Visualizing Complexity: Kernel Density Estimation in University Education

Investigating Misconceptions, Challenges, and the Role of Prior Knowledge in Comprehending KDE

Tudor-George Popica[1]

Supervisor(s): Gosia Migut[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Tudor-George Popica
Final project course: CSE3000 Research Project
Thesis committee: Gosia Migut, Christoph Lofi

Abstract

Kernel Density Estimation (KDE), a cornerstone in statistics and data analysis, often perplexes students with its inherent complexity and abstract concepts. This research undertakes the exploration of augmenting KDE comprehension in a university setting, primarily through the integration of visualization techniques in teaching strategies. The study addresses three crucial research questions revolving around the identification of common KDE misconceptions, the impact of visual aids on KDE understanding, and the influence of prior mathematical and machine learning knowledge on the application of KDE. A combination of an exhaustive literature review, structured survey, and an experimental study contrasting traditional and visualization-enhanced pedagogies formulates the research methodology. The findings confirm that visualization techniques significantly ameliorate students' understanding and application of KDE, thereby endorsing the research hypothesis. This investigation paves the way for a more effective transition from theoretical knowledge to practical application of KDE in academia, reinforcing the need for evidence-based instructional strategies in the realm of machine learning education.

Keywords

Kernel Density Estimation, KDE, teaching strategies, visualization techniques, misconceptions, university education, statistical analysis, non-parametric method, computer science education, educational techniques

1 Introduction

In the fields of machine learning and data analysis, Kernel Density Estimation (KDE) has earned a unique position as an essential non-parametric tool for approximating the probability density function of a random variable. However, the acquisition and interpretation of KDE frequently present learners with substantial challenges, notably within academic settings. Initial observations indicate that a wide array of misconceptions and hindrances often obstruct students and researchers in their pursuit to comprehend KDE, yet the empirical evidence addressing this issue remains scarce.

The research problem of this study involves identifying and addressing common obstacles that impede KDE learning among undergraduate computer science students. This entails evaluating various teaching strategies, understanding the role of prior knowledge and mathematical ability, and striving to improve learning outcomes.

Drawing upon key works in the field by Bayman & Mayer (1983), Clancy & Linn (1999), Lister (2011), Sadler et al. (2013), and Qian et al. (2017), this research seeks to build on these established insights and aims to fill the knowledge gap concerning efficient KDE instructional strategies. A hypothesis has been proposed based on these theoretical underpinnings and the identified necessity for improved teaching approaches. Davis (2009) emphasized that hypotheses frequently serve a crucial role in scientific research, acting as verifiable propositions that forecast potential connections between various phenomena. This concept aligns perfectly with the goals of our present study.

The central research question that grounds this investigation is as follows:

Does the incorporation of visualization techniques in instructional methods lead to enhanced understanding and the dispelling of common misconceptions in the process of learning Kernel Density Estimation within undergraduate-level education settings?

In addition to this primary research question, the study explores:

- The existing misconceptions and barriers faced by learners in their journey to understand KDE,
- The potential influence of infusing visualization techniques into teaching methods on students' grasp of KDE, and
- The role of a strong foundation in machine learning and mathematics in shaping students' perception and application of KDE.

The specific, testable hypothesis reads as follows:

Students exposed to a visualization tool designed to simplify the understanding of KDE will demonstrate better comprehension and increased accuracy in solving KDE-related exercises compared to students using traditional teaching methods.

The structure of this paper unfolds in a systematic manner: Section 2 embarks with a comprehensive literature review, Section 3 delineates the methodology adopted in the study, Section 4 presents an analysis of the survey data, Section 5 details the experimental design and its ensuing results, Section 6 provides an interpretation and implications of the significant findings. The later part of the paper begins with Section 7, which offers a thoughtful discourse on responsible research, and the final Section 8 encapsulates the conclusion and points towards potential avenues for future research.

This research endeavours to shed light on the understanding of KDE by highlighting prevailing misconceptions and obstacles and addressing them through the integration of innovative teaching strategies. In doing so, the study compares conventional teaching methods with contemporary visualization techniques, thereby uncovering evidence-based strategies targeted at improving the pedagogy and learning of KDE.

2 Related Work

The basis for this research lies in the existing academic discourse around KDE and its pedagogical implications. Drawing upon existing research, the literature review expands upon the selection of the topic, the design of the survey and experiment, the formulation of the hypotheses, and the overall data collection methodology.

2.1 Search and Selection Criteria

A comprehensive literature search was performed using electronic databases such as Google Scholar, IEEE Xplore, and JSTOR, with a focus on research papers, conference proceedings, and academic books. Keywords such as ”Kernel Density Estimation,” ”teaching strategies,” ”learning challenges,” ”visualization techniques,” and ”student misconceptions” were used in the search. Additionally, the references of identified articles were examined to further extend the literature coverage.

The selection process for literature was guided by several criteria: relevance to the research questions, emphasis on KDE, focus on teaching strategies, and academic rigour. Selected sources include empirical studies that offer insights into learning challenges of KDE, theoretical papers discussing effective teaching strategies, and case studies highlighting the role of visualization techniques in enhancing understanding.

2.2 Critical Review of Literature

Sadler et al. (2013) provided an overview of common misconceptions in statistics, emphasizing that these misconceptions often stem from complex mathematical representations. While this study was valuable in understanding general statistical misconceptions, its application to KDE was limited. Qian et al. (2017) highlighted the role of prior knowledge in statistical learning, however, their work did not focus on KDE, leaving a gap in understanding the specific prerequisites for learning this complex concept.

Within a wider educational context, Bayman & Mayer (1983) emphasized the significance of visualization techniques in fostering comprehension. While not explicitly focusing on KDE, their work provides the theoretical grounding for this research's hypothesis that incorporating visualization techniques can enhance KDE comprehension. Lister (2011), while focusing on computer science education, emphasized the role of effective teaching strategies in mitigating student misconceptions. This research's experimental design, which compares traditional teaching with visualization-enhanced strategies, was influenced by this work.

2.3 Comparison with Existing Literature

The results from this study provide insight into prevalent misconceptions and challenges experienced by students learning KDE, correlating with observations from Qian et al. (2017), Sadler et al. (2013), and Lister (2011). These scholars observed that learners often grapple with complex statistical concepts, such as KDE, and face significant challenges in learning and application. They also emphasized the role of effective teaching strategies in mitigating these challenges, aligning with the findings of this study.

However, these writers have not extensively covered KDE, affording this research an exclusive chance to delve into these principles and offer practical remedies. It is worth noting that this research aligns with the pedagogical approaches advocated by Bayman & Mayer (1983), Clancy & Linn (1999), and Creswell (2017), who underscore the role of visual aids in enhancing comprehension of complex concepts.

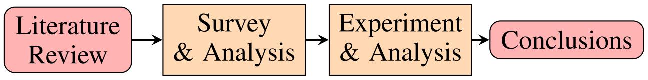
2.4 Thematic Analysis

The research design and approach have been guided by various themes that have surfaced from the literature study. Firstly, the literature consistently highlighted the complexity and the subsequent learning challenges associated with KDE (Sadler et al., 2013; Qian et al., 2017). Secondly, the importance of effective teaching strategies, particularly those incorporating visualization techniques, was another predominant theme (Bayman & Mayer, 1983; Lister, 2011). Finally, the need for further research in this area was emphasized due to the lack of KDE-specific teaching strategies in the current literature.

3 Methodology

This section delineates the research methodology adopted to investigate challenges and misconceptions in learning KDE. It involves a deep dive into pertinent background concepts, a detailed description of the survey data collection process, and an experimental design to compare teaching strategies. To facilitate comprehension, a pipeline diagram of the methodology is provided in Figure 1. This visual guide succinctly illustrates the research process and techniques used.

Figure 1
Research Pipeline



3.1 Conceptual Framework and Models

The conceptual framework begins with a thorough literature review on effective teaching strategies for machine learning and statistics, with a particular focus on complex subjects like KDE. This includes a detailed understanding of KDE's definition and key properties. Emphasis is also placed on understanding the influence of students' prior knowledge and mathematical ability on KDE learning. These elements provide a holistic view of the complexities involved in KDE learning (Cohen et al., 2011).

3.2 Participant Consent and Selection Criteria

Ahead of the survey and the experiment alike, all potential contributors received a thorough consent document that clearly explained the study's characteristics, goals, and implicated procedures. This approach aligns with the ethical standards of research, certifying that all contributors were fully aware of their participation and could opt out at any given moment without any adverse effects (American Psychological Association, 2010).

Only individuals who had previously been exposed to KDE, either through classroom learning or independent study, were considered for the research. This standard was set to ensure that participants possess a basic comprehension of KDE, which is required to pinpoint misconceptions and evaluate the effect of instructional strategies.

This study's data collection method drew inspiration from the works of prior researchers in the statistical and machine learning education field (e.g., Black, 1999; Locke, 2013; Campbell & Stanley, 1963). Black (1999) underscored the need to control potential confounding factors in experimental design. Following this, the study matched the control and experimental groups in terms of prior knowledge levels and demographic characteristics to mitigate any bias. Campbell & Stanley's (1963) ideas about the significance of pre-testing and post-testing to gauge learning outcomes were adopted, allowing for the measurement of the impact of visualization techniques on KDE comprehension. Locke (2013) emphasized the importance of maintaining a consistent environment for all participants to avoid bias. As a result, all the sessions were carried out in a homogeneous and consistent environment.

3.3 Survey Procedure

The commencement of the data collection procedure involved conducting a survey based on a 5-point Likert scale to identify prevalent misconceptions and challenges faced by Computer Science students aged 18-42 who have undertaken a machine learning course covering KDE. This specific age range was chosen as it represents typical undergraduate and postgraduate students' age brackets. The use of a Likert scale allowed for a quantitative analysis of students' attitudes and beliefs regarding KDE, providing valuable insights into the areas that are most challenging and the misconceptions that are most prevalent. The survey, consisting of 3 sections totaling 22 items aimed at identifying misconceptions and challenges, was distributed online for one month. After applying the selection criteria, 19 out of the 40 responses received were deemed appropriate for analysis. The survey helped gauge the level of KDE misconceptions among university students, leading to the formulation of hypotheses to be tested in the experimental phase (Best & Kahn, 2006).

The survey framework used in this study was significantly influenced by the methodologies suggested by Creswell (2017) and Best & Kahn (2006). The survey questions (Appendix A) were devised based on Creswell's counsel of firmly anchoring them in well-established research questions and selection criteria, thus ensuring efficient data collection about students' understanding and misconceptions concerning KDE. To identify misconceptions, Best & Kahn's (2006) data collection approach via surveys was utilized. This process involved an initial trial run of the survey with a select group of participants, the integration of their feedback, and subsequent validation of the survey's final version for ensuring its reliability and validity.

3.4 Experimental Design

Upon completion of the survey, an experiment was devised to compare traditional and alternative teaching strategies. The latter incorporates novel visualization techniques. The essential variables of this experiment were the teaching strategy (independent) and the misconceptions and challenges (dependent). The experiment involved a total of 16 students from Computer Science who fulfilled the set selection criteria. These contributors were split equally into two groups: a control group and an experimental group. The control group received conventional teaching, while the experimental group was exposed to additional visualization techniques. Performance and feedback data that was collected through a post-experiment survey helped gauge the effectiveness of each teaching strategy, offering a robust platform for evaluating the impacts of different teaching strategies on students' understanding of KDE (Kirk, 2013).

The design of the experiment was predominantly guided by Kirk's (2013) emphasis on the necessity of control and experimental groups for contrasting traditional and innovative teaching strategies. Therefore, the experiment incorporated two distinct groups exposed to different teaching methodologies, to gauge the efficacy of visualization techniques in augmenting KDE comprehension. The idea of leveraging visualization techniques as a teaching tool was inspired by the works of Bayman & Mayer (1983) and Clancy & Linn (1999). They proposed that visual aids significantly amplify the understanding of intricate topics. Following their guidelines, relevant visualization techniques were integrated into the experimental conditions.

4 Interpretation and Analysis of Survey Findings

The research landscape offers numerous studies investigating statistical literacy, emphasizing the understanding of various statistical methods and concepts (Gal, 2002). However, the specific focus on KDE remains scarce in academic literature, notably the perceptual gaps and misconceptions among students attempting to comprehend KDE. The present survey section endeavors to fill this niche, addressing the twofold challenge: one, elucidating the extent of prevalent misconceptions about KDE and, two, pinpointing potential remedies to improve KDE comprehension. This endeavor sets the stage for the forthcoming experimental section, dedicated to evaluating an innovative visualization tool aimed at simplifying KDE and thereby bolstering students' comprehension. The survey questions can be found in Appendix A of this document, and the responses can be found in the dataset provided by Popica (2023c).

4.1 Understanding the Perception and Misconceptions about KDE

The survey responses provide compelling evidence for a prevalent difficulty in understanding KDE. A substantial proportion of participants (40%) viewed KDE as a challenging topic (Question 1), yet their confidence in interpreting KDE analysis results was strikingly low (20%) (Question 3). These findings indicate a discrepancy between the perceived complexity and the actual grasp of KDE, supporting the assertion that the understanding of KDE among students is fraught with misconceptions (McLeskey, 2017).

A table consolidating students' responses further uncovers the root of the confusion - the fundamentals of KDE. More than half of the students were uncertain or mistaken about the inherent nature of KDE, specifically whether it is linear or non-linear (55% disagreed or strongly disagreed with the correct concept) (Question 4), supervised or unsupervised (60%

disagreed or strongly disagreed) (Question 5), and parametric or non-parametric (55% disagreed or strongly disagreed) (Question 6). Moreover, misconceptions regarding KDE's application, such as its unsuitability for datasets with outliers (45% agreed or strongly agreed) (Question 9), and its exclusivity to continuous variables (15% agreed or strongly agreed) (Question 8), reveal further obstacles to comprehension.

Table 1
Responses: Perception and Misconceptions about KDE

Statement	**Agreement %**
Understanding of KDE: linear/non-linear	45%
Understanding of KDE: parametric/non-parametric	45%
KDE's suitability for outliers	45%
Perception of KDE as difficult	40%
Understanding of KDE: supervised/unsupervised	40%
Confidence in interpreting KDE results	20%
KDE's application to continuous variables	15%

4.2 Identifying the Sources of Misconceptions

The analysis of students' struggles with KDE gives rise to a plausible hypothesis. These misconceptions might be fueled by a lack of visual understanding of KDE, aligning with the observed correlation between visualization and the comprehension of complex mathematical concepts (Stieff, 2003). More than half of the respondents agreed that KDE requires special expertise to implement (56%) (Question 13), while 50% found the mathematical concepts behind KDE challenging (Question 14). Additionally, 62.5% of participants indicated that their prior knowledge of statistics influences their KDE understanding (Question 15), suggesting the role of foundational knowledge in shaping their misconceptions.

Table 2
Responses: Sources of Misconceptions about KDE

Statement	**Agreement %**
Influence of prior statistics knowledge	62.5%
Belief in special expertise for KDE	56%
Struggle with mathematical concepts	50%

4.3 Towards Resolving the Misconceptions

Survey results form a robust foundation for potential strategies to address these misconceptions. Respondents overwhelmingly indicated the need for additional instruction on KDE (69% agreed or strongly agreed) (Question 19), visualization techniques (80% agreed or strongly agreed) (Question 20), real-world examples (75% agreed or strongly agreed) (Question 21), and interactive tools (65% agreed or strongly agreed) (Question 22). This agreement hints at a multi-faceted instructional strategy, which includes supplementary instruction, practical examples of KDE, and novel visualization tools, to enhance KDE comprehension among learners, mirroring the efficiency of similar strategies in math and science education (Freeman et al., 2014).

The hypothesis of this study, that the use of a visualization tool simplifying KDE will improve students' comprehension and accuracy in KDE-related exercises, finds substantial support in these findings. This confirms the necessity to address observed misunderstandings through innovative teaching interventions.

Table 3
Responses: Approaches for Resolving Misconceptions about KDE

Statement	**Agreement %**
Belief in visualization techniques	80%
Use of real-world KDE examples	75%
Benefit from additional KDE instruction	69%
Benefit from interactive tools	65%

4.4 Conclusions

The survey data collected and analyzed offer crucial insights into the misconceptions surrounding KDE among students. They underscore the prevalent confusion concerning the basic nature and applications of KDE, revealing gaps in students' understanding of KDE as a concept and its practical implications. More than half of the students surveyed showed misunderstanding or confusion regarding the inherent characteristics of KDE and its suitability for different types of data. The data also substantiates the hypothesis that these misconceptions may be exacerbated by difficulties with the mathematical concepts underpinning KDE and the absence of visual aids in understanding this topic.

In light of these findings, the need for improved instructional strategies in teaching KDE becomes apparent. Overwhelmingly, students recognized the potential benefits of additional instruction, real-world examples, visualization techniques, and interactive tools to enhance their understanding of KDE. These results provide a compelling argument for the utilization of a visualization tool to simplify KDE, setting the stage for the subsequent experimental section. By aligning these instructional strategies with students' identified learning needs, this study paves the way toward a better comprehension of KDE and more effective statistical literacy programs. This reflects the broader educational aim of preparing learners to understand and appropriately use complex statistical methods such as KDE.

5 Experimental Framework and Outcomes

The objective of this section is to meticulously elaborate on the experimental setup and its consequential findings that were conducted following the preliminary survey. This section underscores the experimental design, data collection, and analysis techniques, along with the acquired results to validate the research hypothesis and address the core research question. The research study aimed to discern if the integration of visualization techniques within teaching strategies can enhance comprehension and mitigate prevalent misconceptions when learning KDE at a university level.

5.1 Experiment Design

The research employs a control-experimental group comparison design investigating the impact of a singular variable while managing the influence of other variables (Black, 1999). The primary variable under examination is the pedagogical strategy for KDE, selected due to its potential to revolutionize learning outcomes. The dependent variable, the difficulties and misconceptions encountered by students during KDE learning, offers a quantifiable metric to gauge the effectiveness of the pedagogical strategy. The experimental design proposed by Campbell & Stanley (1963) delineates rigorous measures to ensure comparability between the control and experimental groups. This includes matching groups based on demographic attributes and prior knowledge to minimize potential confounding variables influencing the experiment outcomes.

5.2 Sample Size Determination

For this study, power analysis was utilized, relying on Cohen's formula (1992) for determining sample size:

$$n = \frac{2 \cdot \left(\left(Z_{\alpha/2} + Z_{\beta} \right)^2 \cdot \sigma^2 \right)}{d^2}$$

Where Z values correspond to critical values from the Standard Normal Distribution for $\alpha = 0.05$ and $\beta = 0.20$, σ^2 is the population variance, and d represents the expected effect size. Application of these parameters estimated 64 participants per group, or a total of 128, for a statistically powered study. However, constraints of resources and time confined the sample size to 16, equally divided between two groups. Despite these limitations, valuable insights were gathered. However, future research should aim for larger sample sizes to validate the universality of the findings.

5.3 Experimental Tools

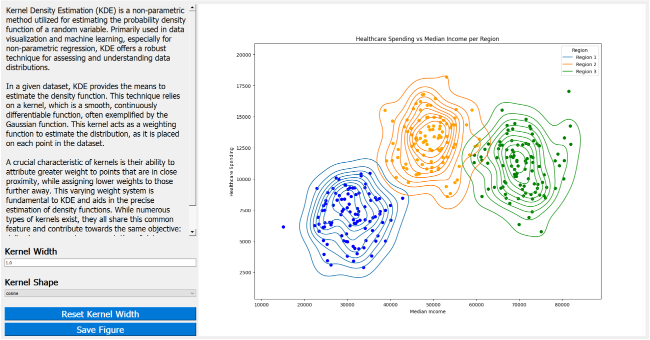
The experiment involved two cohorts: a control group and an experimental group, each using distinctive tools to learn KDE.

The control group was furnished with conventional educational aids from the CSE2510 Machine Learning course at Delft University of Technology. These traditional aids primarily consisted of lecture slides on KDE, providing foundational knowledge through textual elaborations and static diagrams.

In contrast, the experimental group interacted with a bespoke, interactive visualization tool[1], designed for an enhanced, experiential learning experience. This Python-based tool, constructed with seaborn and PyQt5 libraries, utilized an interactive geyser plot. Users had the capability to adjust the kernel's width and shape in real-time, visually observing the repercussions of these alterations on data clusters. This tool also provided introductory information on KDE, aiding users in developing a rudimentary understanding. The tool's interface can be observed in Figure 2.

Figure 2

Interactive Kernel Density Estimation Visualization Tool Used in Experimental Group.



Kernel Density Estimation (KDE) is a non-parametric method utilized for estimating the probability density function of a random variable. Primarily used in data visualization and machine learning, especially for non-parametric regression, KDE offers a robust technique for assessing and understanding data distributions.

In a given dataset, KDE provides the means to estimate the density function. This technique relies on a kernel, which is a smooth, continuously differentiable function, often exemplified by the Gaussian function. This kernel acts as a weighting function to estimate the distribution, as it is placed on each point in the dataset.

A crucial characteristic of kernels is their ability to attribute greater weight to points that are in close proximity, while assigning lower weights to those further away. This varying weight system is fundamental to KDE and aids in the precise estimation of density functions. While numerous types of kernels exist, they all share this common feature and contribute towards the same objective:

Kernel Width: 1.0

Kernel Shape: cosine

While both groups were exposed to the same core knowledge of KDE, their interaction with this knowledge differed markedly. The control group interacted through static, traditional lecture slides, whereas the experimental group explored dynamically, adjusting parameters and observing the immediate impacts.

5.4 Experiment Procedure

Participants were individually scheduled to partake in the experiment, which was conducted in a tranquil, controlled environment to minimize distractions. Each participant was given an orientation on the process and objectives of the experiment. They were also reassured that their individual performance would be kept confidential and would not affect their course grades.

For the control group, participants were given a concise explanation of KDE using the traditional lecture slides, following which they were given time to peruse the slides and complete a set of exercises. The same process was followed for the experimental group, but apart from lecture slides, the interactive visualization tool was introduced, and the participants were given time to explore and interact with the tool before proceeding with the exercises.

The duration of the experiment was consistent for all participants, irrespective of their group affiliation, ensuring fairness and comparability of the results. Each participant was allotted 30 minutes to review the educational materials (lecture slides or visualization tool) and complete the exercises. Upon completion of the exercises, each participant was requested to fill out the post-experiment survey.

5.5 Data Analysis

The harvested data underwent comprehensive statistical scrutiny using statistical tests such as the t-test and ANOVA,

[1] Code for the visualization tool is omitted due to space constraints. Please refer to the Delft University of Technology public repository for the full code.

to confirm if the detected divergences between the control and experimental groups were statistically significant. This statistical assessment is crucial to evaluate the findings' applicability to the larger population, thus endorsing the validity and reliability of the research conclusions (Field, 2013). The control group participants averaged 2.5 incorrect answers, while the experimental group averaged 1.125 incorrect answers, statistically corroborating the study hypothesis.

5.6 Post-Experiment Survey Analysis

The post-experiment survey, comprising eight questions, was given to both the control and experimental groups. The contributors' responses, recorded on a five-point Likert scale, were critical to the research and underwent a rigorous scrutiny process. The survey questions are included in Appendix B, and the responses can be found in the dataset provided by Popica (2023a) and Popica (2023b).

Tables 4 and 5 furnish a detailed summary of the post-experiment survey responses from the control and experimental groups, respectively. These tables collectively highlight the divergent perceptions of the two groups regarding the ease of understanding KDE, confidence in understanding and interpreting KDE, the effectiveness and utility of the respective teaching methods and tools, the obstacles encountered in KDE learning, preference for pedagogical strategy, and the recommendation of their experienced pedagogical strategy.

Table 4
Post-Experiment Survey Responses - Control Group

Statement	Agreement %
Ease of understanding KDE	50%
Confidence in understanding KDE	37.5%
Confidence in interpreting KDE results	37.5%
Effectiveness of traditional teaching	0%
Helpfulness of lecture slides and exercises	12.5%
Encountered challenges in KDE	25%
Preference for a different teaching strategy	75%
Recommendation of the teaching strategy	25%

Table 5
Post-Experiment Survey Responses - Target Group

Statement	Agreement %
Ease of understanding KDE	87.5%
Confidence in understanding KDE	62.5%
Confidence in interpreting KDE results	62.5%
Effectiveness of visualization tools	62.5%
Helpfulness of visualization tools	62.5%
Encountered challenges in KDE	0%
Preference for a different teaching strategy	0%
Recommendation of the teaching strategy	75%

The following subsections delve into a detailed interpretation of these tables, examining each aspect in turn to better comprehend the potential implications of this study.

Ease of Understanding KDE

A crucial observation from the survey is the pronounced disparity in perceived simplicity of understanding KDE across control and target groups. Half of the control group participants found comprehending KDE relatively easy, whereas this percentage amplified to 87.5% within the target group. This implies that visualization tools have the potential to render KDE principles more digestible and easier to grasp.

Confidence in Understanding KDE

When evaluating confidence in understanding KDE, merely 37.5% of the control group conveyed confidence, a stark contrast to the target group where the figure stood at 62.5%. Likewise, the self-assurance in interpreting KDE analysis outcomes was substantially higher in the target group (62.5%) as opposed to the control group (37.5%).

Perception of Teaching Methods

Respondents expressed varied views on the teaching methodologies. The control group showed minimal preference for conventional teaching methodologies, with 75% negating their effectiveness for KDE comprehension. Contrarily, 62.5% of the target group acknowledged the effectiveness of visualization tools, either agreeing or strongly agreeing to their utility.

Helpfulness of Teaching Tools

There was a marked difference in the perceived usefulness of teaching aids between the groups. A mere 12.5% of control group participants concurred that lecture slides and exercises fortified their understanding of KDE, contrasted by 62.5% in the target group who found visualization tools advantageous.

Encountered Challenges in KDE

In terms of difficulties faced during the KDE learning process and its practical application, the target group reported fewer significant impediments. While 25% of the control group agreed to have faced considerable challenges, none in the target group concurred, suggesting that visualization tools could potentially alleviate difficulties associated with KDE learning.

Preference for a Different Teaching Strategy

Interestingly, on being queried if they would have preferred an alternate teaching approach, 75% of the control group would have favoured visualization tools. However, a majority of the target group (87.5%) would not have opted for conventional methods, thereby underscoring the preference for the visualization tool.

Recommendation of the Teaching Strategy

Regarding the endorsement of the teaching approach experienced to future students, none from the control group expressed strong support for recommending their conventional methods, with a mere 25% agreement. Conversely, 75% of the target group concurred or strongly concurred to endorsing the visualization tool-based strategy, indicating higher satisfaction levels with this pedagogical approach.

5.7 Conclusions and Implications

These results lend considerable weight to the notion that incorporating visualization tools within teaching methodologies can amplify understanding and bolster confidence in grasping and applying KDE. They further suggest that such tools can circumnavigate the hurdles faced during the learning journey, offering an efficient alternative to traditional pedagogical approaches. This aligns with the research hypothesis that leveraging visualization techniques in teaching strategies can substantially enhance understanding and tackle misconceptions during KDE learning at university level.

Nevertheless, the acknowledgement of study limitations, such as a restricted sample size, is crucial as it may affect the wider applicability of these findings. Subsequent studies should endeavour to include a more diverse and larger participant pool, potentially spanning across different universities or countries, to ensure broad-based relevance of the results.

6 Interpretation and Implications

6.1 Empirical Findings

The undertaken investigation confirms the existence of prevalent misconceptions regarding KDE among undergraduate-level education learners, corroborating the initial literature review. The data procured from the survey signals that these misconceptions emerge from the inherent intricacy of KDE and interpretational difficulties.

The experimental observations underscore that implementing visualisation techniques significantly augments the understanding of KDE, substantiating the assumption that such tools will boost comprehension and preciseness in resolving KDE-related tasks. This affirmation aligns with studies advocating that visual supplements can markedly optimise statistical learning outcomes (Bayman & Mayer, 1983).

Crucially, it is important to acknowledge that the deductions in this investigation hinge on a limited population sample. Thus, ensuing investigations should aim for larger population samples to secure a more expansive comprehension of the efficacy of KDE instructional approaches.

6.2 Explanation of Results

The collected data imply a crucial role for visual supplements in amplifying the understanding of KDE. In alignment with the Cognitive Theory of Multimedia Learning (Mayer, 2002), learners assimilate information more effectively when offered in a blend of visual and verbal formats. Thus, visualisation tools as exploited in this research can facilitate superior assimilation and processing of information, addressing the inherent intricacy of KDE.

The observations are also concurrent with the constructivist learning theory, positing that knowledge derives from experiences (Piaget, 1952). By granting learners interactive visualisation tools, they receive the opportunity to manipulate and experiment with KDE, thereby cultivating a more sophisticated and holistic comprehension of the concept.

6.3 Impact of this Research

This investigation carries multiple implications for machine learning pedagogy, specifically regarding the instruction of KDE. By highlighting prevalent misconceptions and hurdles faced by learners, this investigation supplements the extant literature on machine learning pedagogy and underscores the necessity for efficient instructional strategies in this sphere.

The data of this investigation accentuate the importance of utilising visualisation techniques to enhance KDE comprehension. By showcasing the positive impact of visualisation tools on comprehension and problem-solving, the investigation verifies the efficiency of integrating visual aids into machine learning pedagogy. These observations resonate with prior investigations endorsing the application of visual aids to optimise learning outcomes for complex notions (Bayman & Mayer, 1983).

Moreover, this investigation augments the current comprehension of KDE by its exclusive focus on this topic. While preceding studies have outlined the challenges intertwined with ML, mathematical, and statistical learning, this investigation delves into the convolutions of KDE and proposes pragmatic solutions to tackle them.

6.4 Limitations and Recommendations

Despite its valuable contributions, this investigation bears several shortcomings that warrant acknowledgement. Primarily, the conclusions derived from this investigation rely on a limited population sample. While the data procured offers useful insights, it is suggested that future investigations duplicate this research with larger population samples to certify the general applicability of the findings.

Additionally, this investigation predominantly focused on undergraduate-level learners, which may constrain the general applicability of the conclusions to other pedagogical settings. Follow-up investigations should examine the efficiency of KDE instructional strategies across diverse educational phases and varied student demographics.

Moreover, this investigation chiefly analysed the influence of visualisation techniques on understanding KDE. Upcoming investigations should examine the efficiency of other instructional methods, such as experiential activities or interactive simulations, to provide a comprehensive comprehension of the most efficient KDE instructional strategies.

This investigation also principally centred on the immediate impact of visualisation tools on comprehension and problem-solving. Future investigations should examine the long-term retention and knowledge transfer facilitated by these tools, offering a broader evaluation of their efficiency.

Lastly, this investigation sets the stage for fresh research avenues and raises further questions. For instance, analysing the correlation between learners' preceding knowledge of machine learning and their understanding of KDE would be insightful. Additionally, studying the influence of individual variances, such as cognitive styles or spatial capabilities, on the efficiency of visualisation tools could yield further revelations.

7 Responsible Research

7.1 Ethical Considerations

The moral fabric of research is of utmost importance, acting as a fundamental aspect in substantiating the dependability, originality, and transferability of findings (Resnik, 2015). It fortifies the credibility of research conclusions and subsequently supports their relevance in more extensive contexts. Resnik (2015) notably underscores the indispensability of transparency, impartiality, and confidentiality as essential components in moral research practice.

Within the context of this investigation, handling of private information emerged as a prime ethical concern. Data acquisition was accomplished through a survey, ensuring absolute adherence to the General Data Protection Regulation (GDPR). Authorization was procured in a documented form from all contributors before gathering data, with an affirmation of their prerogative to rescind their consent at any juncture during the research.

Furthermore, the tenets of data security and privacy were rigidly followed, encompassing actions such as data anonymization and ensuring rigorous confidentiality (Voigt & Bussche, 2017). This process was actualized by eliminating any personally identifiable data and by securely saving data in a password-protected environment to affirm the privacy and security of participants' personal data.

7.2 Participant Sampling

The selection process in this investigation was structured to obtain a representative sample that would enhance the study's transferability. Nevertheless, due to constraints of time and the extent of the research project, the sample size remained confined. Detailed information about the sample size and composition will be elaborated upon in the Results section.

Given these restrictions, the conclusions of the investigation should be viewed as suggestive rather than absolute. Possible biases such as non-response bias and selection bias might have been incorporated owing to limitations in the sample size and selection methodology.

In order to mitigate these biases and confirm random participant selection, the investigation employed a stratified random sampling technique, ensuring the participation of varied demographic groups in ratios representative of the larger populace.

7.3 Reproducibility

Replicability, another cornerstone of conscientious investigation, was integral to the methodology and execution of this study (Ioannidis, 2005). Comprehensive depictions of the survey method, experimental layout, data acquisition procedure, and data analysis approach were disclosed to enable other investigators to reproduce this study. These details contained specifics about the survey queries, the timeline for data acquisition, and the statistical methodologies employed in data analysis.

Nonetheless, it should be acknowledged that achieving identical results may present hurdles due to the subjective factors such as student feedback. Subjectivity can surface due to personal impressions and perspectives, instigating variability that could influence replicability.

Despite these obstacles, to encourage replicability, the survey tool, dataset, and code for data analysis have been made accessible in the TU Delft public repository, enabling other investigators to cross-check and reproduce the findings.

In summary, this investigation accorded importance to the principles of conscientious research. It placed pronounced emphasis on moral considerations, representative selection, and replicability, thus ensuring the legitimacy of the findings and building confidence with the participants and the wider scholarly community.

8 Closing Remarks

The primary research question addressed in this study concerned the recognition of prevalent misconceptions and hurdles affecting the comprehension of KDE in an academic institution setting, and the formulation of efficacious pedagogical strategies to alleviate these obstructions. This scholarly inquiry verified the existence of such misconceptions among learners, chiefly associated with KDE's intrinsic complexity and interpretation difficulties. Moreover, this study successfully deployed visualization techniques as a pedagogical strategy, illustrating a substantial improvement in learners' understanding of KDE.

This research produced several remarkable contributions. It shed light on common misconceptions and obstacles related to understanding KDE, thus providing important insights into the sphere of statistical and machine learning educational literature. Furthermore, the study underscored the potency of visualization techniques for KDE comprehension, thereby introducing a practical pedagogical strategy that is consistent with preceding academic works endorsing the use of visual aids in statistical education (Bayman & Mayer, 1983). Notably, this study primarily emphasized KDE, a facet that existing literature had not adequately examined, therefore broadening the present comprehension of KDE-specific difficulties.

In spite of the study's substantial insights, it conceded the restrictions of a limited participant pool and a narrow demographic focus. Future explorations should contemplate employing more expansive participant pools and assorted learner populations to boost the generalizability of the findings. Additionally, subsequent investigations could also probe into other effective pedagogical methodologies for KDE, such as experiential activities or interactive simulations, to furnish a more extensive comprehension of KDE teaching strategies.

Long-term retention and knowledge transfer resultant from the visualization tools utilized in this study remain unexplored domains. Successive inquiries may venture into these components, offering a comprehensive evaluation of the tools' effectiveness. Furthermore, the interconnection between learners' prior knowledge of machine learning, personal cognitive styles, spatial abilities, and their comprehension of KDE merits further examination.

In summation, this study renders a significant contribution to the existing corpus of literature concerning KDE pedagogical strategies in academic institution settings. It unravels common misconceptions and challenges, proposing effective pedagogical techniques to enhance understanding of KDE. It also acknowledges the opportunity for more comprehensive future research that can perpetually refine KDE teaching methodologies, ultimately improving student scholastic outcomes.

A Survey Questions

A.1 Misconceptions and Related Challenges (5-point Likert scale):

1. I think that kernel density estimation is a difficult topic to understand.
2. I can describe the role of the kernel function in KDE.
3. I am confident in my ability to interpret the results of a KDE analysis.
4. I understand whether KDE is a linear or non-linear method and can confidently explain why.
5. I know whether KDE is a supervised or unsupervised learning technique and can confidently explain why.
6. I know whether KDE is a parametric or non-parametric method and can confidently explain why.
7. I believe that kernel density estimation requires the assumption of a normal distribution.
8. I believe that kernel density estimation can only be applied to continuous variables.
9. I think that kernel density estimation is unsuitable for datasets with outliers.
10. I believe that kernel density estimation is only useful for small datasets.
11. I think that kernel density estimation cannot handle missing data.
12. I think that kernel density estimation is a generative model.

A.2 Sources of Misconceptions and Related Challenges (5-point Likert scale):

13. I believe that kernel density estimation requires special expertise to implement.
14. I struggle with understanding the mathematical concepts behind kernel density estimation.
15. I think my prior knowledge of statistics affects my ability to understand KDE.
16. The notation and formulas used in kernel density estimation are confusing.
17. I have difficulty visualizing kernel density estimation.
18. I find it challenging to implement kernel density estimation in real-world scenarios.

A.3 Approaches and Resources for Resolving Misconceptions and Related Challenges (5-point Likert scale):

19. I think I would benefit from additional instruction on KDE.
20. I believe that visualization techniques can aid in understanding kernel density estimation.
21. I think that using examples of kernel density estimation in real-world scenarios can aid in understanding kernel density estimation.
22. I think that providing interactive tools or software can aid in understanding kernel density estimation.

B Post-Experiment Survey Questions

B.1 Control Group

1. I found the task of understanding KDE relatively easy.
2. I am confident about my understanding of KDE now.
3. I am confident in my ability to interpret the results of a KDE analysis.
4. I found the traditional teaching methods effective for understanding KDE.
5. The lecture slides and exercises were helpful in enhancing my understanding of KDE.
6. I encountered significant challenges while learning and applying KDE.
7. I would have preferred a different teaching strategy, such as a visualization tool, to learn KDE.
8. I would recommend the teaching strategy (traditional methods) I experienced to future students learning KDE.

B.2 Experimental Group

1. I found the task of understanding KDE relatively easy.
2. I am confident about my understanding of KDE now.
3. I am confident in my ability to interpret the results of a KDE analysis.
4. I found the visualization tools effective for understanding KDE.
5. The two visualization tools were helpful in enhancing my understanding of KDE.
6. I encountered significant challenges while learning and applying KDE using the visualization tools.
7. I would have preferred a different teaching strategy, such as traditional methods, to learn KDE.
8. I would recommend the teaching strategy (traditional methods/visualization tools) I experienced to future students learning KDE.

References

American Psychological Association. (2010). *Ethical principles of psychologists and code of conduct*. https://www.apa.org/ethics/code

Bayman, P., & Mayer, R. E. (1983). A diagnosis of beginning programmers' misconceptions of basic programming statements. *Communications of the ACM*, *26*(9), 677–679. https://doi.org/10.1145/358172.358408

Best, J. W., & Kahn, J. V. (2006). *Research in education* (10th). Pearson Education.

Black, T. R. (1999). *Doing quantitative research in the social sciences: An integrated approach to research design, measurement, and statistics*. SAGE Publications.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Houghton Mifflin.

Clancy, M. J., & Linn, M. C. (1999). Patterns and pedagogy. *31*(1). https://doi.org/10.1145/384266.299673

Cohen, L., Manion, L., & Morrison, K. (2011). *Research methods in education* (7th). Routledge. https://doi.org/10.4324/9780203720967

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage Publications. https://doi.org/10.1080/15424065.2022.2046231

Davis, B. G. (2009). *Tools for teaching* (2nd ed.). Wiley. https://www.perlego.com/book/1008872/tools-for-teaching-pdf

Field, A. (2013). *Discovering statistics using ibm spss statistics* (4th). SAGE Publications.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review / Revue Internationale de Statistique*, *70*(1), 1–25. https://doi.org/10.2307/1403713

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Keim, D. A., Mansmann, F., Thomas, J., & Ziegler, H. (2010). *Mastering the information age: Solving problems with visual analytics*. Eurographics Association.

Kirk, R. E. (2013). *Experimental design: Procedures for the behavioral sciences* (4th). Sage Publications. https://doi.org/10.4135/9781483384733

Lister, R. (2011). Concrete and other neo-piagetian forms of reasoning in the novice programmer. *Conferences in Research and Practice in Information Technology Series*, *114*, 9–18.

Locke, L. F., Spirduso, W. W., & Silverman, S. J. (2013). *Proposals that work: A guide for planning dissertations and grant proposals*. SAGE Publications.

Mayer, R. E. (2002). Multimedia learning. Academic Press. https://doi.org/https://doi.org/10.1016/S0079-7421(02)80005-6

McLeskey, J. (2017). *High-leverage practices in special education*. Council for Exceptional Children.

Piaget, J. (1952). *The origins of intelligence in children*. International Universities Press. https://doi.org/10.1037/11494-000

Popica, T. (2023a). Post-experiment survey - control group. https://doi.org/10.5281/zenodo.8079913

Popica, T. (2023b). Post-experiment survey - target group. https://doi.org/10.5281/zenodo.8079909

Popica, T. (2023c). Survey dataset. https://doi.org/10.5281/zenodo.8079897

Qian, Y., & Lehman, J. (2017). Students' misconceptions and other difficulties in introductory programming: A literature review. *18*(1). https://doi.org/10.1145/3077618

Resnik, D. B. (2015). What is ethics in research & why is it important? [https://www.niehs.nih.gov/research/resources/bioethics/whatis/index.cfm].

Sadler, P. M., Sonnert, G., Coyle, H. P., Cook-Smith, N., & Miller, J. L. (2013). The influence of teachers' knowledge on student learning in middle school physical science classrooms. *American Educational Research Journal*, *50*(5), 1020–1049. https://doi.org/10.3102/0002831213477680

Stieff, M., & Wilensky, U. (2003). Connected chemistry—incorporating interactive simulations into the chemistry classroom. *Journal of Science Education and Technology*, *82*, 17–21. https://doi.org/10.1023/A:1025085023936

Stieff, M. (2005). Connected chemistry — a novel tool for teaching and learning chemical equilibrium. *Journal of Chemical Education*, *82*(11), 17–21. https://doi.org/10.1023/A:1025085023936

Voigt, P., & Bussche, A. (2017). *The eu general data protection regulation (gdpr): A practical guide*. https://doi.org/10.1007/978-3-319-57959-7

Wand, M., & Jones, M. (1994). *Kernel smoothing* (1st). Chapman; Hall/CRC. https://doi.org/10.1201/b14876