Size Estimation of the Intersection Join between Two Line Segment Datasets*

Enrico Nardelli[1,2] and Guido Proietti[1,2]

[1] Dipartimento di Matematica Pura ed Applicata, Università di L'Aquila,
Via Vetoio, 67010 L'Aquila, Italy.
{nardelli,proietti}@univaq.it.
[2] Istituto di Analisi dei Sistemi e Informatica, Consiglio Nazionale delle Ricerche,
Viale Manzoni 30, 00185 Roma, Italy.

Abstract. In this paper we provide a theoretical framework for estimating the size of the intersection join between two line segment datasets (e.g., roads, railways, utilities). For real datasets, it has been pointed out that the line segment lengths and slopes are distributed according to specific mathematical laws [14]. Starting from this result, we show how to predict the size of the intersection join between two line segment datasets. We evaluate our formula through several experimentations, showing that the estimation is accurate, as compared to that obtained by using a naive uniform model.

1 Introduction

The *spatial join* between two spatial datasets is one of the most popular spatial operation. It can be defined as follows: Given two datasets $\mathcal{S}$ and $\mathcal{S}'$ of spatial objects and a binary spatial predicate $\theta : \mathcal{S} \times \mathcal{S}' \rightarrow \{\texttt{false, true}\}$, find all pairs of objects $(s, s') \in \mathcal{S} \times \mathcal{S}'$ such that $\theta(s, s') = \texttt{true}$. Among the most common spatial predicates, we recall *intersects, crosses, contains, near, adjacent, northwest, meets* and many others [7]. Among them, the most popular is certainly the *intersects* predicate, since it plays a crucial role for the computation of all kinds of joins [8].

In the past, several processing techniques of the intersection join have been developed. In particular, these techniques deal both when $\mathcal{S}$ and $\mathcal{S}'$ are indexed through an R-tree [2], and when $\mathcal{S}$ and $\mathcal{S}'$ are not indexed [11,12]. Recently, attention has been posed towards the more general problem of optimizing the processing of *multiway spatial join* [13], where the number of datasets involved in the join operation is larger than 2.

In recent years, *line segment* datasets (e.g., roadmaps, drainage systems, railways, utility networks and many others) are appearing more and more frequently in numerous applications involving spatial data, such as GIS [8,10], multimedia [4], and even traditional databases. This is especially true with the advent

* This work has been partially supported by the EU TMR Grant CHOROCHRONOS.

J. Štuller et al. (Eds.): ADBIS-DASFAA 2000, LNCS 1884, pp. 229–238, 2000.

and the rapid growing of spatio-temporal databases, where, for instance, moving points can be represented by means of polylines [5,6]. Therefore, database management systems are usually concerned with intersection join operations involving two datasets of this category, like for instance *"Find all the roads that are crossed by a drain in a given area"*.

Since usually data are stored through their minimum bounding rectangles (MBRs), together with a pointer to the corresponding database entry containing a detailed description of the object, the first step in order to optimize the operation is to retrieve all possible candidates to the output of the join, through a join performed over the MBRs of the objects: this is the so-called *filter step*. Afterward, a *refinement step* takes place, where candidate objects retrieved from the filter step are selected on the basis of effective intersection.

Therefore, to the aim of characterizing the computational effort required by an intersection join and to optimize it as a whole, it is of primary importance to estimate the *size* of the output of the refinement step, that is the number of mutual intersections between the objects in $\mathcal{S}$ and $\mathcal{S}'$. Known techniques for solving this problem generally assume that objects in $\mathcal{S}$ and $\mathcal{S}'$ are *uniformly* and *independently* distributed, although it is well known that this assumption is too restrictive when dealing with real spatial datasets [3]. In this paper we abandon this assumption, and we instead make use of an *exponential law* discovered in the past [14], and concerned with the *complementary cumulative distribution function*[1] (CCDF) of the line segment lengths (expressing the number of line segments $F(\ell)$ having length at least ℓ). We will show that using such a law, we can obtain good estimations, by knowing only few and easy-to-retrieve parameters. More precisely, we will present a large collection of experiments on several line segment datasets, showing that our prediction is usually 40% far from the reality, while the uniform model provides unreliable estimations, with a relative error of up to 5000%. Since the intersection join is the most popular join operation, and given that line segment datasets are among the largest commonly appearing spatial datasets, we conclude that we move an important step forward in the hard task of estimating the size of spatial join operations.

The paper is organized as follows: Section 2 recalls some results achieved in the past on the topic of query optimization and data modelling for multidimensional data, and gives a short insight into the mathematical laws that are used throughout the paper. In Section 3 we develop two formulae that can be used to estimate the size of an intersection join between two line segment datasets: the first one is based on a uniform model, while the second one is based on the above mentioned laws. Section 4 provides a collection of experimental results on real datasets, performed to measure the quality of the estimation provided by our model as compared to the uniform model, and suggests some ideas for a possible future improvement of our model. Finally, Section 5 contains some open problems and concluding remarks.

[1] Remember that the cumulative distribution function of $f(x) : \Re \to \Re$ is defined as $F(x) = \int_{-\infty}^{x} f(t)dt$, while the complementary cumulative distribution function is defined as $\overline{F}(x) = \int_{x}^{+\infty} f(t)dt$.

2 Previous Work

The main topic within the spatial database field which is related to our present work is *query optimization*, and, more specifically, *size estimation* of the intersection join between two line segment datasets. As we will show in the following, we will develop an analytical formula based on a non-uniform distribution of the underlying data. In fact, the uniformity assumption generally lead to pessimistic results [3].

Whereas for one-dimensional data some developed non-uniform distributions (like for example the Zipf distribution [16]) have met with success, for multi-dimensional data difficulties have not been overcome yet. Most of the previous analysis efforts have focused on point data [1]. 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 region data, novel results have been proposed in [15], where the authors developed a realistic statistical model, and showed how to use it to compute the selectivity of window queries.

Concerning *line segment data*, the selectivity of window queries has been estimated making use of an exponential law for the CCDF of the segment lengths [14]. More formally, given two points $p_1 = (x_1, y_1)$ and $p_2 = (x_2, y_2)$ in the Euclidean plane $\mathbb{E}^2$, a *convex combination* of p_1 and p_2 is a point $p = (x, y)$ such that

$$x = (1 - \alpha) \cdot x_1 + \alpha \cdot x_2 \quad \text{and} \quad y = (1 - \alpha) \cdot y_1 + \alpha \cdot y_2,$$

with $\alpha \in \mathbb{R}, 0 \leq \alpha \leq 1$. A *line segment* (or simply segment) $s = \overline{p_1 p_2}$ is the set of convex combinations of its *endpoints* p_1 and p_2. Without loss of generality, we will assume in the rest of the paper that $x_1 \leq x_2$ (if $x_1 = x_2$, then we assume that $y_1 \leq y_2$). The *length* of s is

$$\ell(s) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2},$$

while its *slope* $\theta(s)$ is the angle that s forms with the *horizontal ray*

$$\rho(s) = \{(x, y) \in \mathbb{E}^2 | x \geq x_1, y = y_1\}.$$

Therefore, we have $-\frac{\pi}{2} < \theta(s) \leq \frac{\pi}{2}$. Let $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ be a real dataset of line segments. In [14], it has been observed that the CCDF of the segment lengths, say $F(\ell)$, obeys to the following exponential law

$$F(\ell) = n \cdot \left(n^{\frac{1}{\ell_{\max}}}\right)^{-\ell} \qquad \ell \geq 0, \tag{1}$$

where $\ell_{\max}$ is the length of the longest line segment in $\mathcal{S}$. Hence, the CCDF of the lengths of $\mathcal{S}$ can be synthetically described by means of a mathematical law (named *SLED law*) containing only two constants that can be easily determined: the count of objects n and the length $\ell_{\max}$ of the longest line.

Moreover, in the same paper it has been observed that in many real line segment datasets the orientation of the segments is uniformly distributed. This has been named the *SUD law*. In the next section, we will use both the SLED and the SUD law to predict the size of the *intersection join* between two line segment datasets.

3 Proposed Method

In this section, we first give the problem definition, and we then propose two estimations of the intersection join between two line segment datasets: the first one is based on a naive uniform model, while the second one makes use of the above mentioned laws.

3.1 Problem Definition

Let us rigorously state the problem we are concerned with. For the sake of clarity, we focus on the 2-dimensional space, but all the results can be extended to the d-dimensional space.

PROBLEM: size of the intersection join between two line segment datasets

Given: In the address space $U = [0,1] \times [0,1]$, two line segment datasets $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ and $\mathcal{S}' = \{s'_1, s'_2, \ldots, s'_m\}$, whose longest segments have length $\ell_{\max}$ and $\ell'_{\max}$, respectively;

Find: the size of the *intersection join* between $\mathcal{S}$ and $\mathcal{S}'$, that is the number of mutual intersections between segments of $\mathcal{S}$ and $\mathcal{S}'$, say $\mathit{Size}(\mathcal{S} \cap \mathcal{S}')$.

3.2 A Naive Estimation Based on the Uniform Model

Assuming that $\mathcal{S}$ and $\mathcal{S}'$ obey to a uniform model, we have that each segment in $\mathcal{S}$ has length $\ell_{\max}$, while each segment in $\mathcal{S}'$ has length $\ell'_{\max}$. The following can be proved:

Theorem 1. *Let be given in $U = [0,1] \times [0,1]$ two line segment datasets $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ and $\mathcal{S}' = \{s'_1, s'_2, \ldots, s'_m\}$, whose longest segments have length $\ell_{\max}$ and $\ell'_{\max}$, respectively. If we assume that segments in $\mathcal{S}$ and $\mathcal{S}'$ are distributed according to a uniform model, then we have*

$$\boxed{\mathit{Size}(\mathcal{S} \cap \mathcal{S}') = \frac{2}{\pi} \cdot n \cdot m \cdot \ell_{\max} \cdot \ell'_{\max}.} \tag{2}$$

Proof. To estimate $\mathit{Size}(\mathcal{S} \cap \mathcal{S}')$, we handle the spatial join operation as a sequence of intersection queries posed on $\mathcal{S}'$ of each segment belonging to $\mathcal{S}$. Firstly, observe that given two segments s and s' in U, the probability they intersect is

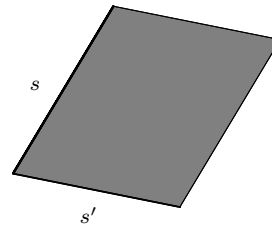

Fig. 1. Two segments s and s' intersect iff the right endpoint of s falls into the grey area.

$$p(s \cap s') = \ell(s) \cdot \ell(s') \cdot |\sin(\theta(s) - \theta(s'))|.$$

In fact, this is the probability that the right endpoint of s falls into the grey polygon depicted in Figure 1.

It follows that the expected number of segments in $\mathcal{S}'$ intersected in U by s_1, say $T(s_1, \mathcal{S}')$, is

$$T(s_1, \mathcal{S}') = \sum_{j=1}^{m} p(s_1, s'_j) = \ell(s_1) \cdot \sum_{j=1}^{m} \ell(s'_j) \cdot |\sin(\theta(s_1) - \theta(s'_j))|.$$

Analogously, when considering the i-th segment in $\mathcal{S}$, we have that

$$T(s_i, \mathcal{S}') = \sum_{j=1}^{m} p(s_i, s'_j) = \ell(s_i) \cdot \sum_{j=1}^{m} \ell(s'_j) \cdot |\sin(\theta(s_i) - \theta(s'_j))|.$$

Therefore, we have that

$$\mathit{Size}(\mathcal{S} \cap \mathcal{S}') = \sum_{i=1}^{n} T(s_i, \mathcal{S}') = \sum_{i=1}^{n} \ell(s_i) \cdot \left(\sum_{j=1}^{m} \ell(s'_j) \cdot |\sin(\theta(s_i) - \theta(s'_j))| \right). \quad (3)$$

Since segments in $\mathcal{S}$ and $\mathcal{S}'$ are oriented according to a uniform model, we have that the average value of $|\sin(\theta(s_i) - \theta(s'_j))|$ equals the average value of $\sin\theta$ in $[0, \pi/2]$, that is $2/\pi$ [14]. Moreover, since segment lengths in $\mathcal{S}$ and $\mathcal{S}'$ are uniformly distributed as well, we have that $\ell(s_i) = \ell_{max}$ and $\ell(s'_j) = \ell'_{\max}$, for all i and j. Hence, we eventually have that

$$\mathit{Size}(\mathcal{S} \cap \mathcal{S}') = \frac{2}{\pi} \cdot n \cdot m \cdot \ell_{\max} \cdot \ell'_{\max}. \quad (4)$$

□

3.3 A More Accurate Estimation

A more accurate estimation for real line datasets can be obtained by assuming that segments in $\mathcal{S}$ and $\mathcal{S}'$ obey to the SLED and to the SUD law. Let $F(\ell)$ and $F'(\ell)$ be the CCDFs associated with $\mathcal{S}$ and $\mathcal{S}'$, respectively. ¿From our assumptions, we have that

$$F(\ell) = n \cdot \left(n^{\frac{1}{\ell_{\max}}}\right)^{-\ell} \qquad F'(\ell) = m \cdot \left(m^{\frac{1}{\ell'_{\max}}}\right)^{-\ell} \qquad \ell \geq 0. \tag{5}$$

The following can be proved:

Theorem 2. *Let be given in $U = [0,1] \times [0,1]$ two line segment datasets $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ and $\mathcal{S}' = \{s'_1, s'_2, \ldots, s'_m\}$, whose longest segments have length $\ell_{\max}$ and $\ell'_{\max}$, respectively. If we assume that segments in $\mathcal{S}$ and $\mathcal{S}'$ are distributed according to the SLED and to the SUD law, then we have that* [2]

$$\boxed{Size(\mathcal{S} \cap \mathcal{S}') \approx \frac{2}{\pi} \cdot \frac{\ell_{\max}}{\ln n} \cdot \frac{\ell'_{\max}}{\ln m} \cdot (m - \ln m - 1) \cdot (n - \ln n - 1).} \tag{6}$$

Proof. The proof is based on the approach used for Theorem 1. Without loss of generality, let us assume that the segments in $\mathcal{S}$ and $\mathcal{S}'$ are sorted in decreasing order according to their length, that is $\ell(s_1) = \ell_{\max}$ and $\ell(s'_1) = \ell'_{\max}$. ¿From the inverse relation[3] of (5), we have that

$$\ell(F) = \frac{1}{\ln\left(n^{\frac{1}{\ell_{\max}}}\right)} \cdot \ln \frac{n}{F} = \frac{\ell_{\max}}{\ln n} \cdot \ln \frac{n}{F}$$

and analogously for $\ell(F')$

$$\ell(F') = \frac{\ell'_{\max}}{\ln m} \cdot \ln \frac{m}{F'}.$$

Following (3), and given that $\mathcal{S}$ and $\mathcal{S}'$ obey to the SUD law, we have that

$$Size(\mathcal{S} \cap \mathcal{S}') = \frac{2}{\pi} \cdot \sum_{i=1}^{n} \sum_{j=1}^{m} \ell(s_i) \cdot \ell(s'_j). \tag{7}$$

To estimate $Size(\mathcal{S} \cap \mathcal{S}')$, we replace the above summation by an integral. This approximation is based on the Euler's summation formula, which for sufficiently smooth functions (like $F(\ell)$ and $F'(\ell)$ are) turns out to be very accurate [9]. Therefore, we have that (7) can be rewritten as follows

[2] In the rest of the paper, all logarithms are natural.

[3] Remember that given a one-to-one function $f(x) : \Re \to \Re$, its inverse $f^{-1}(x)$ is defined by $f(f^{-1}(x)) = f^{-1}(f(x)) \equiv x$.

$$Size(\mathcal{S} \cap \mathcal{S}') \approx \frac{2}{\pi} \cdot \int_1^n \int_1^m \ell(F) \cdot \ell(F') dF \; dF' =$$

$$\frac{2}{\pi} \cdot \frac{\ell_{\max}}{\ln n} \cdot \frac{\ell'_{\max}}{\ln m} \cdot \int_1^n \int_1^m \ln \frac{n}{F} \cdot \ln \frac{m}{F'} dF \; dF' =$$

$$\frac{2}{\pi} \cdot \frac{\ell_{\max}}{\ln n} \cdot \frac{\ell'_{\max}}{\ln m} \cdot (m - \ln m - 1) \cdot (n - \ln n - 1).$$

□

4 Experiments on Real Datasets

To assess experimentally the accuracy of our formula (6), we have tested it on different line segment datasets scattered all around the world (Italy, Germany, Japan, California, Russia, etc.), available at `http://www.gisdatadepot.com`. More precisely, we have downloaded all the line segment datasets available for several different countries. Afterwards, we have computed the intersection join between all pairs of datasets of each country, since it does not make much sense to intersect two datasets from two different countries. Due to space limitations, we here provide a small subset of the experiments, concerned with the following data of North Italy:

- Drainage system (DRAIN), consisting of 18,923 segments;
- Railways network (RAIL), consisting of 4,469 segments;
- Roadmap (ROAD), consisting of 9,732 segments;
- Utility network (UTIL), consisting of 2,070 segments.

All the datasets were stored in vectorial format on a Digital DEC 3000 running UNIX V4.0B. Preliminarily, we have computed all the relevant features needed for checking our results. Such a computation is very fast, since it can be performed by means of a single scan of the datasets. These data are summarized in Table 1. Figure 2 depicts the datasets, along with the CCDF of their segment lengths. Notice that CCDFs are plotted in a log-linear diagram: The pictures confirm that the CCDFs follow very well an exponential law, since they appear as straight lines in the log-linear diagram.

To ascertain the accuracy of our formula (6) as compared with the estimation provided by the uniform model (2), we have opposed them to the real size of the

Table 1. Datasets features.

Dataset	Count	ℓ_{max}	Image Space Area
DRAIN	18,923	0.09961	23.842
RAIL	4,469	0.15468	23.242
ROAD	9,732	0.14578	23.529
UTIL	2,070	0.41221	22.342

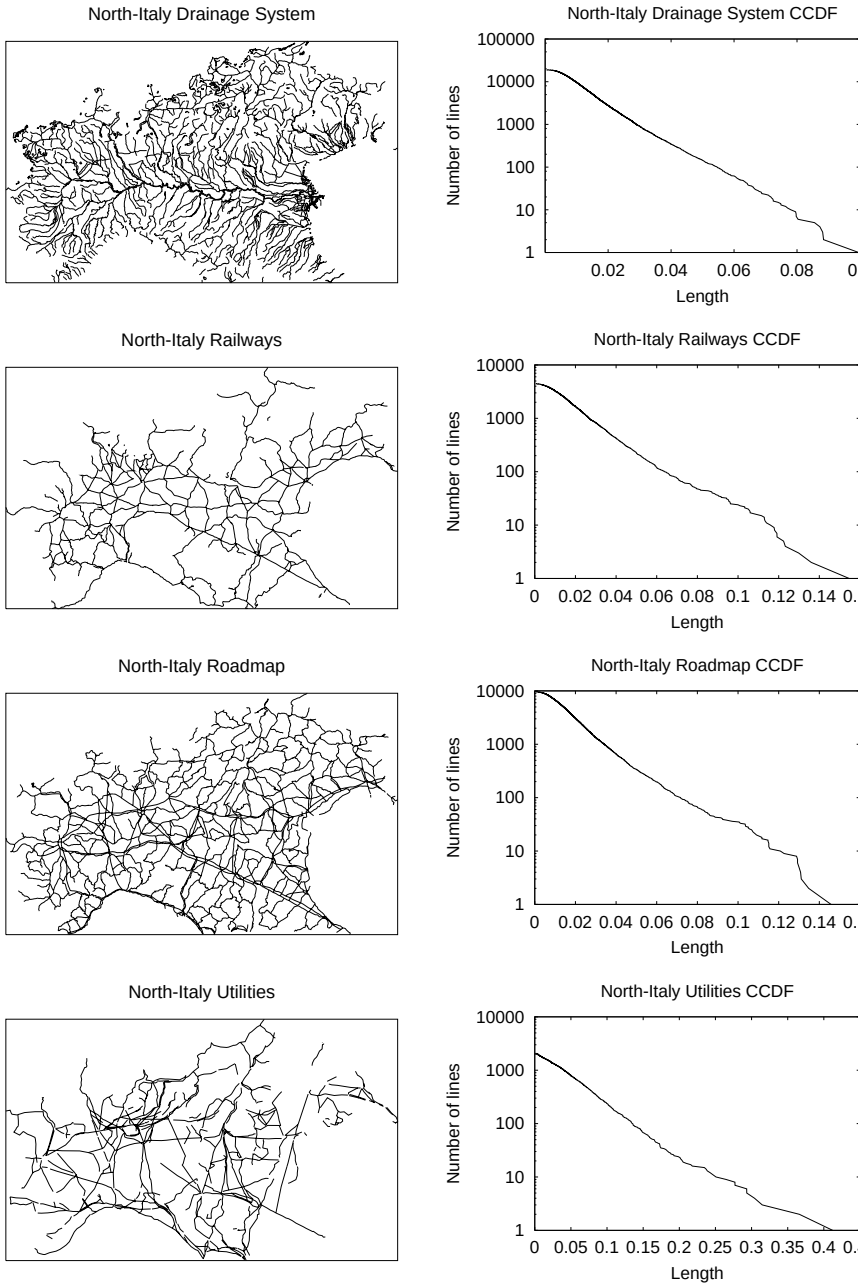

Fig. 2. Used datasets, together with their CCDF plots.

Table 2. Experimental results: real size of the joins versus estimated ones using the proposed technique (NEW) and the uniform model (OLD), together with their respective relative errors.

Dataset	Actual	NEW	OLD	% Err NEW	% Err OLD
DRAIN ∩ RAIL	625	367.1	30482.9	-70.1	4777.2
DRAIN ∩ ROAD	1202	683.9	61956.2	-75.7	5054.4
DRAIN ∩ UTIL	875	510.8	38593.0	-71.2	4310.6
RAIL ∩ ROAD	601	347.0	26861.3	-42.2	4369.3
RAIL ∩ UTIL	385	250.2	16155.2	-38.5	4515.7
ROAD ∩ UTIL	675	489.6	34493.6	-27.5	5010.0

intersection joins between all possible pairs of datasets. The intersection join has been performed by using segment trees implemented in C language. Given the size of some datasets (of the order of several thousands of line segments), the results have been obtained by paying a severe cost in terms of CPU time (up to a hour). This confirms that the size estimation of spatial join operations is a very crucial step in query optimization. Table 2 contains the obtained results and their relative errors, both for our model (NEW) and the uniform one (OLD).

A first comment on the results is that our estimation maintains the error within 75%, achieving an accuracy of 27%, while the uniform model is totally unreliable, with an error in the (over)estimation up to 5,000%. Notice that our model tends to underestimate the actual size of the intersection. Our explanation for this fact is that the datasets from a given country tend to overlap (given that they are defined over the same geographic space and therefore there is a strong correlation among them). In some sense, we could interpret the deviation from the predicted value as a measure of the correlation between the datasets!

We leave as a future study the problem of correcting our formulas so that they take into account from the beginning of the correlation between the two datasets that are going to be joined.

5 Conclusions

The main contribution of this paper is the estimation of the size of the intersection join between two spatial datasets containing line segments.

We showed that very few measures are needed (essentially the count of segments and the length of the longest segment), to achieve quite accurate results. Our experiments on diverse, real datasets, scattered around the world showed that our approach achieves estimates pretty close to the reality, while a straightforward estimation based on a uniform model provides result totally unreliable.

Promising future directions include the study of the intersection join on spatial datasets other than line datasets, the extension to the case of multiway intersection join, and the analysis of other spatial join operations. We also look forward to improve our formulas by taking into account of a *correlation factor* between the datasets.

References

1. A. Belussi and C. Faloutsos. Estimating the selectivity of spatial queries using the 'correlation' fractal dimension. In *21th Conference on Very Large Data Bases (VLDB'95)*, pages 299–310, Zurich, Switzerland, 1995.
2. T. Brinkhoff, H.P. Kriegel, and B. Seeger. Efficient processing of spatial joins using R-trees. In *19th ACM Int. Conf. on Management of Data (SIGMOD'93)*, pages 237–246, 1993.
3. S. Christodoulakis. Implication of certain assumptions in database performance evaluation. *ACM TODS*, 9(2):163–186, June 1984.
4. C. Faloutsos, M. Ranganathan, and Y. Manolopoulos. Fast subsequence matching in time-series databases. In *20th ACM Int. Conference on Management of Data (SIGMOD'94)*, pages 419–429, Minneapolis, MN, May 1994.
5. L. Forlizzi, R.H. Güting, E. Nardelli, and M. Schneider. A data model and data structures for moving objects databases. In *26th ACM Int. Conf. on Management of Data (SIGMOD 2000)*, pages 319–330, 2000.
6. A.U. Frank, S. Grumbach, R.H. Güting, C.S. Jensen, M. Koubarakis, N.A. Lorentzos, Y. Manolopoulos, E. Nardelli, B. Pernici, H.J. Schek, M. Scholl, T.K. Sellis, B. Theodoulidis, and P. Widmayer. Chorochronos: A research network for spatiotemporal database systems. *SIGMOD Record*, 28(3):12–21, 1999.
7. V. Gaede and O. Günther. Multidimensional access methods. *Computing Surveys*, 30(2):170–231, 1998.
8. V. Gaede and W.F. Riekert. Spatial access methods and query processing in the object-oriented GIS GODOT. In *AGDM'94 Workshop*, pages 40–52, Delft, The Netherlands, 1994.
9. R.L. Graham, D.E. Knuth, and O. Patashnik. *Concrete Mathematics.* Addison-Wesley Publishing Company, New York, 1989.
10. R.H. Güting. An introduction to spatial database systems. *VLDB Journal*, 3(4):357–399, 1994.
11. N. Koudas and K.C. Sevcik. Size separation spatial join. In *23th ACM Int. Conf. on Management of Data (SIGMOD'97)*, pages 324–335, 1997.
12. M.L. Lo and C.V. Ravishankar. Spatial joins using seeded trees. In *20th ACM Int. Conf. on Management of Data (SIGMOD'94)*, pages 209–220, 1994.
13. D. Papadias, N. Mamoulis, and Y. Theodoridis. Processing and optimization of multiway spatial joins using R-trees. In *18th ACM Symp. on Principles of Database Systems (PODS'99)*, pages 44–55, 1999.
14. G. Proietti and C. Faloutsos. Selectivity estimation of window queries for line segment datasets. In *7th ACM Conference on Information and Knowledge Management (CIKM'98)*, pages 340–347, Washington, DC, 1998.
15. G. Proietti and C. Faloutsos. Accurate modeling of region data. *IEEE Trans. on Knowledge and Data Engineering,* in press, 2000. Also available as CMU-TR-98-126, Dept. of Computer Science, Carnegie Mellon University, Pittsburgh, PA.
16. G.K. Zipf. *Human behavior and principle of least effort: an introduction to human ecology.* Addison Wesley, Cambridge, MA, 1949.