

Spatio-Temporal GIS Analysis for Environmental Health

David M. Mark, Max J. Egenhofer, Ling Bian, Peter Rogerson, John Vena

1. Specific Aims

Medical researchers need improved tools and analysis methods for examining health-related information in spatial and temporal context. Geographic space is the arena within which all social and environmental processes occur. It follows that location is a fundamental dimension of the phenomena examined by environmental and social scientists. Geographic information thus is a key component of any information system designed to support social or environmental analysis or decision-making, especially for the environmental health sciences. The utility of geographic information systems (GIS) for medical research has long been recognized (Mayer 1983), and use of GIS is now part of standard procedure in epidemiology and related areas (Stallones *et al.* 1992; Briggs and Elliott 1995; Clarke *et al.* 1996; Croner *et al.* 1996). However, for many health conditions, application of GIS has been hampered by the poor ability of commercial GISs to handle multitemporal geographic information or movement. This shortcoming severely impedes the utility for GIS to assist in understanding health problems with long latency periods, such as many forms of cancer, since with mobile populations, the location of the patient at the time of diagnosis or mortality may have little relation to the location of exposure to toxic substances or other environmental risks.

The proposed research will enhance our understanding of geographic approaches to the study of the environmental health sciences by developing new tools for the analysis of geographically and temporally referenced medical information, and new methods for reasoning about environmental exposures and their consequences over space and through time. The methods also will be applicable to hazardous exposures of shorter time periods with more immediate impacts.

This research project focuses on the extraction of health-related information from *geospatial lifelines*, which capture individuals' locations in geographic space at regular or irregular temporal intervals. The objectives of our work are to develop and test the theory of geospatial lifelines in the environmental health sciences by:

- developing methods to trace locations of individual people (patients, cases, or controls) back through time, to discover spatial clusters in the past or to determine past environmental exposures,
- designing, prototyping, and assess computational models that can deal with large sets of geospatial lifelines and environmental information, and
- examining the ethical and legal implications of recording individuals' geospatial lifelines in databases, and establishing procedures for appropriate restrictions on data analysis and dissemination.

Geospatial lifeline data consist of series of discrete space-time samples over the domain of continuous movements, describing an individual's location in geographic space at regular or irregular temporal intervals. In the future, space-time pairs may be measured (e.g., with GPS receivers and digital clocks) or they may be estimated and recorded manually. Geospatial lifeline data may be recorded at different resolutions, but in environmental health applications, we are mainly concerned with data over days to entire lifetimes, with a resolution of hours to years.

2. Background and Significance

2.1. Relevant Prior Research on Sampling and Observing in Space and Time

Berry (1964) presented an organizing framework called "The Geographical Matrix." The two-dimensional version of this is a table, with columns representing *places* and rows representing *characteristics*. Berry added time as a third dimension, consisting of time periods or cross-sectional slices. Sinton (1978) later related this to geographic information systems (GIS) by proposing a scheme that identified six basic kinds of spatio-temporal data generalization. Sinton proposed that, for all spatio-temporal data, one of the three factors of *location*, *time*, and *theme* is fixed, a second is controlled (varied systematically), and the third is observed. For example, the U. S. Census fixes time (the census day), systematically varies space (census tracts and blocks), and takes as its theme the characteristics of the residents of each given area. However, Sinton did not discuss physical objects moving through geographic space. Langran (1992, p. 12) reiterated Sinton's model and provided additional examples for "moving objects," but she described them as fixed attribute (object identity), controlled location, and measured time. This approach would select one object to be tracked, observe one or more locations, and record the time at which the object occupied each observed location—as in the case of time-stamp clocks in offices to record people's attendance. Most geospatial lifeline data, however, fix identity, control time (by establishing recording intervals for a data logging device, an observer, or a diary), and observe locations. Other data fix location (say, a health care facility), vary theme (identities of individual patients), and record or observe the time of use. Since classic data under this scheme have the third (measured) dimension as a single-valued function of the first two, the ordering of these dimensions for data recording or analysis will be an important constraint on methods for analysis or database retrieval.

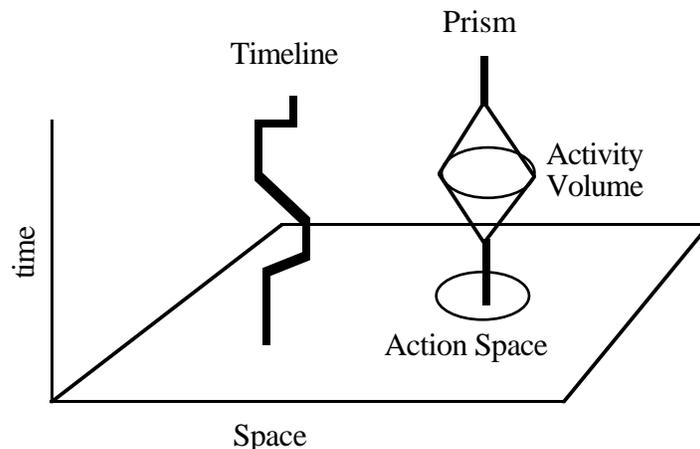


Figure 1: Forer's "space-time aquarium" and basic space-time geometric objects.

Forer (1998) described geometric primitives for what he called a time-space aquarium (Figure 1). One type of object is a *prism* that represents a region of space and that can be reached given a starting or ending point in space-time and a maximum velocity. The other type of geometric object is more relevant to our approach: a continuous line tracing variations in position with time. Such a *timeline* was the core concept of Hägerstrand's (1967; 1970) Lund school of time geography in Sweden; however, the term *timeline* may not be ideal, since timeline is commonly used in an aspatial context to refer to a plot of attributes or events as a function of time. Odland (1998) employs lifeline concepts of a particular type to study demographic processes. His main type of data, a "locational history," is defined as "a complete accounting of an individual's places of residence over some period of time, including the dates of relocation events and the identities of origins and destinations" (Odland 1998, p. 239). At the scale of typical durations of residence,

moves in this type of data are normally essentially instantaneous, taking place in a few days. Thus the geometry of the timelines appears to be vertical lines with horizontal jumps to new locations. We use the term "geospatial lifeline" to refer to the paths underlying these representations.

2.2. Spatio-Temporal Clusters: The Problem of the Interaction of Long Latency Periods and Human Mobility

Often it is important to determine whether observations of some phenomenon are clustered in space or time or both. When trying to determine the causes of some outbreak or chronic pattern of ill health, analysts frequently plot the distributions of cases on maps. This method has been used at least since Dr. John Snow's now famous map of cholera deaths in London, England, which helped identify a particular public water pump as the source of the epidemic (Snow 1936). For infectious diseases with short incubation periods, analysis of the spatial distribution alone may be sufficient; however, there are problems with such methods in the study of environmentally-induced diseases with long latency periods, such as many forms of cancer, since the people could have moved several times since their exposure to environmental hazards, thus breaking up clusters and obscuring patterns. Antó and Sunyer (1992) provide a good description of the problem:

"Although some problems are frequently seen as geographical clusters, whereas others are more frequently identified as temporal clusters, this choice may depend upon the investigator's wishes and upon the model in mind. Rothman (1990) has pointed out that the relative magnitude of space-time clustering is influenced by the length of the induction period and, in case of transmissible diseases, by the pathogenicity of the causative agent. A long induction period allows cases to appear scattered both in time and space, thus attenuating the clustering pattern. By contrast, if the induction period is short, space-time clustering will be easily identified." (Antó and Sunyer 1992, p. 338).

Methods to be developed in this study, based on reasoning about geospatial lifelines of specific cases, will reduce or eliminate this problem of cluster dispersion. If researchers have the data and information manipulation tools, this will allow them to roll cases back to places of residence or travel in the past when they might have been more markedly clustered and to subject such past "snapshots" to two-dimensional cluster detection procedures. Another approach would be to identify clusters directly in three dimensional space-time.

2.3. Significance

The central idea of this proposal is that people's movements through geographic space are a critical factor in exposures to environmental health hazards. Computational models that can account for the fact that people's locations in geographic space are dynamic rather than static will greatly enhance the power and potential of data analysis and reasoning methods for examining environmental exposures or discovering past clusters of currently-ill patients. Individuals navigate through space, they stay at locations where they meet other individuals and they perform regularly reoccurring tasks that involve variable or fixed locations in geographic space. These movements often expose people to environmental factors that can cause health problems at latency periods ranging for seconds to decades. For example, establishing whether a particular U.S. soldier was exposed to hazardous chemicals during Operation Desert Storm requires both a record of the space-time behavior of the soldier, and also the space-time distributions of environmental risk. If the latter are not known, space-time places of high risk might be inferred by comparing the space-time behaviors of soldiers showing symptoms of ill health, with the behaviors of a control group of soldiers not showing symptoms. The data analysis and reasoning methods that will be developed in this study will facilitate such analyses.

3. Research Design and Methods

3.1. The Research Issues

NIH Solicitation PA-95-032 stated that studies to be supported by the program may include, but are not limited to, a set of nine research goals that define the scope of the solicitation (NIH 1995). In this project, we will contribute to the following six of those nine goals, which we number here for subsequent reference:

- Goal 1: Generating hypotheses in environmental epidemiologic studies of spatial and temporal relationships between environmentally-induced diseases and exposures.
- Goal 2: Mapping and/or other visualization techniques for assessing exposure data and disease incidence or mortality data for hypothesis-generating and for purposes of conveying public health information.
- Goal 3: Use of GIS approaches to identify study populations with potential exposure to environmental hazards.
- Goal 4: Surveillance of disease outcomes in populations with exposure to environmental pollution.
- Goal 5: Statistical approaches to assess the validity and significance of apparent time between measured or recorded exposure data and incident cases of cancer and other chronic diseases compared with controls.
- Goal 6: Development of statistical methods to account for uncertainty due to potential confounding factors: e.g., measurement error, repeated measures, and missing data in applications of GIS.

This proposal addresses the above issues through a set of cross-cutting research questions in geographic information science, discussed in the following sections.

3.1.1. Patterns at Different Temporal and Spatial Scales

Although there are many exceptions, a typical human being sleeps in the same place (home) almost every night and makes one particular place home for several years in a row. A typical adult has a job outside the home and spends 35-40 hours per week at some particular work place. Over the course of the week, the person might spend 60% of his or her time at home, 25% at the work place, and the other 15% either at other places or moving between places. All of these sites, including places along a commuting path, provide possibilities for exposure to environmental health risks. The pattern of movement among two regular sites and a variable set of other places repeats through most of the year, but people in developed countries typically have two to four weeks of vacation each year when they do not go to work, and most people go away from home for some or all of that time. At a longer time scale, they may from time to time change their place of residence, or their place of work, or separately or simultaneously. At a shorter time scale, they move about on shorter spatial scales within their work place and in their home.

- Our project will investigate the scales of various behaviors and examine the effects of different patterns and densities of sampling on exposures to health risks, movement after such exposure, and researchers' abilities to make inferences about linkages between environmental exposures and health-related effects

3.1.2. Resolution of Measurements

Human movement has different characteristic spatial and temporal scales, depending on the purpose, nature, and mode of the movement. Data sampled at the wrong scale or interval may completely miss, or drastically distort, the effects of particular processes. For example, since most people move their home location only every few years, a record of home location or regular

sleeping place would typically appear as long intervals of constant location (hundreds or thousands of days), interspersed with jumps that happen on a single day or over periods of just a few days. A one-day sampling interval would capture this scale of spatio-temporal variation and would in fact be highly redundant.

- Our project will establish appropriate spatial and temporal sampling strategies for detecting various processes of response to environmental health risks and will examine the effects of poorly-designed patterns and densities of sampling on our abilities to make inferences about various scales, kinds, and patterns of exposure.

3.1.3. Integration of Multi-Resolution Spatio-Temporal Data

The spatial components of geospatial lifelines for patients or members of control groups will often be recorded as coordinates in two or three dimensions. The analysis of these coordinate values will be significantly more interesting if they can be linked back to a meaningful expression of the places and geographical locations at which the subjects were located, and especially to environmental information. Such *geocoding* requires matching of a subject's path with a georeferenced data set. Geocoded lifelines enable the generation of answers to such queries as, "Where were these cancer patients exactly 15 years ago?" or "Who went through the area of the chemical spill within 6 hours of the spill?" Initial advances in the development of gazetteers for digital geolibraries (Smith 1996) will need to be extended since the geocoding of lifelines must deal simultaneously with various levels of spatial detail. The use of geocoded lifeline information occurs not only in queries, but also in the generation of meaningful answers. A geocoded answer, however, depends upon the scale of the geospatial lifeline activities. For example, the query "Given a geospatial lifeline of an individual patient (e.g., represented by a sequence of time stamped GPS positions), what places did the individual visit?" requires an adjustment of spatial scale as the temporal range of the query domain varies. Over a 1-day interval, the answer may be, "The individual went to the factory, came back home around lunch time, went back to the plant in the early afternoon, and went shopping at the mall before returning home in the evening." Over a six-month period, one may expect something like "The individual stayed most of the time in Orono, with occasional trips to the Maine coast, and a 5-day trip to the greater Boston area." Over a 15-year period, a reasonable answer might be, "The individual lived in London, Ontario, for the first 3 years and then moved to Buffalo, New York."

- Our project will generate new methods for integrating and transforming quantitative and qualitative spatio-temporal descriptions of movements.

3.1.4. Implications of Spatio-Temporal Sampling Protocols

The majority of methods developed for the study of human migration assume that data are available on movements over a fixed time period. For example, most migration data sources are generated via questions such as, "Where did you live last year?" or "Where did you live five years ago?" There are interesting questions regarding the comparability of data based on different time periods, since return and repeat migration are increasingly likely to mask moves made over longer and longer time periods (Rogerson 1990). More fundamental changes need to be made in the analysis of data based upon population registers that monitor the residential locations of individuals. Ledent (1980) has suggested how multistate analyses of population systems can be carried out when register information is available. Another motivation is discussion in the US Bureau of the Census to replace the Census long form with a continuous measurement scheme (known as the American Community Survey) (National Research Council 1994). This survey would begin around 2003; several hundred thousand people would be interviewed each month and a sort of rolling average would be used for small areas analysis. Such rolling averages or sums for small area analysis of flow data would require the development of new methods of analysis. Return and repeat migration would have to be considered and the tradeoffs between temporal and spatial aggregation, versus the desire for information on small-scale, space-time trends, would have to be assessed.

Although many methods have been developed for the detection of spatial, temporal, or space-time clusters, almost all of these assume data that either represent point locations, or locations within

designated areal units. With data on the geospatial lifelines of individuals, the challenge lies in modifying existing methods for cluster detection to accommodate the more detailed data. This would entail focusing upon the appropriate definitions of the "at-risk" populations used in the denominators of rates. Such revised methods will be applicable, for example, to the topic of cluster detection in epidemiology, where there is interest in detecting space-time clusters of high incidence of some health problem.

- Our research under this project will help us to understand implications of particular survey and sampling protocols for human migration studies and for other similar data.

3.2. Research Plan

Our research plan addresses research issues in several logically interrelated components. It starts with the development of a plausible and effective data model for geospatial lifelines, and continues with investigations of methods for querying and presenting lifeline information and for improved analysis. It closes with a description of our plan for evaluation and testing of the geospatial lifeline methods, and for examining implications for privacy and other legal and ethical considerations.

3.2.1. Developing a Data Model for Geospatial Lifelines

Hägerstrand's concept of time geography provides a framework for modeling geospatial lifelines, which this project will extend and evaluate. The basic element of lifeline data is a space-time observation consisting of a triple $\langle \text{ID}, \text{location}, \text{time} \rangle$, where ID is a unique identifier of the individual used throughout all recordings of that individual's movements, location is a spatial descriptor (such as a coordinate pair, a polygon, a street address, a zip code, or some other locative expression), and time is the time stamp when the individual was at that particular location (such as a clock time in minutes or event time in years). Identities of entities may disappear and later reappear (Hornsby and Egenhofer 1997)—for example, a toxic waste site may for a time be considered to be totally remediated and later be found to still be polluted. Identities may also disappear when objects are aggregated (Hornsby and Egenhofer 1998). In most cases, the time stamp will refer to when an event occurred in real time, but not when it was stored in a database (Snodgrass 1992); for some applications, however, such as in assessing liability or fault, it may be important to preserve both sorts of time within the database (Worboys 1998), that is, both when the site was polluted and when the pollution became known to company or government officials.

The data model will be influenced by other factors. For instance, recordings of geospatial lifelines are discrete, while the phenomena they describe are typically continuous. For this reason, different interpolation methods will be needed depending on the ontological characteristics of the movements. Tracking commuters as they pass through polluted areas might require data recorded at 5-minute or even 1-minute intervals, whereas the recording of people's workplaces at 6-month intervals would require a different interpolation method.

- From a study of several types of health-related spatio-temporal information, we will derive a data model for geospatial lifelines that will enable systematic spatio-temporal querying and spatio-temporal analyses. This element of the research plan provides core concepts that are essential to all other aspects of the project.

3.2.2. Exploring Methods for Visualizing Geospatial Lifelines

The visualization of geospatial lifelines for disease cases or other individuals is a very effective means of displaying space-time coincidence and proximity. Consider the hypothetical geospatial lifelines of three people (John, Fred, and Mary) who were found to have developed a rare form of cancer often attributed to exposure to agricultural chemicals, despite the fact that none of them worked in the agricultural or chemical industries. Suppose further that we collect the following information about when and where they lived:

- John was born in 1947 in town "A." In 1978, at age 31, he moved to town "B," where he lived for just 3 years. Then he moved to "D" where he still lives today.

- Fred was born in town "B" in 1921, and at age 22, he moved to town "C" for 5 years, before returning to town "B," where he lived 45 years until his death in 1993.
- Mary was born in town "D," and lived there for 33 years. Then, she lived for 22 years in town "B," before living the last 19 years in town "C."

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—allows one to see that all three people lived in Town "B" for two years, from 1979 until 1981. If their cancer was triggered by a common environmental exposure at place of residence, researchers must examine events in Town "B" in the period 1979-1981.

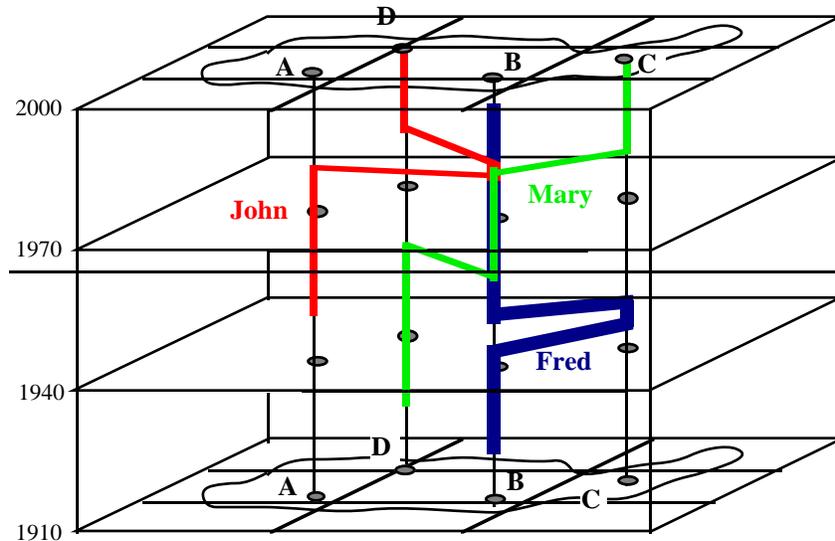


Figure 2: Geospatial lifelines of three individuals over nine decades.

In a monochrome display of lifelines in space-time, relationships among as few as three people's lifelines can be difficult to comprehend. With colors and textures, perhaps a dozen such lines could be drawn and still distinguished, but we are contemplating data sets consisting of hundreds or thousands of geospatial lifelines. The database search procedures to be developed in this project are not directly influenced by this visual clutter, but visualization based on vector plots of geospatial lifelines simply will not be effective or usable. Initially in this project, we will develop and use a raster model of space time (Forer 1998), in which 3-dimensional cells represent aggregates of specific space-time instances with a spatial extent and duration determined by the researcher. For a large collection of geospatial lifelines or of space-time events, a computational model could visit each cell and count the number of geospatial lifelines that passed through it, or count the number of events that occurred in it. Cells with high frequency indicate places in which a large number of subjects stayed or passed through during some short period of time. If the subjects were people diagnosed as having cancer, the highest frequency cells may be the most likely places and times for environmental exposures to carcinogens. Of course, these concentrations could reflect space-time density of the population in general, and the rates would need to be adjusted to account for the non-uniform distributions of human population density in space and over time. Three-dimensional interpolation procedures will be needed for establishing such base populations from census data or other sources, against which lifeline or event densities can be evaluated for anomalies. Spatial statistical methods for detecting spatial clusters can also be extended to search for clusters in space-time. Such approaches will be developed and tested in this project.

- We will study methods for aggregations of geospatial lifelines so that query results can be visualized appropriately. We will use Intergraph's Voxel Analyst software to display space-time densities by density slicing in 3-D, thresholding, and other graphical

techniques. This element of the research plan supports NIH Goals 1 and 2 (see section 3.1 above)

3.2.3. Formalizing a Query Language for Geospatial Lifelines

The concepts of time geography also provide a framework for queries about geospatial lifelines. Possible queries about individual lifelines can be derived from the geospatial lifeline triple <ID, location, time>, distinguishing whether any combination of the three arguments is known or unknown. Examples include "What individuals were at location S at time T?", "Did X stay at location S at time T?", and "Which individuals have ever visited location S?" Through the use of aggregate operators as functions, additional information can be derived about durations of immobility or of trips. Such aggregate operations can then be used in combination with queries across several individuals, as in the case of, "What locations have been visited by more than n members of a given population within any single year?" Queries about geospatial lifelines involve more complex sets of constraints if the different travel speeds of individuals are considered as well. Further examples relevant to environmental risk exposures include "How often did X come to location S?" and "What is the longest time X stayed at location S?"

Some of these queries over space-time prisms resemble operations on OLAP ("On-Line Analytical Processing") data cubes (Gray *et al.* 1996), such as slice, dice, roll-up, and drill-down; however, the semantics of the data cube operations are not (yet) well defined and lack the intersection operation, which is meaningful and important in the analysis of geospatial lifelines. Similar to the development of visual query languages for cubes (Frank 1992a; Richards and Egenhofer 1995), the operations upon geospatial lifeline prisms lend themselves to the design of direct-manipulation user interfaces.

A formal language will provide an organizational framework for queries about lifelines and queries that relate lifelines to other objects. For certain application domains a positive answer to some such queries would indicate a consistency violation (e.g., a physical object cannot be located simultaneously at two different locations).

- We will develop an algebra over geospatial lifelines based on the concept of lifeline prisms. This formalism will lead to a spatio-temporal query language, enabling testing of space-time hypotheses. This element of the research plan also provides core concepts that are essential to all other aspects of the project.

3.2.4. Matching Geospatial Lifelines and Reasoning About Relations Between Lifelines

If large volumes of geospatial lifeline data become available for both cancer patients and for samples of the general population, methods for determining the similarity of two geospatial lifelines could be used to match each cancer patient with someone with a similar history of residential locations. This would allow geographic location to be controlled in a case-control methodology. Exact matches of geospatial lifelines are easy to define, although efficient search for such matches in very large databases will require basic research. More challenging are cases of partial match, since here degrees of match and mismatch would have to be weighted to come up with an over-all index of similarity. The development of similarity measures for geospatial lifelines will require examination of concerns of the end users of information and the reasons that matches are being sought. For example, for cancer case-control sampling, if the latency of the particular form of cancer is known to be around 10 years, then similarity of geospatial lifelines around 10 years ago could be given maximum weight, while similarity the last 3-5 years might be ignored altogether.

An alternative method is the comparison of qualitative descriptions of geospatial lifelines, particularly their qualitative spatio-temporal relations. Qualitative spatial relations and qualitative spatial reasoning methods have shown significant results in 2-dimensional geographic space (Frank 1992b; Hernández 1994; Sharma *et al.* 1994; Hornsby and Egenhofer 1997), with a focus on 2-dimensional, areal objects embedded in a 2-dimensional plane. For example, models for

topological relations between spatial regions (Egenhofer and Franzosa 1991; Randell *et al.* 1992) have been highly successful in GISs as a basis for querying and spatial reasoning. The data model for geospatial lifelines, however, suggests the embedding of a 1-dimensional object (the lifeline) in a multi-dimensional space (location + time); therefore, extensions of the existing models will be needed. For example, the set of topological relations between two lines embedded in 2- or 3-dimensional space comprises 33 relations (Egenhofer *et al.* 1993). When applied to geospatial lifelines, the set of realizable relations is smaller, because geospatial lifelines are monotonic (they do not swing back in time). The significance of the ordering of the time axis, however, requires to consider the direction as an integral part of lifeline relations so that "before" and "after" can be distinguished (Allen 1983).

- We will develop procedures for matching geospatial lifelines and for finding matching cases efficiently in order to identify control populations with similar geospatial lifelines to members of some set of cases. This element of the research plan supports NIH Goal 1.

3.2.5. Processing Incomplete Geospatial Lifeline Data

Often data for a geospatial lifeline are recorded as sets of discrete space-time measurements. In theory, the temporal resolution of the recorded geospatial lifelines could be very dense, such that every time point queried would be available (Tansel *et al.* 1993). There are, however, numerous reasons that speak against such an approach. From a system perspective, we would obtain extremely large data sets for even short timelines. On the other hand, even if the samples are very dense, there may be unforeseen queries that would require a higher resolution. A third argument against the assumption of the availability of dense time recordings is the expectation that occasionally the device for capturing space or time may be unavailable or out of order. Geospatial lifelines relevant to environmental health studies usually will be a series of home addresses or work places, constructed from memory by patients or their relatives, perhaps containing errors or gaps. (For example, relatives may have little or no idea of where loved ones traveled while serving overseas in the military during times of conflict.)

In order to fill such gaps, interpolation methods may be useful to infer locations occupied, answering queries such as: "Given that we have no knowledge of the patient's residence or residences between two points in time, where might the person have been in between?" The larger the time intervals and the faster the movement, the more imprecise we expect such interpolations to be. There are, however, a number of ways to add specificity into the interpolation methods if one has, for instance, some knowledge about the space, but not about the time; some knowledge about the time, but not about the locations; or some information about a possible travel mode. There is a close connection between such differences to the geocoding of geospatial lifelines, since topography and infrastructure often constrain movements. For instance, if a car travels through a city and its locations are tracked every minute, then the interpolation needs to take account of the fact that the car is probably moving along roads in a road network, or through parking places and garages. Such interpolations are ambiguous if there are multiple possible paths between two locations; however, gaps may be filled by analyzing similar paths taken by the same individual at different times, from which a more likely intermediate location can be determined.

Sometimes, it may be appropriate to take a more conservative stance by treating gaps in the record simply as missing data. In order to support such a choice by a researcher, all of the query and analysis methods developed in this project will have to deal properly with geospatial lifelines with gaps, despite the fact that logically the person must have been somewhere at every moment during the gap in the record.

- We will explore interpolation methods based on two particular properties of a geospatial lifeline: (1) the individual's immediate past, as described by the most recent part of the geospatial lifeline, and (2) the individual's typical behavior in the past in similar situations (e.g., whether the individual followed the same route in the past and, if so, whether it consistently led to the same destination). We also will ensure that all methods

developed in this project can deal with gaps in records of geospatial lifelines as missing data when researchers deem that to be a more appropriate response to such gaps. This element of the research plan supports NIH Goals 2 and 6.

3.2.6. Intersecting Geospatial Lifelines with Environmental Data

For health problems induced by short-term exposures to environmental toxins, researchers might need to relate individual geospatial lifelines of people in a population potentially at risk with the spatio-temporal distribution of the hazard. This might even be done in nearly real time in order to warn people about the possible consequences of their exposures. For example, many delivery trucks and emergency vehicles as well as some luxury cars currently are equipped with GPS receivers that continuously monitor their positions. After an airborne toxic release near a highway, methods to be developed in this proposal to identify all trucks that passed through high concentrations of the toxin, allowing drivers to be contacted and advise to undergo check-ups. Research on this topic will have two phases. First, we will implement methods to represent diffusion, movement, and variable concentration of environmental toxins as three-dimensional fields in space time. Basically, a three-dimensional field is a single-valued function of position in a three-dimensional space. (See Scott 1997 for a discussion of geographic analysis in three-dimensional models.) Here, one of the three dimensions actually is time, and the field provides the concentration value of the toxin at each point in space-time. The second aspect of this part of our research plan will involve algorithms for efficient comparison of individual geospatial lifelines, to quickly determine the maximum or cumulative field values that the particular lifeline encountered. Such spatio-temporal intersection procedures may need to be modified if they are to deal efficiently with comparisons of many geospatial lifelines to one spatio-temporal field.

- We will implement methods for interpolating or simulating spatio-temporal fields representing environmental hazards to produce three-dimensional fields of concentration and develop procedures for intersecting large numbers of detailed geospatial lifelines with such fields to estimate maximum and total exposures for individuals who moved through the hazardous area. This element of the research plan supports NIH Goals 1, 2, and 4.

3.2.7. Analyzing Effects of Migration on Environmental Exposure

In demography and the social sciences, migration normally refers to long-term changes in residential location. Human migration is an important process in its own right, but is especially important in its relation to human health. In-migration and out-migration often have greater influence on local population counts than do birth rates and death rates. In the area of health, people may become ill or die far from the location at which they were infected or exposed to environmental health hazards. For chronic diseases or conditions with long latency period, migration is the time scale of human mobility most relevant to the time scale of the health-related processes. For example, a spatial cluster of deaths from a certain form of cancer among retired people in Florida may not be correlated with environmental conditions in that state, but may actually be due to an exposure to carcinogens at work or in the home in the State of New York. In other cases, migration may disperse clusters completely, as might be the case with migrant or casual farm workers, exposed to agricultural chemicals, might settle into sedentary jobs at a wide scatter of places. In such cases, hot spots could only be detected if past locations of the cases can be determined. Also, migration will be the domain within which we will test raster representations of event density for large collections of geospatial lifelines in space-time, described above. Density values will be visualized for exploratory spatio-temporal data analysis purposes and will be used in statistical procedures to detect spatio-temporal clusters.

- We will develop methods for simulating human migration, to provide appropriate background demographic data for use in cases where individual geospatial lifelines of patients are not available. This element of the research plan supports NIH Goals 1, 3, 4, and 5.

3.2.8. Constructing Multiregional Life Tables and Alternative Forms of Migration Data

Life tables summarize the mortality experience of populations by converting age-sex specific death rates into life expectancies and other summary measures of mortality. Andrei Rogers (1973, 1975) and others have developed multiregional generalizations of life table methodology to yield, for example, the expected number of years that individuals of a given age can expect to reside in different regions. These multiregional life tables are based upon both mortality and migration data. Migration data come in many forms, and the form has a significant influence on the subsequent analysis. Common migration questions are "Where did you live one year ago?" and "Where did you live five years ago?". The multiregional life tables that result from these two alternative questions are very different from one another (Kitsul and Philipov 1981; Rogerson 1990). Because life tables are based upon a Markovian model assumption that origin-destination migration probabilities remain constant over time, the five-year data give a better result than the one-year data. This is because it is unrealistic to assume that individuals will continue to move in the future according to the one-year table of origin-destination probabilities, since we know that return migration to a former region of residence is much more likely than the one-year table indicates. The five-year data on migration flows include more return migrants than do one-year data, and are hence "better" for constructing multiregional life tables that yield region-specific "at-risk" population information.

- We will determine whether there is an optimal period length for migration tables, develop methods that would take information from a "rolling census" that surveys a large number of people every month and use it to develop multiregional life tables, and generate methods that would allow one to adjust one-year data for estimated return migration, with the objective of developing better multiregional life table estimates.

3.2.9. Protecting Privacy

Privacy has always been a major concern of the medical profession (Harris-Equifax 1993). Medical researchers are often hampered by a lack of access to detailed information about cases, when access to such data is restricted in order to protect the privacy of the individuals involved. Potential for surveillance of locations and activities of people in space and time will explode over the next decade, as more and more people will carry devices combining the functionality of pagers, cellular telephones, and GPS receivers. If such data are systematically archived, they may be very valuable to the individuals themselves if they become seriously ill at a later time and the geographical source of the causes of the illness is sought. But such massive surveillance of individuals would likely be resisted by people concerned about personal privacy, at least as long as their health were not compromised. How should individual privacies be protected when tracking data on people is made available to researchers in environmental health or other areas? Designers of technologies bear at least some of the social responsibility for the new systems they design. In this case, if geospatial lifelines could be developed with capabilities or design characteristics such that certain intrusive applications could be avoided altogether, we should so design them. One part of our evaluation of methods for analyzing and using geospatial lifeline data in medical research and in general will focus on the ethical issues raised by the ability to create large, geocoded computer-based databases. This study will include examples found in the popular press, government reports, and scholarly publications, and will focus on the way in which standards for data privacy in the medical domain have consequences in the spatio-temporal domain. We also will analyze the ways in which medical and related databases are commonly regulated, both in the United States and elsewhere.

- We will examine the particular issue of privacy concerns about locations or addresses of patients in the context of general issues of geographic information privacy and health information privacy, and recommend guidelines that maximize the quantity and precision of geographically-referenced health information available to bona fide researchers while protecting individual privacy.

3.2.10. Evaluating and Testing

The results of the tasks described in this section will lead to various software prototypes, mostly written in C++, although some will be based on and integrated with existing commercial GIS software. We will integrate these prototypes and develop a comprehensive testbed for the analysis of geospatial lifeline data and their relation to environmental data. Functionality and performance of the prototypes and testbed will be tested with real geospatial lifeline data if such data can be obtained in sufficiently large quantities. Otherwise, test data will be simulated based on known model of human mobility plus aggregate census data. Analytical results, graphic output, and other aspects of the performance of the prototype will be evaluated by the research team member from Social and Preventive Medicine (Vena) and his colleagues, to make sure that outputs of the models will meet the needs of medical researchers.

Data from a recent western New York study of postmenopausal breast cancer will form an initial case study, and other similar and larger data sets will be sought for further model testing. Recent estimates indicate that only about 40 percent of breast cancer cases can be explained by accepted and suspected risk factors. Exposures to chemical and physical agents in the workplace of women have not been studied adequately. Thus, the object of the present protocol is to determine whether occupational exposure to selected chemical and physical agents is associated with increased risks of pre- and postmenopausal breast cancer. A population-based, case-control study of pre- and postmenopausal breast cancer incidence among women aged 40 to 79 years was conducted in western New York State between 1986 and 1991. Only new, histologically-confirmed primary cases of malignant breast cancer from all major hospitals in the region were enrolled (301 premenopausal and 439 postmenopausal). A population-based control group was recruited and was frequency-matched to case subjects on age and county of residence. Face-to-face interviews by trained nurse-interviewers were carried out and information on a wide array of accepted and potential risk factors was obtained. Standardized lifetime occupational histories were also obtained and these included details regarding place of employment, duties and activities associated with each job.

For all 1,550 subjects, the occupational and residence histories will be sent to a team of NCGIA researchers who will then attribute exposure. Statistical analyses, including logistic regression and generalized additive modeling techniques, will be carried out to assess the role of occupational exposures with a focus on polycyclic aromatic hydrocarbons, specific organic solvents and electromagnetic fields. Analyses will be conducted separately for the pre- and postmenopausal groups and will be combined where appropriate. Identification of agents whose exposures are amenable to modification may allow for new public health initiatives to reduce risk.

3.3. Multidisciplinary Research Team

Our investigations will be conducted by a multidisciplinary team of researchers from geography, social and preventive medicine, and spatial information science and engineering. The geographers include two GIS experts, a demographer, and a specialist in privacy and geographic information. The team will collaborate to identify the types of data, sampling methods, spatio-temporal inferences, and analyses that environmental health scientists need to apply to geospatial lifeline information. We will formalize these methods, develop software prototypes, and test these computational tools with domain applications.

These investigations will not be conducted in a vacuum, but will take full account of issues pertinent to privacy laws and other ethical issues related to the potential availability and applicability of geospatial analysis and reasoning methods in environmental health. During the project, the team will go through a full cycle, starting with ontological studies of how well the concept of geospatial lifelines reflects the tracking of individuals' movements through geographic space and ending with the study of policy issues. Constant involvement of researchers from Social and Preventive medicine at Buffalo will keep the project well grounded in relevance to medical applications. We are already aware of the different spatial and temporal scales at which geospatial lifeline data are

collected under current practice, and we know of application-dependent differences in recording protocols. Thus we have time stamped locations in some technologies and application domains, but records of durations that an individual stayed at particular locations in other databases. Based on time geography and constraint databases, the spatial information scientists and GIS experts on the project will define a geospatial query language for the identified ontologies. Issues of ethics and privacy regarding such databases will be a major focus in the last year of the project, and will be used to generate specifications for constraints on database use and access policies.

4. References

- J. Allen (1983) Maintaining Knowledge about Temporal Intervals. *Communications of the ACM* 26(11): 832-843.
- J. Antó and J. Sunyer (1992) Soya Bean as a Risk for Epidemic Asthma. in: P. Elliott, J. Cuzik, D. English, and R. Stern (Eds.), *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*. pp. 323-341, Oxford University Press, Oxford.
- B. Berry (1964) Approaches to Regional Analysis: A Synthesis. *Association of American Geographers* 54: 2-11.
- D. J. Briggs and P. Elliott (1995) The Use of Geographical Information Systems in Studies on Environment and Health. *World Health Statistics Q* 48: 85-94.
- K. C. Clarke, S. L. McLafferty, and B. J., Tempalski (1996) On Epidemiology and Geographic Information Systems: A Review and Discussion of Future Directions. *Emerging Infectious Diseases* 2: 85-92.
- C. M. Croner, J. Sperling, and F. R. Broome (1996). Geographic Information Systems (GIS): New Perspectives in Understanding Human Health and Environmental Relationships. *Statistics in Medicine* 15: 1961-1977.
- M. Egenhofer and R. Franzosa (1991) Point-Set Topological Spatial Relations. *International Journal of Geographical Information Systems* 5(2): 161-174.
- M. Egenhofer and R. Franzosa (1995) On the Equivalence of Topological Relations. *International Journal of Geographical Information Systems* 9(2): 133-152.
- M. Egenhofer and R. Golledge (1994) *Time in Geographic Space: Report on the Specialist Meeting of Research Initiative 10*. National Center for Geographic Information and Analysis, Santa Barbara, CA, Technical Report 94-9.
- M. Egenhofer and R. Golledge, Eds. (1998) *Spatial and Temporal Reasoning in Geographic Information Systems*. Oxford University Press, New York.
- M. Egenhofer and J. Herring (1991) Categorizing Binary Topological Relationships Between Regions, Lines, and Points in Geographic Databases. in: M. Egenhofer, J. Herring, T. Smith, and K. Park (Eds.), *A Framework for the Definition of Topological Relationships and an Algebraic Approach to Spatial Reasoning within this Framework*, NCGIA Technical Report 91-7. National Center for Geographic Information and Analysis, Santa Barbara, CA.
- M. Egenhofer and D. Mark (1995) Naive Geography. in: A. Frank and W. Kuhn (Eds.), *Spatial Information Theory—A Theoretical Basis for GIS*, *International Conference COSIT '95, Semmering, Austria. Lecture Notes in Computer Science* 988, pp. 1-15, Springer-Verlag, Berlin.
- M. Egenhofer, J. Sharma, and D. Mark (1993) A Critical Comparison of the 4-Intersection and 9-Intersection Models for Spatial Relations: Formal Analysis. in: R. McMaster and M. Armstrong (Eds.), *Autocarto 11*, Minneapolis, MN, pp. 1-11.
- A. Frank (1992a) Qualitative Spatial Reasoning about Distances and Directions in Geographic Space. *Journal of Visual Languages and Computing* 3(4): 343-371.
- A. Frank (1992b) Beyond Query Languages for Geographic Databases: Data Cubes and Maps. in: G. Gambosi, M. Scholl, and H.-W. Six (Eds.), *Geographic Database Management Systems. Esprit Basic Research Series* pp. 5-17, Springer-Verlag, New York, NY.
- P. Forer (1998) Geometric Approaches to the Nexus of Time, Space, and Microprocess: Implementing a Practical Model for Mundane Socio-Spatial Systems. in: M. Egenhofer and R. Golledge (Eds.), *Spatial and Temporal Reasoning in Geographic Information Systems*. pp. 171-190, Oxford University Press, New York.

- J. Freudenheim, J. Marshall, J. Vena, R. Laughlin, J. Brasure, M. Swanson, T. Nemoto, and S. Graham (1996) Premenopausal breast Cancer Risk and Intake of Vegetables, Fruits, and Related Nutrients.
- S. Graham, R. Hellmann, J. Marshall, J. Freudenheim, J. Vena, M. Swanson, M. Zielezny, T. Nemoto, N. Stubbs, and T. Raimondo (1991) Nutritional Epidemiology of Postmenopausal Breast Cancer in Western New York. *American Journal of Epidemiology* 134(6), 552-566.
- S. Graham, M. Zielezny, J. Marshall, R. Priore, J. Freudenheim, J. Brasure, B. Haughey, P. Nasca, and M. Zdeb (1992) Diet in the Epidemiology of Postmenopausal Breast Cancer in the New York State Cohort. *American Journal of Epidemiology* 136(11), 1327-1337.
- J. Gray, A. Bosworth, A. Layman, and H. Pirahesh (1996) Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Total. in: S. Su (Ed.), *Proceedings of the Twelfth International Conference on Data Engineering*, New Orleans, LA, pp. 152-159.
- T. Hägerstrand (1967) *Innovation Diffusion as a Spatial Process*. The University of Chicago Press, Chicago, IL.
- T. Hägerstrand (1970) What About People in Regional Science? *Papers, Regional Science Association* 24: 1-21.
- Harris-Equifax (1993) Health Care Information Privacy: A Survey of the Public and Leaders. New York, NY: Louis Harris and Associates, Study No. 934009.
- D. Hernández (1994) *Qualitative Representation of Spatial Knowledge*. Springer-Verlag, New York, NY.
- J.-H. Hong, M. Egenhofer, and A. Frank (1995) On the Robustness of Qualitative Distance- and Direction Reasoning. in: D. Peuquet (Ed.), *Autocarto 12*, Charlotte, NC, pp. 301-310.
- K. Hornsby and M. Egenhofer (1997) Qualitative Representation of Change. in: S. Hirtle and A. Frank (Eds.), *Spatial Information Theory—A Theoretical Basis for GIS, International Conference COSIT '97, Laurel Highlands, PA. Lecture Notes in Computer Science* 1329, pp. 15-33, Springer-Verlag, Berlin.
- K. Hornsby and M. Egenhofer (1998) Identity-Based Change Operations for Composite Objects. in: N. Chrisman (Ed.), *Eighth International Symposium on Spatial Data Handling*, Vancouver, Canada, (in press).
- ISO (1996) *ISO 15046-20 Geographic Information - Spatial Operators*. International Organization for Standardization, Technical Report ISO/TC 211 N298.
- ISO/IEC (1996) *JTC1 SC21 Information Technology—Database Language—SQL/MM—Part 3: Spatial (SQL/MM Spatial)*. International Organization for Standardization, Technical Report ISO/IEC JTC 1/SC 21 N 10441.
- P. Kitsul and D. Philipov (1981) The One Year/Five Year Migration Problem. in: A. Rogers (Ed.) *Advances in Multiregional Mathematical Demography*. pp. 1-34. Laxenburg, Austria: International Institute for Applied Systems Analysis, Research Report 81-6.
- G. Langran (1992) *Time in Geographic Information Systems*. Taylor & Francis, London.
- J. Ledent (1980) Multistate Life Tables: Movement versus Transition Perspectives. *Environment and Planning A* 10: 537-560.
- D. Mark and A. Frank (1992) *NCGIA Initiative 2: Languages of Spatial Relations*. National Center for Geographic Information and Analysis, Santa Barbara, CA, Technical Report.
- D. Mark and A. Frank (1996) *Initiative 13: User Interfaces for Geographic Information Systems*. National Center for Geographic Information and Analysis, Santa Barbara, CA, Technical Report.

- J. D. Mayer (1983) The Role of Spatial Analysis and Geographic Data in the Detection of Disease Causation. *Social Science and Medicine* 17: 1213-1221.
- National Institutes of Health (1995) Geographic Information Systems In Environmental Health Sciences. PA-95-032. NIH Guide, Volume 24, Number 5, February 10, 1995.
- National Research Council (1994) *Counting People in the Information Age: New Approaches to the Census for 2000 and Beyond. Panel to Evaluate Alternative Census Methods, Committee on National Statistics*. National Academy Press, Washington, D.C.
- J. Odland (1998) Longitudinal Analysis of Migration and Mobility: Spatial Behavior in Explicitly Spatial Contexts. in: M. Egenhofer and R. Golledge (Eds.), *Spatial and Temporal Reasoning in Geographic Information Systems*. pp. 238-259, Oxford University Press, New York.
- OGC (1998) *The OpenGIS Abstract Specification Model, Topic 1: Feature Geometry, Version 3*. Open GIS Consortium, Wayland, MA, Technical Report 98-101.
- C. Papadimitriou, D. Suci, and V. Vianu (1996) Topological Queries in Spatial Databases. in: *Fifteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, Montreal, Canada, pp. 81-92.
- D. Randell, Z. Cui, and A. Cohn (1992) A Spatial Logic Based on Regions and Connection. in: B. Nebel, C. Rich, and W. Swartout (Eds.), *Principles of Knowledge Representation and Reasoning, KR '92*, Cambridge, MA, pp. 165-176.
- J. Richards and M. Egenhofer (1995) A Comparison of Two Direct-Manipulation GIS User Interfaces for Map Overlay. *Geographical Systems* 2(4): 267-290.
- A. Rogers (1973) The Multiregional Life Table. *Journal of Mathematical Sociology* 3: 127-37.
- A. Rogers (1975) *Introduction to Multiregional Mathematical Demography*. New York: Wiley.
- P. Rogerson (1990) Migration Analysis using Data with Time Intervals of Differing Widths. *Papers of the Regional Science Association* 68: 97-106.
- P. Rogerson, J. Burr and G. Lin (1997) Changes in the Geographic Proximity Between Parents and their Adult Children. *International Journal of Population Geography*, forthcoming.
- K. Rothman (1990) A Sobering Start for the Cluster Busters' Conference. *American Journal of Epidemiology* 132: S6-S13.
- SAIF (1996) Spatial Archive and Interchange Format.
http://www.env.gov.bc.ca/srmb/fmebc/SAIF_FMEBC.htm.
- M. S. Scott (1997) Extending Map Algebra Concepts for Volumetric Geographic Analysis. Proceedings, GIS/LIS'97.
- J. Sharma, D. Flewelling, and M. Egenhofer (1994) A Qualitative Spatial Reasoner. in: T. Waugh and R. Healey (Eds.), *Sixth International Symposium on Spatial Data Handling*, Edinburgh, Scotland, pp. 665-681.
- D. Sinton (1978) The Inherent Structure of Information as a Constraint to Analysis: Mapped Thematic Data as a Case Study. in: G. Dutton (Ed.), *Harvard Papers on Geographic Information Systems*. 7, pp. SINTON/1-SINTON/17, Addison-Wesley, Reading, MA.
- T. Smith (1996) Alexandria Digital Library. *Communications of the ACM* 38(4): 61-62.
- R. Snodgrass (1992) Temporal Databases. in: A. Frank, I. Campari, and U. Formentini (Eds.), *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space. Lecture Notes in Computer Science* 639, pp. 22-64, Springer-Verlag, Pisa.
- J. Snow (1936) *Snow on Cholera*. Oxford University Press, London.

- L. Stallones, J. R. Nuckols, and B. K. Berry (1992) Surveillance around Hazardous Waste Sites: Geographic Information Systems and Reproductive Outcomes. *Environmental Research* 59: 81-92.
- A. Tansel, J. Clifford, S. Gadia, S. Jajodia, A. Segev, and R. Snodgrass (1993) *Temporal Databases: Theory, Design, and Implementation*. Benjamin/Cummings, Redwood City, CA.
- M. Worboys (1998) A Generic Model for Spatio-Bitemporal Geographic Information. in: M. Egenhofer and R. Golledge (Eds.), *Spatial and Temporal Reasoning in Geographic Information Systems*. pp. 25-39, Oxford University Press, New York.