A Benchmark for Multidimensional Index Structures

Norbert Beckmann
Department of Computer Science
University of Bremen
D-28359 Bremen
Germany

nb@informatik.uni-bremen.de

Bernhard Seeger
Department of Computer Science
University of Marburg
D-35032 Marburg
Germany

seeger@informatik.uni-marburg.de

1. Introduction

In this paper we give an overview of our benchmark for multidimensional index structures. The benchmark consists of 28 data files. From each of them 3 query files were derived, corresponding to 1, 100 and 1000 answers per query on average. In Section 2, we provide a detailed description of the data distributions, Section 3 explaines how the query distributions were derived. In Section 4, the values of important parameters of the data distributions as well as the query distributions are depicted in a table.

2. Data Distributions, Detailed Specification

The data files of our benchmark consist of 28 data sets, 21 of them are artificially generated, whereas the others are from real application scenarios. The artificial data sets come from seven groups. Each of the groups contains a 2D, 3D and 9D data set that follows the same data distribution. The real data sets correspond to 2D, 3D, 5D, 9D, 16D, 22D and 26D distributions.

2.1 Artificial Distributions

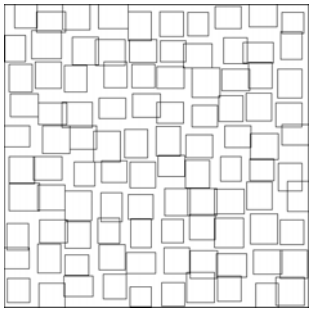
The most important criterion for the generation of the artificial data sets was that the distributions should be difficult to handle for index structures. This implies that some of them should be inserted in a non-random order. The seven distributions, chosen from a larger set, are characterized by their selectivity, the capability to reveal certain weaknesses in the tested algorithms. Data set *Uniform* (uniformly distributed data) is not an exception in this sense.

For the following context we have to define the notions *dithering vector*, *dithering a vector*, and the *density* of a set of objects. A *dithering vector* is a vector D with components d_i ($i = 1, \ldots, dim$), where d_i randomly ranges in $[-c_i,+c_i]$, and $|c_i|$ is the dithering intensity, normally identical for all i. A vector X is *dithered* by simply adding a dithering vector D. The *density* is the sum of the volumes of a set of objects, divided by the volume of the corresponding MBB. Let a rectilinear rectangle be defined by two vectors, one determining its position, the other determining its extensions.

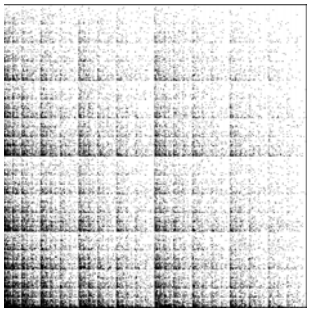
The artificial data sets were generated by algorithms which produce a certain distribution in an arbitrary dimensionality. Thereby (as far as this is possible) the fundamental character of a distribution does not change. All generated data sets consist of at least 1 million objects in a $[0,1]^{dim}$ space, the exact number depending on the applied algorithm. Due to dithering some of the objects may be out of the given space by an epsilon. In the sequel we will provide a description of the artificial distributions, followed by a picture of a 2D sample of each (except *Uniform*).

- **Absolute** (*abs02*, *abs03*, *abs09*) is generated from an equidistant distribution of equal sized squares (density 0.7) which are slightly dithered concerning position and extensions. The input order is row order, left to right.
- **Bit** (*bit02*, *bit03*, *bit09*) is a point distribution. The points, floating point vectors, are generated as follows: Each bit of the mantissa of their components is set with probability 0.3. This is the power-law distribution, closely related to the popular Zipf-distribution. The input order is random.

Absolute (100 rectangles)

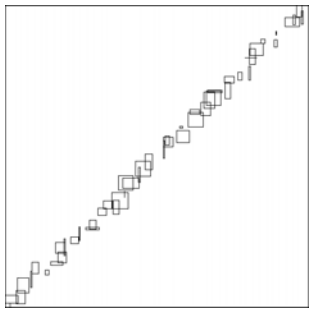


Bit (60,000 points)

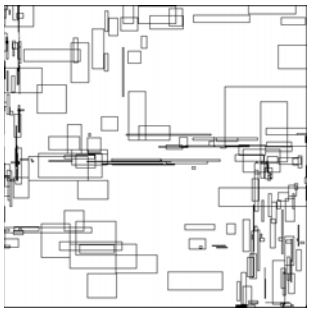


- **Diagonal** (*dia02*, *dia03*, *dia09*) is created from a start distribution of equal sized squares which are equidistantly arranged along the main diagonal of the data space. The input order of the start distribution is 0^{dim} to 1^{dim}. While the input order is kept, dithering of positions and extensions of the rectangles leads to the final distribution.
- **Parcel** (*par02*, *par03*, *par09*) is a rectangle distribution, where the variance of the sizes and shapes of the objects is extremely high. The basic structure is a partition grid, obtained by recursively splitting the data space into two partitions of random size, thereby alternating the split axis after each split. Caused by width-proportional position dithering, huge rectangles often cover tiny ones. The density of the distribution is 0.5, and the input order follows a z-order on the (undithered) rectangles.

Diagonal (50 rectangles)

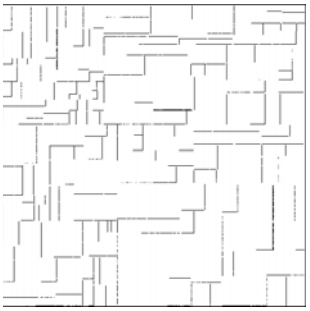


Parcel (256 rectangles)



- **P-edges** (*ped02*, *ped03*, *ped09*) is a distribution of thin stripe-shaped clusters of points. Its production is derived from that of *Parcel*. In an extra step a random input order is obtained.
- **P-haze** (*pha02*, *pha03*, *pha09*) is a distribution of ellipse-shaped clusters of points and also related to *Parcel*. The input order is determined by the distance to a set of anchor points, for which nearest neighbor queries are issued. The result is, that the distribution grows together from multiple points.

P-edges (62,841 points)



P-haze (65,365 points)

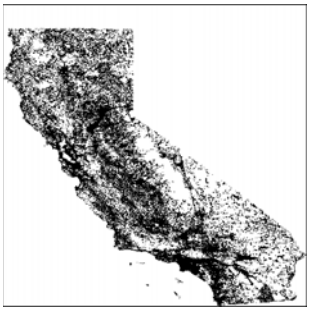
- **Uniform** (*uni02*, *uni03*, *uni09*) is a uniformly distributed set of points.

2.2 Real Distributions

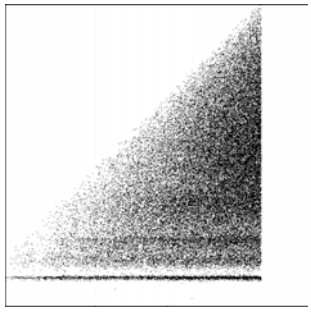
In this section we provide information on the distributions of the real data in our benchmark. In order to give an idea of the distributions with dimensionality greater than 2, we provide a plot of one of their 2D projections.

- **CaliforniaStreets** (*rea02*) is a 2D distribution and consists of 1,888,012 rectangles and points. It is derived from a line segment file of streets in California [Bur 96]. Input order: Sub-regions of roughly 20,000 objects are inserted in random order, partly in multiple layers. Inside the sub-regions the tendency is row order, west to east, the rows inserted north to south.
- **BioData** (*rea03*) is a 3D distribution consisting of 11,958,999 points. From the original file, describing a biological data set, we extracted the 3 floating point attributes. The input order is similar to random order.

CaliforniaStreets: 1,888,012 rectangles and points (each 10th displayed)

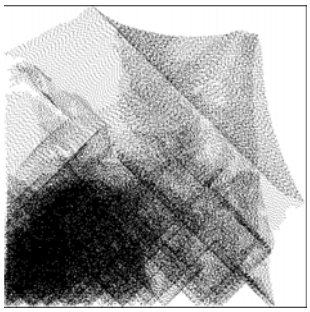


2D view through 3D **BioData**: 11,958,999 points (each 200th displayed)

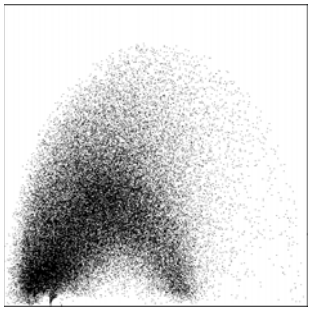


- **CoverType** (*rea05*) is a 5D distribution consisting of 581,012 points. We extracted the 5 meter-attributes from a file of the US Forest Service [UCI 02]. Input order: Seemingly in the order of collection. Not random; directional streams are recognizable.
- **ColorMoments** (*rea09*) is a 9D distribution of 68,040 points. It consists of image features extracted from a Corel image collection [UCI 02]. It has been used in similar experiments. Input order: Similar to random order.

2D view through 5D **CoverType**: 581,012 points (each 3rd displayed)

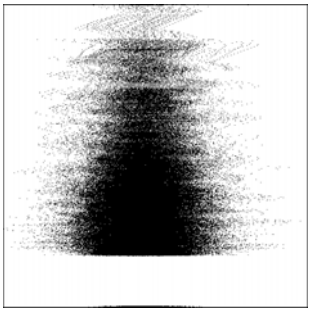


2D view through 9D **ColorMoments**: 68,040 points

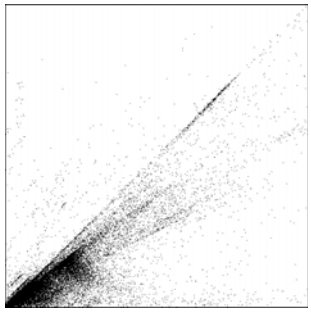


- **Fourier** (*rea16*) is a 16D distribution of 1,312,173 points. It consists of Fourier coefficients from CAD data. We removed 314 points from the original data file due to missing values. Input order: Some 2D views show an insertion order following a curve.
- **ActivityReport** (*rea22*) is a 22D distribution of 500,000 points. It represents 500,000 snapshots (one per second) of the system activity report generated by a Unix kernel. From the more than 30 categories of the original output, we extracted the 22 most active. The input order is chronological.

2D view through 16 D **Fourier**: 1,312,173 points (each 4th displayed)

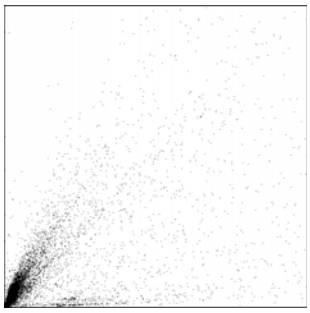


2D view through 22D **ActivityReport**: 500,000 points (each 5th displayed), cutout: $[0\ldots10\%]^2$



- **OccurrenceCount** (*rea26*) is a 26D distribution of 427,060 points. It was generated as follows: A privileged process anonymously counted the occurrence of the letters a…z in the text files of about 2000 users. Each point represents the occurrences count for a certain text file. Since programs create default files, users copy what they utilize, etc., the number of duplicates in this distribution is very high. Some very few extremely large files blew up the data space monumentally. The input order results from the proceeding: While the homes were randomly

mixed, inside the homes an alphabetical depth first search was applied.



2D view through 26D **OccurrenceCount**: 427,060 points, cutout: $[0\ldots1\%]^2$

3. Query Distributions

For each of the data files, we distinguish three query files, *QR0*, *QR2* and *QR3*, with a nominal number of 1, 100 and 1000 objects respectively to be retrieved per query. Apart from a single exception, we discuss below, they are all produced in the same manner.

File *QR0* simply consists of the center-points of the objects of the original data file. The number of answers per query is at least 1. The starting point of the query files *QR2* and *QR3* is an intermediate distribution of points *QP*. Mostly it is created by dithering the centers of the objects of the corresponding data distribution. For the artificial distributions, whose generation process often already contains a dithering step, we sometimes used the undithered prestage. For each point of *QP* a *k*-nearest-neighbors query was performed, with an average number of neighbors $k = 100$ and $k = 1000$, the actual number of neighbors for a single query randomly set to $k_i \in [0.5k, 1.5k]$. Thereby we used the L_∞ metrics. This produces rectangular search regions containing at least k_i objects. In order to limit the query cost, we decided to produce *QR0*, *QR2* and *QR3* only for each 10-th, 100-th and 316-th respectively of the original objects.

The advantage of creating rectangle query files in the described way is, that this allows to determine the performance of an index structure for a fixed (average) number of answers per query, but we barely measure the behavior in the empty parts of the data space. In order to include a test for this intension too, we decided to generate the query files for *P-edges* according to a volume based approach (*P-edges* contains big empty regions).The query distributions for *P-edges* are obtained as follows: First a uniformly distributed set of squares is computed. Each of them covers k / *total* of the entire data space, where *total* is the number of points in *P-edges*. In a second step the squares are dithered concerning their extensions. We consider three files *QR0*, *QR2* and *QR3* that refer to $k = 1$, $k = 100$ and $k = 1000$ respectively. To ensure validity in spite of a very few number of answers, the number of queries for *P-edges* is of the magnitude of the number of objects in the original distribution.

In Table 1 some important properties of the data distributions as well as the query distributions are demonstrated. The columns, left to right, address the distributions in the sequence, as they are introduced in Section 2. The distribution index (1…28) runs over the data files, while the query index (1…84) identifies the derived query files (*QR0*, *QR2*, *QR3*). Top down Table 1 consists of a header part for the data specification (6 rows), followed by three body parts for the query characteristics (each 7 rows). The header part contains the token of the distribution and its dimensionality (2 rows) followed by the distribution index. The 4-th row contains the data kind, i.e. whether the distribution consists of rectangles (rct) or points (pt) or both. The 5-th row specifies the input order, i.e. random (rand) or other than random (oth). The 6-th row specifies whether the distribution contains overlapping rectangles (ovlp) or duplicates (dup) and labels the queries generated by the volume bound approach; a "+" is used as a marker for high degree. Each of the three body parts starts with a query specification heading, top down followed by: The query index for the corresponding query; the fraction of queries with an empty result set in percent; the minimum, maximum and average number of answers respectively per query; finally the standard deviation of the number of answers in percent of the average.

4. Tables

Table 1: Data and Query Properties

distribution	abs	abs	abs	bit	bit	bit	dia	dia	dia	par	par	par	ped	ped	ped	pha	pha	pha	uni	uni	uni	rea	rea	rea	rea	rea	rea	rea
dimensionality	02	03	09	02	03	09	02	03	09	02	03	09	02	03	09	02	03	09	02	03	09	02	03	05	09	16	22	26
distrib. index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
data kind	rct			pt			rct			rct			pt			pt			pt			pt/rct	pt	pt	pt	pt	pt	pt
input order	oth			rand			oth			oth			rand			oth			rand			oth	rand	oth	rand	oth	oth	oth
peculiarities	ovlp						ovlp			ovlp +			**vol. based queries**									ovlp/dup	dup	dup		dup		dup +
QR0 characteristics																												
qu. index (QR0)	1	4	7	10	13	16	19	22	25	28	31	34	37	40	43	46	49	52	55	58	61	64	67	70	73	76	79	82
% empty qu.	0	0	0	0	0	0	0	0	0	0	0	0	94.3	97.8	99.9	0	0	0	0	0	0	0	0	0	0	0	0	0
min. #answers	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1
max. #answers	1	1	1	1	1	1	4	3	2	10	10	5	390	520	1926	1	1	1	1	1	1	9	9	2	1	27	1	4196
avg. #answers	1	1	1	1	1	1	1.26	1.06	1.00	2.11	2.12	1.08	1.02	0.95	0.16	1	1	1	1	1	1	1.20	1.01	1.00	1	1.03	1	87.0
% std. deviat.	0	0	0	0	0	0	36.9	23.1	0.32	52.4	52.7	27.1	681	1030	4670	0	0	0	0	0	0	40.5	30.3	0.41	0	36.6	0	536
QR2 characteristics																												
qu. index (QR2)	2	5	8	11	14	17	20	23	26	29	32	35	38	41	44	47	50	53	56	59	62	65	68	71	74	77	80	83
% empty qu.	0	0	0	0	0	0	0	0	0	0	0	0	46.5	60.7	96.6	0	0	0	0	0	0	0	0	0	0	0	0	0
min. #answers	50	50	50	50	50	50	50	50	50	50	50	50	0	0	0	50	50	50	50	50	50	50	50	50	50	50	50	50
max. #answers	150	150	150	150	150	150	150	150	150	150	150	150	4210	8158	16289	150	150	150	150	150	150	162	156	296	150	696	5086	115993
avg. #answers	99.8	99.8	99.8	99.8	99.8	99.8	99.8	99.8	99.8	99.9	99.9	99.9	102	93.6	13.6	99.8	99.8	99.9	99.8	99.8	99.8	101	100	102	99.4	105	230	13700
% std. deviat.	29.3	29.3	29.3	29.3	29.3	29.3	29.3	29.3	29.3	29.3	29.3	29.3	199	235	1020	29.3	29.3	29.3	29.3	29.3	29.3	29.2	29.2	30.2	29.3	31.9	170	162
QR3 characteristics																												
qu. index (QR3)	3	6	9	12	15	18	21	24	27	30	33	36	39	42	45	48	51	54	57	60	63	66	69	72	75	78	81	84
% empty qu.	0	0	0	0	0	0	0	0	0	0	0	0	11.4	22.0	87.8	0	0	0	0	0	0	0	0	0	0	0	0	0
min. #answers	500	500	500	500	500	500	500	500	500	500	500	500	0	0	0	500	500	500	500	500	500	501	500	501	501	501	501	501
max. #answers	1500	1500	1500	1500	1500	1500	1500	1500	1500	1500	1500	1500	18453	27415	69167	1500	1500	1500	1500	1500	1500	1514	1501	1529	1483	1665	6633	238567
avg. #answers	992	992	992	992	992	992	992	992	992	994	994	994	1000	882	111	992	993	992	992	992	992	999	999	1000	1000	1010	1150	23400
% std. deviat.	29.1	29.1	29.1	29.1	29.1	29.1	29.1	29.1	29.1	29.1	29.1	29.1	142	165	690	29.2	29.1	29.1	29.1	29.1	29.1	29.1	28.9	29.0	28.6	29.0	54.5	183

5. References

[Bur 96] Bureau of the Census: Tiger/Line Precensus Files: 1995 technical documentation, Bureau of the Census, Washington DC 1996.

[UCI 02] UC Irvine KDD Archive, http://kdd.ics.uci.edu/, 2002.