closely with general visualization practices, encoding leverages text to overcome limitations inherent to color or shape in complex spatial contexts.

Annotation of Spatial Features. . Text annotations—such as place names, region labels, legends, and tooltips—anchor geographic features in context, eliminating separate lookup tables and speeding identification. Direct labeling of polygons and lines reduces cognitive load in tasks like point matching and region comparison [3]. In dense or cluttered maps, collision-avoidance algorithms and density-based placement heuristics ensure text remains legible without obscuring critical details [4, 18, 23].

Text as Encoding Channel. . When traditional graphical channels (e.g., color, shape, iconography) hit perceptual or semantic limits—particularly in multi-category or tag-rich maps—text operates as a robust encoding medium. Two main approaches have emerged:

- **Overlay Text Labels**: Text is rendered atop colored regions or map features, using the underlying fill to convey class while the overlaid label ensures immediate category recognition. Examples include intrinsic label-fill strategies that embed words within region shapes

[21, 23], and overlay tag maps that place keywords freely on top of maps [16, 25, 28]. This approach disambiguates hues and reduces legend dependence.

- **Text Glyphs**: The text glyphs themselves carry the class color, and the map background remains neutral or grayscale. Intrinsic tag maps like Taggram [9, 21] and Chroma-Text Field Maps encode term frequency or category importance directly through colored typography, optimizing space and perceptual clarity [33, 34]. Direct text encoding is particularly advantageous when icons or color palettes fail—such as when categories exceed ten hues or when user familiarity with symbols is low—resulting in faster, more accurate interpretation without extensive legend consultation.

In this work, we build on these map-specific text strategies and systematically explore four representative hybrid text–color encoding designs (**Legend-Aside Map**, **Label-Fill Map**, **Label-Fit Map**, and **Colored Label-Fill Map**) to optimize category discriminability and user performance in multiclass map visualization.

3 Visualization

3.1 Categorical Map Visualization

Drawing on available work, we have identified four hybrid visualization strategies (shown in Figure 1) to compare that integrate color and text to depict categorical maps:

- **Legend-Aside Map.** In this classic choropleth approach, each region is filled with a distinct hue, and all class names are listed in a consolidated legend alongside the map. This separation of color (on-map) and text (in the legend) maximizes spatial clarity and keeps the map itself uncluttered.
- **Label-Fill Map.** In this method, text labels are tiled repeatedly over their corresponding regions. The region's fill color both conveys the class and serves as the backdrop for the text, creating an integrated, texture-like effect that tightly couples label and color.
- **Label-Fit Map.** In this strategy, the class name is placed as a single label within each corresponding region.
- **Colored Label-Fill Map.** In this approach, text labels are tiled to cover the region, but the class hue is applied directly to the text glyphs rather than to a background. In other words, the letters become the color carriers while the region behind them remains neutral, yielding a striking, letter-centered encoding of class.

A **Colored Label-Fit** strategy was intentionally excluded from our study. With this method, a single colored label is placed within a neutral region, making it difficult for users to perceive the full spatial distribution of each category without adding explicit boundary lines. However, adding such boundaries would introduce new visual variables (e.g., line color and weight). To avoid the confounding influence of these variables, we excluded this strategy from our experiment.

3.2 Map Generation

All visualizations were implemented in Python, leveraging Matplotlib [15] for map rendering, Numpy [12] for geometric analysis, Wordcloud library [19] for label tiling, and Pillow [8] for image compositing. The source code is available at our GitHub repository . We chose eight publicly available choropleth datasets featuring 8–13 discrete categories—aligned with the approximately 10-hue limit for reliable color discrimination [32]. Source maps with secondary encodings (e.g., rivers, roads) were cleaned in Photoshop to remove extraneous details. And we keep their original color coding. Legends in Legend-Aside Map were standardized in placement (bottom-right), size, and font styling to ensure uniform presentation.

- **Region Orientation and Text Placement:** Yang et al. [20] demonstrated that adapting label direction to region shape allows larger fonts and improves legibility. Therefore, PCA was used to align text along each polygon's dominant axis, maximizing font size in the rotated bounding box.
- **Text Color and Contrast:** To ensure text legibility across all visualizations, we enforced a minimum 3:1 luminance contrast ratio, a standard guideline for readability [32]. Our color control strategy was adapted for the different mapping techniques: For the Label-Fill and Label-Fit maps, where text is overlaid on colored category regions, we rendered all text in black. This choice ensured that the 3:1 contrast ratio was consistently met or exceeded against every category's fill color. For the Colored Label-Fill Map, the text itself carries the categorical identity. Therefore, our approach was to first sample the original hue from each category and apply it to the text glyphs. We then systematically adjusted the luminance of each color until it achieved the required 3:1 contrast ratio against the neutral white background, thereby preserving the intended color association while guaranteeing readability.

The implementation details are as follows:

- **Legend-Aside Map:** Render each region in its category color; add a consolidated legend mapping colors to labels; omit on-map labels.
- **Label-Fit Map:** Place one black text label at each region's centroid, rotate per its PCA axis, and scale to the maximum fitting font size inside the polygon.
- **Label-Fill Map:** Use the Wordcloud library to fill each region's binary mask with repeated black text of its category label, constrained by a maximum font size to ensure a prominent label.
- **Colored Label-Fill Map:** Extend the Label-Fill pipeline by using ImageColorGenerator to sample region fill colors and modify for text glyphs, producing colored text on a white backdrop.

This consistent pipeline—common data sources, preprocessing, orientation, and color treatments—ensures that annotation strategy is the only variable across our comparative evaluation.

4 Experiment

To evaluate the performance and user experience of four visualization strategies, we ran a crowdsourced experiment. Participants completed representative map-reading tasks and provided subjective feedback on each design. We recruited English-speaking users via Prolific [1] and Chinese-speaking users via Credamo [2]–chosen for their complementary reach, and rapid turnaround–to ensure linguistic and cultural diversity. All procedures, including task assignment, data collection, and debriefing, conformed to our university's ethics guidelines and received formal approval from the institutional review board.

[1] https://www.prolific.com/
[2] https://www.credamo.com/

4.1 Participants

A total of 80 participants were recruited for this study (41 female, 39 male; age: M = 31.71, SD = 11.47), with an equal distribution of 40 participants from each platform. Participants were compensated at the minimum fee required platform upon successful completion of the survey. All participants were required to be at least 18 years of age, fluent in English for Prolific participants, and fluent in Chinese for Credamo, free from color vision deficiencies, and prohibited from undertaking the survey more than once. To standardize the viewing experience and ensure consistent map sizes, a strict device requirement was enforced: participants could only use devices with a minimum screen size of 13.3 inches (e.g., laptops or desktops), explicitly disallowing smartphones or tablets. Participants reported frequent use of map-based tools (M = 4.03) and moderate knowledge of data visualization (M = 3.23). Familiarity with Chinese geography (M = 3.08) was similar with U.S. geography (M = 2.93). Eight dataset familiarity scores ranged from 2.48 to 2.86.

4.2 Dataset

We selected eight representative maps from publicly available sources, including academic journals, government agencies, professional databases, and reputable media outlets. These maps encompass diverse thematic areas relevant to regional geography and spatial analysis. The sources and corresponding topics are listed below:

M1) Wikipedia – Distribution of Dialects in China
M2) Visual Capitalist – Most Valuable Agricultural Commodities in the U.S.
M3) Frontiers in Sustainable Food Systems – Ecological Divisions in the U.S.
M4) United States Geological Survey – Farm Resource Regions in the U.S.
M5) International Journal of Environmental Research and Public Health – Major Agricultural Areas in China
M6) Vegetation Classification and Survey – Vegetation Production Units in the U.S.
M7) SAS Blogs – Regional Cultures in the U.S.
M8) Agricultural and Applied Economics Association – Economic Regions in the U.S.

Selection of these maps was guided by two primary criteria: **Number of categories.** Colin Ware [32] suggested no more than ten distinct colors for symbol coding when accurate identification is required. Each selected map contains eight to thirteen thematic categories, deliberately chosen to approach the limits of human color-discrimination and create challenging scenarios for users.
Spatial distribution and difficulty balancing. To control task difficulty and minimize potential biases caused by variations in map structure, we carefully assessed the spatial distribution of categories across regions. Based on the total number of categories, the number of spatial units, and the overall distribution patterns, the selected maps were categorized into two difficulty levels: Simple and Complex. Each participant was assigned one simple and one complex map per map visualization (see arrangement in Table 1).

4.3 Questionnaire

The questionnaire consisted of seven sequential components: (1) introduction, (2) demographic information collection, (3) training trials, (4) main task trials, (5) post-task metric ratings, (6) overall preference selection, and (7) open-ended suggestions for improvement. On average, participants completed the entire questionnaire in approximately 50 minutes. A single attention check question was included to ensure engagement.

Table 1: Map assignments for each visualization type per participant.

PG	N	Legend-Aside	Label-Fill	Label-Fit	Colored Label-Fill
1	20	M5, M2	M6, M3	M7, M4	M8, M1
2	20	M8, M1	M5, M2	M6, M3	M7, M4
3	20	M7, M4	M8, M1	M5, M2	M6, M3
4	20	M6, M3	M7, M4	M8, M1	M5, M2

Each participant group (PG) consisting of N participants completed trials in a within-subject design. Each participant experienced all four visualization types, with each type paired with two maps (e.g., *Legend-Aside* with M5 and M2). The presentation order of the Visualization–Map combinations was randomized across participants.

The study employed a within-subjects design in which each participant completed tasks using four different visualization techniques, with each technique applied to two distinct map datasets. To control for potential learning and order effects, we adopted a Latin Square counterbalancing scheme (see Table 1). This design ensured that (i) every participant was exposed to all four visualization types exactly twice, (ii) the order of presentation was evenly distributed across participants, and (iii) the pairing of visualizations and map datasets remained consistent within each participant group. In total, this procedure produced 32 map stimuli–eight source maps, each rendered in four visualization variants.

Introduction. Participants were provided with a concise explanation of the study's purpose and structure. They were informed of their rights, including voluntary participation, the option to withdraw at any point, and guaranteed anonymity and data protection. Informed consent was obtained prior to continuing.

Demographic information collection. Participants were asked to provide basic demographic information (e.g., age, gender), their familiarity with geographic and data visualizations, and their prior knowledge of the map themes presented in the study.

Training. Then, participants proceeded to the training session, consisting of two map datasets, each accompanied by the full visualization set of four tasks, resulting in eight training trials.

Main task. The core of the study consisted of four map-based tasks designed to evaluate different aspects of categorical map interpretation. The Count task served as the primary measure, introduced specifically to assess the impact of color-code differentiation. We remove the legend on this task on Legend-Aside map. The remaining three tasks–Identify, Compare, and Rank–were adapted from established map-visualization primitives [24] and provided complementary insights into retrieval efficiency, relational analysis, and hierarchical perception.

Each participant experienced all four visualization strategies, with two maps per strategy (one "easy" and one "complex"). For each map, they completed four tasks, resulting in $4 \times 2 \times 4 = 32$ trials per participant.

Post-task metric ratings. After completing the four tasks for each map, participants were required to evaluate the visualization techniques on a 5-point Likert scale (from 1 = strongly disagree to 5 = strongly agree) on four key metrics: Readability, Information Extraction, Aesthetic Appeal, and Complexity. These metrics were carefully selected based on existing literature on infographics [5], and were specifically tailored to the context of geo-information visualization.

- **Readability:** The clarity and legibility of map elements, assessing how easily participants can decode labels and symbols.
- **Information Extraction:** The speed and accuracy with which participants locate and interpret key data points or patterns.
- **Aesthetic Appeal:** The overall visual attractiveness and stylistic coherence of the map design.
- **Complexity:** The extent of perceived visual clutter or unnecessary detail that hinders understanding.

Participants were also asked to rank their overall preference for all presented visualizations. In the end, participants were invited to provide free-text feedback on the four visualizations.

4.4 Measurement

We collected both quantitative and qualitative data to assess each visualizations. For objective performance on the **Count**, **Identify**, **Compare**, and **Rank** tasks, we collected two primary quantitative measures:

- **Accuracy:** We recorded participants' answers for each trial and compared them against the correct answers to derive an accuracy rate.
- **Completion Time:** We measured the time taken by participants from the moment a task was presented until they provided their answer, effectively capturing the task completion duration.

For subjective evaluation, we collected the following:

- **Likert Scale Ratings:** Participants provided ratings for four key metrics–**Readability**, **Information Extraction**, **Aesthetic Appeal**, and **Complexity**–on a 5-point Likert scale (1 = Strongly Disagree to 5 = Strongly Agree).
- **Overall Preference Ranking:** Participants were asked to rank the four visualization strategies (**Legend-Aside Map**, **Label-Fill Map**, **Label-Fit Map**, and **Colored Label-Fill Map**) based on their overall preference for each map visualization.
- **Open Qualitative Feedback:** Participants were encouraged to provide open-ended comments and suggestions on each map visualization, offering valuable qualitative insights into their user experience.

4.5 Hypotheses

Grounded in perceptual and cognitive theory [32] and the nature of these visualization strategies, we hypothesize:

H1 (Count Task): For counting regions by class, in-situ label placements (Label-Fill, Label-Fit, Colored Label-Fill) will yield higher accuracy and faster times than the Legend-Aside Map.

H2 (Identify, Compare and Rank Task): When identifying a single or multiple region's category, the Label-Fit Map will outperform both Label-Fill and Colored Label-Fill designs.

H3 (Increasing Task Difficulty): As task difficulty grows (e.g., from Identify to Compare to Rank), the performance gap of the Legend-Aside Map versus in-situ label placements designs will widen.

H4 (Subjective Preference): Despite possible advantages of hybrid designs, participants will rate the Label-Fit Map highest in subjective preference, balancing readability, aesthetic appeal, and familiarity better than both the legend-based and full-text-fill approaches.

5 Result

All statistical analyses were performed using IBM SPSS (version 27). The Shapiro–Wilk test was used to assess the normality of the data, revealing non-normal distributions across all visualizations. Consequently, the Friedman test was applied to examine the effects of the independent variable. Effect sizes (ES) were reported using Kendall's W to assess the overall consistency across visualizations. For within-subject comparisons, non-parametric Wilcoxon Signed-Rank tests were utilized. To account for multiple comparisons and control the family-wise error rate, Bonferroni corrections were implemented. Descriptive statistics, including means and 95% confidence intervals (CIs), were calculated using 1,000 bootstrap resamples. Statistical significance was indicated as follows: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

5.1 Results for Objective Evaluation:

The objective evaluation results are illustrated in Figure 2.

Count. **Accuracy**: A Friedman test showed significant differences across 4 visualizations ($\chi^2 = 17.547$, $^{***}p < .001$, $ES = .073$). *Colored Label-Fill* (M = 38.8% [30.8, 46.7]) performed significantly worse than *Label Fill* ($^{*}p = .015$) and *Legend-Aside* ($^{*}p = .039$). The remaining three visualization types are *Legend-Aside* (M = 56.3% [48.7, 63.8]), *Label-Fill* (M = 56.9% [49.4, 64.3]), and *Label-Fit* (M = 56.3% [48.1, 64.4]), with no significant differences observed in pairwise comparisons. **Completion Time**: A Friedman test showed significant differences across visualizations ($\chi^2 = 29.355$, $^{***}p < .001$, $ES = .122$). Participants were significantly faster with *Legend-Aside* (M = 39,362.14 ms [31,630.73, 47,093.56]) than with *Label-Fill* (M = 51,101.04 ms [43,190.66, 59,011.41]; $^{***}p = .000$) and *Colored Label-Fill* (M = 56,943.46 ms [48,257.16, 65,629.77]; $^{***}p = .000$). The comparison between *Legend-Aside* and *Label-Fit* (M = 41,844.05 ms [36,691.79, 46,996.31]) did not show a significant difference.

Identify. **Accuracy**: A Friedman test revealed significant differences across 4 visualizations ($\chi^2 = 13.306$, $^{**}p = .004$, $ES = .055$). Significant differences were not found in pairwise comparisons. *Label-Fit* yielded the highest mean accuracy (M = 79.4% [72.5, 85.6]). *Legend-Aside* followed with a mean accuracy of 67.5% [59.4, 75.6]. Both *Label-Fill* (M = 64.4% [56.4, 72.3]) and *Colored Label-Fill* (M = 64.4% [56.6, 72.1]) showed lower performance. **Completion Time**: A Friedman test revealed significant differences across 4 visualizations ($\chi^2 = 16.110$, $^{**}p = .001$, $ES = .067$). *Legend-Aside* had the shortest average completion time (M = 27,538.41 ms [23,852.43, 31,224.39]), which was significantly faster than both *Label-Fill* ($^{*}p = .020$) and *Colored Label-Fill* ($^{**}p = .002$). *Label-Fit* (M = 29,701.38 ms [26,021.20, 33,381.56]) was slower than *Legend-Aside*, but faster than *Label-Fill* and *Colored Label-Fill*, with no significant differences from any pair of them. *Colored Label-Fill* showed

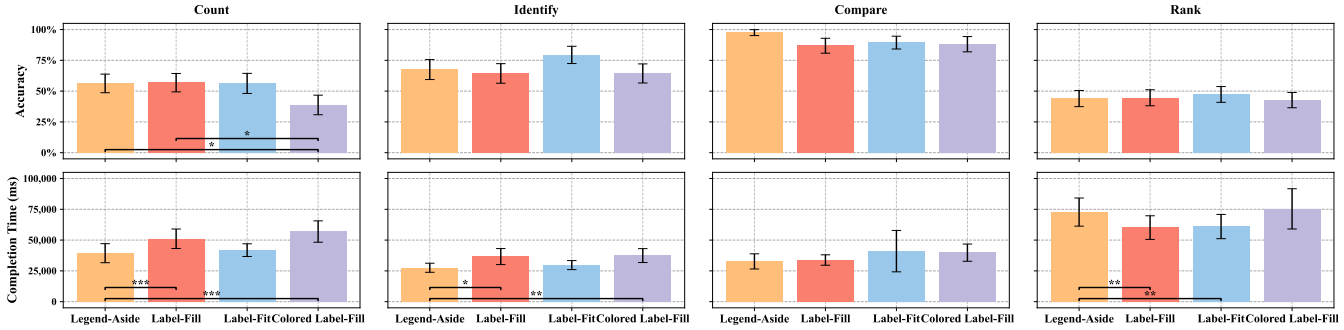


Figure 2: Mean accuracy (%) and completion time (ms) for four visualization types: Legend-Aside, Label-Fill, Label-Fit, and Colored Label-Fill, under task types: Count, Identify, Compare, and Rank. Error bars indicate 95% bootstrap confidence intervals.

the longest average completion time (M = 37,426.03 ms [31,781.19, 43,070.88]).

Compare. **Accuracy**: A Friedman test revealed significant differences across 4 visualizations (χ^2 = 12.118, $^{**}p$ = .007, *ES* = .050). Significant differences were not found in pairwise comparisons. *Legend-Aside* showed the highest accuracy (M = 97.5% [95.1, 99.9]), followed by *Label-Fit* (M = 89.4% [84.2, 94.6]), *Label Fill* (M = 86.9% [80.8, 92.9]), and *Colored Label-Fill* (M = 88.1% [81.9, 94.3]). **Completion Time**: A Friedman test did not show significant differences in Compare time. Completion time intervals ranged from 24,200 ms to 57,800 ms across visualizations. *Legend-Aside* showed the shorest interval (M = 32,696.59 ms [26,532.08, 38,859.11]), while *Label-Fit* had the longest (M = 41,038.74 ms [24,228.84, 57848.65]).

Rank. **Accuracy**: A Friedman test did not show significant differences across 4 visualizations. Ranking accuracy intervals ranged from 36.5% to 53.7% across the four conditions. *Label-Fit* had the highest interval (M = 47.3% [40.9, 53.7]), while *Colored Label-Fill* had the lowest (M = 42.7% [36.5, 48.9]). **Completion Time**: A Friedman test revealed significant differences across 4 visualizations (χ^2 = 13.095, $^{**}p$ = .004, *ES* = .055). The shortest average completion time was observed for *Label-Fill* (M = 60,147.14 ms [50,546.07, 69,748.20]), which was significantly lower than *Legend-Aside* ($^{*}p$ = .035). *Label-Fit* also showed a relatively low completion time (M = 60,999.90 ms [51,150.76, 70,849.04]), significantly lower than *Legend-Aside* ($^{*}p$ = .011). The longest completion time was found for *Colored Label-Fill* (M = 75,357.27 ms [59,011.91, 91,702.63]).

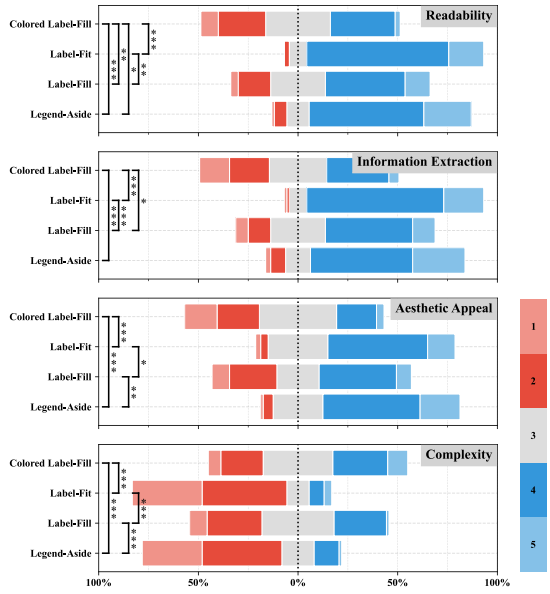


Figure 3: Subjective scores across tasks for Readability, Aesthetic Appeal, Information Extraction, and Complexity, measured on a 5-point Likert scale. Higher scores indicate better performance for all attributes except Complexity, where lower scores reflect more desirable outcomes.

5.2 Results for Subjective Evaluation:

Figure 3 presents the results for four key subjective metrics, and overall preference rankings are shown in Figure 4.

Readability. A Friedman test revealed significant differences across visualizations (χ^2 = 73.175, $^{***}p$ < .001, *ES* = .305). *Label-Fit* received the highest readability rating (M = 4.19 [4.07, 4.32]), significantly higher than both *Label-Fill* (M = 3.66 [3.44, 3.87], $^{**}p$ = .002) and *Colored Label-Fill* (M = 3.16 [2.93, 3.38], $^{***}p$ = .000). *Legend-Aside* (M = 4.09 [3.91, 4.27]) was also rated significantly higher than *Label-Fill* ($^{*}p$ = .024) and *Colored Label-Fill* ($^{***}p$ = .000), while not significantly different from *Label-Fit*. *Label-Fill* was rated significantly higher than *Colored Label-Fill* ($^{**}p$ = .006), which received the lowest rating overall.

Information Extraction. A Friedman test demonstrated a significant differences (χ^2 = 61.296, $^{***}p$ < .001, *ES* = .255). *Label-Fit* received the highest rating (M = 4.23, [4.09, 4.36]), significantly higher than both *Label-Fill* (M = 3.66 [3.45, 3.88], $^{**}p$ = .001) and *Colored Label-Fill* (M = 3.12 [2.87,3.37], $^{***}p$ = .000). *Legend-Aside* (M = 4.08 [3.88,4.27]) was also rated significantly higher than *Colored Label-Fill* ($^{***}p$ = .000), though not significantly different from *Label-Fit* and *Label-Fill*. *Label-Fill* was rated significantly higher than *Colored Label-Fill* ($^{*}p$ = .026), which received the lowest rating.

Aesthetic Appeal. A Friedman test showed significant differences (χ^2 = 52.385, $^{***}p$ < .001, *ES* = .218). *Legend-Aside* (M = 4.00 [3.82, 4.20]) was rated most aesthetically appealing, significantly higher than both *Label-Fill* (M = 3.36 [3.12, 3.60], $^{**}p$ = .005) and

Colored Label-Fill (M = 2.91 [2.67, 3.15], $^{***}p = .000$). *Label-Fit* (M = 3.92 [3.74, 4.10]) was also rated significantly higher than *Colored Label-Fill* ($^{***}p = .000$) and *Label-Fill* ($^{*}p = .039$).

Complexity. A Friedman test showed a significant difference ($\chi^2 = 77.558$, $^{***}p < .001$, $ES = .323$). *Colored Label-Fill* (M = 3.34 [3.12, 3.57]) and *Label-Fill* (M = 3.08 [2.87, 3.28]) were rated as significantly more complex than both *Legend-Aside* (M = 2.33 [2.10, 2.56]), $^{***}p = .000$ and $^{***}p = .000$) and *Label-Fit* (M = 2.15 [1.92, 2.38], $^{***}p = .000$ and $^{***}p = .000$). No other pairwise comparisons reached statistical significance.

Overall Rank. A Friedman test showed significant differences across 4 visualizations ($\chi^2 = 762.388$, $^{***}p < .001$, $ES = .397$). *Colored Label-Fill Map* consistently received the lowest ratings and was rated significantly lower than *Legend-Aside Map* (M = 1.93 [1.86, 2.01]), *Label-Fill Map* (M = 2.79 [2.73, 2.85]), and *Label-Fit Map* (M=1.76 [1.69, 1.82]) (***all $p = .000$). *Label-Fit Map* was rated significantly higher than *Label-Fill Map* ($^{***}p = .000$) and *Colored Label-Fill Map* ($^{***}p = .000$), and showed a marginal advantage over *Legend-Aside Map* ($p = .087$).

Open Feedback. *Legend-Aside* received relatively positive feedback for its clear design, and some users suggested placing region names directly on the corresponding areas. *Label-Fill* and *Colored Label-Fit* were commonly criticized for repetitive text and cluttered layout. Users recommended reducing redundant labels, standardizing text orientation, and incorporating a legend. The main issue with *Label-Fit* was that text in smaller regions appeared too small, affecting readability.

6 Discussion

6.1 Hypothesis Validation

H1 was rejected: *Legend-Aside* achieved similarly high accuracy compared to *Label-Fit* and *Label Fill*, and demonstrated the significant shortest completion time among the four designs in Count. This may be because the number of color categories (8-13) was not so challenging, so the cost of referring to the legend was not high enough to hinder performance.

H2 was basically supported: Except for the Compare, where no significant differences in accuracy or completion time were found, *Label-Fit* showed the highest accuracy and the shortest completion time in Identify within in-situ map visualizations since its single, maximally-sized label maximizes legibility without visual clutter, facilitating rapid recognition. In Rank, although accuracy differences were not significant, *Label-Fit* and *Label-Fill* had comparable completion times, both shorter than *Colored Label-Fill*.

H3 was basically supported: In the Identify task, no significant differences in accuracy were found. However, *Legend-Aside* showed significantly faster completion times than both *Label-Fill* and *Colored Label-Fill*. For the Compare task, neither accuracy nor completion time showed significant differences across visualizations. In the Rank task, accuracy differences remained non-significant, but *Legend-Aside* took significantly more time than *Label-Fill* and *Label-Fit*. These results suggest that *Legend-Aside* performed well in simpler tasks. In complex tasks, increased lookup demands may have slowed performance, while some in-situ designs performed better. Combining text and color can be effective, but the performance did not consistently well for all in-situ designs. Among these tested in-situ designs, *Label-Fit* showed the most consistent and balanced performance.

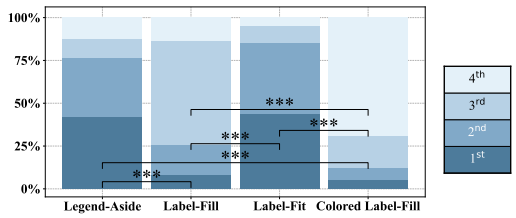


Figure 4: Users' preference on four visualizations.

H4 was basically supported: *Label-Fit* was rated highest for readability, information retrieval, and simplicity. *Legend-Aside* was only preferred for aesthetics. In the overall ranking, *Label Fit* also achieved the highest position.

6.2 Key Findings and Design Recommendations

- **Label-Fit Map as the "All-Rounder."** By placing a single, maximally-sized label within each region, the Label-Fit Map balances legibility and minimal clutter. It led in both task performance and subjective ratings, making it the optimal choice overall.
- **Legend-Aside for Simple Tasks.** Although slower in Rank due to legend lookups, Legend-Aside perform best in Count, Identify and Compare, and received good subjective scores. For category counts, identidy and comparison-focused tasks, it remains a solid design.
- **Caution with Colored Label-Fill.** The Colored Label-Fill Map performed worst in speed and accuracy and was reported as visually confusing. We should avoid pure Colored Label-Fill Map.
- **Label-Fill's Mixed Effects.** Repeating labels across regions (Label-Fill) improved Count and Rank performance but generated visual noise reported by participants. Using adaptive decluttering (e.g., density thresholds) to control repetition and maintain clarity might be a good choice.

6.3 Limitations

We acknowledge several limitations. First, our evaluation was conducted on static maps with 8–13 categories—near the perceptual threshold for hue discrimination—so results may differ when interactive elements (e.g., zooming, filtering) or larger category sets are introduced. Second, we kept original color palettes for all maps; alternative palettes or dynamic color-assignment methods could influence the performance of color-reliant designs. Finally, all trials were run on laptops/desktops under controlled lighting, and findings may not generalize to mobile or small-screen devices.

7 Conclusion

In an 80-participant crowdsourced study on multiclass choropleth maps with 8–13 categories—around the human perceptual limit for reliable hue discrimination—we compared four hybrid text–color annotation strategies: **Legend-Aside**, **Label-Fill**, **Label-Fit**, and **Colored Label-Fill**. Our results indicate that the **Label-Fit** design, which places a single, maximally sized in-situ label in each region, provides the best overall balance of accuracy, speed, and user

satisfaction by minimizing visual clutter while maximizing legibility. The traditional **Legend-Aside** approach matched **Label-Fit** in count, identity and compare tasks but incurred slower performance in rapid rank due to the overhead of legend lookups. **Label-Fill** strengthened category recognition in dense, multi-class scenarios but introduced visual noise that reduced readability, and **Colored Label-Fill** consistently underperformed across both objective and subjective measures. Future work will explore adaptive and interactive labeling techniques, and scalability to larger category sets and diverse device contexts.

Acknowledgments

Yu Liu is partially funded by XJTLU RDF, grant № RDF-22-01-092.

References

[1] Luc Anselin, Ibnu Syabri, and Youngihn Kho. 2006. GeoDa: An introduction to spatial data analysis. *Geographical Analysis* 38, 1 (2006), 5 – 22. doi:10.1111/j.0016-7363.2005.00671.x

[2] Harald Bosch, Dennis Thom, Florian Heimerl, Edwin Püttmann, Steffen Koch, Robert Krüger, Michael Wörner, and Thomas Ertl. 2013. ScatterBlogs2: Real-Time Monitoring of Microblog Messages through User-Guided Filtering. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2022–2031. doi:10.1109/TVCG.2013.186

[3] Richard Brath and Ebad Banissi and. 2017. Multivariate label-based thematic maps. *International Journal of Cartography* 3, 1 (2017), 45–60. doi:10.1080/23729333.2017.1301346

[4] Kevin Buchin, Daan Creemers, Andrea Lazzarotto, Bettina Speckmann, and Jules Wulms. 2016. Geo word clouds. In *2016 IEEE Pacific Visualization Symposium (PacificVis)*. 144–151. doi:10.1109/PACIFICVIS.2016.7465262

[5] Alyxander Burns, Cindy Xiong, Steven Franconeri, Alberto Cairo, and Narges Mahyar. 2022. Designing With Pictographs: Envision Topics Without Sacrificing Understanding. *IEEE Transactions on Visualization and Computer Graphics* 28, 12 (2022), 4515–4530. doi:10.1109/TVCG.2021.3092680

[6] Jiashu Chen, Weikai Yang, Zelin Jia, Lanxi Xiao, and Shixia Liu. 2025. Dynamic Color Assignment for Hierarchical Data. *IEEE Transactions on Visualization and Computer Graphics* 31, 1 (2025), 338–348. doi:10.1109/TVCG.2024.3456386

[7] Wei Chen, Zhaosong Huang, Feiran Wu, Minfeng Zhu, Huihua Guan, and Ross Maciejewski. 2018. VAUD: A Visual Analysis Approach for Exploring Spatio-Temporal Urban Data. *IEEE Transactions on Visualization and Computer Graphics* 24, 9 (2018), 2636–2648. doi:10.1109/TVCG.2017.2758362

[8] Alex Clark. 2015. Pillow (PIL Fork) Documentation. https://buildmedia.readthedocs.org/media/pdf/pillow/latest/pillow.pdf

[9] Davide De Chiara, Vincenzo Del Fatto, Monica Sebillo, Genoveffa Tortora, and Giuliana Vitiello. 2012. Tag@Map: a web-based application for visually analyzing geographic information through georeferenced tag clouds. In *Proceedings of the 11th International Conference on Web and Wireless Geographical Information Systems* (Naples, Italy) *(W2GIS'12)*. Springer-Verlag, Berlin, Heidelberg, 72–81. doi:10.1007/978-3-642-29247-7_7

[10] Cristian Felix, Steven Franconeri, and Enrico Bertini. 2018. Taking Word Clouds Apart: An Empirical Investigation of the Design Space for Keyword Summaries. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 657–666. doi:10.1109/TVCG.2017.2746018

[11] Connor C. Gramazio, David H. Laidlaw, and Karen B. Schloss. 2017. Colorgorical: Creating discriminable and preferable color palettes for information visualization. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2017), 521–530. doi:10.1109/TVCG.2016.2598918

[12] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature* 585, 7825 (Sept. 2020), 357–362. doi:10.1038/s41586-020-2649-2

[13] Manli He, Xi Tang, and Yuming Huang. 2011. To visualize spatial data using thematic maps combined with infographics. In *2011 19th International Conference on Geoinformatics*. IEEE, Shanghai, China, 1–5. doi:10.1109/GeoInformatics.2011.5980880

[14] Tingying He, Yuanyang Zhong, Petra Isenberg, and Tobias Isenberg. 2024. Design Characterization for Black-and-White Textures in Visualization. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (Jan. 2024), 1019–1029. doi:10.1109/TVCG.2023.3326941

[15] J. D. Hunter. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* 9, 3 (2007), 90–95. doi:10.1109/MCSE.2007.55

[16] Alexandar Jaffe, Mor Naaman, Tamir Tassa, and Marc Davis. 2006. Generating summaries and visualization for large collections of geo-referenced photographs. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval* (Santa Barbara, California, USA) *(MIR '06)*. Association for Computing Machinery, New York, NY, USA, 89–98. doi:10.1145/1178677.1178692

[17] Bongshin Lee, Nathalie Henry Riche, Amy K. Karlson, and Sheelash Carpendale. 2010. SparkClouds: Visualizing Trends in Tag Clouds. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (2010), 1182–1189. doi:10.1109/TVCG.2010.194

[18] Chenlu Li, Xiaoju Dong, and Xiaoru Yuan. 2018. Metro-Wordle: An Interactive Visualization for Urban Text Distributions Based on Wordle. *Visual Informatics* 2, 1 (2018), 50–59. doi:10.1016/j.visinf.2018.04.006 Proceedings of PacificVAST 2018.

[19] Andreas C Mueller. 2023. *Wordcloud*. https://github.com/amueller/wordcloud

[20] Alan M. MacEachren Nai Yang and Liping Yang. 2019. TIN-based Tag Map Layout. *The Cartographic Journal* 56, 2 (2019), 101–116. doi:10.1080/00087041.2018.1533294

[21] Dinh-Quyen Nguyen and Heidrun Schumann. 2010. Taggram: Exploring Geo-data on Maps through a Tag Cloud-Based Visualization. In *2010 14th International Conference Information Visualisation*. 322–328. doi:10.1109/IV.2010.52

[22] Dinh Quyen Nguyen, Christian Tominski, Heidrun Schumann, and Tuan Anh Ta. 2011. Visualizing Tags with Spatiotemporal References. In *2011 15th International Conference on Information Visualisation*. 32–39. doi:10.1109/IV.2011.43

[23] Martin Reckziegel, Muhammad Faisal Cheema, Gerik Scheuermann, and Stefan Jänicke. 2018. Predominance Tag Maps. *IEEE Transactions on Visualization and Computer Graphics* 24, 6 (2018), 1893–1904. doi:10.1109/TVCG.2018.2816208

[24] Robert E. Roth. 2013. An Empirically-Derived Taxonomy of Interaction Primitives for Interactive Cartography and Geovisualization. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2356–2365. doi:10.1109/TVCG.2013.130

[25] Aidan Slingsby, Jason Dykes, Jo Wood, and Keith Clarke. 2007. Interactive Tag Maps and Tag Clouds for the Multiscale Exploration of Large Spatio-temporal Datasets. In *Proceedings of the 11th International Conference Information Visualization (IV '07)*. IEEE Computer Society, USA, 497–504. doi:10.1109/IV.2007.71

[26] Terry A Slocum, Robert B McMaster, Fritz C Kessler, and Hugh H Howard. 2022. *Thematic cartography and geovisualization*. CRC Press. doi:10.1201/9781003150527

[27] Rainer Splechtna, Thomas Hulka, Disha Sardana, Nikitha Donekal Chandrashekar, Denis Gračanin, and Kresimir Matković. 2023. Interactive Exploration of Complex Heterogeneous Data: A Use Case on Understanding City Economics. In *VISIGRAPP 2023 - 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, n.n. (Ed.). doi:10.5220/0011787500003417

[28] Dennis Thom, Harald Bosch, Steffen Koch, Michael Wörner, and Thomas Ertl. 2012. Spatiotemporal anomaly detection through visual analysis of geolocated Twitter messages. In *2012 IEEE Pacific Visualization Symposium*. 41–48. doi:10.1109/PacificVis.2012.6183572

[29] Fernanda B. Viegas, Martin Wattenberg, and Jonathan Feinberg. 2009. Participatory Visualization with Wordle. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (2009), 1137–1144. doi:10.1109/TVCG.2009.171

[30] Guangjie Wang, Wenfu Peng, Lindan Zhang, Jiayao Xiang, Jingwen Shi, and Lu Wang. 2023. Quantifying urban expansion and its driving forces in Chengdu, western China. *The Egyptian Journal of Remote Sensing and Space Sciences* 26, 4 (2023), 1057–1070. doi:10.1016/j.ejrs.2023.11.010

[31] Qiming Wang, Kun Yang, Lixiao Li, and Yanhui Zhu. 2022. Assessing the Terrain Gradient Effect of Landscape Ecological Risk in the Dianchi Lake Basin of China Using Geo-Information Tupu Method. *International Journal of Environmental Research and Public Health* 19, 15 (2022). doi:10.3390/ijerph19159634

[32] Colin Ware. 2004. *Information Visualization: Perception for Design* (2nd ed.). Morgan Kaufmann Publishers Inc., San Francisco. doi:10.1016/B978-155860819-1/50001-7

[33] Zhiwei Wei and Nai Yang and. 2024. Using a negative spatial auto-correlation index to evaluate and improve intrinsic tag map's multi-scale visualization capabilities. *Cartography and Geographic Information Science* 0, 0 (2024), 1–18. doi:10.1080/15230406.2024.2391479

[34] Nai Yang, Alan M. MacEachren, and Emily Domanico and. 2020. Utility and usability of intrinsic tag maps. *Cartography and Geographic Information Science* 47, 4 (2020), 291–304. doi:10.1080/15230406.2020.1732835

[35] Xinyuan Zhang, Yifan Xu, Kaiwen Li, Lingyun Yu, and Yu Liu. 2025. ChinaVis24.MapCraft: dissecting and designing custom geo-infographics. *Journal of Visualization* (2025). doi:10.1007/s12650-025-01059-4

[36] Chenghu Zhou, Fenzhen Su, Tao Pei, An Zhang, Yunyan Du, Bin Luo, Zhidong Cao, Juanle Wang, Wen Yuan, Yunqiang Zhu, Ci Song, Jie Chen, Jun Xu, Fujia Li, Ting Ma, Lili Jiang, Fengqin Yan, Jiawei Yi, Yunfeng Hu, Yilan Liao, and Han Xiao. 2020. COVID-19: Challenges to GIS with Big Data. *Geography and Sustainability* 1, 1 (2020), 77 – 87. doi:10.1016/j.geosus.2020.03.005