Interactive visual manipulation of large-scale line data

by

Abel de Bruijn

---

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Friday January 16, 2025 at 9:00 AM.

Faculty:	Electrical Engineering, Mathematics and Computer Science	
Programme:	Master Computer Science	
Research Group:	Computer Graphics and Visualization	
Thesis committee:	T. Höllt, PhD	TUDelft, Supervisor
	J. Urbano Merino, PhD	TUDelft
Project Duration:	April, 2025 - January, 2026	

TU Delft

Interactive visual manipulation of large-scale line data

A. de Bruijn[1]

[1]TU Delft, The Netherlands

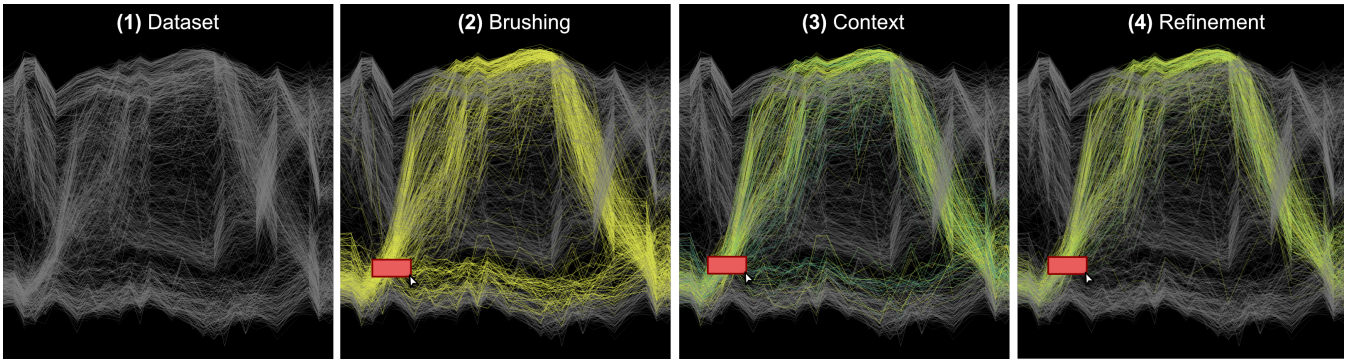

Figure 1: *A complex visualisation of numerous overlapping grey lines is presented, representing raw data (1). A three-stage workflow for analysing large-scale line datasets is used to understand this data: Interactive brushing (2) is shown, where a red rectangular brush is utilised to highlight certain lines that intersect with yellow. A context-aware brush with the same shape is introduced (3). It enables missing lines to be added or non-conforming lines to be removed to correct contextual inconsistencies. A separate refinement technique (4) is able to correct further by removing all outliers.*

Abstract

As line datasets grow larger, the demand for effective visual data analysis becomes increasingly important. Understanding large-scale datasets remains a fundamental challenge. A critical trade-off is presented by existing line selection methods: they either produce efficiency, accuracy, or human interpretability, rarely achieving all three simultaneously. This gap is addressed by the development of human-guided and context-aware brushing techniques, which are supported by manual, semi-automatic and automatic refinement methods. Through empirical evaluation via two user studies, it was found that, whilst context-aware brushes offer theoretical promise, statistical superiority over conventional brushing approaches is not demonstrated. However, selection accuracy is consistently improved by refinement techniques, with manual refinement yielding the highest accuracy gains (12.6%) followed by semi-automatic refinement (9.8%). Notably, efficiency gains from refinement remain dataset-dependent, with no single technique universally dominating across varied data characteristics. Manual and semi-automatic refinements are preferred by users seeking high-accuracy improvements. Although similar efficiency scores are exhibited by manual and semi-automatic refinements, the lowest variance is observed for the semi-automatic method; consequently, it is recommended for users prioritising efficiency. The findings emphasise a fundamental design principle: Interpretability and user agency should be prioritised over full automation.

CCS Concepts

• ***Computing methodologies*** → *Optimization algorithms;* • ***Human-centered computing*** → ***User studies;*** *Web-based interaction; User centered design;*

1. Introduction

Brushing stands as a fundamental and widely adopted interaction technique for visual data analysis. Over three decades ago, Becker and Cleveland [BC87] defined brushing as an interactive method enabling users to select subsets of data points through simple geometric shapes, such as squares, circles, or polygons, directly applied to data visualisations. Since then, numerous efforts are made to extend brushing techniques to more complex data representations, particularly line-based visualisations.

In recent years, datasets have grown substantially in size, often containing thousands or even millions of chart lines. This growth is particularly evident in agriculture [MS25] and art restoration [PGK*22]. Without proper data structure, large datasets quickly become uninterpretable to humans Figure 1(1). Line brushing offers a solution to this problem. Line brushing Figure 1(2) is a common method for selecting data subsets. When users select a subset of data points, the entire line is highlighted, enabling them to interact with data directly. This helps users to recognise patterns, trends and structures.

However, existing line brushing methods present a persistent issue. They typically optimise for either time efficiency, selection accuracy or human interpretability, rarely achieving all. This trade-off significantly limits their usefulness in real-world data analysis. For users to work effectively with large datasets, they need all three qualities simultaneously.

An approach based on context-aware brushing techniques deliberately balances accuracy and efficiency, while it is still understandable to users Figure 1(3). Context is added through line length and the amount of parallel information. It is acknowledged that user errors are inevitable during interactive exploration. Consequently, the brushes are supported by a three-level refinement framework ranging from high user autonomy (manual adjustment) to high automation (machine-guided suggestions) Figure 1(4). Established brushing interactions from Parallel Coordinate Plots (PCPs) [RLS*19, REB*16] are generalised to function-based line data, making them more widely adaptable to line data such as time-series and spectroscopic data.

To empirically validate the approach, a dual user-study is conducted to examine where each context brush and refinement strategy performs optimally across different data characteristics. The findings reveal that whilst context-aware brushes introduce some interaction complexity they could become useful when developed further. The refinement strategies show dataset-dependent performance patterns, suggesting that no single approach universally dominates. Based on these results, a decision framework is defined to identify the most appropriate brushing and refinement combinations for specific dataset characteristics. This offers a concrete guidance for method selection in real-world scenarios.

This paper makes three primary contributions aimed at bridging the divide between automated efficiency, human interpretability, and selection accuracy. First, the characteristics of four established brushes are compared against two novel context-aware brushes. Second, a comprehensive empirical evaluation of multiple refinement strategies across time-series, and parallel coordinate plot data, providing critical insights into the accuracy, efficiency, and user satisfaction of each approach across diverse datasets. Lastly, a novel dataset is introduced comprising over 150 human-generated selections. These selections are made while using the established and context-aware brushes on different line data, serving as a benchmark for training and validating. These can be used to compare future brushing techniques and refinement algorithms. These brushes were evaluated on a interactive platform for rapid prototyping, evaluating, and iterate upon novel brushing and refinement mechanisms for direct line selection. Both the user-selections and platform are publicly accessible at `https://osf.io/tbfmp`.

The remainder of this paper is structured as follows: the necessary background is provided in Section 2 to establish the foundation for the remainder of the paper. Alternatives for making line-data understandable and established brushes are highlighted in Section 3. This is followed by Section 4, where the context-aware brushes are introduced, including the refinement methods. The results and discussion of the preliminary study are detailed in Section 5, which focuses on identifying the best brushes for each scenario. A more in-depth study is presented in Section 6, where the primary objective is to determine which refinement method is preferred by participants across various scenarios. The limitations and future directions of both studies are discussed in Section 7. Finally, the conclusion is presented in Section 8.

2. Background

To limit the scope of the research, a dataset D is introduced where each line $L_i \in D$ is defined by a function $f(x) = y$. Here, x is an independent variable within the interval $[1,N]$ and y is a real-valued dependent variable. To enable effective visual comparison among a diverse range of datasets, all y-values are normalised to a consistent range. Therefore, the entire span of y-values across the dataset D, defined by its global minimum $Y_{\min}$ and maximum $Y_{\max}$, is linearly transformed.

For the method to effectively support visual analysis, a subtle but crucial limiting criterion is considered. The total variation [DJ98] of y along each line should not be excessively large. Specifically, for each line L_i composed of N discrete data points (x_j, y_{ij}) where $j = 1,\ldots,N$, its total variation $TV(L_i)$ is defined as the sum of the absolute differences between consecutive y-values:

$$TV(L_i) = \sum_{j=2}^{N} |y_{ij} - y_{i,j-1}|$$

For most lines L_i in the dataset D, the total variation $TV(L_i)$ is expected to remain below a certain threshold. This limitation is important because as total variation increases, individual lines become jagged, making it harder for humans to detect underlying patterns. When this assumption no longer holds, humans struggle to make accurate manual selections or guide a semi-automatic system in the right direction.

This accommodates various data types, including time-series where x represents time, spectral data where x represents wavelength, or multi-dimensional data where each line represents the trajectory of values across different dimensions, often visualised in Parallel Coordinate Plots (PCPs) introduced by Inselberg [Ins85].

3. Related work

Based on the foundational concepts discussed in the Background section, this section aims to illustrate diverse methodologies and provide an overview of existing strategies. It offers techniques for effective data representation, followed by a showcase of established accurate selection methods.

3.1. Understanding the visualisations

A line plot with many groups (more than six) is commonly referred to as a "spaghetti plot" [Dig13]. With large datasets, the assignment of distinct colours, sizes, or transparencies to individual lines can increase visual clutter, particularly when numerous lines intersect. Extensive enhancements have been developed to facilitate the analysis of patterns in line data. A comprehensive overview is beyond the scope of this paper but may be found in surveys by Ellis and Dix [ED07], Heinrich et al. [HW13], and Behrisch et al. [BBK*18]. The remainder of this section addresses the key papers that motivated the present work.

Using automatic time-series classification (TSC) [RAM05] [JJO11] to group line data into clusters provides less attention to each individual line and provides a way to notice trends within line clusters. While these methods can provide efficiency, they often compromise accuracy [MS25] or human understandability [LPC*24]. This is largely due to limitations in transparency and parameter sensitivity, which makes it challenging to adjust their underlying parameters.

A more direct approach to line grouping is achieved through density estimation. Curve Density Estimates (CDE) [LH11] is described as a smooth, continuous approach grounded in kernel density estimation and effective for high-frequency curves and multimodal data. DenseLines [MF18] is presented as a discrete, binned approach that normalises by arc length for multiple related lines. Density is encoded by both methods through colour intensity, enabling instant pattern recognition, outlier identification and cluster discovery. Fundamentally, differences exist between them. Smooth probability fields suited to single curves are created by CDE, whereas discrete visualisations optimised for comparative analysis across multiple series are produced by DenseLines. For human understanding, both facilitate aggregate trend identification and distribution highlighting. The trade-off is that CDE's smoothness can obscure multiple clusters while DenseLines' binning may complicate within-cluster comparisons.

When line data is pre-clustered, multiple ways are available to showcase each group of lines effectively. Enveloping [Her89] [JSK11, p. 417] is defined as a curve that encloses groups of lines. Instead of showing every individual line, upper and lower boundaries are created around clusters of similar paths by enveloping, thereby revealing the overall structure and flow patterns and reducing visual clutter. Lastly, a tool named Line Weaver [TB21] manages to prioritise clusters of lines that have low spread of standard deviation. These groups are sorted (locally and globally) on the z-axis in ascending order of least standard deviation spread. Both techniques preserve overall structure and flow patterns, but they differ fundamentally in their approach to displaying multiple objects. Enveloping renders only a thickened line trace for each cluster, whereas Line Weaver maintains the visibility of individual lines. Moreover, Line Weaver addresses a critical rendering issue: the order dependency of alpha blending. Rather than being competing approaches, these methods can be complementary. Enveloping could serve as preprocessing for extremely dense datasets, reducing the initial line count before applying Line Weaver's sophisticated rendering. Conversely, Line Weaver help by preserving the visibility of outliers.

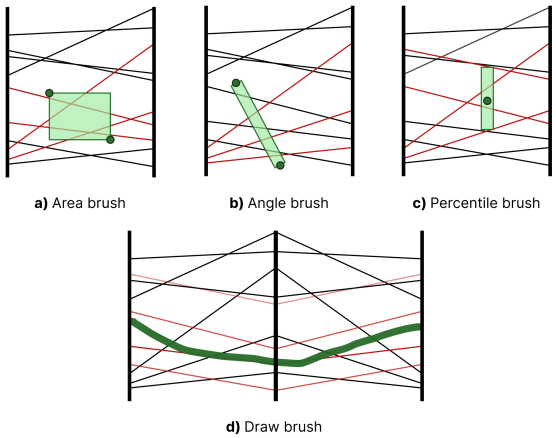

Figure 2: *Four manual brushes.* ***a)*** *rectangle brush selects all lines that intersects one of its borders.* ***b)*** *angle brush selects all lines that are close to perpendicular to the brush.* ***c)*** *Percentile brush is initialized with a single click and grows a vertical line until k% of lines intersect it.* ***d)*** *Draw brush selects lines that are close to parallel to the users selection.*

3.2. Brushing

In 1994, Ward introduced a brush for PCPs in the XmdvTool [War94,MW95]. This tool allows users to "paint" or drag the mouse to select lines of interest between two adjacent axes. Roberts et al. [RLS*19] and Raidou et al. [REB*16] defined an extensive list of brushes for PCPs. To use these brushes in a general $f(x) = y$ setting, some brushes required slight adaptation. This process is described in detail in this section. A summary of these brushes is shown in Figure 2.

A rectangular selection brush for time series is presented by Hochheiser and Shneiderman [HS02]. Lines that intersect the bounding box formed between two points can be selected using this tool. As shown in Figure 2 (**a**), an identical brush implementation is adopted. A similar rectangular selection for parallel coordinates plots is presented by Raidou et al. [REB*16]. This rectangular selection is restricted to allow only a selection boundary to be placed between two consecutive dimensions. Such restrictions prove limiting for line plots in general when plots contain excessively large x-value range, or when lines shifted along the x-direction cannot be captured within a single selection.

The angular brush was originally defined by two lines originating from one of the PCP axes [HLD02,SGMS21]. One line marking the maximum angle and one a minimum angle. All lines that fall in the same range as the two lines are selected. As can be seen in Figure 2 (**b**), the implementation is modified slightly, making it more akin to the rectangle brush. Again a start and end point are first created by the user. All lines whose orientation are perpendicular to the reference line (defined by two points) by no more than an externally defined 'openness' parameter are selected. These changes aim to keep atomic interactions similar, which helps participants establish a common interaction language. Lastly, the brush functionality is extended to work across multiple dimensions.

The percentile brush defined by Radoš et al. [RSM*16] uses one click to specify where the brush should initialise. It operates by selecting a precisely predefined percentage (k%) of data items closest to the users click. When the brush is moved freely across the visualisation, its extent is continuously adapted to maintain the exact k% selection.

A specific trajectory across n-dimensions is selected by the draw brush presented by Roberts et al. [RLS*19]. A poly-line – a line composed of multiple points – is established by drawing from left to right. Once defined, the system compares each data point to the user-drawn poly-line using the squared Euclidean distance at each point. An adjustable openness parameter then determines the acceptable tolerance around this line. Lower openness values produce selections that more closely match the user's original drawing.

4. Methodology

Brushing tools are prone to user error, thereby reducing their accuracy. Lines may be positioned on the wrong side of the selection border, leading to both unwanted inclusions and exclusions. Additionally, lines are often difficult to follow along their entire path. These problems may be addressed by designing brushes with specific line characteristics that guide users toward more accurate selections. Outliers can be filtered, further reducing over-selection issues.

4.1. Context aware brushes

As described in Section 3.2, all brushes except the rectangle brush (Figure 2-**a**) use an external parameter to determine specificity. Two experimental brushes are introduced that exploit context from the line plot to parametrise a rectangular brush. Each brush is subsequently described.

Length. Lines exhibiting similar length or TV, as defined in Section 2, are considered proximate. A rectangle is drawn within the line chart to define a region of interest, and the minimum and maximum lengths of all lines intersecting this region are calculated. These length measurements establish a reference range derived from the total variation characteristics of lines within the selection. Lines whose lengths fall within the average length of this range, adjusted by a user-controlled tolerance parameter, are subsequently selected. By default, the parameter is set to the difference between the minimum length and the average length. Hereby, selecting at least all lines that fall within the users selection.

Parallel. An alternative approach involves the selection of lines that flow in a similar manner to an initial selection. In spectroscopy [UZS*18], for example, two spectral lines may originate from an identical samples yet exhibit differential intensities because a dust particle partially obstructs the sensor. Additional examples exist for different dataset types and measurement approaches.

For each line selected with the rectangle brush, all intersecting lines are averaged into a single mean line LM. The distance between each line in the dataset and LM is then calculated:

$$Distance(L,LM) = |\sum_{j=1}^{N} (LM_j - LM_{j+1}) - (L_j - L_{j+1})|$$

This metric quantifies the cumulative difference in slope vectors between a line L and the mean line LM, thereby measuring structural patterns rather than absolute spatial proximity. The relative delta line distance does not account for the absolute position between the line and LM, enabling the selection of parallel lines even when a cluster contains a large spatial gap. This approach is particularly valuable in applications where physical separation is unrelated to structural similarity.

An openness parameter λ is selected, applying the following function to each line:

$$f_\lambda(L) = \begin{cases} 0, & \text{if Distance(L, LM)} > \lambda \\ 1, & \text{otherwise} \end{cases}$$

This binary function determines whether a line L is retained or discarded based on the proximity of L to the mean line LM in terms of the overall slope pattern. Lines that deviate from the pattern are considered outliers and should consequently be eliminated. When unselected lines with similar patterns are detected, the inconsistency can be rectified.

These contexts could also apply to the angle, percentile, and draw brushes. However, these brushes already possess an external parameter. With too many variables to adjust, users may struggle to identify which parameter to modify. This makes it more difficult to compare how well each brush performs based on each parameter.

4.2. Outlier filtering

Participants are assumed to be well equipped to approach a selection goal, although small mistakes may be made. In selection procedures involving multiple lines, precise manual selection can be challenging. Therefore, a method for identifying outliers within the system is proposed, employing the following approach.

Upon initial user selection, a median line L_M is created to represent the central tendency of the selected trajectories. Let $D_S \subseteq D$ denote the set of lines initially selected by the user. This median line is defined for each x-value within the interval $[1,N]$. For a given x, the set of corresponding y-values from the selected lines is considered: $\{f_i(x) \mid L_i \in D_S\}$. To determine $L_M(x)$, these y-values are first sorted in ascending order. The y-coordinate of the median line, $L_M(x)$, is then computed as the statistical median of this sorted list. If the number of selected lines, $|D_S|$, is odd, $L_M(x)$ is the middle value of the sorted list. If $|D_S|$ is even, $L_M(x)$ is the average of the two central values in the sorted list.

Next, all lines in D_S are are evaluated against this derived median line. The comparison is performed by calculating the squared Euclidean distance between each individual line's trajectory and the median line across their common x-value.

4.3. Levels of refinement

Three outlier filtering techniques are evaluated to determine the most effective implementation, ranging from high user autonomy to extensive automation. The following will present each in detail.

Manual refinement The outlier-filter technique presented in Section 4.2 is not the only method of refinement. The brushes described in Section 3.2 and Section 4.1 can be employed to remove unintentionally added lines through precise movements. Whilst this would increase the accuracy of over-selection, the tool is unable to add new lines. The manual refinement process is expected to require more time than the other methods and is consequently expected to be less efficient.

Semi-automatic refinement Lines are sorted in ascending order based on the distance from the median using the approach from Section 4.2. The lines are binned and presented to the user in the form of a histogram. The range of lines to be selected can be specified by participants using a min-max slider, as can be seen in Figure 3. This creates three clusters: one cluster of interest that should remain, and two on either end that contain the outliers.

Automatic refinement Empirical evidence has suggested that simple nearest neighbour classification is exceptionally difficult to surpass in automated clustering [BWK11]. Consequently, the main thread of lines is identified using the k-means algorithm. N-dimensional lines are reduced to a one-dimensional value by applying the distance function from the outlier-filtering methods. The number of clusters can be adjusted with a slider, the default being set to three. A min-max range, similar to a semi-automatic technique, is thereby generated. Because k-means is non-deterministic, different cluster assignments may be obtained on repeated runs, leading to variability in results. When only the cluster of lines corresponding to the goal is selected, reduced processing time is expected. However, correct boundaries may be misidentified by the automatic k-means algorithm. Consequently, lower accuracy is expected compared with the manual and semi-automatic tool.

4.4. Study metrics

To validate the performance of the brush, a comprehensive evaluation was conducted across three complementary dimensions: selection quality, efficiency and user satisfaction. A holistic view of how well each brush technique performs in practice is provided by these three axes.

The evaluation of **selection quality** was guided by the approach of Fan and Hauser [FH21], wherein four fundamental metrics were employed: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These metrics are classified according to whether a line is present in the selection or not and whether the line is present in the goal or not. This classification is illustrated in Table 1.

	Positive	Negative
True	Selection + Goal	Not selection + Not goal
False	Selection + Not goal	Not selection + Goal

Table 1: *Classification of lines according to their presence in the selection and goal, illustrated through the statistical terms True Positive, True Negative, False Positive, and False Negative*

Accuracy, representing the proportion of correctly selected or excluded items relative to the total, constitutes the primary objective. When such accuracy cannot be achieved, two secondary metrics are employed to diagnose whether the selection is over or under-inclusive. Precision, defined as $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$, measures the proportion of selected items that are also present in the goal, and is diminished when the selection exceeds requirements. Recall, defined as $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, measures the proportion of goal items that have been selected, and increases when few false positives are introduced into the selection.

During correct refinement, a FP is converted to a TN, thereby increasing precision whilst recall is unchanged. During incorrect refinement, a TP is converted to a FN, thereby decreasing both precision and recall. Because refinement methods are constrained to permit only the filtering of selections, no other transformations are possible through refinement. Context-aware brushes extend refinement capabilities by permitting the inclusion of missing lines in the adjusted selection through parametric control. Through the use of such brushes, recall can theoretically be enhanced, providing a mechanism to address under-selection and recover true positive classifications that may have been initially omitted.

Beyond selection quality, **efficiency** was assessed through two complementary metrics. The time required to complete each selection was recorded as the primary measure of speed. Additionally, the number of changes to a selections was tracked, which served as a proxy for selection attempts that did or did not achieve the intended result. Separately, the number of deleted selections are classified. A high number of deletions was interpreted as indicating that the brush technique was challenging and inefficient.

To compare time results for each refinement technique per dataset, selection attempts are captured according to defined criteria. For manual refinement, an attempt is captured each time a selection is added or removed. Conversely, multiple selections are not generated for the semi-automatic and automatic refinement tools. For semi-automatic refinement, attempts are captured each time the slider is adjusted. When a slider is dragged, adjustments are aggregated into a single attempt to ensure comparability with other semi-automatic attempts. For automatic refinement, attempts are captured when a button is toggled. Changes to the number of clusters are similarly classified as attempts, with slider dragging aggregated into a single attempt. Since attempts are fundamentally different across refinement techniques, comparisons between refinement techniques should not be made using this metric.

In addition to these objective measures, **user satisfaction** was also evaluated. User satisfaction was assessed based on the understanding that satisfaction is directly influenced by user com-

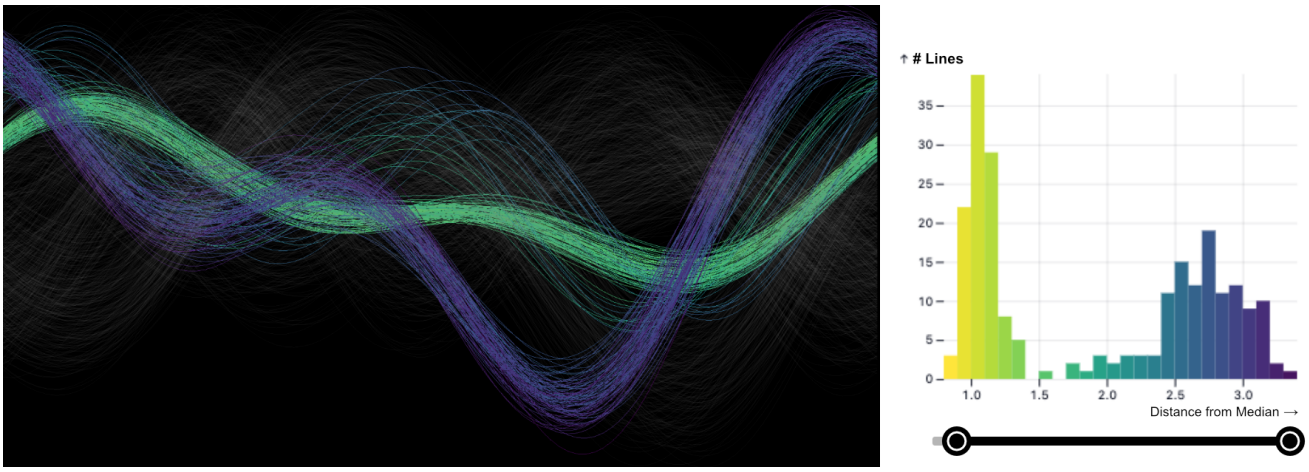

Figure 3: *Histogram of allowed outlier distances from the median, showing the distribution of distance values in the range 0.8 to 3.5. The colour gradient from purple to yellow encodes increasing distance from the median. A min-max slider indicated the range of lines that will be selected.*

prehension. When users understand how to use a technique effectively, a positive experience and high satisfaction are more likely to be achieved. Conversely, when understanding is lacking, users are frustrated, which substantially reduces satisfaction levels. Participants were asked to provide a subjective difficulty score in response to the question: "How difficult or time-consuming was it to make a selection?". In addition, the participants were asked to rate how confident they were during their selection or refinement procedure.

4.5. Dataset difficulty

To evaluate the selection process, participants are tasked to select lines out of a cluster. An image of a cluster is provided, and the selection is required to be recreated using brushing. This is tested on a multitude of datasets to determine whether the selection's properties hold in a more general setting. Datasets are classified as ranging from easy to difficult with respect to the selection process based on the following two criteria: cluster sizes and cluster overlap.

Cluster size indicates how large a selection needs to be in order to correctly select all lines in a cluster. Datasets with extremely small and large cluster sizes present distinct challenges to the selection process. Selection of clusters with small envelopes is inherently difficult because minor brush movements can inadvertently exclude desired lines or include unwanted ones, thereby reducing tolerance for error. Conversely, visual clutter is created by clusters with exceptionally large numbers of lines, making individual line tracking cognitively demanding. Both small and large cluster sizes can be measured by calculating the envelope areas, as discussed in Section 3.1. When the total area of envelopes exceeds the plot size, large overlaps in the clusters can be expected.

Cluster overlap (CO) is quantified in a different way by Maitra and Melnykov [MM10]. Under the Bayes-optimal decision rule, $w_{i|j} + w_{j|i}$ is the pairwise probability that an observation drawn from cluster i would be misclassified as cluster j and vice versa. In

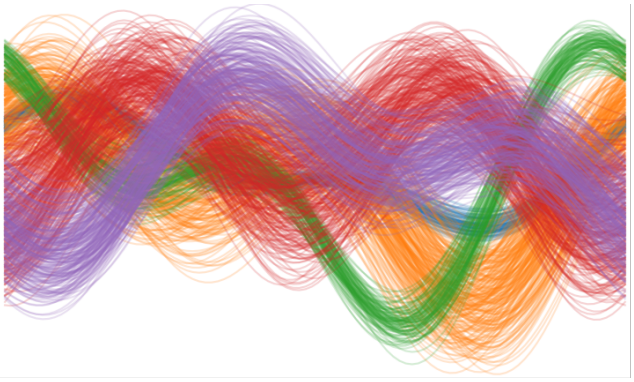

Figure 4: *Syntactic Andrews plot [And72] showcasing five clusters. With two dense cluster of 100 lines (blue), 100 lines (green). Followed by three sparse clusters of lines with 250 lines (Red), 296 (Orange) 250 (Purple). The plot was created by Trautner and Bruckner [TB21].*

the context of line-based selection, cluster overlap becomes particularly critical as it directly impacts the ability to distinguish between clusters visually. When lines from different clusters intertwine or cross extensively, more cognitive load is required during the brushing process. This metric is adapted for line data by computing the overlap at each x-coordinate, measuring the extent to which the range of y-values from one cluster intersects with the range of another cluster. Higher overlap values indicate greater difficulty in achieving precise selections, as careful positioning of brushes is required to avoid unintended lines from adjacent clusters.

5. User Study one: Explorative phase

Two studies have been designed to validate whether the developed tools address the problems outlined in the research objectives of creating an accurate, efficient and human-understandable selection method. An experimental study is conducted with the primary purpose of investigating trends in which methods serve best to increase the level of brush quality. The investigation is refined by a subsequent study in Section 6 which focuses solely on outlier filtering in order to maximise the accuracy gained with as few interactions as possible.

This preliminary study addresses two core research questions: (1) How do the different brushes compare in terms of performance and suitability for the target task? (2) Does the application of outlier filtering enhance selection accuracy and reduce human error without sacrificing efficiency? The following hypothesis is proposed: a tool incorporating selection tolerance with context-aware brushes and outlier-removal refinement reduces human errors caused by imprecise manual selections, thereby improving both the quality and efficiency of desired selections. The remainder of this section presents the study setup, reports the results, and discusses the key implications derived from these initial findings.

5.1. Study setup

The research experiment was conducted using a web tool called ReVISit [CWS*26]. This tool can be used to design, collect data, and reproduce results for many types of online visualisation studies. By using a browser-based user study, participation was made simple, as only a modern web browser was needed. The user study was organised in a group setting where four students and three teachers from the Delft University of Technology were gathered to perform the user test.

A questionnaire was administered, and participants were informed that their responses would be recorded. Identifiable questions were included to examine biases in the study. The questionnaire is provided in Appendix A. The remainder of the study was structured into two separate stages.

In the first stage, an instructional video tutorial was provided to train participants in the user interface. The tutorial demonstrated selection mechanisms, enabling confident use of the brushes. Following the tutorial, participants were assigned a random goal of selecting one of the clusters depicted in Figure 4 using brushing. To minimise bias, each brush was presented in a randomised order. An unlimited number of attempts to add or remove selections was permitted, although a maximum of three final selections was imposed. This cap ensures comparability of results by preventing significant variability in the number of selections.

During the second stage, outlier filtering could be applied to each selection. A tutorial screen explained proper usage of the semi-automatic refinement technique defined in Section 4.3. To limit variability, only the semi-automatic refinement technique was tested in the first exploratory study to assess the effectiveness of the applied refinements. In the second study, all refinement methods were subsequently tested. The re-use of identical brushes with the same dataset introduces a potential memory bias. This risk is acceptable during the exploratory phase but must be eliminated in the final study.

Following each selection, participants were requested to provide self-reported satisfaction scores. Both selection confidence and difficulty were assessed on seven-point scales: difficulty was measured from "Trivial" (0) to "Feels impossible" (6), and confidence was measured from "Not confident at all" (0) to "Extremely confident" (6).

5.2. Essential findings

Several noteworthy findings regarding user interaction with diverse brushing techniques and outlier refinement strategies are revealed by this exploratory study. Given the study's exploratory nature, only essential metrics are reported. Detailed results concerning selection quality (measured via accuracy, precision, and recall), efficiency (expressed in terms of selection time and number of deletions per brush), and self-reported satisfaction (based on confidence and difficulty scores) are presented in Appendix B. The key insights derived from the analysis are discussed below.

A strong implicit preference for the rectangle brush was demonstrated across both quality and temporal metrics. This preference may be attributed to the brush's fundamental status as a geometric selection primitive, a concept likely reinforced through its prevalence in contemporary digital tools. The brush is rated as least difficult (2.14 ± 1.71) and is found to achieve high accuracy (0.96 ± 0.04), without the requirement for outlier removal. Noteworthy is the comparable high accuracy of the angle and percentile brushes relative to the rectangle brush ($\approx$0.95). This is probably attributable to each brush type providing an immediate and unambiguous visual representation of its selection boundary, comparable to the rectangle brush. The resulting selection approximates the expected selection, with only a few trial-and-error iterations required to achieve the target.

A comparison between the parallel brush and the draw-line brush may be considered valuable, as trajectory line representations and distance-based comparisons are relied upon by both approaches. The draw-line brush is designed to compare absolute distances between line points, whilst the parallel brush compares relative distances in slope. When outlier filtering is disabled, higher accuracy is achieved by the draw-line brush (0.80 ± 0.24) compared to the parallel brush (0.70 ± 0.16). This difference is primarily explained by the substantial difference in precision (0.72 ± 0.24 versus 0.37 ± 0.25). The poor precision score for the parallel line brush is attributable to lines from different clusters being inadvertently selected together with the parallel brush, which results in significant variation of the mean line. Consequently, similar lines cannot be identified relative to the initial selection. Since the trajectory line is determined by the user for the draw-line brush, susceptibility to outlier influence is reduced by definition. However, limitations are associated with this control method. Accuracy may be diminished for the draw-line brush when more complex lines are required to be traced, as the trajectories cannot be clearly understood during the drawing process.

Improvement of the parallel brush quality could be achieved

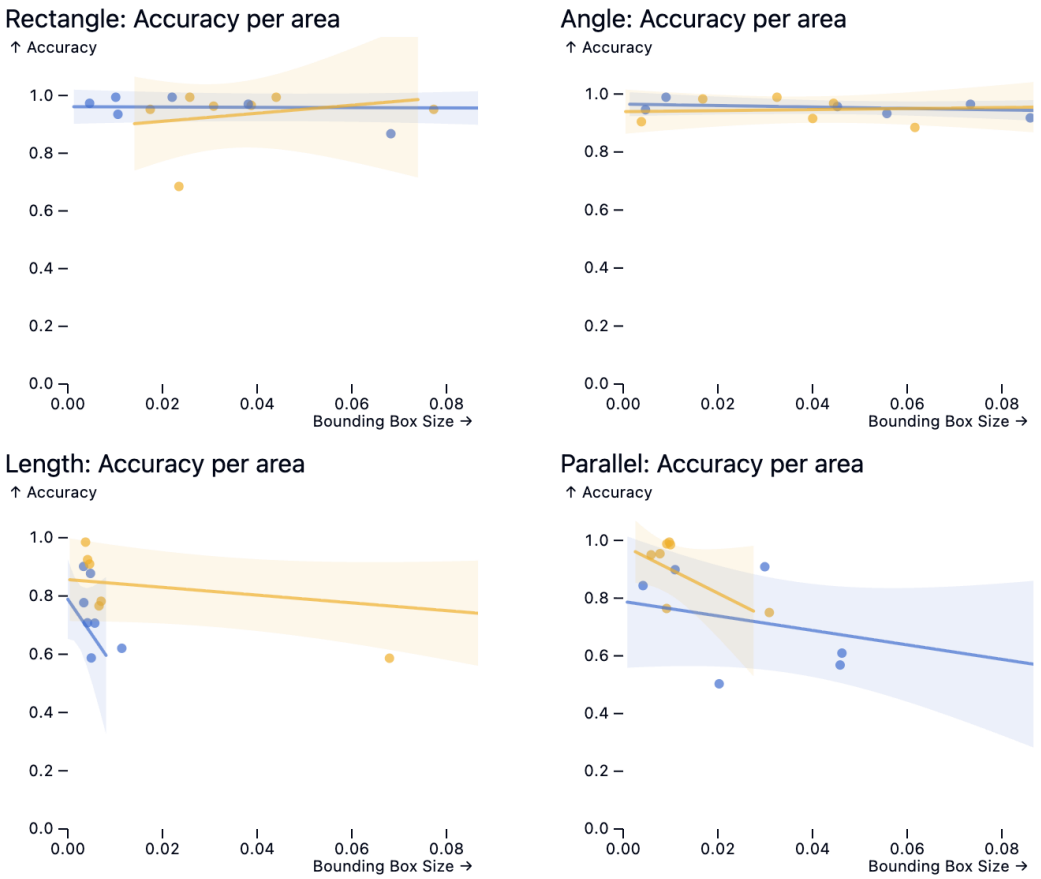

Figure 5: *Accuracy is plotted against the average bounding box area of the selections for each brush type. The selections without an outlier filter are shown in blue, while those with an outlier filter are in yellow. A trajectory line, along with its uncertainty range, is included to highlight a trend.*

by employing a median line, as used in the outlier-removal approach. The median is less susceptible to the influence of a few lines from other clusters. The idea is reinforced by the significant improvements observed when outlier filtering is enabled in the context-aware brushes.

The length brush faced significant limitations owing to the similarity of line lengths in the dataset. Extremely small selections were required to be created, which presented considerable challenges. Precise selection of the desired lines was often difficult to achieve, as minor mouse movements frequently resulted in unintended selections that differed from those intended.

Higher accuracy scores were achieved with the length and parallel brushes for certain participants. A negative correlation between these brush types and accuracy per average bounding box size of selection is demonstrated in Figure 5. For these context-aware brushes, smaller selections with minor line differences were found to outperform larger ones. Accuracy was found to be irrelevant to the areas of the rectangle and angle brushes. Therefore, on average, these context-aware brush sizes are smaller than the regular rectangle brush from which they are derived. This shows that users realised that smaller selections work better for the context-aware brushes. However, even greater accuracy would have been achieved with even smaller brushes. Future performance may be enhanced by the adoption of point-based brush approaches, such as the percentile brush, which uses a smaller area by definition.

A strong desire to combine multiple brushes when making selections was expressed by several participants. Such a feature would render selections more flexible, permitting the choice of the most appropriate tool for each context. The ability to add lines with one brush whilst refining the selection by removing lines with another was requested by other participants. This dual functionality would afford greater control by sketching with an imprecise brush and subsequently refining the selection iteratively with another brush. High confidence was expressed that the addition of this functionality would improve accuracy.

These findings partially validate the hypothesis: a tool incorporating selection tolerance with context-aware brushes and outlier-removal refinement will reduce human errors caused by imprecise manual selection, thereby improving the quality and efficiency of desired selections. It does not hold for context-aware brushing im-

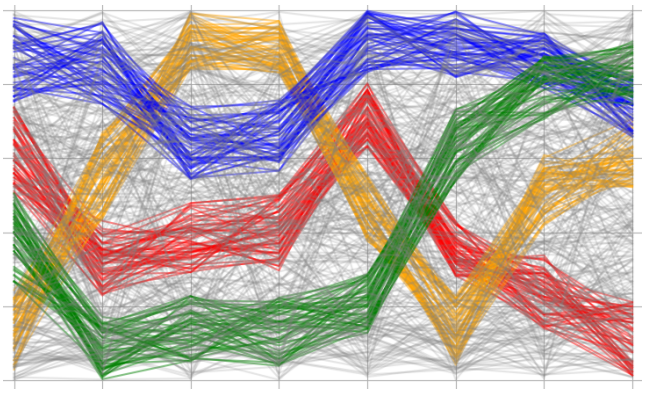

Figure 6: *The 4c.6-300N dataset is an artificial PCP dataset consisting of 4 data clusters, each containing approximately 50 lines. The gray lines represent 300 lines of noise. It was originally created by Blumenschein et al. [BZP*20].*

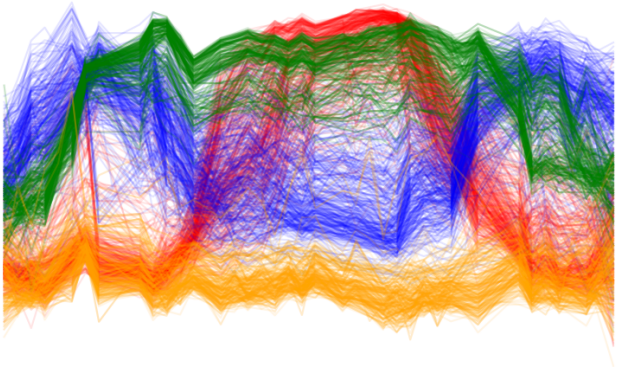

Figure 7: *Time-series with temporal sequence of spectral values that describe how the land surface changes over time created by Tan et al. [TWP]. The red, blue and green lines are each 300 lines, while the yellow lines are 295 lines.*

proving user's selections. However, the outlier-removal refinement was able to reduce user errors efficiently improving the users' accuracy and satisfaction for the draw-line, length and parallel brushes.

In conclusion, context-aware brushes may outperform other brushes for specific tasks or datasets. However, reliable performance across general cases is not achieved. Consequently, further development of context-aware brushes is required to enable effective employment in visual data analysis. In the second study, the refinement techniques are further investigated to determine an optimum that maximises accuracy gain whilst minimising efficiency loss.

6. User study two: Determining the optimal refinement

Based on earlier findings, the refinement methods were prioritised, with reduced emphasis on brush type considerations. The outlier-filter method was demonstrated to significantly improve performance on challenging brushes. A preference for a delete mode was identified. To satisfy this requirement, the semi-automatic refinement technique was compared against a manual refinement method. A drop-down menu was presented for manual refinement, enabling selection of one of the brushes evaluated in Study One. The brushes were ordered from most to least accurate in accordance with the previous study. To evaluate whether the manual or semi-automatic refinement method are optimal in terms of accuracy, efficiency, and user satisfaction, a comparison with a more automated method was conducted in Study Two.

Further investigation of the outlier-filter's properties using real-world, complex datasets was conducted to establish whether performance improvements generalised across a broader range of datasets. To validate this generalisation, two additional datasets were selected to supplement the initial study, each varying substantially in dimensionality, structure, and clustering difficulty.

The first dataset added is the 4C.6-300N dataset, as shown in Figure 6. This PCP plot represents a simulated dataset with a high noise-to-data ratio, in which each cluster contains approximately six times more noise than data. The assumption that outlier filtering remains effective when large amounts of noise are inadvertently selected is examined. The second dataset, the crops dataset, comprises time-series data derived from satellite imagery. A temporal sequence of spectral values describing land-surface change over time is measured by each geographic coordinate. In Figure 7, a preview is shown.

Properties were extracted from each dataset to highlight similarities and differences, as indicated in Table 2. The datasets vary significantly in dimensionality and sample size, ranging from 492×8 to 996×100, thus permitting comprehensive evaluation of method robustness across different data scales.

A critical evaluation dimension is defined through the assessment of envelope sizes and cluster overlap characteristics, which reflect the intrinsic difficulty of each clustering task; both definitions are presented in Section 4.5. Envelope sizes were quantified using an arbitrary function:

$$f(D) = \sum_{c \in D} ES(c)$$

where $ES(c)$ denotes the envelope size of cluster c. By summing these sizes, the percentage of space that could be occupied if the lines are not stacked is indicated. The range extended from values below 0.5 in the artificial datasets (Andrews and 4C.6) to 0.85 in the Crops dataset, indicating considerably larger cluster volumes. Cluster overlap is characterised by substantial variation across datasets, with measurements of 7.75 for Crops, 5.98 for Andrews and 3.09 for 4C.6. A high CO denotes a well-mixed state of lines. Overall, substantially greater difficulty for manual interpretation is expected for the Crops dataset among the three examined.

6.1. Study setup

The second study was conducted with minimal modifications compared to the first study. Initially, consent was obtained and supplementary questions were posed regarding prior participation in the

Name	Type	Shape	Envelope sizes	Cluster Overlap
4C.6	PCP	492×8	0.38	3.09
Andrews	PCP	996×100	0.49	5.98
Crops	Time-Series	996×46	0.85	7.75

Table 2: *A comparative overview of three line datasets: Andrews, 4C.6 and Crops. For each dataset, it specifies its Type (PCP or Time-Series), along with its Shape as $y \times x$ dimensions. Additionally, it quantifies Envelope sizes as the total size of clusters. Total Cluster Overlap signifies how much Gaussian overlap there exist between clusters. Details of this table are presented in Appendix C.*

initial study and other potential biases. Subsequently, the key concepts of the system were detailed in an extensive written tutorial, which outlined each refinement technique and its associated controls. Again, two experimental phases were conducted.

Phase one was designed to assess the extent to which participants could effectively apply the refinement techniques without concern for initial selection creation. Participants were presented with three randomised expert-selections. An expert-led selection represents an initial selection created to exemplify a plausible selection that participants might reasonably make. The expert-selections were established for three primary reasons. First, study duration was minimised, as participants cannot reasonably be expected to invest substantial time in the entire study. Second, uniformity was established among the initial selections to facilitate subsequent comparison. Finally, participant burden was reduced, allowing focus on the refinement stage rather than the initial selection stage. Each participant receives three distinct refinement methods from Section 4.3 in a randomised order.

In the second phase, initial selections were created by participants themselves (self-led) and refined using the same methodology employed in the first phase. The primary objective of the second phase was to examine the extent to which participants could effectively correct selections not initially anticipated.

The study concludes with a questionnaire addressing participant satisfaction regarding the preferred method based on overall confidence and difficulty.

6.2. Risks mitigation

The user study was successfully completed by twenty-eight participants. To ensure the validity and reliability of this user study, several potential biases and confounding factors warrant careful examination. This section addresses three primary sources of risk: memory bias from participant overlap between the first and second studies, the demographic and professional heterogeneity of the participant pool, and variance in the distribution of participants across datasets and refinement methods. By transparently presenting these factors, the extent to which study outcomes may be influenced by such variations can be appropriately assessed.

Memory bias and participant overlap. Participants who had previously participated in the first user study may have possessed an advantage over those experiencing the tool for the first time, as memory bias can increase the performance of the selection procedure. However, only six participants overlapped between the two studies, thereby limiting the magnitude of this potential bias. Although the primary task of both studies is to apply selections, it is considered that the tasks are sufficiently different, particularly with the never-before tested manual and automatic refinement techniques. When comparisons are made between studies, participants from the first user study are to be excluded from the analysis of study two or, at a minimum, marked as special participants.

Participant demographics and professional composition. The study comprised seventeen students and four teachers, whilst the remaining seven participants possessed diverse qualifications. Each participant was either working at, currently studying or had studied at the Delft University of Technology, with the exception of one student affiliated with the University of Amsterdam.

Regarding the fields of study and professional expertise represented within this cohort, thirteen participants specialised in computer science. Seven individuals specialised in mathematics. Four participants were classified as engineers across various specialisations. These three categories of participants were expected to be well equipped with selection tools and may have approached the task with domain-specific advantages. The remaining participants were from diverse fields, comprising three from arts and one from medicine. The heterogeneity of professional backgrounds reflects potential variability in participants' familiarity with data visualisation and selection interfaces, which may influence performance metrics. However, this diversity also enhances the generalisability of findings across different user populations, as results reflect performance across individuals with varying technical expertise.

Distribution of participants across datasets and refinement methods. As written in Section 6.1, datasets and clusters were randomly assigned to each participant. Consequently, variance existed in the distribution of participants across refinement methods. In order to present this variation transparently, the frequency with which each refinement method was selected for each dataset is displayed in Table 3. The self-selection type encompassed manual, semi-automatic, and automatic refinement applied to the same datasets. It should be noted that not all rows sum to twenty-eight. This discrepancy arises from instances in which the ReVISit tool failed to record selections, and instances in which refinement techniques were misclassified by the tool as alternative techniques.

6.3. Brush usage

Different brush usage was de-prioritised to focus solely on refinement techniques. To provide flexibility, multiple brushes could be selected for the self-selection and manual refinement procedures. The number of brushes used provides detailed insight into which brushes demonstrated the highest user satisfaction. Selections can

Type	Andrews	Crops	4C6
Manual	10	8	10
Semi-automatic	9	11	8
Automatic	12	10	4
Self-selection	5	12	10

Table 3: *Number of entries classified by automation level (manual, semi-automatic, automatic, and self) across Archives, Crops, and 4C6 datasets.*

be made during three sections of the study. During phase one, manual refinement can be applied to an expert-led selection. During phase two, the user is instructed to create selections themselves, later on these selections are again refined in the manual refinement stage.

Refinement Type	Rectangle	Angle	Draw	Mean $\pm$ Std
Expert Manual	84	4	0	3.38 ± 1.88
Self Selection	26	4	3	1.38 ± 0.48
Self Manual	63	2	0	2.71 ± 1.81

Table 4: *Counts for three specific interaction types (Rectangle, Angle and Draw) were recorded. These metrics were evaluated across three refinement types: Pre-Manual, Self-Manual and Self-Selection. In addition, the total selections amount per user for each method are presented as mean and standard deviations. A detailed figure of brush-type usage frequency per selection is presented in Appendix D.*

Table 4 presents the brushes included in the final selections. The combination of multiple brushes and sequential selections is permitted by the tool. Manual refinements were not limited by the number of selections, whereas self-selection was limited to two selections. This constraint was imposed to enhance comparability between self-led and expert-led initial selections. Furthermore, the constraint was designed to limit users' initial accuracy scores, thereby ensuring that at least some material for refinement was available to most users after their initial selections. Consequently, manual refinement techniques yielded a greater average number of selections than self-selection.

The default brush type was the rectangle. As few participants modified this setting, the overwhelming majority of refinements were created using the rectangle brush. The angle brush was selected by a smaller number of participants, all of whom were part of the first study. The draw brush in self-selection was utilised by three individual participants, none of whom were part of the first study. Notably, no other brushes were chosen in the final selections of the participants, although some may have been experimented with; they were subsequently removed by these participants.

As the rectangle selection tool was employed by almost all participants. Consequently, statistically meaningful comparisons across brush types cannot be made in this second study. These findings indicate a strong user preference for the rectangle brush across all refinement methods and datasets. Consequently, this consistent application of a single selection method across datasets and participants reduces potential methodological bias that might arise from heterogeneous brush usage, thereby strengthening the validity of comparative analyses regarding refinement method effectiveness. The limited adoption of alternative brushes, despite their availability, indicates that the rectangle brush warrants continued emphasis as the primary interaction mechanism.

6.4. Quality

In Table 5, selection quality is compared across three datasets (Andrews, Crops and 4C.6) and four approaches: manual, semi-automatic, automatic and initial selection as control. Lower envelope sizes and cluster overlaps are closely correlated with higher accuracy scores. The most separated and densely clustered 4C.6 dataset has the highest accuracies, followed closely by the Andrews plot. The Crops dataset exhibits the poorest selection qualities. This discernible trend is already noticeable in the initial selections made, whether by experts or through self-selection methods. Such observations suggest that as the complexity of the datasets increases, users tend to encounter greater difficulties with their initial selections, which may impede their overall accuracy performance.

Examination of refinement techniques indicates that the highest accuracy is consistently attained by manual refinement, followed by semi-automatic refinement. The poorest performance in accuracy is consistently achieved by the automatic refinement approach.

This is supported by Table 6, which shows that the manual refinement has the highest Mean Absolute Change (MAC). This MAC value is the absolute percentage which a participant is able to improve an initial selection using each refinement method. Manual refinement has a MAC of 12,6.% and semi-automatic refinement follows with a MAC improvement of 9,8.%. Notably, a selection can be improved more readily, by at least some amount, using the semi-automatic tool (90,7.%) rather than the manual refinement tool (87,6.%). Although the changes are subtle, it is suggested that the most positive results are yielded by the semi-automatic tool for a greater number of participants.

Validation was undertaken to ascertain whether the effect persists after overlapping participants from study-one are removed from study-two, as a memory bias favouring the semi-automatic technique could be present. After removal, the Success / Total scores of the manual tool decreases to approximately 84,0% whereas the semi-automatic technique remains essentially unchanged at 89,3%. This indicates that the same pattern persists even after the potentially biased participants are removed.

Both MAC and Success/Failed ratio improvements are significantly lower for the automatic refinement tool, with only slightly more than half of the final selections exceeding the initial selection quality. Overall, a notable difference is observed for most participants, with a total average accuracy improvement of 8.4%.

Looking more closely at the results of Table 5. Specifically for automatic refinement, it is demonstrated that self-led selections consistently outperform expert-led selections across the tested datasets. However, the underlying causes of this performance difference remain unclear. Two plausible explanations are proposed to account for this pattern.

Dataset name & type	Phase one (expert initial selection) Accuracy	Precision	Recall	Phase two (self initial selection) Accuracy	Precision	Recall
4C.6 (manual)	0.99 ± 0.04	0.90 ± 0.30	0.89 ± 0.30	0.97 ± 0.05	0.88 ± 0.30	0.80 ± 0.29
4C.6 (semi)	0.98 ± 0.03	0.91 ± 0.17	0.98 ± 0.02	0.93 ± 0.13	0.78 ± 0.27	0.79 ± 0.24
4C.6 (auto)	0.87 ± 0.06	0.21 ± 0.36	0.20 ± 0.35	0.90 ± 0.06	0.48 ± 0.42	0.34 ± 0.31
4C.6 (initial)	0.89 ± 0.03	0.48 ± 0.07	1.00 ± 0.00	0.82 ± 0.16	0.41 ± 0.14	0.91 ± 0.11
Andrews (manual)	0.97 ± 0.03	0.99 ± 0.01	0.88 ± 0.13	0.97 ± 0.03	0.99 ± 0.01	0.87 ± 0.11
Andrews (semi)	0.93 ± 0.06	0.82 ± 0.37	0.67 ± 0.32	0.96 ± 0.02	0.99 ± 0.02	0.85 ± 0.07
Andrews (auto)	0.87 ± 0.09	0.65 ± 0.46	0.40 ± 0.36	0.89 ± 0.09	0.94 ± 0.06	0.59 ± 0.32
Andrews (initial)	0.89 ± 0.02	0.67 ± 0.18	0.88 ± 0.14	0.93 ± 0.04	0.84 ± 0.13	0.95 ± 0.03
Crops (manual)	0.94 ± 0.04	0.94 ± 0.08	0.85 ± 0.08	0.91 ± 0.16	0.89 ± 0.27	0.79 ± 0.27
Crops (semi)	0.93 ± 0.06	0.94 ± 0.03	0.81 ± 0.19	0.92 ± 0.07	0.93 ± 0.08	0.77 ± 0.21
Crops (auto)	0.78 ± 0.16	0.42 ± 0.42	0.38 ± 0.42	0.84 ± 0.12	0.87 ± 0.27	0.49 ± 0.27
Crops (initial)	0.76 ± 0.09	0.51 ± 0.18	0.93 ± 0.03	0.80 ± 0.16	0.68 ± 0.19	0.90 ± 0.13

Table 5: *Statistical means and standard deviations for both phase one and two were calculated for three performance metrics: accuracy, precision, and recall. These metrics were evaluated across three datasets (Andrews, 4C.6, and Crops) and four annotation methodologies (manual, semi-automatic, automatic, and initial selection as control). In all cases, higher is better.*

Accuracy Improvement	Manual	Semi	Auto	Total
Mean Absolute Change	12.6%	9.8%	2.8%	8.4%
Success / Total	87.6%	90.7%	57.4%	78.7%

Table 6: *Accuracy improvements from initial to final selection are presented for manual, semi-automatic, and automatic refinement. Success / Total improvement is defined as the fraction of final selections exhibiting at least 0% improvement compared to initial selections. Details regarding a full distribution of relative improvement for accuracy, precision and recall for each dataset type and expert versus self selections are presented in Appendix E.*

First possible reason: Differences in Expert-Led and User-Led Initial Selections

Expert-led and self-led initial selections could differ fundamentally in their characteristics, thereby influencing the performance of the automatic refinement technique for each. Supporting evidence is provided by the Andrews dataset. Within this dataset, lower accuracy is attained by the expert-led initial selection compared with the self-led initial selection. Notably, lower accuracy is also yielded by the automatic refinement applied to the expert-led selection. This pattern indicates that weaker initial selections produce weaker refined results, irrespective of the refinement method. However, this explanation is not upheld consistently across all datasets. In the 4C.6 dataset, higher accuracy is achieved by the expert-led initial selection than by the self-led initial selection. Nevertheless, following application of automatic refinement, the accuracy of the self-led selection exceeds that of the expert-led selection. This reversal suggests that the automatic refinement technique may be better suited to improving self-led selections than expert-led ones, contradicting the notion that initial selection quality simply propagates through refinement. The Crops dataset presents a more nuanced picture. In this dataset, higher recall is demonstrated by the expert-led initial selection than by the self-led initial selection, yet lower precision is achieved. As recall is generally regarded as more critical than precision during initial selection—because recall establishes the pool of candidates from which precision can be improved during refinement—greater benefit from automatic refinement might be expected for the expert-led approach. Nevertheless, the final automatic selection performs better for self-led selections. This inconsistency indicates that initial selection characteristics alone do not fully explain the observed performance differences.

Method	Expert-led	Self-led
Manual	4.0%	8.3%
Semi	4.0%	0.0%
Auto	39.1%	12.5%

Table 7: *Occurrence of situation that a final selections with 0% precision is reached, indicating that no selected lines matched the goal*

Second possible reason: Memory Bias in Phase Two A second possible explanation is offered by memory bias, whereby participants in phase two may have benefited from experience gained during phase one. The data indicate that accuracies are consistently elevated for automatic refinements in phase two compared to phase one, which may suggest that familiarity with the task enhanced selection quality. Further support for this hypothesis is provided by the observation that precision scores of zero were achieved by a considerable number of users in the automatic refinement condition, indicating that no selected lines matched the goal. This outcome was observed considerably more frequently in automatic refinement than in manual or semi-automatic refinement as indicated by Table 7. For expert-led selections, precision scores of zero were recorded 39.1% of the time, whilst for automatic self-led selections this frequency was only 12.5%. This selective increase in extreme outcomes for automatic refinement in phase one suggests that task-

specific strategies or heightened confidence may have been developed, potentially altering behavioural patterns. This explanation, however, is subject to significant challenges when examined more broadly. If memory bias were the primary driver of improved phase two performance, equivalent improvements would be anticipated across all refinement techniques. Instead, manual refinement techniques demonstrate equal or even higher accuracies in phase one compared to phase two. This pattern contradicts the memory bias hypothesis, as consistent improvements across all methods would be anticipated if experience were the dominant factor.

In summary, while both explanations contribute partially to our understanding of why self-led selections benefit more from automatic refinement, no single explanation fully accounts for the observed patterns across all datasets and refinement methods. The interaction between selection type, refinement technique, and dataset characteristics appears more complex than any individual hypothesis alone can explain.

6.5. Efficiency

As shown in Figure 8, no definitive correlative trend is observed between efficiency and quality results. No consistent increase or decrease is apparent when the three refinement types are examined comparatively. Consequently, the efficiency values must be analysed for each dataset to determine how dataset characteristics influence efficiency outcomes. The usage of an appropriate refinement method for individual use cases is thus informed by this analysis. Because temporal values exhibit considerable variation, all temporal analysis is conducted using the median instead of the mean. The reasons for the large spread are explained in the remainder of this section.

For the 4C.6 dataset, semi-automatic refinement methods were found to be the least time-consuming task (median 35.3 seconds) compared with manual (median 71.4 seconds) and automatic (median 60.7 seconds) refinement methods. Poor average precision (0.48) was exhibited by the initial selections. This indicates that, on average, selected noise exceeded selected lines in the goal. This deficiency was substantially mitigated by the design of the semi-automatic interface. For most users, the large amount of noise situated relatively far from the median of the main cluster was immediately apparent in the slider panel, allowing rapid adjustment of the selection boundaries with minimal deliberation. Manual refinement, by contrast, was found to produce substantially longer completion times, which were largely driven by the observed time variance. It was observed that a participant required over 270 seconds to complete the task. As presented in Figure 9, a total of 17 selection attempts were required before a satisfactory result was achieved by this participant. For this dataset, the average selection count for the manual refinement method was 6.60, significantly higher than that for the Andrews dataset (2.70) and the Crops dataset (3.60). Notably, the accuracy achieved for the manual 4C.6 dataset was the highest among the three methods. This outcome was likely attributable to the relative ease with which clear outliers could be identified. However, the subsequent removal of such outliers proved more difficult, as evidenced by the longer completion times and higher manual selection counts. Automatic refinement demonstrated timings similar to manual refinement, with considerable variance. This variability was most likely caused by the non-deterministic nature of the k-means clustering algorithm. As clustering quality varied between participants, with some receiving well-separated clusters whilst others received poorly separated clusters. Fundamentally, different problem complexities were faced by users despite identical initial selections. Rapid validation and acceptance of the automatic suggestion were possible for users who encountered clean clustering, whereas additional time was invested by users facing ambiguous clusters to toggle cluster numbers and evaluate alternative solutions. This finding underscores a critical insight: automation does not guarantee efficiency when the underlying algorithm produces high-variable quality outputs. This problem will be discussed further in the Limitations section which is in Section 7.

Compared to the 4C.6 dataset, the Andrews plot demonstrates a markedly different efficiency profile. In this instance, median completion times are more closely aligned; however, differences remain observable. The median completion time for the automatic refinement technique (35.2 seconds) was this time the shortest compared to 51.4 seconds for manual selection and 60.5 seconds for the semi-automatic refinement approach. This time performance of the automatic tool is comparable to that observed in the semi-automatic 4C.6 selection. However, quality metrics differed substantially across selection methods. For semi-automatic 4C.6 selection, on average, participants achieved a recall of 0.98. For automatically refined Andrews selection, the recall was considerably lower at 0.40. For the remaining manual and semi-automatic selections, recall values were 0.88 and 0.67 respectively. These results suggest that although quality improvements varied across datasets, participants demonstrated quicker satisfaction with automatic refinements. The similarity in cluster appearances may hinder the identification of remaining lines requiring selection. It is suggested by the temporal data that automatic refinement satisfaction assessments were performed rapidly, being predicated on subjective confidence rather than on objective quality metrics.

The Crops dataset, characterising the most complex challenge (cluster overlap of 7.75, envelope size of 0.85), demonstrates the most instructive efficiency patterns. All three methods required deep engagement with the selection problem and deliberate judgements regarding line inclusion or exclusion. Nearly identical median time metrics were measured by the manual and semi-automatic refinement methods: 51.5 seconds and 51.4 seconds respectively. These time values remain comparable with the more time-consuming selections observed in the other datasets. In contrast, automatic refinement was identified as a substantial outlier (median 122.16 seconds), representing more than double the duration required by the other two approaches. These time metrics prove the opposite of the hypothesis: more automated refinement methods require fewer time to improve an initial selection.

This substantial performance differential warrants closer examination, particularly given that both semi-automatic and automatic refinement methods employ the same median selection line as guiding principle. The superior performance of the semi-automatic method may be attributable to two primary factors. First, the automatic refinement method suffers from a non-deterministic nature, which is already discussed. Second, and more substantially,

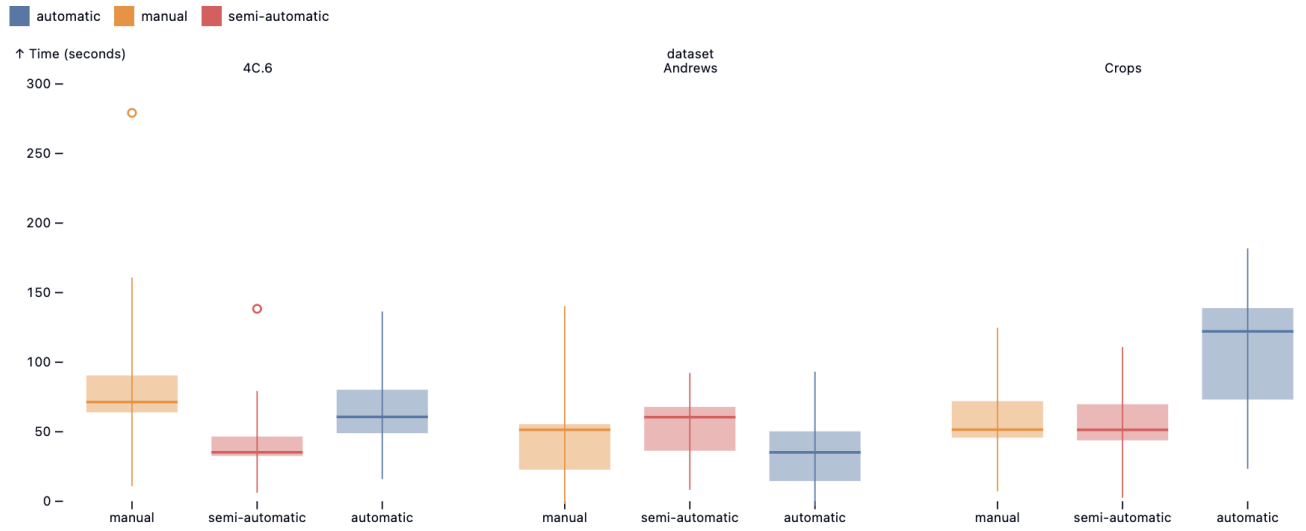

Figure 8: *The time in seconds required to complete the selection goal is presented in a box plot for phase one. The metrics were evaluated across three datasets (Andrews, Crops and 4C.6) and three refinement methodologies (manual, semi-automatic, automatic). In all cases, lower values indicate better performance. Further details concerning medians and interquartile ranges are provided in Appendix F.*

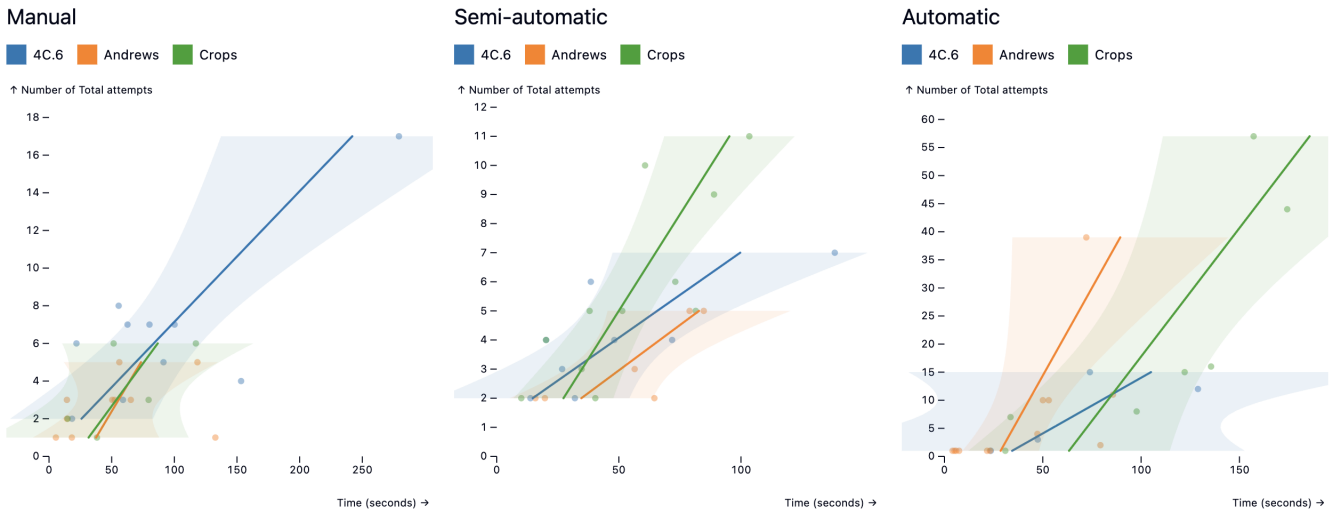

Figure 9: *Time taken for the selection process in seconds versus attempts, grouped by refinement technique. Individual values are depicted by dots, whilst a general trend in the data values is indicated by the line with certainty range. Further details concerning correlation values are provided in Appendix F*

the semi-automatic approach affords enhanced fine-grained control over filtering of the main branch of cluster lines, allowing users to make incremental adjustments without triggering cascading algorithmic re-computation.

The operational complexity of automatic refinement is evident in cluster-size adjustment scenarios. As data become more complex, the likelihood that a desired number of lines can be obtained by selecting from three clusters diminishes. Consequently, an optimal number k of clusters must be identified, a task that proves considerably difficult. Substantial time is expended in experimenting with the cluster-number parameter and in enabling or disabling specific clusters, with at least half of the users requiring more than fifteen attempts in numerous instances for the automatic crops dataset. The complex iterative penalty is avoided by the semi-automatic refinement. When configured with too few bins, some overshoot-

ing or undershooting may be caused, but the effect is considerably less than that observed with automatic refinement. More efficient user-directed refinement is enabled, demonstrating the practicality of interactive control for managing large-scale line-data selections.

The lowest variance in completion times for the semi-automatic method is likely attributable to the limited number of available choices. In manual refinement, selections may be created through multiple brush applications that can be placed on an infinite amount of possible location on the plot. As previously discussed, the automatic method presents more choices relative to the semi-automatic tool. Consequently, when more complex datasets are introduced, completion times for the semi-automatic refinement method are expected to be the closest observed in this study.

Due to limitations in the ReVISit programme, recording the time values for the second phase was made impossible for the second phase. Although some memory bias may have occurred, other efficiency data do not deviate substantially from the results obtained in the two phases. Consequently, it can be assured that trends in time values in phase two are comparable to those obtained in the first phase. The limitations section (Section 7) discusses why recording these metrics proved difficult and how the ReVISit programme can be improved to mitigate this issue.

6.6. Satisfaction

Dataset	Difficulty	Confidence
4C.6 (manual)	2.00 ± 1.26	4.70 ± 1.27
4C.6 (semi)	1.13 ± 0.78	4.38 ± 1.32
4C.6 (auto)	1.00 ± 0.71	4.50 ± 0.50
4C.6 (total)	1.50 ± 1.12	4.55 ± 1.20
Andrews (manual)	1.70 ± 1.27	4.80 ± 1.54
Andrews (semi)	1.17 ± 1.07	5.00 ± 0.82
Andrews (auto)	0.92 ± 0.95	4.17 ± 1.34
Andrews (total)	1.25 ± 1.15	4.57 ± 1.37
Crops (manual)	2.00 ± 0.89	2.60 ± 1.20
Crops (semi)	2.27 ± 1.81	3.91 ± 1.62
Crops (auto)	3.29 ± 1.67	3.29 ± 1.28
Crops (total)	2.52 ± 1.69	3.43 ± 1.53

Table 8: *Difficulty measured from "Trivial" (0) to "Feels impossible" (6), and confidence measured from "Not confident at all" (0) to "Extremely confident". These metrics were evaluated for phase one across three datasets (4C.6, Andrews, Crops) and three refinement methodologies (manual, semi-automatic, automatic).*

Beyond objective time measurements, self-reported difficulty and confident scores provide insight into the user satisfaction. Phase one scores are presented in Table 8. The artificial datasets demonstrate remarkable similarity in overall difficulty perception. Comparable total difficulty scores were obtained for the 4C.6 dataset (1.50±1.12) and the Andrews dataset (1.25±1.15), indicating that the datasets were perceived as similarly manageable. In contrast, the real-world Crops dataset (2.52±1.69) was found to be substantially more difficult, although the average rating remains better than a neutral score of 3, indicating that the task remained manageable. This elevated difficulty aligns with the dataset characteristics presented in Table 2: the Crops dataset exhibited the highest cluster overlap (7.75) and largest envelope sizes (0.85), creating a more visually complex selection environment that naturally demands greater cognitive effort.

Participants rated automatic refinement as substantially more difficult for the Crops dataset (3.29 ± 1.67) compared to manual (2.00 ± 0.89) and semi-automatic (2.27 ± 1.81) methods. This subjective experience correlates with the temporal and quality findings: when algorithmic performance degrades, users experience cognitive strain. Conversely, for the 4C.6 dataset, automatic methods received the lowest difficulty ratings (1.00 ± 0.71), reflecting the alignment between algorithmic output and user expectations when the task itself is relatively straightforward. This pattern indicates that user's experience with the refinement methods is highly dependent on the dataset itself.

Confidence ratings reveal a more subtle landscape without clear directional trends. High and stable confidence was maintained across the artificial datasets for refinement methods: average confidence of 4.55 ± 1.20 for 4C.6 and 4.57 ± 1.37 for Andrews. Confidence remained stable as refinement approaches were transitioned between, suggesting that the consistency of algorithmic output produce trust regardless of automation level. The Crops dataset demonstrated lower absolute confidence (3.43 ± 1.53), exhibiting a slight positive trend from manual (2.60 ± 1.20) through semi-automatic (3.91 ± 1.62) to automatic (3.29 ± 1.28). This non-monotonic pattern is particularly illuminating: confidence peaks with semi-automatic refinement before declining with automation. This suggests that whilst the semi-automatic approach their interpretable visualisation marginally improved confidence, the unreliability of automatic refinement was recognised, leading to appropriately diminished confidence in those recommendations.

When asked verbally, users are greatly satisfied with the tool. This satisfaction is attributed to substantial improvements in selection quality achieved through the refinement techniques employed. Because no single refinement technique demonstrated clear superiority in time-efficiency, participant preferences were not definitively concentrated on any single technique.

This is also illustrated in phase two, when users are asked to rate which refinement method is least difficult and most confident in Table 9. Semi-automatic refinement was selected as the most confident by 12 participants, followed by manual (10 participants), with automatic receiving only 2 votes, a preference distribution that mirrors the actual quality performance (manual: 12.6%, semi-automatic: 9.8%, automatic: 2.8%). Conversely, automatic refinement was overwhelmingly selected as "least difficult" by 13 participants, substantially outpacing manual (6 participants) and semi-automatic (5 participants) approaches. This apparent paradox indicates that the unreliability of automatic refinement was correctly recognised by participants, despite its perception as mechanically simpler. The superficial simplicity of interface interactions and parameter adjustment obscures the underlying algorithmic limitations; when algorithmic outcomes fail to meet expectations, such surface-level simplicity does not engender user confidence.

Metric	Manual	Semi-automatic	Automatic
Least difficult	6	5	13
Most confident	10	12	2

Table 9: *Preferred refinement method based on highest confidence during each selection method and least perceived difficulty for phase two.*

7. Limitations and future work

This paper explores multiple brushing techniques and refinement methods to address the challenges of interactive visual manipulation of large-scale line data. Although findings provide valuable insights into the effectiveness of these approaches, several limitations should be acknowledged, and numerous promising directions for future investigation remain.

The primary limitation in the first explorative study stems from the controlled nature of the study design. The context-aware brushes, specifically the parallel brush, which theoretically outperforms the outlier-removal method when clusters contain internal gaps, were never evaluated under conditions representative of their intended use cases. Participants were not explicitly tasked with selecting clusters exhibiting this property, meaning the theoretical advantages of these brushes could not be fully validated. Future research should deliberately design evaluation scenarios that test context-aware brushes under their optimal operating conditions.

For the second study, a notable technical limitation concerns the incomplete timing data. The ReVISit platform (version 2.0) permits researchers to track responses for embedded views. However, each embedded view can only propagate a single string to the parent programme. Consequently, the initial selection data and all three refinement techniques for the second study were combined into a single JSON response field. This architectural constraint permitted only the determination of when the initial selection phase began and when the final refinement ended. Critical intermediate time points, such as when each individual refinement technique started or concluded, were not recorded. Therefore, the temporal data for the second study phase were excluded from the analysis, preventing a comprehensive comparison of efficiency across both experimental phases.

The selection tools employed in this research were originally developed using the Svelte framework. Because ReVISit officially supports only HTML and React for embedded views, an HTML wrapper was constructed to intermediate message passing between the survey platform, the HTML wrapper, and the Svelte application. This non-standard integration approach likely resulted in some events not being properly recorded by ReVISit, potentially introducing unmeasured data loss. Direct implementation of selection tools in an officially supported language would have mitigated these integration issues and ensured more reliable event tracking. Beyond these project-specific concerns, multiple implementation challenges were encountered related to ReVISit's form caching mechanisms and data storage limitations when handling large provenance graphs. These issues are comprehensively documented in Appendix G and merit attention from the ReVISit development community.

The automatic refinement technique presented a distinct limitation rooted in the inherent non-determinism of the k-means clustering algorithm. Although all participants began with identical initial selections, most received unique refinement problems when applying the automatic refinement method. This variability in algorithmic output introduced confounding factors that complicated the interpretation of results. Alternative automated approaches should be explored, such as Convolutional Neural Network (CNN)-based brushes developed for scatter plots [FH21], which could offer improved performance through learned feature representations. A deterministic approach grounded directly in learned data representations could eliminate this source of variance. In addition, reliance on the median line guidance should be reduced. Thereby, better addressing over-selection problems and moving beyond heuristic-based median line adjustments. The current implementation of refinement techniques is fundamentally limited to filtering, that is, removing lines from an initial selection. It does not support adding previously unselected lines, constraining its applicability for correcting under-selection errors. Early exploratory work demonstrated that augmenting the semi-automatic refinement method with a histogram of all lines in the dataset could theoretically enable line addition. However, this approach introduced substantial visual clutter and revealed many lines lying far from the primary cluster. Resolution would require either an algorithmic preprocessing step to remove distant outliers based on an objective cut-off criterion, or provision of direct user control over this parameter. Neither approach was evaluated in the final study to maintain experimental consistency, but both warrant investigation in future work.

Beyond addressing the aforementioned limitations, several promising avenues merit investigation:

1. **Study cluster-specific characteristics.** With twenty-eight participants and twelve unique cluster goals (three data sets, each containing four clusters), each cluster was tested an average of 2.33 times. This limited sample size results in trend analyses for most clusters being based on only two to three data points, thereby restricting the statistical power of the conclusions. Future work should replicate the study with substantially larger participant cohorts to establish more robust findings regarding cluster-specific performance characteristics.
2. **Iterative Refinement Workflows.** When applying these refinement techniques to complex real-world tasks, users would benefit from an iterative workflow enabling them to add lines through selection, subsequently refine those selections, and repeat this cycle. Future research should systematically investigate the optimal balance between accuracy gains and efficiency losses across successive refinement iterations.
3. Study objective Doing the study in as little time was not the primary objective of the study. Participants were encouraged to attain the highest possible accuracy rather than to complete the process rapidly. A subsequent study could be conducted to assess how rapidly participants can achieve a predetermined score within minimal time.
4. **Extended Application Domains.** The current evaluation was focused exclusively on cluster selection tasks. However, diverse

analytical purposes are served by brushing and refinement techniques—identifying outliers, locating features within specific x-value ranges, and linking selections across linked views. Rich opportunities for future investigation are presented by these alternative use cases, which would substantially broaden the applicability of the findings.

5. **Complex Dataset Evaluation.** It is demonstrated that, even on the most challenging dataset evaluated (the Crops time-series data), mean accuracy levels of approximately 80% were achieved by participants—a surprisingly high performance level. This suggests that the methods may generalise well to complex data. Future work is required to validate these approaches on substantially more challenging datasets, such as painting reflectance data [PGK*22] or other specialised domain datasets. Crucially, these evaluations must involve domain experts who are capable of assessing whether the context-aware brushes and refinement techniques genuinely improve analytical workflows and selection quality, moving beyond controlled experimental metrics to real-world utility.

8. Conclusion

This thesis has addressed the challenge of developing selection techniques for large-scale line data that simultaneously achieve efficiency, accuracy, and human interpretability. Context-aware brushes require substantial improvement before practical deployment, potentially through machine learning approaches that learn contextual features from data rather than hand-crafted definitions.

For the three refinement techniques presented by this paper, the highest absolute accuracy improvements (12,6 percent) were achieved by manual refinement, followed by semi-automatic (9,8 percent) and automatic (2,8 percent). Consequently, manual and semi-automatic refinements are preferred by users seeking high-accuracy improvements. Although similar efficiency scores are exhibited by manual and semi-automatic refinements, the lowest variance is observed for the semi-automatic method; consequently, it is recommended for users prioritising efficiency. The semi-automatic method is also recommended as the most confident method, scoring especially higher than the others the more complex a dataset becomes. The automatic refinement method is rated favourably only in terms of ease of use, whereas its accuracy and overall efficiency remain considerably inferior. Further development of the automatic refinement tool is required before deployment in real-world applications.

The findings support a fundamental design principle: interpretability and user agency should be prioritised over the maximisation of automation. Rather than pursuing fully automatic approaches. Effective systems should combine moderate algorithmic guidance with transparent feedback mechanisms that enable users to understand and control outcomes. Such mechanisms provide a foundation for systems achieving the simultaneous goals of accuracy, efficiency, and human interpretability. As datasets continue to grow in scale and complexity, principled approaches to interaction design become increasingly essential.

References

[And72] ANDREWS D. F.: Plots of high-dimensional data. *Biometrics 28*, 1 (1972), 125–136. Accessed 13 Oct. 2025. doi:10.2307/2528964. 7

[BBK*18] BEHRISCH M., BLUMENSCHEIN M., KIM N. W., SHAO L., EL-ASSADY M., FUCHS J., SEEBACHER D., DIEHL A., BRANDES U., PFISTER H., SCHRECK T., WEISKOPF D., KEIM D. A.: Quality metrics for information visualization. *Computer Graphics Forum 37* (2018), 625–662. URL: https://doi.org/10.1111/cgf.13446, doi:10.1111/cgf.13446. 4

[BC87] BECKER R. A., CLEVELAND W. S.: Brushing scatterplots. *Technometrics 29*, 2 (1987), 127–142. doi:10.2307/1269768. 3

[BWK11] BATISTA G. E., WANG X., KEOGH E. J.: *A Complexity-Invariant Distance Measure for Time Series*. Society for Industrial and Applied Mathematics, 2011, pp. 699–710. doi:10.1137/1.9781611972818.60. 6

[BZP*20] BLUMENSCHEIN M., ZHANG X., POMERENKE D., KEIM D. A., FUCHS J.: Evaluating reordering strategies for cluster identification in parallel coordinates. *Computer Graphics Forum 39*, 3 (2020), 537–549. doi:10.1111/cgf.14000. 10

[CWS*26] CUTLER Z., WILBURN J., SHRESTHA H., DING Y., BOLLEN B., NADIB K. A., HE T., MCNUTT A., HARRISON L., LEX A.: Revisit 2: A full experiment life cycle user study framework. *IEEE Transactions on Visualization and Computer Graphics (VIS) 32* (jan 2026). doi:10.48550/arXiv.2508.03876. 8

[Dig13] DIGGLE P.: *Analysis of Longitudinal Data*. Oxford University Press, 2013. 4

[DJ98] DONOHO D. L., JOHNSTONE I. M.: Minimax estimation via wavelet shrinkage. *The Annals of Statistics 26*, 3 (June 1998), 879–921. Ann. Statist. 26(3). doi:10.1214/aos/1024691081. 3

[ED07] ELLIS G., DIX A.: A taxonomy of clutter reduction for information visualisation. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (2007), 1216–1223. doi:10.1109/TVCG.2007.70535. 4

[FH21] FAN C., HAUSER H.: On sketch-based selections from scatterplots using kde, compared to mahalanobis and cnn brushing. *IEEE Computer Graphics and Applications 41*, 5 (2021), 67–78. doi:10.1109/MCG.2021.3097889. 6, 17

[Her89] HERSHBERGER J.: Finding the upper envelope of n line segments in o(n log n) time. *Information Processing Letters 33*, 4 (1989), 169–174. doi:10.1016/0020-0190(89)90136-1. 4

[HLD02] HAUSER H., LEDERMANN F., DOLEISCH H.: Angular brushing of extended parallel coordinates. In *IEEE Symposium on Information Visualization, 2002. INFOVIS 2002.* (2002), pp. 127–130. doi:10.1109/INFVIS.2002.1173157. 5

[HS02] HOCHHEISER H., SHNEIDERMAN B.: A dynamic query interface for finding patterns in time series data. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems* (New York, NY, USA, 2002), CHI EA '02, Association for Computing Machinery, p. 522–523. URL: https://doi.org/10.1145/506443.506460, doi:10.1145/506443.506460. 4

[HW13] HEINRICH J., WEISKOPF D.: State of the Art of Parallel Coordinates. In *Eurographics 2013 - State of the Art Reports* (2013), Sbert M., Szirmay-Kalos L., (Eds.), The Eurographics Association. doi:10.2312/conf/EG2013/stars/095-116. 4

[Ins85] INSELBERG A.: The plane with parallel coordinates. *The Visual Computer 1*, 2 (aug. 1985), 69–91. doi:10.1007/bf01898350. 3

[JJO11] JEONG Y.-S., JEONG M. K., OMITAOMU O. A.: Weighted dynamic time warping for time series classification. *Pattern Recognition 44*, 9 (9 2011), 2231–2240. doi:10.1016/j.patcog.2010.09.022. 4

[JSK11] JOHNSON C. R., SETHARES W. A., KLEIN A. G.: *Software receiver design: Build your own digital communications system in five easy steps*. Cambridge University Press, 2011. 4

[LH11] LAMPE O. D., HAUSER H.: Curve density estimates. *Computer Graphics Forum 30* (2011), 633–642. doi:10.1111/j.1467-8659.2011.01912.x. 4

[LPC*24] LIU P., PAN Y., CHANG H.-C., WANG W., FANG Y., XUE X., ZOU J., TOOTHAKER J. M., OLALOYE O., SANTIAGO E. G., MCCOURT B., MITSIALIS V., PRESICCE P., KALLAPUR S. G., SNAPPER S. B., LIU J.-J., TSENG G. C., KONNIKOVA L., LIU S.: Comprehensive evaluation and practical guideline of gating methods for high-dimensional cytometry data: manual gating, unsupervised clustering, and auto-gating. *Briefings in Bioinformatics 26*, 1 (12 2024), bbae633. doi:10.1093/bib/bbae633. 4

[MF18] MORITZ D., FISHER D.: Visualizing a million time series with the density line chart, 2018. doi:10.48550/arXiv.1808.06019. 4

[MM10] MAITRA R., MELNYKOV V.: Simulating data to study performance of finite mixture modeling and clustering algorithms. *Journal of Computational and Graphical Statistics 19*, 2 (2010), 354–376. doi:10.1198/jcgs.2009.08054. 7

[MS25] MIKROU I., SAPIDIS N. S.: A systematic evaluation of clustering algorithms against expert-derived clustering. *Oper. Res. Forum 6* (2025), 55. doi:10.1007/s43069-025-00453-w. 3, 4

[MW95] MARTIN A., WARD M.: High dimensional brushing for interactive exploration of multivariate data. In *Proceedings Visualization '95* (1995), pp. 271–. doi:10.1109/VISUAL.1995.485139. 4

[PGK*22] POPA A., GABRIELI F., KROES T., KREKELER A., ALFELD M., LELIEVELDT B., EISEMANN E., HÖLLT T.: Visual Analysis of RIS Data for Endmember Selection. In *Eurographics Workshop on Graphics and Cultural Heritage* (2022), Ponchio F., Pintus R., (Eds.), The Eurographics Association. doi:10.2312/gch.20221233. 3, 18

[RAM05] RODRÍGUEZ J. J., ALONSO C. J., MAESTRO J. A.: Support vector machines of interval-based features for time series classification. *Knowledge-Based Systems 18*, 4 (2005), 171–178. AI-2004, Cambridge, England, 13th-15th December 2004. doi:10.1016/j.knosys.2004.10.007. 4

[REB*16] RAIDOU R. G., EISEMANN M., BREEUWER M., EISEMANN E., VILANOVA A.: Orientation-enhanced parallel coordinate plots. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2016), 589–598. doi:10.1109/TVCG.2015.2467872. 3, 4

[RLS*19] ROBERTS R. C., LARAMEE R. S., SMITH G. A., BROOKES P., D'CRUZE T.: Smart brushing for parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics 25*, 3 (2019), 1575–1590. doi:10.1109/TVCG.2018.2808969. 3, 4, 5

[RSM*16] RADOŠ S., SPLECHTNA R., MATKOVIĆ K., ĐURAS M., GRÖLLER E., HAUSER H.: Towards quantitative visual analytics with structured brushing and linked statistics. *Computer Graphics Forum 35* (2016), 251–260. doi:10.1111/cgf.12901. 5

[SGMS21] SAHANN R., GAJIC I., MOELLER T., SCHMIDT J.: Selective Angular Brushing of Parallel Coordinate Plots. In *EuroVis 2021 - Short Papers* (2021), Agus M., Garth C., Kerren A., (Eds.), The Eurographics Association. doi:10.2312/evs.20211064. 5

[TB21] TRAUTNER T., BRUCKNER S.: Line weaver: Importance-driven order enhanced rendering of dense line charts. *Computer Graphics Forum 40* (2021), 399–410. doi:10.1111/cgf.14316. 4, 7

[TWP] TAN C. W., WEBB G. I., PETITJEAN F.: *Indexing and classifying gigabytes of time series under time warping*. pp. 282–290. doi:10.1137/1.9781611974973.32. 10

[UZS*18] USACHEV A. D., ZOBNIN A. V., SHONENKOV A. V., LIPAEV A. M., MOLOTKOV V. I., PETROV O. F., FORTOV V. E., PUSTYL'NIK M. Y., FINK M. A., THOMA M. A., THOMAS H. M., PADALKA G. I.: Influence of dust particles on the neon spectral line intensities at the uniform positive column of dc discharge at the space apparatus "plasma kristall-4". *Journal of Physics: Conference Series 946*, 1 (jan 2018), 012143. doi:10.1088/1742-6596/946/1/012143. 5

[War94] WARD M. O.: Xmdvtool: integrating multiple methods for visualizing multivariate data. In *Proceedings of the Conference on Visualization '94* (Washington, DC, USA, 1994), VIS '94, IEEE Computer Society Press, p. 326–333. doi:10.1109/VISUAL.1994.346302. 4

Appendix A: Questions to examine biases

Abel master thesis

Help

What is your email? *

We will use this email to contact you in case of any issues with the study. Or click on 'Don't know' if you do not want to provide your email.

Please enter your email

- [ ] I don't know

Would you like to receive updates (up to 5 emails) about the final results, potential publications, further studies, and presentations related to this study? *

- ( ) Yes
- ( ) No
- ( ) Other

What is your profession? *

This information helps us understand the background and mitigate potential biases among the participants.

- ( ) Teacher
- ( ) Student
- ( ) Professor
- ( ) Other

Which field of work best describes your profession? *

This information helps us understand the background and mitigate potential biases among the participants.

- ( ) Coputer Sciences
- ( ) Mathematics
- ( ) Engineering
- ( ) Ecomics
- ( ) Social Sciences
- ( ) Medicine
- ( ) Other

Do you give consent for audio recordings? *

Abel master thesis

Yes

No

Do you give consent for screen recordings? *

Screen recordings will be used to enhance the study and identify areas where participants faced difficulties. If the study is conducted on a computer other than the researcher's, there is a risk that the screen recording may inadvertently capture sensitive information. The recordings will be deleted after the study concludes. The recordings will not be shared with any third parties.

Yes

No

Do you give consent for local data collection? *

Data will be collected locally and not sent to any third parties. The data will contain mouse clicks, movements and timings. This data will be aggregated and anonimised in the final report. Local data will be removed after the study concludes (aproximately in 3-6 months)

Yes

Do you give consent for anonimised data publishment? *

Anonimised data may be published in the future to help other researchers. The data will not contain any personal information only selection information will be published.

Yes

No

Next

Appendix B: Results study one

	No outlier filtering			With outlier filtering		
Brush	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Rectangle	0.96 ± 0.04	0.98 ± 0.04	0.86 ± 0.16	0.93 ± 0.10	0.92 ± 0.18	0.84 ± 0.15
Angle	0.95 ± 0.02	0.91 ± 0.10	0.90 ± 0.06	0.95 ± 0.04	0.93 ± 0.09	0.87 ± 0.11
Percentile	0.96 ± 0.02	0.94 ± 0.05	0.87 ± 0.10	0.94 ± 0.04	0.99 ± 0.02	0.81 ± 0.12
Draw line	0.80 ± 0.24	0.72 ± 0.24	0.73 ± 0.31	0.93 ± 0.04	**0.98 ± 0.02**	0.77 ± 0.16
Length	0.74 ± 0.11	0.42 ± 0.11	0.57 ± 0.21	0.82 ± 0.13	**0.81 ± 0.22**	0.46 ± 0.26
Parallel	0.70 ± 0.16	0.37 ± 0.25	0.66 ± 0.40	**0.91 ± 0.10**	**0.87 ± 0.15**	0.81 ± 0.27

Table 10: *Selection Quality – Statistical mean and standard deviation measurements for each brush were compared between two conditions: selections without outlier filtering and with outlier filtering. Performance was evaluated across three metrics: accuracy, recall and precision. Thickened numbers in the results indicate statistically significant improvements when outlier filtering was applied. This analysis employed a one-sided Wilcoxon signed-rank test with a significance level of $p < 0{,}05$. The null hypothesis (H_0) assumed that performance without outlier filtering would be less than or equal to performance with outlier filtering.*

	No outlier filtering		With outlier filtering	
Brush	Time (min)	Deletions	Time (min)	Deletions
Rectangle	1.78 ± 1.02	2.29 ± 2.66	1.28 ± 0.82	1.14 ± 1.12
Angle	1.64 ± 1.17	1.57 ± 2.44	1.49 ± 0.74	0.71 ± 1.03
Percentile	2.32 ± 1.97	3.71 ± 5.52	1.49 ± 0.71	0.57 ± 0.73
Draw line	2.65 ± 1.03	4.86 ± 4.61	1.50 ± 1.07	**0.71 ± 0.70**
Length	2.46 ± 1.21	7.71 ± 8.03	1.42 ± 0.65	**1.00 ± 0.76**
Parallel	2.88 ± 1.73	6.00 ± 3.89	1.24 ± 0.62	**0.57 ± 0.73**

Table 11: *Selection efficiency – Statistical means and standard deviations were calculated for the time taken to make a selection in minutes and the average number of deletions for each brush. In all cases, lower values indicate better performance, measured across both outlier filtering being disabled versus enabled. Thickened numbers in the results indicate statistically significant improvements when outlier filtering was applied. This analysis employed a one-sided Wilcoxon signed-rank test with a significance level of $p < 0{,}05$. The null hypothesis (H_0) assumed that performance without outlier filtering would be greater than or equal to performance with outlier filtering.*

	No outlier filtering		With outlier filtering	
Brush	Confidence	Difficulty	Confidence	Difficulty
Rectangle	4.43 ± 1.18	2.14 ± 1.46	4.57 ± 1.40	1.71 ± 1.16
Angle	4.00 ± 0.93	2.29 ± 1.46	4.71 ± 0.70	1.86 ± 0.99
Percentile	3.57 ± 1.40	2.86 ± 1.46	4.29 ± 0.88	2.29 ± 0.88
Draw-line	2.86 ± 1.46	3.57 ± 1.50	4.43 ± 1.29	2.14 ± 1.46
Horizontal	1.57 ± 1.76	4.29 ± 1.03	3.14 ± 1.88	3.14 ± 1.36
Length	0.57 ± 0.73	5.29 ± 0.70	3.14 ± 1.88	4.29 ± 0.45
Parallel	0.57 ± 0.73	4.86 ± 0.64	4.14 ± 1.73	2.29 ± 1.16

Table 12: *Selection satisfaction – Statistical means and standard deviations were calculated for difficulty measured from "Trivial" (0) to "Feels impossible" (6), and confidence measured from "Not confident at all" (0) to "Extremely confident" (6). Measured across both outlier filtering being disabled versus enabled.*

Appendix C: Cluster Envelope and Overlap detailed results

Name	Cluster A	Cluster B	Cluster C	Cluster D
4C.6	0.19 ± 0.02	0.19 ± 0.03	0.19 ± 0.03	0.19 ± 0.03
Andrews	0.08 ± 0.01	0.42 ± 0.05	0.16 ± 0.03	0.32 ± 0.04
Crops	0.54 ± 0.06	0.51 ± 0.07	0.37 ± 0.09	0.29 ± 0.07

Table 13: *The Envelope sizes for four datasets (Andrews, Crops and 4C.6) each having four clusters, denoted as C_a, C_b, C_c, and C_d. The values are presented as the mean and standard deviation of the envelope size. The colour coding highlights different magnitudes: red indicates large sizes, orange indicates medium sizes, and blue indicates small sizes.*

	1	2	3	5
1	–	1.24	1.46	0.43
2	1.24	–	1.18	1.08
3	1.46	1.18	–	0.59
5	0.43	1.08	0.59	–

Table 14: *Pair-wise Bayes Cluster Overlap percentages for the Andrews dataset among clusters C1, C2, C3, and C5 are presented*

	0	1	2	3
0	–	0.21	0.41	0.40
1	0.21	–	0.43	0.19
2	0.41	0.43	–	1.45
3	0.40	0.19	1.45	–

Table 15: *Pair-wise Bayes Cluster Overlap percentages for the 4C.6 dataset among clusters C0, C1, C2, and C3 are presented*

	10	12	21	22
10	–	1.30	1.82	2.17
12	1.30	–	2.04	0.26
21	1.82	2.04	–	0.16
22	2.17	0.26	0.16	–

Table 16: *Pair-wise Bayes Cluster Overlap percentages for the Crops dataset among clusters C12, C10, C22, and C21 are presented*

Appendix D: Detailed results brush usage in study two

Figure 10: *Counts for three specific interaction types (Rectangle, Angle and Draw) were recorded. These metrics were evaluated across three refinement types: Expert-select Manual, Self-select Manual and Self-select Selection. Additionally, a selection index is provided (top), indicating indices the brush types originates from. For example, a selection index value of 1 represents the first selection made, etc.*

Appendix E: Detailed quality study two

Dataset	Manual	Semi	Auto	Total
expert 4C.6	87.50%	100.00%	16.67%	79.41%
self 4C.6	100.00%	100.00%	70.00%	90.00%
expert Andrews	100.00%	70.00%	33.33%	66.00%
self Andrews	75.00%	75.00%	75.00%	75.00%
expert Crops	100.00%	94.12%	54.55%	83.78%
self Crops	91.67%	91.67%	41.67%	75.00%
Total	87.63%	90.72%	57.39%	78.69%

Table 17: *Amount of users who were able to score a better accuracy from the initial selection to the final selection. These metrics were evaluated across three datasets (4C.6, Andrews, Crops) and three refinement methodologies (manual, semi-automatic, automatic). Higher percentage is better.*

Dataset	Manual	Semi	Auto	Total
expert 4C.6	9.21%	8.47%	-2.07%	6.96%
self 4C.6	15.36%	11.00%	8.93%	11.84%
expert Andrews	7.57%	2.21%	-4.55%	1.41%
self Andrews	3.42%	2.91%	-4.12%	0.74%
expert Crops	19.54%	15.83%	1.66%	12.52%
self Crops	10.51%	11.76%	3.80%	8.89%
Total	12.62%	9.76%	2.81%	8.39%

Table 18: *Mean absolute change from initial selection to final selection. These metrics were evaluated across three datasets (4C.6, Andrews, Crops) and three refinement methodologies (manual, semi-automatic, automatic). Higher percentage is better.*

Figure 11: *Bar chart of accuracy, precision and recall improvements for manual, semi-automatic and automatic refinement methods. Absolute differences are measured between the initial and final selections. Bars are stacked.*

Appendix F: Detailed efficiency study two

Dataset	Time (seconds)
4C.6 (manual)	71.39 ± [56.44, 97.98]
4C.6 (semi)	35.26 ± [25.14, 54.07]
4C.6 (auto)	607.0 ± [414.7, 877.0]
Andrews (manual)	51.43 ± [15.28, 62.95]
Andrews (semi)	60.52 ± [28.90, 75.36]
Andrews (auto)	35.20 ± [20.78, 57.81]
Crops (manual)	51.53 ± [38.26, 79.48]
Crops (semi)	51.40 ± [36.38, 77.30]
Crops (auto)	122.16 ± [65.66, 146.40]

Table 19: *The statistical median, together with the first and third quartiles for phase-one, time taken to reach the selection goal in seconds. These metrics were evaluated across three datasets (4C.6, Andrews, Crops) and three refinement methodologies (manual, semi-automatic, automatic). Lower values indicate better performance.*

Method	Manual	Semi-automatic	Automatic
4C.6	0.78	0.74	0.76
Andrews	0.31	0.80	0.57
Crops	0.64	0.79	0.80

Table 20: *Correlation values between time taken to reach the selection goal and amount of attempts. The correlation value is quantified by the following mathematical formula:*

$$corr_{xy} = \frac{\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y-\bar{y})^2}}$$

Dataset	Manual	Semi	Auto	Total
4C.6	6.60 ± 3.93	4.00 ± 1.66	7.75 ± 5.89	5.86 ± 4.06
Andrews	2.70 ± 1.42	3.17 ± 1.34	6.83 ± 10.44	4.57 ± 7.19
Crops	3.60 ± 2.06	5.64 ± 2.96	21.14 ± 19.46	9.91 ± 13.27
Total	4.44 ± 3.32	4.52 ± 2.50	11.35 ± 14.84	6.64 ± 9.25

Table 21: *Statistical mean and standard deviations for phase one amount of attempts to reach the final selection final selections manual, semi-automatic and automatic refinements. These metrics were evaluated across three datasets (4C.6, Andrews and Crops)*

Appendix G: Detailed ReVISit limitations

Beyond the ReVISit programme's inability to directly support the Svelte framework and its limitation to a single web response field, three additional issues merit detailed consideration. Comprehensive knowledge of the ReVISit programme is presumed, and technical feedback on the software is provided through this discussion. Where terminology is unclear, further details may be obtained from the ReVISit documentation (https://revisit.dev/docs/introduction/). The primary issues examined herein concern form caching mechanisms, data storage limitations when handling large provenance graphs, and study creation workflows.

Form Caching

A significant advantage of the ReVISit programme is its capacity to store user responses during study participation, which proves beneficial when network connectivity is disrupted. However, this functionality results in the browser retaining submitted data even after study completion, thereby permitting users to modify their answers retroactively. Furthermore, shared computer environments present complications, as subsequent users may inadvertently override the responses of their predecessors. Although the study navigator (https://revisit.dev/docs/analysis/revisit-modes/#study-navigator) provides a "Next participant" button to facilitate participant transitions, this navigator was disabled. This navigator presented a cluttered user experience, whilst simultaneously permitting users to navigate freely between study pages, thus compromising study integrity. It would have been useful for a form to be automatically advanced to the next participant upon completion.

Data Storage

In the first study, seven participants generated 162 megabytes of text data, primarily attributable to provenance tracking (https://revisit.dev/docs/designing-studies/provenance-tracking/) and the logging of mouse and keyboard events. To mitigate this issue, the Window Event Debounce Time setting (https://revisit.dev/docs/typedoc/interfaces/BaseIndividualComponent/#windoweventdebouncetime) was adjusted to a longer time window in the second study, permitting fewer data points to be transmitted to the form. However, the increased debounce time may have resulted in the omission of certain events from the database. In addition, comparisons of time values between studies were no longer possible because the values were inaccurate. Twenty-eight participants yielded 22 megabytes of data in the second study, representing a substantial reduction.

Study Creation

The least limiting factor was that the study creation had to be designed in a JSON file. The methodology employed during local server testing required modification upon deployment to a production environment. As the study configuration had to reference the hosted server rather than the local development server. During local development a separate embedded url had to be used compared to the hosted version of the study. Changing these values proved to be time consuming. Implementation of a dynamic JavaScript or Typescript file would have permitted the insertion of dynamic environment variables to determine the server URL of the software version under evaluation.