# Practical Data Management Techniques for Vehicle Tracking Data[1]

Sotiris Brakatsoulas Dieter Pfoser Nectaria Tryfona
*Research Academic Computer Technology Institute*
*Athens, Hellas*
{*sbrakats|pfoser|tryfona*}*@cti.gr*

## Abstract

*A novel data source for assessing traffic conditions is floating car data (FCD) in the form of vehicle tracking data, or, in database terms, trajectory data. This work proposes practical data management techniques including data pre-processing, data modeling and indexing to support the analysis and the data mining of vehicle tracking data*

## 1. Introduction

In recent years, a new sensor technology is utilized to overcome the problem of costly, stationary sensor networks in traffic assessment and prediction. Floating car data (FCD) refers to using data generated by one vehicle as a sample to assess to overall traffic conditions ("cork swimming in the river"). The essential data component is the position of the vehicle, which recorded in time produces tracking data, or, in database terms, trajectory data.

An essential task in traffic assessment is mining large collections of historic tracking data. For efficient data mining algorithm implementations, basic, efficient data manipulation is needed. The various existing approaches to store and query trajectory data have the distinct and decisive disadvantage that they cannot easily be implemented in practice [2], i.e., by using and extending off-the-shelf DBMSs to manage the data.

This work attempts to show that by exploiting the particularities of the data to create specialized data models, trajectories can be stored efficiently and general purpose access methods can be used to achieve efficient query performance. When movement occurs in networks [4], e.g., cars are bound to a road network, the spatial aspect of the trajectory is captured by this network. Consequently, when storing trajectories as 3D polylines the spatial geometry is stored redundantly. Creating a data model that separates the spatial from the temporal aspect of the data allows us not only to provide efficient query processing capabilities but also reduces the storage requirements considerably (cf. [3][6] for existing indexing approaches). The presented results are based on experiments on real-world trajectory data collected by tracking vehicle fleets in the metropolitan area of Athens, Greece.

## 2. Data Model

Exploiting the network aspect of the data, the *network schema*, a conceptual schema for trajectories [1] as shown in Figure 1 can be defined. The schema models the *spatial aspect* of the trajectories using the network, which consists of edges and nodes. Each edge has a 'from node' and a 'to node'. Additionally, the incident edges of nodes (HAS EDGES) are captured. The geometry of a node is a two-dimensional point. The trajectories are related to the network edges in that a trajectory consists of segments, which in turn relate to the network edges. The *temporal aspect* of the trajectory is captured by assigning two timestamps (time1 and time2) to the segment entity (start and end times). As an alternative to the network schema, a conceptual schema for trajectories not considering any movement constraints was developed, the *3D schema*. Here, the geometry of the trajectory segments is captured by three-dimensional line segments.

## 3. Performance Study

The performance study contrasts the query performance of the network schema to the 3D schema in terms of (i) the storage size and (ii) the I/O cost.

The data set used consists of ca. 26000 trajectories (11 million segments) obtained through GPS vehicle tracking in the municipal area of Athens, Greece (40x40km). A map

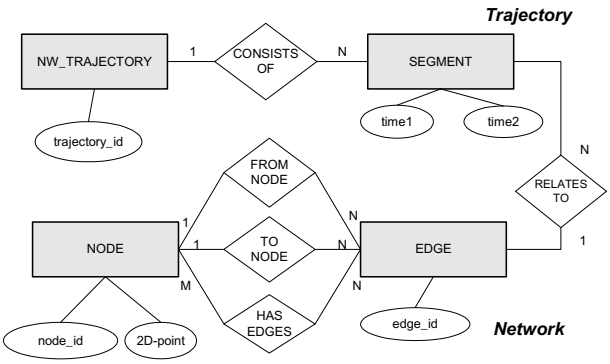


Figure 1: A conceptual schema for modeling trajectories exploiting the network aspect

[1] Work supported in part by the DBGlobe project (IST-2001-32645) and the IXNILATIS project.

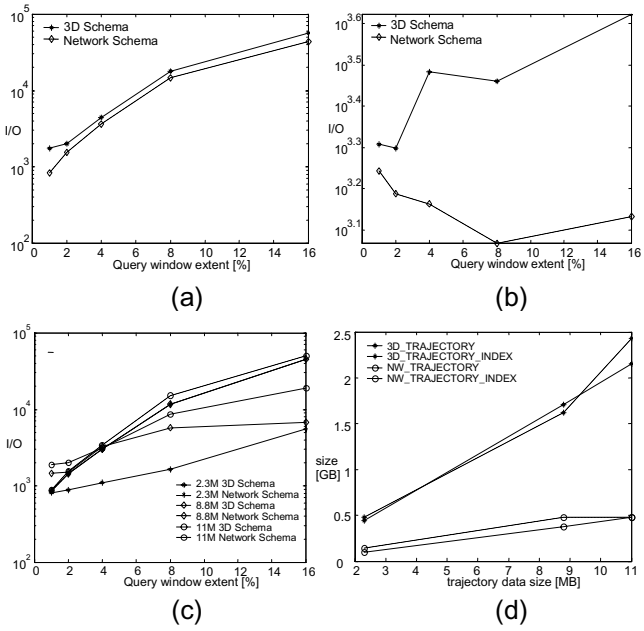


Figure 2: I/O cost for queries with (a) var. spatial and temporal (% of workspace), (b) const. spatial (2%) and var. temporal, (c) var. spatial and const. temporal extent and var. trajectory data set and (d) growth of main table sizes for varying trajectory data sets

matching algorithm tailored to the tracking data was used to map the trajectory data onto vector road map of Athens (150,000 road segments).

The logical schemas derived from the conceptual model of Section 2 were implemented by means of an Oracle DBMS version 9i Spatial. The indexes used were a three-dimensional R-tree index for the trajectory data in the 3D schema and a two-dimensional R-tree index for the road network and a composite B-tree index for the temporal data in the network schema.

The storage requirement for the trajectory data set is (tables + indexes) in the case of the 3D schema 4.6GB, whereas for the network schema only 0.9GB, thus roughly only 20% the size of 3D schema.

As for query processing, a spatiotemporal range query in the 3D schema can be directly executed towards the trajectory data. The network schema requires decomposing the query into a spatial and temporal range query, retrieving first the relevant network edges and based on this intermediate result the trajectory segments that fulfill the overall query predicate.

The query performance is assessed in terms of I/O cost in experiments with (i) a varying query window size and (ii) a varying data set size.

For queries of varying spatial and temporal extent (Figure 2(a)), the network schema outperforms the 3D schema. Further, the I/O cost of the 3D schema increases faster with larger query window sizes. For queries of constant spatial and varying temporal extent (Figure 2(b)) the cost of the 3D schema increases steadily, whereas for the network schema the cost remains more or less constant. The 3D schema treats the spatial and the temporal aspect of trajectories the same way. Thus, increasing either dimension has the same consequence, increased "spatial" search. The network schema separates these aspects and only and increased spatial extent incurs increased spatial search. An increased temporal extent is treated as an extended search over a one-dimensional (temporal) attribute.

The performance study for varying trajectory datasets (2.3M, 8.8M, and 11M segments corresponding to the years 2000-2001, 2000-2002, and all the data, respectively) indicates that the network schema scales better, i.e., the I/O cost of the network schema improves compared to the 3D schema as the time horizon of the data is increased (Figure 2(c)). An increased trajectory data time horizon incurs for 3D schema (i) an increased amount of redundant (spatial) data, (ii) a decreased space utilization in the respective three-dimensional index, and (iii) compared to the network schema, a faster growing database size. Figure 2(d) shows the growth of the main table size and indexes.

Overall, the performance study shows that (i) at considerably smaller storage sizes, the network schema typically outperforms the 3D schema and (ii) the performance gain of the network schema over the 3D schema grows with an increasing time horizon of the trajectory data.

## 4. Future Work

(i) Generalization of the network schema towards normalization that incorporates additional knowledge about the movement (e.g., maximum speed, directions, etc.). (ii) A data mining toolkit for trajectory data to be used in traffic assessment and prediction. (iii) Sensor data from miniaturized (GPS) positioning devices will require research in compression, abstraction, and decision making.

## References


[1] Brakatsoulas, S., Pfoser, D., and Tryfona, N.: Modeling, Storing and Mining Moving Object Databases. In *Proc. IDEAS conference*, pp. 68-77, 2004.

[2] Chakka, V. P., Everspaugh, A., Patel, J. M.: Indexing Large Trajectory data sets with SETI. In *Proc. CIDR*, 2003.

[3] Frentzos, E.: Indexing Objects Moving on Fixed Networks. In *Proc. of the 8th SSTD conference*, pp. 289-305, 2003.

[4] Hage, C., Jensen, C.S., Pedersen, T.B., Speicys, L., and Timko, I.: Integrated Data Management for Mobile Services in the Real World. In Proc. *of the 29th VLDB conference*, pp. 1019-1030, 2003.

[5] Kühne, R., Schäfer, R.-P., Mikat, J., Thiessenhusen, K.-U., Böttger, U. and Lorkowski, S.: New approaches for traffic management in metropolitan areas. In *Proc. of the IEEE conference on ITS*, 2003.

[6] Pfoser, D. and Jensen, C.: Indexing of Network Constrained Moving Objects. In *Proc. 11th ACM GIS symposium*, pp. 25-32, 2003.