ST–ACTS: A Spatio-Temporal Activity Simulator*

Gyozo Gidofalvi†
Geomatic aps
Center for geoinformatik
gyg@geomatic.dk

Torben Bach Pedersen
Aalborg University
Department of Computer Science
tbp@cs.aau.dk

ABSTRACT

Creating complex spatio–temporal simulation models is a hot issue in the area of spatio–temporal databases [7]. While existing Moving Object Simulators (MOSs) address different *physical* aspects of mobility, they neglect the important *social* and *geo–demographical* aspects of it. This paper presents ST–ACTS, a Spatio–Temporal ACTivity Simulator that, using various geo–statistical data sources and intuitive principles, models the so far neglected aspects. ST–ACTS considers that (1) objects (representing mobile users) move from one spatio–temporal location to another with the objective of performing a certain activity at the latter location; (2) not all users are equally likely to perform a given activity; (3) certain activities are performed at certain locations and times; and (4) activities exhibit regularities that can be specific to a single user or to groups of users. Experimental results show that ST-ACTS is able to effectively generate realistic spatio–temporal distributions of activities, which make it essential for the development of adequate spatio–temporal data management and data mining techniques.

Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Database Applications; I.6.3 [**Simulation and Modelling**]: Applications; I.6.8 [**Simulation and Modelling**]: Types of Simulation—*Discrete event*

General Terms

Algorithms

Keywords

spatio–temporal data, data generation, moving object simulation, activity simulator, data mining

*ST–ACTS can be downloaded for research purposes from http://www.geomatic.dk/research/ST-ACTS/ in the form of a MATLAB toolbox. The source data, or simulated data for various predefined parameters are also obtainable.

†Contact author

ACM-GIS'06, November 10–11, 2006, Arlington, Virginia, USA.

1. INTRODUCTION

Simulation is widely accepted in database research as a low–cost method to provide synthetic data for designing and testing novel data types and access methods. Moving objects databases are a particular case of databases that represent and manage changes related to the movement of objects. To aid the development in moving object database research, a number of Moving Object Simulators (MOSs) have been developed [1, 5, 6, 8, 9, 11].

The so far developed MOSs have been using parameterizable random functions and road networks to model different physical aspects of the moving objects–such as their extent, environment and mobility–but they all neglect some important facts. When moving objects represent mobile users, most of the time the reason for movement is due to a clear objective. Namely, users move from one spatio–temporal location to another to accomplish some task, from hereon termed as perform an activity, at the latter location. For example, people do not just spend most of their nights at a particular location, they come *home* to be with their loved ones, to relax, eat and sleep. Similarly, people do not just spend most of their working days at any particular location, they go to a real–world facility, their *work place*, with the intention of working. Finally, based on their habits and likes, in their spare time, people (more or less regularly) go to other real–world facilities, which they like and are nearby.

To model the above mentioned social aspects of mobility is important for two reasons. First, the locations and times where activities can be performed and the patterns in these performed activities define a unique spatio–temporal distribution of moving objects that is essential for spatio–temporal database management. Second, the social aspects of mobility are essential when one wishes to extract spatio–temporal knowledge about the regularities in the behavior of mobile users. The field of spatio–temporal data mining is concerned with finding these regularities or patterns. To develop efficient and effective spatio–temporal data management and data mining techniques, large sets of spatio–temporal data is needed; and while location–enabled mobile terminals are increasingly available on the market, such data sets are not readily available.

Hence, to aid the development in spatio–temporal data management and data mining techniques, this paper presents ST–ACTS, a probabilistic, parameterizable, spatio–temporal activity simulator, which is based on a number of real–world data sources consisting of:

- fine–grained geo–demographic population,
- information about businesses and facilities, and
- related consumer surveys.

The importance of the use of real–world data sources in ST–ACTS lies in the fact, that they form a realistic base for simulation. Concretely, variables within any given data source are dependent, and perhaps most importantly geo–dependent. For example, there is a strong dependence between the education and the personal income of people. The variables are also geo–dependent, due to the fact that similar people or similar businesses tend to form clusters in the geographical space. Furthermore, variables are geo–dependent across the different data sources. For example, people working in bio–technology tend to try to find homes close to work places in that business branch. Using real–world data from various commercial geo–statistical databases and common sense principles, ST–ACTS captures some of the to date not modelled, yet important, characteristics of spatio–temporal activity data.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 defines the objectives of the simulation model. Section 4 describes in detail the source data that forms the basis for the simulation model. Section 5 describes each component of the simulator and how the source data is used in each component. Section 6 evaluates the simulation model in terms of its efficiency and its simulation objectives by examining the characteristics of some simulated data. Finally Section 7 concludes and points to future research directions.

2. RELATED WORK

Due to the short history of spatio–temporal data management, scientific work on spatio–temporal simulation can be restricted to a handful of publications. The first, significant spatio–temporal simulator is GSTD (Generate Spatio–Temporal Data) [11]. Starting with a distribution of points or rectangular objects, at every time step GSTD recalculates positional and shape changes of objects based on parameterized random functions. Through the introduction of a new parameter for controlling the change of direction and the use of rectangular objects to model obstacles, GSTD is extended to simulate more realistic movements, such as *preferred movement*, *group movements* and *obstructed movement* [6]. Since most objects use a network to get from one location to the other, Brinkhoff presents a framework for network–based moving object simulation [1]. The behavior of a moving object in this framework is influenced by (1) the attributes of the object having a particular object class, (2) the combined effects of the locations of other objects and the network capacity, and (3) the location of external objects that are independent of the network. These simulators and frameworks primarily model the physical aspects of mobility. While they can all be extended to model the social aspects, i.e., the objective for movement and the regularities in these objectives, they do not pursue to do so.

Nonetheless, the importance of modelling these social aspects of mobility is pointed out in [1]. In comparison, ST–ACTS focuses on these social aspects of mobility while placing only limited constrains on the physical aspects of mobility. In effect, the problem solved by the above MOSs is orthogonal to the problem solved by ST–ACTS.

In Oproto [8]–a realistic scenario generator for moving objects motivated by a fishing application–the moving behavior of objects is influenced by other, either stationary or moving, objects of various object types. The influence between objects of different types can either be attraction or repulsion. While the repulsive and attractive influence of other objects is an objective for movement, unlike ST–ACTS, Oporto does not allow the modelling of regularities in these objectives.

The GAMMA [5] (Generating Artificial Modeless Movement by genetic–Algorithm) framework represents moving object behavior as a trajectory in the location–temporal space and proposes two generic metrics to evaluate trajectory data sets. The generation of trajectories is treated as an optimization problem and is solved by a genetic algorithm. With appropriately modified genetic operators and fitness criteria the framework is used to generate cellular network trajectories that as frequently as possible cross cell boarders, and symbolic location trajectories that (1) exhibit mobility patterns similar to those present in a set of real–life sample trajectories given as input, (2) conform to real–life constraints and heuristics. Based on sample activity trajectories, the GAMMA framework can be configured to generate activity trajectories that contain real–life activity patterns. While the generated trajectories will be similar to the input trajectories, since they are symbolic, they will, as the input trajectories implicitly assume a location–dependent context, (see third and fourth principle in Section 3). To simulate spatio–temporal activities of an entire population, a representative sample of context–dependent trajectories is needed, but is hard to obtain. In comparison, ST–ACTS, based on intuitive principles and a number of real–life geo–statistical data sources, is able to generate realistic, spatio–temporal activity data that takes this location–dependent context of activities into account.

Time geography [4] is a conceptual basis/paradigm for human space–time behavior which considers (1) the indivisibility or corporeality of the human condition; (2) that humans typically operate over finite intervals of space and time; (3) the natural laws and social conventions that partially constrain space-time behavior; and (4) that humans are purposive. ST–ACTS models some aspects of this paradigm in a concrete, implemented data generator.

3. PROBLEM STATEMENT

Existing MOSs capture only *physical* aspects of mobility, i.e., the *movement* of the objects, adequately. However, to aid the development of spatio–temporal data management and data mining methods, *social* aspects of mobility that arise from human behavioral patterns should be captured by a model. The most important principles that govern these social aspects of mobility are:

First Principle: People move from a given location to another location with an *objective of performing some activity* at the latter location.

Second Principle: Not all people are equally likely to perform a given activity. The *likelihood of performing an activity* depends on the interest of a given person, which in turn depends on a number of demographic variables.

Third Principle: The *activities performed by a given person are highly context dependent.* Some of the more important parts of this context are: the current location of the person, the set of possible locations where a given activity can be performed, the current time, and the recent history of activities that the person has performed.

Fourth Principle: The *locations of facilities*, where a given activity can be performed, are *not randomly distributed*, but are influenced by the locations of other facilities and the locations of the users those facilities serve.

The first principle can be thought of as an axiom that is in relation to Newton's first law of motion. Movement that is motivated by the sole purpose of movement and does not obey this principle–for example movement arising from outdoor exercise activities–are not modelled.

The second principle can be rectified by many examples from real life. Two of these examples are that elderly people are more likely to go to a pharmacy than younger people and younger people are more likely to go to a pop or rock concert than elderly people.

The third, perhaps most important principle, is due to several factors. First, movement is a necessary (not always pleasurable) requirement to perform some activity, and hence in most cases the amount of movement required to do so is minimized by the actor, i.e., people tend to go to a café that is near by. Second, activities are not performed with equal likelihood at different times. For example, most people tend to go to work in the morning hours as opposed to other parts of the day; consequently the likelihood of performing that activity during in the morning is higher than during other periods of the day. Furthermore, due to their nature, different activities have different durations. The duration of a given activity puts a natural constraint on the possibility of performing another activity while the previous activity lasts. For example, people tend to start to work from the morning hours for a duration of approximately 8 hours; consequently the likelihood of grocery shopping during the same period is lower than otherwise. Finally, while a person may perform an activity with a very high likelihood, the activities performed by the person are not temporally independent. For example, it is very unlikely that even a person who likes pop and rock concerts a lot, goes to several performances during the same Saturday evening.

The fourth principle is mainly a result of the supply–and–demand laws of economics. Locations of facilities are mainly influenced by competition, market cost, and market potential. For example, eventhough the cost of establishing a solarium salon on the outskirts of town might be low, the market potential might not even compensate this low cost. Hence it is very unlikely that one will finds several solarium salons on one city block. The spatial process that gives rise to locations of facilities is a complex, dynamic process with feed–back, which is governed by the laws of competitive markets. Hence, using a snapshot of the spatial distribution of real–world facilities as contextual information forms a reasonable basis for constructing a realistically model of spatio–temporal activities that can be performed at those facilities.

The primary *qualitative* objective of the simulation model is to capture the above described governing principles of human behavioral patterns and is referred to as the *validity* of the simulation model. In addition, the simulation model has to achieve a number of *quantitative* objectives. First, the simulation model has to be *effective*, i.e., it has to be able to generate large amounts of synthetic data within a reasonable time. Second, the simulation model has to be *parameterizable*, i.e., based on user–defined parameters it has to be able to generate synthetic data sets with different sizes and characteristics. Finally, the simulation model has to be *correct*, i.e., the synthetic data produced by model has to have the same statistical properties with respect to patterns as it is defined by the model parameters and inputs.

referred entity	conzoom® variable	categories
person	person count	1
	age	9
	education type	9
	employment status type	12
	employment branch type	12
housing unit	unit count	1
	house type	6
	house ownership type	4
	house area	5
household	household count	1
	family type	5
	fortune	6
	personal income	5

Table 1: Variables in conzoom®.

4. SOURCE DATA

The source data used in the simulation model are commercial products of Geomatic, a Danish company specializing in geo–demographic data and analysis for market segmentation, business intelligence, and direct marketing [3]. Due to the commercial nature of these data sets, the methods of their exact derivations are not to be described herein. Nonetheless, concepts and principles used in the derivation process and the resulting relevant contents of the databases are explained below.

conzoom® Demographic Data: conzoom® is a commercial database product that contains fine–grained, geo–demographic information about Denmark's population [3]. The variables that describe the statistical characteristics of the population can be divided into three groups: *person*, *housing unit*, and *household* variables. These variables and the number of categories for each is shown in Table 1.

In Table 1, variables that have "type" in their names are categorical variables; variables that have "count" in their name are counts of the corresponding entities within a 100–meter grid cell; and finally, the rest of the variables are continuous variables that have been categorized into categories that are meaningful for market segmentation.

Since, for example in the countryside, the number of persons, households or units could be very low in a 100–meter grid cell, grid cells are grouped together into meaningful, large enough clusters to comply with social and ethical norms and preserve the privacy of individuals. The basis for clustering is twofold: geography and the publicly available one–to–one housing information. The intuition behind the basis is also twofold. First, people living in a given geographical region (be that a state, a county, a postal district) are similar in some sense; for example, they might have a more similar political orientation from people living in another geographical region. Second, people living in similar houses are likely to be similar in other demographic variables; for example an established family with a stable source of income is more likely to be able to buy a larger, more expensive house than a person who just started his/her career. As mentioned earlier, to preserve the privacy of individuals, the clusters are constrained to contain at least some fixed number of households. Statistics for the variables, depending on the sensitivity of the information contained in them, are obtained from Statistics Denmark [10] for clusters con-

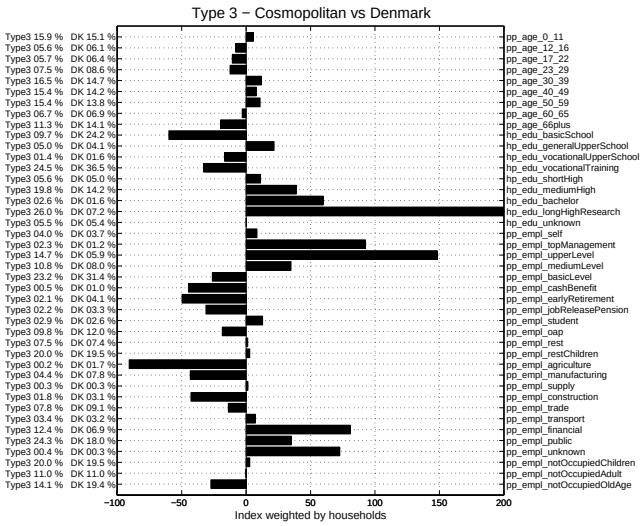

Figure 1: Partial profile of conzoom℗ type 3

structed at an appropriate level of cluster size constraint, for example 20, 50, 100, and 150 households per cluster. In case of a continuous variable, for example age, counts of the corresponding entities (in this case persons in the cluster) are obtained for the categories of the given variable.

Due to this constrained geo–clustering method, the conzoom® clusters obtained comply with the social and ethical norms and preserve the privacy of the individual, yet the statistics obtained are accurate enough for effective market segmentation. This segmentation results in grouping the Danish population into 29 conzoom® types, which are defined for each 100–meter grid cell. Cosmopolitan (type 3) is one example of the 29 conzoom® types. Comparing the demographics of type 3 to the demographics of the rest of Denmark's population gives the *demographic profile* of the type. This profile is partially shown in Figure 1. It roughly describes individuals that are more likely: to be middle aged (30–59 years old), to live in larger cities in larger, multi–family houses that are either owned by them or are private rentals, to be mostly couples with children, to have a medium to long higher education, to hold higher level or top management positions in the financial or public sector, and to have a better household economy (both in terms of wealth and income) than the average Dane.

mobidk™ Daily Movement Data: mobidk™ is an upcoming, commercial database product that contains detailed information about the daily movement of the Danish population between home and work [3]. Again, to preserve the privacy of users, the movement data is aggregated to non–overlapping and connected geographical regions. It is represented in a relational database format as: $\langle from\_region, to\_region, count \rangle$, meaning that from the geographical region *from_region*, *count* number of people move on a daily basis for work to the geographical region *to_region*. In ST–ACTS, these geographical regions are parishes, which on average contain 1176 households, and 195 100–meter grid cells [1].

[1]The commercial version of mobidk™ contains the same information for smaller, neighborhood clusters that on average contain 230 households and 38 100–meter grid cells.

bizmark™ Business Data: bizmark™ is a commercial database product that contains detailed information about Danish businesses both in the public and the private sector [3]. Some of the one–to–one information that is available about businesses is their location, the number of employees working in them, the physical size of the business facility, and the international branch codes the businesses fall under. Detailed but aggregated information about the employees within businesses is also available for appropriate bizmark™ clusters, which are constructed taking into account geography, business branch, business size in term of number of employees and physical size of the business facility, and various other descriptive business variables.

GallupPC℗ Consumer Survey Data: GallupPC® is a commercial database product and as the name suggests, it contains detailed survey responses of consumers about their demographics; interests such as culture, hobbies, and sports; household consumptions, purchasing habits; transportation habits; views on various subjects; attitudes and exposure to various advertisement media [2]. The questions in the surveys are yes/no questions. To measure the magnitude of the response of an individual survey subject to a specific question, the original yes/no question is re–phrased with a reference to a time–frequency interval. For example the original yes/no question "Do you go to the library?" is re–phrased to 7 yes/no questions using the following time–frequency intervals: daily / almost daily; 3-4 times a week; 1-2 times a week; 1-3 times a month; 1-5 times every 6 month; seldom, and never.

5. ST–ACTS: SPATIO-TEMPORAL ACTIVITY SIMULATOR

In this section, main components of ST–ACTS and their use of the source data is described. In the description a simulated person, who performs activities in time and space, will be abbreviated as a simperson. A MATLAB toolbox for ST–ACTS can be downloaded for research purposes from http://www.geomatic.dk/research/ST–ACTS/.

Drawing Demographic Variables for Simpersons: The conzoom® source data contains accurate, detailed demographic information about the population aggregated to a cluster level. As described in Section 4, continuous variables are discretized into categories. Clusters contain counts for all categories for all variables. Having the exact number of persons, housing units, and households at a grid cell level, and assuming the same distribution of variables in the individual grid cells as in the cluster they belong to, counts for all categories for all variables are calculated at a grid cell level. A simperson is assigned a category for a given variable proportional to the counts of the categories for the given variable in the grid cell the simperson lives in. In short, a category for the variable is assigned to the simperson according to the distribution of the variable. To draw assign categories for variables without replacement, corresponding counts in the given grid cell are decremented. Since counts of some of the variables in the grid cell refer to entities other than persons, but are variables that are part of the demographic variables that describe a person, these counts are adjusted to sum to the number of persons in the cell.

Skewing Distributions based on Correlations: The above described method for assigning categories for demo-

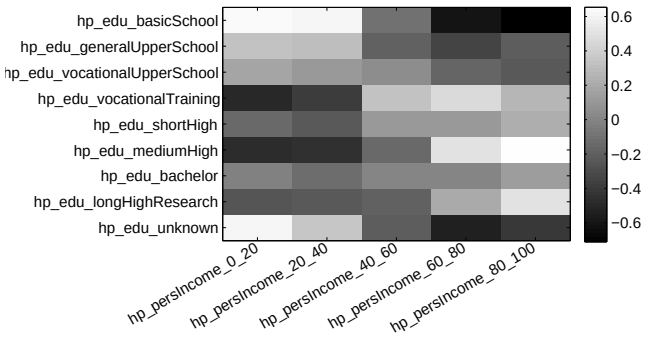

Figure 2: Correlation between education and income

graphic variables has one major flaw: demographic variables are not independent. For example the education type variable has a strong correlation with the personal income variable. This correlation is illustrated in Figure 2. Correlations are calculated between the percentages of the categorized variables, and samples are weighted by the number of persons in the cells. From the colorbar on the side one can see that darker shades mean stronger negative correlations and lighter shades mean stronger positive correlations. The correlations support the common knowledge that people having higher education levels tend to have better paying jobs. Similar correlations exist between other variables.

To remedy the above described flaw, which could result in unrealistic assignment of categories for variables to simpersons, the assignment is modified by drawing categories from skewed variable distributions that try to embed the correlations between the variables as follows. For a given simperson, the category for the first variable, age, is drawn without replacement from unskewed distribution of the age variable. An example of this distribution and the result of the draw is shown in the top most left subgraph of Figure 3, where for the age variable the category 5 was drawn, which represents that the simperson is in the age group 30–39. The distribution of the second variable, education, is shown in the second–from–top left subgraph of Figure 3. Given this distribution, categories 4, 6 and 8 are most likely to be assigned to the simperson for the education variable. However, the correlations (shown in the third left subfigure of Figure 3) between the age category 5 and education variable reveal positive correlations for categories 1 and 4, and a negative correlation for category 8 for the education variable. After normalizing (shifting to mean 1) the correlations, the original distribution of the education variable is skewed by pair–wise multiplying the raw counts of categories of the education variable and the normalized correlations for the education variable given that the age category of the simperson is 5. This skewed distribution is shown in the bottom left subgraph of Figure 3 and is used for sampling the education variable, resulting in the education category 4, vocational training. Values for further variables are drawn from skewed distributions that take into account the categories for the previously drawn variables, by skewing the distribution of the current variable by the average of the normalized correlations for the so far drawn categories. This process is shown from top to bottom on the right subfigures of Figure 3, where given that the age category a the simperson is 5 and the education category is 4 for the third variable, employment state, the category 11 is drawn.

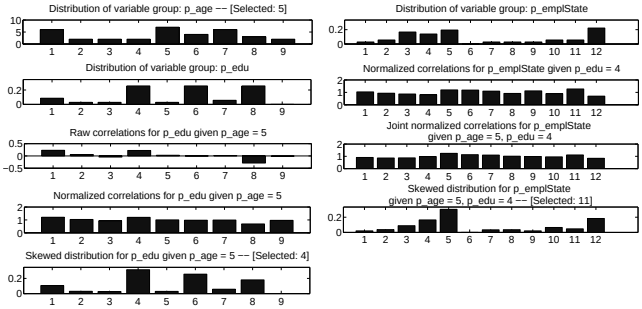

Figure 3: Drawing samples without replacement from correlated, multivariate distributions

Assigning Simpersons to Work Places / Schools: Activities can be divided into two groups: *free time activities* and *mandatory activities*. While the notion of "mandatory" activity may differ from person to person, for the purposes of simulation, ST–ACTS considers going to *school* and *work* as mandatory activities. The rest of the activities in ST–ACTS are considered free time activities.

With respect to mandatory activities, simpersons can be divided into three groups: retired, worker, and student. For the retired simpersons, it can be assumed that they enjoy the fruits of a hard–working life and have no mandatory activities. Consequently, they spend the majority of the time either at home or performing free time activities. The following paragraphs describe the methods in ST–ACTS (and their usage of the base data) for assigning simpersons in the worker and student groups to their work places and schools respectively.

Assigning Worker Simpersons to Work Places: Simpersons in the worker group are assigned to *work places* in two steps. In the first step, given the *home parish* and employment branch of the simperson, the parish–to–parish commuting probabilities, and the spatial distribution of businesses in branches, a *work parish* is assigned to the simperson. In the second step, given the employment branch that the simperson works in, businesses in the same branch that are located in the work parish of the simperson are retrieved from bizmark^{TM}. Finally, proportional to the number of employees that work in the selected businesses, the simperson is probabilistically assigned to one of the businesses / work places.

Assigning Student Simpersons to Schools: Simpersons in the student group are assigned to schools in two steps. In the first step, depending on the age group of the simperson, he or she is assigned to either one of the four educational institution types, or is assigned to be "not in school" and hence is considered to a member of the worker group. In the second step, educational institutions of the simpersons's educational institution type are retrieved from bizmark^{TM}, and the simperson is assigned to the institution that is closest to the simperson's home[2]. The following paragraph explains the first of these steps in more detail.

Simpersons in the student group can be divided into four subgroups based on which one, if any, of the four educational institution types they attend: kindergarten, primary

[2]The Danish public school system is controlled by the municipalities, which assign students to educational institutions that are nearby. Locations of these institutions are carefully planned to meet the needs of the population.

school, secondary school, or college / university. As described above, each simperson below age 30 is assigned to one of the four age groups: $[0, 11]$, $[12, 16]$, $[17, 22]$, and $[23, 29]$. Assuming all simpersons up to age 5 or 6 go to kindergarten (or daycare centers), simpersons in the $[0, 11]$ age group are assigned with equal likelihood to either a kindergarten, or a primary school. For each of the remaining three age groups, based on information obtained from Statistics Denmark [10], the probabilities of attending one of the four education institution types are derived, which are shown in the table:

	[12-16]	[17-22]	[23-29]
primary school	0.9198	0.0235	0.0002
secondary school	0.0654	0.4639	0.0552
college / university	0.0000	0.1194	0.2365
not in school	0.0148	0.3933	0.7081

Then, given the age group of the simperson and the corresponding probabilities, the simperson is assigned to either one of the three educational institution types, or to be "not in school" and is considered to be a member of the worker group.

Daily Activity Probabilities: A subset of the GallupPC® consumer survey questions, described in Section 4 represent activities that require the movement of the consumer. Some of these activities are shown on the y–axis of Figure 4. To preserve space and clarity, the following, additional activities are included in the model, but are excluded from the figure: art exhibition, church, pop/rock concert, museum, post office, theatre, solarium, hairdresser, and shopping. The shopping activity is further subdivided into 22 subtypes of shopping that are tied to a particular brand or type of store.

Using the geo-demographic parts of the surveys, each survey subject is assigned to one of the 29 conzoom® types. To derive a single indicator for how likely a given conzoom® type is to perform a given activity, the answers to the re–phrased time–frequency questions are normalized and averaged as follows. First, every time–frequency interval for an activity is normalized to represent the probability of performing the given activity on an average day. For example, a subject's positive reply to the question "Do you perform activity a n times during a period Δt?" equivalently means that the probability of that subject to perform activity a on any given day is $P(a) = n/day(\Delta t)$, where day is a function that returns the number of days in period Δt. $P(a)$ is equivalently referred to as the *Daily Activity Probability* (DAP) of activity a. Second, these daily activity probabilities of individual subjects of a given conzoom® type are averaged. Figure 4 shows a sample of these daily activity probabilities for a subset of the conzoom® types. From the figure it can be seen, for example, that a college student is most likely to go to a library, a cinema, a discotheque, or a fitness center; while a retired farmer is the least likely to perform these activities. Since the figure has the same probability scale, it also reveals that, depending on type, going to a fitness center is about a 7 to 22 times more frequent or popular activity as going to classical concerts. As mentioned before, ST–ACTS includes the daily activity probabilities of 35 activities for 29 conzoom® types.

Activity Simulation with Spatio–temporal Constraints: A simple, random, discrete event activity simulator can be constructed as follows. At every time step, a random subset of the simpersons is chosen to perform an activity. Then, for each selected simperson, given his/her conzoom® type and the associated daily activity probabilities, an activity is assigned. Then, each selected simperson is moved to the closest facility, where his/her assigned activity can be performed. This simple simulator does not model several spatio–temporal constraints on the activities. In the following, these constraints are discussed, and for each, the proposed modelling solution that ST–ACTS implements is presented.

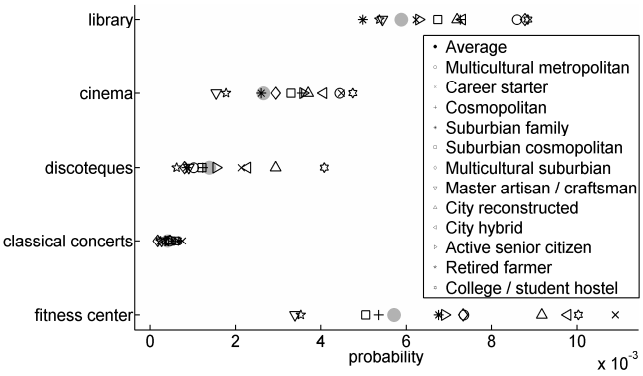

Figure 4: Sample daily activity probabilities

Temporal Activity Constraint: Certain activities are more likely to be performed during specific periods than others. For example, people in the work force tend to leave their homes for work at the beginning of a workday. Consequently, the same people are less likely to go to a discotheque, which is presumably closed, during the same period. To model the *Temporal Activity Constraint* (TAC), ST–ACTS allows the user to define for each of the three population groups the probabilities for each of the activities for every hour of every day of the week. These probabilities are used to limit the ability of the simperson to perform certain activities during certain time periods. They are not to be confused with the conzoom® type dependent *daily activity probabilities*, which encode the activity preference of each type. Through the TACs ST–ACTS allows the modelling of opening hours, and to some degree sequential patterns. The TACs of an activity are defined by a 7 by 24 matrix, where columns represent hours of the day, and rows represent days of the week.

Activity Duration Constraint: Not all activities take the same amount of time. For example people usually work 6-10 hours, spend about 2 hours in a cinema, and 30 minutes in a grocery store. To model this, from the starting timestamp of an activity a that is assigned to a simperson s, s becomes *occupied* for $\delta_{occupied}(a)$ time steps. During this period s is not assigned any other activities. In ST–ACTS, *Activity Duration Constraint* (ADC) for each activity are probabilistically drawn from the user–defined activity duration distributions, which is normally distributed with mean $\mu_{\delta_{occupied}(a)}$ and variance $\sigma_{\delta_{occupied}}(a)$.

Minimum Elapsed Time Between Activity Repetition Constraint: While people prefer some activities over others, it is very unlikely that they would repeat the same, even if preferred, activity many times, one–after–the–other within a short period. For example, it is very unlikely, that even a very active simperson, right after finishing his workout at the fitness center, decides to go to a fitness center

```
(0) //geo–demographic data (conzoom®): D
(0) //population movement data (mobidk™): M
(0) //business data (bizmark™): B
(1) ST–ACTS (T,ΔT,DAP,TAC,ADC,METC,MDC)
(2)   s.dem ← drawDemographicVariables(D)
(3)   s.work ← simpsToWork(s,B,M)
(4)   s.acts ← initSimpsActs(s,t=1)
(5)   for t = 1..T
(6)     free ← unoccupiedSimps(s.acts,t)
(7)     a ← validActsToFreeSimps(s,DAP,TAC,METC,t)
(8)     [f, d] ← facilitiesForActs(a,s,MDC,B)
(9)     δ_occupied ← durationsOfActs(a,ADC)
(10)    δ_trans ← transitionTimes(d,speed(d))
(11)    s.acts ← updateSimps(a,f.loc,δ_occupied,δ_trans,,t)
(12)  end for
```

Figure 5: Discrete event simulation in ST-ACTS

again. This constraint is modelled in ST–ACTS through the user–defined $\delta_{elapsed}(a)$, activity–dependent *Minimum Elapsed Time Constraint* (METC). The constraint is enforced by maintaining a recent history of activities for each simperson and validating newly drawn activities against it.

Maximum Distance Constraint: For most activities there is a *maximum distance* a person is willing to travel. This maximum distance represents a spatial constraint on the activities that a simperson s will perform, given the current location of s and the locations of facilities, where a selected activity a can be performed. Hence, during the simulation if there is no suitable facility for a within maximum distance of the current location of s, the activity is considered invalid for s, and s becomes idle. The *Maximum Distance Constraint* (MDC) is controlled by a user–defined, activity–dependent parameter in ST–ACTS.

Physical Mobility Constraint: To move from one location to another takes time. While detailed simulation of this movement is not an objective of ST–ACTS, basic physical mobility constraints are modelled. After a facility f for an activity a is selected for a simperson s, s is moved after δ_{trans} time steps to the new location. δ_{trans} is calculated based on the Euclidian distance d in km between the current location of s and the location of facility f, assuming a constant speed. This constant speed, in km/h, is probabilistically drawn from the distribution $speed(d) = \max(5, N(3d, d^2))$. $speed(d)$ assigns lower speeds to shorter, and higher speeds (with larger variance) to longer distances. It, to some extent, captures common modes of transportation, i.e., people tend to walk on shorter trips, use public transportation or bicycle on slightly longer trips, and use a car or commuting train on even longer trips.

Discrete Event Simulation: Using the conceptual building blocks presented so far, the discrete event simulation performed in ST–ACTS can be summarized as shown in Figure 5. The first three comments indicate that named data sets are used in the simulation, but are not user–defined parameters of it. Arguments to ST-ACTS, shown on line 1, are the user–defined parameters that have been described in the previous paragraphs. On line 2 demographic variables are assigned to simpersons based on skewed variable distributions. On line 3 simpersons are assigned to work places and schools. On line 4, at time step $t = 1$ (Monday, 00:00) all simpersons are initialized to be at "home" doing activity "home to stay" until the early morning hours. Following these preprocessing steps, at every time step t, on line 6, currently unoccupied (free) simpersons are found. Then, on line 7 for each free simperson a valid action is found according to the daily activity probabilities (DAP) of actions for the conzoom® type of the simperson. Actions are valid, if they both meet the temporal activity constraint (TAC) and the minimum elapsed time constraint (METC). On line 8 valid facilities are found for these valid activities. Facilities are valid if they meet the maximum distance constraint (MDC). On line 9, activity durations are drawn that meet the activity duration constraint (ADC). On line 10, according to the distances to the assigned activities, transition times are calculated. Finally, on line 11, information about the newly assigned activities are stored and the activity histories are updated for the affected simpersons.

6. EVALUATION OF THE SIMULATION

ST–ACTS was implemented and tested in MATLAB running on Windows XP on a 3.6GHz Pentium 4 processor with 1.5 GB main memory. The geographical extent of ST-ACTS was restricted to the municipalities of Copenhagen and Frederiksberg in Denmark. In this extent, the number of simpersons is 590,050 (178,826 retired, 268,615 workers, and 142,609 students), the number of working places is 1,264,129 in 193,299 businesses, and the number of facilities is 10,544. Simulation experiments were performed for a time step length of $\Delta T = 5$ minutes. To test the performance of ST–ACTS, in all experiments "strict" TACs were set on the two most likely activities, go "home to visit" and go "home to stay". TACs of other activities were set to model opening hours of corresponding facilities. As a result a simperson performs on average 9.6 ± 3.2 activities per day.

To evaluate the effectiveness of ST-ACTS, simulations were performed for varying sizes of randomly selected subsets of simpersons during the course of a single day (24 hours). Figure 6 shows both the CPU times (right y–axis) and the I/O time for logging the events (left y–axis). Both of these quantities scale approximately linearly with the number of simpersons. In short, the simulation is fast and scales well.

In a larger experiment activities of the total population for the course of a full week have been simulated. The table below shows the output of ST-ACTS for a cosmopolitan type simperson during the course of a day.

a.begin	a.loc(x)	a.loc(y)	a.end	a.name
8:35	722941	6172634	15:50	work / school
17:05	720408	6173933	17:45	Fakta
18:55	721350	6177550	20:20	home to visit
20:45	723555	6175390	21:10	solarium
21:50	723483	6175299	23:30	cinema
23:40	721350	6177550	8:25	home to stay

The simulation, without logging the individual events and only keeping statistics about activities, took 98 minutes. To evaluate the validity of ST-ACTS, the gathered statistics have been analyzed. Due to space limitation, only some results of this analysis are discussed in detail, while others are only summarized.

To evaluate ST-ACTS's ability to generate the correct distribution of activities, the input DAPs have been compared to the simulated DAPs, shown in Figure 7. While, due to the previously mentioned "strict" TACs, the simulated DAPs are about 4 times higher than the input DAPs,

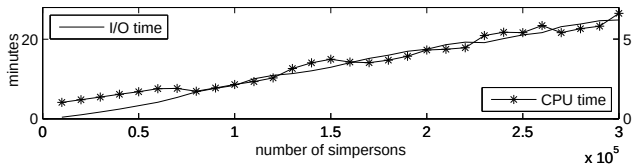

Figure 6: CPU and I/O times for simulations

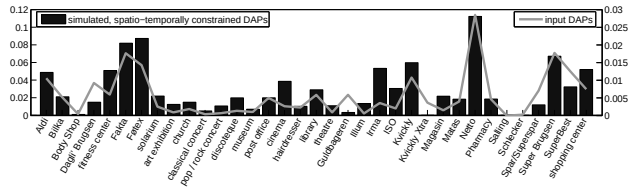

Figure 7: Input and simulated DAPs

the relative simulated DAPs among activities is similar to the input DAPs. By using less "strict" TACs, i.e.: allowing simpersons to go home earlier after work, the scale of simulated DAPs match that of the input DAPs. Differences in the relative DAPs can be explained by the effects of spatio-temporal constraints on activities.

To evaluate ST–ACTS's ability to control temporal constraints on activities, counts for each assigned activity for every hour–of–day and day–of–week were maintained. Figure 8 shows the average number of assigned activities for each hour–of–day averaged over the days–of–week. Due to the large variation in frequency counts for different activities in different periods of the day, the base 2 logarithm of frequency counts are shown. From the figure it can be seen that certain groups perform certain activities at certain times of the day more frequently than other groups. For example, it can be seen that the retired group is more likely to perform activities during working hours, simply because they are free to do so. Opening and closing times of facilities is also controlled by the parameters. For example, no-one goes to discotheques during the day, and no-one goes to shopping centers in the middle of the night.

To evaluate ST–ACTS's ability to control spatial constraints on activities, the daily distance travelled to work by an average simperson (2.7 ± 2.3 km) was compared to the total daily distance travelled by an average simperson (8.3 ± 3.6 km). While for the same numbers no ground truth was available to evaluate against, considering the average 9.6 activities per day the numbers seem reasonable. The simulated data has also been verified that no trips violate the activity–dependent maximum distance criteria.

7. CONCLUSIONS

Realistic models that simulate the spatio–temporal activities of users, and hence the distribution of moving objects, are essential to facilitate the development of adequate spatio–temporal data management and data mining techniques. In this paper, ST-ACTS, the first of such simulators is presented. Experimental results show that, using a number of real–world geo–statistical data sources and intuitive principles, ST–ACTS is able to effectively generate realistic spatio–temporal activity data. It is also demonstrated that the generated data has the same characteristics as it is defined by the user–controllable model parameters. ST-ACTS has been implemented in MATLAB and is available for research purposes.

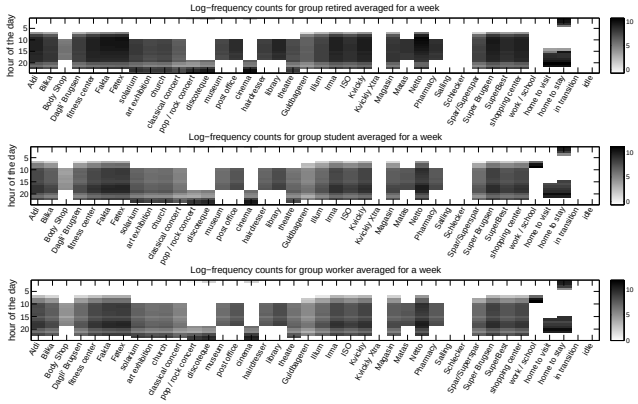

Figure 8: Validity of ST–ACTS in terms of TACs

While the correspondence between the characteristics of the generated data and the model parameters is demonstrated, the accuracy of the simulation has to be necessarily affected by the limited modelling of physical aspects of mobility. Hence in future work, integrating the output of ST-ACTS as an input to sophisticated network–based moving object simulation as in [1] is planned. Such a more complex simulator will provide synthetic data sets that can aid the development in telemathics, intelligent transportation systems, and location–based services.

Acknowledgments

This work was supported in part by the Danish Ministry of Science, Technology, and Innovation under grant number 61480. Making conzoom®, bizmark™, and GallupPC® available for research purposes is gratefully acknowledged to Geomatic ApS., and The Gallup Organization. Providing population movement data at a parish level, is gratefully acknowledged to Thomas Nielsen from the Danish Center for Forest, Landscape and Planning. The help from co–workers, Susanne Caroc, Esben Taudorf, Jesper Christiansen, and Lau Marcussen is also gratefully acknowledged.

8. REFERENCES

[1] T. Brinkhoff. A Framework for Generating Network–Based Moving Objects. *Geoinformatica 6(2)*, pp.153–180, 2002.
[2] The Gallup Organization: http://www.gallup.com/
[3] Geomatic ApS – Center for Geoinformatik: http://www.geomatic.dk
[4] T. Hägerstrand. "Space, time and human conditions." In *Dynamic allocation of urban space*, ed. A. Karlqvist et. al. Lexington: Saxon House Lexington Book, 1975.
[5] H. Hu and D. L. Lee. GAMMA: A Framework for Moving Object Simulation. In *Proc. of SSTD*, pp. 37–54, 2005.
[6] D. Pfoser and Y. Theodoridis. Generating Semantics–Based Trajectories of Moving Objects. In *Proc. of Emerging Technologies for Geo–Based Applications*, pp. 59–76, 2000.
[7] J. F. Roddick, M. J. Egenhofer, E. Hoel, D. Papadias, and B. Salzberg. Spatial, Temporal and Spatio–Temporal Databases–Hot Issues and Directions for PhD Research. In *Proc. of SSTD*, pp. 1–6, 2003.
[8] J.-M. Saglio and J. Moreira. Oporto: A Realistic Scenario Generator for Moving Objects. In *Proc. of DEXA* Workshop on Spatio–Temporal Data Models & Languages, pp. 426–432, 1999.
[9] T. Sellis. Research Issues in Spatio–temporal Database Systems. In *Proc. of SSD*, pp. 5–11, 1999.
[10] Statistics Denmark: http://www.dst.dk
[11] Y. Theodoridis, J. R. O. Silva, and M. A. Nascimento. On the Generation of Spatiotemporal Datasets. In *Proc. of SSD*, pp. 147–164, 1999.