Exploring How SE Academic Papers Use Figures

Marco Torchiano
Computer and Control Engineering
Politecnico di Torino
Torino 10129
Italy
marco.torchiano@polito.it

Lorenzo Laudadio
Computer and Control Engineering
Politecnico di Torino
Torino 10129
Italy
lorenzo.laudadio@polito.it

***Abstract*—The usage of figures to represent data or concepts in scientific articles is a common practice. We aim to understand what figures are used for in SE articles and in particular how quantitative data is represented. For this purpose we analyzed 865 articles published in leading software engineering scientific conferences and journals and classified 6342 figures and their contents. 47% of the figures are used to convey quantitative information and the rest depict more abstract non-quantitative information. The most common types of quantitative diagrams are bar plots, box plots, and line plots, accounting for 75% of the quantitative figures. We also found that each figure contains 1.6 errors, although 75% of them do not contain any critical error. Critical blatant errors are found in less than 5% of the figures.**

***Index Terms*—visualization, software engineering**

I. Introduction

In software engineering (SE) research, the effective communication of ideas, methodologies, and results is paramount. Papers in this field often employ a variety of visual aids, including figures, tables, and code fragments, to complement the textual content and enhance clarity. While there exist a large research community that explores innovative visualization techniques, the majority of visualizations in SE papers remain relatively mundane, relying on basic, well-established methods. This reliance on standard visual aids reflects both the practical constraints and the entrenched practices within the field.

Practitioners who regularly writes papers and supervises younger researchers, can often find themselves providing guidance on the effective use of visualizations. This necessity arises from observing a recurring issue: poorly designed diagrams that fail to effectively convey the intended information. Additionally, as reviewers, we frequently encounter diagrams that are not only visually unappealing but also ineffective in communicating complex ideas. These experiences underscore a broader concern within the community regarding the presentation quality of SE papers. A critical question emerges: is it worthwhile to invest substantial effort in designing more sophisticated and effective diagrams? This paper aims to address this question by examining the current state of visualization practices in SE research.

The goal of this paper is to advance the discourse on visualization in software engineering. The main contributions consist in:

- Overview of the State of the Practice: We provide a comprehensive analysis of the current visualization practices within prominent SE venues, identifying prevailing trends and common practices.
- Purpose of Figures: An examination of the various roles that figures play in SE papers, shedding light on how they contribute to the overall narrative and understanding of the research.
- Catalogue of Common Diagrams: A detailed catalog of the most frequently used diagrams, offering insights into their typical applications and variations.

In particular we will focus on quantitative visualization, i.e. charts and diagrams that encode quantitative measures. This is opposed to non-quantitative visualization that represent conceptual aspects in a more or less formal way, e.g., software representations such as UML diagrams.

Overall by surveying the current practices and identifying most common errors, this paper seeks to elevate the standard of visual communication in software engineering research, ultimately contributing to the clarity and impact of published work.

II. Background

Data visualization is a critical component in the communication of scientific research, enabling the distillation of complex data into comprehensible visual formats. By transforming raw data into graphical representations, researchers can more effectively highlight patterns, trends, and anomalies, thereby facilitating better understanding and interpretation. In the context of academic publications, well-designed visualizations are not merely supplementary but essential elements that enhance the clarity and impact of the presented research.

A. Visualization of Quantitative Information

The seminal book *The Visual Display of Quantitative Information* [1] revolutionized the way quantitative data is presented. Tufte emphasized principles such as lie factor and data-ink ratio, and laid the groundwork for a set of best practices that prioritize proportionality, clarity, and utility in the visual representation of data.

Building on the principles established by Tufte, subsequent scholars have further refined the guidelines for effective data visualization. Notably, Tamara Munzner's *Visualization Analysis and Design* [2] offers a comprehensive framework for creating insightful and effective visualizations. Munzner's work

addresses the complexity of visual encoding and interaction, providing a systematic approach to design that is applicable across various disciplines. Her guidelines cover a broad spectrum of visualization types, from simple charts to complex, interactive graphics, emphasizing the importance of aligning visualization techniques with the specific needs and goals of the analysis.

In the domain of software engineering (SE), tailored guidelines for visualization have been proposed to address the unique challenges and requirements of the field. One notable contribution is the Empirical Standards for Software Engineering Research, as articulated by [3]. These standards provide a robust framework for conducting and presenting empirical research in SE, including specific recommendations for the use of visualizations. The guidelines emphasize aim to ensure that dagrams and charts effectively support the communication of research findings.

B. Graph Error Taxonomy

On the basis of the extensive literature and the personal experience in reviewing papers we defined a taxonomy of errors that include three levels of severity:

- Critical: the error can severely impact the correct understanding of the quantitative message;
- Major: the error can significantly impact the ease of understanding, demanding more effort than strictly required;
- Minor: the error can moderately impact the ease of understanding.

A summary of the errors we identified, divided by severity is reported in table Table I.

Table I: Summary of errors by severity level.

Severity	Errors
Critical	Cropped, Deformed, DoubleScale, MissingAxesRef, NonZeroBased, SimilarColors, TooManyCats
Major	3D, GridDistinctRanges, InterruptedScale, Legend, Mislabeled, Overplotting, RotatedXLabels, SilentLog, Unlabeled
Minor	ColorsUncoded, HeavyBackground, HeavyGrid, LegendBorder, LegendInside, OverlappedLabels, PatternFill, Raster, Shadow, TooMuchPrecision, WasteSpace, WrappedXLabs

The detailed descriptions of the are reported below:

1) Critical:

- **Cropped** The graph is cropped and some details are not visible since they lay outside. This impedes the vision of part of the diagram.
- **Deformed** The graph is somehow deformed with proportion not consistente across the area. Often this is – wrongly – used to emphasize some detail instead of using more suitable techniques, the reuslt is a deception of the observer [4].
- **DoubleScale** The graph uses a double (vertical) scale. Research in cognitive science [5] suggests that people may misinterpret trends if they do not notice the differing scales, potentially leading to incorrect conclusions about the correlation between the two datasets.
- **MissingAxesRef** The whole axes or the tick marks are missing. This make the quantitative information contained in the diagram fuzzy and imprecise failing the goal of conveying accurate information.
- **NonZeroBased** A bar diagram with truncated bars, axis not starting at 0. This is a serious problem concerning the visual integrity of the diagram, in particualar it affects proportionality [1], thus falsifying the quantitative message of the diagram.
- **SimilarColors** Graph uses very similar colors, hard to tell apart. Research indicates that low contrast between colors (especially in terms of luminosity) can reduce the speed and accuracy of information interpretation [7].
- **TooManyCats** The graph encodes too many categories (usually more than six) with attributes like color or shape. For instance, [8] recommends that the number of colors used to represent nominal data should be restricted to seven or less, while sudies by [6] found that people can typically distinguish 4 to 6 distinct shapes effectively.

2) Major:

- **3D** There are 3D effects. The additional cognitive effort required for 3D interpretation often results in slower reaction times and higher error rates [6].
- **GridDistinctRanges** When multiple graph are present (grid of chars, a.k.a. small multiples), the corresponding axes have different intervals. [9] shows that when the scales are consistent across all plots, viewers can quickly and accurately compare values; while inconsistent axes may lead to misinterpretation.
- **InterruptedScale** One of the axes is interrupted and restarted after a discontinuity to show extremely spread values. Axis breaks can affect the perceived effect size, leading viewers to believe that the differences between data points are more significant than they actually are [4].
- **Legend** There is a legend that could have been turned into direct labeling. Direct labeling (i.e., placing labels close to the data points) instead of using a separate legend reduces the need for eye movements between the legend and the chart elements, thus it leads to faster and more accurate interpretation [10].
- **Mislabeled** Labels are obscuring or get confused with data. A key recommendation when creating a diagram is 'above all show the data' [1].
- **Overplotting** There are too many overplotted points that prevent distinguishing individual points. It leads to misleading interpretations, particularly in scatter plots used to show relationships or correlations; since the points are not distinguishable, viewers may fail to detect underlying relationships or trends [5].
- **RotatedXLabels** The labels on the X axis are rotated (not horizontal). Rotated labels can interfere with the viewer's ability to quickly interpret the data, as they require additional time for visual decoding; such effect is particularly pronounced in bar charts with long category labels [5].

- **SilentLog** One of the axes uses a log scale but this is not clearly mentioned. While log scales help in visualizing both small and large values in the same plot without extreme compression or skewing of the data points [11], viewers often misinterpret log scales if the axis labels do not clearly indicate the logarithmic nature of the scale [5].
- **Unlabeled** The axes are not labeled. While the meaning of the axes can be inferred from the caption of the main text of the article, the lack of labels makes understanding the graph much less immediate.

3) Minor:

- **ColorsUncoded** There are many color not corresponding to an explicit coding (legend or other). If colors are used purely for decorative purposes without encoding information, they may inadvertently create a false visual hierarchy, drawing attention away from the key data elements [5].
- **HeavyBackground** The background is heavy. If the background is so intense to almost obscure the data or to draw attention away from the data, the visual message is lost [1].
- **HeavyGrid** The grid of the graph is heavy. If the visual presence of grid is so strong to almost obscure the data, the visual message is lost [1].
- **LegendBorder** The legend has a border (preventing free eye scan)
- **LegendInside** The legeng is inside the graph area
- **OverlappedLabels** Labels are overlapping with each other. Overlapping labels are difficult to read thus they require additional effort or make impossible understanding.
- **PatternFill** Area fill is using a parttern (e.g., lines or dots) instead of diferent hues or gray levels. Excessive use of patterns can create a 'busy' appearance, making it harder for viewers to focus on the main data trends [5].
- **Raster** Usage of low resolution raster images instead of vectorial format. When a figure uses a raster image with poor resolution it appears unpleasantly grainy – espectially when zoomed in – and possibly difficult to read.
- **Shadow** The graph makes use of shadows. Using shadows as well as other decorative visual effects reduces the data-to-ink ratio and makes the diagram cluttered, thus weakening the visual message [1].
- **TooMuchPrecision** The graph reports values that have a too high precision (decimal digits) for the intended purpose. Any additional information that is not necessary increases the cognitive load and makes the graph less understandable [5].
- **WasteSpace** A large portion of graph area is empty. This is a bad us of space that dilutes the visual message in the graph.
- **WrappedXLabs** Labels on the x axis are wrapped due to limited space. Wrapped labels are more difficult to read, often they can be solved with a simple graph redesign.

III. Experimental design

The general goal of the study can be formulated using the GQM template [12]:

Table II: GQM definition

Analyze	the usage of figures articles
For the purpose of	understanding
With respect to	the type, technique, and mistakes
From the viewpoint of	paper authors and reviewers
In the context of	SE conferences and journals

A. Research questions

In order to achieve the above goal we define the following research questions:

- **RQ1**. Mode: how are figures used in SE articles?
  To have an initial assessment of the phenomenon we consider important to understand how many figures are used in SE papers. Also it is interesting to observe if there is a trend in time concerning the usage of figures.
- **RQ2**. Content: what are figures used for?
  Figures are used to convey many different types of information. The most general distinction is between quantitative and non quantitative information. A further step is to look into the different type of contents shown in the figures.
- **RQ3**. Type: what types of diagrams are used to convey quantitative information?
  Focusing on quantitative diagrams we investigate the type of diagrams used in the papers.
- **RQ4**. Mistakes: What are the errors committed in quantitative diagrams?
  We focus on all diagrams and then we analyze the specific mistakes committed in the most used diagram types.

For all the question above we aim to investigate whether a change can be observed for different venues and if time affected any aspect.

B. Variables

To investigate the above research questions we collected a set of variables that are described in Table III.

In particular we collected measures on two type of entities: the articles and figures that appear in them.

Table III: Variables

Entity	Variable	Description
Article	VenueType	categorical: { Conference, Journal }
	Venue	string: name of conf. or journal
	Year	integer: year of publication
	Pages	integer: pages of the article
	NumFigures	integer: number of figures
	FigDensity	derived: NumFigures/Pages
Figure	Category	categorical: {Q, NonQ}
	Type	categorical: type of content
	Mistake	set of categorical: the errors found

Concerning the figures, in addition to the main distinction between quantitative (Q) and non-quantitative (NonQ), we

categorized the type of content. The list of types of figure contents are reported in Table IV divided into quantitative and non quantitative.

Table IV: Types of figures

Category	Types
NonQ	Code, Graph, Picture, Schema, Screenshot, Table, Wordcloud
Q	Alluvial, Area, Bar, Bar Grouped, Bar Stacked, Bar Stacked Diverging, Beeswarm, Boxplot, Bubble, Bump, Dendrogram, Donut, Dot, Forest, Heatmap, Line, Pie, Radar, Scatter, Slope, Sunburst, Surface, Treemap, Venn, Violin

The taxonomy of graph type was initially formed on the basis of the main graph types described in the literature, e.g. [5]. Then it was updated when, during the analysis of the articles, a diagram that was impossible to classify appeared.

C. Procedure

We selected a set of recent issues of two leading SE Journals – IEEE Transactions on Software Engineering (TSE) and Empirical Software Engineering Journal (EMSE) – and SE conferences – ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM) and ACM/IEEE International Conference on Software Engineering (ICSE) –. We downloaded all the articles in 2022 issues of the two journals, those appearing in the year 2017, 2019, 2021, and 2022 of ESEM and a sample of those appearing in years 2018, 2019, 2021, and 2022 of ICSE[1]. For each article we went though it and for each figure we classified it using the taxonomy presented in the section above. In addition, for the quantitative figures, we applied the error taxonomy described in Section II-B to identify errors present in the figure.

The person performing the analysis started with a limited number of articles then they discussed all the collected data with the leading researcher to define the correct application of the taxonomy. After the processing of a whole venue a second round of discussion focused on the dubious cases that emerged during the analysis.

IV. Results

Overall in our study we analyzed a total of 865 articles that included 6342 figures. The detailed counts for the different venues we took into consideration are reported in Table V.

A. RQ1: Mode

The first RQ focuses on how much figures are used in SE papers to convey information.

To address this question we looked at the number of figures that are used in the articles. Since the length of the articles can be quite diverse, we computed a derived measure that is the density of figures, i.e. the number of figures per page. Fig. 1 reports the figure density for the four different venues, i.e. the conferences ESEM and ICSE as well as the journals TSE and ICSE. The figure shows the distribution of density using a boxplot and reports the mean value as a cross.

[1]Due to the large number of articles in ICSE, only a sample of the total article was analyzed except for year 2019.

Table V: Summary of articles and figures analyzed

VenueType	Venue	Year	Figures	Articles	TotalArticles
Conference	ESEM	2017	180	63	63
		2019	183	48	48
		2021	143	24	24
		2022	123	24	24
	ICSE	2018	95	14	152
		2019	641	109	109
		2021	209	38	138
		2022	584	69	99
Journal	EMSE	2022	1918	192	192
	TSE	2022	2266	284	284
	Total		6342	865	1133

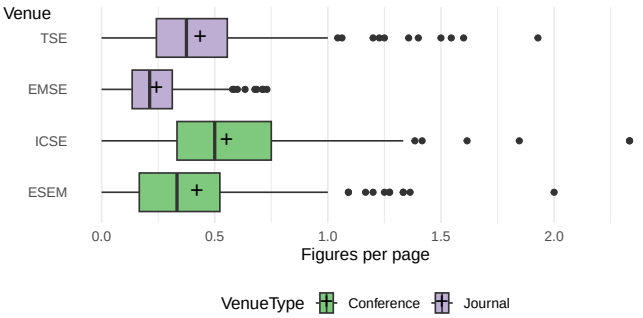

Fig. 1: Figure density (figures per page) of papers in different venues

We observe that on average the papers have a mean density of 0.42 figures per page and the median is 0.35. The details are reported in Table VI. The proportion of articles that have no picture is 2.9%, with the highest percentage for ESEM conference (9.4%).

Table VI: Summary of figure density by venue

Venue	Figures	Pages	Mean F/P	Median F/P	w/o Fig
ESEM	3.9	9.0	0.42	0.33	15 (9.4%)
ICSE	6.6	12.0	0.55	0.50	5 (2.2%)
EMSE	10.0	41.0	0.24	0.21	3 (1.6%)
TSE	8.0	18.3	0.44	0.38	2 (0.7%)
Total	7.3	20.0	0.42	0.35	25 (2.9%)

By looking at the figure we can observe significant differences among the four venues. Such visual assessment is confirmed by the result of an ANOVA analysis of figure density vs. Venue that are reported in Table VII. Considering as the reference level for the Venue variable the ESEM conference, by looking at the coefficient estimates we observe a significantly lower higher value for the ICSE conference and lower for EMSE journal, while the density of TSE is not different.

Table VII: ANOVA results of Figure density vs. Venue

Coefficient	Estimate	Std. Error	t value	p.value
(Intercept)	0.421	0.023	17.967	<0.001 ***
VenueICSE	0.131	0.030	4.309	<0.001 ***
VenueEMSE	-0.177	0.032	-5.596	<0.001 ***
VenueTSE	0.016	0.029	0.534	0.593

Since the length of articles we report in Fig. 2 a scatter plot of figure density vs. number of pages.

Fig. 2: Figure density vs. pages in article for different venues.

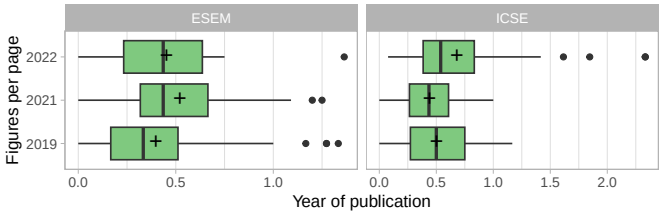

Fig. 3: Variation of figure density in different years

We observe a small negative correlation (Pearson r=-0.273) between the number of pages and the figure density.

To understand if a change in time occurred we compared the two conferences over the same years. Fig. 3 reports the boxplot of density over three years.

From the figure we can observe a negligible (r=0.138) correlation of density with year of publication.

B. RQ2 Content

Fig. 4 reports the proportion of quantitative (Q) vs. non quantitative (NonQ) figures in the article, divided by venue.

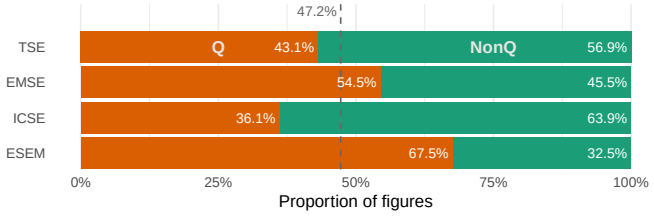

Fig. 4: Category of figures

Roughly half (47.2%) of the figures are used to convey quantitative information while the remaining are used to represent other types of information. The proportion varies notably among the four venues considered, with ESEM conference articles having an average of two thirds figures being quantitative, while ICSE articles invert the proportion with around one third of quantitative figures.

The detailed distribution of the proportion of quantitative figures per paper is reported in Fig. 5. We observe that in

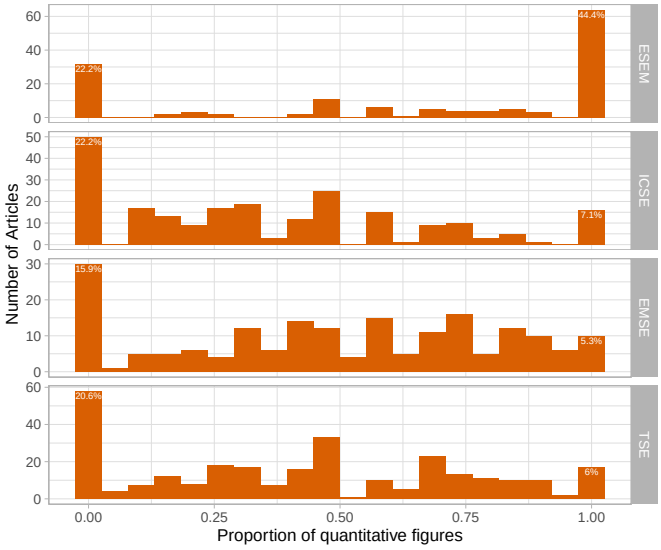

Fig. 5: Distribution of quantitative figure proportions

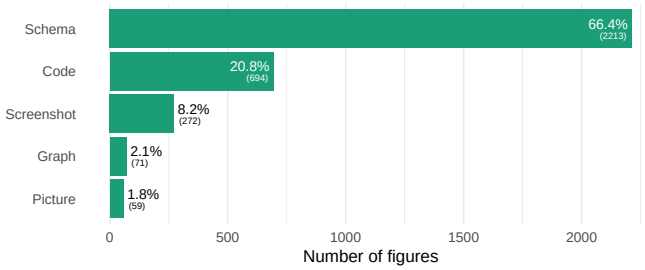

Fig. 6: Type of non quantitative figures

the case of ESEM the distribution is very extreme with 22% of articles that have no quantitative figure at all and 44% that contain only quantitative figures. For the other venues the proportion articles with only quantitative figures is much smaller with proportions ranging from 5% to 7%, while the proportion of articles with only non quantitative figures is close to 20%.

Focusing on the non quantitative figures only, Fig. 6 reports the numbers as well as the proportion of the five different non quantitative types defined in our taxonomy. Two out of three non quantitative figures fall into the broad *Schema* category that includes all the diagrams use to explain something, including also software architecture or UML diagrams. Around one in five *NonQ* figures are used to depict code, this is common practice used instead of the *Listing* environment. A smaller proportion (8.2%) is represented by screenshots. Eventually, we observed a few figures depicting graphs – set of nodes and edges possibly labelled – and just 59 figures reporting pictures or photographs.

C. RQ3 Types

We report in Fig. 7 the overall proportion of the different types of quantitative figures present in our taxonomy.

Overall one of every three quantitative figures contains a bar plot The two other common used plot types are boxplots

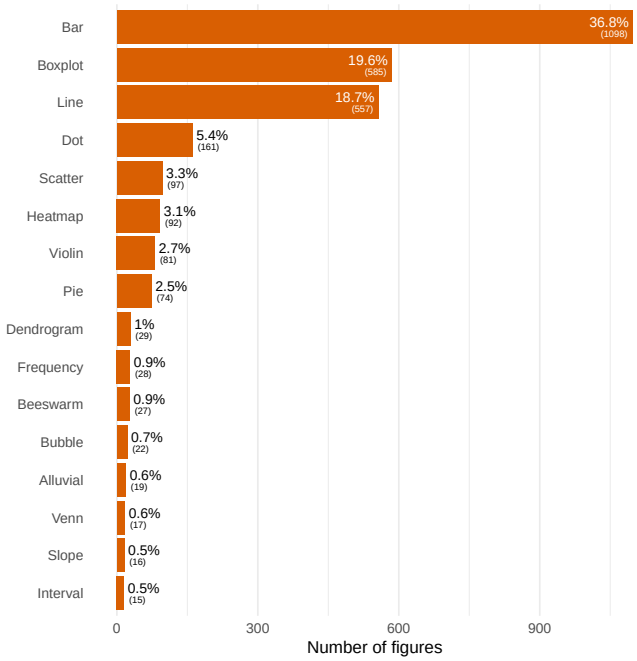

Fig. 7: Type of quantitative figures

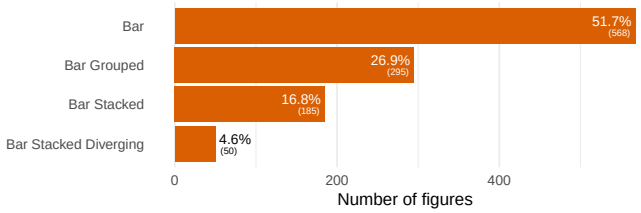

Fig. 8: Specific sub-types of bar graphs

(19.6%) and line plots (18.7%). Dot plots and scatter plots together make 8.8%, heatmaps account for 3.1% of quantitative figures, the remaining types overall account for 13% of quantitative diagram, any of them accounting for less than 3% individually.

In the taxonomy we used for quantitative diagram types – reported in Section III-B –, we have different types of bar plots. Fig. 8 reports the number and the proportion of the different type of bar charts. Apart the simple bar charts that are used in half of the cases, one in four bar diagram use grouped or clustered bars, and 17% use stacked bars. A small minority of bar diagram are diverging diagrams.

D. RQ4: Errors

On the basis of the error taxonomy described in section Section II-B we detected the errors committed in the quantitative figures. The violin plots in Fig. 9 describe the distribution of errors per quantitative figure detected in the articles appearing in the four considered venues.

We observe the average number of errors per figure is 1.6 (median 1). Overall half of the figures exhibit at least one error, with half the EMSE articles showing at least 2 errors. Concerning the proportion of figures where no error

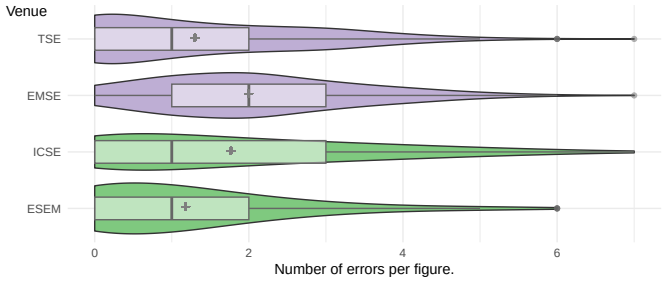

Fig. 9: Number of errors per figure by venue

was detected, they are 37.4% for ESEM,30.6% for ICSE,12.2% for EMSE,40.1% for TSE.

A more detailed picture can be gained by looking at the frequency of error free figures by severity level and venue reported in Table VIII.

Table VIII: Proportion of error free figures per severity and venue

Venue	Critical	Major	Minor
ESEM	79.3%	64.6%	62.0%
ICSE	77.9%	49.2%	54.4%
EMSE	74.3%	40.3%	41.1%
TSE	77.5%	65.2%	60.1%

Overall we observe that one every four pictures contains at least a critical error, with similar condition across the four venues. Concerning the major errors, the journal TSE and the conference ESEM contain at least one error in one out of three figures, while the proportion of figures with at least one major error is 50% for ICSE and 60% for EMSE. As far as minor errors are concerned, we observe a similar amount of defective figures for all venues except EMSE where 60% of figures contain at least one error.

Focusing down more, from the severity levels to the specific types of errors, Fig. 10 reports the occurrence numbers and proportions of all the error types defined in the error taxonomy we defined in Section II-B. The errors types are reported divided by severity levels.

As far as the major errors are concerned, we observe that the most common error – affecting 9.6% of figures – is the adoption of a colour palette that does not enable an easy distinction of categories. The next error consists in missing references on the axes (7.2%), and eventually the attempt to encode too many distinct categories in a single plot (6.6%).

The most common major error is the use of rotated labels on the x axis (16.2%), next is the use of a legend instead of direct labelling (14.4%). The next error affects figures containing multiple diagrams and consists in using different scales on aligned diagrams (9.9%).

The two most common minor errors regard the use of legends, often (20% of figures) the legend is placed inside the plot, which reduces the clarity, also (17.8%) the legend is surrounded by a border that affects the ease of scanning back and forth between plot and legend itself. Another common error is the use of raster images (15.1%) that affects the overall visual quality of the diagram.

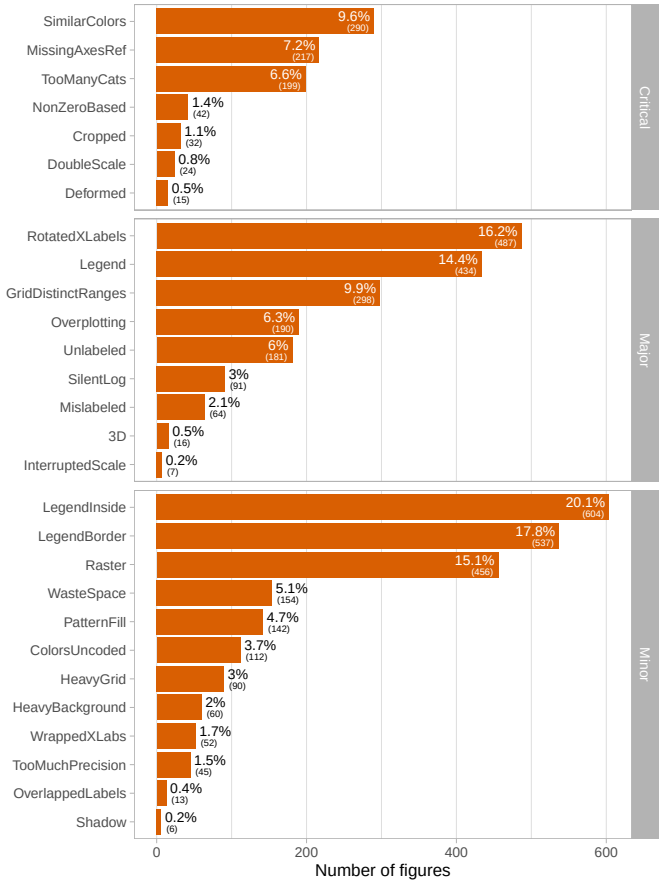

Fig. 10: Overall frequency of error committed in quantitative figures

V. Discussion

Based on the findings reported in the previous section we can answer the original research questions of the research and outline a few additional considerations.

A. Answers to Research Questions

As far as the mode of use of figures (RQ1) is concerned, we observed a relatively large adoption of figures with an average of 0.42 figures per page. As a comparison the present paper contains 12 figures in 8 pages, that is 1.5 figures per page, a clear outlier if we look at Fig. 1. Two venues are aligned with the average (ESEM and TSE) while the two other depart significantly: ICSE has higher density while EMSE has lower density. Concerning this latter difference, it can be explained by the larger number of pages of the articles published by EMSE.

RQ1: how are figures used in SE articles?

On average, there are two figures every five pages with some difference among venues, partly due to the length of articles.

Concerning the category of figures – quantitative vs. non-quantitative –, we found that 47% of figures is quantitative but there are large difference among the three considered venues. Articles appearing in ICSE have just 1/3 of Q figures while ESEM papers have 2/3. A possible explanation for such difference is that since ESEM is hosting mainly empirical studies, the authors often need to show quantitative information and Q figures are widely – 44% of papers contain only quantitative figures – used for this purpose. On the other hand, ICSE hosts papers that have less empirical content and thus employ Q figures less, moreover the wide spectrum of topics present at ICSE require to provide context and details that are better explained with diagrams; such use of overview diagrams could explain the higher number of figures in ICSE papers, in fact the proportion of articles without quantitative figures is simila in ICSE and ESEM.

RQ2: what are figures used for?

A bit less than half of the figures are used to convey quantitative information, with wide variation from 1/3 to 2/3. Overall one in five papers have no quantitative figure at all; while only 6% of papers contain only quantitative figures, with the notable exception of ESEM conference where 44% articles do.

When looking at the type of quantitative diagrams used, we observe that more than one in three is a bar plot (36.8%). The second most used diagram type are boxplots (19.6%), in fact showing the distribution of a set of values is a common necessity in empirical studies. Other means to show distributions are violin plots (2.7%) and beeswarm (0.9%), in addition to histograms that in our taxonomy are conflated with bar plots. Line plots represent the third most common type of quantitative diagram (18.7%). They address the common requirement of showing trends and relationships between pairs of variables. Scatter plots that have similar use are much less common (3.3%). Other types of diagram that are known for perceptual issues, that is pies and bubbles, are rarely used: 2.5% and 0.7% respectively.

RQ3: what type of quantitive diagrams are used?

The most common type is bar plot, used in one out of three figures, followed by boxplots and line plots, each used in 20% of figures.

Talking about errors, we found 1.6 errors per figure on average (median 1). Anyway, across all venues, one in three figures showed no errors according to our taxonomy, while considering only critical errors, three out of four figures are fine. Focusing on major and minor errors, we observed that half of the figures showed at least one.

The use of colours too similar to each other is the most common critical error, this is often due to little care spent in selecting an appropriate palette. The lack of values on the axes is another relatively common error that has a heavy impact on the ability to understand the diagram. Often in graphs that have been generated by basic tools – e.g. spreadsheet programs –, the categories are automatically encoded using

many different variation of a single visual attribute – e.g. shape or color – that require a lot of cognitive effort to be discerned. A little more design effort should be spent in finding alternative representations that make understanding more immediate. Another common result of using unaltered graphs produced by e.g. spreadsheets is having rotated or slanted labels on the x axis, making reading the graph more difficult than required. Most of the time a simple solution is to swap x and y axes; this should be the immediate choice in presence of long labels. Another common mistake that is worth mentioning is the use of a legend instead of direct labelling the series of data in the plot area. This is probably due to the default in most tools as well as the difficulty in implementing direct labelling in those tools.

Most of the times errors could be solved with a small effort. We believe they are introduce because authors do not pay enough attention to the quality of the figures and, in the case of journals, neither the reviewers do. Some basic and easy to follow guidelines are presented in the Empirical Standards for Software Engineering Research [3].

We note that very few figures show evident mistakes such as using non-zero based bars, copping figures, adopting double scales or deforming the graph area; still almost 5% of figures are affected.

RQ4: what errors are found in quantitative figures?

The average figure in SE articles contains one or two errors, although on average 1/3 of figures show no error. A common critical error is the use of hard to discern palettes. The use of rotated labels on the x axis is the most widespread major error. Common minor errors include a less than ideal use of legends.

B. Limitations

The study presented in this paper is exploratory and therefore presents several limitations; we highlight the main ones.

Only a few selected SE publication venues have been considered. Although the venues have a very good reputation, they do not represent the whole publication spectrum in SE. Moreover the articles have been drawn from a limited time span, they represent a sample. This might affect the external validity of the study. The graph error taxonomy has been built based on the literature in the area of data visualization, although it has not been empirically validates. It could have omitted important errors or, vice-versa, considered error that are not such in the view of other researchers. In addition the severity level has been assigned to each error on the basis of the author expert judgment. In the classification of the type of quantitative diagrams we considered only the main type, although there are several cases of diagram that mix multiple types, e.g. bars + lines, that were not considered.

VI. Conclusions

In this work we collected articles from four leading SE publication venues, two conferences and two journals, and analyzed how figures are used. We found the average density is 0.42 figures per page and half of them are used to convey quantitative information. The most common quantitative graph types are bar plots, boxplots, and line plots. We also classified the errors committed in the quantitative graphs and we found that on average every figure has 1 or 2 errors.

Most errors are relatively easy to address and we advocate for a wider diffusion of visual literacy in the SE research community.

As future work we would like to extend the survey to a wider time span as well as other venues. In addition we should take into consideration figures with multiple types. Eventually, a pragmatic guide to assess and improve quantitative diagrams and help researchers avoid the most common pitfalls, could be defined based on this work.

References

[1] E. R. Tufte, *The visual display of quantitative information*. Graphics Press, 1983.

[2] T. Munzner, *Visualization analysis and design*. CRC Press, 2014.

[3] P. Ralph *et al.*, "Empirical standards for software engineering research." 2021 [Online]. Available: https://arxiv.org/abs/2010.03525

[4] A. V. Pandey, K. Rall, M. L. Satterthwaite, O. Nov, and E. Bertini, "How deceptive are deceptive visualizations? An empirical analysis of common distortion techniques," in *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 2015, pp. 1469–1478.

[5] S. Few, *Show me the numbers, second edition*. Analytics Press, 2012.

[6] C. Ware, *Information visualization: Perception for design*. Elsevier Science, 2013.

[7] S. Silva, B. S. Santos, and J. Madeira, "Using color in visualization: A survey," *Computers & Graphics*, vol. 35, no. 2, pp. 320–333, 2011.

[8] L. W. MacDonald, "Using color effectively in computer graphics," *IEEE Computer Graphics and Applications*, vol. 19, no. 4, pp. 20–35, 1999.

[9] J. Heer, F. B. Viégas, and M. Wattenberg, "Voyagers and voyeurs: Supporting asynchronous collaborative visualization," *Communications of the ACM*, vol. 52, no. 1, pp. 87–97, 2009.

[10] R. M. Ratwani, J. G. Trafton, and D. A. Boehm-Davis, "Thinking graphically: Connecting vision and cognition during graph comprehension," *Journal of Experimental Psychology: Applied*, vol. 14, no. 1, p. 36, 2008.

[11] W. S. Cleveland, *The elements of graphing data*. AT&T Bell Laboratories, 1994.

[12] R. van Solingen and E. W. Berghout, *The goal/question/metric method: A practical guide for quality improvement of software development*. McGraw-Hill, 1999.