Return of the Lernaean Hydra: Experimental Evaluation of Data Series Approximate Similarity Search

Karima Echihabi
ENSIAS, Mohammed V Univ.

Kostas Zoumpatianos
Harvard University

Themis Palpanas
Université de Paris

Houda Benbrahim
ENSIAS, Mohammed V Univ.

ABSTRACT

Data series are a special type of multidimensional data that are present in numerous domains. A key operation in data series analysis pipelines is similarity search, which has been extensively studied in the data series literature, leading to the development of efficient exact indexing methods. In parallel, the multidimensional community at large has studied approximate similarity search techniques. In this paper, we propose a comprehensive taxonomy of similarity search techniques that reconciles the terminology used in these two domains, and we conduct a thorough experimental evaluation to compare approximate similarity search techniques under a unified framework, on synthetic and real datasets in memory and on disk. Although data series differ from generic multidimensional vectors (series usually exhibit correlation between neighboring values), our results show that simple modifications to exact data series indexing techniques enable them to answer approximate similarity queries with strong guarantees and an excellent empirical performance, on data series and vectors alike. These techniques outperform the state-of-the-art approximate techniques for vectors when operating on disk, while remaining competitive in memory.

PVLDB Reference Format:
Karima Echihabi, Kostas Zoumpatianos, Themis Palpanas, and Houda Benbrahim. Return of the Lernaean Hydra: An Experimental Evaluation of Data Series Approximate Similarity Search. *PVLDB*, (): xxxx-yyyy, 2019.
DOI:

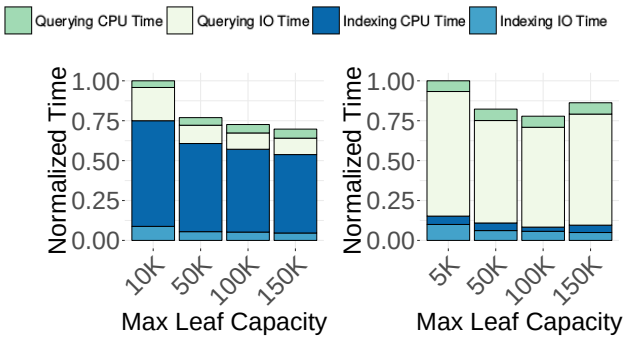
1. EXPERIMENTAL EVALUATION

This document contains the complete results for the experimental evaluation of data series approximate similarity search.

Proceedings of the VLDB Endowment, Vol. , No.
ISSN 2150-8097.
DOI:

1.1 Parametrization

The optimal leaf size for the DSTree and iSAX2+ methods is chosen based on the results in Figure 1.

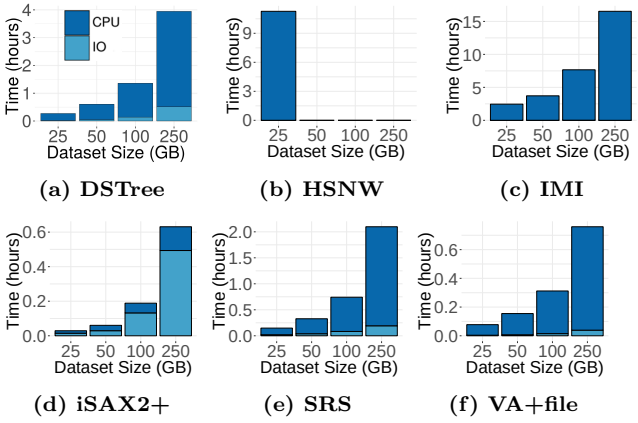


(a) DSTree (b) iSAX2+

Figure 1: Leaf size parametrization

1.2 Indexing Scalability

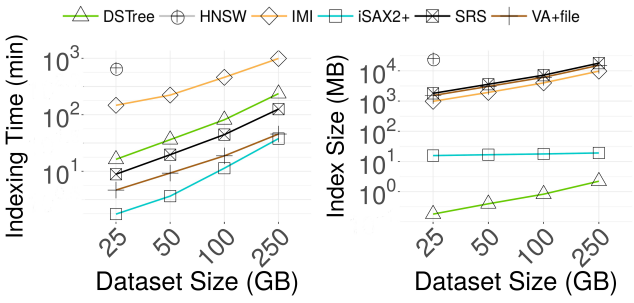
Figure 2 depicts the indexing scalability for each technique, while figure 3 compares the techniques together.



(a) DSTree (b) HSNW (c) IMI

(d) iSAX2+ (e) SRS (f) VA+file

Figure 2: Indexing scalability

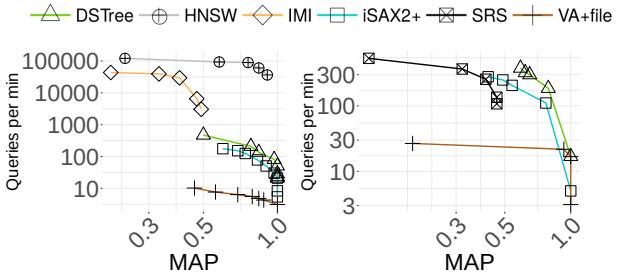


(a) Indexing Time (b) Size in Memory

Figure 3: Comparison of indexing scalability

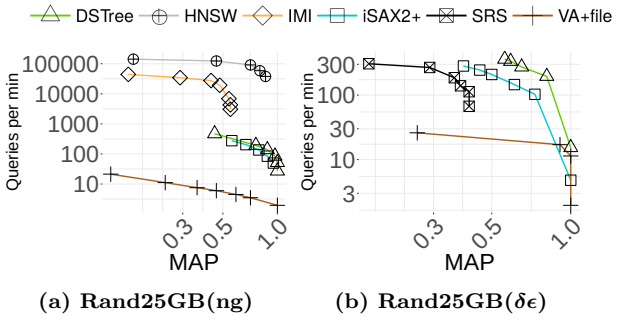
1.3 In-memory Experiments

Throughput vs. MAP. Figures 4, 5 summarizes the results for 100 1NN and 10NN queries on the Rand25GB dataset with length 256.



(a) Rand25GB(ng) (b) Rand25GB($\delta\epsilon$)

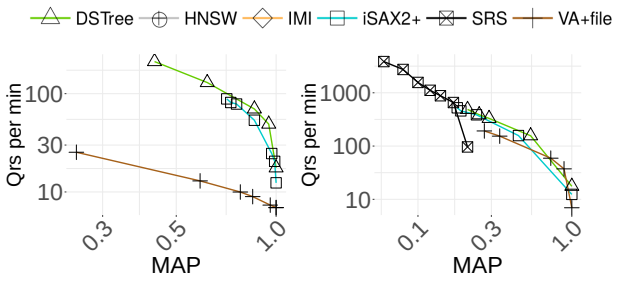
Figure 4: Efficiency vs. accuracy (MAP) in memory (1NN queries)



(a) Rand25GB(ng) (b) Rand25GB($\delta\epsilon$)

Figure 5: Efficiency vs. accuracy (MAP) in memory (10NN queries)

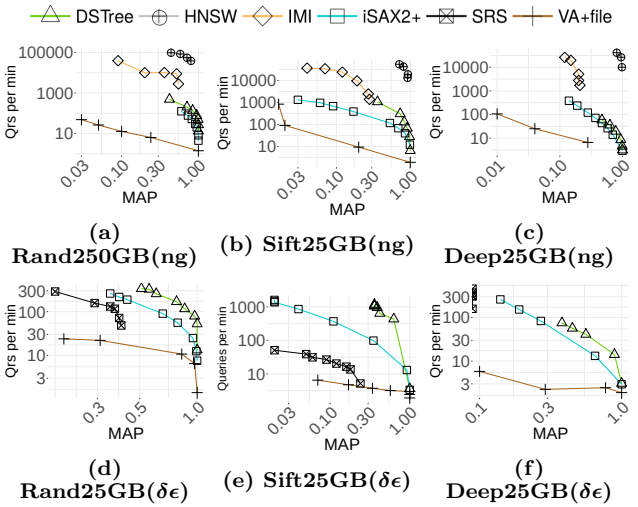
Figure 6 reports the results for the Rand25GB dataset with length 16384.



(a) Rand25GB(ng) (b) Rand25GB($\delta\epsilon$)

Figure 6: Efficiency vs. accuracy (MAP) in memory (Series Length = 16384, 100-100NN queries)

Figures 7 summarizes the results for 100 100NN queries on the Rand25GB, Sift25GB and Deep25GB datasets.

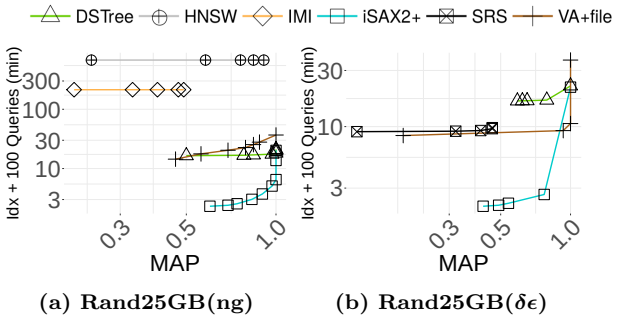


(a) Rand250GB(ng) (b) Sift25GB(ng) (c) Deep25GB(ng)

(d) Rand25GB($\delta\epsilon$) (e) Sift25GB($\delta\epsilon$) (f) Deep25GB($\delta\epsilon$)

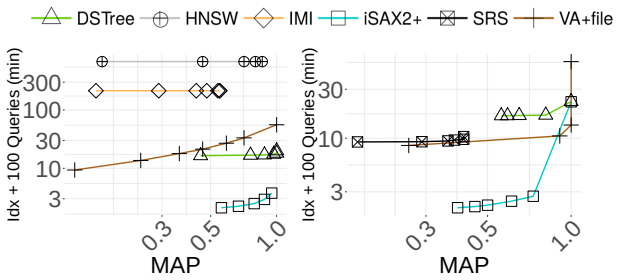
Figure 7: Efficiency vs. accuracy (MAP) in memory(100NN queries)

Indexing and Answering 100 Queries vs. MAP. Figures 8 and 9 report the total time it takes to build an index on the Rand25GB dataset and answer 1NN and 10NN queries respectively.



(a) Rand25GB(ng) (b) Rand25GB($\delta\epsilon$)

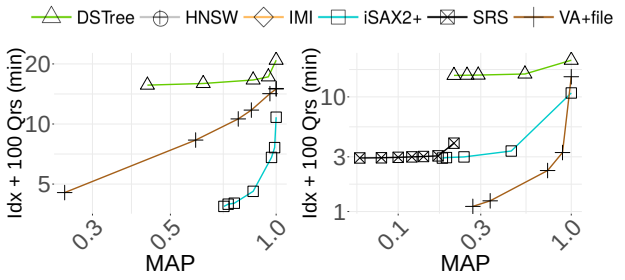
Figure 8: Efficiency vs. accuracy (MAP) in memory (Indexing + 100-1NN queries)



(a) Rand25GB(ng) **(b) Rand25GB($\delta\epsilon$)**

Figure 9: Efficiency vs. accuracy (MAP) in memory (Indexing + 100-10NN queries)

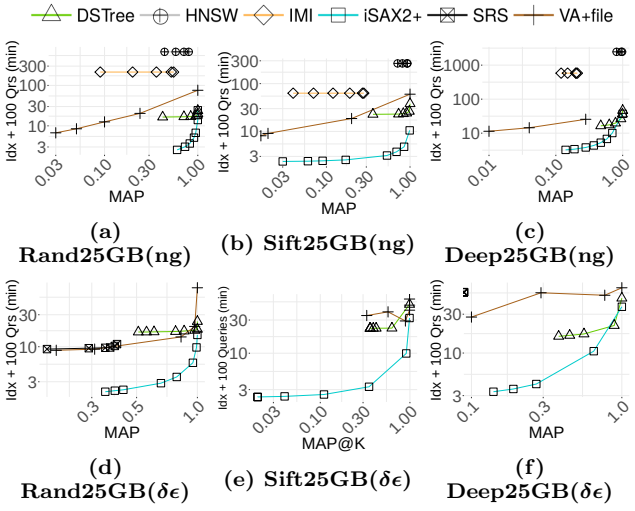
Figure 10 reports these numbers for the Rand25GB dataset with length 16384.



(a) Rand25GB(ng) **(b) Rand25GB($\delta\epsilon$)**

Figure 10: Efficiency vs. accuracy (MAP) in memory (Series Length = 16384, Indexing + 100-1NN queries)

Figure 11 reports the results for indexing and answering 100 100NN queries for the Rand25GB dataset with length 256 and the real datasets Sift25GB and Deep25GB.



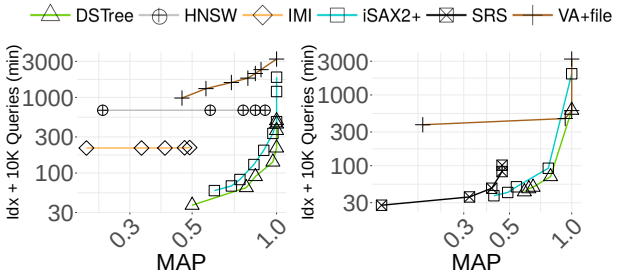
(a) Rand25GB(ng) **(b) Sift25GB(ng)** **(c) Deep25GB(ng)**

(d) Rand25GB($\delta\epsilon$) **(e) Sift25GB($\delta\epsilon$)** **(f) Deep25GB($\delta\epsilon$)**

Figure 11: Efficiency vs. accuracy in memory (Indexing + 100-100NN queries)

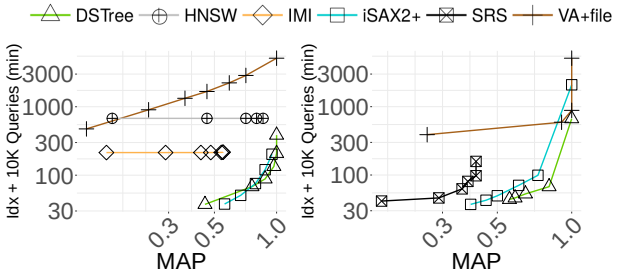
Indexing and Answering 10K Queries vs. MAP.

Figures 12 and 13 report the results for indexing and answering 10K 1NN and 10NN queries for the Rand25GB dataset with length 256..



(a) Rand25GB(ng) **(b) Rand25GB($\delta\epsilon$)**

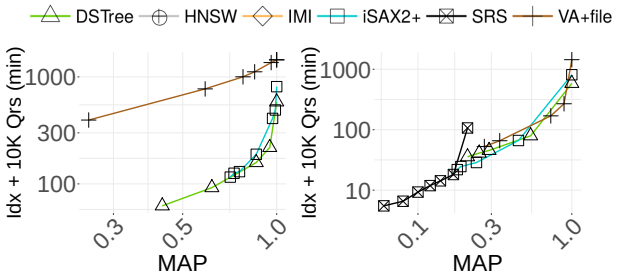
Figure 12: Efficiency vs. accuracy (MAP) in memory (Indexing + 10K-1NN queries)



(a) Rand25GB(ng) **(b) Rand25GB($\delta\epsilon$)**

Figure 13: Efficiency vs. accuracy (MAP) in memory (Indexing + 10K-10NN queries)

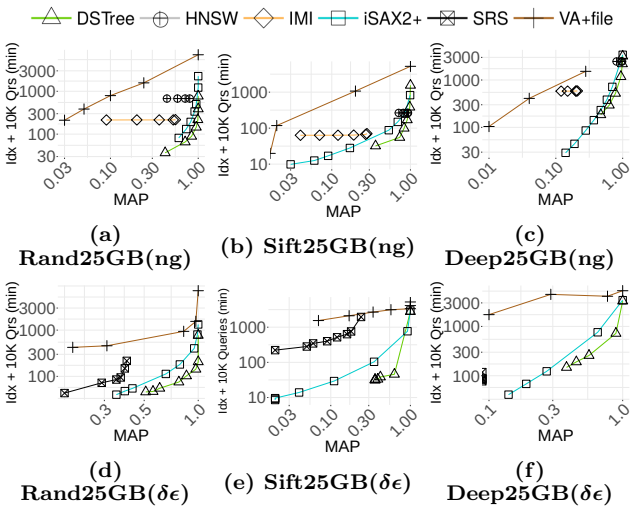
Figures 14 reports on the time it takes to build an index for the Rand25GB dataset with length 16384 and answer 10K 100NN queries.



(a) Rand25GB(ng) **(b) Rand25GB($\delta\epsilon$)**

Figure 14: Efficiency vs. accuracy (MAP) in memory (Series Length = 16384, Indexing + 10K-100NN queries)

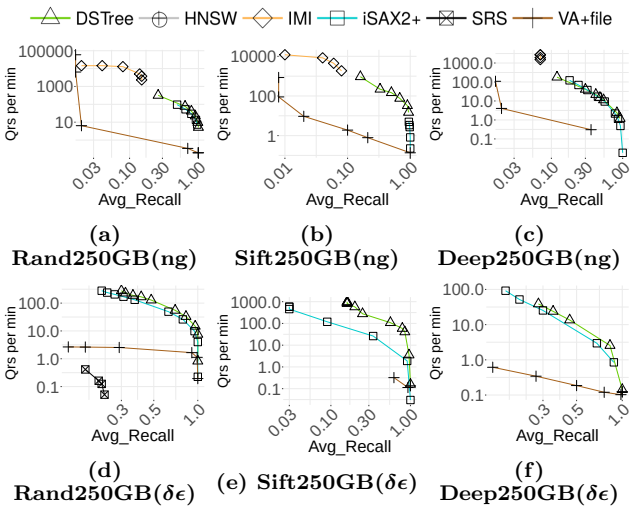
Figures 15 reports the numbers for answering 10K 100NN queries and indexing the Rand25GB, Sift25GB and Deep25GB datasets.



(a) Rand25GB(ng) (b) Sift25GB(ng) (c) Deep25GB(ng)

(d) Rand25GB($\delta\epsilon$) (e) Sift25GB($\delta\epsilon$) (f) Deep25GB($\delta\epsilon$)

Figure 15: Efficiency vs. accuracy in memory (Indexing + 10K-100NN queries)



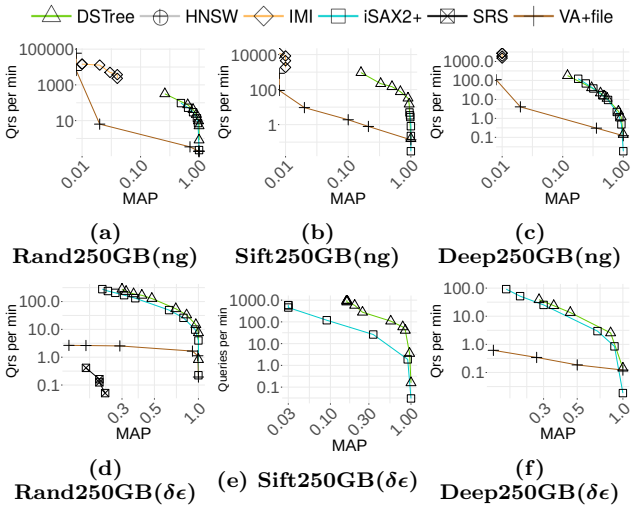
(a) Rand250GB(ng) (b) Sift250GB(ng) (c) Deep250GB(ng)

(d) Rand250GB($\delta\epsilon$) (e) Sift250GB($\delta\epsilon$) (f) Deep250GB($\delta\epsilon$)

Figure 17: Efficiency vs. accuracy (Avg_Recall) on disk (100NN queries)

1.4 On-disk Experiments

Throughput vs. MAP. Figure 16 summarizes the results for 100NN queries on Rand250GB, Sift250GB and Deep250GB.



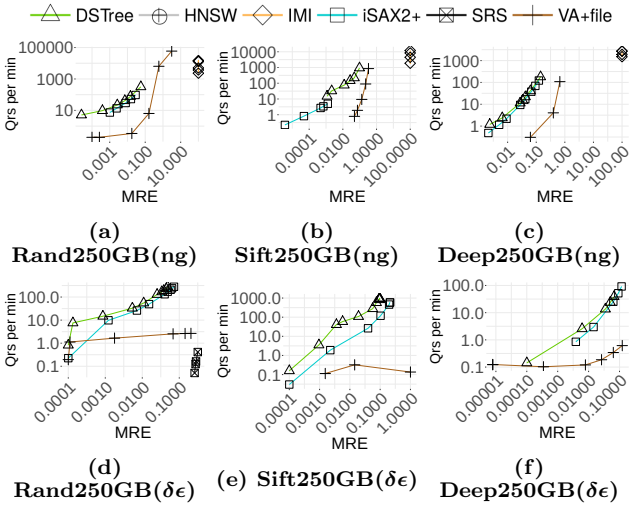
(a) Rand250GB(ng) (b) Sift250GB(ng) (c) Deep250GB(ng)

(d) Rand250GB($\delta\epsilon$) (e) Sift250GB($\delta\epsilon$) (f) Deep250GB($\delta\epsilon$)

Figure 16: Efficiency vs. accuracy (MAP) on disk (100NN queries)

Throughput vs. Avg_Recall. Figure 17 summarizes the results for 100NN queries on Rand250GB, Sift250GB and Deep250GB.

Throughput vs. MRE. Figure 18 summarizes the results for 100NN queries on Rand250GB, Sift250GB and Deep250GB.

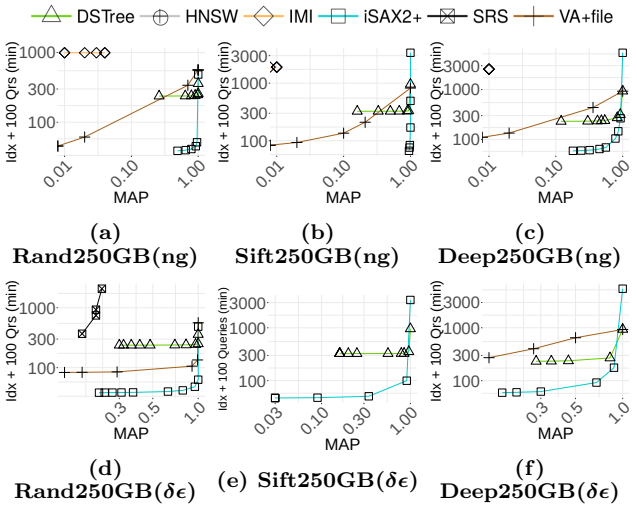


(a) Rand250GB(ng) (b) Sift250GB(ng) (c) Deep250GB(ng)

(d) Rand250GB($\delta\epsilon$) (e) Sift250GB($\delta\epsilon$) (f) Deep250GB($\delta\epsilon$)

Figure 18: Efficiency vs. accuracy (MRE) on disk (100NN queries)

Indexing and Answering 100 Queries vs. MAP. Figure 19 summarizes the results for 100NN queries on Rand250GB, Sift250GB and Deep250GB.

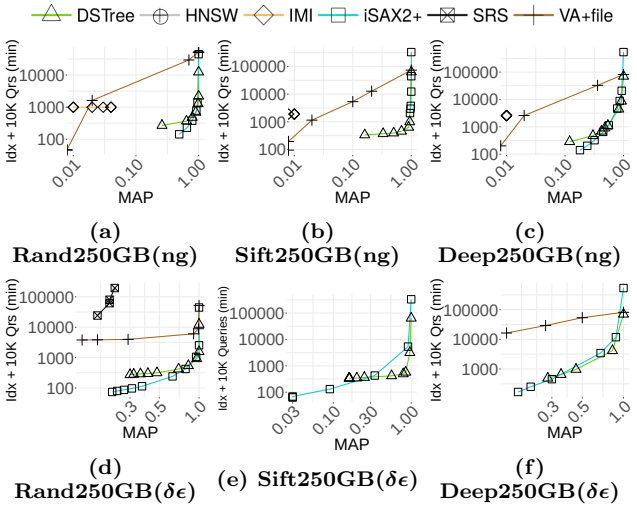


(a) Rand250GB(ng) (b) Sift250GB(ng) (c) Deep250GB(ng)

(d) Rand250GB($\delta\epsilon$) (e) Sift250GB($\delta\epsilon$) (f) Deep250GB($\delta\epsilon$)

Figure 19: Efficiency vs. accuracy on disk (Indexing + 100-100NN queries)

Indexing and Answering 10K Queries vs. MAP. Figure 20 summarizes the results for 100NN queries on Rand250GB, Sift250GB and Deep250GB.

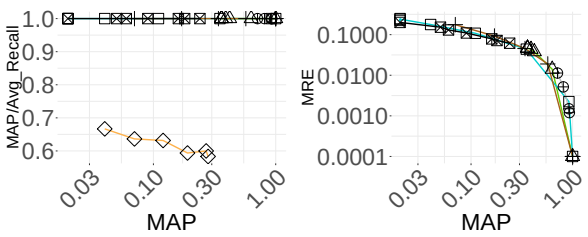


(a) Rand250GB(ng) (b) Sift250GB(ng) (c) Deep250GB(ng)

(d) Rand250GB($\delta\epsilon$) (e) Sift250GB($\delta\epsilon$) (f) Deep250GB($\delta\epsilon$)

Figure 20: Efficiency vs. accuracy on disk (Indexing + 10K-100NN queries)

Comparison of Accuracy Measures. Figures 21a and 21b compare all three measures for the popular real dataset Sift25GB (we use the 25GB subset to include in-memory methods as well).

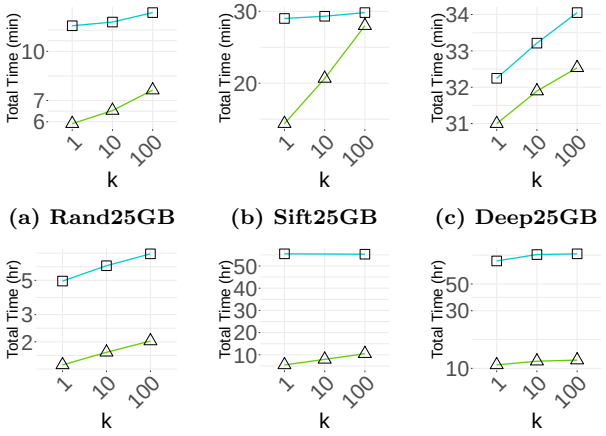


(a) Recall vs. MAP (b) MRE vs. MAP

Figure 21: Comparison of measures (Sift25GB)

1.5 Further Experiments with Best Methods

Effect of k. Figure 22 show the of k on the performance of the best methods.

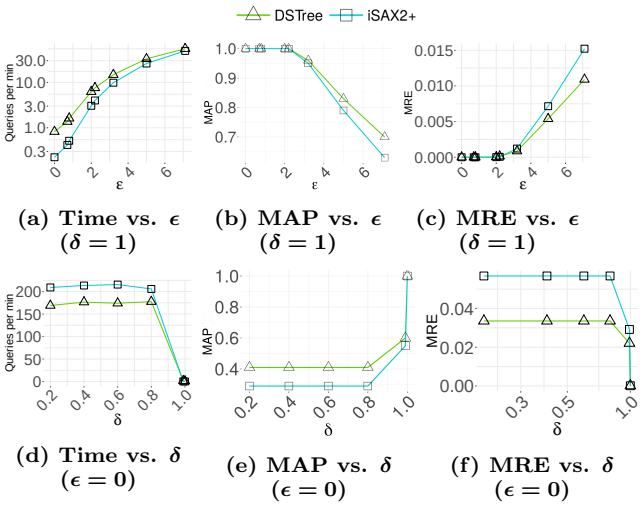


(a) Rand25GB (b) Sift25GB (c) Deep25GB

(d) Rand250GB (e) Sift250GB (f) Deep250GB

Figure 22: Efficiency vs. k (ϵ-approximate)

Effect of δ and ϵ. Figure 23 describes the effect of delta and epsilon on the performance of the best methods.

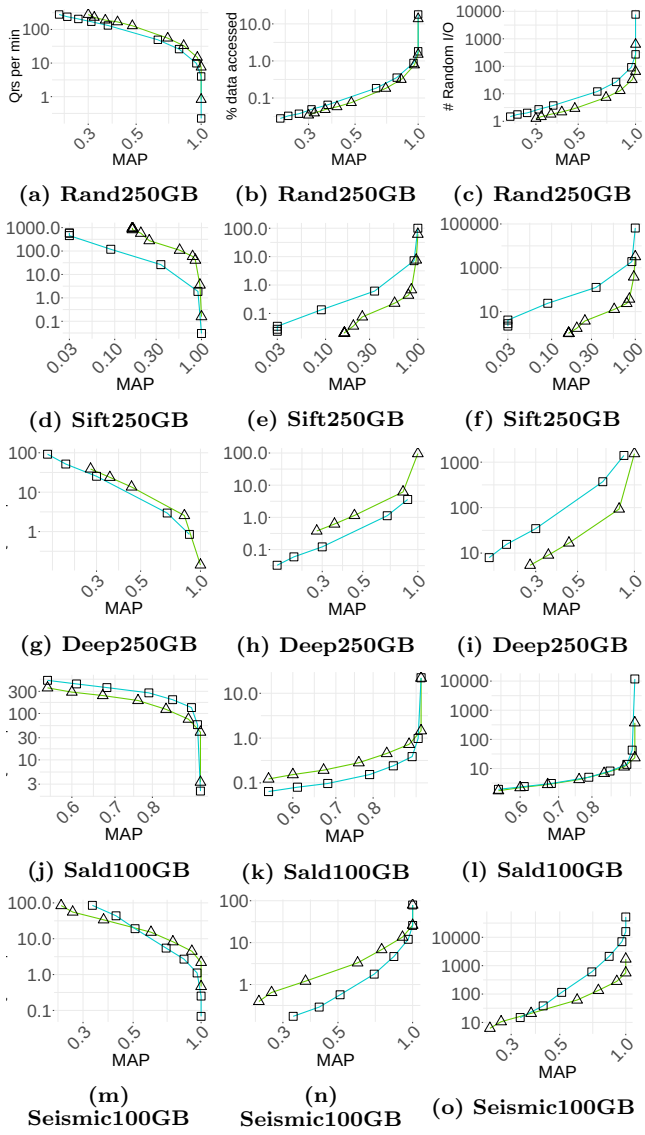


(a) Time vs. ϵ ($\delta = 1$) (b) MAP vs. ϵ ($\delta = 1$) (c) MRE vs. ϵ ($\delta = 1$)

(d) Time vs. δ ($\epsilon = 0$) (e) MAP vs. δ ($\epsilon = 0$) (f) MRE vs. δ ($\epsilon = 0$)

Figure 23: Accuracy and efficiency vs. δ and ϵ

Additional Datasets. Figure 24 summarizes further experiments on the additional real datasets Sald100GB and Seismic100GB and reports the number of random I/O operations and the percentage of data accessed for all disk-based datasets.



(a) Rand250GB (b) Rand250GB (c) Rand250GB

(d) Sift250GB (e) Sift250GB (f) Sift250GB

(g) Deep250GB (h) Deep250GB (i) Deep250GB

(j) Sald100GB (k) Sald100GB (l) Sald100GB

(m) Seismic100GB (n) Seismic100GB (o) Seismic100GB

Figure 24: Efficiency vs. accuracy for the best methods (ϵ-approximate)