Synthetic and Real Spatiotemporal Datasets

Mario A. Nascimento[1] Dieter Pfoser[2] Yannis Theodoridis[3]

[1]Department of Computing Science, University of Alberta, Canada, mn@cs.ualberta.ca

[2]Computer Technology Institute, Greece, pfoser@cti.gr

[3]Department of Informatics, University of Piraeus, Greece, ytheod@unipi.gr

Abstract

In the context of a spatiotemporal research environment, it is very important to be able to systematically generate data with predictable characteristics. For instance, it allows one to use the same datasets, or others similarly characterized, for benchmarking access structures or mining techniques. This paper presents a survey of existing generators of synthetic spatiotemporal data. It also covers a few real datasets, which are (at the time of this writing) publicly available for research use.

1 Introduction

While spatial data management and temporal management have been researched since more than 20 years ago (e.g., [6, 13]), the combination of both as a research topic is younger although just as strong in terms of interest (e.g., [7]).

Among the many topics which have been explored recently, such as spatiotemporal data modeling and query languages (e.g., [5]), spatiotemporal data mining (e.g., [11]) and spatiotemporal indexing (e.g., [8]), many (notably the former two) require the use of datasets in order to be evaluated. Hence the need for an automatic means to generate datasets in a systematic way and with predictable characteristics. Interestingly, despite the same need exists for "purely" spatial and temporal data, little work on data generation can be found, e.g., the *a La Carte* environment for benchmarking spatial joins (http://www.infres.enst.fr/˜bdtest/sigbench/) [4] and the *SpyTime* environment for temporal data (http://www.cs.nyu.edu/cs/faculty/shasha/spytime/spytime.html).

Although we also cover some real spatiotemporal datasets, this paper deals mainly with the issue of generating synthetic spatiotemporal data, with a focus on non-networked based data. (Network-based data generators are covered in [1] and elsewhere in this issue.) Towards this goal, the paper is structured as follows. Section 2 covers GSTD, to our knowledge, the first web-based, spatiotemporal data generator and its enhancements over time. Two other systems, G-TERD and Oporto are also reviewed, and all three are compared among themselves. Next, Section 3 presents some real datasets one could also use. Finally, we give directions for future work in Section 4.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

2 Spatiotemporal Data Generators

2.1 GSTD

GSTD [16] was initially built upon a few basic yet general principles discussed in [14]. As a result GSTD currently supports the generation of both points and MBRs (Minimum Bounding Rectangles). The generated datasets are transaction-time oriented and memory-less (i.e., future events do not depend on past states). Further, the cardinality of the dataset is assumed to be constant throughout the data generation process.

The following three parameters control the data generation process and allow the generation of a wide variety of scenarios (we use the same terminology as in [16]):

- The *duration* of an object, i.e., how often (time-wise) a change of its position occurs.
- The *shift* of an object, i.e., how fast (or slow) it will move.
- The *resizing* of an object (applicable only to objects of type MBR), i.e., the shrinking/enlargement of objects.

For each of those parameters the user can chose a statistical distribution to be followed; the current implementation supports Uniform, Gaussian and Skewed (Zipfian) distributions. In addition the user can also specify upper and lower bounds for each of the three parameters.

Finally, GSTD also provides three different ways one can handle the case of points leaving the dataspace of interest (the unit square): (i) in the *radar* approach, objects may leave the dataspace of interest and while not displayed are still considered since they can eventually return (and be re-displayed); (ii) objects can also "bounce off" the space coordinates in the *adjustment* approach; and (iii) in the *toroid* approach, as the name suggests, the data space is assumed to be toroidal, hence objects never leave it.

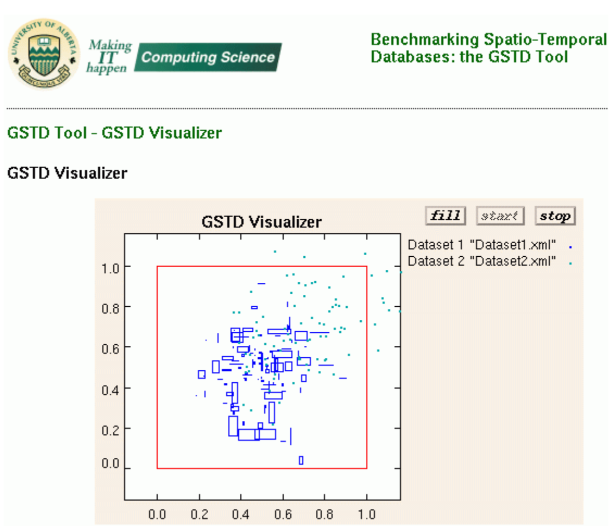


Figure 1: Snapshot of two (animated) datasets being displayed concurrently

Some enhancements over the original GSTD algorithm were introduced in [10]. First the idea of *nervousness* is introduced, i.e., varying the object's *shift*. In GSTD's initial design the changes in the objects' shift were to take effect during the whole simulation lifetime. The introduction of the new parameter allows it to change its behavior (again in a systematic way). A second change was the notion of an *infrastructure*, i.e., objects which obstruct movement. Infrastructure can be composed of real objects or synthetically generated MBRs. In the latter case, MBRs could change their shape/size and move as well.

Initially developed as a stand alone application, GSTD was improved and re-implemented as a web-based application (available via http://db.cs.ualberta.ca:8080/gstd; the site also provides source code for the data generator, so that it can be run locally.) [15]. Its current version allows one to generate and to store on the Web server several datasets in each run. One or more of those datasets can be visualized (in an animated manner) at the same time. The user can download the dataset (in XML format) for future use and/or distribution. Note that as long as the users publish the values of the GSTD parameters they used, anyone can reproduce (and use) exactly the same dataset – this is the chief goal of GSTD, namely, removing the *ad hoc* nature of evaluating and comparing different systems.

To illustrate some of the GSTD features from above, Figure 1 shows a single snapshot of two datasets (generated separately) being displayed concurrently. One of the datasets exhibits points moving freely (radar approach) from a central cluster (Gaussian) towards the upper left corner of the dataspace, whereas the other dataset is a set of moving MBRs, which change shape and size in time.

2.2 G-TERD

The Generator for Time-Evolving Regional Data, G-TERD, (http://delab.csd.auth.gr/stdbs/g-terd.html) differs somewhat from GSTD in that it generates sequences of raster images [17]. As a separate paper in this issue is devoted to G-TERD, we cover only its relation to GSTD.

Whereas GSTD is web based, G-TERD is an MS-Windows based application; its source code for the (stand-alone) data generator is publicly available through the web. The generated data can be visualized (although not animated as for GSTD) using an accompanying application.

G-TERD allows the user to set more parameters than does GSTD. It supports the statistical distributions supported by GSTD and a few additional ones. While GSTD generates moving points and MRRs, G-TERD is able to generate regions of more general shapes, which may, e.g., rotate, enlarge, or shrink. The coloring of regions is also supported. Like GSTD, G-TERD allows for the specification of obstacles to movement.

GSTD's *radar* approach allows objects to leave the dataspace; the viewable area in GSTD is fixed and cannot be changed. In G-TERD, the dataspace is typically larger than what the user sees, and a so-called *scene-observer* capability allows the user to change point of view, e.g., follow a particular object's path in time or "fly" over the dataspace.

2.3 Oporto

The Oporto generator (http://www-inf.enst.fr/∼saglio/etudes/oporto/) [12] was not designed to be as general as GSTD or G-TERD; instead, it mimics a very specific scenario: fishing at sea. In a nutshell, it models fishing ships, which leave harbors following shoals of fish while at the same time avoiding storm areas. The shoals of fish themselves are attracted by plankton areas.

Harbors are static objects, while ships, storms and plankton areas, so-called *bad and good spots*, are dynamic ones. Ships and harbors are modeled as moving and static points, respectively, while spots are MBRs, which can vary in shape and size, but do not move. In addition, they always grow and subsequently shrink (which may not be exactly a very realistic assumption). Shoals of fish, on the other hand, can change size, shape and position over time. The user can model a shore line along with the location of harbors on it.

Unlike GSTD and G-TERD, the underlying model of the Oporto generator is based on the notion of attraction and repulsion. That is, ships (fish) are attracted by fish (plankton), whereas storm areas repel the ships.

While the authors argue that Oporto is capable of generating datasets representing several scenarios, it seems to be quite limited when compared to GSTD and G-TERD. Nevertheless, one can argue for the value of being based on a well known real application. Another limitation when compared to the other generators is its limited capability of generating data according to different distributions – only the Uniform distribution is supported.

Oporto allows the user to generate and visualize animated datasets using the web (like GSTD) and is also available as a MS Windows stand alone application (actually two, one for the the generator and another for the visualizing the results). In Figure 2, the two consecutive snapshots illustrate the motions of two moving objects (ships), with the former (latter) being attracted by a gray (white, respectively) shoals of fish.

3 Real Spatiotemporal Datasets

Data generators can produce datasets of any size and kind. To empirically evaluate algorithms size is of foremost importance, but the kind of data eliminates final doubts about the suitability of a method.

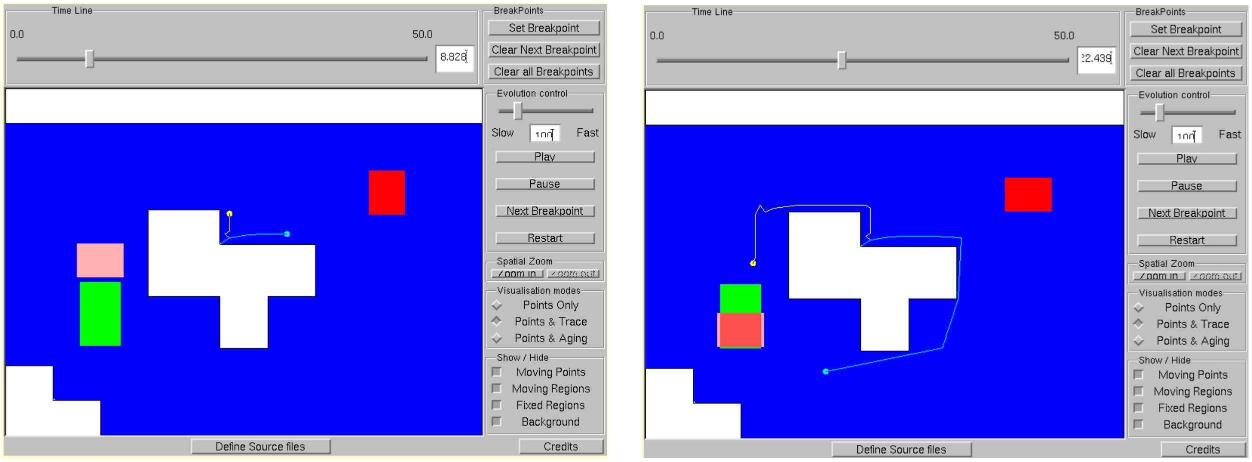


Figure 2: Snapshots of Oporto's interface

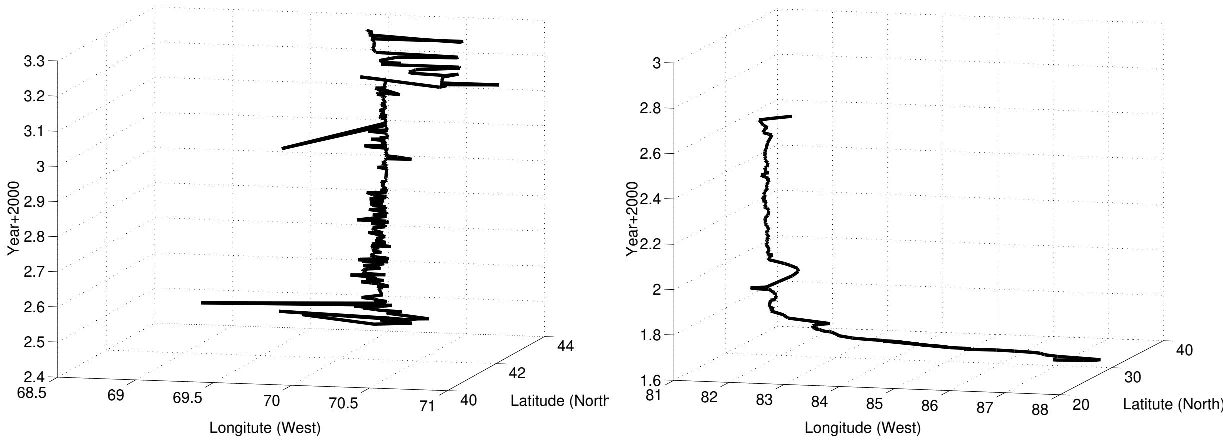
In the following, we survey a number of available datasets of varying size and kind. All datasets comprise position samples of moving point objects. They are characterized by the parameters (i) number of moving objects, (ii) number of position samples, (iii) spatial and (iv) temporal extent. Additional datasets can be found on the homepage of the author [3]. The visualization of the datasets uses a three-dimensional spatiotemporal representation [9].

Animal Tracking The tracking of animals is common for many scientific purposes. Two of the larger datasets that exist are the tracking of seals [19] and turtles [2]. The seal dataset (cf. Figure 3(a)) was obtained by tracking one animal ("Louise"). It consists of 261 position samples. The spatial extent of the data is 2 and 3.5 degrees of Longitude and Latitude, respectively. The temporal extent is from May 2002 to March 2003. The position samples in the dataset are of varying precision. Various degrees of goodness values in the dataset indicate the reliability of the positional fix. The same site features several other, although smaller datasets from seals, whales, pinnipeds, etc.

The tracking of a turtle resulted in the dataset visualized in Figure 3(b). It consists of 155 data points. The spatial extent of the data is 6 and 7 degrees of Longitude and Latitude, respectively. The temporal extent is from July 2001 to August 2003. No positional precision is indicted in the dataset. The same site contains a total of 11 turtle datasets of similar size.

Hurricanes A large meteorology database provides hurricane tracking data [18]. Figure 4(a) and (b) visualize the traces of 12 storms recorded in the year 2002. The dataset consists of 365 data points. The spatial extent of the data is 70 and 50 degrees of Longitude and Latitude, respectively. The temporal extent is from July to October 2002. The site contains overall storm tracking data from the years starting in 1996 up until the present.

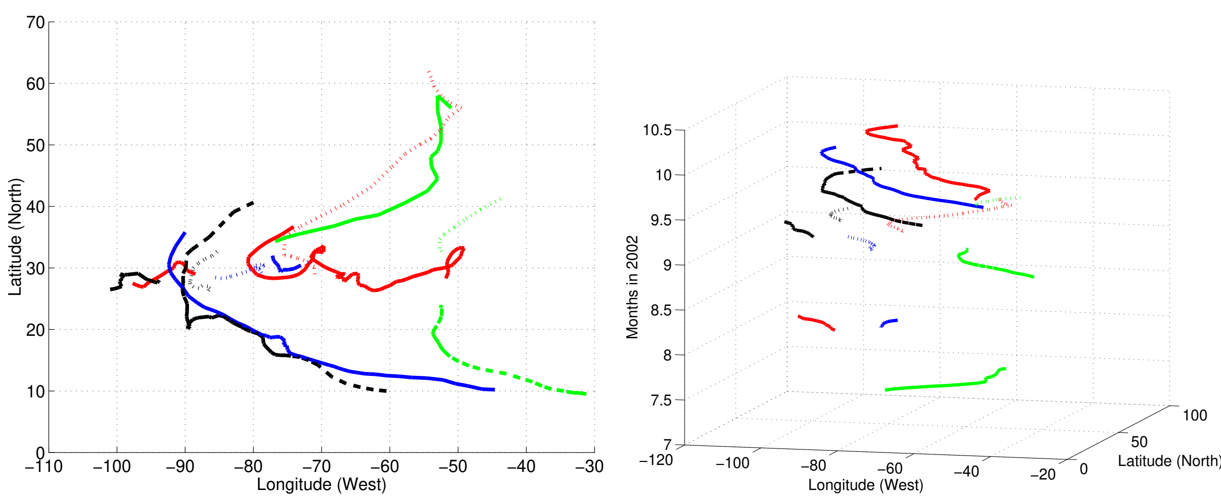
Public Buses The largest dataset in this survey stems from the tracking of public transport buses in the urban area of the city of Patras, Greece. The dataset is a result of tracking 13 buses using GPS receivers. The dataset consists of 28619 entries which were obtained by sampling the position of the vehicle at a regular interval of 30 seconds. The spatial extent of the data is 16 and 20 kilometers of Longitude and Latitude, respectively. The temporal extent is a 24 hour interval. To obtain the dataset, please contact the second author.



(a) seal

(b) turtle

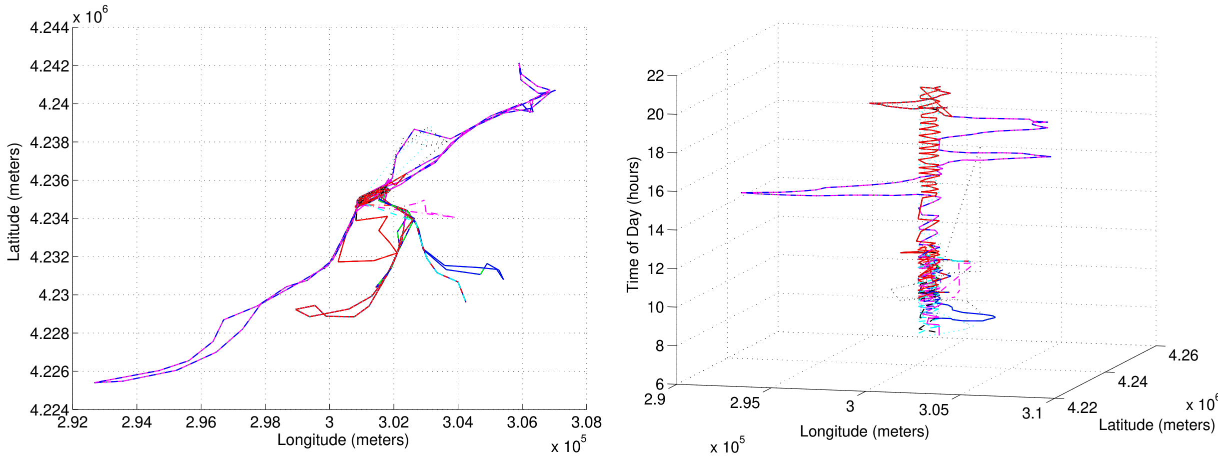
Figure 3: Animal tracking datasets



(a) spatial projection

(b) spatiotemporal representation

Figure 4: Hurricane dataset



(a) spatial projection

(b) spatial projection

Figure 5: Bus dataset

4 Future Work

One clear shortcoming common to all of the above tools is that they can only generate 2D spatiotemporal data. Although one would not be able visualize the generated data, it would be useful (and not as intuitive) to be able to generate datasets in higher dimensional spaces.

Such tools could be further improved to allow maintaining a (likely moderated) database of datasets generated, specially those used in publications. Some published papers simply mention the use of those tools without specifying details, which makes it hard (if not impossible) for someone to duplicate their datasets, defeating the very purpose of such tools.

GSTD and Oporto could be extended to allow the user to import real datasets to serve as the data space's infrastructure (G-TERD does allow this) and/or allow the user to create those by sketching them in the interface itself. Another useful enhancement could be to have objects aware of each other, e.g., one cannot get closer (or farther) than a predetermined distance. (Oporto allows this in the special case of objects belonging to different classes only.) Note that this would require some kind of embedded spatiotemporal indexing, which could be a plug-in method provided by the user him/herself.

Indexing trajectories seems to be a topic of growing interest, as such, the above tools could also be extended to generate trajectories following some particular specification, e.g., be contained within a pre-defined corridor.

Acknowledgments The authors would like to thank and acknowledge the following people who took part in GSTD's development over the past years: Jefferson R. O. Silva, Aggelos Kokorogiannis, Giannis Poulakis, Victor Salamon, and Daniel Mallett. The development of GSTD has been partially supported (at different times and by different means) by: Chorochronos Project (European Union); FAPESP, CNPq and FINEP (Brazil); and TRC, NSRC and Nykredit Corp. (Denmark). M. A. Nascimento is currently supported by NSERC Canada. D. Pfoser's research is supported in part by the Information Society Technologies programme of the European Commission, Future and Emerging Technologies under the IST-2001-32645 DBGlobe project, and the IXNILATHS project funded by the Greek General Secretariat of Research and Technology. Y. Theodoridis is also with the Data and Knowledge Engineering Group at the Computer Technology Institute, Greece (`ytheod@cti.gr`).

References

[1] T. Brinkhoff. A framework for generating network-based moving objects. *Geoinformatica*, 6(2):153–180, 2002.

[2] Caribbean Conservation Corporation/Sea Turtle Survival League. Sea turtle activity data, Web site: http://www.cccturtle.org/sat3.htm, 2003.

[3] D. Pfoser. Spatiotemporal datasets, Web site: http://dke.cti.gr/people/pfoser/data.html, 2003.

[4] O. Guenther et al. Benchmarking spatial joins *à la carte*. In *Proc. of the 10th Intl. Conf. on Scientific and Statistical Database Management*, pages 32–41, 1998.

[5] R.H. Gueting et al. A foundation for representing and querying moving objects. *ACM Trans. on Database Systems*, 25(1):1–42, 2001.

[6] A. Guttman. R-trees: A dynamic index structure for spatial searching. In *Proc. of 1994 ACM SIGMOD Intl. Conf. on Management of Data*, pages 47–57, 1984.

[7] C.S. Jensen et al., editors. *Proc. of the 7th Intl. Symp. on Advances in Spatial and Temporal Databases*, volume 2121 of *Lecture Notes in Computer Science*, 2001.

[8] G. Kollios et al. Indexing animated objects using spatiotemporal access methods. *IEEE Trans. on Knowledge and Data Engineering*, 13(5):758–777, 2001.

[9] D. Pfoser and C. S. Jensen. Capturing the uncertainty of moving-object representations. In *Advances in Spatial Databases, 6th International Symposium, SSD'99, Hong Kong, China, July 20-23, 1999, Proceedings*, pages 111–132, 1999.

[10] D. Pfoser and Y. Theodoridis. Generating semantics-based trajectories of moving objects. *Intl. J. of Computers, Environment and Urban Systems (Special issue on Emerging Technologies for Geo-Based Applications)*, 27(3):243–263, 2003.

[11] J.F. Roddick and K. Hornsby, editors. *Proc. of the 1st Intl. Workshop on Temporal, Spatial, and Spatio-Temporal Data Mining*, volume 2007 of *Lecture Notes in Computer Science*, 2001.

[12] J.-M. Saglio and J. Moreira. Oporto: a realistic scenario generator for moving objects. *Geoinformatica*, 5(1):71–93, 2001.

[13] R.T. Snodgrass and I. Ahn. Temporal databases. *IEEE Computer*, 19(3):35–42, 1986.

[14] Y. Theodoridis et al. Specifications for efficient indexing in spatiotemporal databases. In *Proc. of the 10th IEEE Intl. Conf. on Scientific and Statistical Database Management*, pages 123–132, July 1998.

[15] Y. Theodoridis and M.A. Nascimento. Generating spatiotemporal datasets on the WWW. *SIGMOD Record*, 29(3):39–43, 2000.

[16] Y. Theodoridis, J.R.O. Silva, and M.A. Nascimento. On the generation of spatiotemporal datasets. In *Proc. of the 6th Intl. Symp. on Advances in Spatial Databases*, pages 147–164, July 1999.

[17] T. Tzouramanis, M. Vassilakopoulos, and Y. Manolopoulos. On the generation of time-evolving regional data. *Geoinformatica*, 6(3):207–231, 2002.

[18] Unisys Weather. Atlantic hurricane data, Web site: http://weather.unisys.com/hurricane/index.html, 2003.

[19] WhaleNet. Satellite tagging data, maps and information, Web site: http://whale.wheelock.edu/whalenet-stuff/stop_cover.html, 2003.