A Performance Comparison of General Purpose Multi-Dimensional In-Memory Indexes – All Results

Revision 1.3 – 4th April 2018

Tilmann Zäschke
zoodb@gmx.de
zaeschke@inf.ethz.ch

1. INTRODUCTION

This document contains all TinSpin[1] test results from the test runs between November 2016 and January 2017.

1.1 Revisions

- Rev. 1.0 2017-01-28 Initial version.
- Rev. 1.1 2017-09-18 Added brief section on data.
- Rev. 1.2 2018-02-26 Fixed labels in Fig. 16 and 17.
- Rev. 1.3 2018-04-04 Numerous textual improvements.

2. OVERVIEW

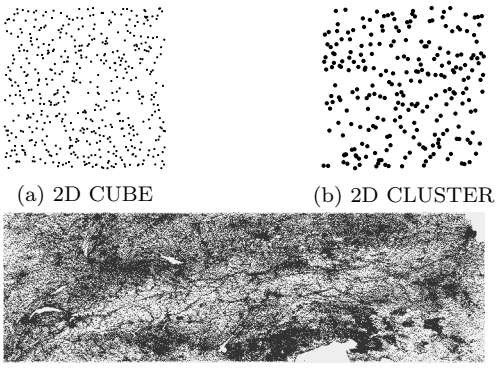
The following index implementations were tested:

- CBF CritBit tree by J. Fager[2]
- CBZ CritBit tree by T. Zäschke[3]
- KDL KD-Tree by Levy[4]
- KDS KD-Tree by Savarese[5]
- PH/PHM PH-Tree by T. Zäschke et al.[6]
- QTZ Quadtree by T. Zäschke[3]
- RSS R*Tree by N. Beckmann et al[7], optimized for in-memory use by T. Zäschke
- RSZ R*Tree by T. Zäschke[3]
- STRZ Sort-tile-recursive loaded R-Tree by T. Zäschke[3]
- XTS X-Tree by S. Berchtold et al[7], optimized for in-memory use by T. Zäschke

2.1 Terminology

- d: Number of dimensions
- N: Size of the dataset
- k: Number of requested nearest neighbors

[1] http://www.tinspin.org
[2] https://github.com/jfager/functional-critbit
[3] https://github.com/tzaeschke/tinspin-indexes
[4] http://home.wlu.edu/~levys/software/kd/
[5] http://www.savarese.com/software/libssrckdtree-j
[6] http://www.phtree.org
[7] http://chorochronos.datastories.org



(a) 2D CUBE (b) 2D CLUSTER

(c) OSM Alps

Figure 1: The CUBE, CLUSTER and OSM Alps datasets

3. TEST SET-UP

3.1 Test Data

The OSM-P (points) and OSM-R (rectangles) datasets are extracts from OpenStreetMap.org representing the European Alps[8], extracted on 2016-11-09. It ranges from Vienna in the north east to almost Grenoble in the south west, thus including major point clusters (cities) such as Vienna, Munich and Zurich (Fig. 1c). The dataset consists of $\approx 2.1 \times 10^8$ points. Geographically it extends between about min/max longitude=3.931094/20.2583918 and latitude=37.7126446/49.1369103. The rectangles (OSM-R) are bounding boxes for all line segments in the dataset.

The synthetic CU-P/CU-R datasets (Fig. 1a), have the shape of a cube filled with up to 50,000,000 elements that are distributed uniformly at random between 0.0 and 1.0 in every dimension. Each element has unique coordinates.

The synthetic CL-P/CL-R datasets (Fig. 1b) consists of 1000 clusters that are distributed uniformly at random between 0.0 and 1.0. In each cluster, elements follow a Gaussian distribution with standard deviation $\sigma = 0.001$. The CLUSTER dataset contains up to 50,000,000 elements.

All data in CU and CL is generated randomly, however all tests use the same sets of randomly generated data. All datasets have duplicate points/rectangles removed.

3.2 Test Execution

All tests were executed with the TinSpin framework. The frameworks executes all tests three times with different datasets, the graphs show the averaged results.

Tests were executed while varying dataset size between $N = 5 \times 10^5$ and $N = 5 \times 10^7$ using 2D data (OSM) or 3D data (CU, CL). We also varied dimensionality while testing CU and CL datasets with $N = 10^6$ and $2 \leq d \leq 40$ for point data and $2 \leq d \leq 28$ for rectangle data. Most window queries were created such that they return on average 1000 entries. The only exception are the tests with varying window query size, which were done with $N = 10^6$ and $d = 2$ (OSM) or $d = 3$ (CL and CU).

The experiments were executed on a desktop PC with 32GB RAM and an Intel i7-4790K 4.00GHz CPU with 4 cores (8 logical processors). All algorithms are implemented in Java and ran on Oracle JDK 1.8.0_51 64bit with -Xmx28G -XX:+UseConcMarkSweepGC.

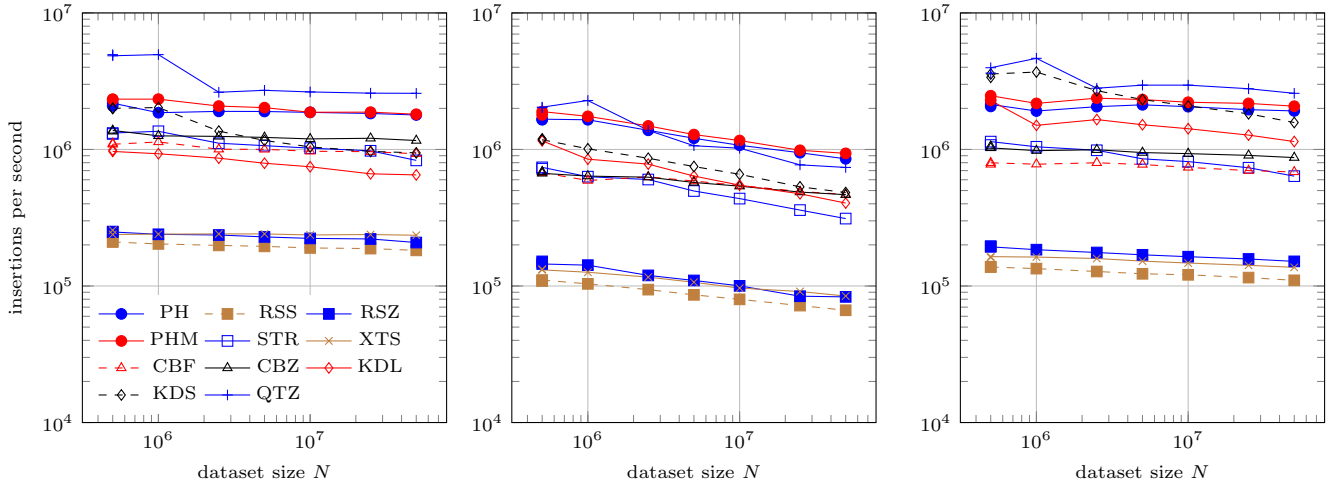
[8] `http://download.geofabrik.de/europe/alps.html`

For a detailed description of the tests please contact the author via zoodb@gmx.de.

4. RESULTS

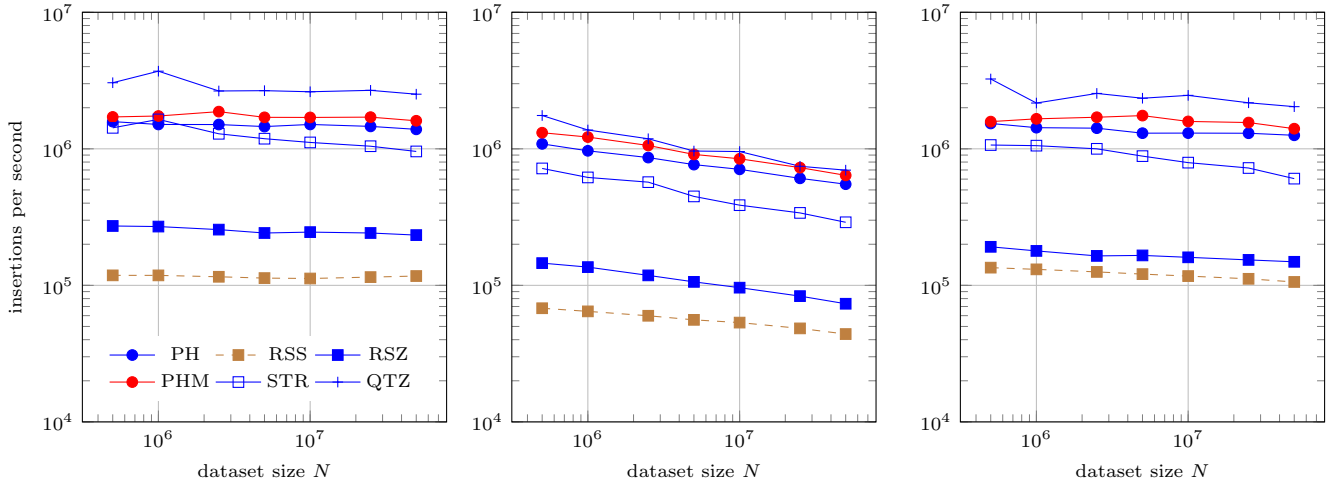
Results are shown on in the following order:

- Insertion
  - Dataset size N: Figures 2 – 3
  - Dimensionality d: Figures 4 – 7
- Memory usage
  - Dataset size N: Figures 8 – 9
  - Dimensionality d: Figures 10 – 13
- Window queries
  - Dataset size N: Figures 14 – 15
  - Query result size: Figures 16 – 17
  - Dimensionality d: Figures 18 – 21
- Exact match queries (point queries)
  - Dataset size N: Figures 22 – 23
  - Dimensionality d: Figures 24 – 27
- kNN queries
  - Dataset size N: Figures 28 – 31
  - Dimensionality d: Figures 32 – 39
- Update
  - Dataset size N: Figures 40 – 41
  - Dimensionality d: Figures 42 – 45
- Remove
  - Dataset size N: Figures 46 – 47
  - Dimensionality d: Figures 48 – 51



(a) Alps OSM-P (b) CUBE-P (c) CLUSTER-P

Figure 2: Point insertion rates for the OSM ($d = 2$), CUBE ($d = 3$) and CLUSTER ($d = 3$) datasets



(a) OSM-R (b) CUBE-R (c) CLUSTER-R

Figure 3: Rectangle insertion rates for the OSM ($d = 2$), CUBE ($d = 3$) and CLUSTER ($d = 3$) datasets

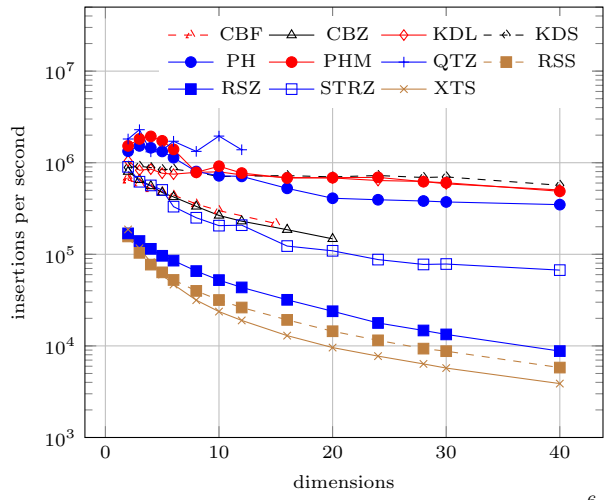


Figure 4: DIM: Insertion rates for CU-P with $N = 10^6$

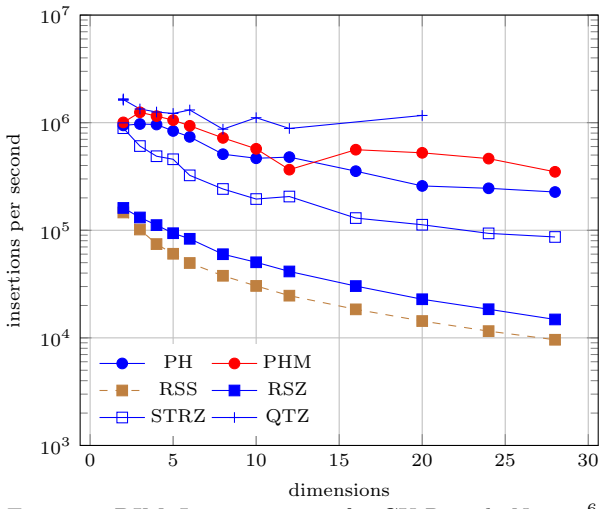


Figure 5: DIM: Insertion rates for CU-R with $N = 10^6$

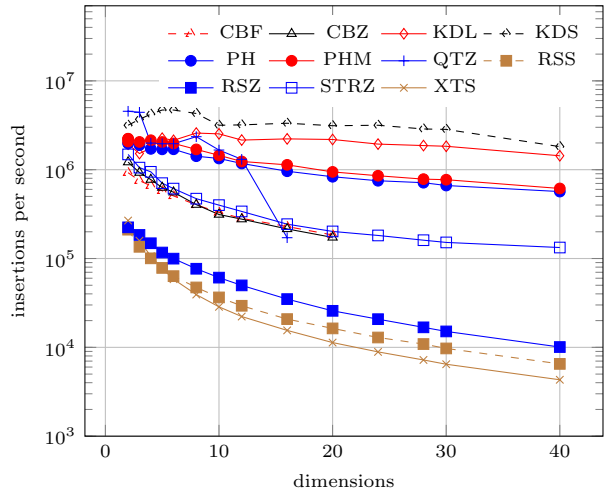


Figure 6: DIM: Insertion rates for CL-P with $N = 10^6$

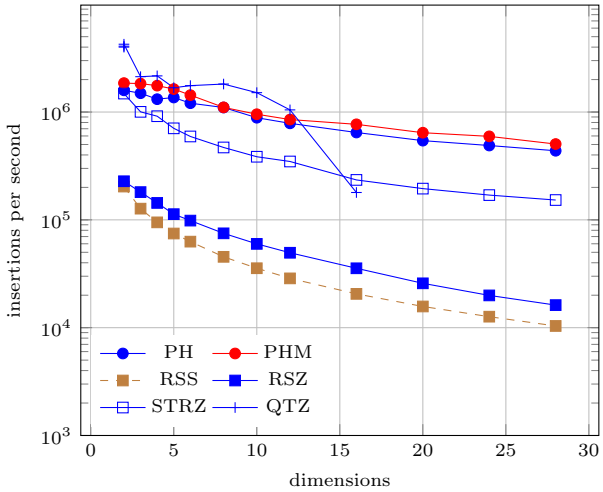
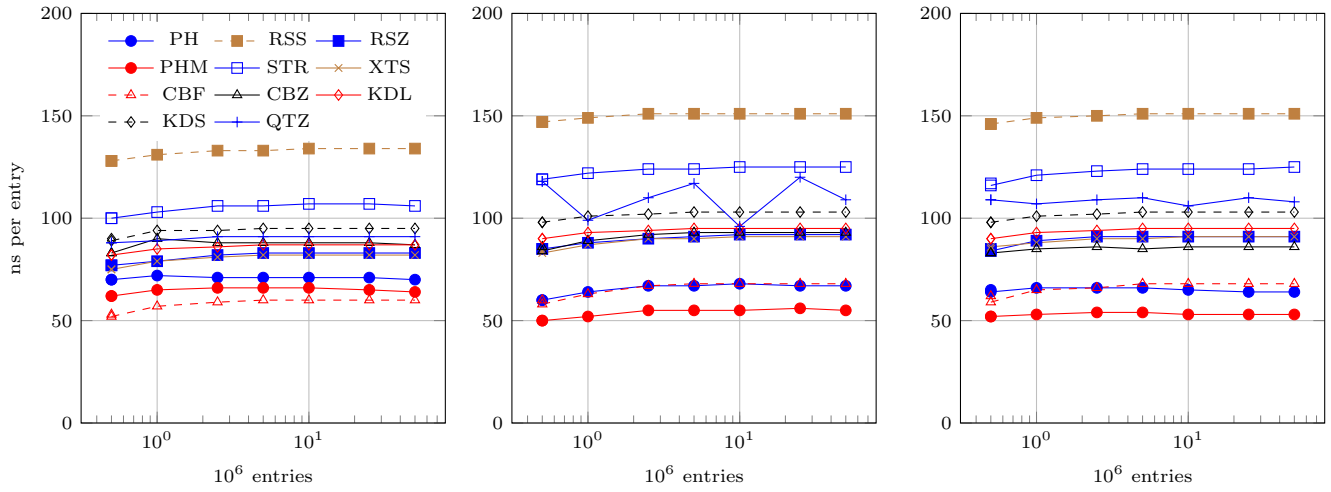
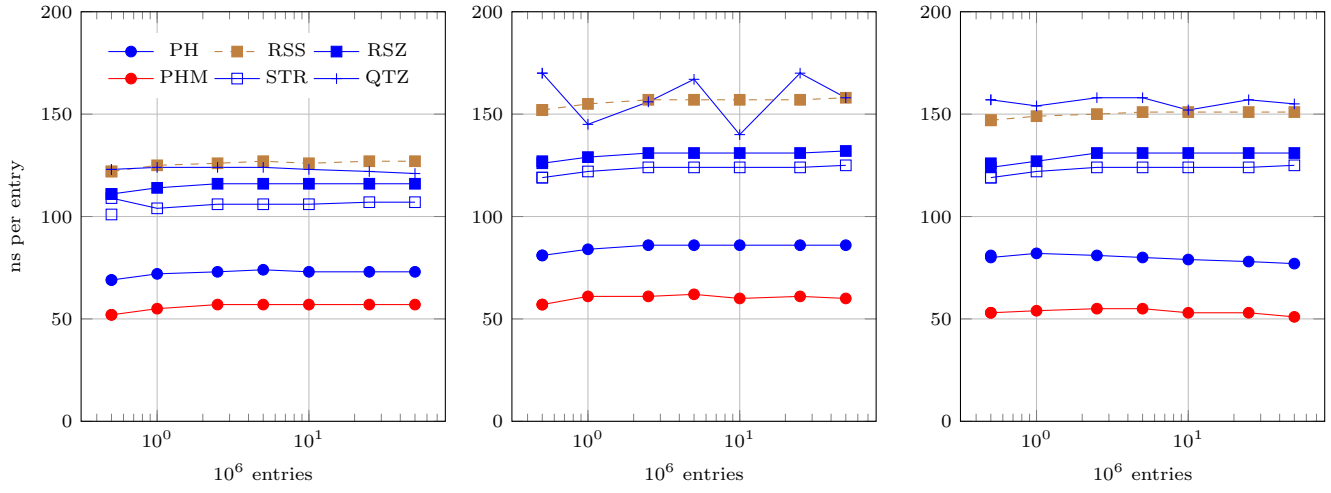


Figure 7: DIM: Insertion rates for CL-R with $N = 10^6$



(a) Alps OSM-P (b) CUBE-P (c) CLUSTER-P

Figure 8: Memory usage per point entry for the OSM ($d = 2$), CUBE ($d = 3$) and CLUSTER ($d = 3$) datasets



(a) Alps OSM-R (b) CUBE-R (c) CLUSTER-R

Figure 9: Memory usage per rectangle entry for the OSM ($d = 2$), CUBE ($d = 3$) and CLUSTER ($d = 3$) datasets

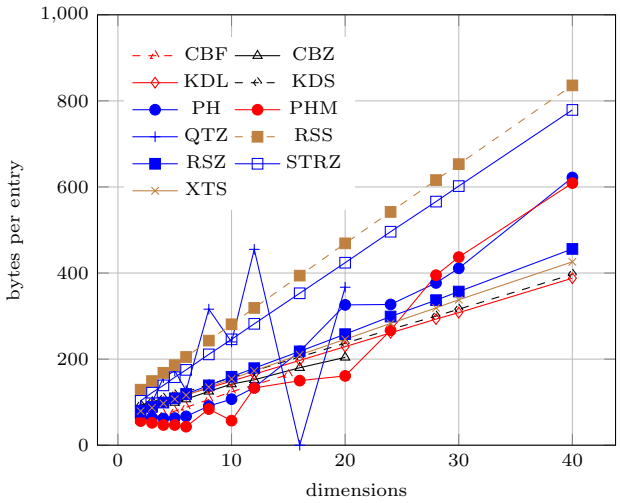


Figure 10: DIM: Memory usage per point for CU-P with $N = 10^6$

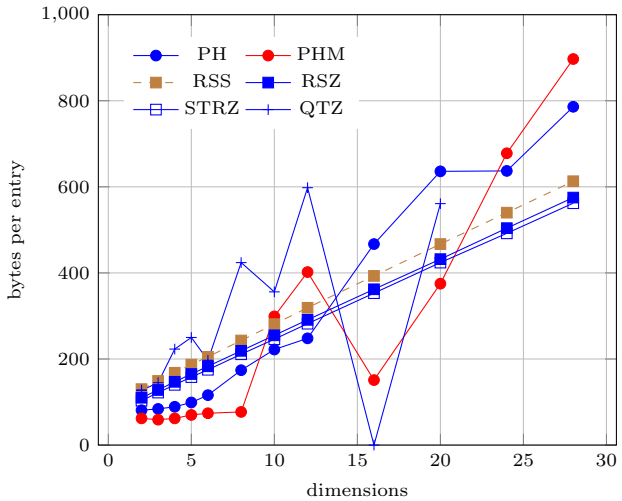


Figure 11: DIM: Memory usage per rectangle for CU-R with $N = 10^6$

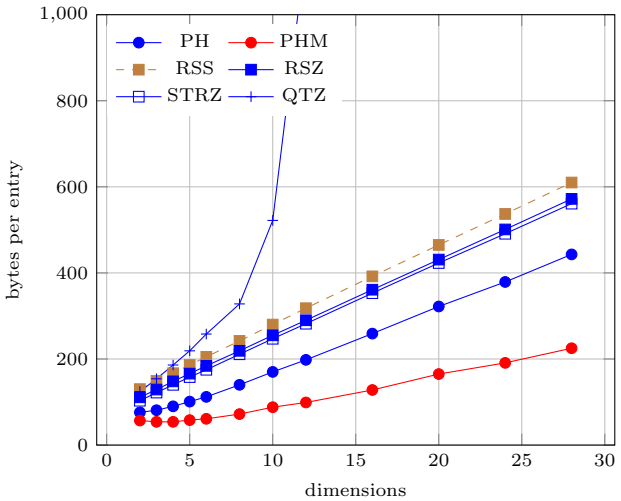


Figure 13: DIM: Memory usage per rectangle for CL-R with $N = 10^6$

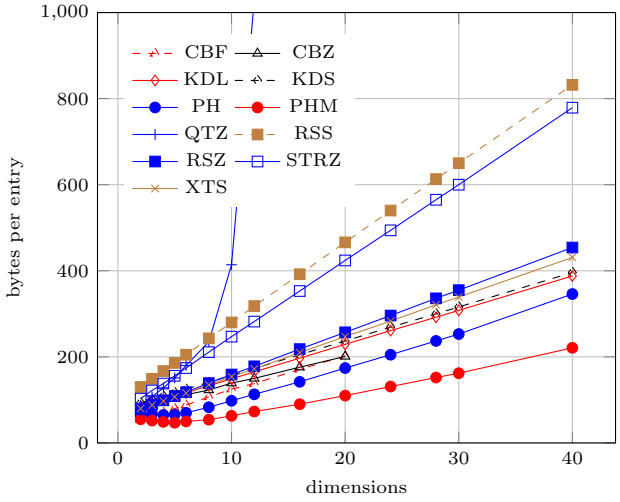
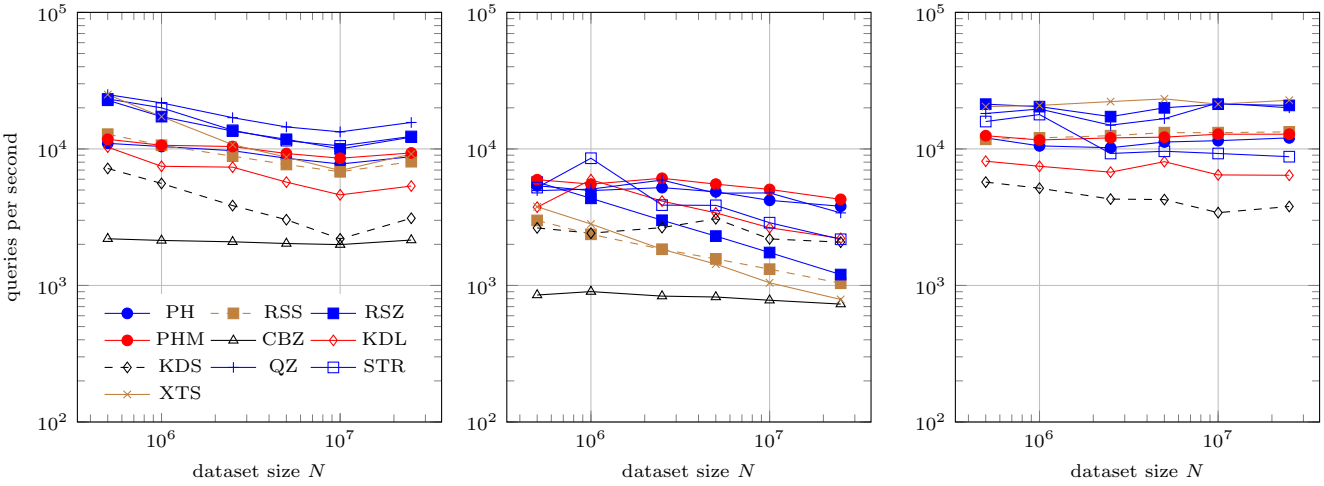
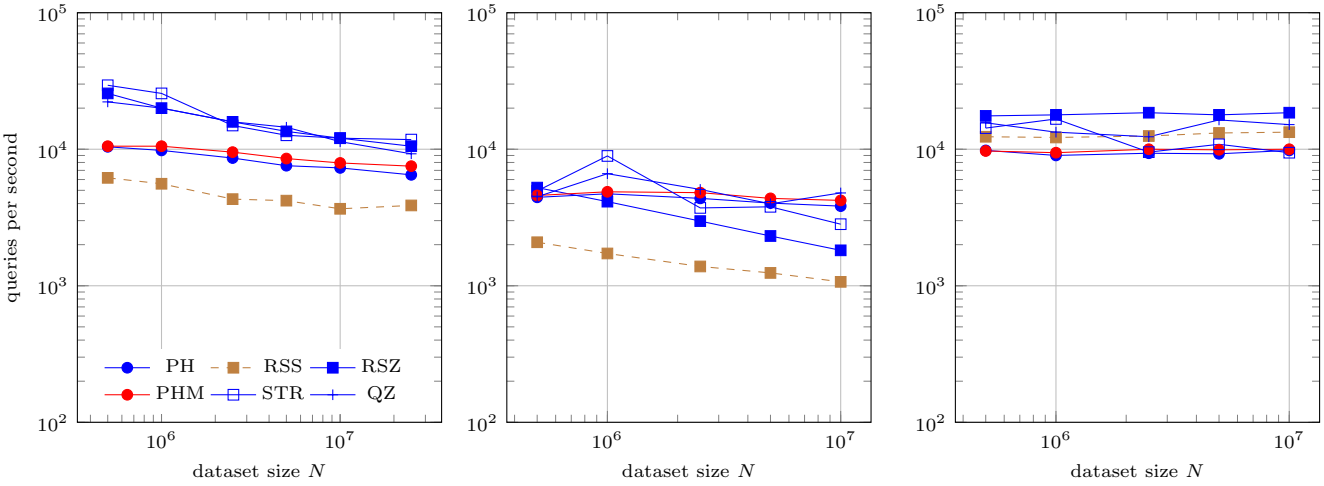


Figure 12: DIM: Memory usage per point for CL-P with $N = 10^6$



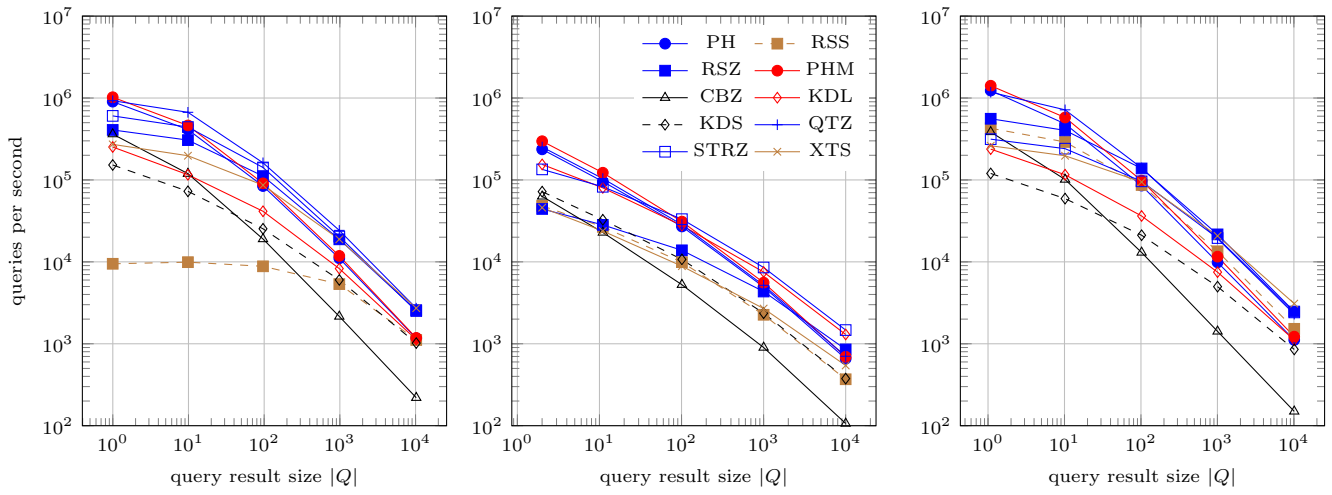
(a) 2D OSM-P (b) 3D CUBE-P (c) 3D CLUSTER-P

Figure 14: Window query rates for the OSM ($d = 2$), CUBE ($d = 3$) and CLUSTER ($d = 3$) datasets



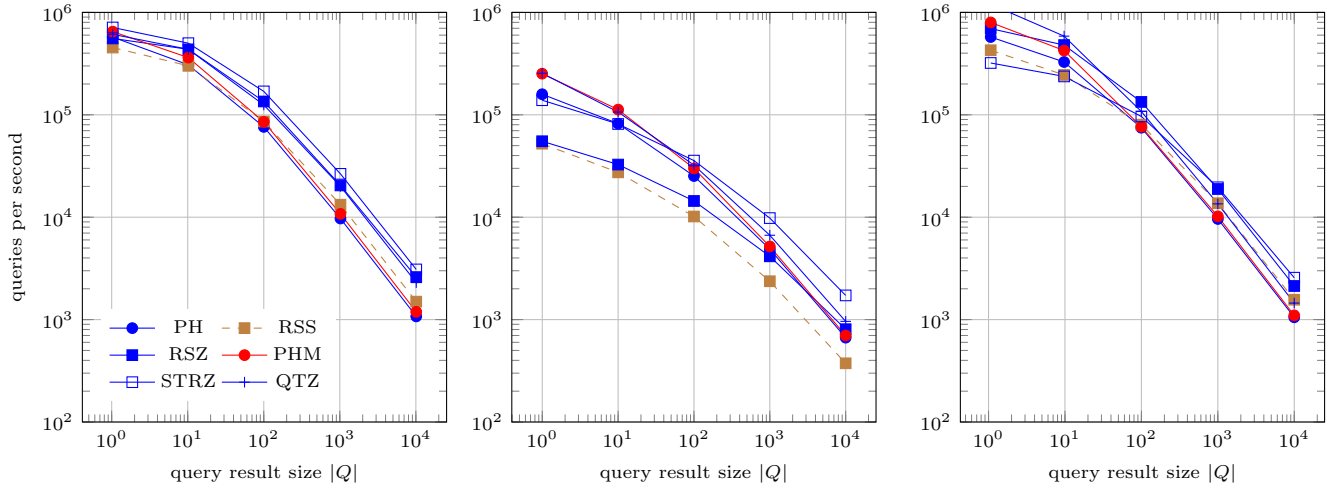
(a) Alps OSM-R (b) CUBE-R (c) CLUSTER-R

Figure 15: Window query rates for the OSM ($d = 2$), CUBE ($d = 3$) and CLUSTER ($d = 3$) datasets



(a) 2D OSM-P (b) 3D CUBE-P (c) 3D CLUSTER-P

Figure 16: Varying query result size with $N = 10^6$ with the OSM ($d = 2$), CUBE ($d = 3$) and CLUSTER ($d = 3$) datasets



(a) Alps OSM-R (b) CUBE-R (c) CLUSTER-R

Figure 17: Varying query result size with $N = 10^6$ with the OSM ($d = 2$), CUBE ($d = 3$) and CLUSTER ($d = 3$) datasets

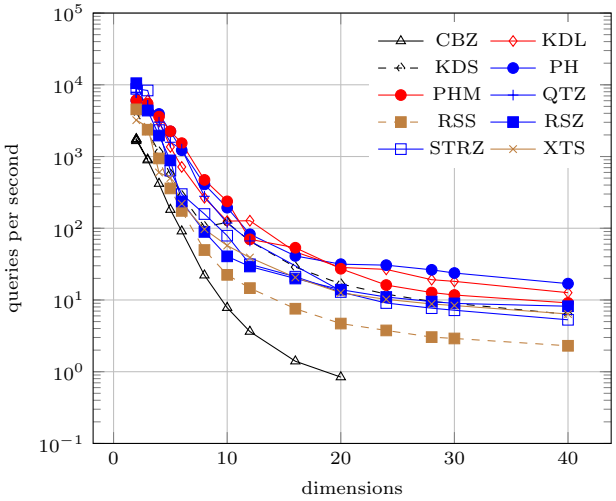


Figure 18: DIM: Window query rates for CU-P with $N = 10^6$

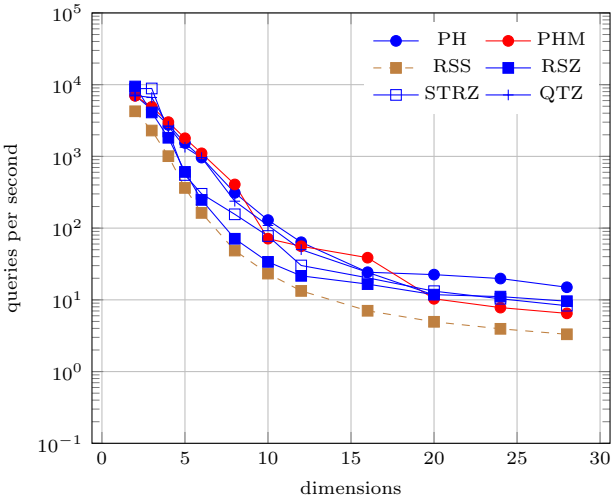


Figure 19: DIM: Window query rates for CU-R with $N = 10^6$

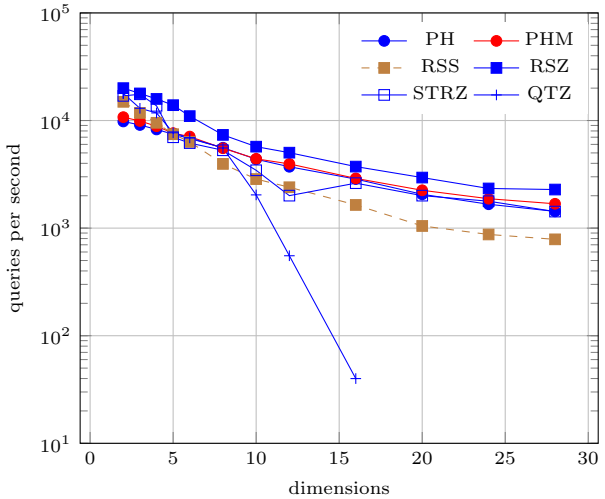


Figure 21: DIM: Window query rates for CL-R with $N = 10^6$

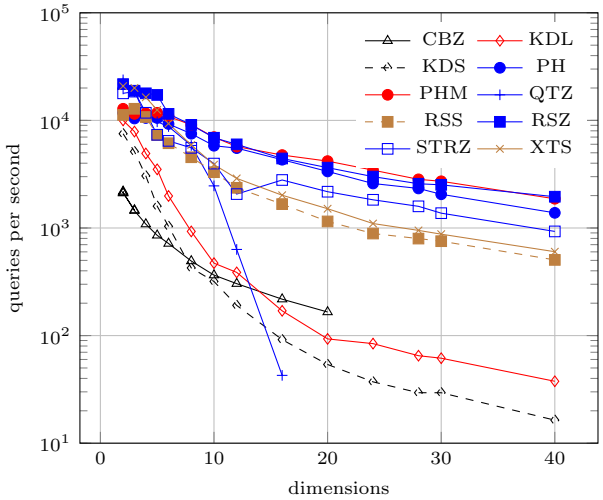
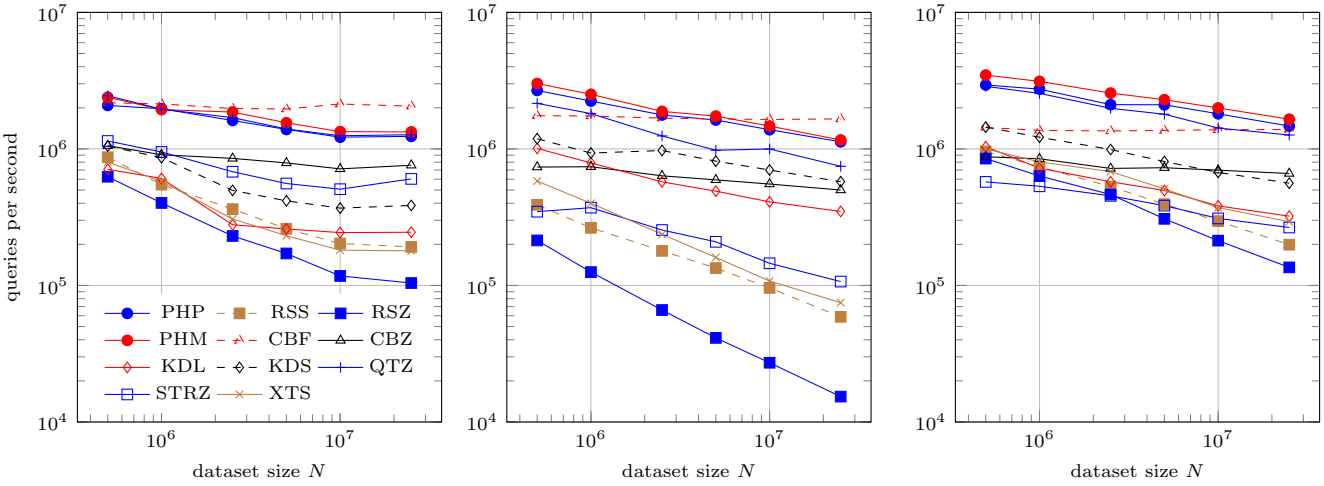
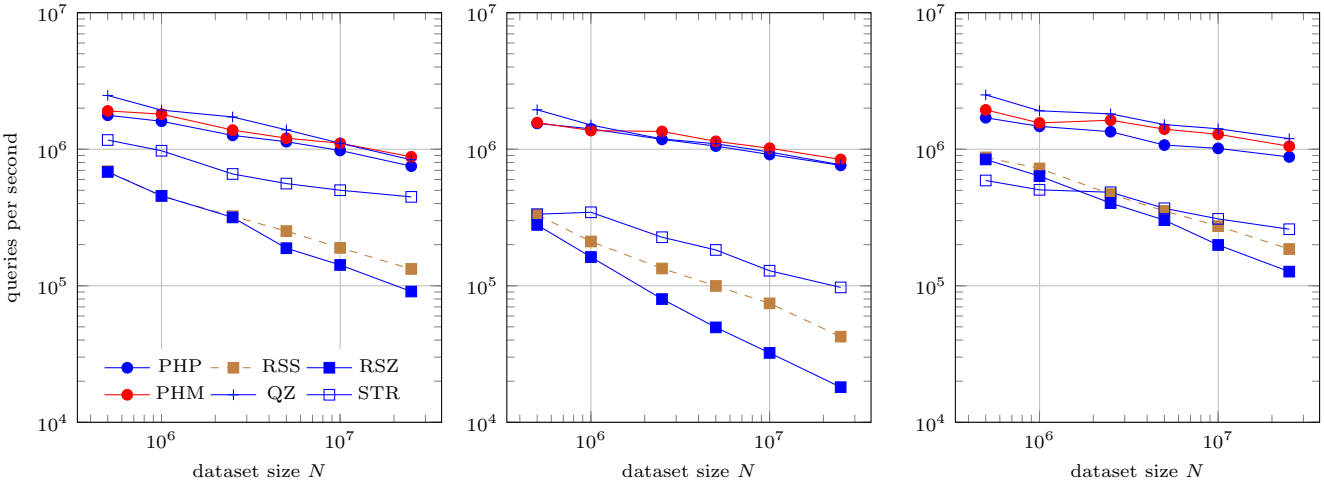


Figure 20: DIM: Window query rates for CL-P with $N = 10^6$



(a) 2D OSM-P (b) 3D CUBE-P (c) 3D CLUSTER-P

Figure 22: Exact match query rates for the OSM ($d = 2$), CUBE ($d = 3$) and CLUSTER ($d = 3$) datasets



(a) 2D OSM-R (b) 3D CUBE-R (c) 3D CLUSTER-R

Figure 23: Exact match query rates for the OSM ($d = 2$), CUBE ($d = 3$) and CLUSTER ($d = 3$) datasets

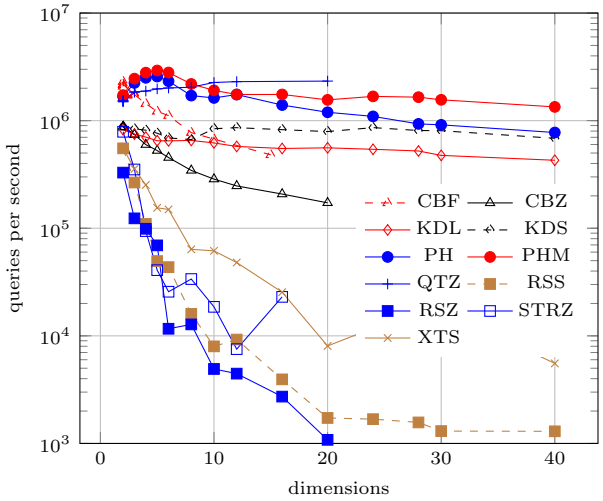


Figure 24: DIM: Exact match query rates for CU-P with $N = 10^6$

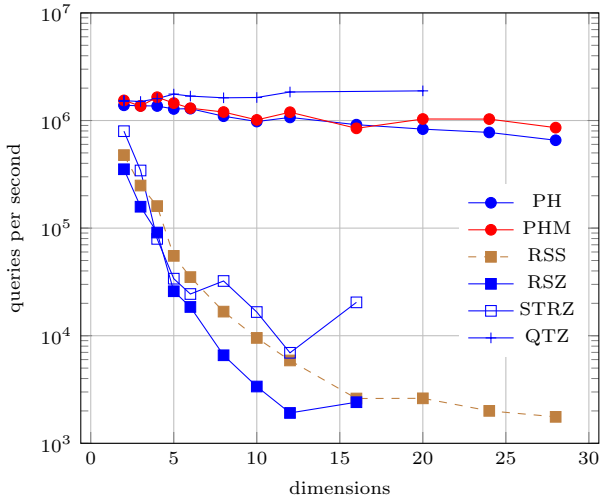


Figure 25: DIM: Exact match query rates for CU-R with $N = 10^6$

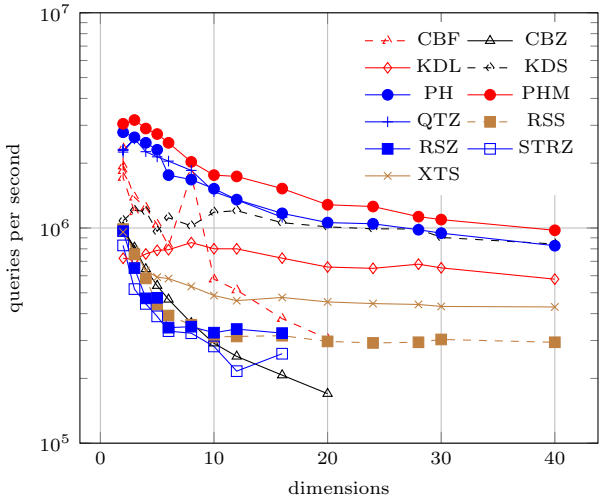


Figure 26: DIM: Exact match query rates for CL-P with $N = 10^6$

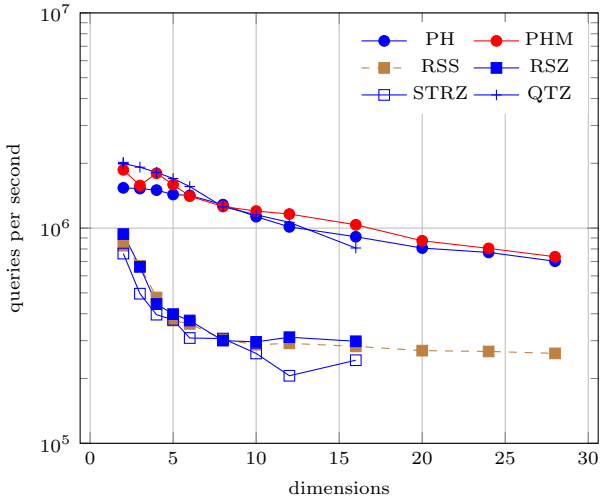
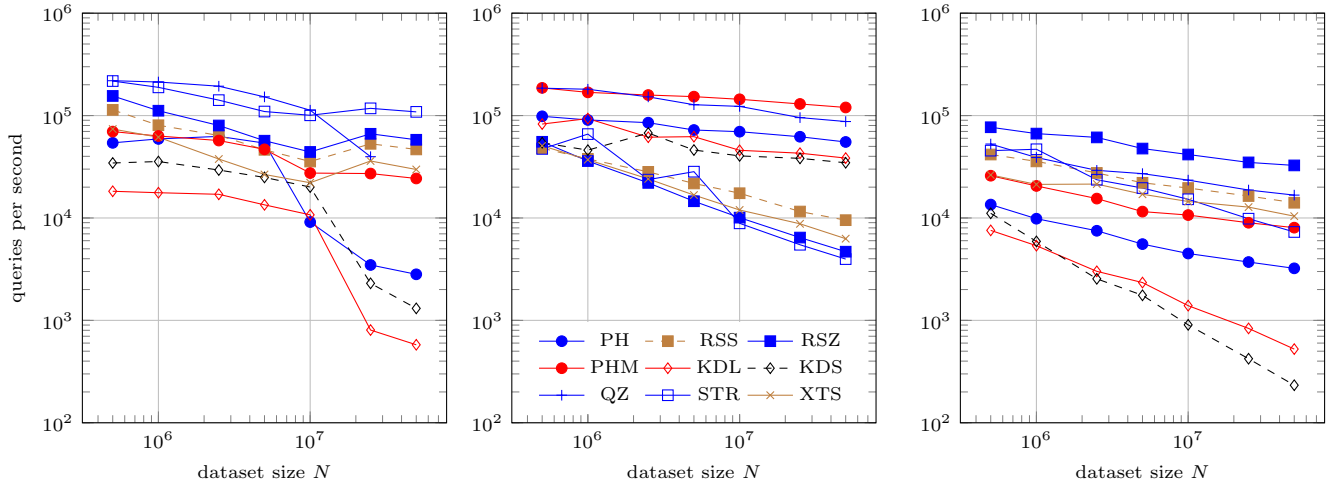
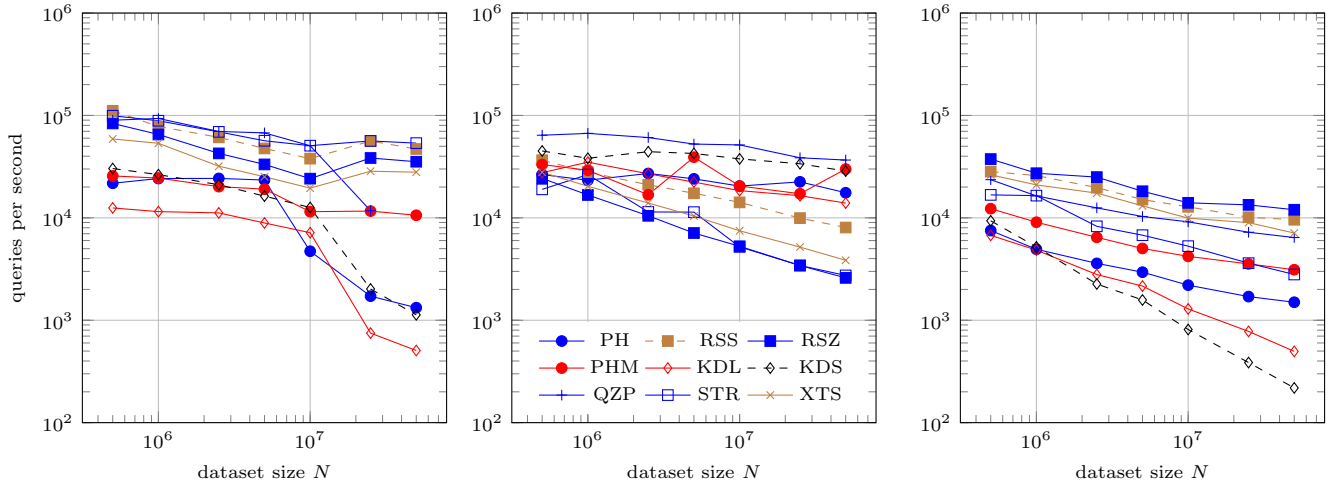


Figure 27: DIM: Exact match query rates for CL-R with $N = 10^6$



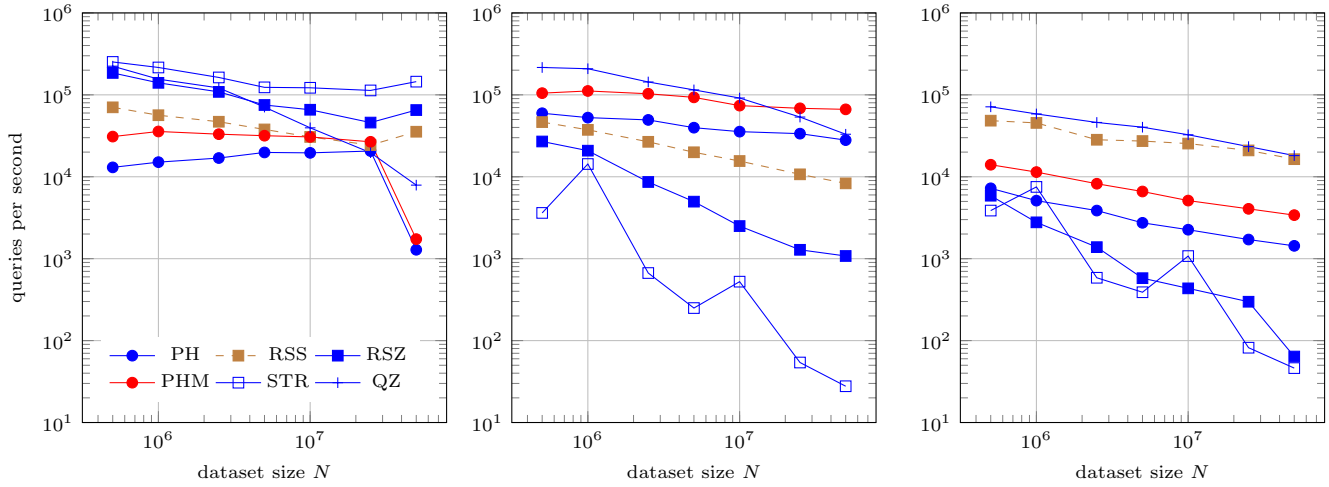
(a) 2D OSM-P (b) 3D CUBE-P (c) 3D CLUSTER-P

Figure 28: 1NN query rates for the OSM ($d = 2$), CUBE ($d = 3$) and CLUSTER ($d = 3$) datasets



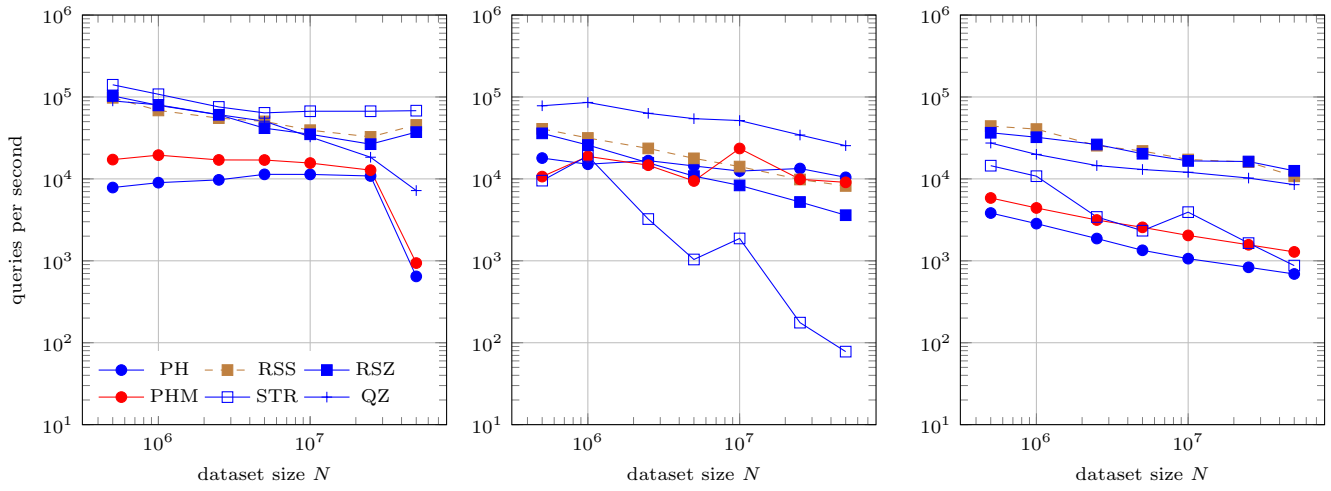
(a) 2D OSM-P (b) 3D CUBE-P (c) 3D CLUSTER-P

Figure 29: 10NN query rates for the OSM ($d = 2$), CUBE ($d = 3$) and CLUSTER ($d = 3$) datasets



(a) OSM-R (b) CUBE-R (c) CLUSTER-R

Figure 30: 1NN query rates for the OSM ($d = 2$), CUBE ($d = 3$) and CLUSTER ($d = 3$) datasets



(a) OSM-R (b) CUBE-R (c) CLUSTER-R

Figure 31: 10NN query rates for the OSM ($d = 2$), CUBE ($d = 3$) and CLUSTER ($d = 3$) datasets

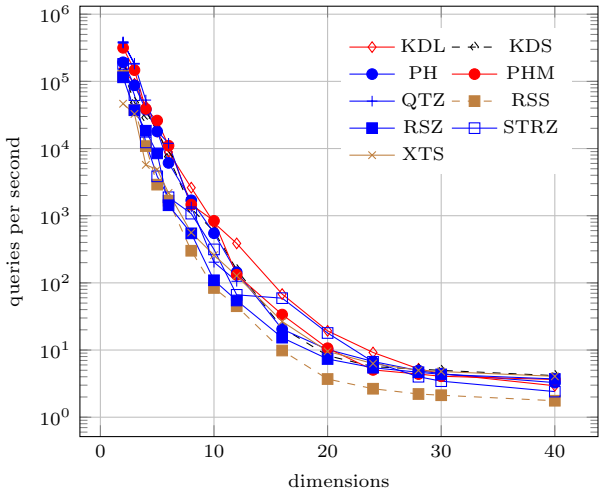


Figure 32: DIM: 1-NN query rates for CU-P with $N = 10^6$

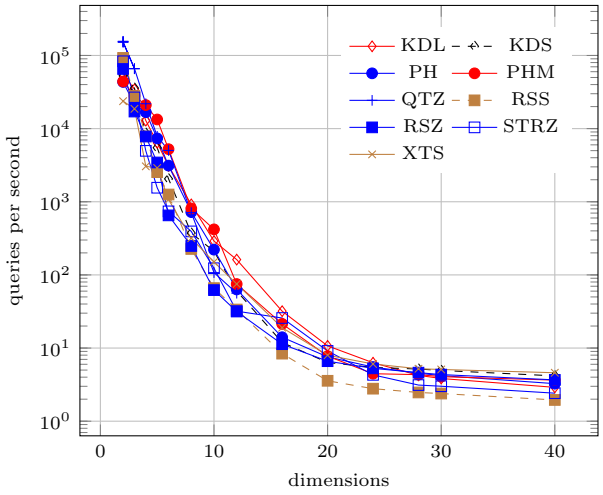


Figure 33: DIM: 10-NN query rates for CU-P with $N = 10^6$

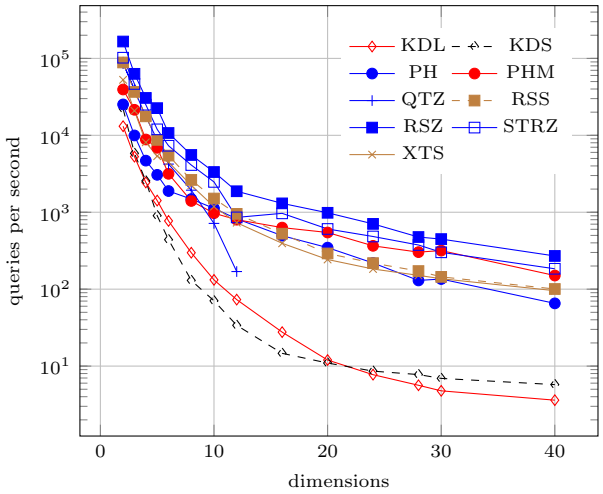


Figure 34: DIM: 1-NN query rates for CL-P with $N = 10^6$

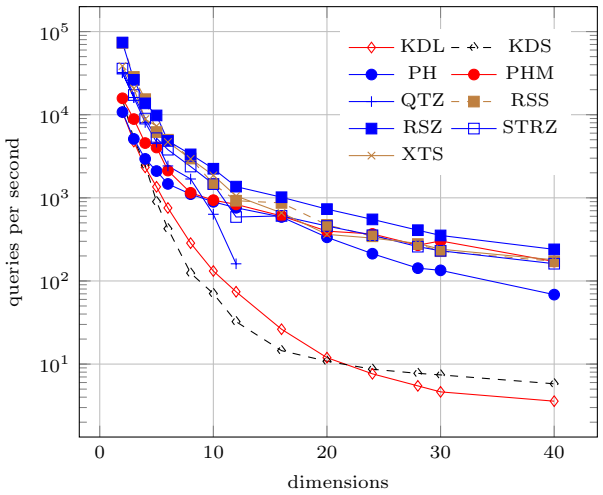


Figure 35: DIM: 10-NN query rates for CL-P with $N = 10^6$

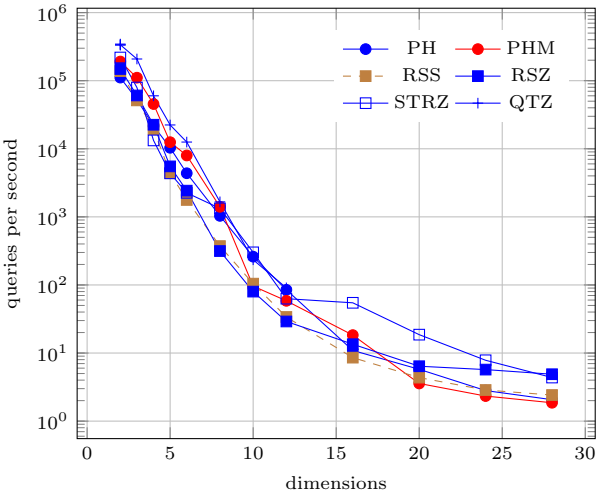


Figure 36: DIM: 1-NN query rates for CU-R with $N = 10^6$

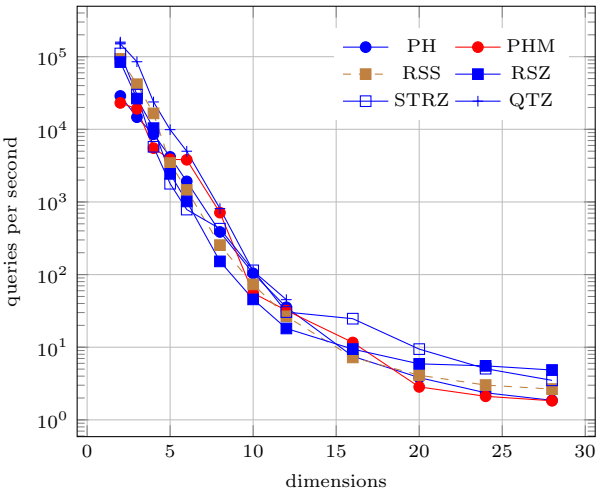


Figure 37: DIM: 10-NN query rates for CU-R with $N = 10^6$

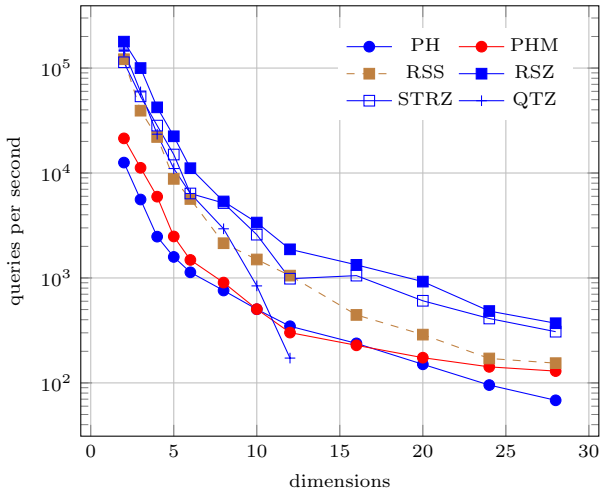


Figure 38: DIM: 1-NN query rates for CL-R with $N = 10^6$

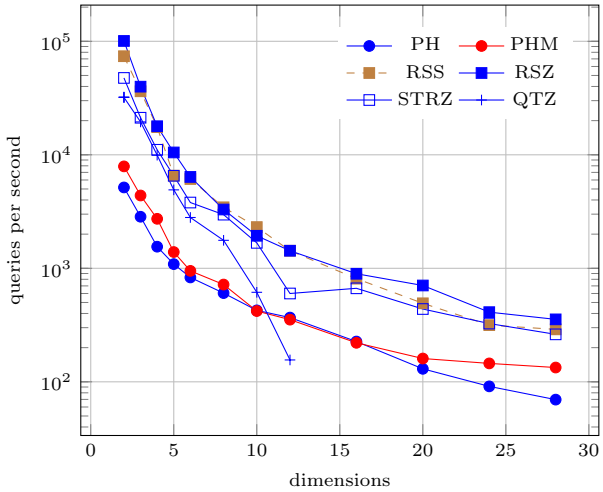
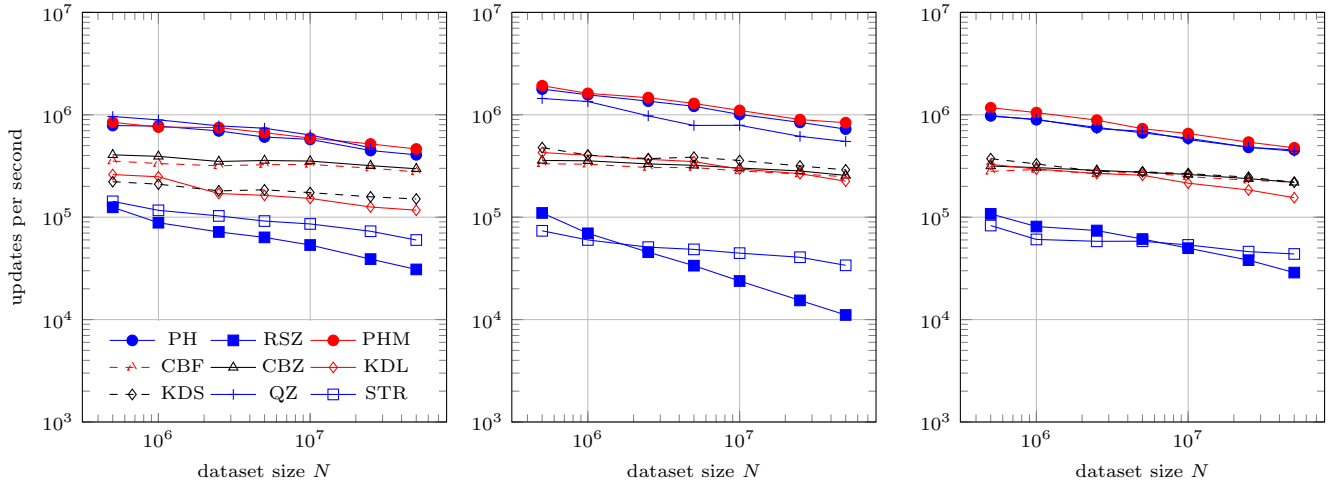
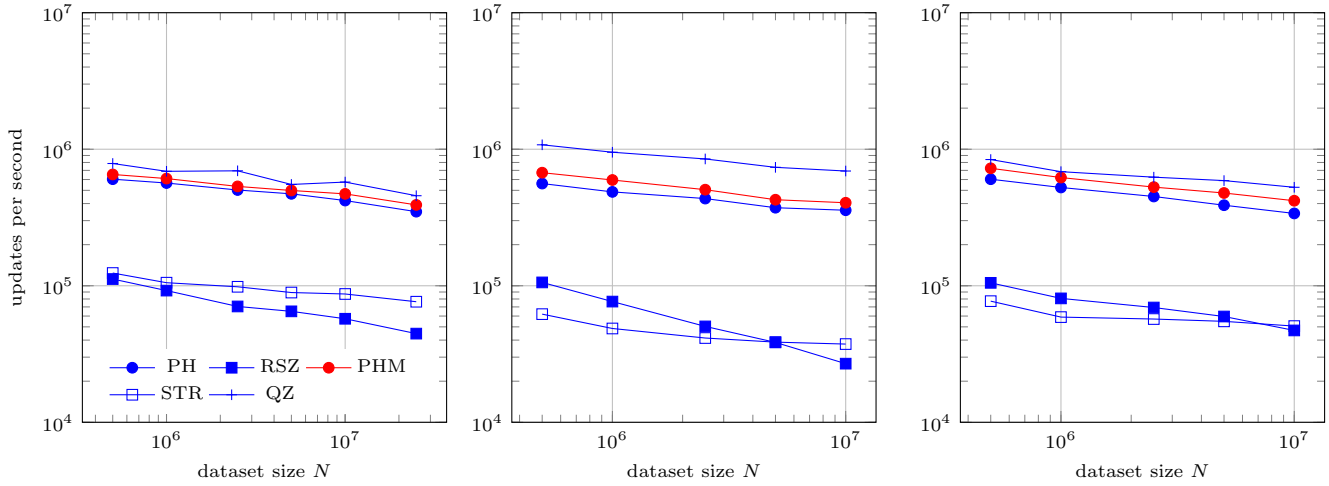


Figure 39: DIM: 10-NN query rates for CL-R with $N = 10^6$



(a) 2D OSM-P (b) 3D CUBE-P (c) 3D CLUSTER-P

Figure 40: Update rates for the OSM ($d = 2$), CUBE ($d = 3$) and CLUSTER ($d = 3$) datasets



(a) 2D OSM-R (b) 3D CUBE-R (c) 3D CLUSTER-R

Figure 41: Update rates for the OSM ($d = 2$), CUBE ($d = 3$) and CLUSTER ($d = 3$) datasets

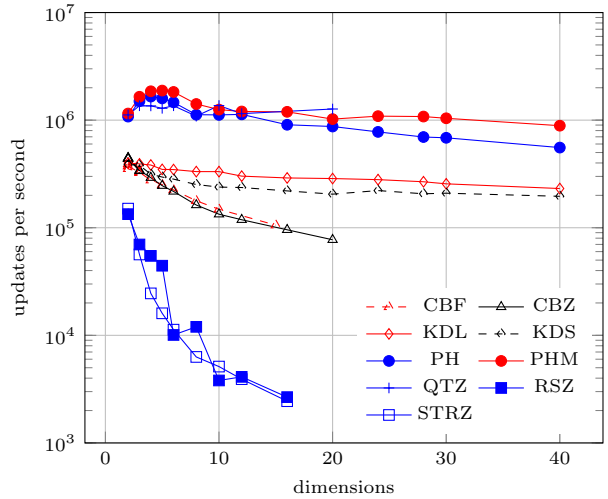


Figure 42: DIM: Update rates for CU-P with $N = 10^6$

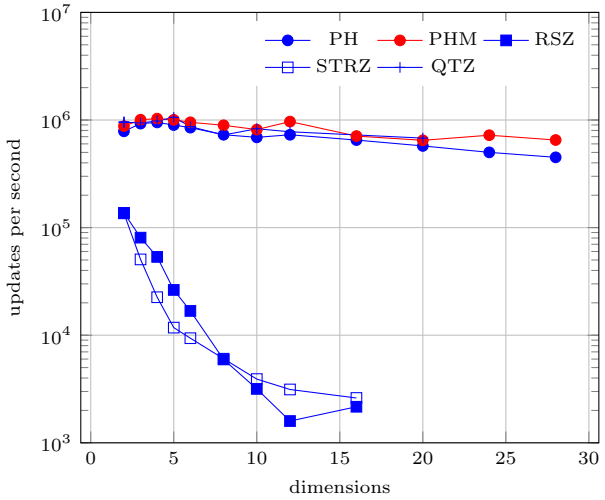


Figure 43: DIM: Update rates for CU-R with $N = 10^6$

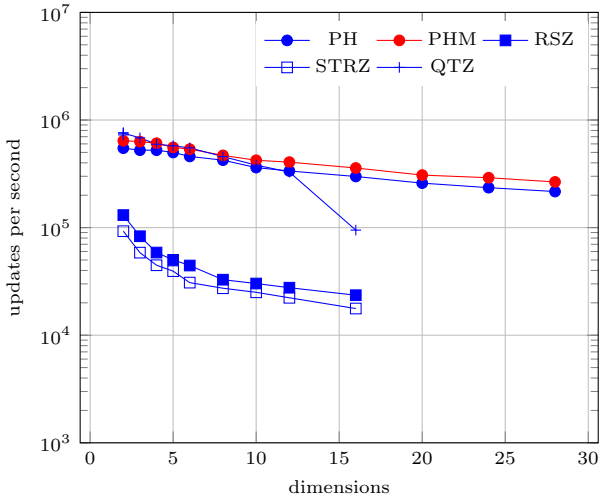


Figure 45: DIM: Update rates for CL-R with $N = 10^6$

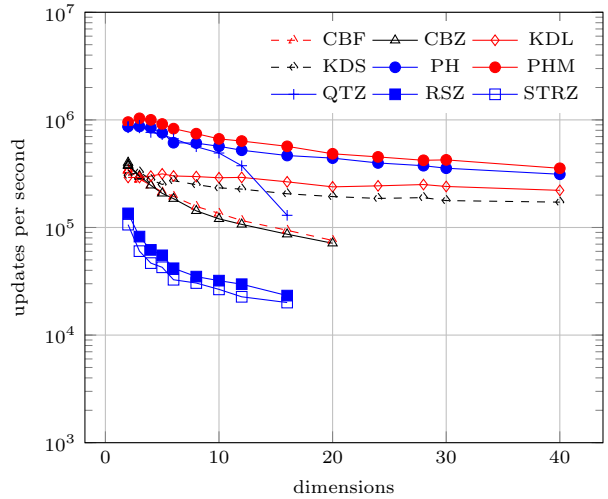
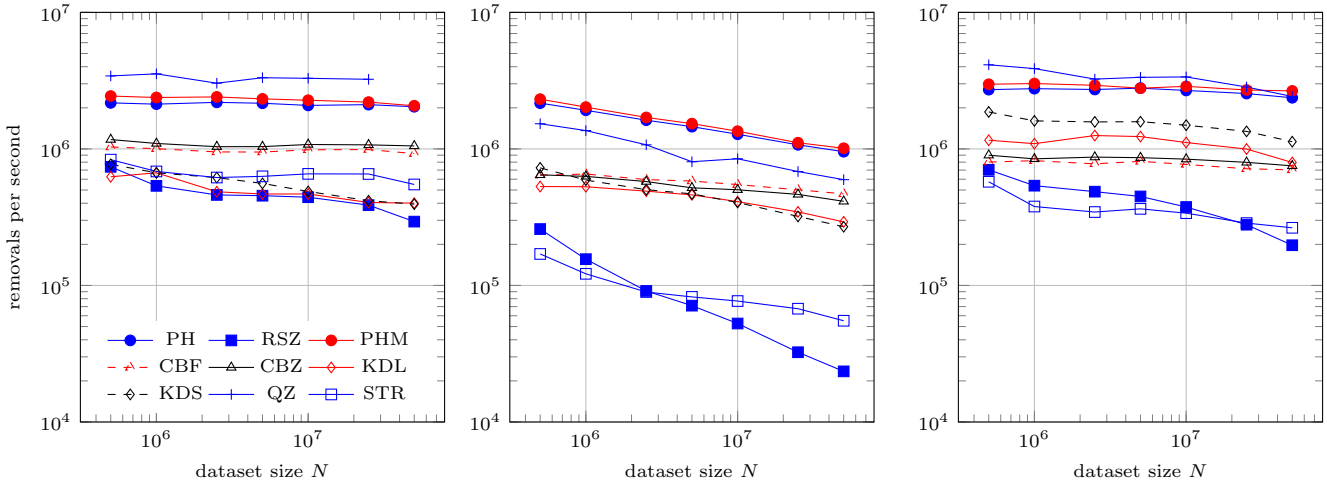
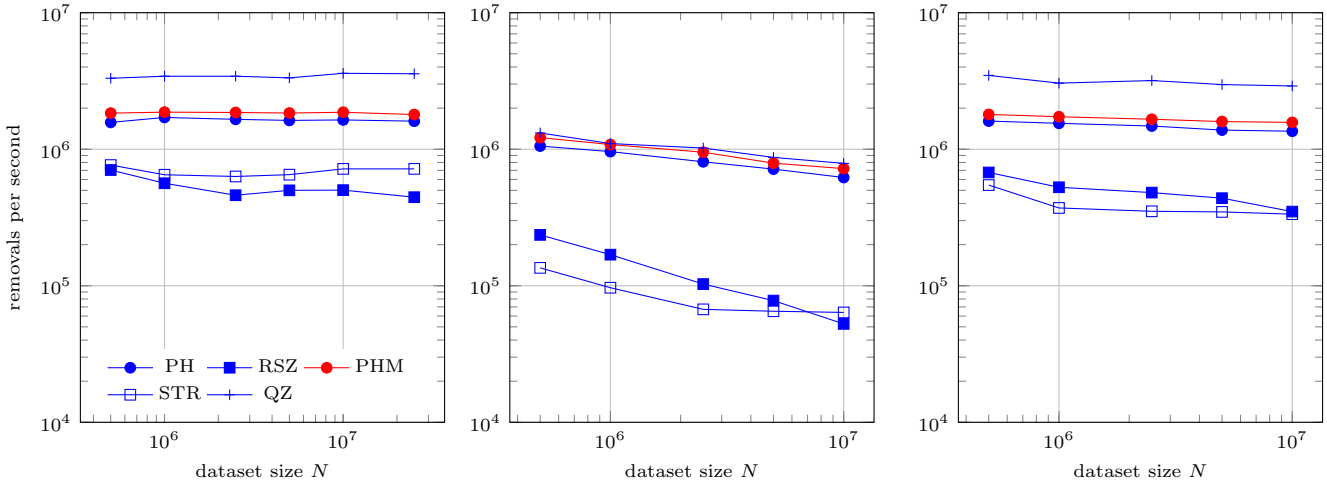


Figure 44: DIM: Update rates for CL-P with $N = 10^6$



(a) 2D OSM-P (b) 3D CUBE-P (c) 3D CLUSTER-P

Figure 46: Removal rates for the OSM ($d = 2$), CUBE ($d = 3$) and CLUSTER ($d = 3$) datasets



(a) 2D OSM-R (b) 3D CUBE-R (c) 3D CLUSTER-R

Figure 47: Removal rates for the OSM ($d = 2$), CUBE ($d = 3$) and CLUSTER ($d = 3$) datasets

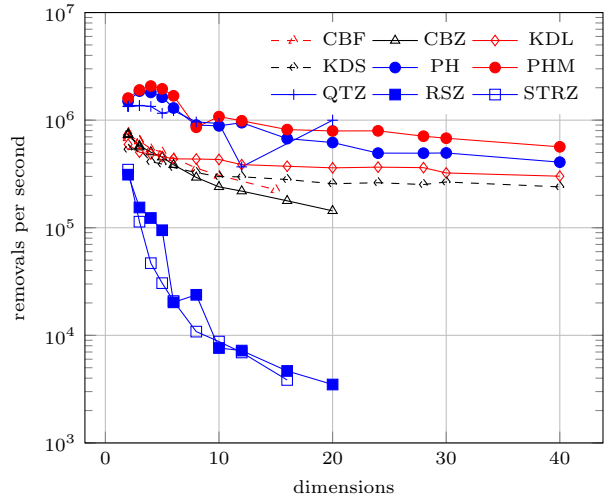


Figure 48: DIM: Removal rates for CU-P with $N = 10^6$

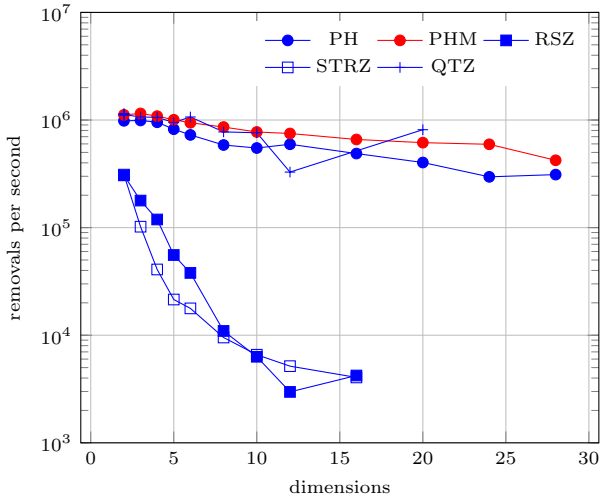


Figure 49: DIM: Removal rates for CU-R with $N = 10^6$

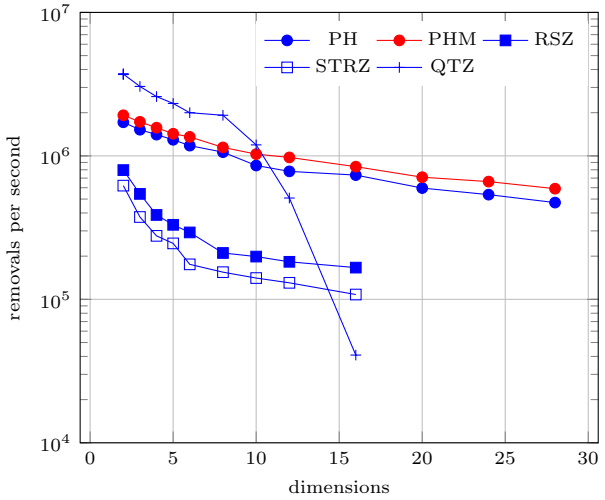


Figure 51: DIM: Removal rates for CL-R with $N = 10^6$

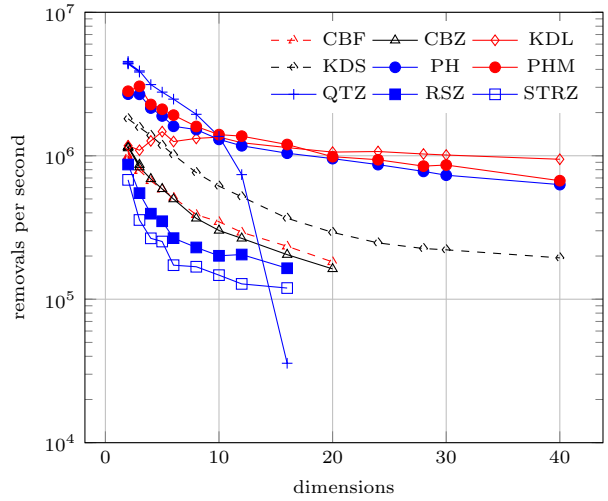


Figure 50: DIM: Removal rates for CL-P with $N = 10^6$