Selectivity Estimation of Window Queries for Line Segment Datasets

Guido Proietti[1] Christos Faloutsos[2]

May 1998

CMU-CS-98-137

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

Despite of the fact that large line segment datasets are becoming more and more popular, most of the analysis for estimating the selectivity of window queries posed on spatial data –the most important parameter for query optimization– has focused on point or region data only.

In this paper we move one significant step forward in line segment datasets theoretical analysis. We discovered that real lines closely follow a distribution law, that we named the *SLED law* (Segment LEngth Distribution). The SLED law can be used for an accurate estimation of the selectivity of window queries. Experiments on a variety of real line segment datasets (hydrographic systems, roadmaps, railroads, utilities networks) show that our law holds and that our formula is extremely accurate, enjoying a maximum relative error of 4% in estimating the selectivity.

[1]On leave from Dipartimento di Matematica Pura ed Applicata, University of L'Aquila, Via Vetoio, I-67010, Italy. His research was partially supported by the Italian National Research Council (CNR) under the fellowship N.215.29 and by the EU TMR Grant CHOROCHRONOS. E-mail: `proietti@cs.cmu.edu`.

[2]The research of this author was partially supported by the NSF under Grant IRI-9625428, by the CMU's InforMedia, by the NSF, ARPA and NASA under NSF Cooperative Agreement No. IRI-9411299. E-mail: `christos@cs.cmu.edu`

Keywords: Spatial databases, GIS, line segment data, range queries, selectivity estimation.

1 Introduction

Spatial data appear in numerous applications, such as GIS [8, 9] multimedia [6] and even traditional databases. Statistical modeling of real data involves the concise description of a dataset with a few parameters (e.g., total count, area, length etc.), so that we can obtain accurate estimates. Such a concise description is useful for at least the following settings:

- selectivity for window queries, k nearest neighbor queries, spatial joins etc.
- analysis of spatial access methods (SAM). For example, how many nodes will an R-tree or quadtree require to store such a dataset, how many such nodes a query will touch, etc.

Although some statistical models have been developed in the past for points, rectangles and regions, as we describe in detail in Section 2, no theoretical results exist for line segment data. Previous analysis are limited to empirical comparisons of the performances of various spatial indexing methods (see [7] for a comprehensive survey on the topic).

In this paper we move one significant step forward in line segment datasets theoretical analysis. We focus on large collections of line segments, like for instance roadmaps, hydrographic systems, railways, utilities networks and so, and we show that they can be efficiently modelled by means of a novel distribution law. Such a model will reveal its usefulness in predicting the selectivity of window queries posed on the dataset. Moreover, we show that a similar law holds for any window subset of a given line segment dataset. This is important for two reasons: (1) from a theoretical point of view, since it allows to predict the length of the longest line segment in a query window; (2) from a practical point of view, since we can quickly estimate the length of the longest line segment of the whole set by sampling from a query window.

The remainder of the paper is organized as follows: Section 2 gives a brief description of previous work on the topic. In Section 3, we provide the theoretical basis of our paper and we give the distribution law of line segment datasets. In Section 4 we show how such a law can be used to estimate selectivity of window queries on line segment datasets and we provide a method for a fast estimation of the length of the longest line segment of the dataset. Section 5 presents a large collection of experimental results on real line segment data (rivers, roadmaps, railroads, utilities networks) which give empirical evidence of the theoretical analysis. Finally, Section 6 contains concluding remarks and future work.

2 Survey

The main topic within the spatial database field which is related to our present work is *query optimization*, and, more specifically, *selectivity estimation* in window (or range) queries, which are the most popular spatial access operation [12].

In [10, 12], an analytical formula to compute selectivity for a window query as a function of the underlying data morphology and distribution is given. To apply such a formula when these parameters are unknown, one typically makes the *uniformity* and *independence* assumption on them. Unfortunately, these assumptions do not hold for real datasets and generally lead to pessimistic results [3].

Whereas for one-dimensional data some developed non-uniform distributions (like for example the Zipf distribution [14]) have met with success, for multi-dimensional data difficulties have not been overcome yet. In fact, some proposed non-uniform model (such as, for instance, clustering ad-hoc methods [11, 1]), are not flexible enough to be applied to a large variety of data. Recently, the introduction of the concept of fractal dimension of a set of spatial data (e.g., points, regions, etc.) has allowed to better describe the structural properties of the data themselves and to precisely analyze space and time performances of spatial data structures generally used to store them. Most of the analysis efforts have focused on point data [5, 2]. In fact, for point data, the count and the fractal dimension of the dataset are sufficient to accurately estimate selectivities for window queries, spatial joins and nearest neighbor queries. For non-point data, the most relevant results achieved are related to optimal packing for R-trees construction [10] and to the estimation of the number of quadtree blocks that are needed to store a spatial dataset consisting of a single region [4]. Recently, novel results for *region data* have been proposed in [13], where the authors developed a realistic statistical model, and showed how to use it to compute the selectivity of window queries.

However, all these works focus on point, rectangle or region datasets only. Therefore, to the best of our knowledge, this is the first attempt to model accurately line segment datasets.

3 Fundamental laws: SLED and SUD

Length of segments in real line segment datasets do not obey a uniform distribution. Rather, it turns out that the complementary cumulative distribution function (CCDF) of the lengths follows the *SLED law* (Segment LEngth Distribution), that is:

Conjecture 1 (SLED law) *The number of line segments $C(\ell)$ of length greater than or equal to ℓ, follows the law*

$$\boxed{C(\ell) = k \cdot \alpha^{-\ell} \qquad k, \alpha > 0, \qquad \ell \geq 0.} \tag{1}$$

Moreover, it turns out that the *slope* (i.e., the non-oriented acute angle between a segment and the horizontal axis) distribution of real line segment datasets obeys the *SUD law* (Slope Uniform Distribution), that is:

Conjecture 2 (SUD law) *The number of line segments $T(\theta)$ having slope equal to θ is*

$$\boxed{T(\theta) = constant \quad 0 \leq \theta \leq \tfrac{\pi}{2}.} \tag{2}$$

Based on the above laws, in the next section we show how to estimate the selectivity and the length of the longest line segment for window queries. Table 1 gives a list of symbols used throughout the paper.

Symbol	Definition
$\mathcal{L}$	Dataset of line segments
N	Total number of line segments of $\mathcal{L}$
$L(\mathcal{L})$	Total length of $\mathcal{L}$
ℓ_i	Length of the i-th line segment in $\mathcal{L}$
θ_i	Slope of the i-th line segment in $\mathcal{L}$
$\ell_{\max}$	Length of the longest line segment in $\mathcal{L}$
$\overline{\ell}(\mathcal{L})$	Average length in $\mathcal{L}$
$C(\ell)$	Number of segments of $\mathcal{L}$ having length at least ℓ
$\mathcal{S}$	Subset of line segments
N'	Total number of line segments of $\mathcal{S}$
$\ell'_{\max}$	Length of the longest line segment in $\mathcal{S}$
$\overline{\ell}(\mathcal{S})$	Average length in $\mathcal{S}$
$C'(\ell)$	Number of segments of $\mathcal{S}$ having length at least ℓ
$\vec{q} = (q_x, q_y)$	Query window of sides q_x, q_y
$Sel(\mathcal{L}, \vec{q})$	Selectivity for window queries of sides q_x, q_y

Table 1: Symbol table

4 Analysis

For clarity of presentation, we will first define a preliminary problem where line segments are supposed to be parallel, giving the exact selectivity and providing an accurate approximation of it. After, we will analyze the general case where line segments are arbitrarily oriented, providing also in this case an exact and an extremely accurate solution to the selectivity problem. Since such an estimation assumes the knowledge of the length of the longest line segment $\ell_{\max}$ and the number of line segments N of the dataset, we will lastly show how to quickly estimate $\ell_{\max}$ once the longest line segment of a subset of the whole dataset is known. This latter result is of particular interest for practical cases, since it allows to extrapolate the SLED law of the dataset by sampling from a subwindow.

4.1 Preliminary problem: selectivity of parallel line segments

Let us first focus on parallel line segments. After, all the results will be extended to arbitrarily oriented line segments.

PROBLEM 1: selectivity of parallel line segments

Given:

- A set $\mathcal{L} = \{l_1, l_2, \ldots, l_N\}$ of parallel line segments having slope $0 \le \theta \le \pi/2$, embedded in the unit square $U = [0,1] \times [0,1]$.
- The length $\ell_{\max}$ of the longest line segment in $\mathcal{L}$.
- A $q_x \times q_y$ *window query* $\vec{q}$.

Find the *selectivity* $Sel(\mathcal{L}, \vec{q})$ in $\mathcal{L}$ of the window query $\vec{q}$, that is, the number of line segments in $\mathcal{L}$ intersecting $\vec{q}$.

Let ℓ_i be the length of the segment l_i. To compute $Sel(\mathcal{L}, \vec{q})$, we adapt the formula in [10, 12] to manage line segments rather than rectangles. In fact, since the rectangular queries are uniformly distributed in the unit square address space, then the probability that a window intersects a line segment equals the probability that a point falls onto the line segment of $\mathcal{L}$ 'inflated' as shown in Figure 1.

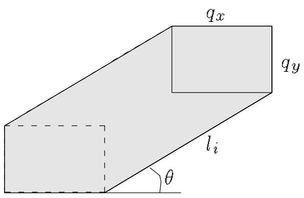


Figure 1: The 'inflated' line segment segment (shaded area)

Thus, a line segment of length ℓ_i behaves like a polygon of area

$$\ell_i \cdot (q_x \cdot \sin\theta + q_y \cdot \cos\theta) + q_x \cdot q_y.$$

Summing over all the inflated line segments we therefore obtain

$$Sel(\mathcal{L}, \vec{q}) = \sum_{i=1}^{N} \Big(\ell_i \cdot (q_x \cdot \sin\theta + q_y \cdot \cos\theta) + q_x \cdot q_y \Big) = L(\mathcal{L}) \cdot (q_x \cdot \sin\theta + q_y \cdot \cos\theta) + q_x \cdot q_y \cdot N \qquad (3)$$

where $L(\mathcal{L})$ is the total length of the set of line segments.

However, the question is to estimate the selectivity without knowing $L(\mathcal{L})$ (and, as we will see later, to estimate the selectivity for arbitrarily oriented line segments). Given Eq.(1), we show that we can compute an accurate estimation, once we fix k and α. We prove the following:

Theorem 1 *Given a set $\mathcal{L} = \{l_1, l_2, \ldots, l_N\}$ of parallel line segments embedded in the unit square $U = [0,1] \times [0,1]$, having a fixed slope $0 \leq \theta \leq \pi/2$, having lengths obeying to the SLED law and whose longest line segment has length $\ell_{\max}$, the selectivity of a rectangular window query $\vec{q}$ is*

$$\boxed{Sel(\mathcal{L}, \vec{q}) = \ell_{\max} \cdot \left(\tfrac{N-1-\ln N}{\ln N}\right) \cdot (q_x \cdot \sin\theta + q_y \cdot \cos\theta) + q_x \cdot q_y \cdot N.} \tag{4}$$

Proof. We start with Eq.(3). We need to estimate $L(\mathcal{L})$. By assumption, $\mathcal{L}$ obeys to the SLED law (Eq.(1)). Hence, from the initial conditions $N \equiv C(0) = k$ and $1 \equiv C(\ell_{\max}) = k \cdot (\alpha)^{-\ell_{\max}}$ it follows that $\alpha = \sqrt[\ell_{\max}]{N}$ and therefore

$$C(\ell) = N^{1 - \frac{\ell}{\ell_{\max}}} \tag{5}$$

and from the inverse relation we have

$$\ell(C) = \ell_{\max}(1 - \log_N C).$$

Therefore, it follows

$$L(\mathcal{L}) = \sum_{i=1}^{N} \ell_i \approx \ell_{\max} \int_1^N 1 - \log_N C \; dC = \ell_{\max} \cdot \left[C - \frac{1}{\ln N}\left(C \ln C - C\right)\right]_1^N = \ell_{\max} \cdot \left(\frac{N - 1 - \ln N}{\ln N}\right) \tag{6}$$

from which the thesis follows. □

In the next section, we relax the assumption of parallelism, to front real instances of line segment datasets.

4.2 Real problem: selectivity of arbitrarily oriented line segments

However, real line segment datasets are far to contain only parallel line segments. Therefore, the next step is to solve the following realistic and more general problem:

PROBLEM 2: selectivity of arbitrarily oriented line segments
Given:

- A set $\mathcal{L} = \{l_1, l_2, \ldots, l_N\}$ of line segments having slopes $\theta_1, \theta_2, \ldots, \theta_N$, embedded in the unit square $U = [0,1] \times [0,1]$.
- The length $\ell_{\max}$ of the longest line segment in $\mathcal{L}$.
- A $q_x \times q_y$ *window query* $\vec{q}$.

Find the *selectivity* $Sel(\mathcal{L}, \vec{q})$ in $\mathcal{L}$ of the window query $\vec{q}$, that is, the number of line segments in $\mathcal{L}$ intersecting $\vec{q}$.

For the above problem, Eq. (3) becomes

$$Sel(\mathcal{L}, \vec{q}) = \sum_{i=1}^{N} \Big(\ell_i \cdot (q_x \cdot \sin\theta_i + q_y \cdot \cos\theta_i) + q_x \cdot q_y\Big) = \sum_{i=1}^{N} \ell_i \cdot (q_x \cdot \sin\theta_i + q_y \cdot \cos\theta_i) + q_x \cdot q_y \cdot N. \quad (7)$$

The question here is to estimate the selectivity without knowing ℓ_i and $\theta_i, i = 1, \ldots, N$. In this case, assuming the dataset obeys the SLED law (Eq.(1)) and the SUD law (Eq.(2)), we can prove the following:

Theorem 2 *Given a set $\mathcal{L} = \{l_1, l_2, \ldots, l_N\}$ of line segments embedded in the unit square $U = [0,1] \times [0,1]$, having slopes $\theta_1, \theta_2, \ldots, \theta_N$ obeying to the SUD law, having lengths obeying to the SLED law and whose longest line segment has length $\ell_{\max}$, the selectivity of a rectangular window query $\vec{q}$ is*

$$\boxed{Sel(\mathcal{L}, \vec{q}) = \tfrac{2}{\pi} \cdot \ell_{\max} \cdot \left(\tfrac{N-1-\ln N}{\ln N}\right) \cdot (q_x + q_y) + q_x \cdot q_y \cdot N.} \quad (8)$$

Proof. Since segments have slopes uniformly distributed, independently of the length of a line segment, we can substitute the term $\sum_{i=1}^{N} q_x \cdot \sin\theta_i + q_y \cdot \cos\theta_i$ of Eq. (7) with its average value over the interval $[0, \pi/2]$, that is

$$\frac{\int_0^{\pi/2} q_x \cdot \sin\theta + q_y \cdot \cos\theta \; d\theta}{\pi/2} = \frac{\Big[-q_x \cdot \cos\theta + q_y \cdot \sin\theta\Big]_0^{\pi/2}}{\pi/2} = \frac{q_x + q_y}{\pi/2} \quad (9)$$

Then, the proof immediately descends from Theorem 1. □

The above theorem will provide a good estimation for window selectivity on real line segment datasets, since, as we show next experimentally, the assumptions that line segment datasets obey the SLED and the SUD law are realistic.

4.3 Practical issue: fast estimation of $\ell_{\max}$

Sometimes we do not have at disposal $\ell_{\max}$ directly from the dataset. Therefore, computing it requires to scan the entire dataset and this could be very time consuming. However, we conjecture that subwindows of the dataset will follow the SLED law as well. Moreover, we also conjecture that the average length of a line segment will be the same in the whole dataset and in a subwindow of it. More formally, if we focus on a subset $\mathcal{S} = \{s_1, s_2, \ldots, s_{N'}\}$ of $\mathcal{L}$, with $N' < N$, having the longest line segment of length $\ell'_{\max}$, we are conjecturing:

Conjecture 3 $C(\ell) = N^{1-\frac{\ell}{\ell_{\max}}} \Rightarrow C'(\ell) = N'^{\,1-\frac{\ell}{\ell'_{\max}}}$.

Conjecture 4 $\overline{\ell}(\mathcal{L}) = \overline{\ell}(\mathcal{S})$.

In the experimental section, we will show that these conjectures are altogether realistic. Hence, $\ell_{\max}$ can be inferred in the following way: from Conjecture 3 and from Eq. (6), the average length $\overline{\ell}(\mathcal{S})$ of a line segment in $\mathcal{S}$ is

$$\overline{\ell}(\mathcal{S}) = \ell'_{\max} \cdot \left(\frac{N' - 1 - \ln N'}{N' \cdot \ln N'} \right) \tag{10}$$

and therefore, from Conjecture 4, it will be

$$\boxed{\ell_{\max} \cdot \left(\frac{N-1-\ln N}{N \cdot \ln N} \right) = \ell'_{\max} \cdot \left(\frac{N'-1-\ln N'}{N' \cdot \ln N'} \right)} \tag{11}$$

from, which, knowing $\ell'_{\max}$, N' and N we can easily compute $\ell_{\max}$. In the experimental section, we will show the accuracy of our estimation.

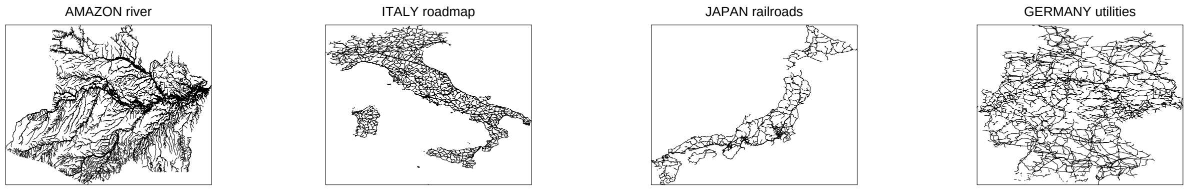
Observation 1 On the contrary, if $\ell_{\max}$ is given in advance, we can use the above relation to estimate the length of the longest line segment in a subwindow of the image space. This can be useful in answering to a query like: "Given a point in the two-dimensional space containing the line segments, which is the longest line segment within a radius of x?", which usually occurs in GIS applications.

5 Experiments on real datasets

To assess experimentally the accuracy of our analysis, we have used four line segment datasets of completely different nature, shown in Figure 2. All of them are available at `http://www.maproom.psu.edu/dcw/`, and they are:

- The Amazon River (RIVER): This consists of $N = 150241$ line segments, embedded in a 17.43×11.89 image space, having a total length $L(\mathcal{L}) = 1457.7$ and such that $\ell_{\max} = 0.100853$.
- The roadmap of Italy (ROAD): This consists of $N = 28534$ line segments, embedded in a 11.85×11.59 image space, having a total length $L(\mathcal{L}) = 459.273$ and such that $\ell_{\max} = 0.165347$.
- The railroads of Japan (RAIL): This consists of $N = 17836$ line segments, embedded in a 16.01×14.23 image space, having a total length $L(\mathcal{L}) = 259.87$ and such that $\ell_{\max} = 0.127677$.
- The utilities of Germany (UTIL): This consists of $N = 17790$ line segments, embedded in a 9.01×7.48 image space, having a total length $L(\mathcal{L}) = 494.053$ and such that $\ell_{\max} = 0.220543$.

The code for the window queries has been written in C under UNIX and the simulation experiments ran on a SUN SPARC station. In the following subsections we present experiments for: (a) verifying the SLED law (Eq. (1)); verifying the SUD law (Eq. (2)); (c) verifying the accuracy of our formula in estimating $Sel(\mathcal{L}, \vec{q})$ (Eq. (8)); (d) verifying the accuracy of our formula in estimating $\ell_{\max}$ (Eq. (11)).

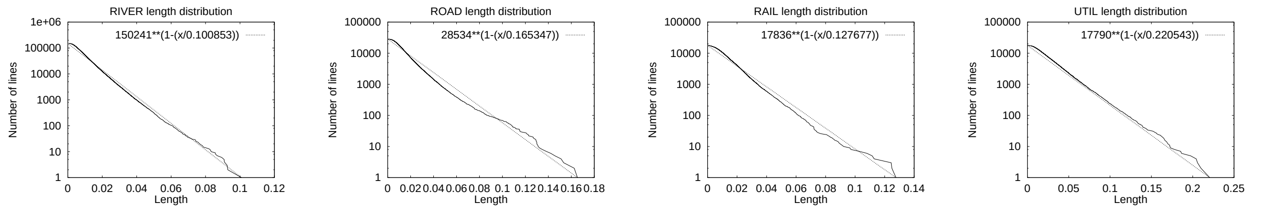


(a) RIVER (Amazon River) (b) ROAD (Roadmap of Italy) (c) RAIL (c) (Railways of Japan) (d) UTIL (Utilities of Germany)

Figure 2: Used datasets: (a) RIVER, (b) ROAD, (c) RAIL, (d) UTIL.

5.1 Verifying the SLED law

To assess the SLED law, we have computed the CCDF of the line segment length for each dataset. Figure 3 shows in a log-linear diagram the obtained results (solid line), along with the theoretical expected distribution given by Eq. (5), appearing as a straight line in the log-linear diagram (dotted line).



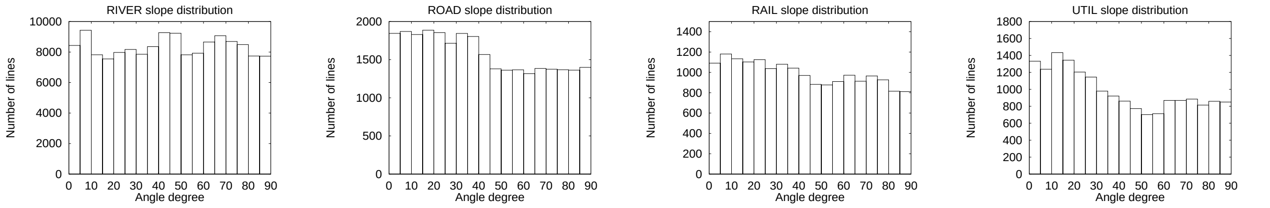
(a) RIVER (b) ROAD (c) RAIL (d) UTIL

Figure 3: CCDF of the lengths (solid line) for the used datasets in a log(count) *vs* length diagram, along with the theoretical expected distribution given by the SLED law (dotted line): (a) RIVER, (b) ROAD, (c) RAIL, (d) UTIL.

It is impressive that all four datasets, even if their characteristics are so different, obey almost perfectly to the SLED law. We have also tested the SLED law on other datasets, obtaining similar results, which are here omitted for space constraints.

5.2 Verifying the SUD law

Moreover, we have computed the distribution of the slopes of the line segments, to ascertain its uniformity (SUD law). We have divided the interval $[0, \pi/2]$ in 18 subintervals of width $\pi/36$, i.e., each interval corresponds to an angle of $5°$, and we have computed the frequency of each subinterval. Figure 4 shows using histograms the obtained results for each dataset: note that all these graphs (apart from a slight deviation in the UTIL dataset), show a uniform distribution of the line segment slope.



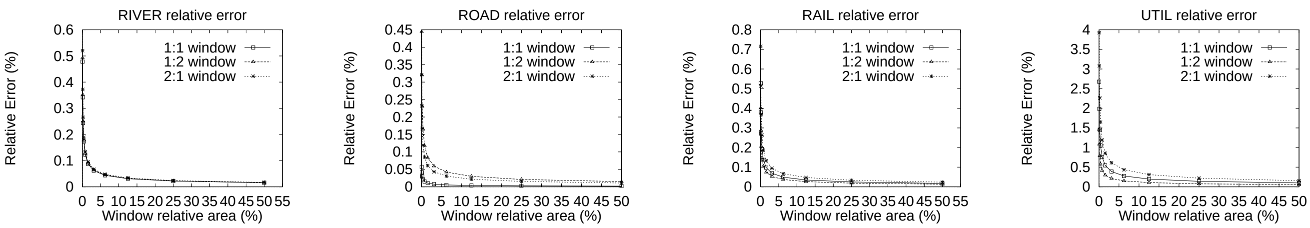
(a) RIVER (b) ROAD (c) RAIL (d) UTIL

Figure 4: Line segment slope distribution, using an interval range of 5°, for each used dataset: (a) RIVER, (b) ROAD, (c) RAIL, (d) UTIL.

5.3 Verifying our formula for selectivity

We used Eq.(7) to compute the exact selectivity on each dataset for query windows of relative area ranging from 0.05% to 50% of the image space, and we compared it with the prediction provided by Eq. (8). We examined three types of queries, depending on the aspect ratio of the query window: 1:1 (square), 1:2 and 2:1. Figure 5 shows the percentage relative error of our approach, for the RIVER, ROAD, RAIL and UTIL dataset, respectively. Note that for each dataset, our approach is usually within 1% to the reality, and never exceeds a 4% of relative error. Results appear to be independent from the window aspect ratio. The slight degradation of the accuracy for the UTIL dataset can be ascribed to the not perfect according of such dataset to the SUD law.

Finally, following the recommendations from statistics, we have also computed the *geometric average* of relative errors, for each dataset and for each different window aspect ratio, summarized in Table 2. Even in this case, it is clear the accuracy of our predictions.



(a) RIVER (b) ROAD (c) RAIL (d) UTIL

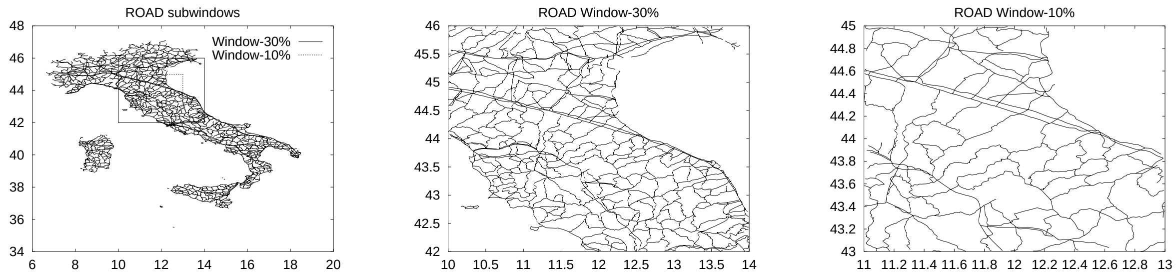
Figure 5: Relative error (%) *vs* query window relative area, for square, 1:2 and 2:1 window queries: (a) RIVER, (b) ROAD, (c) RAIL, (d) UTIL.

Geometric avg. rel. error (%)			
	Aspect ratio		
Dataset	1:1	1:2	2:1
RIVER	0.08	0.09	0.09
ROAD	0.01	0.08	0.06
RAIL	0.09	0.07	0.13
UTIL	0.53	0.29	0.83

Table 2: Geometric average relative error (%) in estimating $Sel(\mathcal{L}, \vec{q})$ for each dataset and for each aspect ratio of the query window

5.4 Verifying the estimation of $\ell_{\max}$

Finally, to check that the estimation of $\ell_{\max}$ from sampling works well, we considered two subwindows of the ROAD dataset, as shown in Figure 6, where Window-30% consists of 8983 line segments (i.e., a 30% of the total number of line segments) having $\ell'_{\max} = 0.145298$, and Window-10% consists of 2535 line segments (a 10% of the total) having $\ell'_{\max} = 0.130847$.



(a) ROAD dataset (b) Window-30% (c) Window-10%

Figure 6: Zooming into ROAD dataset: (a) the whole set with two subwindows; (b) the largest subwindow (Window-30%); (c) the smallest subwindow (Window-10%).

As a preliminary check, we have verified the truthfulness of our Conjectures 3 and 4. To verify Conjecture 3, we have computed the CCDF of the line segment length for the above windows, to verify they obey to the SLED law. This produces the graphs shown in a log-linear diagram in Figure 7. Afterwards, to verify Conjecture 4, we have computed the average length for each dataset, obtaining the results summarized in Table 3, where we also show the percent relative error with respect to the average length of the ROAD dataset. From the obtained results, we can conclude that both the conjectures hold.

Dataset	$\overline{\ell}$	Relative error (%)
ROAD	0.016111	–
Window-30%	0.015943	1.04
Window-10%	0.016635	3.25

Table 3: Average length (first column) and relative percent error (second column) for the ROAD dataset and its subwindows.

Observation 2 Note that the theoretically expected SLED laws of the ROAD dataset and of its subwindows appear as lines almost perfectly parallel. In fact

$$\overline{\ell}(\mathcal{L}) = \ell_{\max} \cdot \left(\frac{N - 1 - \ln N}{N \cdot \ln N} \right) \approx \frac{\ell_{\max}}{\ln N}$$

that is, $\overline{\ell}(\mathcal{L})$ is almost equal, in a log-linear diagram, to the negated inverse of the slope of the line corresponding to the theoretical SLED law of $\mathcal{L}$. Since from Eq. (11) we have

$$\frac{\ell'_{\max}}{\ln N'} \approx \frac{\ell_{\max}}{\ln N} \tag{12}$$

it follows that in a log-linear diagram, the SLED law graphs for the whole set and for its subsets will appear as almost perfectly parallel lines (with slope $\overline{\ell}(\mathcal{L})$). Therefore, Figure 7 gives a visual proof of our conjectures.

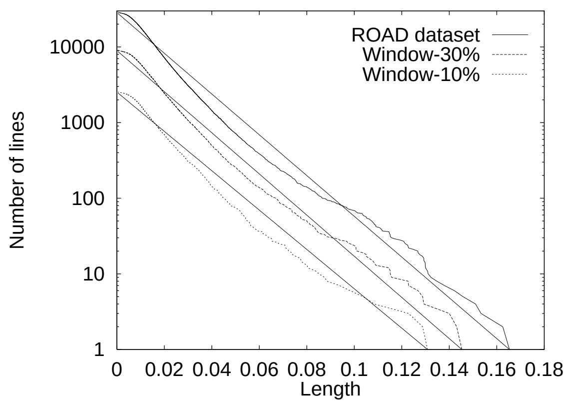


Figure 7: Comparison of the CCDF of the segment length for ROAD dataset and Window-30% and Window-10% (dotted lines), along with the theoretical expected distributions given by the SLED laws (solid lines).

Finally, we used Eq. (11) to estimate ℓ_{max}, and we obtained the results summarized in Table 4. Again, the error is extremely law (3.25% maximum), which confirms the accuracy of our approach.

Dataset	Estimation of $\ell_{\max}$	Relative error (%)
Window-30%	0.163626	1.04
Window-10%	0.170732	3.25

Table 4: Estimation of $\ell_{\max}$ (first column) and relative percent error (second column) for the two subwindows of the ROAD dataset.

6 Conclusions

The main contribution of this paper is the discovery of a law that governs real line segment datasets, such as rivers, roadmaps, railroads, utilities networks and many others. We showed that the complementary cumulative length distribution of the line segments follows a law, that we named SLED law. Thus, only two measures are needed (the total count of objects and the length of the longest line segment), to achieve extremely accurate estimation for selectivity of window queries. Our experiments on diverse, real datasets, showed that our approach achieves selectivity estimates within 4% for the maximum relative error, and usually performs within 1%. Additional contributions are:

- A formula for computing the exact selectivity for a line segment dataset, given the length and the slope of each segment.
- A fast estimation of the length of the longest line segment of a dataset by sampling from a subwindow. This is especially important for a practitioner, since it allows to estimate the selectivity without scanning the entire database when $\ell_{\max}$ is not known in advance.

Promising future directions include the study of additional query types (nearest neighbor etc.) and the analysis of SAMs on real line segment data.

References

[1] N. Beckmann, H. Kriegel, R. Schneider, and B.Seeger. The R*-tree: an efficient and robust access method for points and rectangles. In *Proc. of the 9th ACM-SIGMOD Symposium on Principles of Database Systems*, pages 322–331, Nashville, TN, 1990.

[2] A. Belussi and C. Faloutsos. Estimating the selectivity of spatial queries using the 'correlation' fractal dimension. In *Proc. of the 21st VLDB Conference*, pages 299–310, Zurich, Switzerland, 1995.

[3] S. Chistodoulakis. Implication of certain assumptions in database performance evaluation. *ACM TODS*, June 1984.

[4] C. Faloutsos and V. Gaede. Analysis of n-dimensional quadtrees using the hausdorff fractal dimension. In *Proc. of the 22nd VLDB Conference*, pages 40–50, Bombay, India, 1996.

[5] C. Faloutsos and I. Kamel. Beyond Uniformity and Independence: Analysis of R-trees Using the Concept of Fractal Dimension. In *Proc. ACM SIGACT-SIGMOD-SIGART PODS*, pages 4–13, Minneapolis, MN, May 1994. Also available as CS-TR-3198, UMIACS-TR-93-130.

[6] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos. Fast subsequence matching in time-series databases. In *Proc. ACM SIGMOD*, pages 419–429, Minneapolis, MN, May 1994. 'Best Paper' award; also available as CS-TR-3190, UMIACS-TR-93-131, ISR TR-93-86 .

[7] V. Gaede and O. Günther. Multidimensional access methods. 1998. Submitted for publication, available at http://www.wiwi.hu-berlin.de/~gaede/vg.pub.html.

[8] V. Gaede and W.F. Riekert. Spatial access methods and query processing in the object-oriented GIS GODOT. In *Proc. of the AGDM'94 Workshop*, pages 40–52, Delft, The Netherlands, 1994.

[9] R.H. Güting. An introduction to spatial database systems. *VLDB Journal*, 3(4):357–399, 1994.

[10] I. Kamel and C. Faloutsos. On packing r-trees. In *Proc. of the 2nd ACM Intl. Conf. on Information and Knowledge Management*, pages 490–499, Washington, DC, 1993.

[11] J. Orenstein. Spatial query processing in an object-oriented database system. In *Proc. of the 5th ACM-SIGMOD Conference*, pages 326–336, 1986.

[12] B. Pagel, H. Six, H. Toben, and P. Widmayer. Towards an analysis of range query performances. In *Proc. of the ACM-SIGMOD Symposium on Principles of Database Systems*, pages 214–221, Washington, DC, 1993.

[13] G. Proietti and C. Faloutsos. Accurate modeling of region data. Technical Report CMU-TR-98-126, Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 1998.

[14] G.K. Zipf. *Human behavior and principle of least effort: an introduction to human ecology.* Addison Wesley, Cambridge, MA, 1949.

Contents