J Geograph Syst (2005) 7:115–136
DOI: 10.1007/s10109-005-0153-8

Journal of Geographical Systems

Measuring similarity between geospatial lifelines in studies of environmental health

Gaurav Sinha and David M. Mark

Department of Geography and National Center for Geographic Information and Analysis (NCGIA), University at Buffalo, Buffalo, NY 14261, USA (e-mail: gsinha@acsu.buffalo.edu)

Received: 13 November 2003 / Accepted: 25 October 2004

Abstract. Many epidemiological studies involve analysis of clusters of diseases to infer locations of environmental hazards that could be responsible for the disease. This approach is however only suitable for sedentary populations or diseases with small latency periods. For migratory populations and diseases with long latency periods, people may change their residential location between time of exposure and onset of ill health. For such situations, clusters are diffused and diluted by in- and out-migration and may become very difficult to detect. One way to address the problem of diffused clusters is to include in analyses not only current residential locations, but all past locations at which cases might have been exposed to environmental hazardous. In this paper, we assume that a person's residential history provides such information and represent it through a discrete geospatial lifeline data model. Clusters of similar geospatial lifelines represent individuals who have similar residential histories—and therefore represent people who are *more likely* to have had similar environmental exposure histories. We therefore introduce a lifeline distance (dissimilarity) measure to detect clusters of cases, providing a basis for revealing possible regions in space-time where environmental hazards might have existed in the past. The ability of the measure to distinguish cases from controls is tested using two sets of synthetically generated cases and controls. Results indicate that the measure is able to consistently distinguish between populations of cases and controls with statistically significant results. The lifeline distance measure consistently outperforms another measure which uses only the distance between subjects' residences at time of diagnosis. However, the advantages of using the entire residential history are only partly realized, since the ability to distinguish between cases and controls is only moderately better for the lifeline distance function. Future work is needed to investigate

This project is supported by grant number 1 R01 ES09816-01 from the National Institute of Environmental Health Sciences, NIH. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIEHS or NIH. We wish to thank Peter Rogerson for helpful discussions of the migration models, and the anonymous reviewers for pointing out areas where the paper could be improved.

modifications to the inter-lifeline distance measure in order to enhance the potential of this approach to detect locations of environmental hazards over the lifespan.

Key words: Geospatial lifelines, Mobile objects, Spatio-temporal clustering

1 Introduction

Many aspects of human health are related directly to exposures to environmental toxins or hazards. Some of these environmental factors, such as food or tobacco smoke, relate to individual lifestyle and habits. Others, such as radon or air pollution, however, relate to the external human environment, either in buildings or outdoors in geographic space. Often, the locations or spatial concentrations of these environmental health hazards are not known, but spatial clusters or "hot spots" may nevertheless be found through spatial analysis and mapping of cases. When place of residence at the onset of ill health is used as the basis for mapping, the residential location is in effect being used as a surrogate for environmental exposure. However, attempts to relate disease clusters to fixed geographic environmental hazards are weakened by the fact that many people change their residential location through their life course: current residence may not provide the best estimates of lifetime exposure to environmental risks. Many health problems require long exposures to risk factors, and many, especially numerous forms of cancer, have long latency periods between exposure and onset of symptoms or health problems (Rogerson and Han 2002).

Discovering the environmental factors responsible for hot-spots and clusters for such diseases (e.g. many forms of cancer) is difficult because mobile populations tend to break up clusters of and obscure patterns of observable cases on account of different mobility patterns (Mark et al. 2000). Since human mobility rates influence exposure estimates and risks, mobility at different time scales (hour, day, year, life) corresponds to different categories of health problems. For example, for an emerging infectious disease such as SARS, details of the mobility of an infected person are needed on very fine temporal and spatial scales. On the other hand, for forms of cancer exhibiting relatively longer latency periods, hour by hour mobility may be irrelevant; broader patterns of residential or work-place history may be more critical to the identification of relevant environmental risk factors.

As can be deduced from the above discussion, the spatial distribution of cases of diseases at the time of diagnosis within non-sedentary populations does not provide sufficient information to estimate locations and times of environmental exposures. But in practice, most attempts to estimate environmental health factors related to cancer hot-spots or clusters still rely solely on place of residence at the time of disease onset and neglect the impact of the rate of mobility. A better alternative is to use the complete residential histories (if available) of people—by recording a person's residential history, it becomes possible to account for health risk exposures at past residential locations. The basic hypothesis then is that compared to snap-shot information of residences at time of diagnosis, accounting for the full spatio-temporal history of cases infected with diseases with long latency periods, is a more informed analytical framework for migratory populations.

If cases were clustered in the past, the locations of those clusters can be inferred from residential life-histories, even if most of the people moved away from the area of exposure before becoming ill.

In this paper, we use geospatial lifelines (Mark and Egenhofer 1998) to model residential histories within a geographic information systems framework. Clusters of geospatial lifelines represent individuals who have similar residential histories—and therefore represent people who are *more likely* to have had similar environmental exposure histories. A lifeline data model allows direct comparison of the space-time behavior of cases; measuring distances between lifelines will lead to the inference of potential clusters in the past which can be then be subjected to more rigorous two-dimensional cluster analysis techniques (Besag and Newell 1991; Kulldorff 1997). Any measure of distance applicable to lifelines, if correlated with health outcomes, may reveal locations of common risk factors that could in turn help reveal causal factors. Such space-time cluster analysis of lifelines based on residential address histories will be useful in the epidemiological investigation of environmentally induced diseases that need cumulative exposure or exhibit long latency periods.

In the following sections, we first discuss the ontological underpinnings of geospatial lifelines and residential histories. Next, we present a lifeline distance function that we subsequently use for a synthetic dataset to measure the similarity of geospatial lifelines. Since this is an experimental study to assess the efficacy of a new statistic (namely lifeline similarity), we do not use authentic data; instead we rely on simulations to both produce populations with realistic residential mobility and to expose them to highly-simplified spatially-localized risk factors. Evaluation of the distance function involves consideration of its ability to detect differences between populations of cases and controls, which would result in the real world due to different kinds, configurations and parameters of *environmental* risk factors; without real-world test data that includes a known source of exposure, the effectiveness of the method can be most readily assessed using simulation.

2 Ontology of geospatial lifelines

The lifeline data model (Fig. 1) is inspired by time-geography (Hägerstrand 1970, 1976; Parks and Thrift 1980); it is a representation of an individual's movement pattern in geographic space; Miller (1991) implemented many time-geography principles within a GIS environment. The term "geospatial lifeline" has recently been proposed to refer to the type of data that may be modeled using time-geography principles:

> "A geospatial lifeline is here defined to be the continuous set of positions occupied by an object in geographic space over some time period. Geospatial lifeline data consist of discrete space-time observations of a geospatial lifeline, describing an individual's location in geographic space at regular or irregular temporal intervals." (Mark and Egenhofer 1998).

Most of the work on moving object trajectory data models in the past (Vlachos et al. 2002a, b; Yanagisawa 2003) has been limited to the technical considerations of storage, retrieval and computational complexity, with little consideration of the actual phenomena being modeled. In this paper,

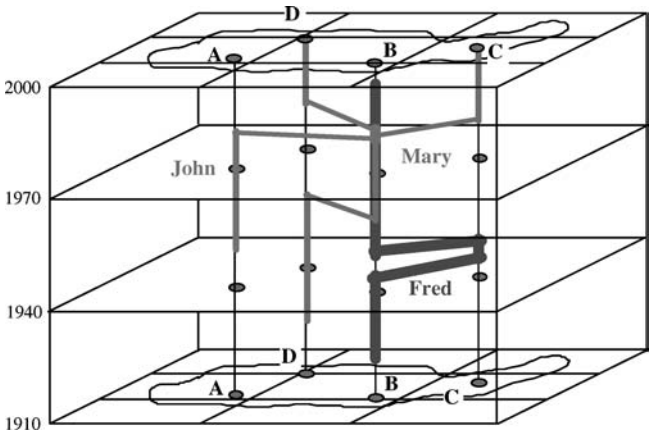

Fig. 1. Visualization of 3 lifelines in the space-time cube—with the x- and y-axes representing a 2-dimensional projection of geographic space and the oriented z-axis representing the progressing time

however, explicit attention will be accorded to the ontology of the phenomena being represented by the geospatial lifeline. This provides insights into the exact nature of an address history and how it is related first to a person's movement in space-time and then in turn to environmental exposures.

2.1 Objects

Objects comprise a very important part of our world. Although one might question the existence of an object at the molecular level, objects often exist unambiguously at the perceptual level of everyday action and reasoning (Gibson 1979). Detached objects have complete closed boundaries that separate them from their environments. Individual organisms, including people, are detached objects in this sense. Attached objects may be conceptualized as being objects, but can be modeled and represented as parts of the objects to which they are attached.

2.2 Mobile objects

Physical objects normally are considered to exist continuously in space and time. This means that if they move, they are assumed to occupy a connected series of intermediate positions between any two observed locations. If the object more or less maintains its shape as it moves, then we can separate the form from the location, and represent the locations of the object by the location of its centroid. As the object moves, the centroid must also occupy a continuous sequence of positions in space-time.

2.3 Addresses

It is important to note a number of things about the ontological status of a person's legal address. A postal address in the United States and in

many other countries is often, but not always, a unique identifier of a dwelling place or a building. In many societies, people normally have legal or home addresses; but since people do not normally spend all of their time at home, a person's address is not a perfect surrogate for his location. People's absences from their legal addresses occur at a variety of time scales. On a daily basis, many people in developed countries go to work or school five days a week, shop regularly, visit places for vacation or business purposes, etc.. They may also go away from home for extended periods, such as for higher education or military service, without changing their legal addresses.

2.4 Residential history data

Some comment is needed on the actual residential history data that are being collected in cancer studies. For example, our colleague Professor J. Freudenheim and her research group are conducting studies of breast cancer, and are including potential impacts of environmental exposure through spatial and spatio-temporal analysis (cf. Han 2002; Bonner et al. 2003). In these studies, cases and controls were asked to list all of their past residences along with other background information. Addresses were recalled from memory, and the quality of early life addresses might be somewhat suspect. Subjects were asked to list the start year and end year for each residence, which means that two (or more) addresses appear to apply to the entire year during which the move occurred. Such data characteristics would have to be taken into account when applying the measures presented in this paper to real data.

2.5 Differences in ontology of mobile objects and legal addresses

The ontology of address or legal address histories is not the same as the ontology of moving objects. This is not just a matter of granularity, scale, or resolution as some have framed the problem in the past (Hornsby and Egenhofer 2002). The geospatial lifeline for a real, bona fide, continuously existing object must be continuous (connected) in space-time. But a person's legal home address history might have gaps and perhaps even overlaps (two legal addresses at the same time). Different authorities might have different standards for legal address. Whereas moving objects move at speeds limited by the laws of physics and in practice by transportation technologies, a legal address can move great distances instantaneously. When a person moves as a physical object, he or she must occupy, however briefly, a connected set of places in between; whereas it would be inappropriate to think that a person's home address occupies positions in between the end points when the person move their residence. The address therefore refers to a particular place, but the legal address is a fiat entity and a virtual place. A residential address history is not merely a discretization or sample of their history as a mobile physical object.

3 Measures of similarity between geospatial lifelines

The main goal of the research reported in this paper is to develop a similarity measure for geospatial lifelines, and to test the power of that measure to detect differences between cases and a control group. The measures selected should be consistent with the ontology of residential histories of mobile populations. As noted earlier, trajectories are a special case of lifelines with a particular ontology, based on continuous motion, a property not found in residential histories. Hence it is not always appropriate to employ existing trajectory similarity operators for residential history data. Nevertheless, research on trajectory similarity is a good starting point for understanding the advantages and disadvantages inherent in different kinds of distance or similarity functions for spatio-temporal sequences.

A measurement theory for time geography has been proposed by Miller (2005). Yanagisawa et al. (2003) and Vlachos et al. (2002a, b, 2003) have previously introduced similarity operators for trajectory data. Such operators have been used to measure similarity in diverse contexts: stock market indices, animal movements, vehicular navigation paths, mobile phone or credit card usage, and many other kinds of temporally varying data. The design of similarity operators for trajectory data is motivated by one of the fundamental issues in data mining using time-series data: finding sequences which partially or fully match other sequences generally provided by the query (Park et al. 2000). In fact, time-series based similarity operators have been designed not only for spatio-temporal data but also for sequences defined in multidimensional attribute space (Lee et al. 2000; Keogh et al. 1999; Keogh, 1997; Das et al. 1997). In a data mining context, the design and success of such similarity functions is contingent not only on their ability to return the nearest neighbors (closest matches), but also on their algorithmic complexity and processing time.

3.1 Distance functions

While similarity can be measured directly, it is often more intuitive to measure first the distance (conceptual or physical) and then obtain a similarity measure through an inverse function. However, measuring distance between complex objects is often itself a complicated process in the context of data models used in modern information systems (cf. Okabe and Miller 1996). Many desirable properties for distance or similarity measures have been suggested by respective authors, but any one function can generally satisfy only some and not all of those properties. Hence the application, more than anything else, decides the design and choice of a particular distance function.

One property that is always desired, but sometimes difficult to achieve, is that of a metric—a function that gives a generalized scalar distance between two objects. With a metric distance function, it is possible to distinguish objects on an interval scale of measurement and then develop indexing schemes for databases. For a mapping $d : U \times U \rightarrow \mathbb{R}$, where U is the set of objects or data vectors, $\mathbb{R}$ is the set of real numbers, and $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ are data vectors defined in $\mathbb{R}^k$, the metric function $d(.,.)$ must satisfy the following four properties (Duda et al. 2001):

(i) Non-negativity: $d(\boldsymbol{x}, \boldsymbol{y}) \geq 0$;
(ii) Reflexivity (uniqueness): $d(\boldsymbol{x}, \boldsymbol{y}) = 0$, iff $\boldsymbol{x} = \boldsymbol{y}$;
(iii) Symmetry: $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$;
(iv) Triangle Inequality: $d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z}) \geq d(\boldsymbol{x}, \boldsymbol{z})$.

If d is such a distance metric, $\ln(d)$ and $-\ln(\max(d) - d)$ can also be treated as (metric) distance functions; $1/d$ and $\exp(-d)$ assume the status of (metric) *similarity* functions; if d is limited to finite values only, then $1 - d/\max(d)$ and $\sqrt{1 - d/\max(d)}$ are also similarity functions.

There have been many measures of similarity suggested in the past to calculate $d(.,.)$—the best known of all metrics is the Minkowski metric.

If $\mathbf{x} = (x_1 \ldots, x_i, \ldots\ldots, x_m)^T$ and $\mathbf{y} = (y_1 \ldots, y_i, \ldots\ldots, y_m)^T$ are two real vectors in $\mathbb{R}^k$, the Minkowski metric is calculated as follows:

$$M_p\langle x, y\rangle = \sum_{i=0}^{m} \left(|x_i - y_i|^p\right)^{1/p} \tag{1}$$

For $p = 2$, this yields the familiar Euclidean distance between vectors; for $p = 1$, the Manhattan or city-block distance is obtained and for the asymptotic case (e.g. $Lt(p) \rightarrow \infty$), we get the maximum value metric: $\max_i (|\, x_i - y_i \,|)$. For $p < 1$, this measure ceases to be a metric, because the triangle inequality is no longer valid. Veltkamp and Hagedoorn (2000) review many other properties and measures of similarity for pattern matching, most of which have generic appeal and can be adapted for a wide variety of application contexts.

3.2 Measuring lifeline distances

In this paper we introduce a lifeline distance function based on the Minkowski metric for measuring the distance between two lifelines. *Proximity in space* and the *temporal duration* of the proximity are the two essential parameters for the lifeline distance function. To incorporate both, a time-weighted distance function is described below.

$$d_1(L_1, L_2) = \frac{\sum_{t_i = t_0, t_j = t_1}^{t_1 = t_{n-1}, t_j = t_n} (t_j - t_i) * M_2\langle s_i, s_j\rangle}{\sum_{t_i = t_0, t_j = t_1}^{t_i = t_{n-1}, t_j = t_n} (t_j - t_i)} \tag{2}$$

where, t_i and t_j refer to times of successive moves by *either* lifeline; t_0 is the time at which the lifelines started overlapping; t_n is the time of diagnosis; $t_j - t_i$ is the duration that two people were separated by the distance between two location vectors $s_i = (x_i, y_i)$ and $s_j = (x_j, y_j)$; M_2 is the Euclidean distance operator (or a special case of the Minkowski metric for $p = 2$).

This distance function d_1 is defined such that it provides an intuitive way to measure the distance between two lifelines. The function essentially is a weighted average of successive separation distances between two residences, where the weights are the durations a particular separation distance was maintained before either one or both residences were changed. The range of this function spans from a minimum of *0* to the maximum physical distance d_m manifestable in the geographic domain of interest. Researchers interested

in generic similarity measures, rather than in distances, can additionally apply mapping functions like $\exp(-d)$ or $1 - d/d_{\max}$ to force the function to evaluate always between *0* and *1*.

It is necessary to point out here that the distance function introduced in Eq. 2 is *not* a metric. For example, if this function is used to measure distances between lifelines, two separate individuals (e.g. parent and child) with different movement histories before start of overlap at $t = t_0$ but identical movement histories after start of overlap, will have zero lifeline distance between them; hence the reflexivity (property (ii) above) is not satisfied.

Hence, Eq. 2 can be further modified, if desired, to penalize the time of non-overlap to distinguish between identical lifelines and partially identical lifelines (e.g. identical after start of overlap). The modified version of this distance function is defined in Eq. 3 below.

$$d_2(L_1, L_2) = \frac{\left[\sum_{t_i=t_0, t_j=t_1}^{t_i=t_{n-1}, t_j=t_n} (t_j - t_i) * M_2\langle s_i, s_j\rangle\right] + T * E_{M_2}[L_1, L_2]}{\left[\sum_{t_i=t_0, t_j=t_1}^{t_i=t_{n-1}, t_j=t_n} (t_j - t_i)\right] + T} \tag{3}$$

T is the total duration of time for which either of the lifelines existed but did not overlap with the other and $E_{M2}[L_1, L_2]$ is the expected Euclidean distance between any two randomly generated points in the study area. In future, the measure could be further modified to give differential weights to different time periods.

The expected distance between lifelines is generally difficult to calculate theoretically if the study area is not of a regular geometric shape. Even for regularly shaped areas such as rectangles, the mathematical formula for the expected distance between points is difficult to solve exactly (Lazoff and Sherman 1994). Hence, in practice, the expected distance can be calculated through Monte Carlo simulations, by randomly generating a large ($N > 1000$) number of points in the study area and creating a distribution of the distances between pairs of randomly selected points; the mean of the distribution will approximate the expected distance closely if the sample size is large. Note that if lifelines belong to the same cohort, the second terms in both the numerator and denominator evaluates to zero (i.e. $T = 0$) in which case d_2 resolves to d_1.

4 Simulation of lifelines

4.1 Why simulation?

While using authentic real-world data would have benefits, synthetic (simulated) data offer a better basis for understanding the behavior of a new method of analysis, because they can be generated for a wide range of parameter values and forced to operate in specific regions in parameter space. Simulation allows the researcher to control the 'true' pattern of environmental influences on health, thus enabling the verification and calibration of methods to infer such differences given only residential history data on cases and controls. Similarly, the efficacy of the lifeline distance function in clustering similar lifelines (i.e., cases due to the same environ-

mental hazard) as opposed to dissimilar lifelines (i.e., controls who have not been exposed to the same hazard), must be evaluated rigorously for a wide range of space-time configurations of exposures and lifelines, and this would not be possible with real data. Hägerstrand (1970) himself suggested that "reasonably good simulations should improve our ability to survey whole systems and help to reduce the considerable trial and error component in applications," and Parks and Thrift (1980) stated that "simulation is seen as a means to a sharper appreciation of the possibilities open to individuals and population aggregates and is generally preferred to inductive, sample survey techniques".

There are many other practical constraints that make use of simulated data attractive to us. Residential histories generated through interviews and questionnaires may have missing components due to failing memories or incomplete surveys, in which case either the distance function has to be modified to reason with incomplete histories, or residential histories would have to be interpolated before the distance function can be applied without modification. Obtaining samples may also be prohibitive economically, since considerable resources are required to conduct surveys of sample sizes appropriate for statistical analysis. Also, the measure of similarity would have to be evaluated with many different samples from different mobile populations to achieve significant confidence about predictions made regarding the likelihood of a test subject as a potential case.

Lastly, in the bio-medical domain, real case-control data carry serious confidentiality constraints. Colleagues conducting medical research prefer not to provide access to such data until the potential utility of analysis methods has been demonstrated. Consequently, we use only synthetic data to specifically control where to evaluate the lifeline distance function in the exposure-lifeline parameter space.

4.2 Simulation procedure

In this study, we employ two different methods for simulating populations of lifelines, one based on successive residential locations being completely independent and random within the study area (a typical 'null hypothesis' approach), and the other based on a modified random walk model based on an exponential model of 'migration move distances' observed in the US for migratory populations (Rogerson et al. 1993). We generated training datasets for both kinds of lifeline patterns.

In order to test the similarity measures, we also must simulate environmental exposures. In this paper, we have adopted an extremely crude and unrealistic model of exposure and ill health. Exposures at work place and from foods and similar sources are ignored, and all exposures to the simulated risk are assumed to be at the place of residence. Furthermore, for this initial study, the risk area is considered to be a circle of fixed location and radius. Lastly, we assume a deterministic relation between exposure and ill health—all simulated individuals who live in the risk area for more then the arbitrary threshold of years are classified as cases, and all those individuals who for any reason do not meet the criterion are classified as controls.

Each model of residential mobility was then combined with three different variations on the exposure model. The lifeline distance function defined in Eq. 3, was then used to measure the expected case-to-case, case-to-control, and control-to-control lifeline distances. Results are analyzed to show that the lifeline distance function is more efficient in distinguishing cases from controls, than snap-shot based two-dimensional analysis of distances at the time of diagnosis.

4.2.1 Random-positions (RP) migration model

This method of simulation assumes that with each change in residence, a person moves to any other point in the study area with equal probability. All individuals are equally likely to move in a given time period; all lifelines and all moves within a lifeline are independent of each other and the direction and distance of the move is constrained only by the extent of the study area. In the simulation, the probability of a move is evaluated only once each year; hence the minimum temporal granularity is one year. The pseudo-code in Fig. 2 was used to generate random lifelines for a rectangular shaped study area. The pseudo code ensures that no location or move time is recorded twice for a lifeline, that the moves are ordered in increasing temporal order and that all locations are generated from a uniform random distribution that produces all abscissa and ordinate values with equal probability for the defined abscissa and ordinate range. Obviously this is an unrealistic model of actual human behavior, but it provides a sort of null hypothesis against which to evaluate results obtained from a more realistic model of residential histories, discussed next.

4.2.2 Exponential-distance (ED) migration model

This method of simulating lifelines is more realistic; it also uses a random walk model of residential mobility but chooses move probabilities and move distances from empirically-derived patterns of residential mobility in the United States taken from Plane and Rogerson (1993). The implementation uses a exponential model of move distances for individuals in different age-groups described in Yang (2001). The method involves certain assumptions for making the simulation tractable:

(i) individuals either belong to the same age-cohort or their age composition has the same cumulative distribution as published by the Census Bureau of the United States;
(ii) the annual probability of movement is dependent on age only; (the probabilities are obtained from Plane and Rogerson, 1993);
(iii) for each individual's movement, the moving distance is calculated as in Eq. 4; this is derivable through simple algebraic manipulation from Eq. 5,

$$x = -\ln(1-y)^{*}b \tag{4}$$

$$y = F(x) = 1 - e^{\frac{-x}{b}} \tag{5}$$

where b is the median moving distance, y is the cumulative probability of a move, and x is the moving distance (Rogerson, et al. 1993);

(iv) the moving direction is randomly distributed between 0 to 360 degrees.

All moves are restricted to end within a rectangular study area, so that if a simulated move would have taken the person outside the study area, that move is not made and another potential move is simulated. The minimum time interval between moves was set to one year, as for the random positions simulation. The pseudo-code that was used for simulation for this case study is provided in Yang (2001).

4.2.3 Environmental exposure

For purposes of this study, we generated environmental health risks through a deterministic model. Simulated people were labeled cases if they lived for sufficient time within an area of risky environmental conditions. We further simplified the model by assuming that the risk area was a circle based on a fixed distance from a point source of risk. The environmental risk region thus was modeled as a static, 3-dimensional space-time cylinder of constant radius r and height h; the radius r of the cylinder determines the spatial extent of the exposure at any given time and the height determines the duration of the exposure. Simulated people whose residential lifelines were generated by each of the above methods were then classified as cases or controls depending on their spatio-temporal interaction with the risk area. Any simulated person that spends cumulatively more time (not necessarily continuously) than the minimum required for contracting the disease is labeled as a case; a simulated person whose lifeline spends *less* than the threshold time inside the exposure cylinder is treated as a control, unaffected by the exposure.

It is important to state the assumptions we have made in designing the study—we assume that the exposure affects all individuals similarly, and that the intensity and spatial dimensions of the environmental hazard remain unchanged throughout its existence. In the real world, these assumptions may be violated, but these assumptions greatly simplify the investigation of the lifeline similarity measure, without compromising much on the generalizability of the measure.

```
for i = 1; i < #moves {
  do {
    for (x_i, y_i ϵ study_area(min_x, min_y, max_x, max_y)) and (t_0 < t_i < t_n)
                                                    {
      x_i = rand(min_x, max_x);
      y_i = rand(min_y, max_y);
      s_i = (x_i , y_i) ;
      generate t_i-1 < t_i < t_n;
    }
  } while  (s_i, t_i) != unique((s_i, t_i)
```

Fig. 2. Pseudo Code for generating lifelines by the Random-Positions (RP) method

4.2.4 Monte Carlo simulation of cases and controls

The Monte Carlo simulation method is applied here to generate a large number of cases and controls, based on each of the two migration models (random positions and exponential) for statistical evaluation of results. The threshold time for becoming a case could be varied depending on which disease is being investigated. The generation of a large number of controls and cases may become an issue if the ratio of the volume of the cylinder and the space-time study area is very small. This is because the Monte Carlo simulation method employed here generates a cohort of lifeline cases and controls in exactly the same way; only the random interaction of the lifeline with the exposure when evaluated after the full generation of the lifeline is used to classify it is as a case or control. Since normally the area of high risk is significantly small compared to the study area, the requisite number of controls can be generated with relative ease; on the contrary, many lifelines generally will have to be simulated and rejected as controls before the requisite number of cases can be generated.

4.2.5 Simulation parameters

Both the random position and exponential-distance random walk migration models were used to generate lifelines for three different exposures. To simplify the modeling in this proof-on-concept study, we chose to model only a single cohort (age at conclusion = 70 years) for all simulations to avoid complications arising due to inaccurate population composition when all age groups would be considered. Six situations were run, based on three different exposure risk region sizes, for each of the two migration models were generated, and the case-case and control-control distributions of lifeline distances were calculated using Eq. 2. The details of all the simulations are as follows:

(i) *Simulation Models*: Random Positions (RP) and Exponential-Distance (ED) Random Walk models.
(ii) *Distance Function*: d_1(Eq. 2).
(iii) # *Cases*: 2000.
(iv) # *Controls*: 2000.
(v) *Study Area*: 200 X 200 km^2 rectangle.
(vi) *Age of all cases and controls at end of study period*: 70 years.
(vii) *Median Distance b (for ED migration model)*: 11.2 km.
(viii) *Exposure threshold for contracting disease*: 10 years.
(xi) *Exposure cylinders* $< r, h >$: < 5 km, *70* years$>$; $<$10 km, 70 years$>$; $<$10 km, 20 years$>$

5 Results

5.1 Comparing distributions

Both the Random-Positions (RP) and Exponential-Distance (ED) methods of lifeline simulation were used to evaluate the lifeline distance function d_1 for a cohort of simulated subjects who all were born in the same year and

who lived to be 70 year olds. Figure 3a, b display the distribution of the lifeline-distance statistic for 1000 pairs of cases and controls generated by the RP method; Figs. 3c, d display the same for cases and controls generated by the ED method. From visual inspection, the distributions in Fig. 3a, b closely resemble a normal distribution while that in Fig. 3d do not. The distribution in Fig. 3c resembles a normal distribution based on visual analysis, but given the large sample size, statistics indicate that the distribution in Fig. 3c also is significantly different from normal. This can be verified from Tables 1 and 2, where the skewness of the RP distributions is almost zero, but is much higher for cases under the ED model of migration.

The one sample Kolmogorov-Smirnov (K-S) composite goodness-of-fit (GOF) test was used to test the null hypothesis that each empirical distribution was similar to a normal distribution with mean and standard deviation as estimated from the samples for an exposure of 10 km radius and that had a 70 years long presence from 1930 to 2000. The null hypothesis

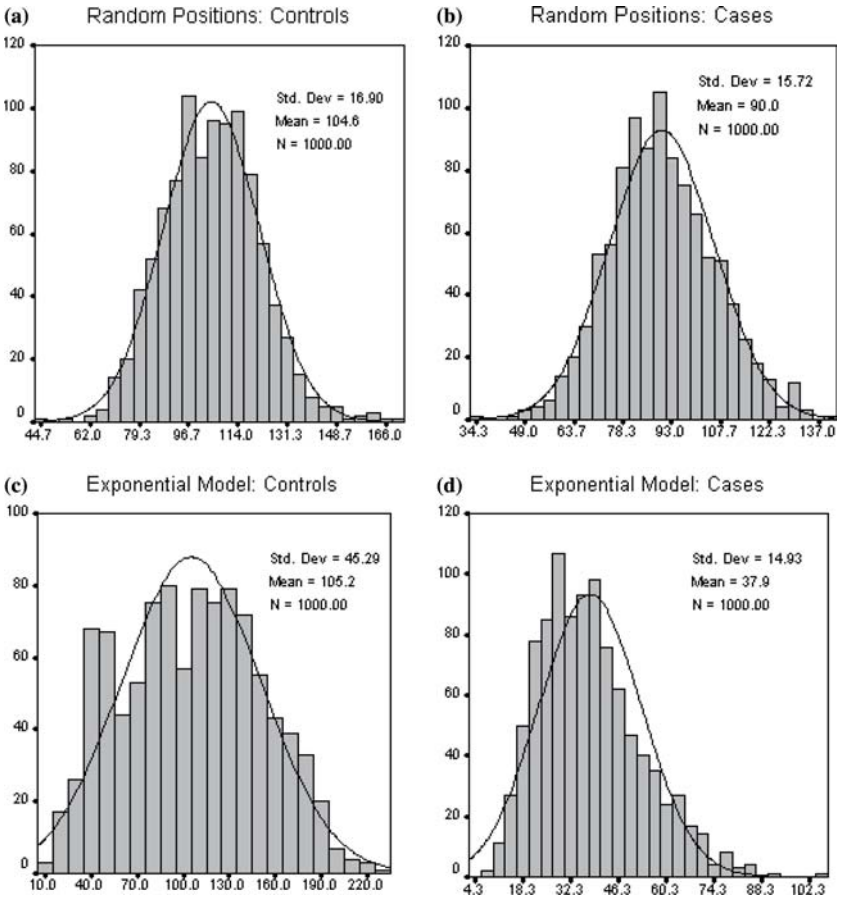

Fig. 3. Distributions of lifeline distance statistic for cases and controls for RP and ED migration models for simulating lifelines for an exposure cylinder of 10 km radius and 70 years duration (1930–2000)

Table 1. Statistics for lifelines generated by the RP migration model

Exposure type	Controls (RP)			Cases (RP)		
	Mean (km)	Std. Dev. (km)	Skewness	Mean (km)	Std. Dev. (km)	Skewness
1930–2000; 10 km radius	104.62	16.90	0.250	90.03	15.71	0.175
1930–2000; 5 km radius	104.79	16.35	0.115	88.83	15.66	0.011
1950–1970; 10 km radius	104.80	17.04	0.112	80.64	16.56	0.097

Table 2. Statistics for lifelines generated by the ED migration model

Exposure type	Controls (ED)			Cases (ED)		
	Mean (km)	Std. Dev. (km)	Skewness	Mean (km)	Std. (km)	Dev.Skewness
1930–2000; 10 km radius	105.21	45.29	.098	37.92	14.93	.719
1930–2000; 5 km radius	102.96	44.99	.177	33.58	13.52	.898
1950–1970; 10 km radius	105.45	44.67	.206	28.36	10.61	.822

could not be rejected for either of the two RP distributions (p-value = 0.5) but was strongly rejected for the two ED distributions (p-value $\approx$ 0) [1]. Repeating the test for the other two exposures and for other cohorts (age = 30, 50, 90) reproduced these results, albeit at slightly different levels of significance (*p*-values). The null hypothesis that RP and ED models generate sampling distributions from the same population was also rejected by the two-sample K-S test (*p*-value $\approx$ 0), thus additionally verifying that the difference between the (mean) lifeline distances calculated for RP and ED migration models is statistically significant.

These results show that the lifeline distance function d_1 can be used to distinguish between different migration patterns, just from an examination of the distributions of case-case or control-control lifeline distances.

5.2 Comparing statistics

Tables 1 and 2 show 12 different sets of statistics that were calculated for control-control and case-case distance measurements ($n = 1000$) for 3 different exposure cylinders for the RP (Table 1) and ED (Table 2) migration models.

[1]The mean and standard deviations were obtained from the samples. This is allowed in the K-S test, if the reference distribution is normal (MathSoft, 2000).

The mean lifeline distance for controls in both tables is very similar for all exposures and is actually very close to the expected value of 104.28 km between two randomly generated points in a 200 km $\times$ 200 km rectangle (we used Lazoff and Sherman's (1994) formula for calculating this expected distance). This result is to be expected because of the following. Our simulation procedure randomly distributes in the study area the origins of all lifelines for both simulation methods (RP and ED). The lifelines then 'evolve' (non-randomly) in space-time as controlled by the parameters of the particular migration model. As explained earlier, the lifeline distance d_l can be interpreted as duration-weighted average of Euclidean distances between pairs of residences. The mean lifeline distance d_l between pairs of randomly selected control lifelines (these will have a random separation vector throughout on account of their random separations at birth) should therefore approximate the theoretically expected Euclidean distance between any two points, randomly located in the same area as used for constraining the sample space of spatial locations for lifeline nodes.

The real power of the lifeline distance is revealed when statistics for cases and controls are compared. Foremost, it can be seen that mean lifeline distances for cases are always *lower* than that for controls for both RP and ED migration models. The statistics for controls and cases, for lifelines generated by the ED migration model, give a strong indication that the lifeline distance measure is able to distinguish clearly between lifelines which have had similar exposure history compared to those who have not been exposed to the same environmental hazard (exposure cylinder in our case study). The Kolmogorov-Smirnov two sample test (MathSoft 2000) for comparing empirical distributions indicates (p-value ≈ 0) that cases and controls have different empirical distributions for all exposures and for both simulation methods.

For the ED migration model, the mean lifeline distance for cases is much smaller than that for controls for all exposures—this is expected intuitively because the median distance of any move for a lifeline is only 11.2 km (i.e., $b = 11.2$ in Eq. 4) and therefore lifelines diverge relatively more slowly (and more realistically). Hence, if lifelines have to have similar exposure histories, then they will also have to originate relatively close in geographic space. Thus cases for the ED simulations tend to cluster strongly *over their entire lives* and not just during exposure. Controls also diverge equally slowly, but since they will tend to originate anywhere in the study area, their mean lifelines distances are much higher than that of controls.

It can also be observed, from comparing distribution means from Tables 1 and 2 that, for the RP migration model (Table 1), the distribution for both cases and controls are characterized by higher mean lifeline distances as compared to ED distributions. For example, for the same exposure (1930–2000; 10 km radius) the difference of means for cases and controls is only 14.59 km for the RP model, while it is 67.29 km. This can be explained briefly thus: if a lifeline has to be exposed for at least 10 years to be labeled as a case, it tends to spend much more time near the exposure than a corresponding RP lifeline—this is because a lifeline from the RP simulation can migrate far away from the exposure with one move, while a lifeline from the ED simulation will generally take several moves to migrate to large distances away from the exposure site.

Hence, it is relatively more difficult to distinguish between RP cases and controls than between ED cases and controls. This means that for populations, which are sufficiently mobile and whose successive moves can be characterized as practically random (and hence can be simulated by the RP method), the distinction between cases and controls will be masked (if we were to use only the lifeline distance function for distinguishing between them). The degree of masking will be dependent on the extent of the randomness, the move vectors and the rate of movement. However, the distinction between cases and controls gets stronger as the cylinder gets smaller, since the 'opportunity' to be exposed decreases and cases must be closer in space-time to share the smaller exposure cylinder volume.

5.3 Use of lifeline distances at time of diagnosis only

It is also important now to re-assess the need for the lifeline distance operator—it was suggested because the spatial distribution of the cases at the time of diagnosis is not useful in retrodicting the exposure clusters in the past. To verify this, for the same set of simulated data as simulated by the ED method, we used a new time-of-diagnosis distance function, described in Eq. 6, to distinguish cases from controls,

$$d_{t_n}(L_1, L_2) = M_2 < s_{n1}, s_{n2} > \tag{6}$$

where t_n is the time of diagnosis; s_{n1} and s_{n2} are the locations of the individuals at t_n, as obtained from lifelines L_1 and L_2 respectively and M_2 is the Euclidean distance operator. This function considers only the last pair of distances (i.e., at the time of diagnosis) from the pair of lifelines, to create a distribution for controls and cases; statistics generated for this function are compared to that from the lifeline distance function from Eq. 2.

In this section we will compare results only for the Exponential-distance migration method of simulating lifelines. Initially all parameters (e.g. study area, exposure cylinders, mobility rate, cohort age, number of cases and controls) for the simulation remained the same, except the use of the new distance function. The first set of simulations was conducted for all three exposure cylinders used earlier in Tables 1 and 2. The statistics obtained are shown in Table 3 and the sample distributions of cases and controls for one exposure (70 years, 10 km radius) are shown in Fig. 4 (below). The distribution is obviously not normal and visual inspection of the cases

Table 3. Statistics for d_{tn} for the ED migration model lifelines

Exposure type	Controls (d_{tn})			Cases (d_{tn})		
	Mean (km)	Std. Dev. (km)	Skewness	Mean (km)	Std. Dev. (km)	Skewness
1930–2000; 10 km radius	104.54	47.91	0.112	38.55	26.85	0.98
1930–2000; 5 km radius	103.29	48.11	0.167	33.77	24.76	1.08
1950–1970; 10 km radius	102.41	48.30	0.159	27.08	18.87	1.51

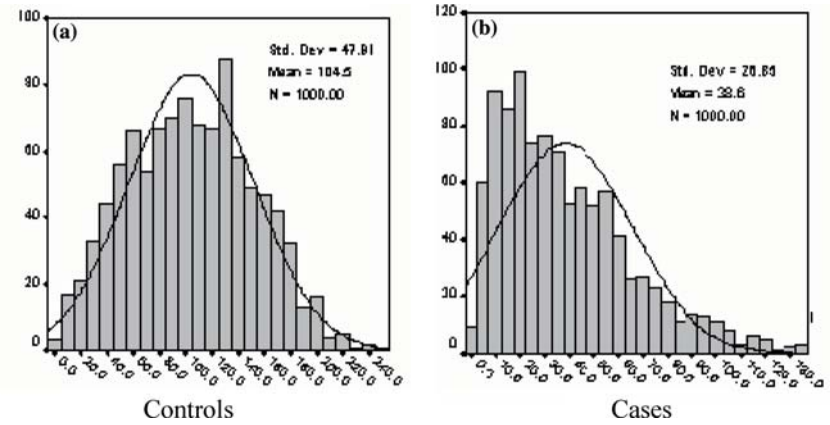

Fig. 4. Distributions for d_{tn} (distance at diagnosis) for cases and controls for ED migration model lifelines for an exposure cylinder of 10 km radius and 70 years duration

distribution indicates similarity with some form of an exponential distribution. The statistics and distribution for controls are almost identical to that obtained for the lifeline distance function d_1 (Table 2). For cases, the means are almost identical for the two distance measures (Eqs. 2 and 6), but the dispersion (std. dev.) and skewness is higher for d_{tn} (Table 2).

Thus results indicate that the two distance functions d_1 and d_{tn} *differ only in distribution shape*; they are *almost identical in terms of central location statistics*. This would mean that the distance function designed to exploit complete residential histories is after all just as good as the distance at time of diagnosis—which defeats the whole purpose of using a lifeline based statistic. To investigate further, we simulated a new set of lifelines, with the median move distance doubled to 22.4 km. The cohort age was maintained at 70 years; three exposure cylinders were used, two of which were the same as before. 100 unique pairs of comparisons (100 controls and 100 cases) were made, statistics for which are shown in Table 4. (The statistics and distributions for controls were similar to that observed in Table 2 and are omitted here).

Table 4. Comparison of lifeline distance function and the time of diagnosis distance function for the ED migration model lifelines for median move distance = 22.4 km

Exposure type	Cases (d_1)			Cases (d_{tn})		
	Mean (km)	Std. Dev. (km)	Skewness	Mean (km)	Std. Dev. (km)	Skewness
1930–2000; 10 km radius	59.52	19.12	0.27	55.46	36.76	0.80
1950–1970; 10 km radius	64.94	18.05	0.45	85.00	44.55	0.53
1955–1960; 2 km radius[a]	55.95	18.75	0.29	68.24	33.25	0.11

[a]The threshold for getting ill is set to 2 years for this small exposure cylinder

Table 4 indicates that if the exposure has a large space-time presence, such that people can get affected over a wide range of time, the lifeline based distance function is almost just as bad (or good) as the distance at time of diagnosis. On the other hand, as the exposure gets relatively smaller, such that simulated cases have to be more clustered to become exposed, the lifeline distance case-case distribution does center itself farther away from the control distribution and is easier to distinguish. Table 4 suggests that d_l is only marginally better than d_{tn}. Comparison of standard deviations for the two statistics indicates that d_l has much less uncertainty associated it than d_{tn}. This might be a crucial fact, as this increases the power of d_l to reject false cases.

Before concluding, we briefly discuss the results from another set of simulated data for the same study area and the median move length of 11.2 km for the ED model of migration. We measured the means and standard deviations for the two distance functions for 5 different age cohorts (Table 5). Now, just as is common for a random walk phenomenon (the exponential distance model of simulation is similar to a floating random walk process as opposed to grid based equal length walks) we see in Table 5 that with increasing age people tend to be farther from each other (all lifelines for all cohorts started at the center of the study area). The rate of separation is the maximum in the twenties and then slows down considerably around 50 years of age (this is due to the particular age based mobility rates used for the ED model simulation from Plane and Rogerson 1993). If we were to interpret the cohort ages in Table 5 as the ages of diagnosis, we can make two observations:

i) the mean for the distances at diagnosis is always larger than the mean lifeline distances for the same pairs of lifelines and
ii) with increasing cohort age, the merit of the time-averaged distances as used in the lifeline distance function becomes more pronounced because the distinction between cases and controls is more difficult to make due to similar central location values.

6 Discussion of results

The use of two different methods of simulation in this case study is important when evaluating the robustness of the lifeline distance function. The ED model, although much simplified through its assumptions, is still based on actual data regarding human residential mobility in the United States (Rogerson et al. 1993)—it therefore should simulate the actual distribution of residential history lifeline distance statistics much more faithfully than the RP migration model, which imposes no realistic constraints on the migration of an individual, except that they must stay in the 200 by 200 km square study area. Due to the lack of constraints, and as statistics indicate (Tables 1 and 2), the RP model makes it much more difficult to distinguish between cases and controls, especially as the exposure cylinder grows in volume. Therefore, if the lifeline-distance statistic is robust enough to detect differences for this model, then it is likely to be that much more effective in the case of real data as well.

Our results indicate that the lifeline distance measure d_l is able to consistently distinguish between the distributions of cases and controls for

Table 5. Comparison of lifeline distance function and the time of diagnosis distance function for the ED model based lifelines for different age cohorts

Cohort age (years)	Mean (d_1) (km)	Std. Dev(d_1) (km)	Mean (d_{tn}) (km)	Std. Dev(dtn) (km)
10	21.52	12.66	34.52	19.66
30	37.78	16.65	55.63	29.14
50	47.56	22.17	63.73	34.52
70	53.19	23.94	68.65	36.10
90	55.35	24.52	69.16	35.72

both simulation methods. However, the efficiency varies with the exposure size and duration. Results are much more promising for populations generated by the more realistic simulation in which people migrate relatively shorter distances than in the case of a purely random migration process. It is also true that the performance of the lifeline distance function deteriorates as the movement pattern becomes randomized because cases become less clustered relative to controls as the spatio-temporal extent of the exposure risk area becomes larger. Interestingly, the standard deviations for RP lifelines (Table 1) are considerably smaller than those obtained for the ED Model lifelines (Table 2). Thus, in the case of the ED lifelines, the statistical power of the lifeline distance statistic is reduced somewhat because of the higher variability.

When we compared the lifeline distance function with the time of diagnosis distance function, we found *no significant benefits* for the former—until we increased the median move length to twice what it was before and reduced the critical exposure time from 70 years to 20 years. From this result, it would seem reasonable to propose that the lifeline distance function will be better than the distance at diagnosis function at *higher mobility rates*. This would be especially true if lifelines diverged substantially after exposure; this scenario is observable for small exposure cylinders experienced sufficiently early on in life when mobility rates were still high so that lifelines could separate far from each other by the time of diagnosis. However, there is a caveat to be raised here—for the RP method (randomized movement histories) higher mobility and smaller exposure cylinders make the difference between cases and controls distributions more difficult to detect for the lifeline distance statistic. Hence for d_1, *there is an inverse relationship between power to distinguish between cases and performing better than* d_{tn}.

7 Conclusions and future work

The lifeline distance function was designed in the hope that it would be able to detect similarities in the patterns of residential histories of people and find groups of people who have clustered sometime in the past, well before time of diagnoses. Combining the information from Table 5 with the prior discussion on the performance of the lifelines, we can offer two favorable situations in which the lifeline distance function is a better function than the simple distance at time of diagnosis, for classifying cases and controls:

(i) Cases cluster sufficiently early on in their lives near an exposure and then continue spreading far from each other after exposure (to diffuse the cluster with increasing time). In such a scenario, mean time of diagnosis distance (d_{tn}) between cases is significantly larger when compared to mean lifeline distance (d_l). This is because d_{tn} in this case will be reflective of the increased distances between cases at the final time of diagnosis, whereas d_l, because of its time-weighted distance averaging characteristic, is a relatively more robust measure in the case of diffusion of cases, subsequent to clustering near an exposure. Note that if the exposure is long persistent, such that cases can be exposed over a wide duration, the suggested lifeline distance function is not as effective.
(ii) Highly mobile populations will make the case-control distributions more similar but they will also make the use of the lifeline distance function more effective relative to simple distance at time of diagnosis

However, a good lifeline distance function must be generalizable to all scenarios. Hence, the lifeline distance function must be modified to make it much more robust to case-control distinctions for a wide variety of exposure histories. For example, Vlachos et al. (2003) use a modified Euclidean-distance-based similarity measure which makes the similarity measure more robust to unusually large distances (outliers) between lifelines. Another way to discount these outliers would be to use the geometric or the harmonic means instead of the arithmetic means used to average the distances for a pair of lifelines. The geometric mean is however not applicable if the lifelines intersect, unless the minimum distance is set to a non-zero threshold.

Yet another way to modify the lifeline distance function is to limit the time-weighted averaging to only the lower quartile of distances generated during a lifeline, thus eliminating separation vectors larger than a threshold magnitude. Thus we can preferably use only those distances which characterize the scale of clustering for an exposure (given that we have a hunch of the size of the exposure). Similarly, lifeline distances could be calculated only for specific time windows, if some information is available about when likely exposures might have been encountered or some temporal regions can be eliminated with certainty.

It may also be useful to develop visualization schemes for lifelines that can afford visual detection of similar lifelines. This will be difficult when the number of lifelines is large. One could also count the number of years for which the lifelines were within threshold of each other, where the threshold distance might relate to typical sizes of exposure areas. In this paper we also did not present the results of validation by testing with new cases after building characteristic distributions for cases and controls for a given exposure.

Finally, one thing should be kept in mind: if lifelines do not tend to diverge much, i.e., the sampled population has been not very mobile, the simple distance between cases at the time of diagnosis will be as good an indicator as the lifeline distance function.

Future work therefore must include more new lifeline distance functions and must simulate new cases for a given exposure and then set up a misclassification matrix. The function that classifies new cases most

accurately will be the best measure. It may be the case that different measures on account of different distribution shapes and statistics might exhibit different classification accuracies for different exposure histories. Exploring classification accuracies of different functions for different exposure types, population types, migration vectors and mobility rates will help us significantly improve our capability to reason about how environmental hazards affect different cohorts.

References

Besag J, Newell J (1991) The Detection of Clusters in Rare Diseases. *Journal of the Royal Statistical Society,* A154(Part 1):143–155

Bonner MB, Han D, Nie J, Rogerson P, Vena J, Freudenheim JL (2003) Positional Accuracy of Geocoded Addresses in Epidemiologic Research. *Epidemiology*, 14:408–412

Das G, Mannila H (1997) Finding Similar Time Series. In: *Proceedings of the Conference on Principles of Knowledge Discovery and Data Mining*.

Duda RO, Hart PE, Stork DG (2001) *Pattern Classification*. 2nd ed John Wiley and Sons, New York

Gibson JJ (1979) *The Ecological Approach to Visual Perception,* Houghton-Mifflin, Boston

Hägerstrand T (1970) What About People in Regional Science?. *Papers of the Regional Science Association,* 24:1–21

Hägerstrand T (1976) The Time-Space Trajectory Model and its Use in the Evaluation of Systems of Transportation. In: *International Conference on Transportation Research,* Vienna, Austria

Han D (2002) *Geographical epidemiology of breast cancer in Western New York: Exploring spatio-temporal clustering in GIS*. Unpublished PhD dissertation, Department of Geography, University at Buffalo, December 2002

Hornsby K, Egenhofer M (2002) Modeling Moving Objects Over Multiple Granularities. *Annals of Mathematics and Artificial Intelligence,* 36:177–194

Keogh E (1997) A Fast and Robust Method for Pattern Matching in Time Series Databases. In: *9th International Conference on Tools with Artificial Intelligence (TAI '97)*

Keogh E, Pazzani MJ (1999) Scaling Up Dynamic Time Warping to Massive Datasets. In: *3rd European Conference on Principles and Practice of Knowledge Discovery in Databases,* pp 1–11

Kulldorff M (1997) A Spatial-Scan Statistic. *Communications in Statistics: Theory and Methods,* 26:1481–1496

Lazoff DM, Sherman AT (1994) An Exact Formula for the Expected Wire Length for Two Randomly Chosen Terminals. *Technical Report TR CS-94-08*, Computer Science Department, University of Maryland, Baltimore County (July 20, 1994). 14 pages. http://citeseer.ist.psu.edu/david94exact.html

Lee SL, Chun SJ, Kim DH, Lee HJ, Chung CW (2000) Similarity Search for Multidimensional Data Sequences. In: *16th International Conference on Data Engineering*; Feb 28 – Mar 03, San Diego, USA: IEEE: 2000

Mark D, Egenhofer MJ, Bian L, Hornsby KE, Rogerson PA, Vena J (2000) Spatiotemporal GIS Analysis for Environmental Health: Solutions Using Geospatial Lifelines. In: Flahaut A, Toubiana L, Valleron AJ (eds) *Geography and Medicine: GEOMED '99,* Elsevier, Paris. pp 65–78

Mark DM, Egenhofer MJ (1998) Geospatial Lifelines. In: Guenther O, Sellis T, Theodoulidis B (eds) *Integrating Spatial and Temporal Database, Dagstuhl Seminar Report No. 228*

MathSoft (2000) *S-PLUS 2000 Guide to Statistics, Volume 1*, Data Analysis Products Division, MathSoft, Seattle, WA

Miller HJ (1991) Modeling Accessibility Using Space-Time Prism Concepts within a GIS. *International Journal of Geographical Information Systems*, 5(3):287–301

Miller HJ (in press) A measurement theory for time geography. *Geographical Analysis*, forthcoming

Okabe A, Miller HJ (1996) Exact computational methods for calculating distances between objects in a cartographic database. *Cartography and Geographic Information Systems*, 23:180–195

Park S, Kim S, Chu WW (2000) Segment Based Approach for Subsequence Searches in Sequence Databases. In: *16th ACM Symposium on Applied Computing,* Las Vegas, Nevada, USA. pp 248–252

Parkes D, Thrift N (1980) *Times, Spaces, and Places: A Chronogeographic Perspective*. John Wiley and Sons

Plane DA, Rogerson PA (1994) *The Geographical Analysis of Population Geography: With Applications to Planning and Business,* John Wiley and Sons, Inc

Rogerson P, Han D (2002) The effects of migration on the detection of disease clusters. *Social Science and Medicine*, 55:1817–1828

Rogerson PA, Weng RH, Lin G (1993) The Spatial Separation of Parents and their Adult Children. *The Annals of the Association of American Geographers,* 83(4):656–671

Veltkamp RC, Hagedoorn M (2000) Shape Similarity Measures, Properties, and Constructions. In: *Advances in Visual Information Systems, Proceedings of the 4th International Conference, VISUAL 2000; Lyon,* Springer, France. pp 467–476

Vlachos M, Hadjieleftheriou H, Gunopulos D, Keogh E (2003) Indexing MultiDimensional Time Series with Support for Multiple Distance Measures. In: *9th ACM SIGKDD,* ACM Press, New York, Washington, DC, USA 2003. pp 216–225

Vlachos M, Kollios G, Gunopulos D (2002b) Discovering Similar Multidimensional Trajectories. In: *18th International Conference on Data Engineering (ICDE),* San Jose, California, USA. pp 673–684

Vlachos M, Gunopulos D, Kollios G (2002a) Robust Similarity Measures for Mobile Object Trajectories. In: *5th International Workshop on Mobility in Databases and Distributed Systems (MDDS),* Aix-en-Provence, France. pp 721–726

Yanagisawa Y, Akahani J, Satoh T (2003) Shape-Based Similarity Query for Trajectory of Mobile Objects. In: Chen M-S, Chysanthis PK, Sloman M, Zaslavsky A (eds) *Mobile Data Management,* Springer, Berlin. pp 63–77

Yang Z (2001) *Modeling and Reasoning with Geospatial Lifelines in Geographic Information Systems*. Unpublished PhD Dissertation, Department of Geography, University at Buffalo, January 2001