Title
Activity-Based Travel Analysis in the Wireless Information Age

Author
Marca, James E

Publication Date
2002

UNIVERSITY OF CALIFORNIA,
IRVINE

Activity-Based Travel Analysis in the Wireless Information Age

Dissertation

submitted in partial satisfaction of the requirements
for the degree of

Doctor of Philosophy
in Civil Engineering

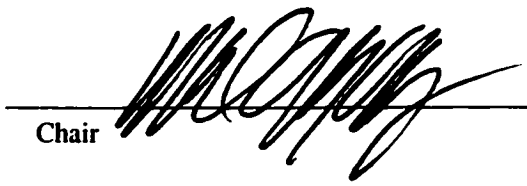
by

James Edward Marca

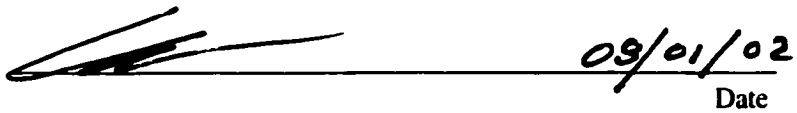
Dissertation Committee:
Professor Michael G. McNally, Chair
Professor Wilfred R. Recker
Professor R. Jayakrishnan

2002

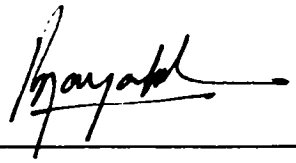
The dissertation of James Edward Marca
is approved and is acceptable in quality
and form for publication on microfilm:



Chair Date



Date



Date

University of California, Irvine

2002

Dedication

To my wife, Brooke Tomblin, whose love, encouragement, patience, and support made this dissertation possible.

And to my daughter Grace, who cheerfully distracted me.

Contents

List of Figures

List of Tables

Acknowledgments

I want to thank my committee chair, Dr. M^{c}Nally, for inspiring me to think about activities and travel. I also want to thank my committee members, Dr. Jayakrishnan, and Dr. Recker, whose advice and comments pushed me to clarify and expand many of the ideas in this dissertation. Finally, I am indebted to Craig Rindt for reading and commenting on early drafts of this work.

Financial support was provided primarily by University of California Transportation Center (UCTC) graduate fellowships.

Curriculum Vitae

James Edward Marca

2002 Ph.D. in Civil Engineering, University of California, Irvine

1993–1996 Associate, Charles River Associates, Boston, MA

1994 M.S. in Civil Engineering, University of California, Irvine

1991 Programmer, Sparta Corporation, Laguna Hills, CA

1989–1990 Systems Engineer, Rocketdyne Div. of Rockwell Intl., Canoga Park, CA

1989 B.S. in Engineering, Harvey Mudd College, Claremont, CA

Field of Study

Activity-based analysis of transportation demand.

Abstract of the Dissertation

Activity-Based Travel Analysis in the Wireless Information Age

by

James Edward Marca

Doctor of Philosophy in Civil Engineering

University of California, Irvine 2002

Professor Michael G. McNally, Chair

One of the main barriers to a better understanding of activities and travel patterns is the difficulty in collecting long-duration data. Previous studies have examined computer-aided interview techniques. Others have researched the potential for global positioning system (GPS) antennas to collect more accurate travel data. This dissertation combines these two techniques by adding the use of wireless communications technology to integrate streaming, real-time GPS data with a dynamically generated, web-based activity survey. In addition, three separate analysis techniques are examined using the results of an informal pilot test. The purpose of these analysis techniques is to weave together the large set of GPS data that can be collected with the much smaller set of activity responses that can

be expected. The net result represents both an advance in data collection techniques, as well as a new, peer-to-peer approach to gathering and sharing experiential transportation information, an approach that should be incorporated into future Intelligent Transportation Systems designs.

Chapter 1

Activity analysis and data collection

This dissertation examines the collection and processing of activity data. The central tenet of activity-based approaches to travel demand analysis is that the demand for travel is *derived* from a superseding demand to perform activities. When the activity under consideration cannot be performed at the place the actor finds himself or herself, then the person must travel to a destination at which the activity can be performed. In theory then, to understand travel one must understand the activities that lead to travel, and by extension, to simulate realistic travel patterns, one must be able to simulate activities.

Our current understanding of activity engagement and the limited data that can be collected restrict simulations to generating a simplified set of synthetic activities. These simplifications stretch the hypothesis that the activities so generated are actually demands that supersede the demand to travel. In other words, the simulated activities have only a weak ability to generate travel patterns.

For example, one of the better sources of activity data is Portland METRO (1994). This data set contains over 100,000 activities recorded over a two-day period. The data

are clean, well coded, and fully cross-referenced with the characteristics of the person and the household. A reasonable model could be estimated from these observations for generating two-day activity sequences. But that model would not contain any information on the dynamic behavior of the individual. If one were evaluating travel patterns for a largely static population, then one could feel confident that the activities generated are in turn generating a demand for travel. But if one were evaluating a dynamic problem, such as the impact of gradually increasing population levels, there are neither mechanisms nor data available to describe how activity patterns change in response to changing resources or environments.

In order to estimate a true, activity-centric model of travel behavior, we need data that captures the medium and long range dynamics of individuals. Thus this dissertation focuses on new techniques for gathering and processing activity data that will enable the estimation of models that can generate realistic synthetic activity patterns.

This work took two tacks. First, steps were taken to integrate the hardware and software necessary to gather many real-time activity data streams, as described in chapter 3. The hardware was developed by a team consisting of the author and other researchers at UCI. The software to collect and store GPS data was primarily designed and implemented by the author. This data collection system collects positions, not activities. To annotate positions with activity descriptions, the near real-time stream of GPS points transmitted from the data collection unit was integrated into a web-based activity survey. This survey is documented in chapter 4.

Second, since streaming, automatic collection of positions and activities may produce huge amounts of data, it is important to integrate data analysis into the data collection pro-

cess. This work originated by applying sequence alignment techniques (Gusfield, 1997) to the Portland data. This analysis demonstrated the deficiencies of that data set for generating longer sequences of activity patterns. Sequence alignment methods and spatial statistical analysis techniques have been combined into a small set of classification and analysis tools. Chapter 5 documents the analysis of activity destinations as a spatial point process. Chapter 6 describes a method for predicting activity names for new, as yet unlabeled data points. And chapter 7 documents the application of the optimal local alignment problem to identifying possible activity patterns in a long string of activities. All of these techniques have been integrated into the on-line survey website.

These chapters also apply these data analysis techniques to the data collected in the course of an informal pilot survey. The point process analyses of chapter 5 suggest that the level of spatial randomness in activity destinations is quite low. That is, the collected destinations are all tightly clustered, rather than being randomly distributed in space. Chapter 6 explains a method for leveraging that clustering behavior to predict the relative frequencies of the activities that have been observed so far at nearby points. This method would be useless if people did not travel to the same destinations for their activities. The pilot data showed that the unlabeled points are often close enough to labeled points to be assigned an activity name with a high degree of confidence.

The development and application of sequence alignment techniques for detecting activity patterns within the activity data streams are documented in chapter 7. Patterns of activities are detected in the unaugmented activities, but the low numbers of activities entered in the on-line survey limit the usefulness of this technique. Three different strategies were tested to augment the raw sequences. First, travel activities were assigned to the

travel periods. If no activities were entered for a day, this would make a sequence of travel acts, which detects patterns based on the amount of daily travel. Second, the results of the probabilistically assigned names were added to label a subset of the unlabeled points. And finally, a brute force association of space and activities was attempted by tagging all of the activities with an extra name based on their longitude and latitude. The results compare the effect of these different approaches on the size and extent of the resulting activity patterns. A larger, more formal survey is needed to exploit the sequence alignment techniques and compare commonalities and differences in the resulting patterns across individuals.

The usual stated purpose for pursuing activity-based approaches to travel analysis has been to improve the accuracy and benefits of planning models. While not disparaging this goal, more useful applications of the principles of activity analysis may be to enrich and inform the development of on-line, near real-time traveler information devices. While the web-based activity survey developed here has many similarities to traditional data collection tools, with a modest level effort the web site could be converted into a distributed, decentralized advanced traveler information system. These ideas are expanded in chapter 8. Finally, chapter 9 concludes and summarizes the contributions of this dissertation, and give a road map for future research.

The next chapter introduces activity-based travel analysis. While not an exhaustive review of the literature, it covers the major themes and directions of research, with an eye towards identifying concepts which apply to a dynamic activity simulation context. The core ideas and research results of the activity-based literature are then used to explain the rationale behind the major contributions of this dissertation. Chapter 2 concludes with a short description of my approach to the collection of activity data.

Chapter 2

Literature review

2.1 Activity constraints

Arguably, Hägerstrand (1970) started the field of activity analysis by proposing that people as social creatures be explicitly included in regional science. Hägerstrand theorized that the web of interactions that defines the modern social system is the result of the interactions of individual strands of time and space representing the life-arcs of individuals. In saying this, he cautioned that "one risks becoming lost in a description of how aggregate behavior develops as a sum total of individual behavior, without arriving at essential clues toward an understanding of how the system works as a whole. It seems to be more promising to try to define the time-space mechanics of constraints which determine how the paths are channeled or dammed up. ... I am going to look at the matter entirely from the point of view of constraints." To this end, Hägerstrand introduced the concept of time-space prisms limiting behavior. The idea is that people must travel forward in time, and cannot skip or repeat segments of time. Furthermore, travel through space must be contiguous, is limited

by the maximum speed of travel, and takes place on the available transportation system.

Objective constraints such as speed limits and road designs are not enough to adequately describe a population's movements. Despite the confines of the time-space prism, the number of options available to people are still essentially infinite. In addition to the constraints on time and space, Hägerstrand observed that there are other "capability constraints," such as needing to go home at night, needing to sleep, needing to eat, and so on. These constraints refer to limitations due to biology as well as to the capabilities of the individual (ability to drive, etc).

Individuals also often need to interact with others, either directly or indirectly, in order for an activity to take place. This requirement gives rise to a class of constraints that Hägerstrand called "coupling constraints," in which activities can occur only if the prerequisite space-time paths intersect (in space and time) in appropriate ways (people arrive on time for a meeting, trucks deliver groceries to a store, and so on). Finally, there are constraints which place limits on the range of activities that are possible. Hägerstrand called these "authority constraints." They capture the influence of others upon the individual (speed limits, operating hours, etc.).

Although Hägerstrand argued that the individual person was important to regional science, he understood that importance to cover the individual's interaction with the real world, *not* the individual's internal motivations and desires. His approach, which generally goes by the name *time-geography* (Lenntorp, 1976), stresses the examination of what is possible for an individual within space and time. Thus it is a modeling structure of constraints. Lenntorp (1976) describes the output of this research effort as "an instrument ... for assessing what is, and what is not, geometrically possible ('for assessing which fu-

tures have a geometric possibility of becoming history')." The internal motivations of an individual are largely irrelevant to this tool—the past is understood to influence the individual, which in turn may cause echoes of past time-space paths in future activity programs, but the mechanisms governing this linkage are viewed as an unfathomable black box by Hägerstrand's time-geography.

The contribution of this school of thought to travel modeling is to consider what is and isn't possible and why. For example, in a strict trip-based methodology, the benefits from improving a freeway link are assessed by measuring the time saved by trips that use the freeway link. The time-geography approach would use the improved speeds on the link in question to formulate a different constraint space, which implies an entirely new action space for all affected individuals. These revised action spaces enable a different set of activity programs which take advantage of the improved travel times, either directly by enabling longer trips at higher speeds, or indirectly by freeing up time and other resources for engaging in other activities.

The approach that perhaps takes time-geography's enumeration of constraints the furthest is the HAPP model (Recker, 1995). Here the time-space prisms of real observations, taken from the Portland two day activity diary (Portland METRO, 1994), are manipulated in an optimization framework that also incorporates some known traffic conditions and other features. However, the formulation of Recker (1995) is quite complex as it stands. It remains to be seen whether or not other constraints and objectives, such as social norms or societal obligations, can also be incorporated successfully into this framework.

2.2 Theories of intentions

The complement to the analysis of activities via constraints is to examine the intentions for engaging in activities. The most common theory of personal intentions is that of choosing the utility maximizing behavior from amongst several alternatives. However, utility maximization is not the only model of individual intentions. Chapin (1974) summarizes empirical work he conducted in the previous decades. His approach is probably best described as social-behavioral. He examined the impacts of city form at the macro level upon the residents of the city, and developed a model of relationships between the city characteristics and the activities and opportunities of its residents. While founded upon empirical studies, most notably a large study of the time-use of Washington D.C. residents, Chapin's work is probably more influential for its theoretical offerings.

For a simple set of activities with a limited number of characteristics, specifying and estimating a choice model is significantly less computationally intensive than enumerating and examining the all of the constraints that bound that simple set of activities. Documenting perhaps the most fully developed, practical example of a choice-based activity model, Ben-Akiva and Bowman (1995) build directly upon the highly effective application of utility maximization logit models to mode choice modeling. In their theoretical framework, they build a nested, sequential picture of the activity choice process. Instead of merely modeling the choice of travel mode, the choices expand to include primary activities, secondary activities, primary tours, and secondary tours, as well as the mode and destination choices embedded in each of these choices.

This effort is a good example of both the practical benefits of applying activity-based

modeling concepts to travel demand forecasting, as well as the inherent limitations of a continued expansion of strictly choice-based models. The benefits are obviously the ability to link a large scale travel demand model to activities. Planners can explore the impact of simple, aggregate behavioral changes, such as increasing the percentage of people who prefer to telecommute. Instead of just eliminating a trip to and from work, the activity choice model can explore the possibility that telecommuting individuals will trade their eliminated work trip for other trips and errands during the day.

On the downside, while the model is much more complicated to estimate and run than a conventional four-step planning model, it does not capture the full influence of activities upon travel. For example, the choice-based framework presumes that people make decisions about multiple activities and travel behaviors simultaneously, prior to ever engaging in any real action. Other models of decision making or even a mixture or decision styles are not allowed in the theoretical framework, in order to make the model tractable. This means that asynchronous interrupts, such as an unforeseen incident or problem, are nearly impossible to incorporate into a person's travel pattern once the "choice" of a daily pattern has been made. Ideally, one would want an activity model to allow the actors to continuously monitor their environment and the demands placed upon their time, in order to dynamically shuffle activities and destinations.

Other limitations also exist for the implemented model. The primary activities listed by Ben-Akiva and Bowman (1995) include only *home, work, school,* and *other*. Further, the decisions of urban actors are modeled as having distinct time scales, so that choices about residential location and work site are made independently and separately from choices about daily activities. The day is divided into periods rather than simulated minute by

minute. Finally, they state that the secondary destination choices were eliminated from the model in order to simplify its implementation.

The reductions to the complexity of the model have been made so as to limit the geometric explosion of the choice space. The model does run and is a significant improvement on the previous generation of travel forecasting models. But the ultimate activity-based model that expands the activities, combines long-term and short-term behaviors, and models time either continuously or at very small time steps has yet to be realized. It is likely that pursuing the choice-based framework will not produce this model.

Note that the individual-level constraints proposed by time-geography cannot really be applied to the aggregated people, zones, and time-scales used by Ben-Akiva and Bowman (1995). The personal and coupling constraints proposed by Hägerstrand require keeping track of individuals, not aggregating individuals across zones. This leads to a computational double-whammy—one has to generate and solve an extremely large choice problem on the one hand, while on the other an interesting technique that would place realistic boundaries on the choice problem will probably take as long or longer to compute.

2.3 Mixing intentions and constraints

Jones et al. (1983) developed an extensive home interview technique, called HATS. HATS framed the household activity and travel scheduling problem as a game to be solved during the interview by all members of the household. The optimization of travel patterns and activity destinations for each individual was set within a context of household-level constraints. In addition to significant results relating life cycle stage (a capability con-

straint) to activity engagement (Clarke and Dix, 1983), they also were able to develop theories about the internal motivations of individuals, and the larger needs and drives of the household.

Vaughn et al. (1997) propose a method that all but eliminates any choice process in the generation of activities for a household. Instead, the authors map the characteristics of a simulated household onto a set of real household characteristics collected in a survey, where all of the choices have already been made by real people. The real household closest to the synthetic household is selected, and then that real household's activity patterns are assigned to the synthetic household, except for the geographic information, which is discarded. The sampled pattern of activities forms a so called "skeletal" pattern for the synthetic household. After iteratively incorporating the particular constraints faced by the synthetic household, such as the local geographic conditions, the skeletal pattern is turned into the actual patterns performed by each member of the synthesized household.

The strength of this approach is that it is computationally tractable. However, this tractability comes at some cost. First, by breaking the link between the observed households' spatial environments and their own activity patterns when creating their synthetic, skeletal patterns, Vaughn et al. (1997) miss important motivating context behind the observed activity pattern (Harvey, 1997). Second, the approach is limited by observed activity patterns.

An approach that combines time-space constraints with a sequential model of activity generation and then activity engagement is the STARCHILD model and its descendents (M^{c}Nally, 1997; M^{c}Nally and Recker, 1986; Recker et al., 1986a,b). This research effort initially produced a theoretical framework for viewing activities in the context of complex

household and spatial constraints, and an operational simulation that put the theoretical ideas into practice. The more recent work on this approach has addressed placing the generation of activity patterns in a more tractable computational format, as well as incorporating geographic information system (GIS) technologies into the simulation of activity destination choice.

STARCHILD's focus was on enumerating complete patterns of activity that loop out from the simulated person's home and return some time later. The household was taken to be the fundamental activity-generating entity. Individuals belonging to each household were assigned activity patterns based on the needs of the household. This led to a sequential process of estimating and generating household needs and activities, then individual needs and activities, and then fitting in any possible "unplanned" activities. Most of the theoretical and practical work was devoted to specifying and estimating this process of activity sequence generation, and to mapping the many possible activity sequences for each person onto the available transportation and land-use resources.

Another theme in the activity analysis literature is to study the process of selecting an activity outside of a discrete choice based approach. These approaches use collected activity and time use data to build models of an individual's cognitive processes, such as scheduling activities, and choosing paths between activities. They generally build upon the work done in artificial intelligence on devising computational process models (CPMs). The most advanced of these is the *ALBATROSS* framework, which is discussed in the next section.

2.4 The *ALBATROSS* system

ALBATROSS , A Learning BAsed TRansportation Oriented Simulation System, is the modeling system developed by researchers at the University of Eindhoven. This section focuses on the general framework of *ALBATROSS* given in Arentze and Timmermans (2000b). The concepts behind the *ALBATROSS* model are similar to the agent-based approach adopted by the author (Marca et al., 1999), and since it is much more complete than the author's framework, it is worth examining in detail. The important innovations include the acknowledgment that activities and travel take place in a complex, dynamic environment, and in devising a computational mechanism that allows agents to adapt, improve, and abandon choice heuristics used to negotiate the environment. The *ALBATROSS* use of agent refers more to software agents that govern the simulation, rather than microscopic agents who mimic the behavior of real people, as in Marca et al. (1999). At this stage, the difference is immaterial. When our simulation approach matures enough to produce results, then the differences in the concept of an agent will become more relevant.

The general framework of the *ALBATROSS* system is reproduced in figure 2.1. This is not significantly different from the many other frameworks that have been proposed. The main difference is that the *ALBATROSS* project is perhaps the only one that has come close to implementing this large scale framework. Personal and household characteristics, the cognitive environment, situational conditions and constraints, and system properties and policies are all modeled as influencing to some degree or other the short, medium, and long-term dynamics of the activity and travel behavior of individuals and households. Within the decisions themselves, long-term decisions are placed at the root of the decision tree,

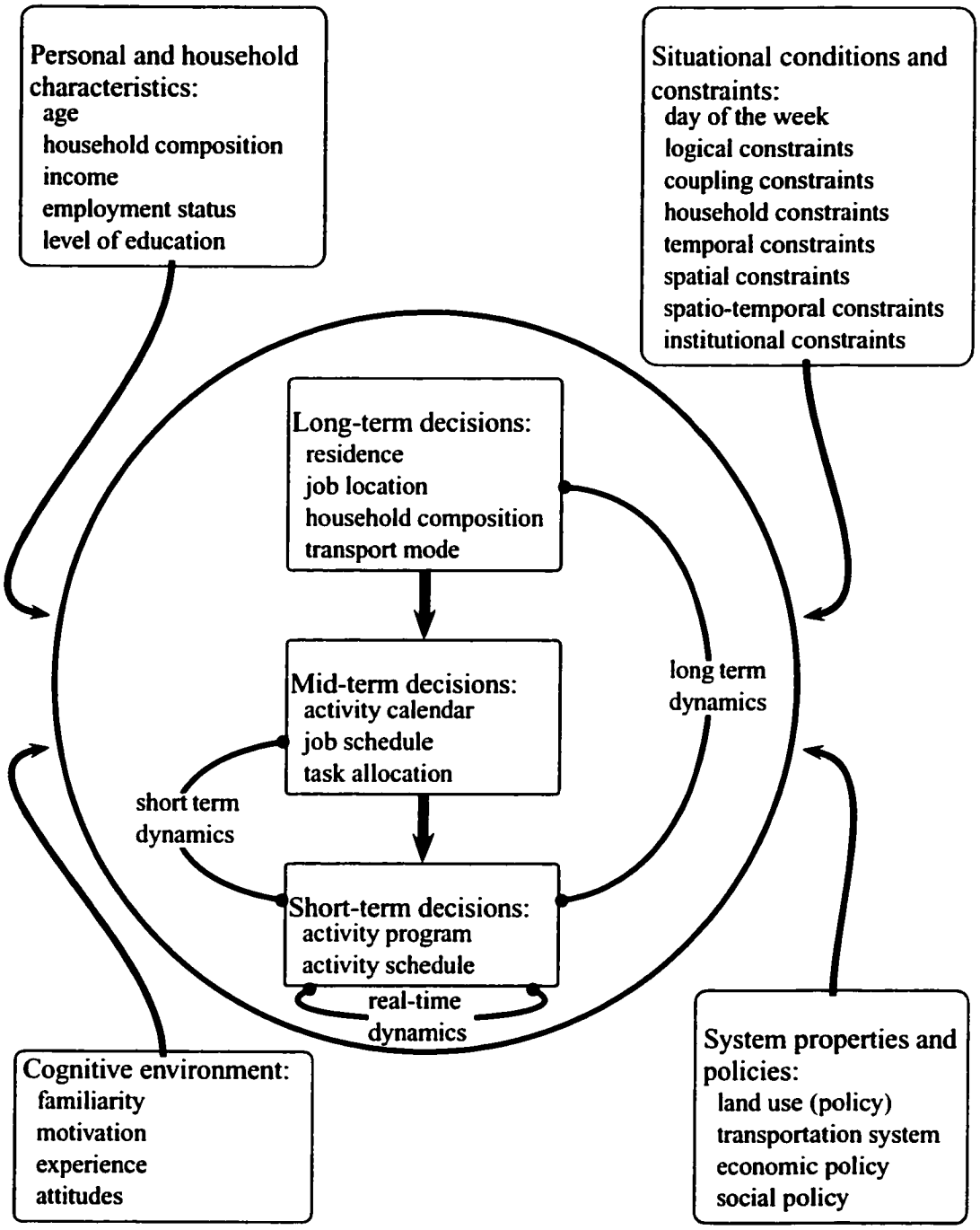


Figure 2.1: Conceptual framework of the *ALBATROSS* system, after Arentze and Timmermans (2000b)

owing to their importance and relatively infrequent and inflexible nature. The selection of residence, job type, and job location, and so on, are made once and then influence medium and short range choices. These shorter term actions in turn influence the long term choices, or more appropriately, influence the evaluation of the long-term choices and the frequency and type of the next long-term choice decision.

As can be seen in figure 2.1, $\mathcal{ALBATROSS}$ is a model of households and household interactions, similar to other activity based simulations such as STARCHILD. Arentze and Timmermans (2000b) state explicitly that "activity participation, allocation and implementation fundamentally take place at the level of the household." $\mathcal{ALBATROSS}$ assumes that household needs are accomplished through the actions of the household members—individuals in the simulation. That is, the set of actions of one household member, A_{hi}, are a subset of the entire set of household activities, A_h. There is no mechanism that explicitly generates an individual's activities independent of the household.

The long-range set of household activities is called a calendar, and the daily supply of household activities is called the activity program. This program consists of activities that are *planned* to be executed in the coming day. This program is then split amongst the household members and sequenced in time and space, resulting in the activity schedule. The realization of the activity program as actual activities and travel is called the activity pattern. The pattern may end up being different from the planned schedule, and may or may not include all of the activities listed in the daily activity program. The agenda–pattern–realization sequence is a common design pattern in activity modeling (Bhat and Koppelman, 1993; Doherty and Axhausen, 1998). This design pattern guides the approach taken by two recent computerized survey tools, CHASE (Doherty and Miller, 2000) and

REACT! (Lee and McNally, 2001), which are discussed in section 4.2.

The mid-term decisions within the *ALBATROSS* framework include devising the household calendar and formulating activity schedules. Allocation of the tasks on the household calendar amongst the household members is also viewed as a mid-term decision. That is, tasks are viewed as being relatively stable assignments, rather than dynamically shifted between household members on a daily basis. It is not clear whether there exist mechanisms to adjust how tasks are allocated between household members when the actual simulation model is run. One would expect that if this were a variable part of the model, it would be mentioned in the conceptual framework. Our agent-based framework explicitly allows activities to be shifted between actors based on communication and negotiation mechanisms.

The short term decisions in the *ALBATROSS* framework involve devising a daily activity program (what each actor will attempt to do) and then planning out the day's program in an activity schedule. The actors then attempt to perform their schedules within the simulated environment. This generates experimental knowledge of the environment that propagates up to longer time scales. Short term activities also allow feedback between the program and the schedule determination, based on the ability to perform activities within the dynamic environment.

ALBATROSS is based on a model of individual cognition. The approach views individuals as resource limited; they exist within an environment that both offers them opportunities and imposes constraints, but they are unable to consider all possible choices, nor are they able to deterministically choose the best alternative. *ALBATROSS* uses a computational process model (CPM) that incorporates these limitations, instead of relying

upon the more common utility maximization choice model. The CPM methods and terminology are borrowed from the artificial intelligence and data mining communities. A good introduction to this approach, although unrelated to the *ALBATROSS* work, is given by Thill and Wheeler (2000). The general approach is to derive a decision tree which classifies a population of observed events. The goal of the tree-building algorithms is to deduce a parsimonious tree which maximizes the information gain at each branching. The observed events consist of the choice process of interest—choosing a destination, selecting an activity sequence, and so on. The decision space consists of all possible choices. Using the characteristics of the alternate choices in question combined with the observed outcomes, the decision tree uses as little information as possible to sort the observed outcomes into similar categories.

The *ALBATROSS* approach has adapted these techniques to estimate a CPM from collected activity diary data (Arentze and Timmermans, 2001a). But instead of relying strictly upon observed survey data, it appears that they have also devised a technique for choosing sequences of production rules, called scripts, which optimize the local outcomes for the individuals in the simulation Arentze and Timmermans (2001b). The optimization has been demonstrated for goal seeking and information seeking behavior. In theory at least, the software agents controlling the simulation could modify CPM trees during the simulation to reflect changing situations. These revised CPMs capture the local choices made by individuals in the simulation who might be engaging in goal seeking or information seeking behaviors.

This framework has relevance for the process of data collection and analysis. First of all, it requires longer sequences of observations to identify periods of information gathering

behavior versus activity and travel pattern improvement behavior. Second, if people are in fact using different high-level heuristics to govern their behavior patterns, then a snapshot activity diary will capture not only good strategies, but also bad ones, situations that are in flux, environments that are new to the actor, and so on. It is these dynamic aspects that offer the most insight into how people learn about their environment and tailor their activities accordingly, but it is just these aspects which are hardest to identify in a *short-term* survey of one or two days. Longer time periods should be studied to better understand the activity and travel patterns that are stable, versus those that are dynamic adjustments, mistakes, or merely initial searches.

2.5 Towards a multiagent simulation framework

Despite a complex model of the individual's computational process, the $\mathcal{ALBATROSS}$ implementation is firmly rooted in the household. This is a curious dichotomy. The approach guiding this dissertation research differs from the $\mathcal{ALBATROSS}$ point of view in that the activity system is perceived as the result of interactions of individuals, not of households. The assumption of household primacy cannot experiment with the conditions that might lead to different kinds of households, with behavior that exists outside of the household, or with activities by an individual that run counter to the needs of a household. In contrast, an individual-centric modeling framework should define a household as the result of individuals cooperating with each other in the pursuit of household goals. Households should be emergent properties of the model, rather than explicitly described. The difficulty is that this requires the development of a model of cooperation. But this model

brings with it the ability to model any cooperation and interaction between any two agents, not just households.

A *motivation* for cooperation is needed in order to make it happen. At first this motivation will be most likely be definitional—household members cooperate with other household members because they are household members. As such the practical differences between *ALBATROSS* and this approach at this time are slim. However, as the model becomes more sophisticated, the intention is to explore the impact of different models of cooperation and motivation, so that household members cooperate based on the internal models of each interacting agent. This approach would allow households to form and dissolve, and for co-habitating adults to behave more like independent individuals than a household. Most importantly, this approach defines household formation as an emergent property of the rules that govern individuals, and allows other groups to emerge as well, such as friends or coworkers. If the gradual development of this approach is successful, it should produce an agent based simulation framework in which the macro-scale behaviors such as travel patterns and timings are the result of micro-scale behaviors that are directly programmed into the individual agents. If the model can produce interesting, emergent behavior from low-level parameter specifications, then we feel it will contribute significantly to our understanding of the impacts of the wide range of human behavior on transportation facilities in particular, and the urban environment in general.

In this aspect Hägerstrand's admonition cited at the beginning of this chapter against becoming lost in an exploration of the interactions is being ignored. The most interesting features of multi-agent simulation models arise not from explicitly formulating the higher level boundaries and limits of behavior, but by modeling the details of the interactions of

individuals, and observing what aggregate properties of the population emerge from the simulation. The insights into the fundamental properties of interacting agents derived from these simulations will inform our understanding of interacting people. This in turn will allow us to design transportation systems which enhance travel *and* activity patterns, rather than just speeding up trips from A to B.

2.6 Activities: real, measured, and simulated

In order to devise models of individual behavior, better data collection techniques are needed to observe the long term dynamics of real individuals. Chapter 1 stated that the primary contributions of this dissertation relate to data collection. This section provides the motivation for this work, moving from the theory of activity based travel analysis to the practical aspects of measuring activities.

There is a difference between activities that are performed, the names and attributes recorded in a survey, and the activities that are generated in a simulation. The actor in the real world doesn't care what name might be assigned, or even how the name will be assigned or what will be named. Things get done, and that's that. Academic analyses of activities focus on common features across individuals, and devise relatively precise definitions accordingly. The data are analyzed relying upon these precise definitions to form conclusions about how travel is dependent upon activities. Then within the context of a simulation, an activities are used to generate the stratified types of travel behavior.

Examples abound in the literature of analysts trimming the many categories in a survey set into just a handful of analyzed activity names. For example, (Golob and McNally, 1997)

reduce the 28 different activity names assigned to the Portland activity survey (Portland METRO, 1994) into 3 broad categories: subsistence activities, non-discretionary activities, and discretionary activities. Golob (2000) provides a concise synthesis of the theory behind dividing activities into types of activities. Harvey (1997) summarizes many of the advances and techniques of the time use literature for the activity analysis audience. These techniques include ways to measure and assign primary and secondary activities to a particular block of time, for example.

The crux of the problem from a measurement standpoint is that in order to compare activities across individuals, the activity names must be consistent. The ideal situation would be to have all activity dimensions orthogonal, especially the name of the activity since names are often used as shorthand for all other dimensions. The usual three classes of activities—subsistence, non-discretionary, and discretionary activities—do not overlap. The goal of common activity names sets up a conflict within the survey instrument. On the one hand, the survey designer tries to offer the respondent a complete set of activity names, but using language that can be translated into a small set of canonical activity names. On the other hand, the survey respondent is forced to fit his or her activities into one or two of these descriptions.

The goal of the survey instrument documented in chapter 4 is to collect as long a sequence of activities as possible from survey participants. The primary goal is to find repeated patterns *within* a single individual's activities. In this type of analysis, the only requirement is that the activity names be consistent within a single individual's responses. Instead of focusing on the name, it is easier to focus on other dimensions that are more precisely measured. Specifically, with the aid of a GPS unit, one can measure the loca-

tion and start time of an activity quite well, with the caveat that the current vehicle-based measurement of GPS positions measures a trip end rather than an activity start. In the next chapter, a GPS measurement system is described that collects such data and streams it back in real time to a database. Chapter 4 describes the companion web-based activity survey that allows survey participants to annotate their travel destinations with the names of the activities they perform using their own words and phrases.

The resulting system records activities whose names are unique to each respondent. Stripping the activity name of its meaning *across* individuals does not mean that no name can be attached to common activity and travel behavior within a simulation or policy analysis context. The representative activity pattern approach (McNally, 1997; McNally and Recker, 1986) also breaks the link between the actual activities and the name assigned. The "representative" part of the name refers to the central activity pattern within the activity pattern class. In the same way, one can examine behaviors of individuals, determine common patterns, and then assign common names to activities associated with a particular vector of activity and travel characteristics. These common names can be used to label the activity sequence in a simulation or policy review context.

Chapter 3

Automating activity data collection

3.1 Introduction

This chapter documents the development and implementation of a wireless data collection system, referred to as the Tracer system. Tracer was developed for a number of end applications. As one of the principal developers of Tracer, the author deliberately implemented functionality that relates to the ideas explored in this dissertation.

As the name suggests, the Tracer system traces the movement of equipped vehicles through space and time. The hardware component consists of the extensible data collection unit (EDCU) used to collect GPS data and transmit it wirelessly. The software component consists of programs to enable the collection, transmission, storage, processing and use of the GPS data. The implementation of a web-based activity diary using the Tracer position data is documented in chapter 4, and chapters 5 through 7 describe some preliminary analysis of the pilot data, using data analysis tools and techniques that have been integrated into the on-line survey web site.

Collecting GPS data for use in a travel analysis context certainly is not a new idea. The next section documents other approaches that have been reported in the literature. There are undoubtedly many more practical implementations, as this exciting technology has a great many researchers hopping on the bandwagon. Among travel behavior researchers, our implementation, described in section 3.3, appears to be unique in its usage of a real-time wireless data transmission technology. The description of the EDCU and the Tracer system includes the client and server programs that enable the actual data collection and storage.

3.2 GPS data collection

The Tracer system arose from a concept of an *extensible data collection unit* that could be used for a variety of purposes. A major inspiration was the success of the Federal Highway Administration's Lexington Area Travel Data Collection Test (Battelle, 1997; Murakami and Wagner, 1999). This project demonstrated some of the advantages of using a GPS device to collect data automatically. While GPS antennas have errors associated with them that must be handled appropriately (Wolf et al., 1999), these errors are largely mechanical and quantifiable. This is a tremendous improvement over the variable and intractable human errors related to personal reporting of travel times, routes, and destinations.

Most of the initial implementations of GPS data collection devices have focused on quantifying the errors arising with the use of this new technology, and establishing contexts for their use. For example, the Lexington study compared reported trip rates with GPS measurements, discovering among other things that while reported trip start and end times

are rounded off to 5 or 15 minute intervals, the GPS records show an a uniform distribution of departure times (Murakami and Wagner, 1999). Using the same data, Yalamanchili et al. (1999) provide some descriptive statistics of the differences between GPS and recall data. They also focus extensively on evidence of trip-chaining detection and propensity. They observe that the households that participated in the GPS/handheld computer subsample did not enter trip purpose information for all of the trips that the GPS device recorded, with information provided primarily for "significant" trip purposes.

Wolf et al. (1999) explore the use of a GPS device combined with a personal digital assistant (PDA) for collecting travel data in the Atlanta metropolitan area. They make several recommendations for reducing and managing the errors related to the use of such a device, with a focus on the integration of GPS points into a geographic information system. Later work by these authors explored using the GPS device as the exclusive means of collecting travel and activity data and doing away with the traditional travel survey in favor of a shorter follow-up telephone interview (Wolf et al., 2001). Rather than relying on the travelers as the primary explanatory source for the measured GPS points, the approach relies upon geographic information system (GIS) data to provide a first cut at establishing trip purposes, routes and destination. The authors compare traditional travel data collected by respondents in a small feasibility study, to data that can be inferred from entering the GPS measurements into a GIS. They claim that the comparison shows that the purely mechanical approach is often superior to relying on respondent data, primarily because of the human errors and underreporting of the respondents.

A large-scale application of GPS data collection is taking place in the city of Atlanta, in association with the SMARTRAQ project at the Georgia Institute of Technology which

produced the above research. This travel and activity survey includes a subsample of approximately 500 individuals who are recording their travel using a PDA with a GPS antenna. Unfortunately, the survey is ongoing as of this writing, and so no reports have been published documenting its methods or results. This project is one of the largest applications to date, and its results should be very informative.

Bachu et al. (2001) also test relying exclusively upon the GPS unit to collect travel data. Their method uses a commercial, battery-powered GPS recording device to collect travel data for several days. Their data collection instrument does not include a PDA, as they deliberately wanted to avoid the battery-life penalties as well as the potential user interface biases. The GPS recorders are collected from the respondents after a few days, and the data are entered into a GIS to generate printed maps. These maps are then shown to the respondents in a face to face follow-up survey to verify the accuracy of the GPS tracking, and to prompt the respondents for information that cannot be recorded by the GPS devices. The authors explain that their methods try to reduce the reliance upon hand-held computers, devices whose very newness is perhaps a barrier to use by survey respondents.

Draijer et al. (2000) examine the use of GPS records in multiple travel mode studies, not just in personal vehicles. For this purpose the designed a data logging platform called GeoMate, weighing approximately 2 kg and integrating a battery, an embedded computer, and a GPS antenna into a standard video-camera shoulder bag. The GPS antenna was attached to the shoulder strap of the camera bag to afford it a better view of the sky. While the GeoMates were reportedly somewhat temperamental, losing data completely from 22 of the 151 respondents, in general the device was able to be used by respondents on any type of travel mode, including walking, bicycling, public transit, and car. The study focused

both on the technical ability to use the GeoMate on different travel modes, as well as on the willingness of respondents to use the device. On the technical side, respondents indicated that the device was too bulky or otherwise inconvenient for non-motorized travel modes, and also often inconvenient for trip purposes such as shopping and visits. In terms of the willingness of the respondents to use the device, it was found that privacy was an important concern, but did not significantly decrease the response rate.

At the time of the design of the EDCU, we were aware of the Lexington study, and of encouraging reports of the benefits of GPS measurements. We drew up a specification and RFP for a small embedded system, with requirements that it fill a variety of different functions (M^{c}Nally, 1999). The primary goal was to collect and save GPS trip data, which would allow us to analyze traffic conditions and travel behavior. The secondary goals outlined in the RFP were to transmit data in real-time, and integrate the collected data with more detailed survey tools. This functionality would enable dynamic travel behavior studies which used GPS data both as a memory jogger, and as a means to tailor specific questions. The real-time transmission capability was also vital to the use of this device in a traffic probe function, which was important to the needs of the funding project. Interestingly, no other GPS devices used for travel behavior research seem to be leveraging wireless technology for real-time transmission of the collected data, although there are several examples in fleet management and freight operations. Finally, the specification noted that the ideal implementation should be easily extended, both in hardware and software, so that the device could be outfitted with newer technology rather than becoming obsolete in six months. The actual implementation of this specification is discussed in the next section.

3.3 The Extensible Data Collection Unit (EDCU) and supporting software

Aero Data (2002) was selected to build the EDCU. They designed and delivered a unit, shown in figure 3.1, which consists of

- an external GPS antenna;
- 16MB CompactFlash removable storage (easily upgradable);
- 32MB RAM, providing a 16MB ramdisk for the runtime filesystem;
- an internal CDPD wireless data modem with an external antenna;
- an uninterruptible power supply (UPS);
- a standard ethernet connection; and
- PC104 board risers, plus space for 2–3 expansion boards.

The EDCU does not have a built-in user interface at this time. This was a deliberate design choice, both to reduce the cost of the device, and because the technology of handheld computers is rapidly evolving. Indeed, at the present time we are exploring a future user interface expansion exploiting local area wireless technology such as 802.11b, which has become common and inexpensive only recently. These kinds of expansions can be done thanks to the unit's expansion board slots as well as its open operating system.

The EDCU has been designed to be used in an automobile. The power is supplied by the 12V cigarette lighter that is standard equipment in cars, or by the so-called utility port

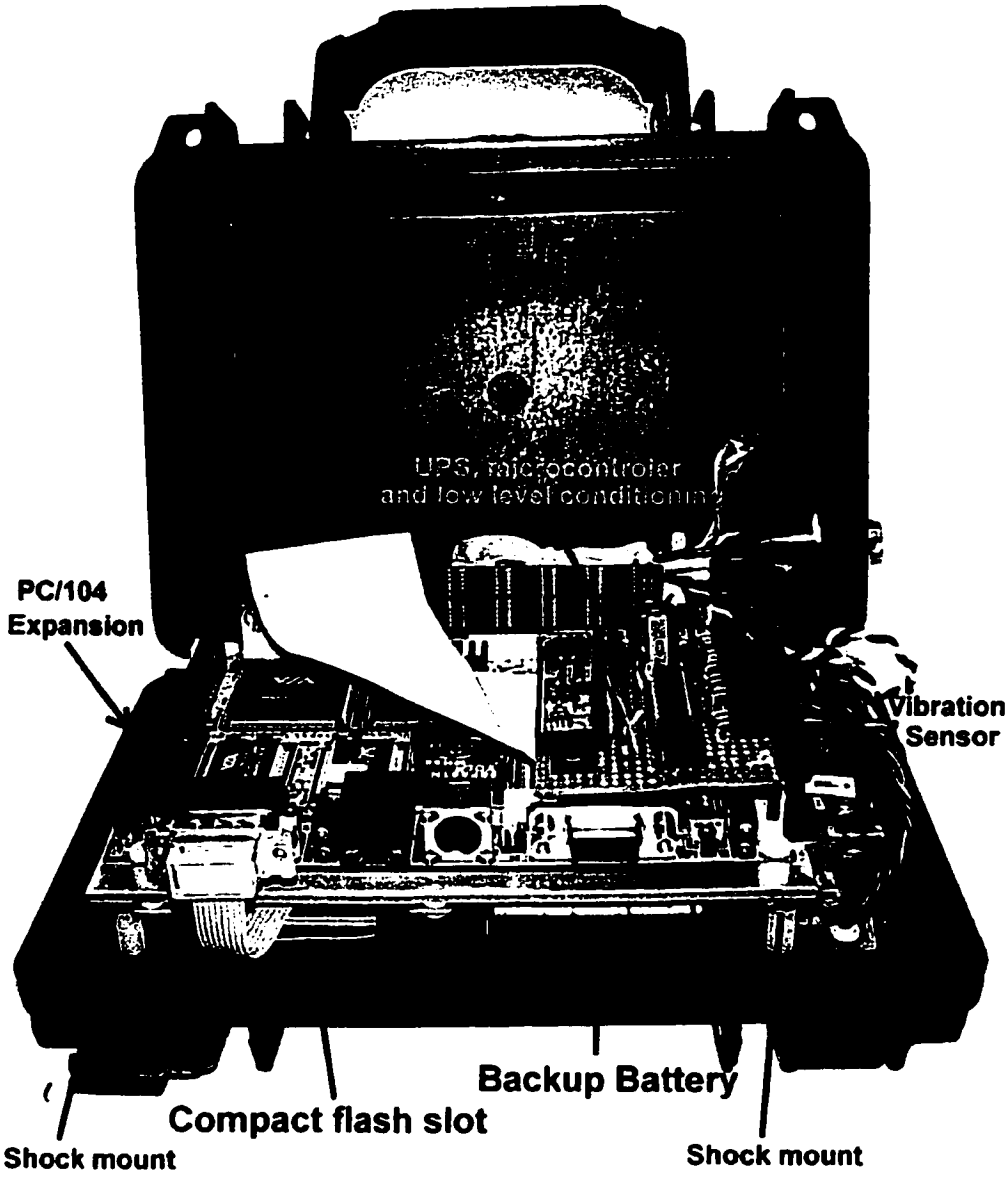


Figure 3.1: The Extensible Data Collection Unit (EDCU)

which is beginning to replace the cigarette lighter. The difference seems to be that the cigarette lighter is always powered by the car battery, whereas the utility port switches off when the car is turned off. To accommodate use in a car in which the power is always on, the EDCU has a motion sensor switch, and will shut itself down after 30 minutes of inactivity. To accommodate vehicles which kill power when the car is shut off, the EDCU has an uninterruptible power supply (UPS), using a small rechargeable battery to provide power to the unit as it shuts itself down gracefully. In practice, neither solution was foolproof, but they generally worked as intended.

The EDCU software runs on the Linux operating system—specifically an older version of the Linux Embedded Appliance Firewall (LEAF) distribution (Linux Embedded Appliance Firewall (LEAF), 2002). On top of the operating system, the unit runs an HTTP (web) server, and telnet and FTP servers. The web server and FTP capabilities enable remote control of the EDCU units, in case it should be necessary to change their behavior in the field, as well as the ability to download specific log files. Since Linux is a network aware operating system, it was straightforward to develop wireless client and server programs for the EDCU. These programs automatically log data in a central database while the EDCU is active.

The EDCU can collect and save GPS data in multiple ways, as described in figure 3.2. An internal GPS monitor listens to the signals from the GPS antenna, and saves the data to a files in a directory on the CompactFlash memory card. This memory card is the primary repository of collected data. The GPS data can be "hand carried" to the database by plugging this CompactFlash memory card into a card reader attached to a laptop or desktop computer. If the cellular modem is up and working, then data are transmitted auto-

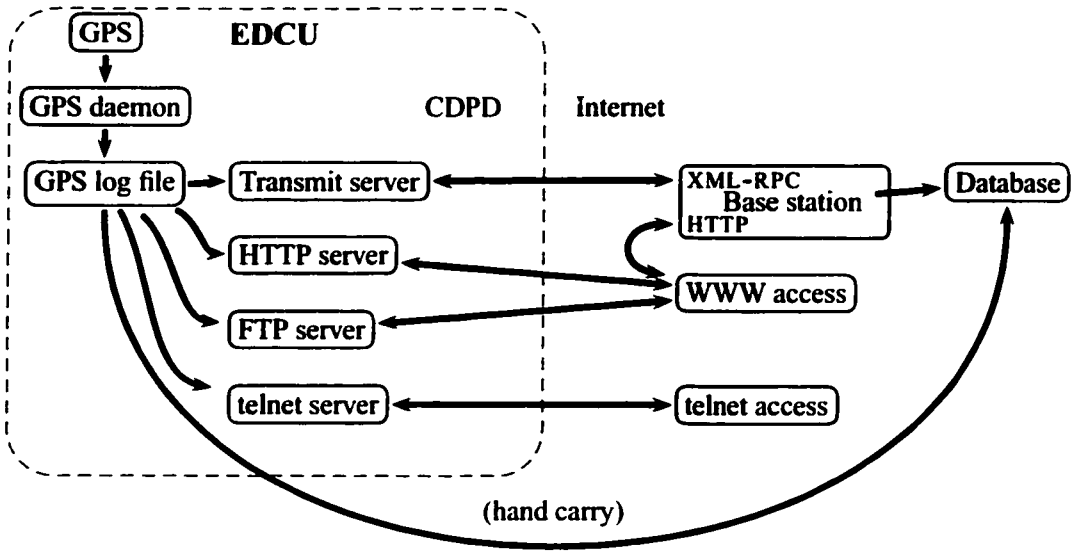


Figure 3.2: Flow of information from the GPS antenna to the database. Data are transmitted automatically while the vehicle is moving via the Transmit server. Alternately GPS log files can be downloaded via the HTTP and FTP servers. One can also login directly via the telnet server. Finally, the GPS data can be "hand carried" to the database by removing the CompactFlash memory card.

matically while the vehicle is moving via the Transmit server, programmed by the author. This server watches the log file for new data. Every few seconds, it transmits the latest set of GPS records to the Base Station, using XML-RPC (UserLand Software, 2001; Warner, 2001) calls over the standard HTTP Internet protocol. Log files can be downloaded via the HTTP and the ftp servers. One can also login directly via the telnet server, which is useful both for developing applications as well as for changing parameters in the field. In practice, the automatic real-time wireless connection via the Transmit server has been used almost exclusively, due to its ease of use.

The base station operations are filled by web and database servers running at UCI. The functions filled by the base station computers include receiving and logging real-time data to a database, and providing a web interface server. The web interface allows analysts and/or drivers to get information on deployed EDCUs, and view a sketch graphic of the

data collected by each unit. These operations are discussed in subsequent chapters.

3.4 Operational experience

The primary advantages of the EDCU, as built and delivered, are its openness and its flexibility. While the basic function of collecting GPS data will likely remain the same, we can add functionality to the unit to suit our needs. The Linux operating system enabled the easy implementation of network protocols, such as the XML-RPC client and server programs that automatically log data. We were also able to setup the EDCU to serve as a firewall and router for any computers (such as a laptop) connected to its ethernet port. This allows these computers to get wireless Internet access via the EDCU's cellular data connection.

The unique feature of the Tracer system, compared to other uses of GPS units in travel behavior research, is the use of wireless data technology to store travel data to a database as soon as it is recorded by the mobile units. This eliminates the need to physically interact with the units in the field to collect data, which in turn makes long, open-ended surveys possible. A small pilot study with the Tracer system has successfully tracked three EDCUs for over 100 days each without ever directly accessing the units to collect data. The ability to collect data for long durations without the need for analyst intervention (and the associated time and monetary costs) is an important step towards being able to field a full-scale survey designed around a long period of GPS data collection.

The most surprising feature of the EDCU, which may turn out to be its most important contribution to transportation engineering, is that this small, embedded device is a wire-

less web *server*. While the content is rather limited (collected GPS positions, speeds, and times), the unit is serving original content to the broader Internet. The usual paradigm for automotive telematics is that the wireless device in the vehicle is a client for some centralized product, be it advertising or route guidance information. But in our case, the EDCU almost exclusively performs as a server, not a client. This point forms the basis for the discussions of chapter 8.

Chapter 4

A web-based activity diary

4.1 Automated travel and activity surveys

This chapter discusses the steps required to study *activities* using the *travel* data collected automatically via the Tracer system: the GPS enabled EDCU device and the supporting software described in chapter 3. The next section reviews some recent applicable technological advances. Section 4.3 describes a new survey instrument which uses the Tracer GPS data to solicit activity information. The author is not aware of any other real-time implementation of a web-based activity survey that integrates streaming GPS data. Section 4.4 presents some preliminary implementation lessons gleaned from a very small pilot application. The pilot study has identified several areas which can and should be improved prior to a larger implementation of this survey system.

The results demonstrate that near real-time GPS data can be successfully integrated into an activity survey. However, it is not clear that this integration will result in the long-duration activity surveys as intended. Each survey volunteer soon stopped entering descrip-

tive activity data, while continuing to contribute long periods of GPS data collection. This indicates that the website interface needs to be even easier to use. This is especially telling in that the pilot study volunteers were close associates of the author.

More detailed analyses of the data collected with the pilot study are presented in chapters 5 through 7. These chapters also document the various analysis tools that have been incorporated into the survey web site. Another exciting future application is the potential for using the activity survey as a springboard for enabling peer-to-peer activity and travel information sharing. This is the topic of chapter 8.

4.2 Collecting activity data

As was stated in the introduction to this dissertation, activity models demand a much richer data set than that required for the traditional trip-based travel planning process. An excellent example is contained in the *ALBATROSS* book (Arentze and Timmermans, 2000a), which presents the model framework alongside an extensive discussion of the model's data needs, and the survey diary that they developed to collect that data. *ALBATROSS* intends to model every aspect of urban travel and activity explicitly, building upon a base of household activities. Therefore every household activity must be recorded, as well as every aspect of each activity, to the extent that such collection is practical. The trade-off is a reduction in the maximum duration of the survey. People aren't willing to fill out extensive time use diaries for long periods of time.

Much of the research in activity survey design is directed towards increasing the richness, accuracy, and duration of the collected data. Three technological enhancements to

travel surveys have been developed recently which all contribute to these goals. First is the computer aided survey instrument. Second is the GPS data collection device. Third is the ability to transmit data wirelessly, in near real-time, using common cellular phone technology. The contributions of each of these are discussed in the following paragraphs.

Computers accurately record and remember facts. This has been used to design surveys that can prompt a user to fix holes and inconsistencies in their responses. The labor saving aspects of using a computer in a survey have resulted in more questions and/or more detail being added to activity surveys, instead of attempting to simplify a survey to extend its duration (Kalfs and Saris, 1997). This is largely due to the goals of the data collection projects, which place a premium on learning about the details of activities their planning. Kalfs and Saris (1997) give a thorough accounting of the ways in which computer interview techniques improve upon pencil and paper diaries. The findings demonstrate that errors decline in the computer aided surveys, but that more can be done to reduce under-reported trips.

Two recent computer-aided activity surveys are described next. CHASE (Doherty and Miller, 2000) studies the scheduling process for one week of activities. Respondents were asked on Sunday evening to enter their planned activities for the week into CHASE. Every night during the week, they were asked to update their actual activities, and fill in any new planned activities. The program stored each action the respondent took, enabling an analysis of the household planning process for the week. REACT! (Lee and M^{c}Nally, 2001), a second-generation version of the CHASE approach, preserves most of the good features of CHASE while adding several enhancements. Most notably, a point-and-click mapping capability was added to aid in geolocating activities, and steps were taken to

reduce the opportunity for respondents bias the results by using the survey instrument as a personal scheduling tool.

The goal of both REACT! and CHASE is to gather data that will enable the estimation of a computational process model of how people choose and schedule their activities. Doherty and Axhausen (1998) describe an activity modeling framework, similar to that proposed by Bhat and Koppelman (1993), which requires a detailed model of the activity planning and scheduling process in order to turn an activity agenda into an activity program. The data collected from REACT! and CHASE would be used to calibrate the scheduling module in this framework. The scheduling module ideally would mimic how real people go about scheduling activities.

Measuring and modeling the mental processes underlying human behavior is a difficult problem. A different approach, advocated here, is to streamline the measurement of revealed behavior so that longer durations of activities can be observed and examined for spatial and temporal patterns. Rather than modeling thought processes, one can model observed point processes and activity sequencing patterns. This requires a very simple subset of activity data—just the name, the place, the start time, and the duration of the activity. Computer enhancements to such a survey should be designed to simplify the survey instrument and minimize respondent burden, so as to lengthen the survey period.

The second important technological advance is the GPS data logger (Battelle, 1997; Murakami and Wagner, 1999). Kalfs and Saris (1997) conclude their review of computer aided survey capabilities by recommending that more be done to automate the data collection process. They identify GPS technologies as a (then) new and promising way to automate time and position recording of trips, eliminating many recall errors and time in-

accuracies. Similarly, many of the theoretical problems with surveys and diaries (discussed in detail in Timmermans (2000)) revolve around eliminating missing trips. CHASE and REACT! try to use the respondent's reported durations and start times to check the completeness of the response, but this approach still relies upon a possibly inaccurate recollection of time, and the associated respondent burden. Incorporating a GPS device into the activity survey can solve most of these problems. Section 3.2 provides some citations of recent projects that have studies the use of GPS devices and data in a travel survey context.

The approach of Bachu et al. (2001), which only collects GPS measurements, followed later by an activity survey, is very close to the approach taken here. The difference is that the Tracer system takes advantage of the third technological innovation, wireless data communication. A GPS device by itself cannot collect activity data. Its output must be combined with an activity diary in some way, and the incorporation should not cause more problems than it solves. Manual integration steps, such as requiring the respondent to plug the data collection unit into a computer, or using some removable media to transfer data back and forth, are likely to be more burdensome and error prone than a pencil and paper diary. Handheld computers are an interesting option, but are still somewhat expensive and difficult to use, and still require the physical collection of the unit to gather the data. Other options such as mailing in the survey device for processing or sending a trained technician with a laptop for a face to face interview are very expensive.

The design and implementation of the Tracer system is somewhat unique in the activity survey world in that it enables the integration of GPS data by making use of a cellular digital packet data (CDPD) technology. CDPD modems operate nearly everywhere a cell phone does, and the cost of such service, while perhaps too high for the average person, is

within reach of a survey budget at $30 to $50 per month per unit.

An early report on the design of the Tracer system was criticized by an anonymous reviewer for using CDPD technology, rather than something faster and more advanced. CDPD is indeed a slow technology, barely able to achieve a sustained transfer rate of 10kbps. On the other hand, CDPD is an available solution and usable now in many major cities. At the time we made the decision, Metricom had a faster technology, but it would not be available for another few months in Orange County, CA. In hindsight, the decision to go with the slow but available CDPD service was the right thing to do, as Metricom has subsequently gone out of business.

That being said, we are wireless technology agnostics. There is nothing special about CDPD, and there are many disadvantages, such as the price of service and the slow speeds. We are actively researching a wireless ethernet 802.11b solution, but that would not provide the nearly uniform coverage of CDPD. The CDPD wireless link allows the transmission of GPS records every few seconds to a central database. The ability to install a device, turn it on, and forget about it until the end of the survey period is quite appealing. Any replacement technology should have the same feature.

4.3 Web-based, EDCU-enhanced activity survey

4.3.1 Activity survey goals

The preceding section described some thoughts behind the design of this web-based survey. The purpose of the survey is to gather activity information about destinations,

assuming the location and time of the destinations are already known. The primary design criterion for the survey instrument is to be easy on the user. That coupled with the fact that the location and time are already known means that many of the detailed questions that are common in time use and activity surveys are not asked here. Instead, the user-friendly focus requires that all unnecessary questions be stripped away. The minimum data that are of interest are the activity at the destination, and comments on the route to the destination.

4.3.2 Pilot survey design

The discussion in section 2.6 observed that what people do may not fall into a neat description of behavior that fits into the categories of a static survey form. In order to gather evidence primarily to examine the activity-location pairing conjecture, it is perfectly reasonable to let the respondents themselves describe their activities in their own words. This is easily done on a web-based form with free text entries. The unique entries are stored in a database, and used to dynamically generate a survey which, over time, will become tailored more and more to the past entries of each respondent.

The first generation web survey displayed a map of the travel in question, followed by a blank text block. The respondent could type whatever he or she wished to properly describe the activity and the destination. In addition, another text block was included to elicit comments on and description of the travel to the destination, as shown on the map. The survey itself was composed of simple HTML elements such as an image tag for the map, and two form tags for the text entries, plus a submit button. The HTML was stored as a template in the database. When a user accessed the system and selected a particular

period of travel, a Perl program fetched the proper data from the database, generated the map, and completed the HTML template to generate a dynamic survey. Each separate trip and destination in the travel period were displayed one after the other.

This simple design was tested internally. Several deficiencies came to light over time. First, the map generator was unacceptably primitive, and had to be significantly expanded. Second, displaying the travel plus destination alone were not enough of a memory jogger. Instead, it quickly became apparent that other information should be provided, such as the day and date, the time, the average travel speed, and so on. The map did not provide enough prompting because it did not indicate travel directionality, and without any time information it was often hard to tell which end was the activity destination. To solve this, an arrow is now rendered at the destination to make it easy to spot.

The third deficiency related to the entry of travel and activity annotations. The form itself and the free text aspect solicited a lot of descriptive information. However, there was too little organization to the responses, making post processing nearly impossible even within a single individual. Furthermore, repetitively typing the same thing every day for the same activity is not a user-friendly interface. Some further refinement of annotation information was needed, as well as a way to offer past responses as options for new activities.

The internal review of the first generation survey resulted in minor changes to the survey form itself, but required some rather extensive changes to the back end processing. The revised survey and the internals required to make it happen are described in the following sections.

4.3.3 Displaying a map

As the system was expanded in response to initial comments and with an eye to displaying travel from many different trips on one map, there was a need to formulate algorithms for aggregating the information from possibly overlapping trips. As solutions to this problem were tested, it became obvious that aggregation of travel data for one person across several trips was more or less the same problem as aggregating information across people. Thus the simple web survey took on more and more of the back-end functions one might expect to find in a GIS system, albeit with a static, non-panning, non-zooming image.

The result is a lightweight, object oriented map generator, written in Perl. The core classes represent the map image to be displayed, the travel to the destination, and the activity at the destination. A day might produce several trips, each represented by a unique `traceid` database key. The user selects the day or days to be viewed, and the program fetches the corresponding list of `traceids` from the database. Each `traceid` is used to create an instance of the trace image class. These objects grab their GPS points from the database, compute averages, durations, and other data-driven annotations, and of course generate a plot of the trip layered on top of an appropriate map of the travel area. The first time a trace object is created, it obviously must compute its state based on the raw data, and generate an image. This state information is saved to the database, so that subsequent object instantiation and web page rendering are faster.

The plots are static images, but based on the preliminary trials, the image can be drawn at three scales. The shortest trips are plotted on a neighborhood scale, in which features like individual blocks are readily apparent. Medium length trips are plotted on a scale at

which local streets appear to be a dense mesh, while any trips longer than the extent of the medium map are drawn on an enormous map scale which can fit most of the Los Angeles metropolitan area. Maps are rendered as mosaics when necessary, so that mid-length trips can span multiple maps at the more detailed scales, rather than occupying a few inches in the larger scales. A survey screen-shot, with a map drawn at the neighborhood scale, is shown in figure 4.1. A map rendered at the mid-level scale and using the mosaic feature is shown in figure 4.2.

Each trace image object can also be instantiated within an image composite class. The image composite object uses its component trace image objects to generate a map of all of the travel during the period being displayed. Since each trace image object knows its own data points, as well as the background maps it requires at each of the three scales, a composite rendering is as simple as first getting a list of all of the background maps from each trace, forming the combined map image, and then instructing each trace object to render itself on that image. This can produce very large graphics, with the largest image so far being over half a megabyte and spanning several screens of a web browser's page. The maps can be quite interesting, as any number of EDCU data tracings over any time period can be displayed at once for the survey operator. A 50% scaled version of a composite map is shown in figure 4.2. Note that the black and white rendering makes it difficult to see the several traces on the map. The website displays color travel traces on top of grayscale maps, so the images are very easy to read.

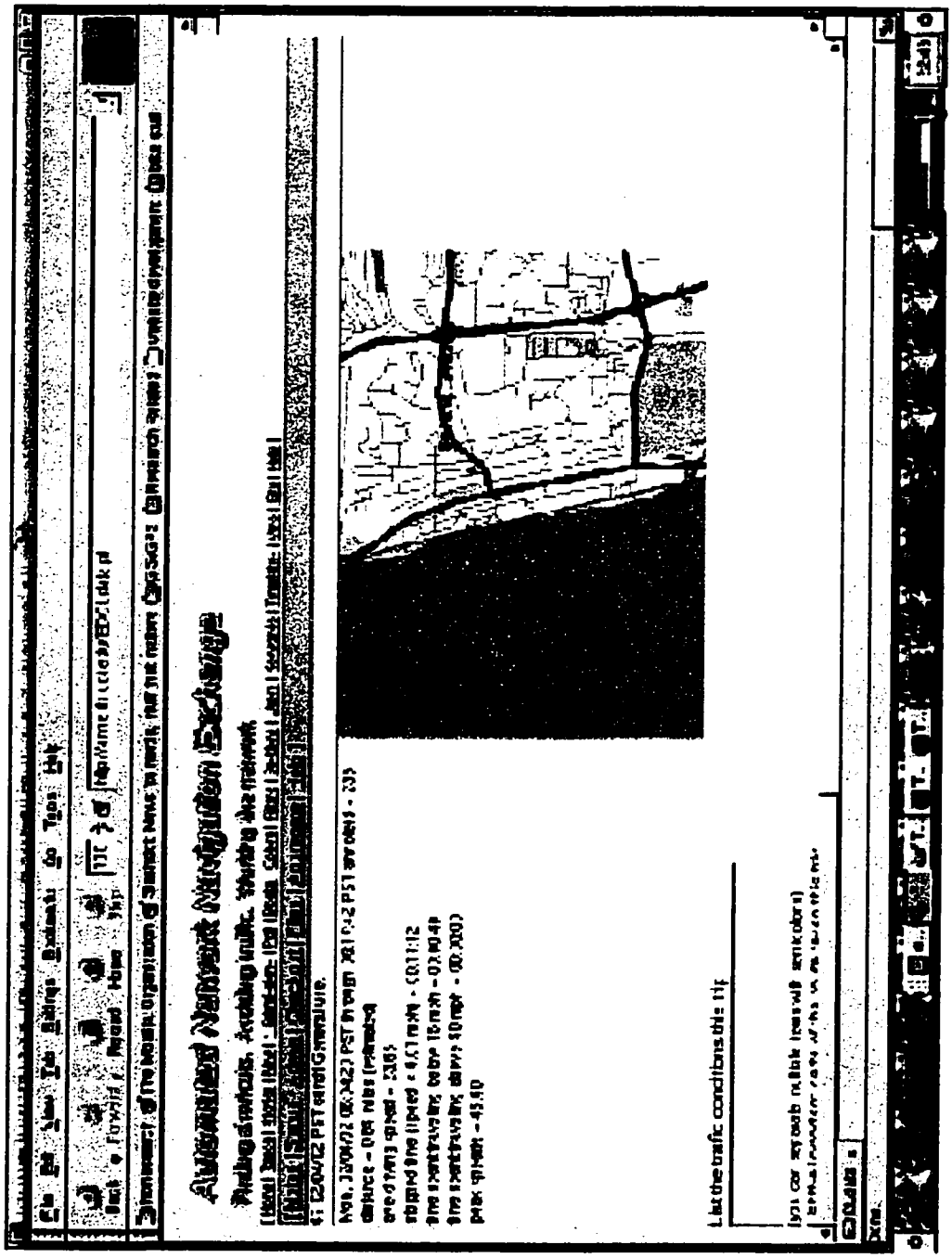


Figure 4.1: The initial screen of a survey. Other trips are visible by scrolling down. The web version renders the travel path in red, yellow, or green, depending upon the travel speed, on a grayscale map.

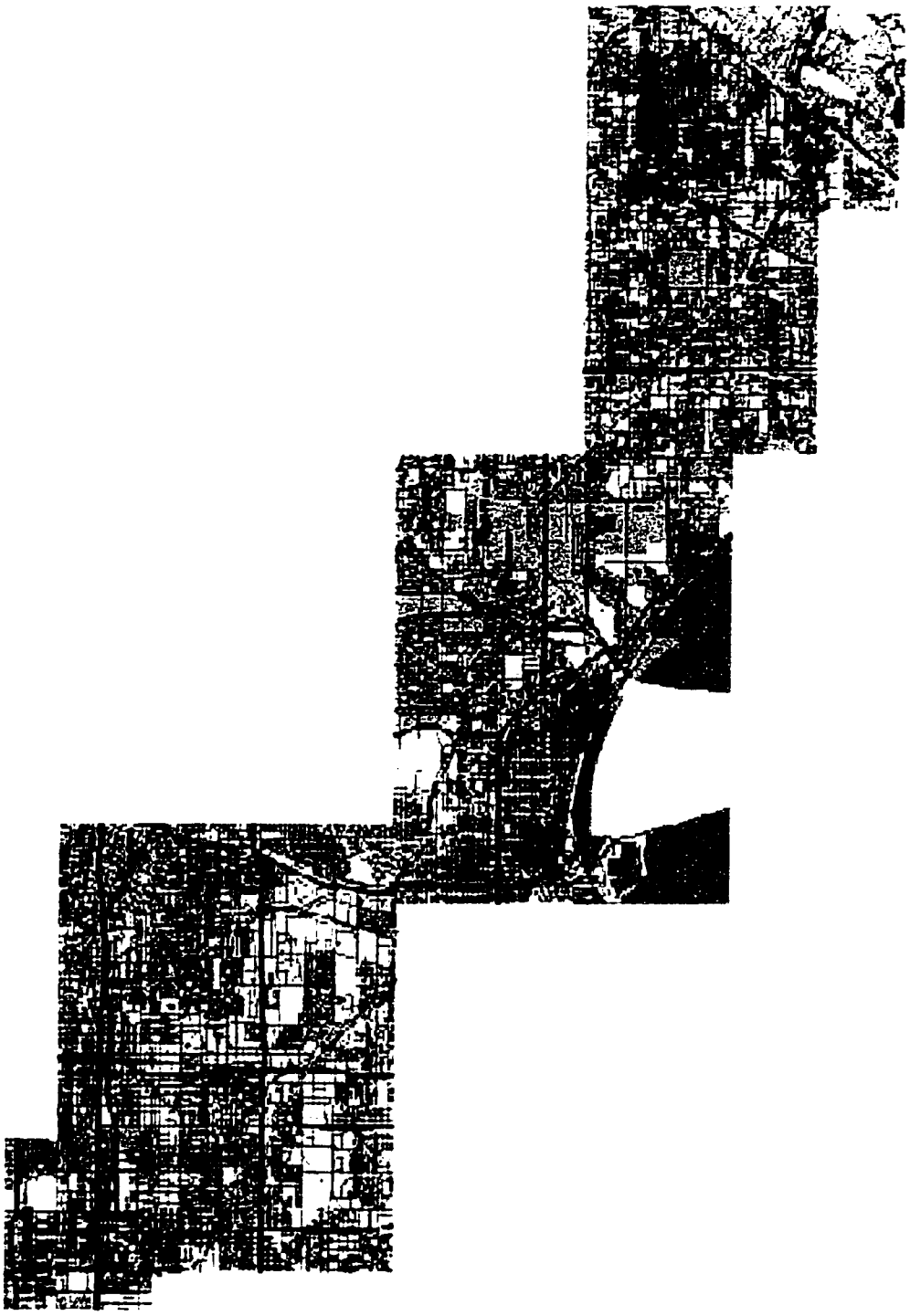


Figure 4.2: An example of a composite image of a day's travel. This image is one-half the size of the image on the website. The real map shows a color trace on a grayscale map.

4.3.4 Travel and activity annotations

The second generation survey contains text forms for the travel conditions, the route, the destination name, the activity at the destination, and the people involved in the activity. A screenshot sequence similar that which would be seen by a first-time user is shown in figure 4.1 and figure 4.3. None of the questions are required in any way. The user can skip some or all of the forms for some activities and fill out others, and the web server will not attempt to inquire about the non-responses.

The HTML page is generated piece by piece for each trip being shown. First the `traceid` is used to create a trace object, which is used to generate a map of the travel as explained above. Other classes are instantiated as needed to fetch the numerical information about the trip, and any past text entries relating to the traffic conditions, the trip route, the activity, the destination, and the people involved. These components are stored in a large list, and then passed to a rendering program which fills out a blank HTML template with the information. The templatization allows the user interface and the general layout of the survey to change quite easily without needing to modify any of the underlying programs. Further, most of the sub-components of the web page themselves use templates in their creation.

Figure 4.1 is a screenshot of the first page of the website, at 50% reduction from the original. To the left of the map is the numerical summary, dates and times that characterize the travel. Below the map and its annotations are the survey questions, shown in figure 4.3. This pattern is repeated for each separate trace.

When information is entered into the forms, the response is saved to the database and

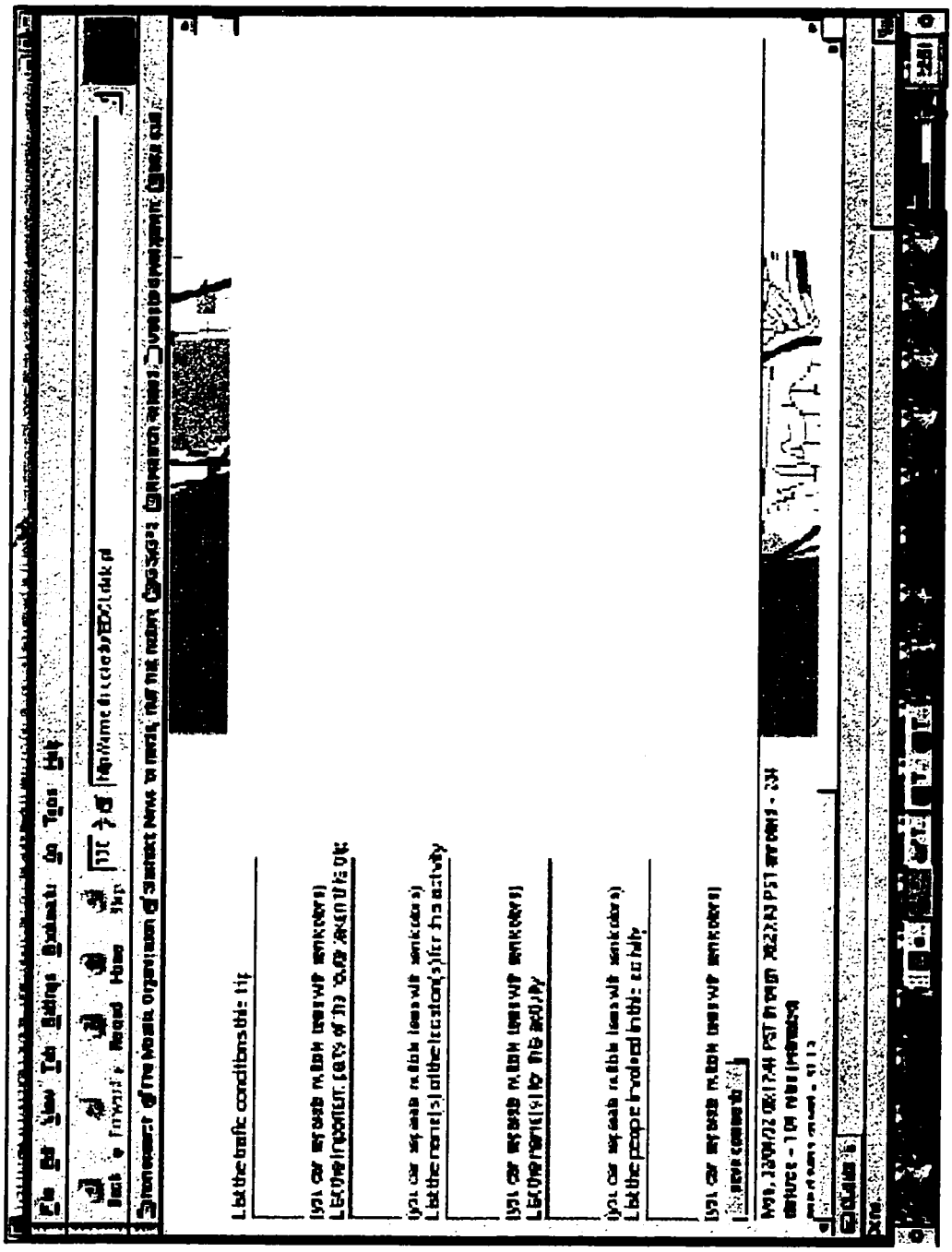


Figure 4.3: Text forms for the travel conditions, the route, the destination name, the activity at the destination, and the people involved in the activity. This user has not yet entered any information whatsoever.

used to create checkboxes for future entries. So for example, if the user staring at figure 4.3 entered "my house" for the destination, then the next time the survey is generated it will display a checkbox option for "my house" to the right of the location text box. Over time, as more and more locations are visited, the list of checkboxes can get quite large. To aid in finding an alternative, the options are sorted based on their closest previously observed distance to the destination in question. An example of a survey form after several weeks of data entry is shown in figure 4.4. The success of the sorting option reflects the repeated nature of activities, with simple knowledge of the time and place of the destination producing a reasonable approximation of what the respondent is most likely to enter. This is discussed in further detail in chapter 6.

It should also be noted that the respondent can check or enter as many options as apply. Although the open-ended text block has an unlimited potential length, after a few uses the respondent can figure out a balance between repeated items and the length of the entry. For example, instead of typing "went home and had dinner", which creates a single checkbox entry for an activity, the respondent can type "went home; had dinner" which will create two checkboxes for future use. By prompting the respondent with his or her own entries, it was hoped that accurate data could be collected about the number of different activities performed at each location, rather than the number of predefined activity categories the respondent felt related to his or her actions.

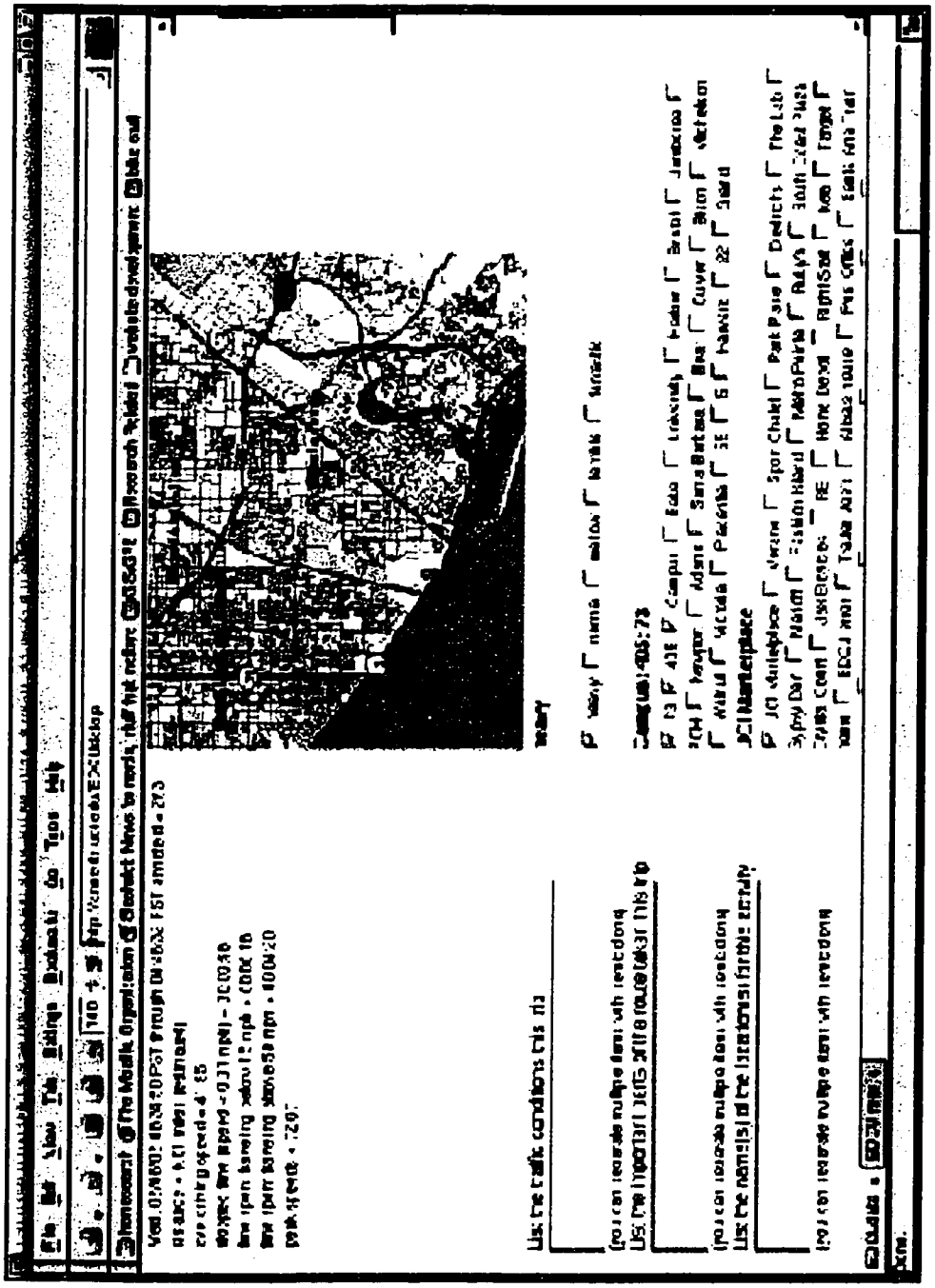


Figure 4.4: The survey dialog with several weeks of travel options entered and available as checkboxes. The web version is in color.

4.3.5 Aggregating information from many trips

Aside from the composite image generating class, the side goal of presenting aggregated travel information to the survey respondent is largely unmet by the current implementation of the survey. The beginnings of the data aggregation pages are reached by clicking the link called "Aggregation" at the top of the page. This page presents the respondent with a potentially huge list of all of the options and text entries he or she has made. A composite map can be created by building queries using the same checkbox data the user entered in the survey form. For example, one might select all travel to or from "home" with a "meal" activity name between certain dates. The compositing class uses all of the traces that qualify to generate a single image. This serves no purpose for the activity survey, but it does allow the respondent a useful tool for viewing his or her own activities and travel. The potential exists for a much more thorough and useful aggregation function. This future work is discussed in chapter 8.

4.4 Survey field test and lessons learned

The Tracer system has been installed without problems in several different kinds of vehicles. The only requirement is that the vehicle have a working cigarette lighter to provide the unit power. If the cigarette lighter does not shut down when the vehicle ignition is turned to off, the EDCU should automatically shut itself down after approximately 30 minutes. According to the EDCU vendor, with normal lead acid batteries, the EDCU should never cause a battery to drain within that 30 minute period.

Linking an on-line survey with the Tracer data requires only that the survey administra-

tor activate a user id (uid) for the prospective survey participant. As a security precaution, only authorized users can access the portion of the web site that displays Tracer data. Further, a user will only see his or her own travel when accessing this part of the website, unless the user has survey administrator security permissions.

Most EDCU/Tracer installations to date have been for demonstration purposes and did not include any attempt to collect activity data as well. Four people volunteered to test the Tracer and on-line survey over an extended period of time. These volunteers placed an EDCU in one of their vehicles, and were asked to annotate their activities and destinations. Invariably the respondents stopped annotating their activities after a short period of time, but they all continued to use the EDCUs in their vehicles for much longer periods. The main result of the pilot survey was to gather important information on what can be improved in the on-line survey and the Tracer system. In addition, the different numbers of annotated trips combined with the different trip patterns from each volunteer offers an opportunity to test some data analysis tools. Chapter 5 examines modeling the GPS recorded destinations as a spatial point process. If at least 15 repeated observations of a single activity are collected, then this same analysis can be performed for the particular activity. Chapter 6 tests a method for estimating the spatial frequency of activities in space. This model offers a way to assign the most likely activity to a particular unlabeled destination, based only on the recorded coordinates. The more activity notations a respondent enters, the more this will help the categorization of future, unlabeled destinations. Finally, chapter 7 applies sequence alignment techniques (Gusfield, 1997) in an attempt to find self-similar patterns within each volunteer's recorded activities.

The remainder of this section focuses on the lessons learned regarding the mechanics

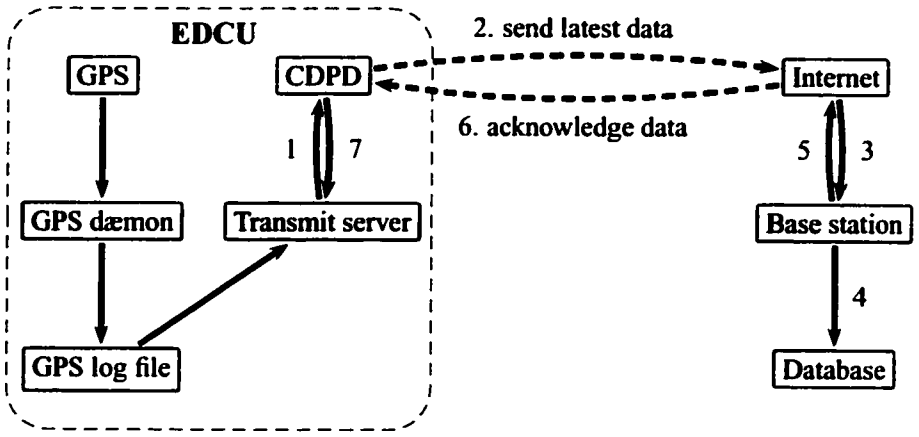


Figure 4.5: Flow of information from the GPS to the database. If the CDPD link fails to deliver message 2 or message 6, then the transmit server does not receive acknowledgment that the data has been stored to the base station database.

of the data collection process. The Tracer system produced errors, both missing real stops and generating false stops. In addition, the activity survey turned out to be much more burdensome than anticipated. The following discussion documents what appeared to work, and what needs improvement, based on the results of the pilot survey.

4.4.1 Detecting a stop

The operational logic of the wireless EDCU data collection process is shown in figure 4.5. There are two sources of errors with the way this process is currently implemented. First, the current implementation of the transmit server is a traditional serial program flow. Referring to the numbered messages in figure 4.5, the transmit server sends message 2 and then waits to receive message 6. If the CDPD link fails for message 2 or 6, then message 6 will not be received by the transmit server, and it will block indefinitely. In practice, the transmit server is restarted every 5 minutes, so the effect of a CDPD transmission drop is to cause a 5 minute gap in the collected data.

Second, the base station has no way of knowing whether or not a mobile unit is powered up and collecting data or stopped. Rather than having the base station continuously probe all of the mobile units (which is only possible if the base station knows the IP addresses of the units), the mobile units were instead programmed to send a "shut down" message to the base station when the units turned off. This strategy turned out to be unsuccessful, as the battery backup uninterruptible power supply (UPS) often did not have enough power to gracefully shut down. Another problem with this strategy was that about half of the various test vehicles did not power down their cigarette lighters when the ignition was switched off. Thus the EDCUs did not power down at all, but instead continued to transmit (stationary) points. The base station was reprogrammed to attempt to automatically detect a stop. The algorithm was only partially successful, with a bias towards adding too many stops rather than too few.

The cumulative result of these two sources of error is that the collection of destinations is much larger than the actual number of destinations. Further, due to the gaps in transmission, sections of travel between stops are often missing. Finally, if the CDPD connection is weak and/or a trip takes less than a few minutes, the system may entirely miss a trip and a stop.

All of these problems have fairly easy solutions. The solution to the transmit server problem is to fork the process with each transmission, so that the parent continues to monitor the log file and spawn new transmission processes. A hiccup in the wireless link will stall only the child process, not the parent or the other transmit processes. After a certain time, the persistent child processes can be rounded up by the transmit server dæmon into a new process that will combine and resend their messages. This solution has not been im-

plemented yet on the EDCU, but a similar forking scheme was devised for the base station XML-RPC dæmon to allow multiple simultaneous EDCU connections.

The algorithm to detect stops on the server side has been improved, as part of the effort to analyze the data. The new algorithm checks for time periods of travel data that have zero speed. The current algorithm also attempted to do this, but the implementation seemed to fail for technical reasons relating to how the persistent data was being stored in the database. Second, the new algorithm examines gaps in travel data for spatial continuity before inserting a stop. If the power down and power up points are separated by an appreciable distance, and if the time difference implies a reasonable speed between those points, then it is assumed that the vehicle has not stopped, but rather simply had a transmission gap for some reason. When the data transmission server is improved on the EDCUs, the need for this check should diminish.

4.4.2 User interface

The user interface for the on-line survey has both strengths and weaknesses. The interface is easy to use at first, and presents information clearly. However, aside from very short trips, the automatically generated maps are usually mosaics of two or more maps. This means that they extend beyond the design map size of 400×300 pixels. While the maps are quite clear and easy to read, the respondent must scroll down and right in order to see them, and then scroll back left to enter the survey information. This may cause confusion for users who are unaccustomed to information that extends beyond the boundaries of a browser window. The solution to this problem is not obvious, as it is endemic throughout

the Internet whenever a web page does not fit easily within the lowest common denominator 640×480 pixel screen. Frames are often used to address this problem, but they bring their own cross-browser compatibility issues.

A more serious problem is the potentially infinite growth of the checkbox lists of past entries. As more and more entries are recorded, the lists begin to dominate the survey form, and become more of a hindrance than an aid. The obvious solution is to limit the numbers of checkboxes that are shown, with a bit of Javascript embedded to enable an option to see more choices if desired. The current option sorting algorithm, based on distance and time, is reasonably good at putting the most likely options at the top of the list, but it has some problems and could be made much more reliable. Intelligently limiting the options is a topic that is covered in chapter 6.

Chapter 5

An analysis of activity destinations as random point processes

5.1 Are activities and destinations linked?

One of the goals of the activity based approach to travel analysis is the ability to model travel entirely as a demand derived from the need to perform activities. This requires a comprehensive model of activity generation and scheduling, combined with simultaneous models of the impacts of the environment upon the activity choices of the actor. For the moment, this goal is still out of reach. This chapter addresses the relationships between activities and the destinations at which activities are performed.

Recent practical applications of activity based concepts to travel modeling assign an individual or household a fixed agenda of activities, and then simulate the attempted execution of those activities. The *ALBATROSS* system, as described in section 2.4, simulates the execution of activities by assigning heuristic scripts to each actor (Arentze and Timmer-

mans, 2000b). McNally (1997) presents a framework in which activities are assigned locations based on land use patterns, with the constraint that the travel distance must agree with the constraints of the particular sojourn's representative activity pattern (RAP). Vaughn et al. (1997) sample a sequence of activities from a pool of observed patterns. The original destinations are discarded and replaced with new destinations that correspond to the simulated household's simulated environment. These approaches and many others all presume that each destination choice for an activity is independent of past choices. While some exceptions might be made for situations such as work or home locations, for any activity the selection of the destination essentially comes down to independent random selections.

Within an aggregate model, it is perhaps good enough to presume that each destination choice is independent of previous choices. Often these aggregate models are only used to generate travel patterns for a single, isolated, representative day. But many activity based models are microsimulation models, in which discrete individuals are modeled over time. Further, activity based methods claim to be able to examine the impacts of policy decisions such as encouraging carpooling or promoting telecommuting. In such cases, the presumption of spatial randomness for activity destinations has not been verified. While there has been some discussion in the literature about destinations that can serve multiple activities (Cullen and Godson, 1974; Damm, 1982) there has been little discussion about the more general question of whether activities and destinations are strongly or weakly linked.

This chapter does not solve this problem. However, it does provide preliminary evidence that the destinations an individual chooses for his or her activities are strongly clustered in space. The Tracer/ANNE pilot study documented in chapter 4 collected data from

four volunteers for over one month each. While the respondents quickly tired of completing the activity survey, they all voluntarily kept the data collection units in their vehicles for extended periods of time. The actual durations were 35 days, 70 days, over 100 days, and over 140 days. This should be interpreted as a critique of the on-line survey, not as an indication of the expected duration of a GPS-based survey.

While one might not be able to collect this long a period of GPS data in a real survey, the extended period of data collection gives some weight to an examination of the spatial variations of the destinations. That is, one can be reasonably confident that this period has captured a significant proportion of all of the possible travel behavior within each individual. The next section examines all destinations for each of the survey participants. Section 5.3 then examines the spatial dispersion of the destinations associated with particular activities for those few activities which were entered often enough via the on-line survey.

5.2 Analysis of destinations

The field of spatial analysis has existed for several decades. As with most computationally intensive disciplines, the recent explosion of desktop computing power has spurred an active development of new statistical spatial analysis techniques. Within the R statistical language and environment (Ihaka and Gentleman, 1996), there are at least eight different spatial statistics packages that are useful and under active maintenance, as well as a dedicated project to integrate R packages into the GRASS GIS (Baylor University, 1999). The following analysis makes extensive use of the R packages `spatstat` and `spatial`.

An important question when examining a point process is whether the process exhibits spatial order, randomness, or clustering. From a simple scatter plot, such as figure 5.1, the destinations appear to be clustered in the lower right corner of the region. But it is hard to say if within that clump of points the destinations are randomly distributed or further clustered.

Complete spatial randomness is defined as a homogeneous Poisson process with some intensity λ (Cressie, 1993; Ripley, 1981, 1988). Spatial order is defined as a regular pattern, such as a grid of dots. Spatial clustering is defined as a process which has a density of events that is much higher than one would expect from a Poisson process operating at the same length scales. One important way to establish clustering and order relative to the baseline Poisson process is to estimate Ripley's K function, which approximates the second moment of the spatial random process. This function is defined by Ripley (1988) as:

$$\lambda K(h) = \mathrm{E}[\text{number of points with distance } \leq h | \text{ point at } \mathbf{x}]. \tag{5.1}$$

A Poisson process (complete spatial randomness) should produce K values that are proportional to πh^2 for any h. Processes with K functions that are significantly greater than this theoretical measure indicate a higher degree of clustering, while processes with lower K functions are characterized by regular dispersion and order. The results are typically shown graphically—the Poisson point process lies along the sloping parabola in the K versus h plot, and along a diagonal if K is divided by π and square-rooted.

Figures 5.2 through 5.5 plot the empirical estimate of K versus the theoretical Poisson process curve for each of the survey participants. All plots indicate that destinations are clustered for as long as the estimate of K is reliable. The sharp increase in K at very short

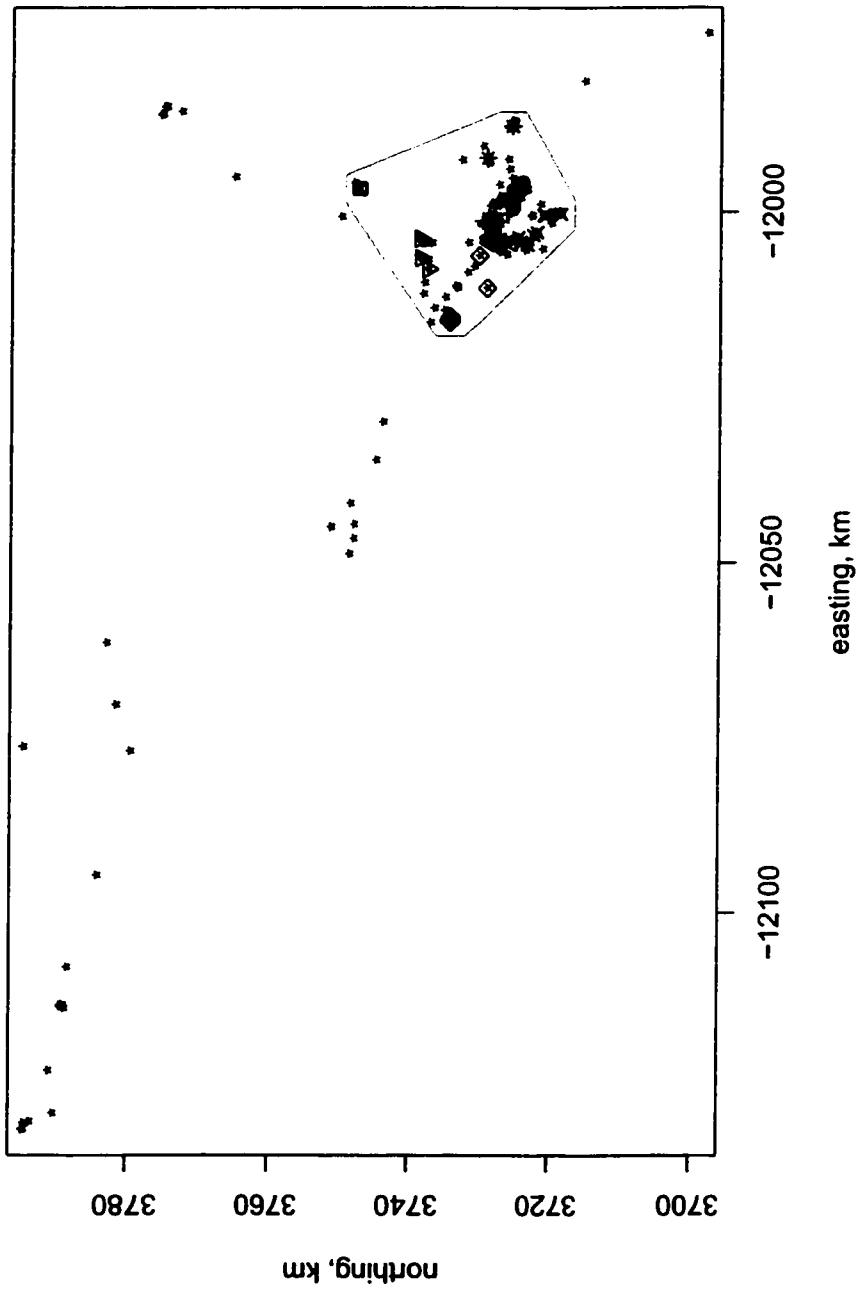


Figure 5.1: All destinations recorded for user id 2. The small stars are unnamed activities, while the larger symbols indicate the location of named activities. The boundary box shown encloses all named activities.

distances is most likely the result of repeated visits to approximately similar destinations. There is a natural scattering of GPS points at each destination, due to random errors and varying parking spots, but even restricting the space of the analysis to areas immediately surrounding the destinations and examining the the point processes at distances under 500 meters did not indicate that the point process was random. This means that for short distances, the high K function curves can be attributed almost entirely to the high density of destinations within a neighborhood of other destinations, and that the randomization effects in the system are outweighed by the clustering effects even at very short distances. The most likely explanation for this is that the final destination points are clustered along the streets leading up to the destinations, or otherwise approached their parking spots from the same direction.

The K functions tend to level off a bit prior becoming erratic at the limit of their estimation bound, but they generally remain well away from the Poisson curve. Note that K should never be a decreasing function, although its estimator might become unstable as h approaches the dimensions of the bounding region. It is possible that continued measurement would show a natural limit to K. This would be the result of both repeated visits to locations, and the limit aspect of the maximum distance it is possible to travel in a day. One should definitely expect that as the distance rises (and as the bounding region increases), the number of new observations at greater distances should taper off to zero. This is because the distance on the horizontal axis of the K plots represents the distance between any two randomly chosen events. A highly ordered or a random activity pattern would need to maintain its rate of spatial regularity or scattering at greater and greater distances. The practicality of this drops to zero for any real activity process, because of the growing

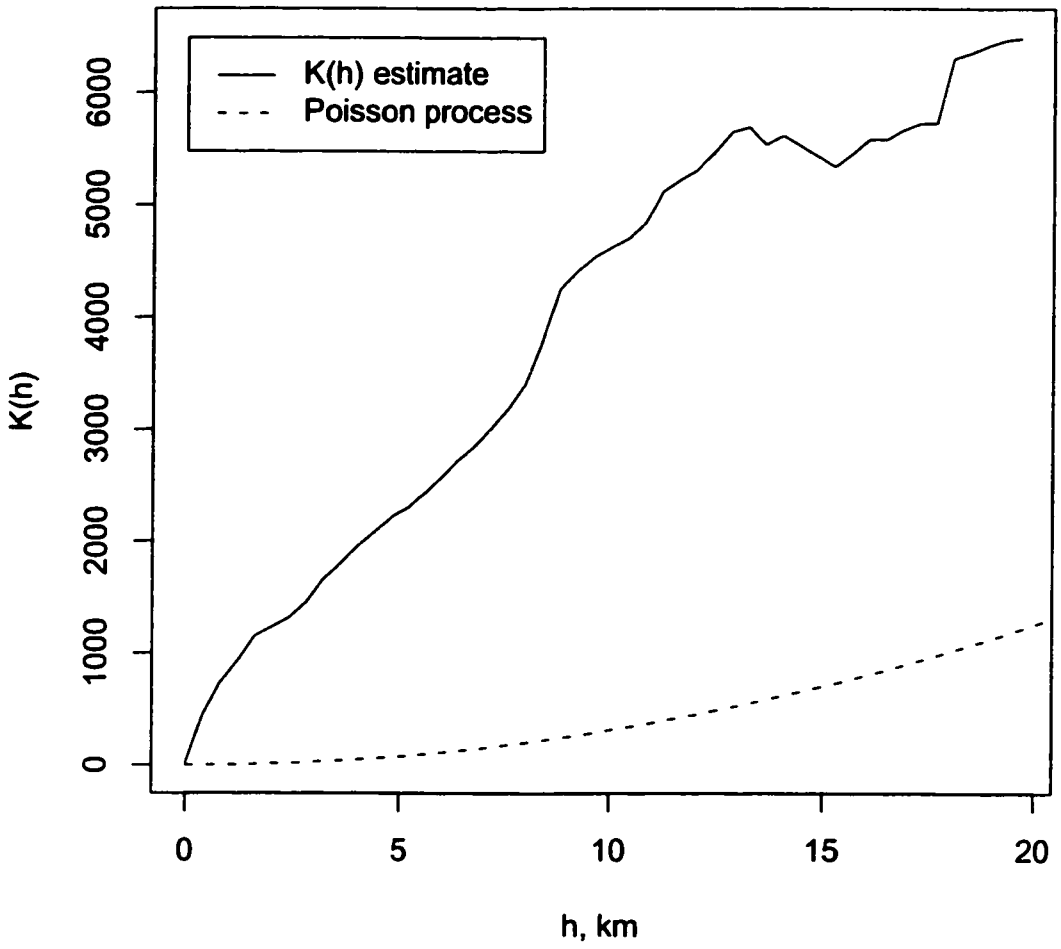


Figure 5.2: Estimated K function for all user id 2 destinations

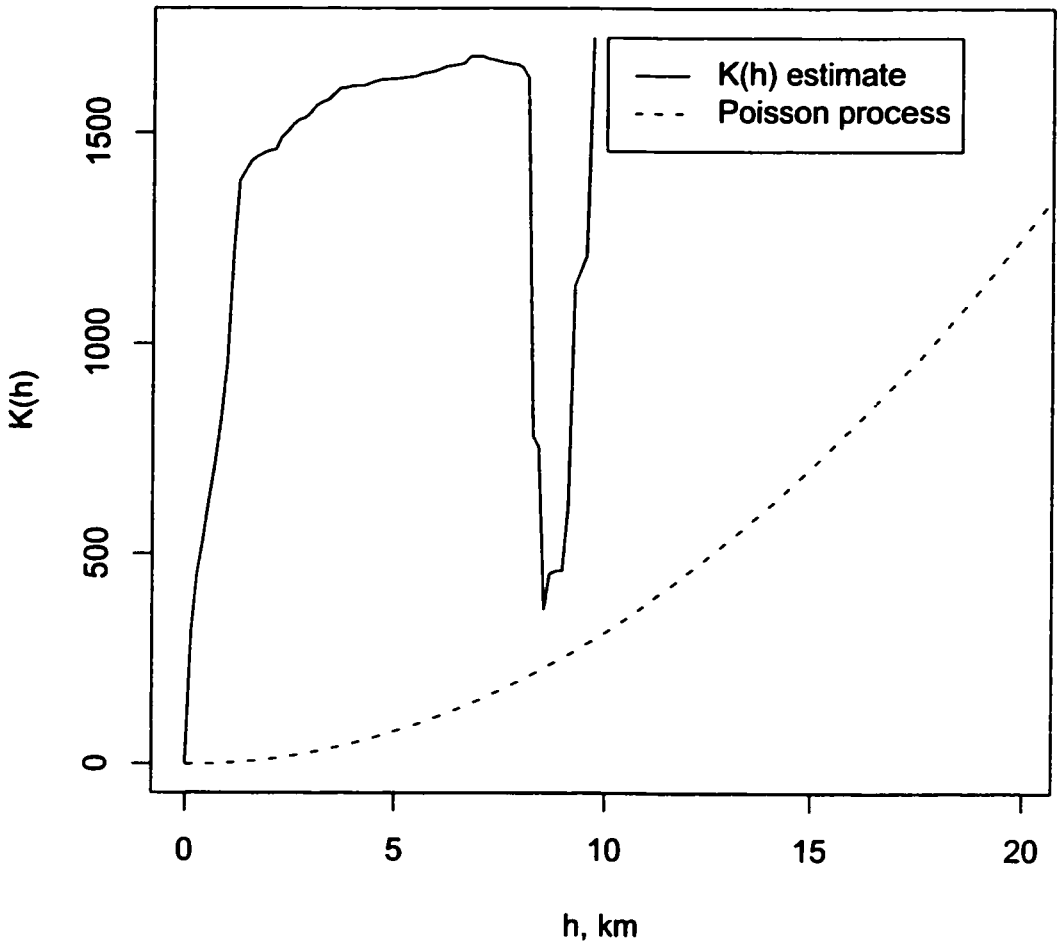


Figure 5.3: Estimated K function for all user id 4 destinations

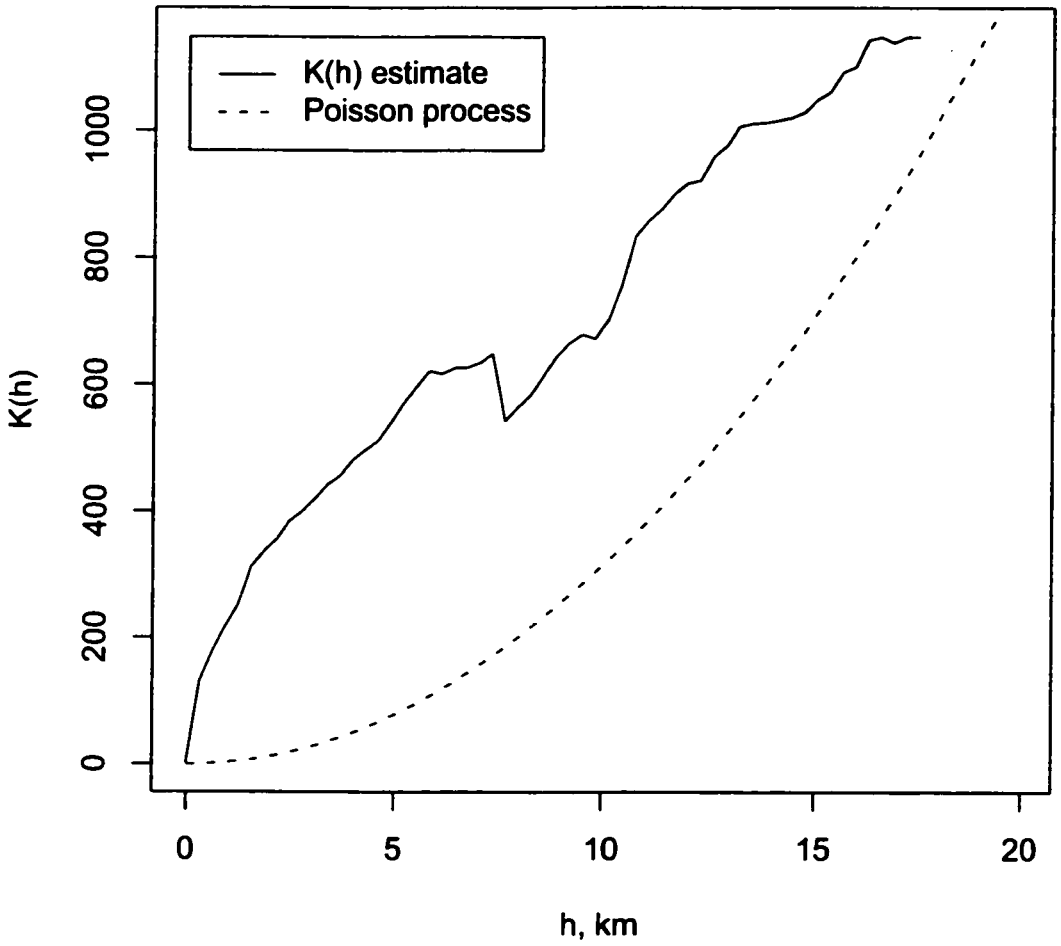


Figure 5.4: Estimated K function for all user id 5 destinations

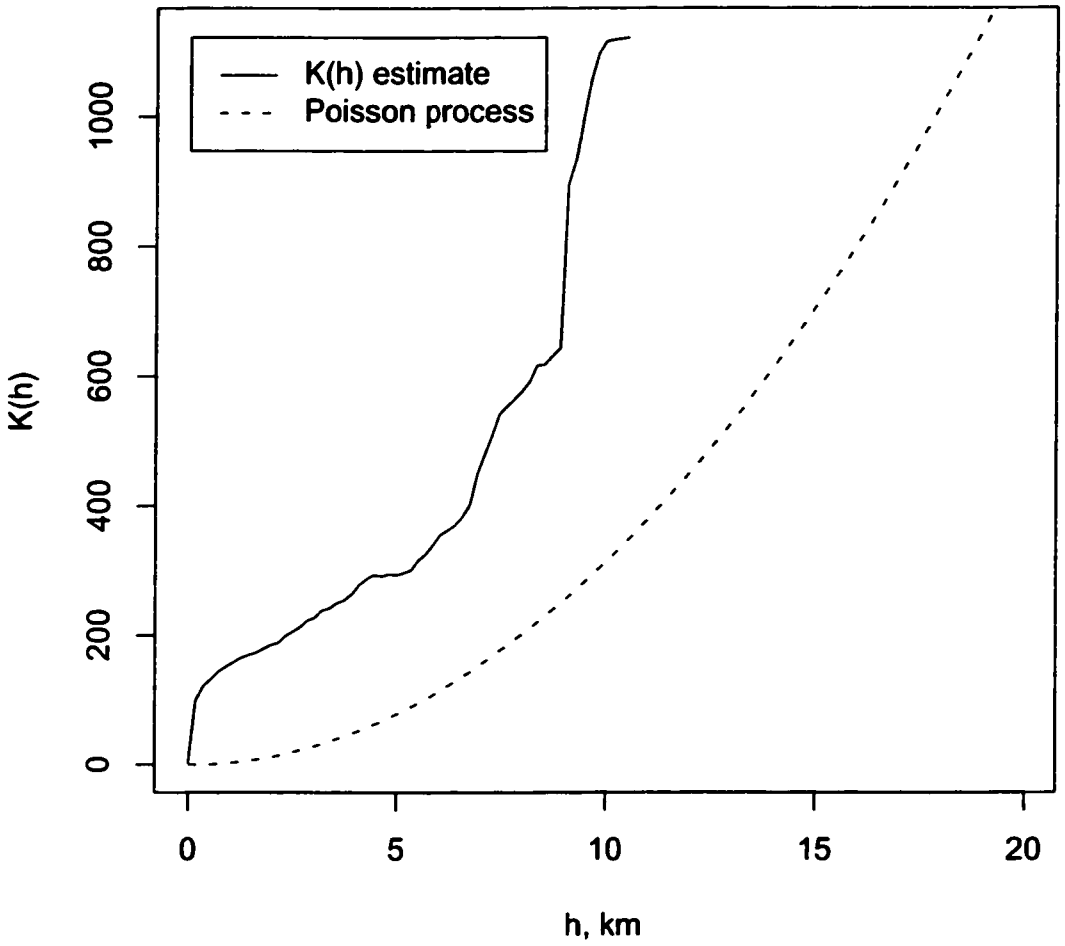


Figure 5.5: Estimated K function for all user id 6 destinations

(infinite) travel time required to visit all of the points.

Figures 5.2 through 5.5 depict empirical K functions which have been computed over an area defined by a convex hull bounding all observed points, using a 2 km buffer in most cases. Alternatively one could use the rectangle surrounding the points, but that would add large empty areas. Further, using the convex hull pushes the empirical K curve somewhat closer to the Poisson curve, but the location of destinations still exhibited strong clustering tendencies.

In an attempt to further refine the areas, a series of spatial clustering runs were performed, attempting to group together isolated clusters of points which could then be analyzed for clustering independently. Distance based clustering did not produce anything useful, however. An example of a clustering of all of the named activity points for user id 6 is shown in figure 5.6. While one might be able to iteratively determine the optimal clusters to use, the goal of this chapter is to develop *automatic* analysis tools which do not need to be adjusted by hand. For want of something more sophisticated, the convex polygon surrounding the named points was used to form a tighter region for testing clustering effects. Including unnamed points within this region theoretically should boost the degree of randomness, if indeed more points translate into increased spatial randomness.

The responses of user id 6 were processed to demonstrate the effect of the differently sized regions on the estimate of spatial randomness. User id 6 was chosen because the initial plot in figure 5.5 was the closest to the Poisson curve. Figure 5.7 reproduces figure 5.5 in the upper right corner, and compares it to three other empirical estimates of K. The upper left curve was estimated using all of the points within an area defined by a bounding rectangle. The lower left curve was estimated from all points, named and unnamed,

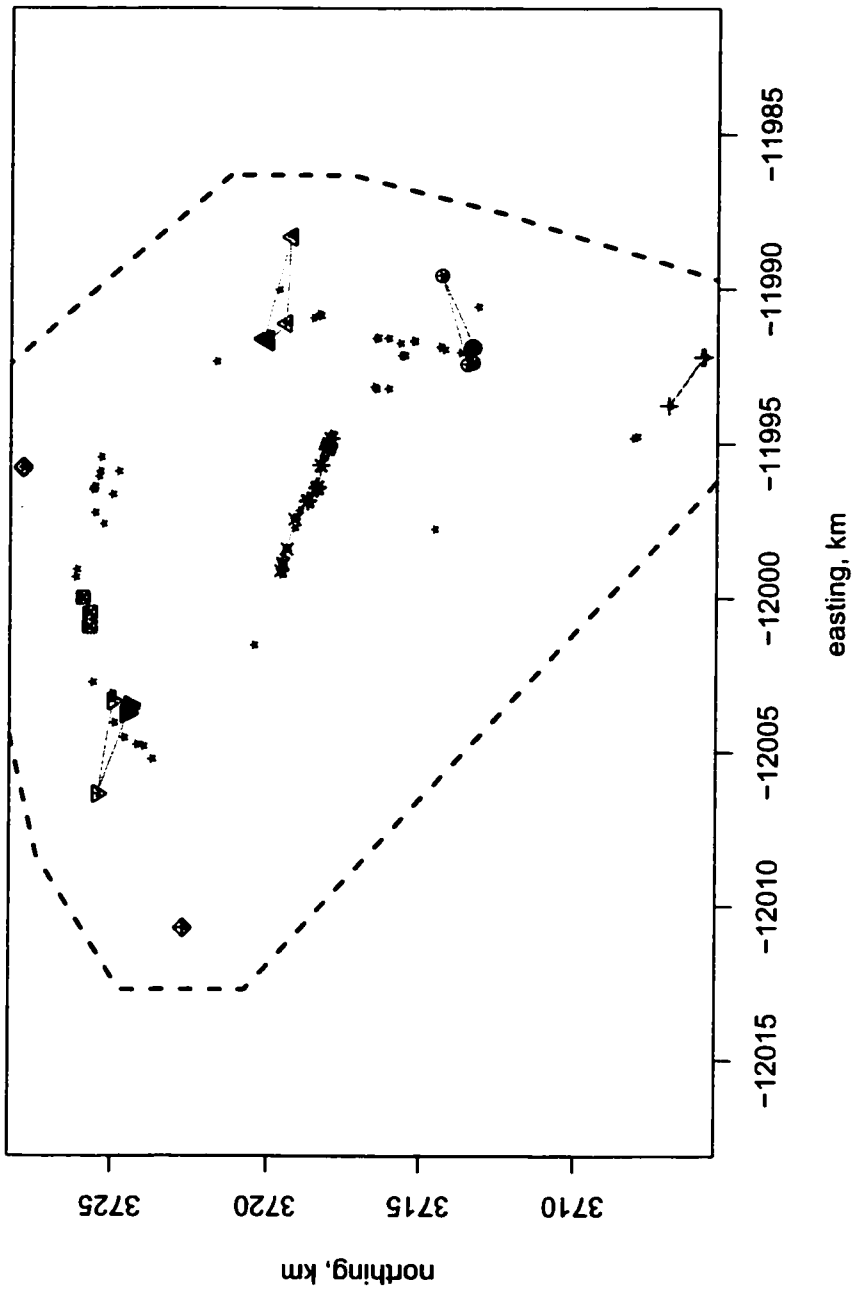


Figure 5.6: A spatial clustering of all named activities for user id 6. Named acts are shown as large symbols. Unnamed acts are shown as small stars. The dashed line represents a convex hull containing all named points. Note that unnamed activities typically do not fall inside of the cluster boundaries. Also note that many destinations appear to be clustered linearly, as if the EDCU transmitted its final point along a street leading to the actual destination.

within the small convex hull that bounds the named points. Finally the lower right curve was estimated using just the named points located within the small convex hull.

The smallest area with the greatest number of points (the lower left plot of figure 5.7) should produce the estimate of K which is closest to the ideal Poisson curve, *if* observing the process for a longer period of time adds randomness. However, this curve moves away from the Poisson curve when compared to the lower right (named points only). This shows that adding more points from the set of activity destinations isn't likely to result in an increase in apparent spatial randomness of the activity destination process. This effect was noticed for all four sets of data. The conclusion is that for these individuals, a short period of observation (using only the named activities) was not significantly changed, and if anything became more clustered, less random, upon further observation (using all of the unlabeled destinations). A reasonable conclusion is that, taken together, destinations are clustered, not random, and that no amount of further observation will reveal more random behavior.

As was said at the outset, this is an extremely small sample of individuals, and so the results should not be generalized. However, for each individual, one can argue that the long durations of the survey have captured a representative sample of each respondent's travel patterns. It is unlikely that continued observation will reveal enough new destinations to move the process closer to the Poisson process curve, especially when more observations appear to do the opposite. If this result can be repeated for a larger sample, then it would suggest that microsimulations of travel patterns should consider modeling destinations as being spatially clustered and even repeated, rather than randomly distributed from activity to activity and between time periods.

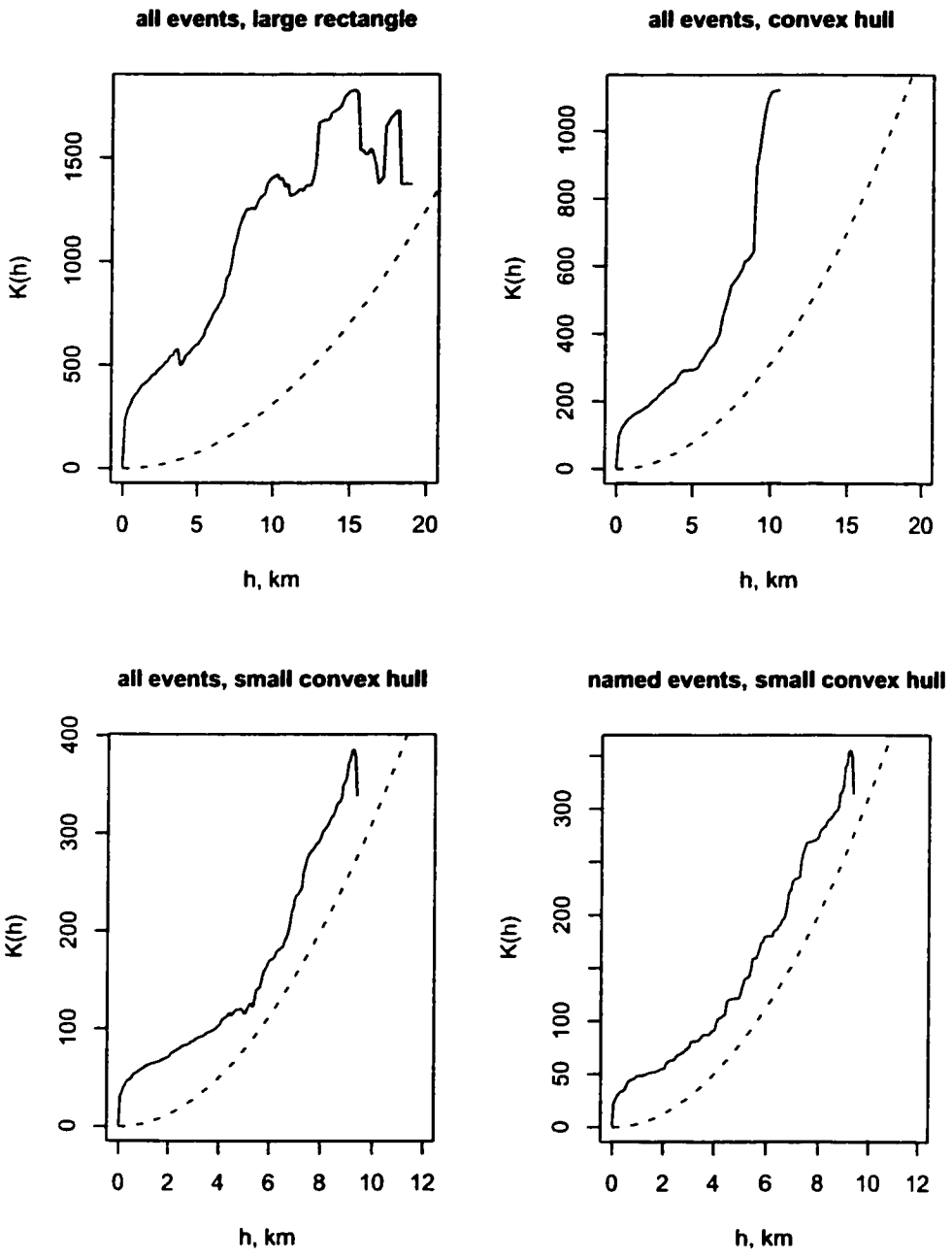


Figure 5.7: Effect of different boundaries on empirical estimate of K function, using observations from user id 6.

5.3 Analysis of activity-labeled destinations

The previous section showed that the observed destinations are clustered in space by computing the empirical K function. The K functions can also be calculated for each of the sets of locations that correspond to a particular activity type, or between one activity type and any other. In keeping with the comments in the preceding section on the effect of the spatial area on the estimate of K, the named activity destinations were evaluated using only the area of the convex hull surrounding these named points. The smaller area slightly reduces the relative degree of clustering, as shown in figure 5.7.

In the documentation for R's `spatstat` library, it is recommended that h be kept below $1/4$ of the length of the smallest side of the polygon bounding the data to reduce the impact of edge effects on the estimated K function. It is also noted that "bias may become appreciable" for sets with fewer than 15 observations. Using this as a guide completely ruled out three of the four participants, since none of their labeled activities were repeated more than 15 times. This is due to data entry fatigue, rather than non-repeating activities. It is likely that if the respondents were encouraged to fill out more of the on-line survey, the number of repeated activities would increase much faster than the number of new activity names. Alternately, one could apply the techniques that will be described in chapter 6 to boost the numbers of assigned activity names, albeit with some decline in the confidence one should place in those observations. However, as this is only a pilot study intended to guide future research and develop analysis techniques, the activity observations were not "enhanced." Using just the raw pilot data, only uid 2 had more than 15 occurrences of any single named activity, with three activities passing this threshold.

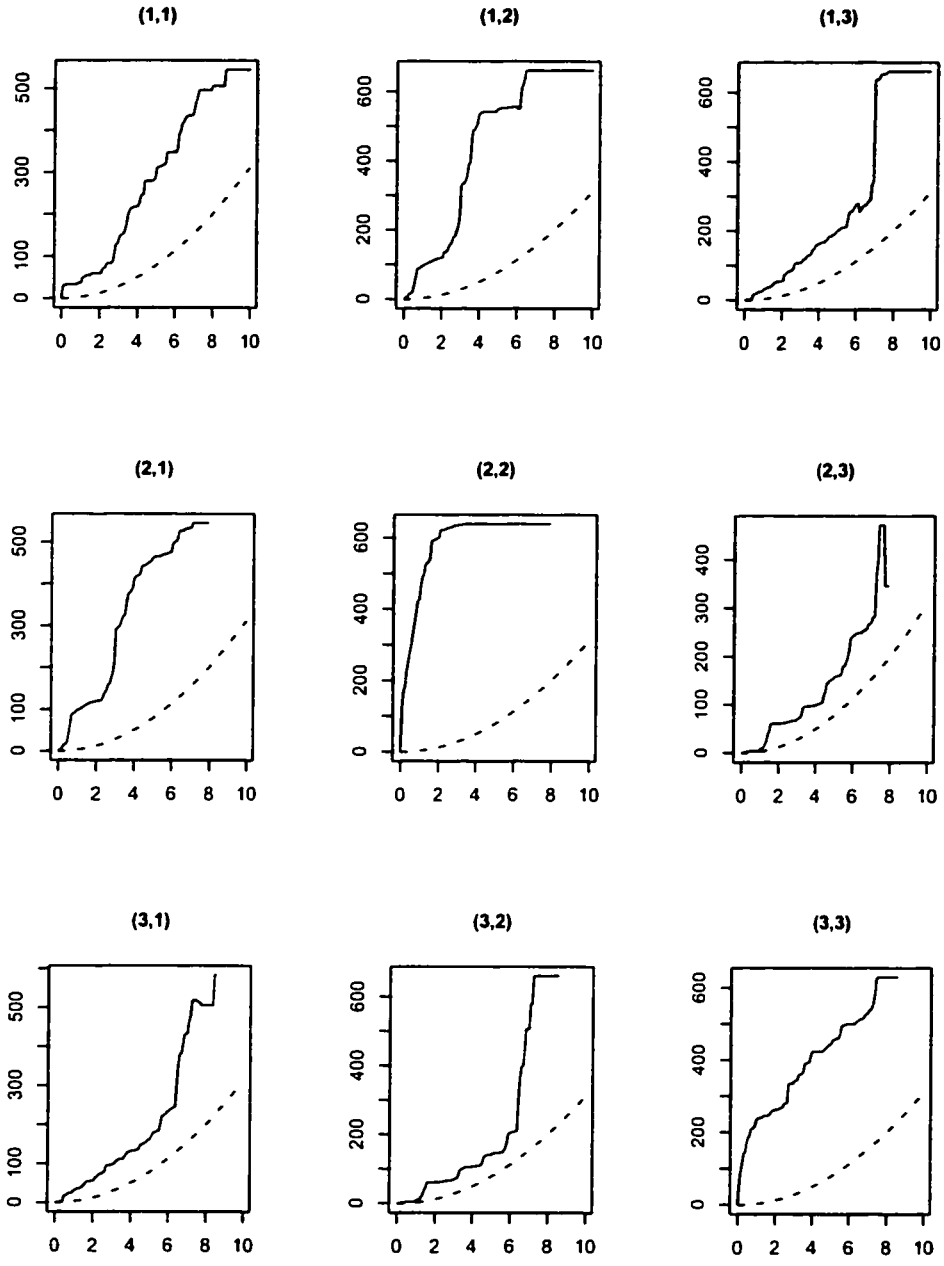


Figure 5.8: Estimated crossed K function versus distance h for the most frequent user id 2 named activity destinations. K is calculated within the convex polygon.

Figure 5.8 shows the own and cross estimates for K between the three activity types. Given that this is just one observation and the survey is just a pilot survey conducted informally, no general conclusions can be drawn from the results. However, the plots indicate that a full scale survey would do well to examine the effect of activity type and the relative effects of spatial randomness. Specifically, the relative locations of activities two and three lie close to the theoretical Poisson process curve. This result was quite surprising. Even though it may not hold up with more data collection, in the abstract it makes sense. The approximately Poisson distribution of the curve means that for any randomly chosen point labeled with activity 2, the observed points labeled with activity 3 are approximately distributed spatially randomly. The plot of activity 2 against itself exhibits a steep sharp vertical curve, which implies that activity 2 probably happens in a single location. Looking at a scatterplot of the destinations in figure 5.9 supports this speculation.

Figure 5.9 also demonstrates how hard it is to visually detect clustering. Activity 3 is quantitatively closer to randomly distributed than activity 1, but at first glance they look as if they are roughly coincident. A closer look shows that activity 3 events occur at several different distances from activity 2 to the maximum distance at which activity 2 is defined. With its outlying points, activity 1 does not have a continuous scattering of events across all different values of h. Further, the destinations of activity 1 that lie closer to the activity 2 location tend to be clustered together. This means that for those distances, there are *more* activities than would be expected from a Poisson distribution. Finally, the scatterplot can be slightly misleading in that there are some activity 1 events that occur at nearly identical locations, thus increasing the frequency of events even more.

Finally, note that the crossed K relationships are not necessarily reflexive. Any ran-

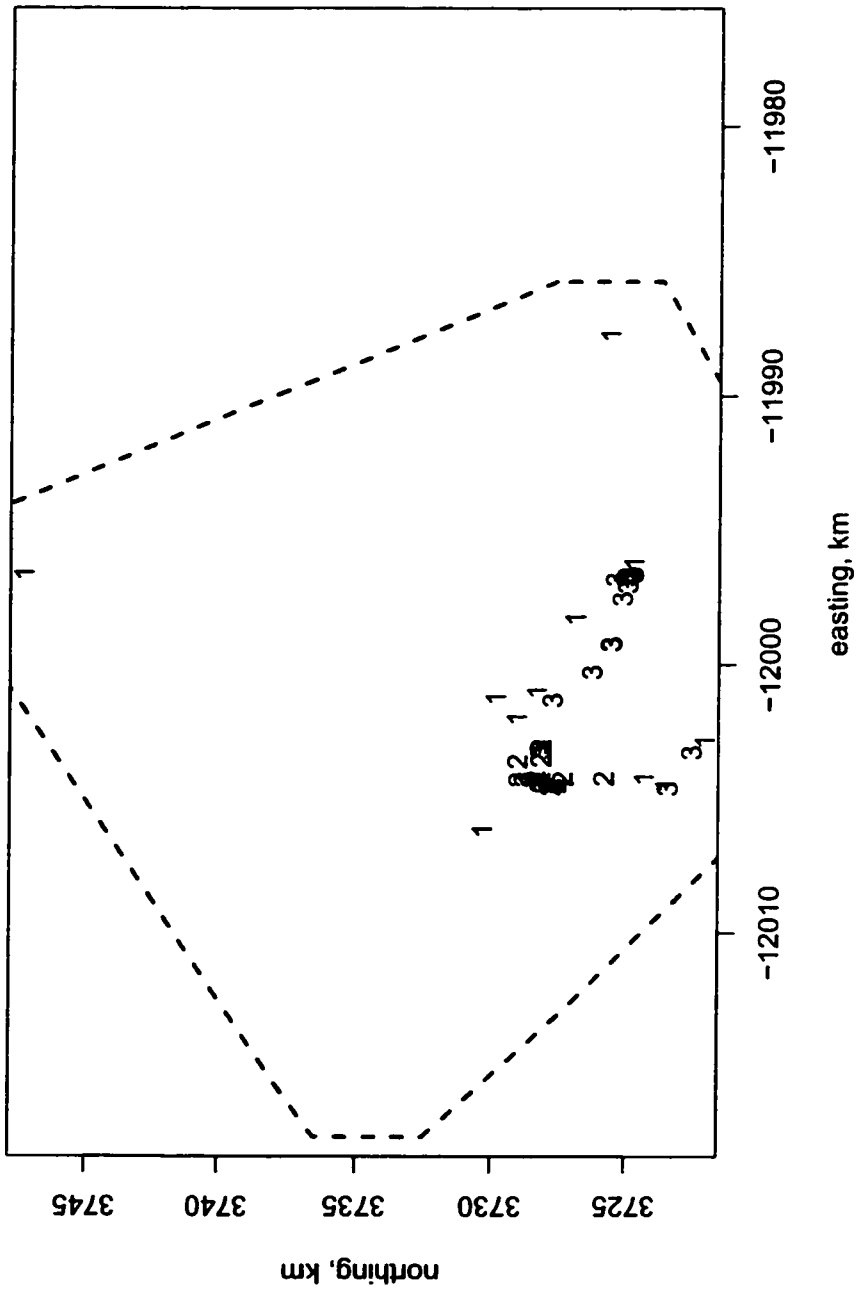


Figure 5.9: Scatterplot of activities 1, 2, and 3 for user id 2.

domly chosen activity 3 event finds activities 1 and 2 to be nearly Poisson distributed for very short length scales, but then after about 4 kilometers the destinations of the other activities become far more clustered than the Poisson curve.

5.4 Conclusions and areas for further research

The data used in the analyses reported here were gathered from a very small, informal test application of the Tracer data collection system. As such, this closing section provides more areas of future research than general conclusions. Nevertheless, two features stand out in the data. First, within the context of all unlabeled destinations, destinations appear to be clustered in space. At the very least this represents four data points in solid opposition to the presumption that destinations are chosen according to a spatially random process. The long durations of the observations further strengthen this counter-evidence result.

Second, the spatial randomness of an activity's destinations may be dependent upon the location of other activities' destinations. It was shown that one activity which is spatially clustered relative to itself is almost as random as a Poisson process when viewed from the locations observed for another activity. This is a piece of evidence going the other way, stating that perhaps some activities are spatially randomly distributed about other (perhaps more important) activity destinations.

There are several areas where the analyses performed here could and should be improved. First, one obvious critique is that destinations should follow land use, not planar space. Thus simulation frameworks which purport to choose their destinations randomly from within the space of "suitable" land uses may be accurately modeling a real destina-

tion point processes. Given the strong clustering shown in the data, this is probably not the case. However, given the impact of the shape of the activity space on the estimates of the K function, a more precise analysis of spatial randomness should take into account developed areas and obvious holes in the landscape, such as mountains, lakes, or vacant lots. Further, distances should be based on a transportation system metric rather than on straight line distances. Because the Tracer system collects travel data, the distances between destinations can actually be computed using the respondents' *own* travel patterns. This suggests an extreme case of limiting activity space—to estimate the spatial randomness only within the sausage of space surrounding observed travel.

Another area that should be addressed is to improve the on-line activity survey to reduce respondent fatigue and thereby obtain more activity labels than in the pilot study. Alternately, one could assign approximate activities to unlabeled destinations with some probability score, based on past entries, using the method documented in chapter 6. Increasing the numbers of labeled activities is important for understanding the relative spatial distributions, such as in figure 5.8. Another practical improvement would be to reduce and minimize the randomizing effect of the EDCU measurement error, so as to better observe the actual dispersion of activities. While this would increase the relative clustering, study of the point processes could focus on modeling the *process*, not the random measurement errors.

Further research into developing analytical models of activity destinations as point processes could also be fruitful. One candidate that was briefly explored with the data at hand was the Neyman-Scott process model (Cressie, 1993). This model defines a clustering process by first defining a Poisson point process, and then expanding each of the Poisson

points. The points can be expanded with any other process, such as a randomly filled disc, another Poisson process, or even another Neyman-Scott process in a recursive cascade. These kinds of models are easy to specify and tweak by hand, but are rather difficult to parameterize and fit to observed events. The effort required for this kind of analysis is best applied to a more rigorously collected data set.

The application of such activity point process models would be to provide a destination generating mechanism for an activity-based microsimulation model. A similar area of research would be to examine the point process implications of the various activity modeling frameworks that have been proposed. For example, the *ALBATROSS* model's reliance upon heuristically improved scripts could generate repeated spatial behavior. It is reasonable to suppose that a script for solving a particular activity scheduling problem could stabilize on visiting only a small set of destinations. There is also some evidence, provided by Recker's (1995) Household Activity Pattern Problem formulation, which indicates that perhaps people's real schedules are so constrained that the only way they can do them is to find one really good solution and stick to that. Similarly, McNally (1997) proposes a framework which maintains information on past events when determining the location of subsequent activities. This self-referential process may also generate clustering behavior when executed within a realistic model of activity opportunities.

Finally, a long-term goal of research in this area should be to discover whether it is possible to classify activities across different individuals solely according to the intrinsic spatial randomness of the locations used to engage in the activities. For example, it is often claimed that a work activity is similar to a school activity in many ways. Instead of blindly grouping these activities, one could sift through observed data and pull out all

activities that have similar spatial clustering or randomness with respect to themselves and to other activities. This would provide an objective, data-driven way to combine activity travel patterns across individuals.

In conclusion, the evidence presented in this chapter, sparse as it is, should at least spur more research into the relationships between activities and destinations. A greater theoretical understanding of how and why activities are scattered in space will enhance the practical linkage between activities, travel, and transportation.

Chapter 6

Collecting activity data from GPS readings

6.1 Introduction

The previous chapter showed that activities are clustered in space. This chapter attempts to convert that observation into a more quantitative association of activities and places. This chapter documents a method for using whatever activity data has been entered to try to tag unlabeled activity destinations. One clear lesson to be learned from the pilot study is that filling out an activity survey is more onerous than driving around with a GPS recorder. So a practical survey should expect to get more days of GPS records than days of activity data. Thus it is important to develop a method to link the reported activity names and other details with the destinations of those activities. One obvious application of this method is to generate longer sequences of activities, albeit ones with less certainty than the actual entries of the survey participant. This data can also be used to further streamline the

activity survey by suggesting only the most likely activities for an unlabeled destination. Finally, the evaluation of the most likely activities associated with unlabeled destinations also identifies those destinations about which very little is known. One could design a short follow-up survey that asks for more information on these points, rather than squandering the respondent's good-will collecting data on well-known destinations.

6.2 Association of activities with locations

Chapter 5 has shown that the destinations recorded in this pilot survey are clustered in space. This was shown by computing the empirical K function for each set of destinations, and demonstrating its pronounced deviation from a Poisson process. Further evidence of the coincidence of activities and destinations was apparent during the course of the survey. As documented in section 4.4, the on-line survey sorted the activity options and other input checkboxes according to a simple distance and time based metric. The previously entered activity name that occurred closest to the point in question was presented first. In the case of ties, the closest time was used to induce a sub-ordering. The pilot survey volunteers commented that the first option was often the one chosen, unless all of the past entries were irrelevant to the activity in question.

The sort algorithm worked well in that it usually placed first or second the most relevant location name, activity name, and people involved. But there were some annoying quirks that were difficult to eliminate. First, if the respondent decided to change the name of an activity over time, say from one long entry to multiple short entries, or if the EDCU missed a stop once and the respondent felt compelled to document the missed activity at an incor-

rect location, then the sorting algorithm would offer the "wrong" options. Further, as time goes by, one would want the most frequently chosen options to be weighted higher than simply the options that are geographically closest—something a distance-based sort cannot do. A more complete solution is to develop a probability distribution for the likelihood of engaging in different activities at different locations. That is the topic of this chapter.

There do not seem to be any similar efforts in the activity analysis literature. This is probably due to the short durations of most GPS-related surveys, and the focus on using GPS devices to verify the accuracy of larger activity survey samples, as discussed in section 3.2. Wolf et al. (2001) discuss eliminating the travel diary by carefully merging GPS records with a detailed geographic information system (GIS), but they do not discuss searching for or relying upon repeated spatial or temporal patterns within the GPS data itself. Again, this may be due to the shorter durations of GPS data collected.

A naive approach to associating observed activities with locations is to first bin the data by imposing a grid or tessellation on the observation area, and then use the frequency of observations in each bin to compute an empirical distribution. This is easy to do, but misses relationships between neighboring bins. Especially given the clustered nature of destinations, the likelihood of engaging in an activity at any given longitude and latitude is likely to be correlated with the likelihood at all nearby points. A technique called kriging has been developed in mining exploration, meteorology, and other disciplines where spatial processes are natural and unavoidable (Cressie, 1993). Kriging incorporates point estimates of spatial correlation, as well as global and local trends, when estimating a prediction surface.

6.3 Estimating a semi-variogram

In order to estimate a kriging prediction surface, one must first estimate what is known as the semi-variogram (Cressie, 1993; Venables and Ripley, 1999). The semi-variogram approximates the spatial variance of the observations. In a mining application, one would measure variables relating to the porosity of the rock, and so on. In a forestry application, one might measure the diameter of trees or the types of plants. In this application, the goal is to associate activities with destinations, and so the best variable to use is the frequency of observations—the numbers of unique activities (including repeats) at a particular longitude–latitude pair.

Within the observed data from the pilot survey, the frequency of visits to an exact latitude and longitude ranged from 1 to 4. It is to be expected that the frequency of any given location is likely to be low, since longitude and latitude are accurate to 6 decimal places. However, the general vicinity of a destination will have a relatively high concentration of visits if a repeat activity occurs there. This is shown graphically in figures 6.1 through 6.4. As more data is collected, one would expect the observation of unique points to steadily increase in the neighborhood of all frequently visited points, since there are an infinite number of points available as measurements become more precise. If the precision of observation is reduced, the effect of the noise diminishes and the underlying spatial clustering can be observed.

The semi-variogram is designed to characterize autocorrelation in spatial data. For small distances data tend to be autocorrelated. If the exact destination could be measured each time, then one would observe a uniform plane with sharp spikes of high activity at

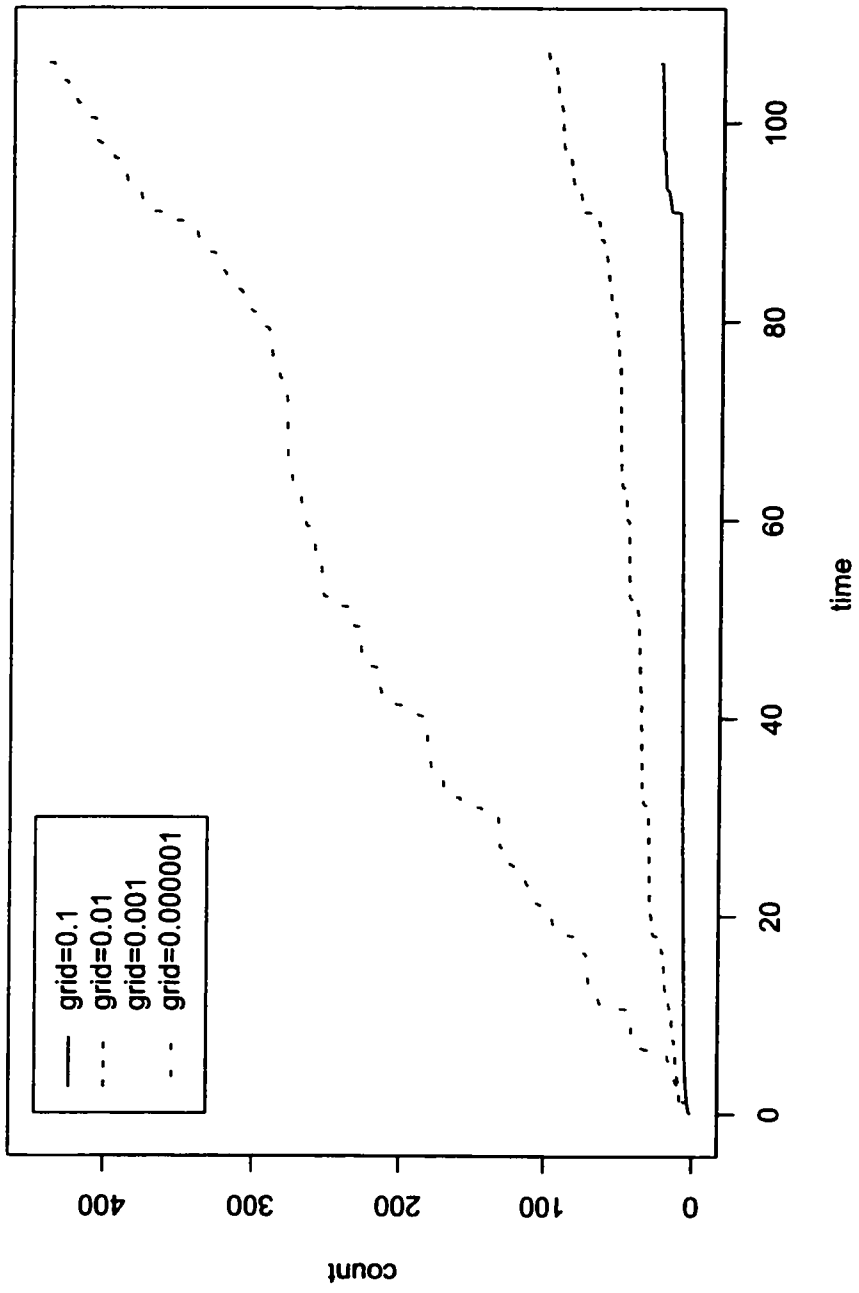


Figure 6.1: Growth of new destinations versus elapsed survey days, for user id 2. The tapering off of low precision measurements shows that destinations are repeated over time. The constant rise of finer precision measurements demonstrates the natural variability of GPS measurements.

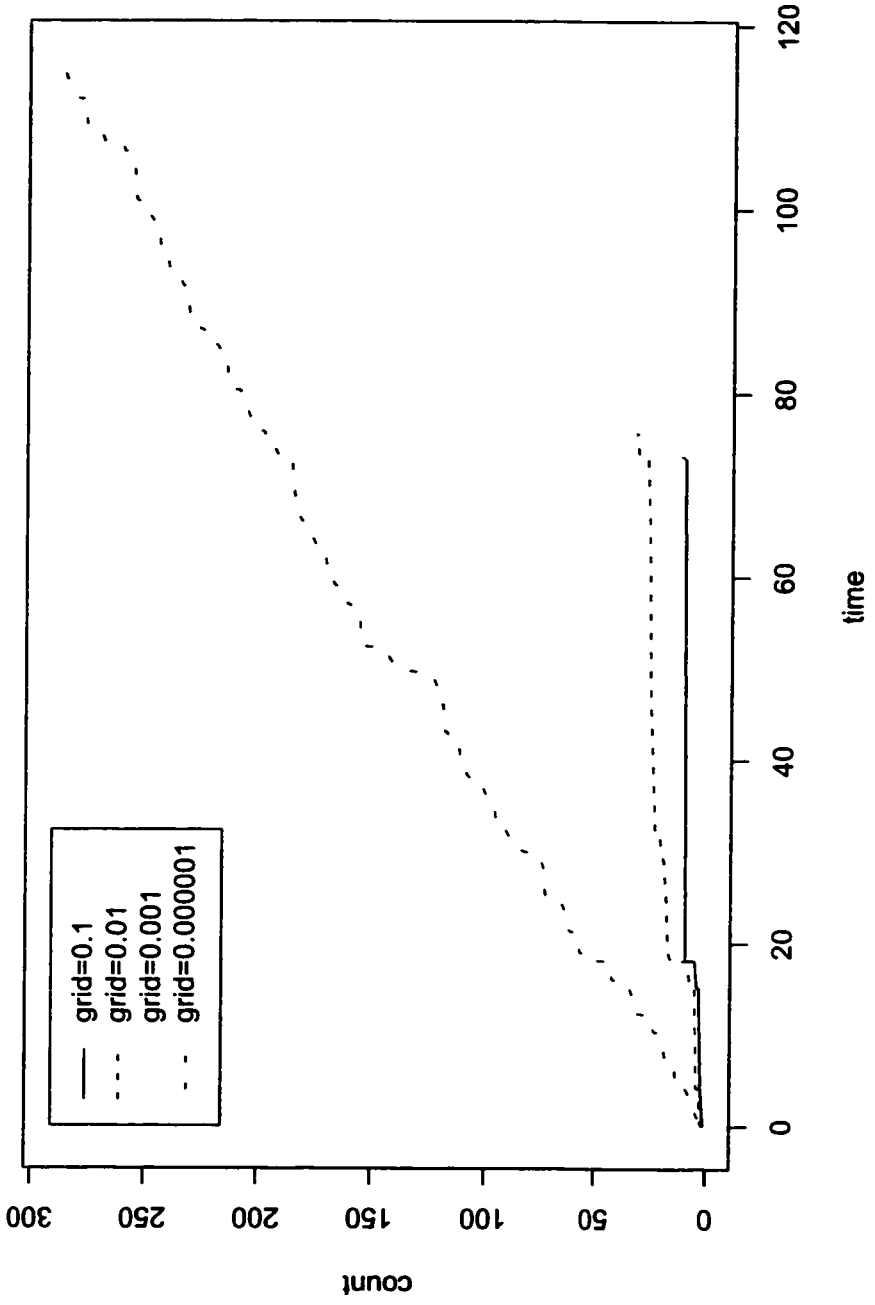


Figure 6.2: Growth of new destinations versus elapsed survey days, for user id 4. The tapering off of low precision measurements shows that destinations are repeated over time. The constant rise of finer precision measurements demonstrates the natural variability of GPS measurements.

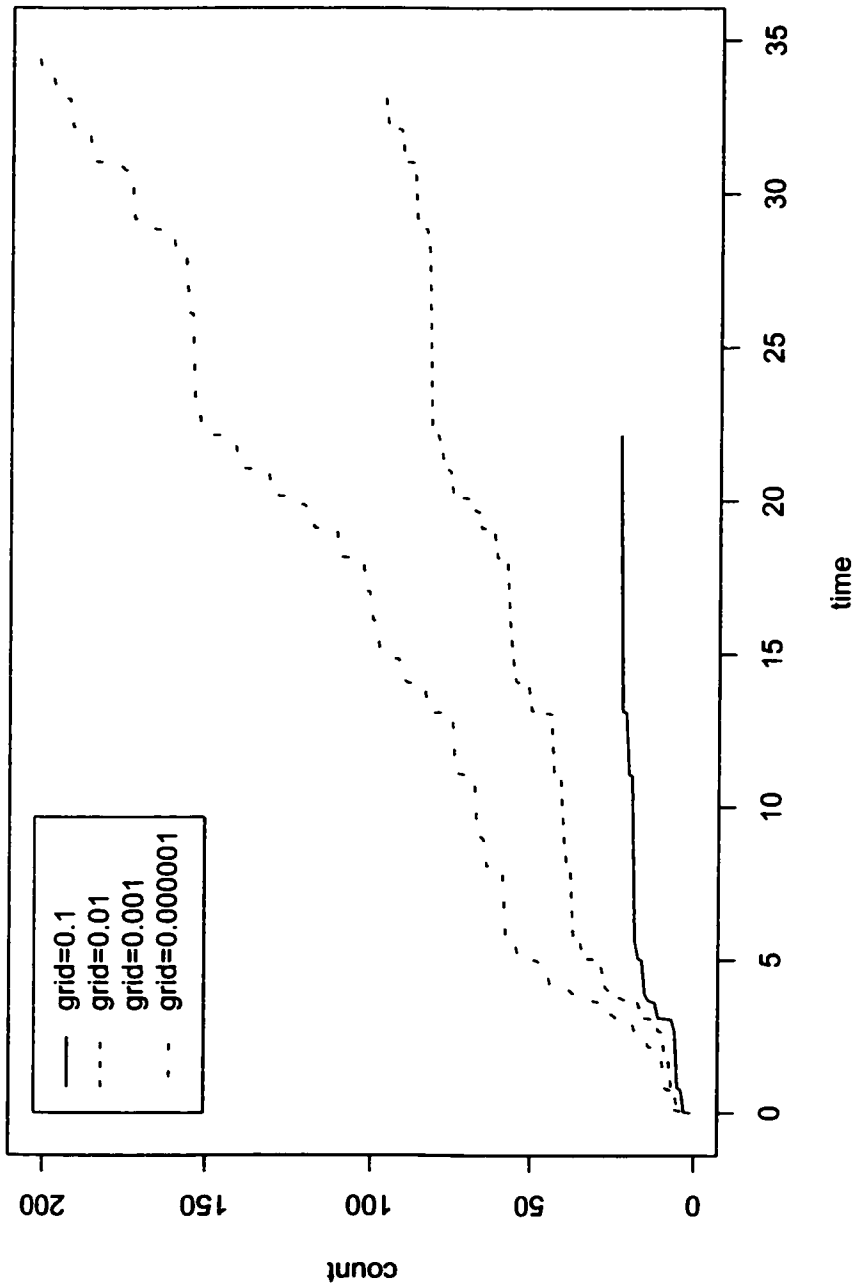


Figure 6.3: Growth of new destinations versus elapsed survey days for user id 5. The tapering off of low precision measurements shows that destinations are repeated over time. The constant rise of finer precision measurements demonstrates the natural variability of GPS measurements.

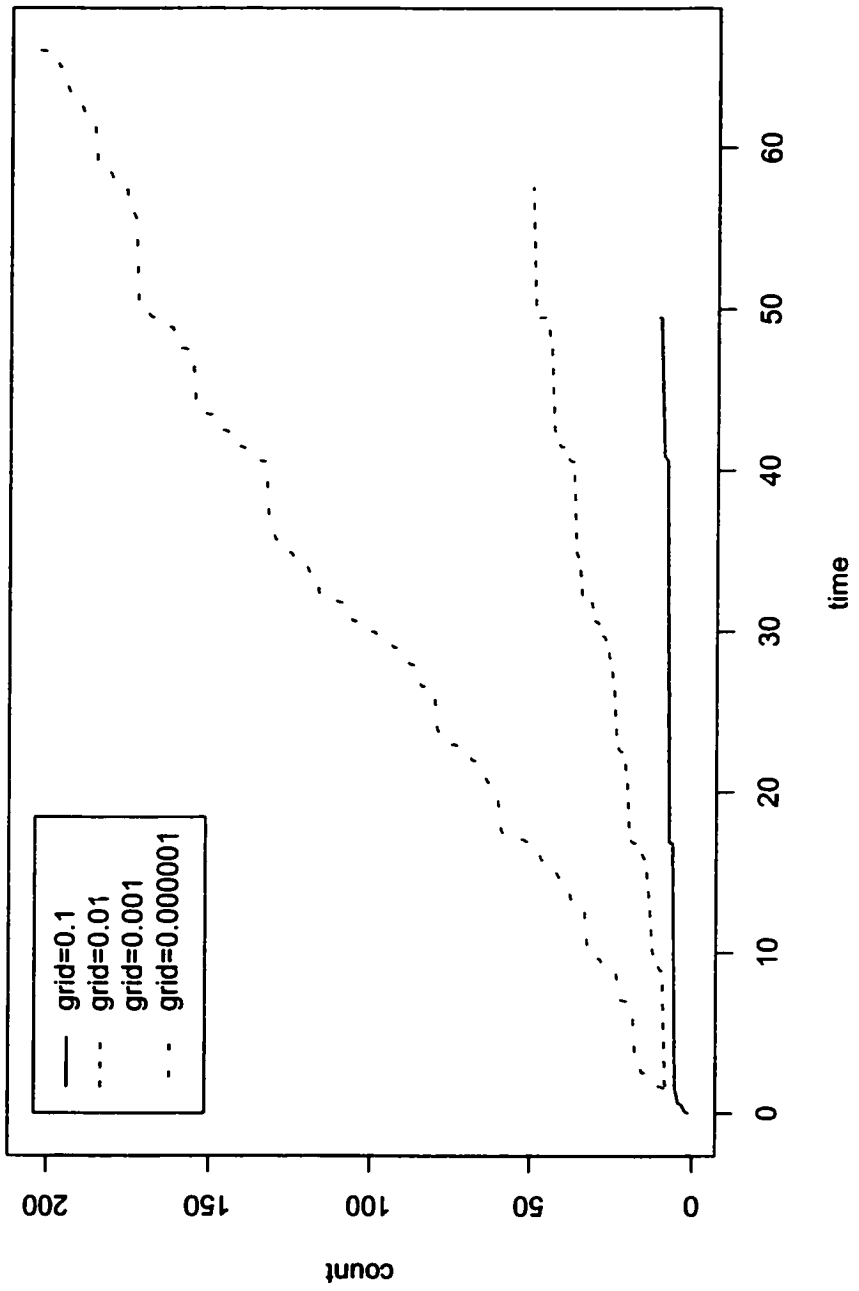


Figure 6.4: Growth of new destinations versus elapsed survey days, for user id 6. The tapering off of low precision measurements shows that destinations are repeated over time. The constant rise of finer precision measurements demonstrates the natural variability of GPS measurements.

various activity destinations. Instead, the spikes are flattened and spread out by a number of dispersing factors. These include the inherent error of GPS measurements, the variable polling rate of the EDCUs, the aliasing of the last reported time versus the actual time the vehicle was stopped, and so on. Many of these factors are described in more detail in chapters 3 and 4. In addition to these data collection sources of variability, the exact destination itself—in this case, where the vehicle is parked—is subject to a natural spreading in space. Factors such as the availability and location of a dedicated parking lot, versus the potentially more spatially distributing effect of first-come, first-served on-street parking will cause the final, measured destination to vary even though the "actual" destination of the activity is the same.

At larger distances, this effect declines and the variance stabilizes, as can be seen from the low precision curves in figure 6.1 through figure 6.4. At the coarsest granularity, the numbers of new destinations increase quickly at first, and then grow at a slower rate. This is no surprise, as the results of the previous chapter demonstrated that activity destinations are clustered. To give an idea of the scales involved, in Orange County, California 0.01 degrees of longitude by 0.01 degrees of latitude is roughly 0.9 km by 1.1 km, or 1.44 km^2. This area defines a rectangle that approximately bounds South Coast Plaza, a large regional shopping mall. In short, the travel patterns observed are clustered at the scale of a shopping mall, which may be peculiar to Southern California.

In terms of the mechanics of estimating a semi-variogram, there are many factors that one should consider. For example, there may be significant directional or systematic variations. This is a natural result of the fact that vehicles tend to drive on linear roads, although it is softened somewhat by the existence of parking lots at destinations. Examining a scat-

ter plot of recorded destinations for known points show that the EDCU device would often record its last point on streets leading up to the final destination. As an example, the scatterplot shown in figure 5.6 on page 67 has a handful of tightly clustered lines of points. This was common for all four vehicles. This will result in a large variation in one or two directions, and very little variation in other directions. But it is difficult to specify these local spatial and directional effects in general terms. The goal of this analysis is to be able to estimate a kriging surface for any respondent's data, without having to individually tweak the local parameters. For a general method that is applicable to any individual traveling in any area, it is safer to presume that the autocorrelation is uniform in all directions. However, a large scale study would do well to consider the local effects of the transportation network on the expected variance of destination measurements.

It is also theoretically possible to estimate separate variograms for each activity type. However, this was not possible due to the small numbers observed of each kind of activity. In a larger study, it is likely that the numbers of activities will be the same or fewer per person, given realistic expectations of the maximum duration of the survey. On the other hand, it is reasonable to assume that the spatial autocorrelation is largely independent of the activity at the destination. Effects such as small GPS errors, the scattering of parking locations around the destination, and the slight aliasing introduced by the periodic polling and updating scheme of the GPS data collection unit occur independent of the destination or the activity, although some factors such as the availability of a dedicated parking space are possibly related to the actual activity. Further analysis of these effects should be conducted with a larger, more formally collected data sample.

For the above reasons, the semi-variogram was estimated for each respondent using all

labeled locations, regardless the name of the activity assigned. Unlabeled activities were not used, as the intent is to try to use only the labeled activities to infer information concerning the unlabeled destinations. For three of the four sets of points, the semi-variogram was estimated to a maximum distance of 5 km, ensuring that local effects are captured, but not extending the influence of a point too far. For user id 5, whose trips covered very long distances and who labeled only 12 destinations, the estimation of a variogram had to be performed with a distance limit of 15 km. A robust estimator from Cressie (1993) implemented in the geoR library package of the R statistical language and environment (Ihaka and Gentleman, 1996) was used to estimate the semi-variogram. These estimates curve over a largely horizontal line through the observed data, with a short initial curve capturing the local spatial auto-correlation. The semi-variograms were then used to estimate kriging surfaces for each of the labeled activities. Again the library geoR was used, with the estimate computed using a weighted least squares (WLS) estimator, with the so-called nugget effect (spikes at data points; see Venables and Ripley (1999, p. 441)) allowed to be non-zero and free to vary in the estimation. The resulting surfaces are discussed in the next section. Once again, note that a much more extensive treatment of the estimation of the semi-variogram is warranted, and should be performed when a larger survey sample is collected. As this is only a pilot study, the most important contribution is to demonstrate the potential applications of this data.

6.4 Tagging unknown activities using estimated kriging surfaces

Some of the estimated surfaces are shown in the following figures. For uid 4, only a small number of recorded events were labeled with activity names. However, only a few activities took place outside of a small geographic area. Figure 6.5 shows a kriging surface estimated for a rectangle encompassing all of the named points. The surface shows the locations where any activity might be performed (ignoring the activity names). This plot (as well as all of the following surfaces) is rendered such that black areas indicate higher predicted frequency of observations, while white areas indicate a predicted value of very few events. The figure also shows the named activities as stars, and the unnamed activities as points.

Similar kriging surfaces were generated for all activities entered. All of the activity-specific kriging surfaces were used to predict the expected number of events of each type at all of the unlabeled locations. When these values are summed and used to normalize the various predictions, the result approximates the probability that a certain kind of act might have occurred at a particular location, given the past survey entries. As shown in figure 6.6, this procedure assigns labels with a 90% or greater probability to 79 unlabeled activities. This is only a small number of the 296 unlabeled activity destinations, but it is somewhat remarkable in that only 20 activities were provided labels in the survey. Again, this result only captures the relative predicted frequency of activities. For the 79 destinations above 0.9, one activity is predicted to occur more than 9 times out of 10, or alternately, the other activities are known from past observations to be highly unlikely to occur at that point.

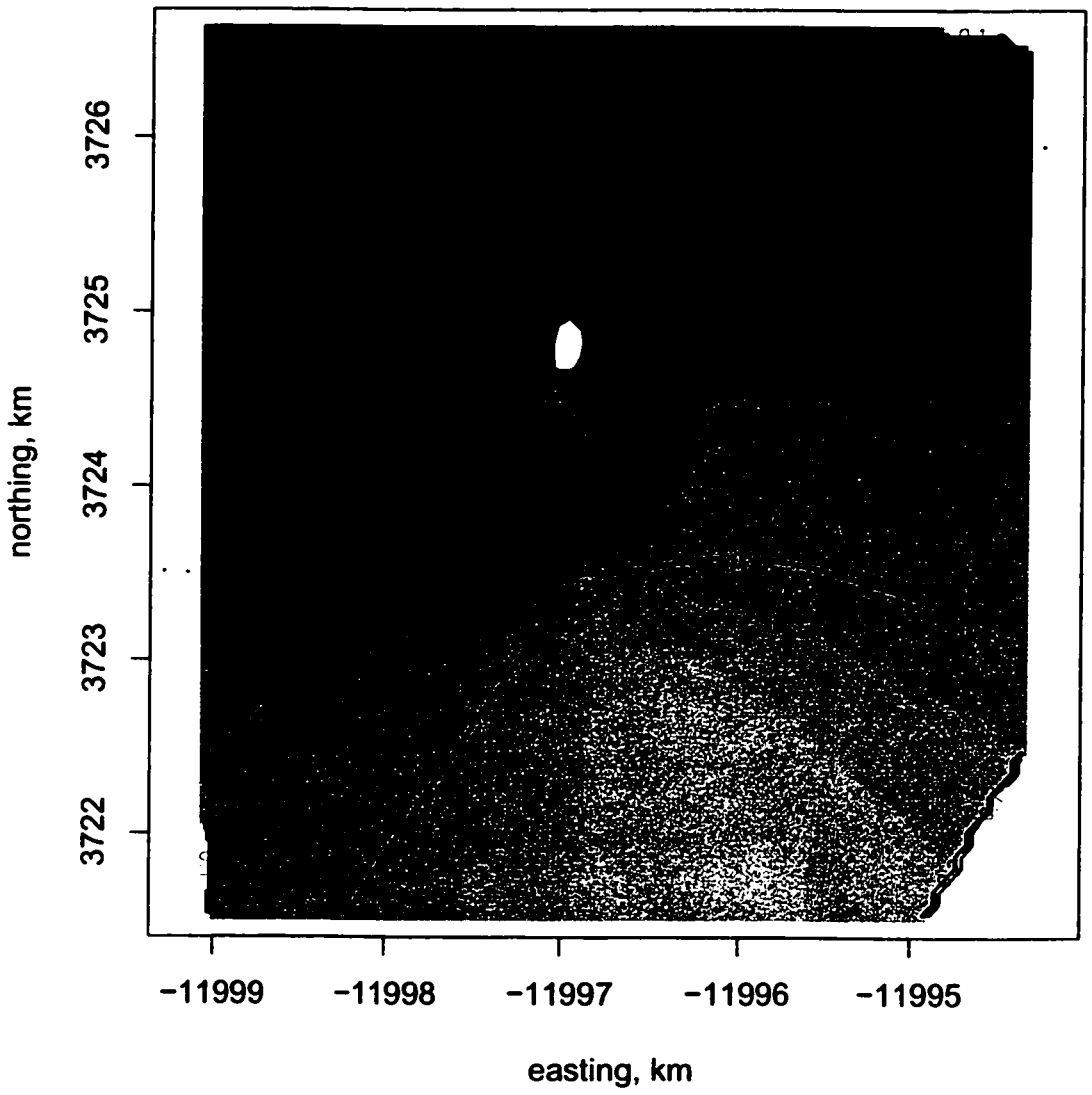


Figure 6.5: Kriging surface for user id 4 named activities. The surface was computed using the named (starred) activities. The small dots represent unlabeled points. The dark areas represent higher expected frequencies, and the white areas lower frequencies.

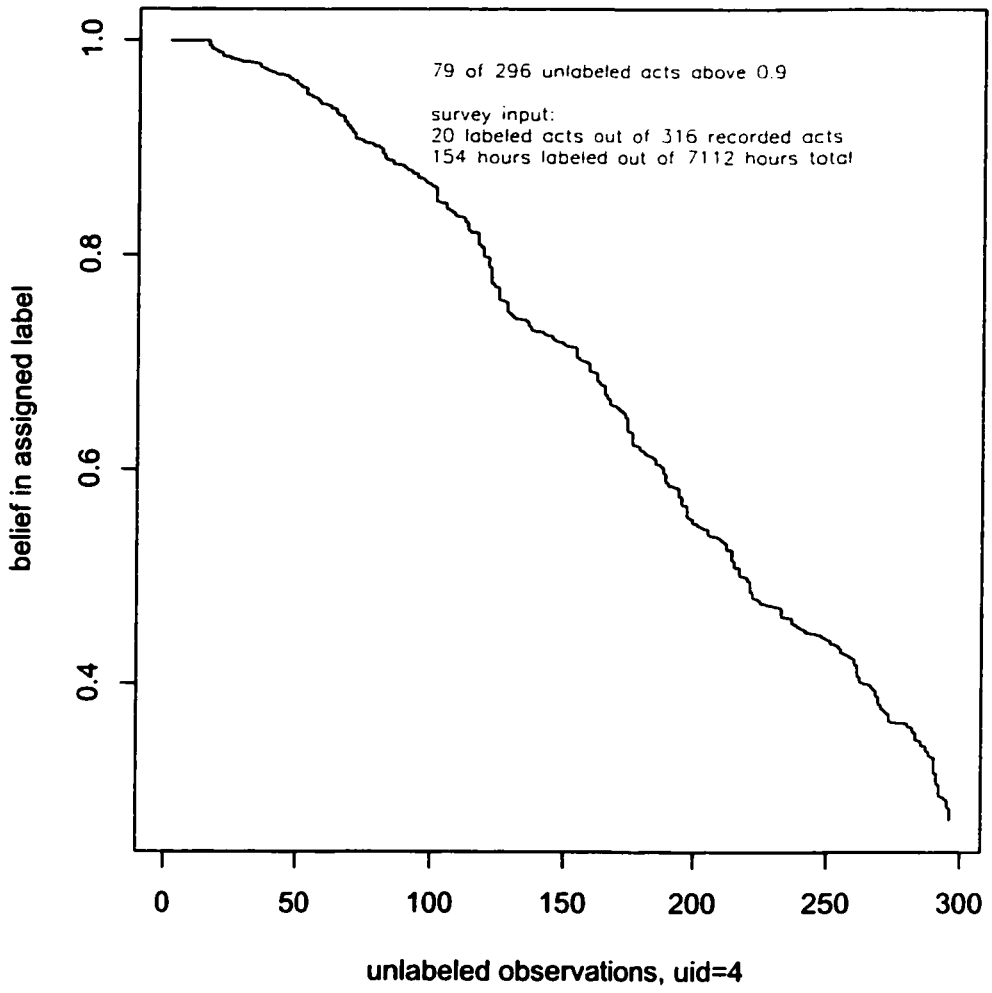


Figure 6.6: Probabilities of the most likely activity labels assigned to each unknown activity for user id 4, given the location of the activity and past survey entries.

A different low-response case is given by the set of observations from uid 5. Here a post-mortem analysis revealed that most of the recorded stops were actually mistakes. The cellular data antenna had been placed inside the car, which likely caused a a degradation in the signal quality, which in turn triggered a bug in the data transmission and stop recording process (see section 4.4.1 for details). The volunteer stated that most of the correct stops are located in the neighborhood of the few labeled activities. As can be seen in figure 6.7, the estimated surface is mostly a uniform gray, indicating no real decidability. Following this trend, figure 6.8 shows that the most likely activity for most of the unlabeled activities cannot be assigned with anything greater than a 70% probability. Even though many of the unlabeled activities are the result of a data collection error, the responses of this uid are similar to what one would expect for any respondent with only a few labeled activities and many unlabeled activities spatially distant from those labeled events.

Respondents uid 2 and uid 6 both have a much higher percentage of labeled versus unlabeled destinations, as well as more recorded activity categories, when compared to respondents uid 4 and uid 5. The higher response rates result in generally higher probabilities assigned to estimated activity labels when the locations are well known. The high probabilities occur both because the particular activity is by itself more likely to occur, and because all of the other activities are less likely to occur.

This effect is demonstrated by comparing figure 6.9 and figure 6.10. The surfaces shown in figure 6.5 and figure 6.7 plot the total destination frequency, rather than the relative likelihood of any one named activity. That is, surface corresponds to the likelihood of engaging in *any* named activity somewhere on the plane. Figure 6.9 shows the same plot for uid 6. In contrast, figure 6.10 estimates the relative probability of the volunteer

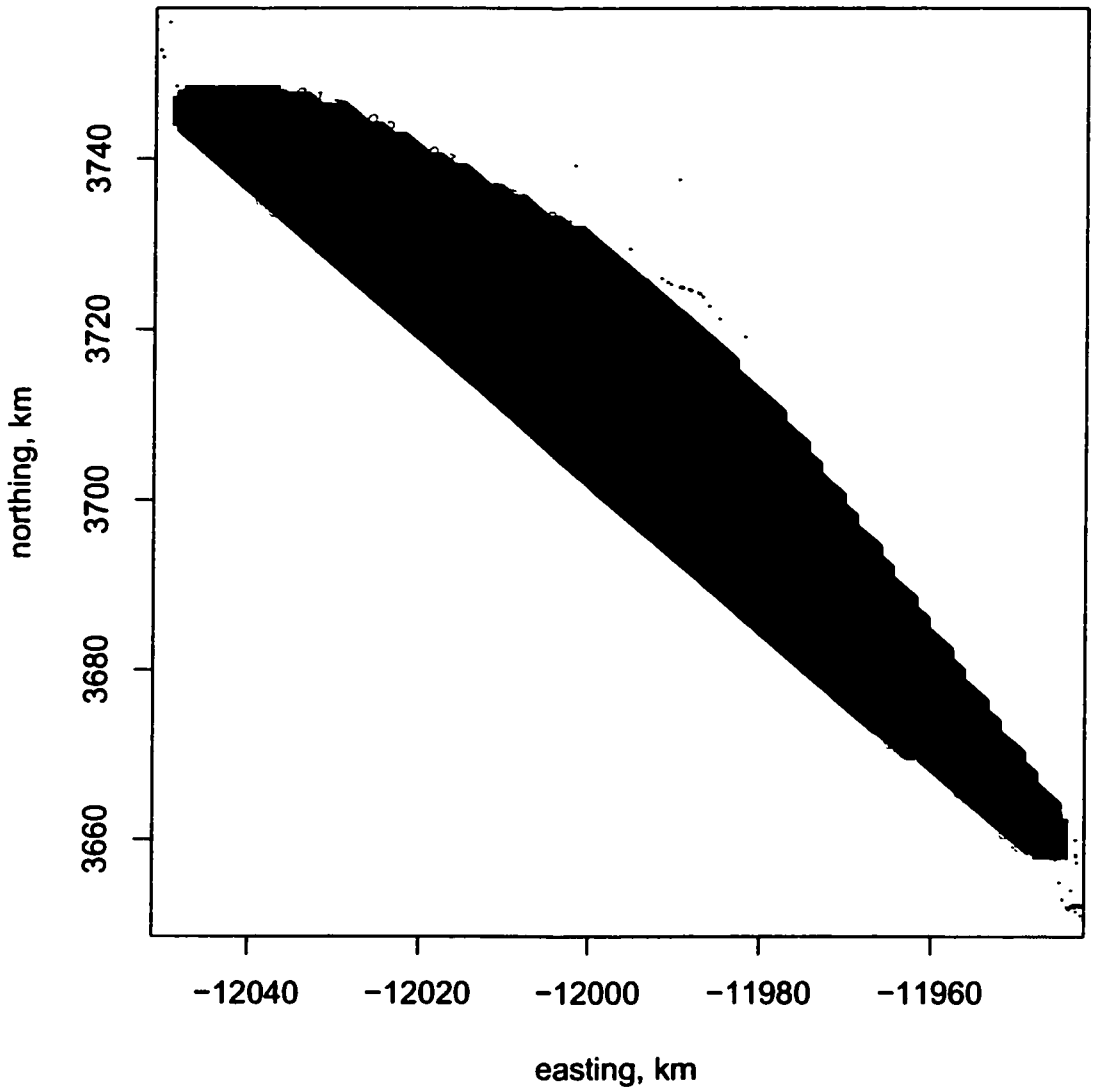


Figure 6.7: Kriging surface for all user id 5 activity destinations. The surface was computed using the named (starred) activities. The small dots represent unlabeled points. The dark areas represent higher expected frequencies, and the white areas lower frequencies. Most of the unlabeled point are in fact uncorrected data collection errors along a daily travel route.

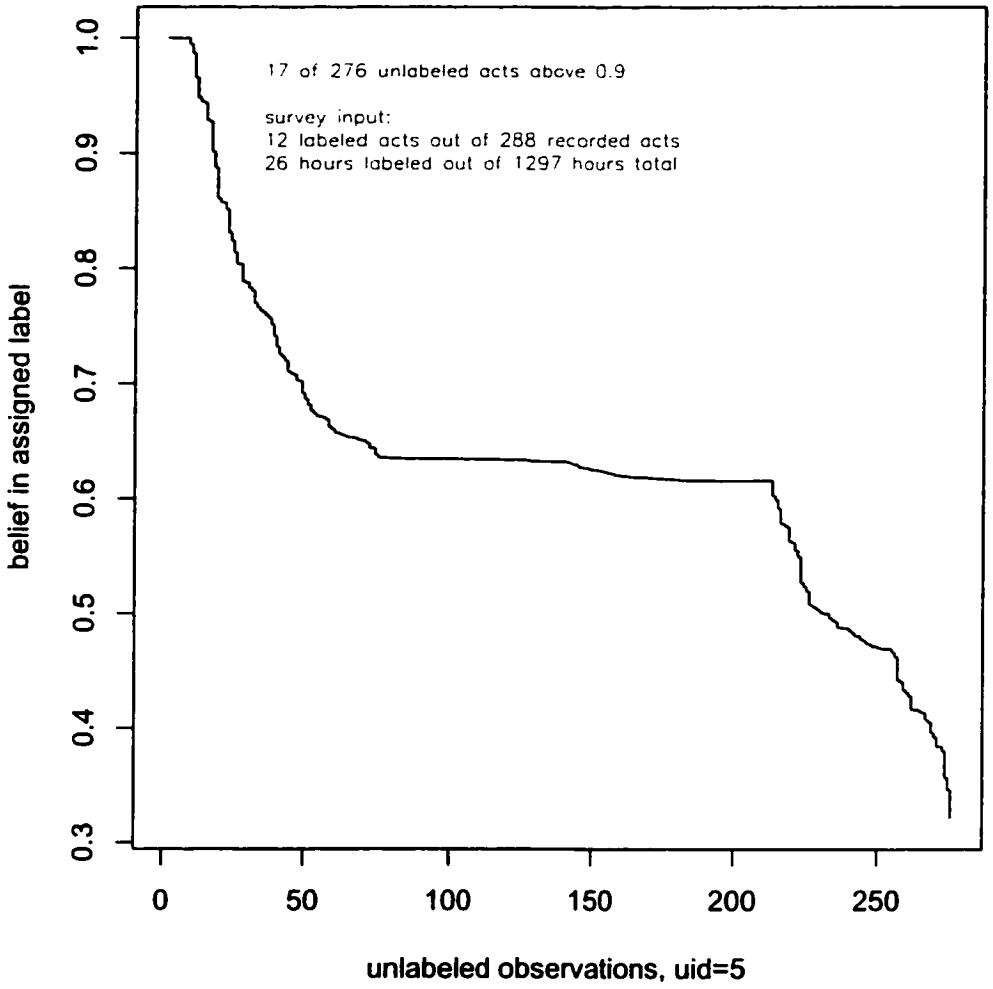


Figure 6.8: Probabilities of the most likely activity labels assigned to each unknown activity for user id 5, given the location of the activity and past survey entries. While the unlabeled activity destinations were reportedly errors, this curve is similar to what one should expect from a low response rate with a large spatial distance between points. Compare to figure 6.6, which has less spacing between points.

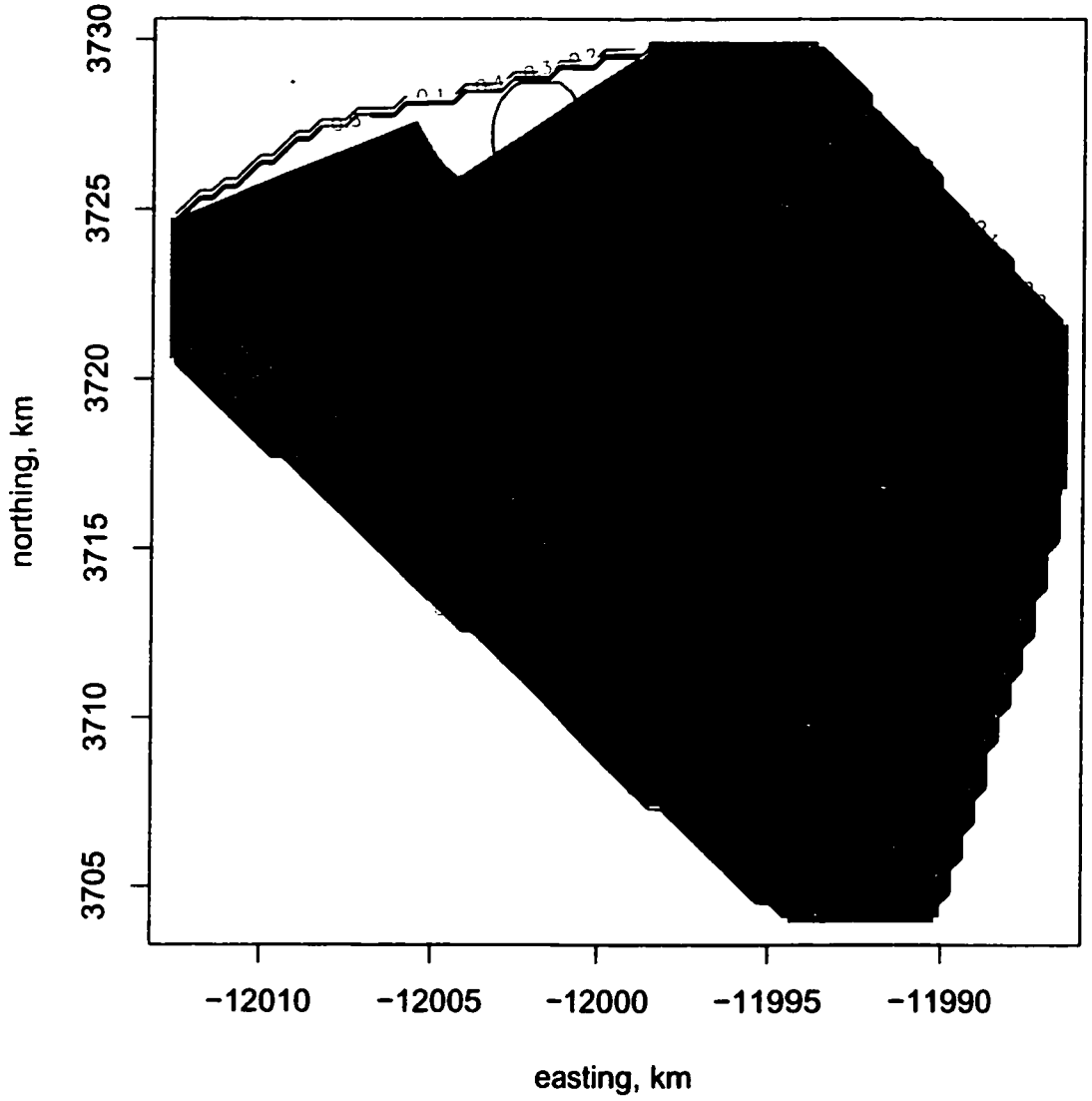


Figure 6.9: Kriging surface for user id 6. The surface was computed using the named (starred) activities. The small dots represent unlabeled points. The dark areas represent higher expected frequencies, and the white areas lower frequencies.

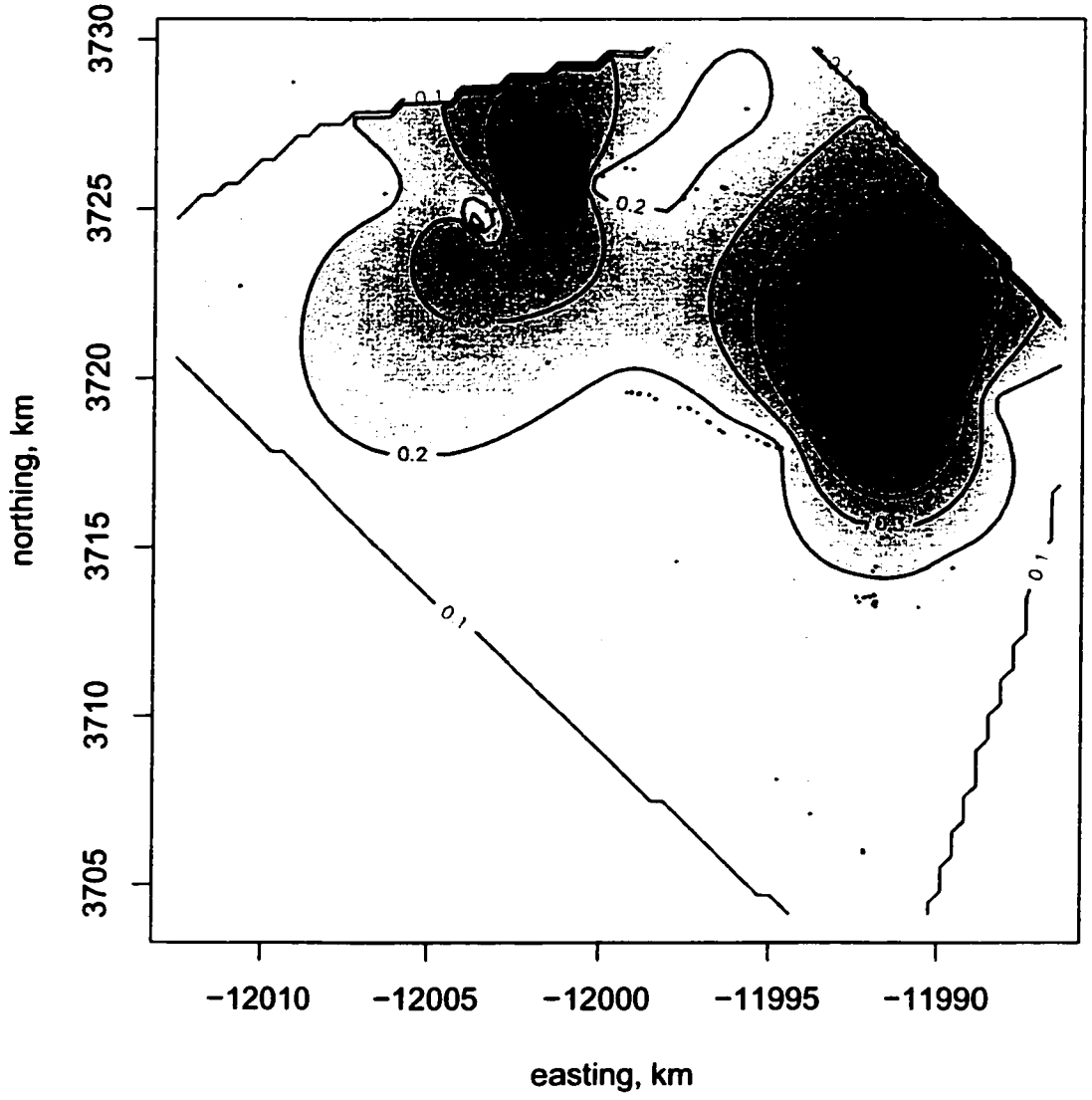


Figure 6.10: Kriging surface for user id 6 activity 1. The stars are activity 1 observations, while the points are both labeled and unlabeled destinations. Compare areas of high and low expected frequency to the "any activity" surface of figure 6.9.

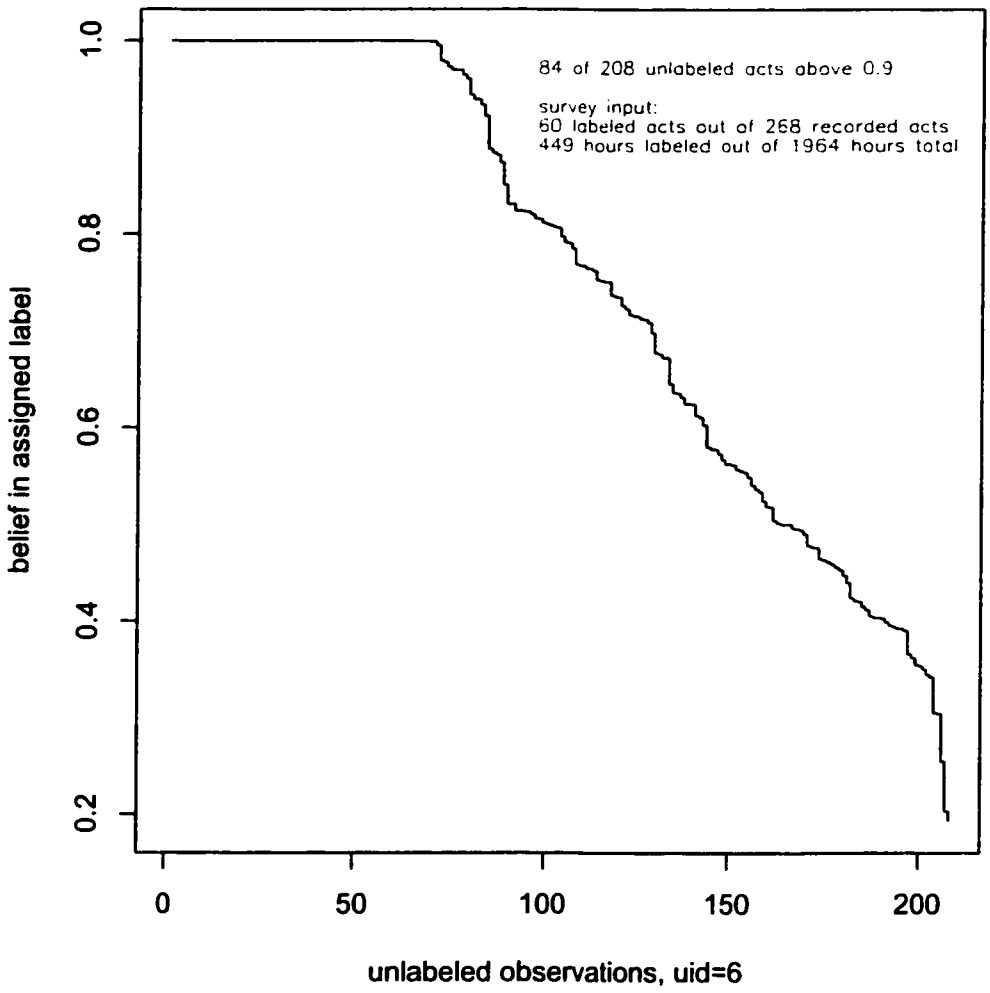


Figure 6.11: Probabilities of the most likely activity labels assigned to each unknown activity for user id 6, given the location of the activity and past survey entries.

engaging in activity 1 at any point in the region. Careful comparisons between the two figures will show that the white points in figure 6.10 (indicating a very low likelihood of engaging in activity 1 at that point) correspond to points that have been given some other activity name (shown as stars in figure 6.9).

As the numbers of repeated observations goes up for a particular activity in an area, so too does the size and extent of the surrounding neighborhood. But if another activity is also observed at or near the same point, the spreading effect is reduced. Finally, when all of

these labeled activities are used to develop a kriging surface for some *other* activity, the net effect is to have a tight cluster of zero frequency observations incorporated into the kriging surface. This will generally improve the ability of the various kriging surfaces to assign a single activity name with high confidence to the unknown observations that take place within well known regions, as shown in figure 6.11. Finally note that while the number of activities "labeled" by the kriging surface is even higher than the number actually labeled in the on-line survey, this effect will decrease with a shorter period of GPS data collection, although the proportion of labeled versus unlabeled may remain relatively constant.

The last set of plots show the impact of larger numbers of observations on the kriging surface. Three of the most frequent activities for uid 2 are shown, with activity 1 in figure 6.12, activity 2 in figure 6.13, and activity 3 in figure 6.14. These are the same activities examined in section 5.3, figure 5.8 on page 71. The locations for activities 1 are distributed throughout the observation area. Activity 2 has the highest frequency of any activity, and all of the recorded events seem to be clustered in one small area. Thus there is an extremely low probability that activity 2 will occur at any other location, as shown by the large area of white in figure 6.13. In between these is activity 3, which is somewhat spread in space, but not as dispersed as activity 1. The higher spatial dispersion of activities one and three give their kriging surfaces a higher expected frequency of engaging in the activities at any given point, resulting in the grayer tinge to those plots. As more data are collected, if it turned out that only these locations are used for these activities, then the peaks would get higher and higher relative to other areas, which would increase the relative contrast between high (black) and low (white) probability areas.

One potential problem that has not been addressed is that the kriging surface has prob-

lems beyond the boundaries of the observed data. Consider the surface shown in figure 6.12. The upper right hand corner is a dark black, even though only one observation is in the vicinity. The efficacy of this activity labeling approach is linked to the degree of repetition and clustering in the respondent's behavior. However, if in this case the respondent decided to travel to a new location in the northeast corner of figure 6.12, that new destination would be assigned a high frequency for activity 1. If some other activity has an equally high edge effect, then this won't matter to the prediction, since no activity will dominate relative to other activities. Of course relying on edge effects canceling each other out is error prone. A simple correction is to bound the valid range of the kriging prediction surface by the observed, labeled destinations, which would assign equal probability to all activity types if an activity occurs outside of the current known area.

Despite the dispersion of these plots, the underlying activity processes are still spatially clustered. An assignment of activities to uid 2's unknown points generates a high probability of assignment for a large number of the unlabeled points, as shown in figure 6.15. Again, the absolute numbers of labels assigned is directly related to the length of the GPS survey period. The proportion of the assigned to unknown is likely to be stable, however, assuming stability in the general activity pattern of the respondent.

The kriging surfaces and the resulting high numbers of high probability labels reinforces the result of chapter 5 that destinations are clustered. Even though activities in general *and* activities that have been assigned a particular activity name are dispersed in space, the destinations tend to be tightly clustered around previous destinations, which results in the generally high probabilities shown in figure 6.15. While any randomly chosen location might have a roughly equal probability of being assigned to activity 1 or activity 3, in fact a

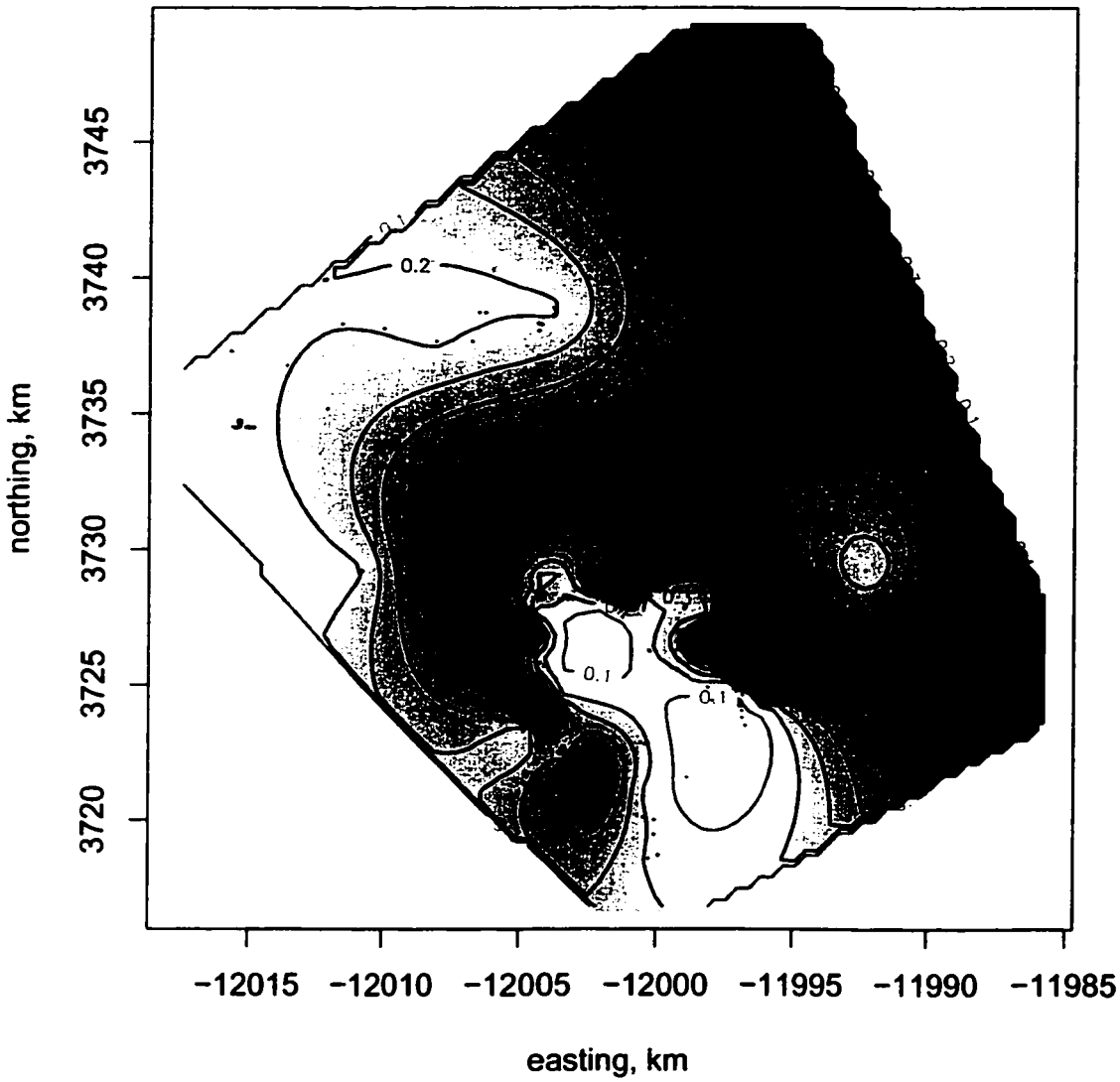


Figure 6.12: Kriging surface for user id 2's activity 1. See also figure 5.9.

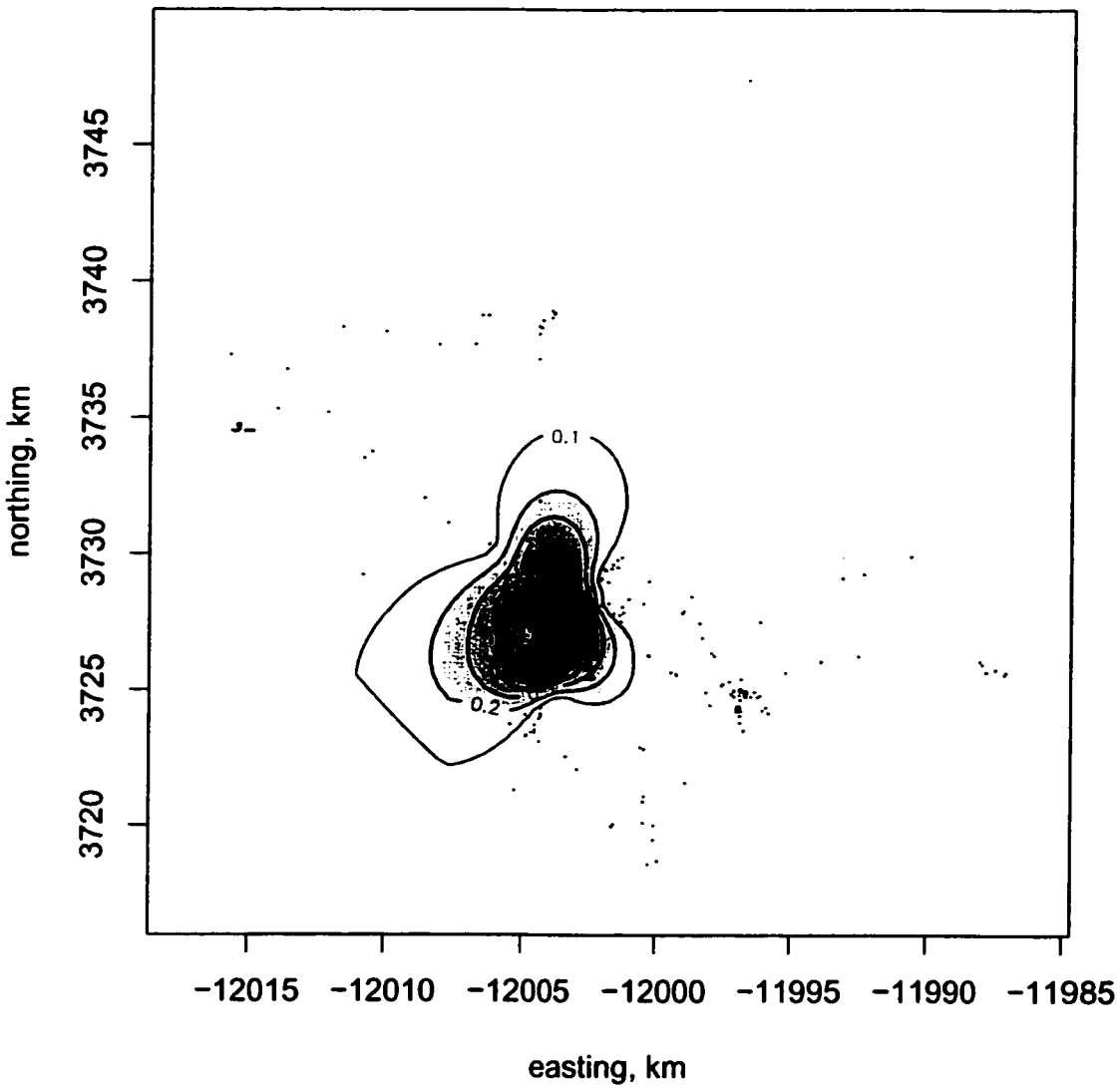


Figure 6.13: Kriging surface for user id 2's activity 2. See also figure 5.9.

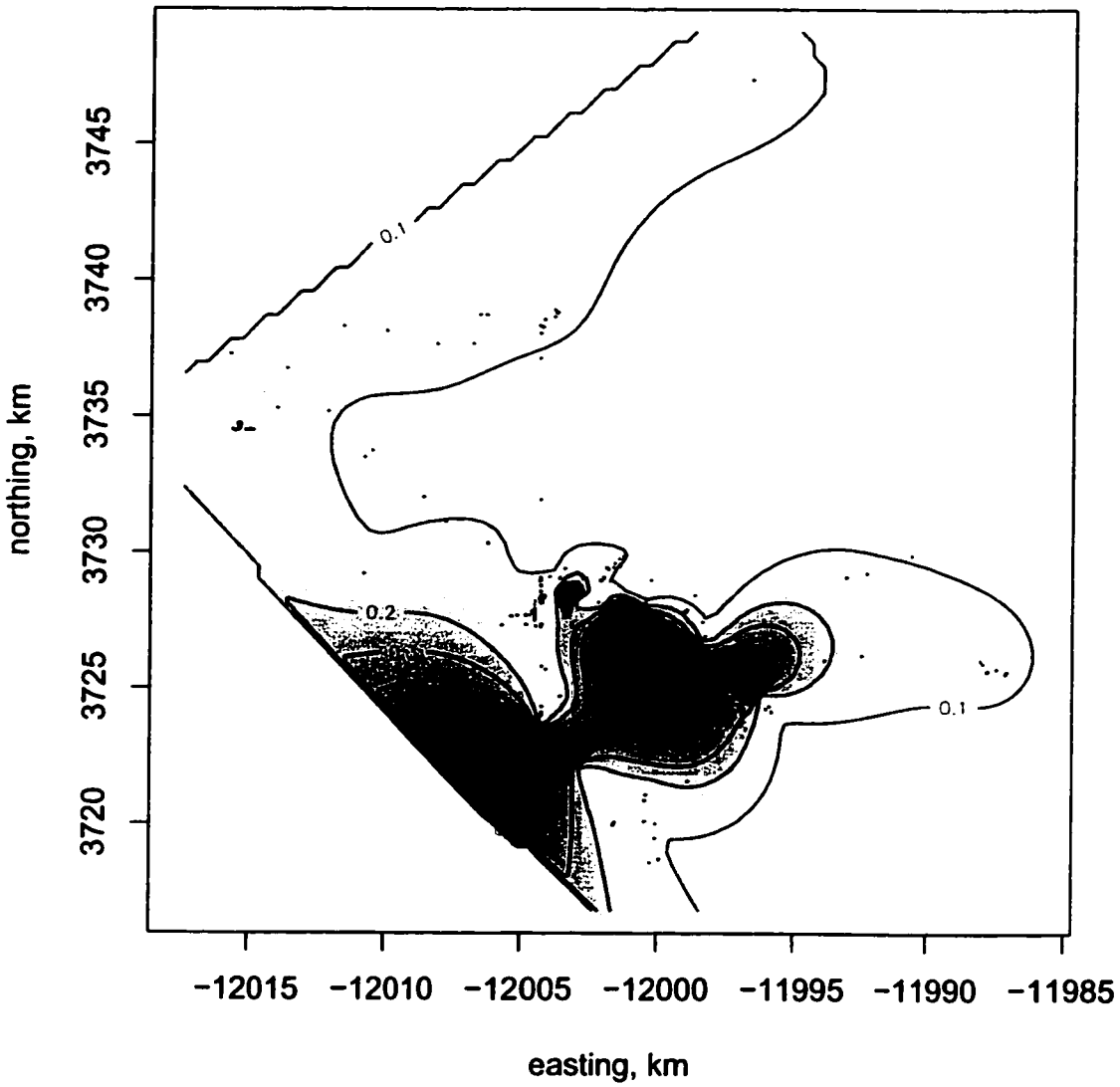


Figure 6.14: Kriging surface for user id 2's activity 3. See also figure 5.9.

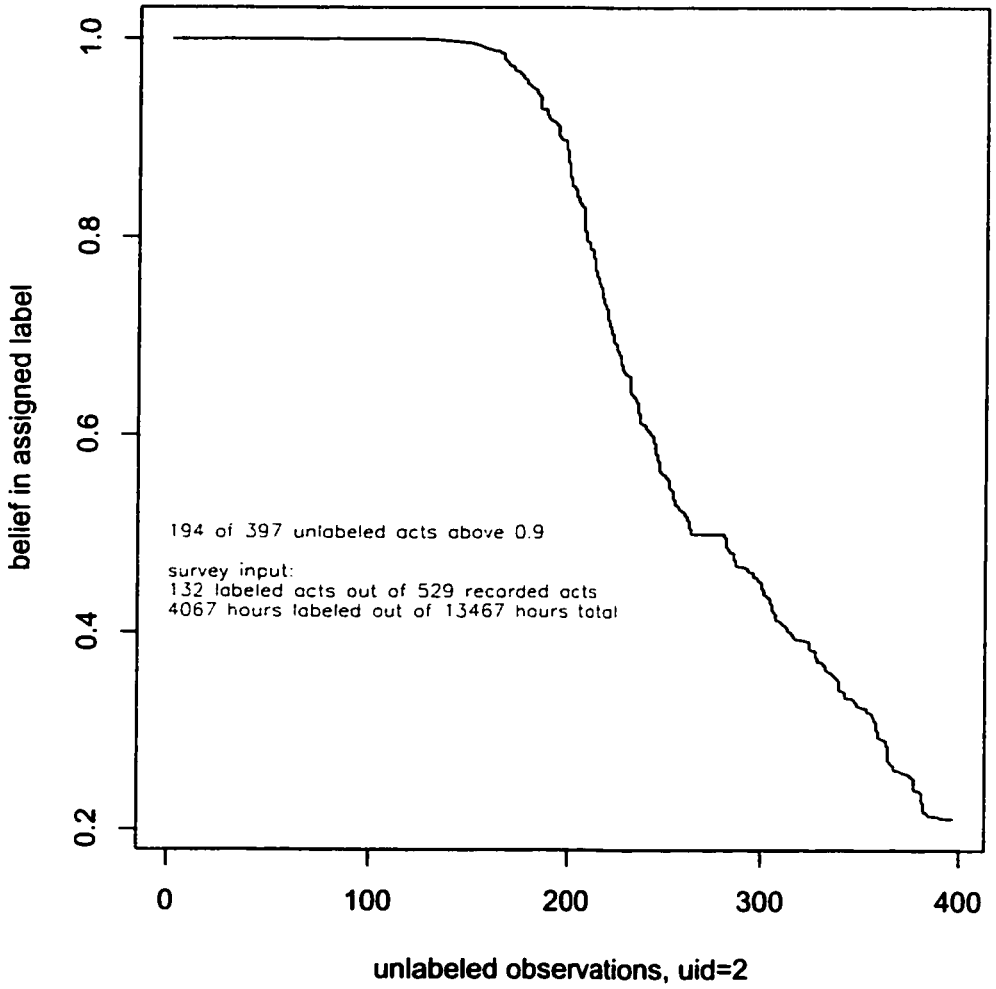


Figure 6.15: Probabilities of the most likely activity labels assigned to each unknown activity for user id 2, given the location of the activity and past survey entries.

large percentage of unlabeled destinations are so close to previously labeled activities that they are assigned an activity name with probability 1.

6.5 Conclusion

This chapter presents the results of a small pilot study exploring the use of a GPS device used in an activity survey. While the respondents quickly tired of entering data in the activity survey, they continued using the GPS data collection devices for much longer periods of time. One can imagine similar scenarios in real surveys, perhaps where respondents are instructed to respond to an on-line survey for only a subset of days, or only at the beginning and end of a survey period.

The GPS measurements provide accurate time and position data for the parking spot of the vehicle at the start of an activity, which approximates the actual time and place of the activity. This chapter demonstrates that it is possible to estimate activity-specific kriging surfaces for these data, using a semi-variogram estimated from all of the observed and labeled destinations. The kriging surface gives the expected frequency of events for each named activity at any given longitude and latitude. The combination of the frequencies from all kriging surfaces (one for each reported activity) can be used to generate a probability for the activity names at each point.

It was shown that this method can generate most likely activities that have a high probability, relative to other entries. The net conclusion is that activity destinations, at least for the four individuals observed, are *not* random spatial processes, which means that a small set of labeled destinations can be used with some confidence to estimate the activities at

unlabeled points. Further refinement of the technique is also possible by including the time of day of the activity, which was not considered here.

The proposed initial application of this technique is to further streamline the activity survey. Instead of presenting the respondent with a list of *all* past activities, appealing to the kriging surfaces as a synthesis of all past observations allows the survey to generate a list of only the most likely activities. This will reduce the respondent burden slightly by eliminating the visual clutter caused by irrelevant activity choices.

Another application that could result from further development of this technique is to augment a short activity survey with an approximate activity survey, based only on GPS data. This application would pair a GPS data collection device and an on-line survey, as in this pilot survey example. However, the activity survey would be designed to end prior to the GPS data collection period. The extended time period of GPS data collection (which is relatively painless to the respondent) can then be post-processed using the activity-specific kriging surfaces to determine the most likely activity at each destination. Those activities which are assigned with some high degree of certainty can be used to approximate an extended duration activity survey. This technique is applied in the next chapter when trying to identify activity patterns.

Alternately, the areas of low or medium predicted frequency can be the subject of focused questions. For example, suppose an on-line survey is designed to bracket a period of GPS data collection. The initial survey can collect activity data for all destinations visited, as with this survey. Then the second survey can be designed to prompt the respondent to fill in activity data for days which have high numbers of unknown destinations. Alternately it might be more important to further refine some area of space which has multiple kinds of

activities situated next to each other. In any case, the goal is to use the results of the kriging analysis to get the most useful information possible from a respondent, rather than wasting the respondent's good will answering questions about destinations which are already well known.

An interesting extension of this work that would have important ramifications for survey design would be to establish a relationship between the point pattern of destinations, the numbers of activities labeled via the on-line survey, and the fraction of unknown points that can be assigned an activity label with a high degree of confidence. As the point process becomes more random, the power of the kriging surface will diminish as the peaks become hills or slight ripples. As was shown in the comparison of user id 4 and user id 5 responses, a very small number of responses has a much greater impact when those responses are near all or most of the respondent's activities. But just gathering more activity data is not likely to guarantee a good match to every unknown point, as there are always new destinations or activities. Determining a good balance is the key to a cost effective survey design which gets the most mileage out of a respondent's limited patience.

Chapter 7

Identifying activity patterns within a long string of activities

7.1 Introduction

In recent years, computational biology has applied string matching and sequence alignment algorithms to the problem of determining DNA sequencing from observed snippets of DNA, RNA and amino acid proteins. This field has benefited tremendously from these algorithms. Bioinformaticists based their work on the fundamental assumption that "biologically meaningful results could come from considering DNA as a one-dimensional character string, abstracting away the reality of DNA as a flexible three-dimensional molecule, interacting in a dynamic environment with protein and RNA...." (Gusfield, 1997). As yet, we do not have a similar assumption in activity analysis. One of the goals of this chapter is to open up the discussion regarding the correct interpretation of sequences of activities, and activities themselves.

The main difference between this chapter and other applications of sequence alignment in activity analysis is that the data analyzed consist of very long, multiday sequences of activities. The sequence alignment methods are being used not to compare across people, but rather to ascertain patterns within the long sequence observed for each individual. As documented in previous chapters, the data are the result of an informal pilot survey, and as such no general conclusions can be drawn from the analyses. However, the data can be used to design and implement a new analysis technique, and to identify issues that would be well served by further data collection.

7.2 Important sequence alignment concepts

The basics of sequence alignment are covered in detail elsewhere (Cormen et al., 1990; Gusfield, 1997; Joh et al., 2001b; Wilson, 1998). The core aspects of sequence alignment are summarized in the following paragraphs. First, at its heart the algorithm is a dynamic program which computes some recurrence relationship between all of the members of the two strings, and as such, the full table must be computed in order to obtain a result (Cormen et al., 1990). For two strings of size n and m, the algorithm requires an $\mathcal{O}(nm)$ running time. Typically, n and m are roughly the same length. Computing the optimal alignment requires a back-trace through the table, taking an additional $\mathcal{O}(n)$ operations.

Second, the central recurrence can take several different forms. That used by Wilson (1998) and Joh et al. (2002) is a *differencing* recurrence, resulting in what is known as the edit distance between two strings. The relation assigns zero to matches, some positive number to mismatches, and (optionally) positive scores to space insertions and gap initia-

tions (described below). At each step of the recurrence, the total score is minimized. At the conclusion of the algorithm, the final value represents the minimum cost of operations to convert one string into the other. The back-trace step will generate the longest common subsequence—the longest subsequence of matched characters between the two sequences which generates the minimum edit distance.

An alternate recurrence relation is to compute the *similarity* between two strings. This is the usage of sequence alignment that this paper adopts. The similarity score increases with increasing matched activities, while the difference increases with increasing mismatched activities. Further, the way the recurrence is defined in this paper, the final result of the dynamic program gives the score of the optimal subsequence similarity, ignoring the effects of "edge" activities. An additional back-trace step gives the optimally aligned subsequence, as before.

The third fundamental characteristic of sequence alignment algorithms is that the basic operations for both types of recurrence are a match, a mismatch, and inserting a space. In some cases, as in Wilson (1998) and Joh et al. (2002), a change from any activity to any other costs the same as one insertion and one deletion operation, leading to the notion of an *indel* operation. This approach is not adopted here, as it is more flexible to allow the insert, delete, and change costs to vary with the objects being compared. In the bioinformatics literature, the DNA and protein weight matrices have been refined so as to represent as best as possible the true biochemical and/or evolutionary cost of substituting different elements of the alphabet in a sequence. No such level of refinement exists within activity analysis to date, and so assuming *indel*-type operations are valid is somewhat premature.

In addition to these basic costs, there is a gap cost associated with starting a gap (insert-

ing a space character next to an activity character) in the optimal subsequence. Spaces and gaps are similar to temporal wildcards—once created, they can be changed to anything. Imagine the mechanism that might cause the transformation of several short activities into a single long activity. The deletion of a single activity has some fixed cost, but deleting several activities to free up a block of time could be modeled to cost less than deleting and replacing each of the activities individually. Once a gap of any length is created, it isn't as expensive to add to that gap as it is to create another gap somewhere else in the sequence. The gap cost ideally should reflect the initial penalty for changing from a particular activity to this wild card state. In practice, the net effect of a high gap cost is to promote subsequences that are very similar to the original strings, with very long gaps, while cheap or free gap creation allows optimal matched sequences with many "holes" scattered throughout.

7.3 Use of sequence alignment in activity analysis

Wilson (1998) first introduced sequence alignment algorithms into the time use and activity analysis literature. He applied the standard bioinformatics program ClustalW to an activity survey taken in Reading, England in 1973. In the process of demonstrating some of the potential for sequence comparison algorithms, he also raised many questions about the applicability of these techniques to activity sequence analysis.

One such question stems from the usual notion that activities have multiple dimensions. In contrast, DNA sequences are lists of one-dimensional entities—the names of the nucleotides. Wilson (1998) worked with at most two dimensions, the name of an activity and its duration. These fit easily into a string representation if one incorporates duration

by using a time-slicing technique. Wilson (1998) discusses time-slicing, and notes that it has a side-effect of giving more weight to longer duration activities when comparing sequences. For example, a one-hour activity split up into 5-minute intervals would become 12 activities in a row. If activities are important to activity patterns for reasons other than their duration, then time-slicing will obscure important information.

To address other features (more measurement dimensions) of an activity, Wilson (1998) proposes using tree alignment techniques, or developing an expanded alphabet of activities. In the alphabet approach, an activity is defined not just by its name, but also by discrete identifiers for all of its other dimensions. In further work with sequence alignment techniques, Wilson and his co-researchers extended the biology-specific ClustalW program to a more general version, called ClustalG, to enable the application of the expanded alphabet approach (IATUR, 2001; Wilson, 2001; Wilson et al., 1999).

In a series of related papers in support of the *ALBATROSS* research effort (Joh et al., 2002, 2001a,b), Joh and his co-authors take another approach to applying string comparison techniques to a K-dimensional activity. Instead of developing multidimensional objects (as in Wilson's enlarged alphabet approach) or using a tree matching algorithm, they suggest breaking up the dimensions (name, duration, people involved, mode, etc.) and aligning K independent sequences, one for each dimension ascribed to an activity.

It is difficult to explain the differences between the two approaches without an example, and so the example given in Joh et al. (2001a) will be used. However, Joh et al. (2001a) did not adequately explain some of the assumptions behind the example. From a careful reading of the article, it appears that the following assumptions are in force.

g

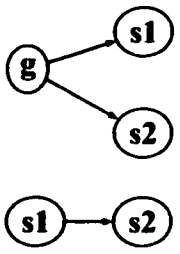
Activity 1
Work–Rotterdam–Car–Alone

s1

Activity 1	Activity 2
Shopping–Amsterdam–Car–Alone	Work–Rotterdam–Bus–With Others

s2

Activity 1	Activity 2
Shopping–Amsterdam–Car–Alone	Social Visit–Maastricht–Bus–With Others



chg (1) + chg (1) + delete Activity 2 (4) = 6 *or*
delete Activity 1 (4) + chg (1) + chg (1) = 6

chg (1) + chg (1) + delete Activity 2 (4) = 6

match Activity 1 (0) + chg (1) + chg (1) = 2

Figure 7.1: Comparison of multidimensional travel patterns via an expanded alphabet, after Joh et al. (2001a, figure 2a)

- operation costs: 1 insert = 1 delete = 1 mismatch = 1 change;
- all dimensions are of equal importance;
- a change of one aspect of the expanded alphabet representation is equal to a change in any other aspect;
- the ability to insert gaps is ignored; and
- the cost of an operation is independent of the activity or activity dimension being changed.

All of these assumptions could be modified, which would of course produce different results for the example (and invalidate the conclusions drawn from this example).

In figure 7.1, a single-activity target sequence, *g*, is compared against two other activity sequences, *s1* and *s2*. This example applies the concept of an expanded alphabet, and is used by Joh et al. (2001a) to point out an apparent short-coming of the method. In order to convert activity sequence *s1* into the target sequence *g*, one would either have to delete the first activity at a cost of 4 operations, and then change the mode and the people descriptions, at a cost of 2 operations, or else modify the first activity and delete the second. Both options give the identical result that *g* and *s1* are 6 operations apart. Similarly, to convert activity sequence *s2* into the target sequence *g*, one would change the activity name and location in the first activity at a cost of 2 operations, and then delete the second activity at a cost of 4 operations, again producing a distance of 6 operations between the sequences.

The same activity sequences are compared in figure 7.2, but this time using the simultaneous alignment of each different dimension of the activities. With the independent alignment of the different dimensions, it is possible to "shift" the different parts of activity sequence *g* to match exactly one of the two activities in sequence *s1*, resulting in a distance of 4, rather than 6. Joh et al. (2001a) claim that this is a desirable feature of their multidimensional encoding and analysis scheme, claiming that the smaller distance satisfactorily captures the intuitive closeness of *g* and *s1* relative to *g* and *s2*.

In this I feel Joh et al. (2001a) are not quite correct. While one might interpret *s1* as being closer to *g* than *s2*, the requirement that attributes must be reshuffled to create a "match" implies that *s1* should be farther away from *g* than a true match. To illustrate this point, consider two other sequences, *s3* and *s4*, as shown in figure 7.3. Sequence *s3* has an identical match in the second activity, and sequence *s4* has an identical match in both activities. In both cases, the multiple, independent alignment of dimensions results in a distance

g

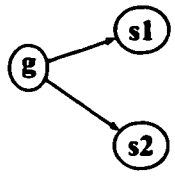
Activity 1
Work
Rotterdam
Car
Alone

s1

Activity 1	Activity 2
Shopping	Work
Amsterdam	Rotterdam
Car	Bus
Alone	With Others

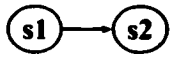
s2

Activity 1	Activity 2
Shopping	Social Visit
Amsterdam	Maastricht
Car	Bus
Alone	With Others



name align (1 del) + place align (1 del) + mode align (1 del) + people align (1 del) = 4

name align (1 del, 1 chg) + place align (1 del, 1 chg) + mode align (1 del) + people align (1 del) = 6

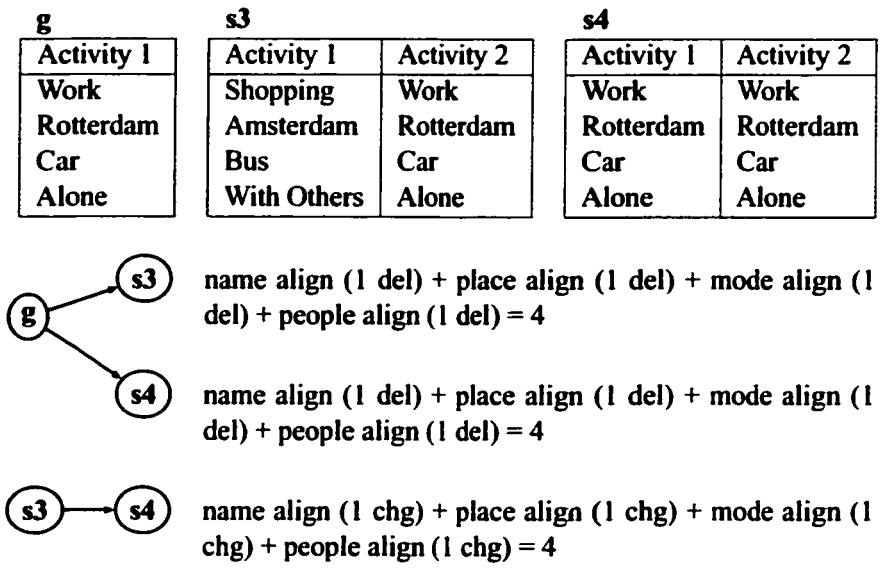


name align (1 chg) + place align (1 chg) + mode align (0) + people align (0) = 2

Figure 7.2: Comparison of multidimensional travel patterns via independent alignment of multiple dimensions, after Joh et al. (2001a, figure 2b)

of 4 operations, the same as that between *g* and *s1*. Rather than gaining specificity in this case, the multiple dimensional alignments have instead blurred the ability to distinguish the difference between pure matches (figure 7.3) and an artificially recombined match (*g* to *s1*). While the expanded alphabet approach would also fail to distinguish between *s3* and *s4* relative to *g*, this method puts *s1* further from *g* than the natural matching activities.

The idea of splitting up the dimensions of an activity and reshuffling them is also somewhat disconcerting. The authors have chosen to do this—and have chosen to ignore the within-activity dependencies—in an attempt to accommodate between-activity interdependencies. There is no theoretical reason to break up the characteristics of an activity. This is especially important given the preliminary findings presented in chapter 5 and chapter 6 that activities and destinations are closely related, and demonstrably dependent for each of the pilot survey participants. If the activity names, space, and time were independent of each other, then one could presume that splitting up these dimensions could be done "for



g

Activity 1
Work
Rotterdam
Car
Alone

s3

Activity 1	Activity 2
Shopping	Work
Amsterdam	Rotterdam
Bus	Car
With Others	Alone

s4

Activity 1	Activity 2
Work	Work
Rotterdam	Rotterdam
Car	Car
Alone	Alone

Figure 7.3: Lack of specificity in comparison of multidimensional travel patterns via independent alignment of multiple dimensions, after Joh et al. (2001a, figure 2b)

free." But knowledge of any of the location, time, and/or the name of an activity for a particular person changes the probabilities of the remaining variables, as shown in the various kriging surfaces plotted in section 6.4. In short, the multiple dimensions of an activity should not be treated as independent events.

Finally, the multiple dimension alignment technique does not allow for the specification of unique measures of the difference between the (multidimensional) activities themselves. The method cannot be used to state that *work—Rotterdam—car* is closer to *work-related—Rotterdam—bus* than it is to *shopping—Rotterdam—car*. Wilson's progress on using an expanded alphabet (Wilson, 2001; Wilson et al., 1999) *can* make these sorts of distinctions. The expanded alphabet approach relies on a look-up table to define inter-object differences (or similarity). This could be easily replaced with an even more general function call.

Joh et al. (2001a) and the related papers do make a good point that one should strive to capture aspects of sequential similarity when comparing sequences. While the features of

the sequence should be subordinate to comparing the individual activities themselves, there are some sequencing effects that cannot be captured by the standard sequence alignment algorithm. For example, looking again at figure 7.3, one would like to capture the idea that *s4* is closer to *g* than *s3* because of the multiple exact matches. And from figure 7.1, that *s1* is closer to *g* because all of the aspects of *g* do exist in the sequence *s1*, compared to *s2*. But these notions possibly can be captured by examining the multiplicity of same-cost matches that are possible between the sequences. Once again, being able to all a custom lookup table or a difference function that takes the activities in question as arguments is invaluable, and ultimately much simpler than the multiple dimension alignment method.

7.4 Analysis of streaming activity patterns

The most important implementation hurdle is that the recurrence relation speed and space requirements scale with the square of the size of the input data. If there are n activities at the start of a day, and if m activities are observed over the course of the day, then a full alignment of all the data against itself will take $\mathcal{O}((n+m)^2)$ compared to $\mathcal{O}(n^2)$ the day before. So it is imperative *not* to have to perform these full alignments continuously. Similar constraints are placed upon the space required for the algorithm, as each cell of the table must be maintained until the optimal backtrace is identified. Technically, new data received should only require the outer edge of the table to be recalculated. The inside portion (comparing the n activities already observed and processed) will remain the same.

Compare the analysis of streaming activity sequences to the *ALBATROSS* application of sequence alignment techniques (Arentze and Timmermans, 2000a). In the

ALBATROSS system, sequence alignment is used to compare generated tours to a database of known tours. While the core recurrence relation has been complicated with the inclusion of multiple dimensions, the fundamental sequences being compared are relatively short. Within the context of an on-line data collection tool where the sequences are growing longer all the time, an $\mathcal{O}(n^2)$ running time will eventually become unacceptable.

The sequence alignment routines have been partially integrated into the suite of tools accessible through the ANNE/Tracer web survey administrator interface. The eventual goal is to embed the sequence alignment process into the data collection process, so that results are determined as new data are received. At the moment, the implementation processes all of the observed activities for a respondent at once. So, with a click of a mouse, the survey administrator can process all of a respondent's observed activities, destinations, and travel to look for apparent patterns of behavior. Note that for a 100-day test sequence, the full alignment operation, when run in this "off-line" mode, fills up about 300 MB of RAM and hogs the CPU for long periods of time. In a way, running this program in off-line mode is like a self-imposed denial of service attack.

7.5 An algorithm to find patterns of activities

The general strategy taken here is to look for repeated subsequences of activities. Ideally, one would want the recurrence relationship to detect similarities rather than differences, and to exclude areas of obvious mismatch in favor of subsequences that are well aligned. In sequence alignment parlance, this is called the optimal local alignment problem (Gusfield, 1997, chap. 11). As with the global alignment problem applied by Wilson

(1998) and Joh et al. (2001b), there are typically multiple solutions that give the same similarity scoring for a given pair of sequences. Further, it is *not* the case that the optimal local alignment subsequence found for a pair of sequences is also the optimal for a group of sequences. In fact, the problem of finding the best local alignment for multiple sequences is NP-complete. An accepted heuristic to the global alignment problem in the biological world is to use the optimal subsequence pairing between two sequences, and then to look for other, non-overlapping subsequences that also have high similarity scores. While this technique will not identify *all* of the "good" subsequence alignments between two sequences, let alone identify the ones that are best for a global partitioning of the set, it is much cheaper than complete enumeration.

The problem of finding a set of activity patterns has differences from the biological case. Unlike DNA sequences, we can easily observe the correct ordering of activities, but that ordering is not guaranteed to repeat, and the string of activity sequences for each person will grow without bound. Aligning a sequence to itself to look for regions of overlap in the off-diagonal areas is a reasonable strategy for an initial starting point, but the algorithm should allow for new information to be received which might end up changing the results. This does not require recomputing the sequence alignment table for data that has already been processed, but it will required some flexibility in defining the activity patterns (the locally optimally aligned subsequences of activities).

The heuristic strategy proposed is that data be passed against the master sequence in chunks or slices that are as long as the expected pattern. Suppose one expects patterns to extend to at most a 24 hour period. Then using that period to define a slicing granularity, new data is only processed when the accumulated new activities span more than 24 hours.

With each new period of data, a very short sequence will be compared against a very long sequence. This should allow for a running time which will allow constant updating of the table with each elapsed period, rather than waiting for system resources to become available. Past passes through the data are stored in the database as a table, and in existing activity patterns.

Although strictly speaking a dynamic programming algorithm requires the full table to be computed to verify the solution, it may be okay to only compute the lower or upper triangle for two reasons. First, given that the sequence is being compared to itself, it is reasonable to expect some symmetry in the table—although removing the results along the diagonal might ruin this symmetry. Second, if there *are* repeated patterns, then by definition they should repeat. This means that while a particularly good pairing of activity sequences might be locked away in the upper triangle of the dynamic programming table, if the activity sequence is important, it should repeat itself, thereby making optimal matches more likely.

The details of the algorithm are as follows.

Step 1 Initialize parameters.

The parameters currently include include:

- the user id of the respondent being examined;
- the numerical id of the EDCU unit if more than one unit has been used and if only one unit's data is of interest;
- the expected time period for the pattern;

- the method for handing identified and candidate patterns in the recursion

Step 2 Retrieve activity observations from the database.

Since the data collection programs had various misfeatures and idiosyncrasies, the activities are given a thorough review, including an analysis of the travel in between activities for any "missed" activities due to the EDCU not shutting down.

Step 3 Construct a complete, sequential list of activities.

This list of activities includes the activity start and stop time, the location coordinates, and any names (possibly multiple) assigned to the activity by the user. In addition, the names assigned by the user may be augmented depending upon the settings of three parameters. If the *travel activity* flag is set, then each period of travel is used to create a travel activity in the sequence. If the *lon-lat proxy* flag is set, then each observed destination, including named destinations, is assigned an additional name consisting of the rounded longitude and latitude values. For example, the coordinates $-117.147323, 32.912110$ would be assigned a proxy name of "-117.147,32.912." Finally, if the *kriging estimates* flag is set, then those names assigned a probability of 90% or higher by the estimated kriging surface (see chapter 6) are used for each of the unnamed activities. The probability tolerance level can also be changed by passing a different value to the method.

Step 4 Examine travel records, and expand activity list if necessary.

After the list of activities is constructed, the travel data between activities is examined, and any possible stops (defined as a period of no movement for a minimum of 5 minutes) are inserted into the sequence.

In addition, if the appropriate parameter is set, the travel itself can be viewed as an activity. The reasoning behind this is that travel takes up a finite amount of time, and is one of the few activities that can be detected automatically by the GPS recording device. Therefore even if no other activity name has been entered, it still might be possible to compare activities based on periods of travel alone.

Step 5 Form an all-to-all lookup table of activity similarities.

After the complete sequence of activities, travel acts, and inserted acts is created, a lookup table is created containing all comparison costs. Since multiple activities are being allowed, and since activity durations are being handled directly as a weighting factor rather than with a time-slice approach, the table is quite large. It consists of all possible entries into any given activity slot, from a simple longitude latitude duration entry generated by the parsing step, to a multi-named activity entered by the respondent. Since in the best case all activities will have to be compared to all other activities at least once, and since I expect to make many such comparisons in a recursive search for a good pattern, it make sense to calculate the distance matrix up front rather than compute and store it as the algorithm proceeds.

Note that the fluid nature of an on-line survey can be tricky. If the survey respondent goes back and edits any past responses, then this operation not only throws off one of the cells in this distance matrix, but it also completely invalidates the alignment process. Dynamic programming requires that every cell must be visited, as it might be in the optimal path. If a cell changes, then the table must be recomputed starting from that cell. For this reason, examining all of the activities at the start of the algorithm can identify any past

entries that might require recomputing, by noting if any past information has been edited since the last update of the algorithm.

Step 6 Slice the complete sequence into candidate patterns of total duration equal to or greater than the time period parameter.

The next step is to split the complete sequence into segments that are equal to the specified length of the pattern. If one is searching for a 12 hour pattern, for example, the sequence is broken up into chunks that are at least 12 hour long. I say at least because it is quite likely that there are activities that will exceed the pattern length. These are a sequence of length one.

Step 7 Execute the dynamic programming recurrence for each candidate sequence against the full sequence.

Each sliced period is passed against the full sequence, less the candidate sequence itself. In an off-line application, this will compute an entire horizontal block of the dynamic programming table, less a small rectangle centered about the diagonal where the pattern is not matched to itself. Instead either a single space character with zero duration, or a space character with duration equal to that of the slice is inserted. The actual behavior can be switched by the a parameter passed to the initialization routine. The default is to insert a blank activity with a duration equal to the slice period.

The standard recurrence relation for optimal local alignment described in Gusfield (1997) has been implemented in Perl. This recurrence relation includes the ability to assign a cost to gap creation, since the identification and removal of long patterns of activities is

similar to the biological motivation for dealing with gaps. Since I am allowing multiple names, the standard comparison of names has been replaced by a functional lookup, which compares name objects. In this case, those name objects are simply indices that point to the aforementioned distance matrix's rows and columns. These entries could just as easily be pointers to more complete objects.

Step 8 Store the optimal subsequence alignment result for each candidate sequence.

When the optimal subsequence alignment is complete, the solution is stored in a result pointer. This pointer identifies the cell with the highest score, which is equivalent to the end point of the optimal subsequence at which the two sequences are most aligned (see Gusfield (1997) for a proof and more detail on the algorithm). To recreate the optimal subsequence, one simply has to crawl along the path from this pointer until the score drops to zero.

Step 9 Choose the best matching pair of sequences to cluster next.

In offline mode, this scanning is repeated for each slice. After all of the best subsequences have been computed for each slice, that best subsequence is declared the primary activity pattern and the algorithm enters its clustering stage. The sequence that matched the candidate pattern is removed from further consideration and is assigned to the first cluster along with the sequence that generated the match.

Step 10 Recompute the dynamic programming table for the best candidate pattern with the matched sequence removed from the master sequence, and again look for the best alignment.

The matching sequence is then passed along the original sequence again, without the matched sequence, and is assigned a new optimal subsequence. Because the table prior to the extraction of the sequence is the same for this second pass, the algorithm simply starts off from the activity prior to the gap that was inserted for the matched sequence. This step could be done for all sequences, but it is postponed until it is absolutely necessary (say if the "best" pairing contains a previously clustered off sequence.)

Step 11 Go to step 7, until all subsequences of the master activity sequence have been matched.

The process repeats until all sequences have been matched, or until a stopping criterion limiting the lower bound for the sequence similarity score is met.

In on-line mode, which has not yet been fully implemented, the new slice of data would be passed against the full sequence. Its best match with an opposing subsequence would be assigned a score, as above. Obviously the opposing sequence would have already been assigned to an existing cluster, based on some other match score. If the new score is better than the old score, then the opposing sequence is removed from its old cluster, and a new cluster is started. The new slice is then passed again against the sequence from the match point forwards, and a new optimal subsequence is calculated, pointing to a new opposite subsequence. Again this subsequence will already belong to another cluster, and it is removed only if the new score is better than the old score.

The recursion stops when there are no more sequences left to match, or when the best match possible by the new sequence fails to extract a sequence from an existing cluster. If this stopping condition occurs for the first match, then the new sequence is added to that

existing cluster.

In practice, all of this is hidden behind a simple interface. The website sets up the alignment object with one call, starts the sequence alignment process with a second call, then accesses a moderately readable, HTML-formatted result with a third call. The resulting HTML page can be very wide and very long. For example, one run of the algorithm on a 100-day sequence of activities spanned seven screens wide at small but legible font size. The intent of the web interface is to allow the analyst to experiment with the impact of different parameter values, although in practice this will require a much faster web server.

7.6 Analysis of pilot data

One of the main difficulties with analyzing the long duration Tracer pilot data is that the participants quickly tired of entering activity data. Step 3 of the algorithm described how three different methods can be used to augment the raw sequence of activity names. First, following the method developed in chapter 6, estimated activities can be inserted for many of the GPS-recorded destinations by accepting all activity names assigned a probability greater than 90% (see figure 6.6, figure 6.8, figure 6.11, figure 6.15). The second alternative is to recognize the linkage between destinations and activities by assigning each activity a longitude-latitude proxy name. Specifically, as figure 6.2 and similar figures showed, destinations observed at the 0.001 level of precision for longitude and latitude tend to taper off over time. Finally, even without any modifications to the activity names, the inclusion of *travel* activities may help to identify similar daily patterns. The impact of these modifications in various combinations are examined below.

The sequence alignment process generates a very long, very wide HTML table for each person. Each pattern is described by listing the full seed sequence, followed by the various optimal subsequence alignments with all of the other periods that belong to the group. These tables are not reproduced here. As demonstrated by Wilson (2001), displaying long sequences of activities is difficult to do well, and is effort wasted on the data gathered in this pilot survey. If the sequences or patterns were important or demonstrated some generalizable characteristics, then a listing of the matched sequences would be warranted.

Instead only a simple example sequence is shown in detail. Table 7.1 shows the effect of changing name-augmenting strategies on the similarity score of an aligned pair of 24-hour sequences. Three options are explored. The first flag governs whether longitude and latitude are used as a proxy name for all of the activities. The second sets whether or not the results of the kriging surfaces are used to name unlabeled destinations. The third flag controls the inclusion of travel activities in the sequence. For the simple example chosen, the kriging flag has no effect, and so its entry is omitted.

The pattern used in table 7.1 is simple, but it demonstrates many of the variables that affect the sequence similarity score. The first pattern match results from the raw sequence. The first sequence has double names for the two labeled activities, while the second sequence has only single names. The mismatch of the third activity, indicated by a space character in both sequences, results in a small drop in the similarity score, roughly equal to $31/2$ minutes of mismatch penalty. The time-weighting heuristic weights matches by the shortest overlapping time, and mismatches by the longest mismatching time, so as to prevent alignments which promote long to short matches and mismatches. The various activity and mismatch durations are approximately the same, but they would not be if the

Table 7.1: Impact of parameters on a single pattern match. Note that no effect is seen from the insertion of kriging names because all destinations were already labeled. The "_" character indicates a space inserted into the sequence for an unknown or unmatched activity.

lonlat proxy	travel acts	activity sequence			
0	0		Work; Lunch 421.40 min	– 3.45 min	Home; dinner at home 967.72 min
			252,840	252,619	790,809
			Work 551.90 min	– 3.68 min	dinner at home 896.98 min
0	1	GPS Travel 4.18 min	Work; Lunch 421.40 min	GPS Travel 3.45 min	Home; dinner at home 967.72 min
		1,630	254,470	256,540	794,730
		GPS Travel 2.72 min	Work 551.90 min	GPS Travel 3.68 min	dinner at home 896.98 min
1	1	GPS Travel 4.18 min	Work; Lunch; 33.648,-117.839 421.40 min	GPS Travel 3.45 min	Home; dinner at home; 33.638,-117.838 967.72 min
		1,630	507,310	509,380	1,585,760
		GPS Travel 2.72 min	Work; 33.648,-117.839 551.90 min	GPS Travel 3.68 min	dinner at home; 33.638,-117.838 896.98 min

time weighting heuristic were to be reversed.

The second row shows the effect of inserting travel activities into the sequence. The unknown activity from the first alignment is discovered to be a travel activity in both sequences. The small negative mismatch penalty is replaced by a positive match score. Further, both sequences are preceded by a travel activity, which adds another small boost to the similarity score, and increases the length of the alignment. The third row adds the lon-lat name to each activity (but not the travel activities). Since the locations match, the sequences are still aligned, with the net effect be an approximate doubling of the similarity score due to the additional matched activity name. If the location did not match, the score would be the same as the second row.

Table 7.2 shows the net impact of the augmenting the activity names on the resulting numbers and sizes of activity patterns. Since inserting the lon-lat proxy names adds a new name to every activity, it tends to have the most significant effect. However, it is not necessarily the most desirable effect. In addition to driving up the similarity scores for each matched sequence, it also can increase the number of clusters that are generated. This happens for two reasons. First the clustering heuristic attempts to create groups with proximate similarity scores. For very long duration matches, such as that shown in table 7.1, the addition of the lon-lat proxy name effectively doubles the similarity score for some pairings, but not necessarily for all parings in a set. As the kriging surfaces of chapter 6 showed, activities are clustered in space but aren't necessarily located at a single point. Therefore it is entirely probable that adding the lon-lat proxy will split up a group of which contains the same activity names into as many groups as there are locations. Second, if sequences only match based on the proxy, perhaps because the activity name is missing, then they will

Table 7.2: Differences in patterns arising from sequence alignment by varying different parameters.

user id	lon-lat proxy	kriging names	travel acts	number of patterns	matched days	$\sum$ similarity scores
2	0	0	0	5	26	16,988,615
2	0	0	1	5	63	17,736,291
2	0	1	0	6	38	23,092,935
2	0	1	1	5	66	23,800,184
2	1	0	0	12	50	27,140,513
2	1	0	1	14	60	27,167,957
2	1	1	0	12	52	30,382,444
2	1	1	1	14	62	30,422,208
4	0	0	0	1	5	3,150,556
4	0	0	1	4	70	3,402,589
4	0	1	0	4	48	24,297,439
4	0	1	1	5	78	24,365,246
4	1	0	0	15	81	103,480,577
4	1	0	1	15	87	101,308,356
4	1	1	0	13	84	111,804,563
4	1	1	1	14	88	108,911,801
5	0	0	0	1	1	10,736
5	0	0	1	3	30	675,963
5	0	1	0	1	7	569,887
5	0	1	1	4	29	1,151,356
5	1	0	0	11	29	17,843,908
5	1	0	1	11	28	16,424,606
5	1	1	0	11	29	17,986,057
5	1	1	1	10	30	17,056,537
6	0	0	0	4	17	5,539,778
6	0	0	1	4	45	10,875,803
6	0	1	0	8	29	10,181,495
6	0	1	1	8	43	15,544,601
6	1	0	0	7	43	29,034,684
6	1	0	1	8	48	34,440,527
6	1	1	0	8	41	29,001,212
6	1	1	1	10	46	33,855,576

create a separate group.

Finally, adding the lon-lat proxy can also create undesired matches. Looking again at table 7.1, if the first activity in the second sequence had been something like "recreation" instead of work, the insertion of the location proxy would have created a match in the first activity where before there was a mismatch. While the strength of the match (the similarity score) would only be half as strong as the one shown, it still seems better to rely upon the kriging surface to estimate the similar activities by the relative locations of the destinations.

The kriging flag generally increased the numbers of patterns and the numbers of matched days over the raw sequences, but not necessarily over adding travel activities alone. In most cases the net effect of adding the kriging flag and the travel activity flag separately was the same as adding the flags together. The effect of the lon-lat proxy was less stable. For the two sets of observations with very low numbers of labeled activities (user ids 4 and 5), the lon-lat proxy tended to be the dominating force behind the matches versus mismatches. Therefore there are tremendous jumps in the numbers of patterns and the absolute similarity scores when this flag is added. For the other two sets of observations, the insertion of the flag tends to differentiate between sequences. Rather than be the only reason for matching, in many cases the lon-lat proxy adds another factor to differentiate otherwise identical scored alignments. Thus it is probably a good idea to insert this flag when the respondent has been conscientious about reporting activities, but not such a good idea otherwise. Since in a practical survey one should expect user fatigue rather than consistent attention to detail, the net recommendation should be to drop this flag. Spatial effects are better handled in other ways.

7.7 Conclusion

In terms of practical advances, this chapter has extended the sequence alignment technique to analyzing long sequences of activities. An algorithm has been developed and implemented for processing activity data. This algorithm has been implemented within the ANNE/Tracer data collection system, which has the ability to continuously collected travel and activity data. The sequence alignment technique advocated is similar to Wilson's expanded alphabet approach, although the alphabet in this case are programmatic objects with uniquely defined distance functions, rather than simple letters. The approach taken by Joh of splitting up the activity dimensions is not used because it appears to be incorrect, offers no advantages within the context of this application, and is limited to using fixed, uniform insertion, deletion, space insertion, and gap creation costs. As a theory of activity sequences is developed and expanded, it will increasingly important to be able to modify the space and gap costs.

Several areas still require significant effort before this technique becomes a practical tool. First, the large HTML tables are very difficult to use. One possible solution is to develop activity-specific icons and graphics, so that more information can be packed into a smaller display. It is distracting to have to scroll left and right to see the full extent of a sequence. In the same vein, the patterns generated by the alignment require individual inspection in order to interpret them. The goal of this process is to be able to automatically extract meaningful patterns from within a person's activity sequence. Capturing the "meaningful" part of this is difficult.

Second, the process so far has only searched for patterns that are contained within a

24-hour period. A more realistic approach would be to expect patterns of several different periods to overlap. For example, a 24-hour pattern might consist of working and meals, while a weekly pattern might contain shopping trips or recreational activities. Shorter patterns might exist for navigating particular circumstances, such as running quick errands en route to another activity. Devising a method for efficiently finding patterns at different length scales are the subject of future work.

Finally, another topic of future research is to attempt to merge the sequence alignment approach with the point-process techniques described in chapter 5, by either inserting time as a third dimension, or analyzing activities as a point process in time, as Cressie (1993) does in one of his examples. This analysis would require true activity names, or the estimated names from the kriging surfaces of chapter 6, rather than the pseudo longitude-latitude names used in this chapter, as longitude and latitude are guaranteed to vary perfectly with themselves! The goal integrating the point process analysis would be to identify activities and time periods where strong clustering or possibly even strong temporal regularity is apparent, as well as other activities and periods that are closer to a random process. This might help filter out noise and possibly define the shape or length of the activity sequence patterns, which could in turn improve the search for similar subsequences.

Chapter 8

Implementing distributed data collection and peer-to-peer data sharing

8.1 The potential of wireless data collection

The previous chapters have focused on collecting and analyzing long-term GPS and activity data. The conclusion of chapter 4 suggests that the most likely use for the propose survey system is to sandwich two periods of activity survey around a longer period of GPS data collection. Then chapter 5 observed that the activity destinations are highly clustered in space, and chapter 6 proposed a mechanism for linking activities and places based on the traveler's own responses. Chapter 7 showed that it is possible to detect patterns of daily activities within long sequences of GPS data augmented using the kriging technique of chapter 6. This chapter turns from the analysis of the individual to examining all of the four pilot survey participants together. The perspective is *not* to aggregate activity names or destination point patterns, but rather to observe areas where information might be usefully

shared between the individuals.

One of the paradigm shifts that the EDCU should produce is to view telematics technology as creating mobile content producers, not consumers. Too often the terms of the debate over traveler information systems and in-vehicle telematics are framed by commercial content providers who hope to monetize the traveler as a consumer. The streams of GPS data flowing out of EDCU-equipped vehicles are unique, original sources of information. The EDCUs are collecting information on real traffic conditions, preferred routes, and other travel information, as well as the destination point patterns studied in prior chapters. Collecting and aggregating this information could be useful to the traveler, allowing her or him to study accurate measurements of travel times and speeds rather than relying on often erroneous estimates of the same. In addition, this information could be useful to *other* travelers. Section 8.2 documents the destinations and travel paths of the four pilot survey volunteers, and discusses the benefits of information exchange.

Mechanisms for exchanging data between individuals are explored in section 8.4. The easiest solution is to maintain a centralized database, but travel destinations are highly personal, and should be kept private and under the travelers' control. Section 8.4 also discusses how the Tracer system can be decentralized, and some of the modifications that will be necessary to make the wireless data collection cheaper to the individual, and to make the sharing of information over the Internet practical. Finally, the current National ITS Architecture (ITERIS, 2002) is reviewed in section 8.5 for evidence of and ability to integrate peer-to-peer information exchange.

8.2 Overlapping destinations and paths

There isn't much point to sharing information if each traveler already has perfect information. As there are no accepted sources for acquiring perfect information, one would have to suppose that the mechanism for becoming informed is experience—knowledge gathered from past trips and destinations. Figure 8.1 shows the various destinations recorded for each of the four pilot survey volunteers. Note that the destinations for user id 5 are slightly different than prior figures in that only destinations greater than 10 minutes were allowed, in order to filter out the false stops generated by the EDCU transmission problem. The different point sets have areas of overlap—most notably the area surrounding UC Irvine—as well as areas of unique destinations. Each person has a different but overlapping set of travel experiences, and therefore a different level of knowledge about the travel and activity space.

While the pilot survey is by no means an exhaustive record of all destinations and all paths taken by the four vehicles, it does capture a great deal of recent history. If one presumes that the degree and applicability of travel and activity knowledge are directly proportional to the frequency and recentness of a visit to an area, then the non-overlapping portions of figure 8.1 represent areas where one traveler has certainly more recent and possibly more complete information than the others.

A more convincing diagram comes from examining the travel segments. Figure 8.2 shows the average travel speeds at all points recorded and transmitted to the base station for user id 6. The average was calculated very simply by summing up the speed observations at all repeated longitude and latitudes and dividing by the number of observations at

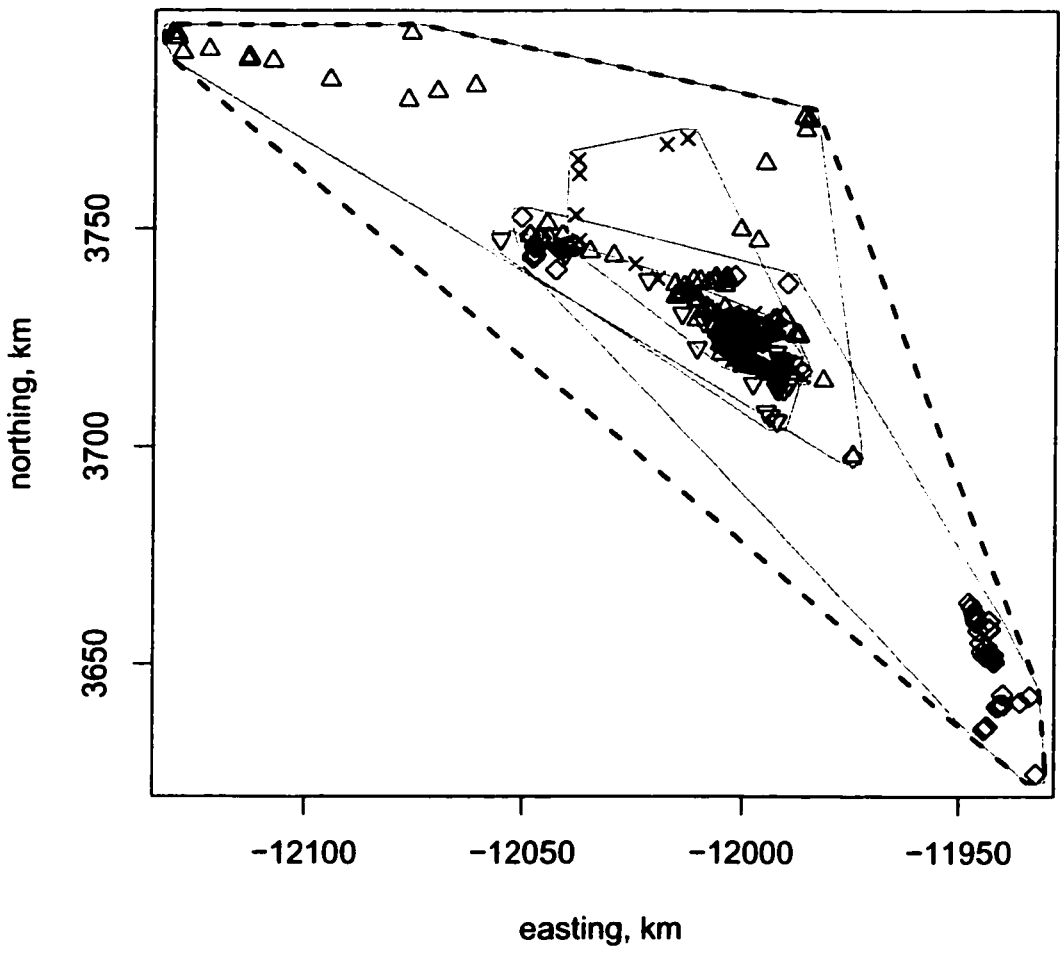


Figure 8.1: All destination points for all vehicles with durations longer than 10 minutes (see text). All vehicles have many destinations in the vicinity of UC Irvine, as well as unique information about other destinations and routes.

that point. More complicated calculations of average speed are possible, for example those that take into account the speeds of each trip rather than viewing the GPS records as separate measurements. Nonetheless, it demonstrates that even with this coarse computation, different average speeds can be observed over time (and visualized in a plot) for different routes.

In the same travel area, figure 8.3 plots the average speeds observed by all four Tracer vehicles. Obviously a great many new links are added to the plot. On the one hand, one would expect that a person would have a feel for the expected speeds on various links by knowing what kind of facility it is (freeway, arterial, and so on). However, one can never be sure about the expected volume on a link, until one had driven on it a number of times and has built up a body of experience about the conditions on the link at different times of day.

What figure 8.3 shows is that user id 6 could learn quite a bit from the information collected from the other 3 vehicles. The speeds shown are the averages over all observations, but if each user had access to the full set of travel observations, queries could be set up by the time of day and the day of the week, and along an expected path or set of paths. Obviously this is not as good as full information, but it is the next best thing—broad practical experience.

8.3 Using Tracer to share traveler-collected information

As it is implemented, there are no provisions within the Tracer system or the on-line survey to allow individuals to examine other travelers' destinations or travel patterns. Within

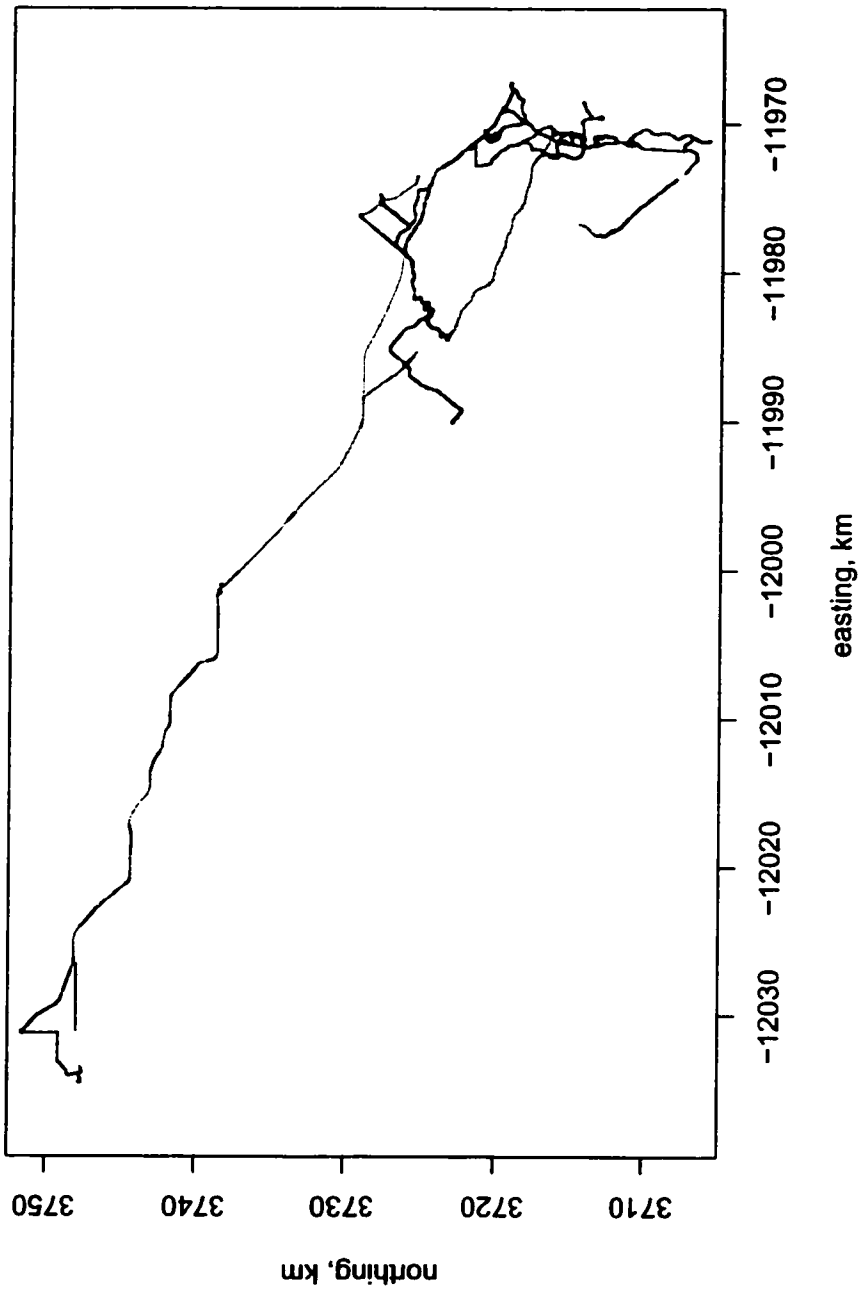
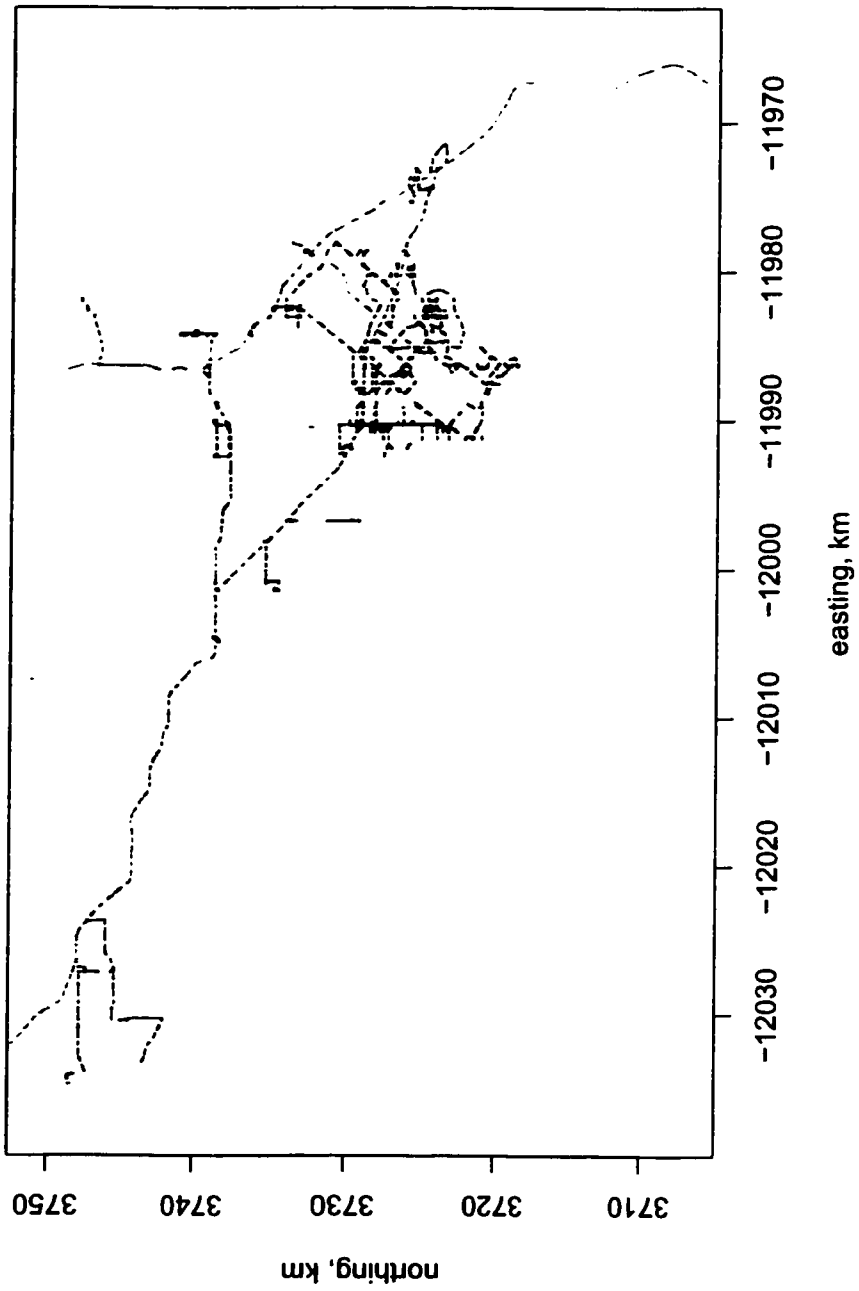


Figure 8.2: User id 6 observed average travel speeds at all travel points. The dark points represent slower average speeds while the lightest points are average speeds greater than 70 mph. GPS records with speed=0 are not included.


northing, km
3710
3720
3730
3740
3750
easting, km
-12030
-12020
-12010
-12000
-11990
-11980
-11970

the context of an activity survey such functionality is not necessary. However, moving beyond the survey, it is possible to implement a "publish" feature, in which participants in the system could tag either individual trips or trips that fall into broad categories as publishable. While preserving each user's privacy is especially important when travel monitoring is involved, releasing time and speed data on common arterials and freeways would perhaps be acceptable, as long as the exact criteria could be controlled by the person sharing the information.

Once the data is published by the traveler, then it becomes available to the other travelers in the system for queries. The current mechanism for queries within the survey web site is a very simple plotting tool which builds a plot based on checkboxes corresponding to the previous entries of the respondent. So for example, one click on "work," "meal," and the "or" checkbox to see a plot of all trips to and from activities that are labeled "work" or "meal." If other travelers' link speeds in and around these destinations are included in this plot, then one might see a more complete picture of route choices.

8.4 From centralized Tracer towards decentralized information exchange

The main barrier to implementing data sharing within the on-line web site is that the data are collected centrally. Privacy will be an increasing concern as wireless devices become more common. Travelers should be suspicious of other people looking over their electronic shoulders, and they should have ultimate control over what is done with infor-

mation that they themselves have collected. The best way to do this is to decentralize the Tracer system and its on-line web site component. While this is not yet an option for most of the traveling public, the steady rise in cable and DSL modems means that more and more people have the ability to save and retrieve their own data on their own web server. This section comments on some of the work that will be necessary to allow this decentralization to occur while still maintaining the ability to query all travelers' published experiences.

As far as the hardware cost is concerned, the natural growth rate of computer technology should make the required hardware essentially free of charge. The current incarnation of Tracer is hosted on computers that were cutting edge two or more years ago. The hardware was available for free to the project in that it was obsolete. A Pentium III-400 is incredibly slow compared to today's $200 beige box computers, but with sufficient RAM and disk space it is perfectly adequate for running a small website. The EDCU is currently rather expensive to obtain, but the rising popularity of GPS-enabled devices means that the prices should fall to under $100 soon. The generic GPS logger probably will not have the ability to transmit data wirelessly, but given the terrible price to speed ratio of CDPD service, it is probably best to expect that individual Tracer users would prefer to hand carry their data to their PC, or rely on local area wireless options.

Once the travel data is captured by the in-home Tracer unit, it is ready to be displayed with the on-line web site. The user would add notations as with the current version, and then choose what data to publish. Unlike the current system, publishing data is not as simple as setting a "publish" flag in a common database. The published data must really be publicized to the broader Internet. Similarly, the traveler must be able to find other Tracer users and their published data.

The best way to accomplish this is to rely on web standards, specifically XML and related technologies. A document type definition (DTD) will need to be designed which will allow users to know what to expect when they request data. Similarly, a technology call resource description format (RDF) has been developed which allows websites to succinctly describe their latest headlines and content. These technologies are being actively used and developed by the weblogging community, people who I view as the early adopters of the Tracer system in its web publishing incarnation.

There is little benefit to be gained by the initial adopters and publishers of travel information. As more and more people set up their own Tracer systems, of course jumping on the bandwagon is likely to provide much more information than the user will ever contribute. But at the beginning there are likely to very few network effects. This is why I see today's on-line diarists and webloggers as the natural early adopters of the system. These individuals have already demonstrated that they are willing to publish content regardless of how many people are reading their material. The Tracer system will bring the real world to the fingertips of the weblogger. Instead of typing text and hyperlinks, the user can add longitude and latitude coordinates, and links to maps. And of course some individuals might even compulsively try to offer the most complete sets of trip data, in the hopes of becoming the launch point for all travel-related queries.

Finally if and when Tracer and similar systems become widely used, the most likely mechanism for transferring data from the vehicle to the home computer is likely to be a wireless local area connection, such as the 802.11b standard or its successors. This means that many vehicles will be traveling with local area wireless devices and GPS data collectors, with perhaps a local cache of historical information relating to the day's planned

travel. A clear area for future research is to explore how one would get vehicles to share this information with each other as they are traveling, rather than waiting for the trip to be completed.

8.5 Review of peer-to-peer components in National ITS architecture

The National ITS Architecture (ITERIS, 2002) describes a centralized system, in which travelers and their vehicles are largely viewed as end consumers of information. The overall architecture is shown in figure 8.4 with the poor image quality a result of using the official architecture image files. Although it appears that the architecture encourages vehicles to talk to each other over vehicle to vehicle wireless connections, in fact the architecture specifies that these connections are to be used for automated highway systems and automatic vehicle control. The official summary descriptions of the three relevant subsystems are excerpted below from ITERIS (2002).

Vehicle subsystem This subsystem resides in an automobile and provides the sensory, processing, storage, and communications functions necessary to support efficient, safe, and convenient travel by personal automobile. Information services provide the driver with current travel conditions and the availability of services along the route and at the destination. Both one-way and two-way communications options support a spectrum of information services from low-cost broadcast services to advanced, pay for use personalized information services.... Ultimately, this subsystem supports completely automated vehicle operation through advanced communications with other vehicles in the vicinity and in coordination with supporting infrastructure subsystems....

Personal information access subsystem This subsystem provides the capability for travelers to receive formatted traffic advisories from their homes, place of work, major trip generation sites, personal portable devices, and over multiple types of electronic media. These capabilities shall also provide basic routing information and allow users to select those transportation modes that allow them to avoid congestion, or more advanced capabilities to allow users to specify those transportation parameters that are unique to their individual needs and receive

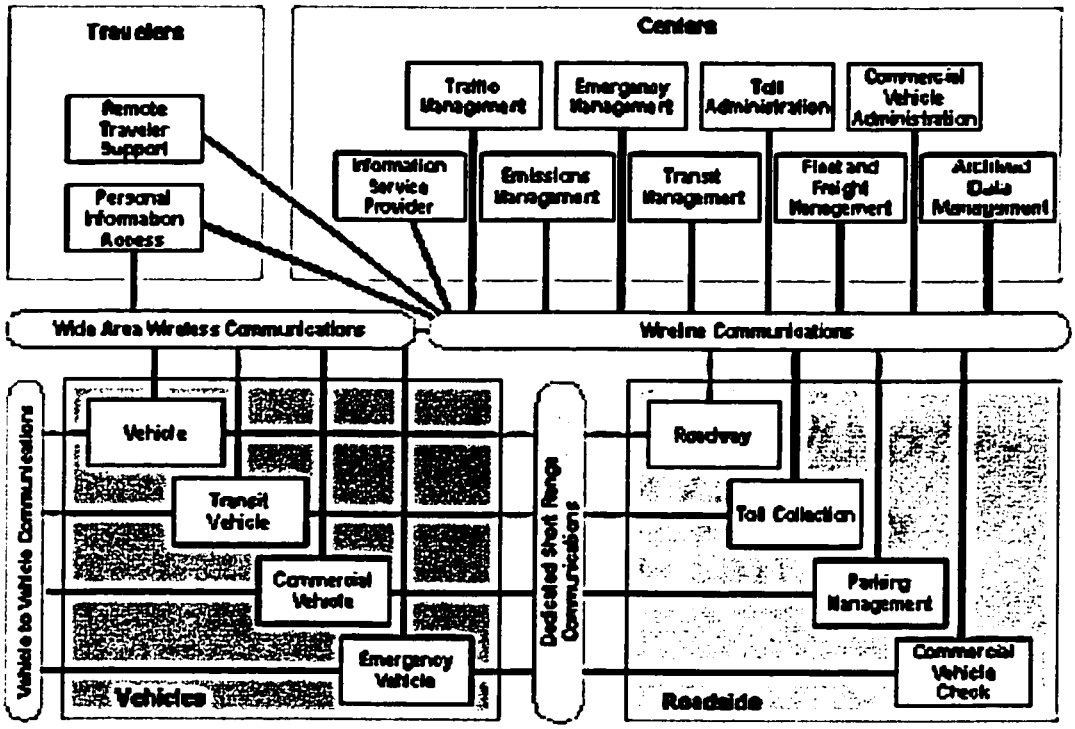


Figure 8.4: The National ITS Architecture usage diagram. From ITERIS (2002).

travel information. This subsystem shall provide capabilities to receive route planning from the infrastructure at fixed locations such as in their homes, their place of work, and at mobile locations such as from personal portable devices and in the vehicle or perform the route planning process at a mobile information access location. This subsystem shall also provide the capability to initiate a distress signal and cancel a prior issued manual request for help.

Remote traveler support subsystem This subsystem provides access to traveler information at transit stations, transit stops, other fixed sites along travel routes, and at major trip generation locations such as special event centers, hotels, office complexes, amusement parks, and theaters. Traveler information access points include kiosks and informational displays supporting varied levels of interaction and information access....

The diagrams associated with the above systems support the conclusion that the ITS architecture does not acknowledge peer-to-peer information sharing. Specifically, figure 8.5 shows personal information requests routing to map update providers, information service providers, emergency management, and transit management. Similarly, vehicle requests are routed to those resources, as well as to the roadway subsystem, and to other vehicles. However, the other vehicles link content only contains vehicle to vehicle coordination, and the vehicle to roadside link contains only messages pertaining to automated highway systems and vehicle probe data.

The National ITS Architecture may still be transformed by market forces. Peer-to-peer communication and distributed data collection and storage will be driven from the bottom up, as travelers adopt these new technologies for other reasons. The main impediment is the focus on vehicle to vehicle control messages, rather than traveler to traveler information passing. In the short term there will be a limit to the available bandwidth for vehicle to vehicle communications via local area wireless networking. The ITS focus on vehicle to vehicle control is likely to fill the channel without leaving any room for other, consumer-driven applications. This is especially true when there is a centralized information service provider (ISP) function described, linked to a rational revenue model based on premium

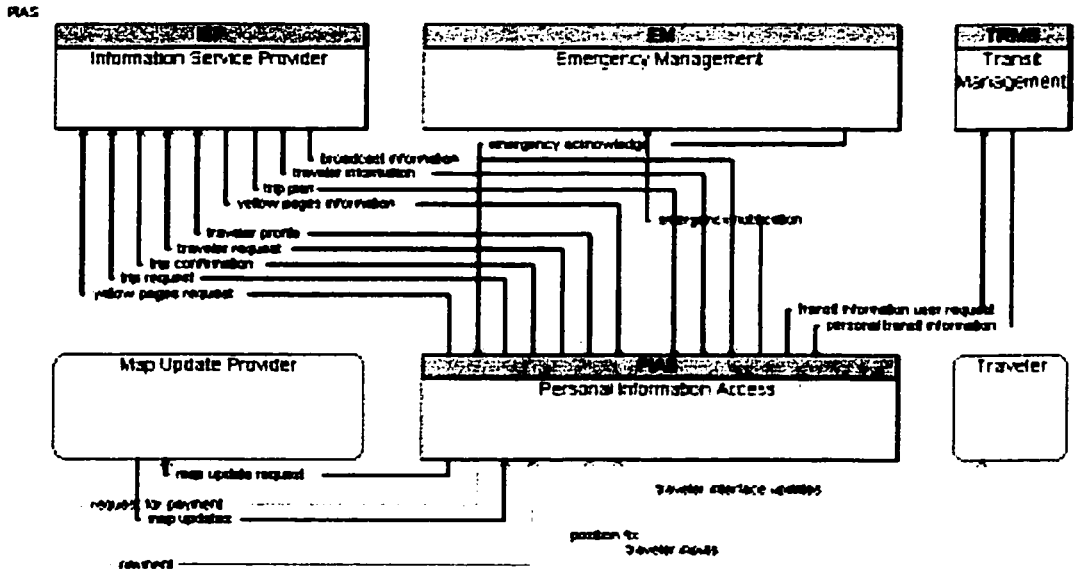


Figure 8.5: Detail from the personal information access subsystem. From ITERIS (2002).

subscription services.

The danger from an institutional perspective of taking this radical approach is that the technology is as yet unproven. Local-area wireless modems are only now being widely deployed. There is no real evidence concerning their practical bandwidth when being used on a crowded highway environment. This in turn limits the theoretical analysis and design of a distributed database. The safer option for ITS is to design around wire lines collecting data from infrastructure (road detectors, video devices, etc.), and then rely on private industry to solve transmitting that data back to the travelers, packaged as answers to travel queries.

As an early adopter of the wireless data technology, however, the Tracer system is exposed to the biggest problem with the current cellular technology approach to sending information (whether the technology is the slow but working CDPD modems, or the fast but not yet available 3G technologies). Specifically, cellular technology requires that users

pay a license fee to a corporation in order to send and receive data. In the case of CDPD, the fee is astronomically high compared to the actual speed of data transfers. This fee is a significant barrier to adoption.

In contrast, the spectrum allocated to the 802.11 family of protocols, to Bluetooth, and to other less developed wireless networking technologies is unlicensed. As long as individuals adhere to the low power broadcast requirements, the FCC allows the spectrum to be used free of charge. Free is not a barrier to use, and in fact will probably encourage use to the detriment of the system as a whole. The vision offered here is a free advanced traveler information system. Collecting one's own data incurs only hardware costs, and then sharing that travel information with others incurs no resource depletion costs. Being able to share the information via free local area wireless spectrum is especially attractive.

Finally, a centralized ISP vision is much more difficult to scale than a peer-to-peer vision. Imagine a sudden accident on the freeway. Under the ISP model everybody will be dialing up to ask their own unique questions on how this accident affects them, and will be getting variations on the same answer. The central server will have to handle a sudden onslaught of queries, and the wireless transmission capacity will have to be designed to handle huge surges in traffic. The more popular the client-server model gets, the worse its performance will be under peak loads, and the more over-designed the network in general will have to be to handle these peak loads. One should expect nothing less in a system designed by traffic engineers.

In contrast, a peer-to-peer information sharing approach improves with increased adoption. While there are plenty of issues that still need to be solved, in principle the accident in the above example will cause a flood of information to propagate outward from the ac-

cident, with data hopping from one vehicle to the next, and from vehicles to infrastructure nodes where it can travel even faster. Rather than having to frame a specific query and wait for an answer, one can instead listen for new information and pass that information along, reacting appropriately if the information is relevant to the planned travel. More peers means more eyes and sensors gathering information, more aggregate storage capability, more total processing power, more geographically continuous reach, and more independently developed algorithms and ideas.

Finally, if there is one thing that the Napster phenomenon should demonstrate to those who would sell digital content, it's that information (digital content) will be shared readily by people if it is easy to do so and represents no loss to themselves. I claim that information service providers will provide services that will not be significantly differentiable from what will be common knowledge once ITS systems are pervasive. That is, there are only so many unique traffic situations that might arise, and only so many donut shops that have never been visited before. If information is freely shared between vehicles, then once one vehicle knows about a good routing strategy or a local donut shop, that information will propagate outward to every other vehicle who might ask for that information. Similarly, Napster demonstrated that people like sharing data, and like broadening their minds with new information. The vast majority of music available for download on Napster at its height was not pop, but rather interesting odds and ends that gained a new audience. It did not diminish the quality of the music to share the music files. In fact, sharing files helped everybody on the system since there were more servers to choose from.

Similarly, it will not diminish the quality of the collected traffic information—whether real-time or historic—to share that data with others. While one might argue that a selfish

hoarding of knowledge of the best route or system instabilities will allow an individual to switch unilaterally and be better off, the plots at the beginning of this chapter for just the four pilot survey participants made it clear that for every route that a traveler knows well, there are many more that are completely unknown. Sharing an "expert" route might make a commuter slightly worse off in the one case, but as an expert on the route, the commuter can probably adjust with little effort. In contrast, the occasional foray into unknown territory is likely to be fraught with bad choices and no information of local dynamics. Having access to a pool of "expert" routes on all possible paths would be a huge benefit.

8.6 Conclusion

This chapter presented some early notions of how to move from the Tracer survey system to a decentralized, peer-to-peer advanced traveler information system. The core argument for the usefulness of such a system is that many of the routes and destinations of the four individuals who participated in the Tracer pilot study did not overlap. It was argued that individuals are likely to have a great deal of recent experiential knowledge about certain routes and destinations, but little to no expertise for the vast majority of potential paths and activity sites.

A discussion of the adaptations needed to convert the Tracer system into a peer-to-peer information system came next. The main conclusion is that a data type definition (DTD) or XML Schema is needed to codify what information might be shared between independent users of the Tracer system. Further, a short discussion of the likelihood of adopting the equipment ensued, focusing on personal Tracer implementations. The biggest change to

the system will have to be removing its reliance upon licensed wireless data services, such as CDPD, which are likely to be too expensive for the average consumer. Instead, a more likely option is for the user to have a local area wireless hub located in the garage, say, for uploading past data, and possibly downloading the day's travel plans and related data.

This chapter closed with a review of the National ITS Architecture, looking for peer-to-peer information exchange. No such provision was found, aside from vehicle to vehicle automatic control messages. It was argued again that peer-to-peer information dissemination is likely to scale better and be more useful in the long run than a centrally controlled traveler information service.

Chapter 9

Conclusion

This dissertation contributes important advances to the state of the art of transportation engineering. First, this dissertation describes the design and implementation of a GPS enhanced data collection unit. While the idea is not original with this dissertation, this GPS unit was successfully combined with a simple on-line survey that elicits activity descriptions and other annotations of the mechanically observed and recorded travel. The author is unaware of any other real-time integration of an activity survey with GPS data. This activity survey tool has been designed with the intention of being used to conduct very long term data collection efforts. The design goal of simple annotations to travel patterns, rather than detailed collection of multiple dimensions of activity data, is relatively unique. The breadth of data is sacrificed in the hopes of gaining a longer survey duration, which in turn will allow new kinds of activity sequencing research.

An informal pilot survey of four volunteers was conducted. Each individual soon stopped entering descriptive activity data, while continuing to contribute long periods of GPS data collection. The activity notation stopped much sooner than hoped, which indi-

cates that the website interface needs to be even easier to use. This is especially telling in that the pilot study volunteers were close associates of the author.

The pilot data did suggest two ways to improve that interface. First, a kriging surface was developed for the relative frequency of each named activity in space. This can be used to shorten and improve the listing of checked options related to each destination. As more data is collected and annotated, the the kriging surface will suggest the activities that are most likely to have taken place at a destination. Apparently irrelevant past entries can be hidden from view behind a "more options" button interface. Second, whenever a new destination is noted, that fact is apparent from its low or zero probability on the activity frequency surface. This could lead to a very sparse "new information" survey page, in which the respondent's time will be used to provide data on those points about which the least is known.

The analysis of survey data that was collected suggested that activities may be strongly linked to specific destinations. Some evidence of this is implied by the non-uniform kriging surfaces that were estimated. Stronger evidence is provided when the destinations for activities are examined as the result of some unknown spatial point process. For each survey volunteer, all destinations (ignoring any activity labels) are not not the result of a spatially random point process, but rather are strongly clustered. Second, it may be the case that specific activities have unique spatial distribution characteristics, relative to their own locations, and relative to the locations of other activities. These preliminary findings must be followed up with a larger data sample, coupled with improvements to the Tracer system and the website that were noted in the text.

In addition to the kriging surfaces linking activity names with locations and the point

process analysis indicating that activities are clustered in space, contributions were also made in the area of examining sequences of activities. In contrast to prior applications of sequence alignment techniques to activity analysis, the work reported here focused on the similarity between sequences, rather than differences, and attempted to align optimal local subsequences, rather than searching for a global alignment between sequences. The results indicate that patterns can indeed be determined with this approach, but that the raw reported activity names are not likely to generate many different patterns. It is recommended that travel activity names be added to the sequences, and that the reported activity names be augmented by assigning highly likely names to unlabeled destinations, based on the results of the kriging analysis.

This data collection tool, given its real-time characteristics, spurred an entirely different line of thinking. Instead of just collecting travel and activity data for research purposes, it appears that another useful application for this tool is to provide travelers with a way to collect, study, and learn from their *own* travel. Further, the integration of the real-time data stream with an on-line web site means that it will be a simple matter to implement a data sharing feature. Individuals who participate in the system can choose to publish their own trip and activity data, and potentially anybody with a web browser can access this information. The data will be available to all users of the system, for applications such as learning about new travel destinations and routes, or building a more detailed understanding of a well known route. A more difficult step will be to allow entirely independent Tracer systems to cooperate, but it is technologically feasible right now.

In the very near future, this kind of information could be collected by many simultaneous participants, all with automatic data publishing rules implemented. In such a world,

near real time traffic conditions could be available to all, based solely on the data collection efforts of the travelers themselves. This vision is markedly different from the National ITS Architecture approach, which discounts peer-to-peer information sharing to such an extent that it isn't even mentioned as a possibility. And yet, the technology for the system described in chapter 8 exists now. The only real barrier to implementation is the cost of wireless bandwidth, and the cost of the GPS data collection equipment. If prices continue to drop, and if free broadband spectrum can be used for these kinds of products, then this system might become a model for widespread adoption of distributed, peer-to-peer ITS. If this vision becomes a reality it will have profound consequences upon the way we live and move in the years ahead—all because of a simple desire to collect long term activity information.

Bibliography

Aero Data. Aero Data home page. published on the Internet, 2002. URL `http://www.aerodata.net`.

Theo Arentze and Harry Timmermans. *ALBATROSS: A Learning Based Transportation Oriented Simulation System*. CIP-DATA Koninklijke Bibliotheek, Den Haag, Netherlands, 2000a. ISBN 90-6814-100-7.

Theo Arentze and Harry Timmermans. Conceptual framework. In *ALBATROSS: A Learning Based Transportation Oriented Simulation System* Arentze and Timmermans (2000a), chapter 2, pages 71–80. ISBN 90-6814-100-7.

Theo Arentze and Harry Timmermans. Coevolutionary approach to extracting and predicting linked sets of complex decision rules from activity diary data. *Transportation Research Record*, (1752):126–132, 2001a.

Theo Arentze and Harry Timmermans. Inductive learning approach to evolutionary decision processes in activity-scheduling behavior. *Transportation Research Record*, (1752): 1–7, 2001b.

Prashanth K. Bachu, Trisha Dudala, and Sirisha M. Kothuri. Prompted recall in global

positioning system survey: Proof-of-concept study. *Transportation Research Record*, (1768):106–113, 2001.

Battelle. Lexington area travel data collection test. Technical report, Federal Highway Administration, 1997.

Baylor University. Geographic Resources Analysis Support System: Official GRASS GIS Web Site, 1999. URL `http://www.baylor.edu/~grass/`.

M. E. Ben-Akiva and J. L. Bowman. Activity-based disaggregate travel demand model system with daily activity schedules. In *EIRASS Conference on activity-based approaches: Activity scheduling and the analysis of activity patterns*, Eindhoven University of Technology, Eindhoven, The Netherlands, May 25–28 1995.

Chandra R. Bhat and Frank S. Koppelman. A conceptual framework of individual activity program generation. *Transportation Research A*, 27(6):433–446, 1993.

F. Stuart Chapin, Jr. *Human activity patterns in the city; things people do in time and in space*. Wiley, New York, 1974.

Mike Clarke and Martin Dix. Stage in lifecycle—a classificatory variable with dynamic properties. In Susan Carpenter and Peter Jones, editors, *Recent Advances in Travel Demand Analysis*, chapter 11. Gower Publishing Company Limited, Gower House, Croft Road, Aldershot, Hampshire GU11 3HR, England, 1983.

Thomas H. Cormen, Charles E. Leiserson, and Ronals L. Rivest. *Introduction to Algorithms*. MIT Press, 1990.

Noel A. C. Cressie. *Statistics for Spatial Data*. John Wiley & Sons, Inc., New York, revised edition, 1993.

Ian Cullen and Vida Godson. *Urban Networks: The structure of Activity Patterns*, volume 4 of *Progress in Planning*. Pergamon Press, Oxford, 1974.

David Damm. Parameters of activity behavior for use in travel analysis. *Transportation Research A*, 16:135–148, 1982.

Sean T. Doherty and Kay W. Axhausen. The development of a unified modeling framework for the household activity-travel scheduling process. *Stadt Region Land*, 66:45–59, 1998.

Sean T. Doherty and Eric J. Miller. A computerized household activity scheduling survey. *Transportation*, 27:75–97, 2000.

Geert Draijer, Nelly Kalfs, and Jan Perdok. Global positioning system as data collection method for travel research. *Transportation Research Record*, (1719):147–153, 2000.

Dick Ettema and Harry Timmermans, editors. *Activity-Based Approaches to Travel Analysis*. Elsevier Science Inc., Tarrytown, NY, USA, 1997. ISBN 0-08-042584-4.

Thomas F. Golob. A simultaneous model of household activity participation and trip chain generation. *Transportation Research B*, 34:355–376, 2000.

Thomas F. Golob and Michael G. McNally. A model of activity participation and travel interactions between household heads. *Transportation Research B*, 31B(3):177–194, 1997.

Dan Gusfield. *Algorithms on strings, trees, and sequences.* Cambridge University Press, New York, 1997. Reprinted 1999 (with corrections).

Torsten Hägerstrand. What about people in regional science? *Papers of the Regional Science Association*, 24:7–21, 1970.

A. S. Harvey. From Activities to Activity Settings: Behaviour in Context. In Ettema and Timmermans (1997), chapter 11. ISBN 0-08-042584-4.

IATUR. ClustalG program, version 1.2. available on the World Wide Web, 2001. URL `http://www.stmarys.ca/partners/iatur/clustalG/`.

Ross Ihaka and Robert Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

ITERIS. National ITS architecture, version 4.01. on the Internet, July 2002. URL `http://itsarch.iteris.com/itsarch/index.htm`.

Chang-Hyeon Joh, Theo Arentze, Frank Hofman, and Harry Timmermans. Activity pattern similarity: a multidimensional sequence alignment method. *Transportation Research B*, (36):385–403, 2002.

Chang-Hyeon Joh, Theo A. Arentze, and Harry J.P. Timmermans. Multidimensional sequence alignment methods for activity-travel pattern analysis: A comparison of dynamic programming and genetic algorithms. *Geographical Analysis*, 33(3):247–270, July 2001a.

Chang-Hyeon Joh, Theo A. Arentze, and Harry J.P. Timmermans. Pattern recognition in complex activity-travel patterns: Comparison of Euclidean distance, signal-processing

theoretical, and multidimensional sequence alignment methods. *Transportation Research Record*, (1752):16–22, 2001b. Paper No. 01-0188.

P.M. Jones, M.C. Dix, M.I. Clarke, and I.G. Heggie. *Understanding Travel Behavior*. Oxford Studies in Transport. Gower Publishing Company Limited, Gower House, Croft Road, Aldershot, Hampshire GU11 3HR, England, 1983.

N. Kalfs and W. E. Saris. New Data Collection Methods in Travel Surveys. In Ettema and Timmermans (1997), chapter 13, pages 243–261. ISBN 0-08-042584-4.

Ming-Sheng Lee and Michael G. McNally. Experiments with computerized self-administrative activity survey. *Transportation Research Record*, (1752):91–99, 2001. Paper No. 01-3092.

Bo Lenntorp. *Paths in Space-Time Environments: A Time-geographic Study of Movement Possibilities of Individuals*. Lund Studies in Geography. Royal University of Lund, Lund, Sweden, 1976.

Linux Embedded Appliance Firewall (LEAF). Linux Embedded Appliance Firewall (LEAF), 2002. URL `http://leaf.sourceforge.net`. This is the successor to the Linux Router Project.

James E. Marca, Craig R. Rindt, and Michael G. McNally. A simulation framework and environment for activity based transportation modeling. Technical Report UCI-ITS-AS-WP-99-2, University of California, Irvine, Institute of Transportation studies, 1999. CASA working paper.

Michael G. McNally. An activity-based microsimulation model for travel demand forecasting. In Ettema and Timmermans (1997), chapter 2. ISBN 0-08-042584-4.

Michael G. McNally and Wilfred W. Recker. *On the formation of household travel/activity patterns: A simulation approach*. U.S. DoT Contract No. DTRS-57-81-C-0048, Final Report. Department of Civil Engineering and Institute of Transportation Studies, University of California, Irvine, 1986.

Michael G. McNally. GPS/GIS technologies for traffic surveillance and management: A testbed implementation study. Submission for PATH RFP FY 1999-2000, February 1999.

Elaine Murakami and D. P. Wagner. Can using global positioning system (GPS) improve trip reporting? *Transportation Research C*, 7(2/3):149–165, April/June 1999.

Portland METRO. Oregon and Southwest Washington 1994 Activity and Travel Behavior Survey. 600 Northeast Grand Avenue, Portland, OR 97232-2736, 1994.

Wilfred W. Recker. The household activity pattern problem: General formulation and solution. *Transportation Research B*, 29B(1):61–77, 1995.

Wilfred W. Recker, Michael G. McNally, and Greg S. Root. A model of complex travel behavior: Part I—Theoretical development. *Transportation Research A*, 20A(4):307–318, 1986a.

Wilfred W. Recker, Michael G. McNally, and Greg S. Root. A model of complex travel behavior: Part II—An operational model. *Transportation Research A*, 20A(4):319–330, 1986b.

B. D. Ripley. *Spatial Statistics*. John Wiley & Sons, Inc., New York, 1981.

B. D. Ripley. *Statistical Inference for Spatial Processes*. Cambridge University Press, 1988.

Jean-Claude Thill and Aaron Wheeler. Tree induction of spatial choice behavior. *Transportation Research Record*, (1719):250–258, 2000.

Harry Timmermans. Data requirements and data collection. In *ALBATROSS: A Learning Based Transportation Oriented Simulation System* Arentze and Timmermans (2000a), chapter 5, pages 136–151. ISBN 90-6814-100-7.

UserLand Software. XML-RPC home page, 2001. URL `http://www.xmlrpc.com/`.

Kenneth M. Vaughn, Paul Speckman, and Eric Pas. Generating household activity-travel patterns (HATPs) for synthetic populations. In *76th annual meeting of the Transportation Research Board*, Washington, D. C., January 12–16 1997. TRB.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S-Plus*. Statistics and Computing. Springer-Verlag, New York, third edition, 1999.

Dave Warner. XML-RPC: It works both ways. davewarner98@yahoo.com, Jan 2001. URL `http://www.oreillynet.com/pub/a/python/2001/01/17/xmlrpcserver.html`.

Clarke Wilson. Activity patterns of Canadian women: Application of ClustalG sequence alignment software. *Transportation Research Record*, (1777):55–67, 2001.

Clarke Wilson, Andrew Harvey, and Julie Thompson. ClustalG: Software for analysis of activities and sequential events. In *Paper presented at the 1999 IATUR annual conference*, University of Essex, Colchester, UK, October 1999. IATUR.

W. C. Wilson. Activity pattern analysis by means of sequence-alignment methods. *Environment and Planning A*, 30:1017–1038, 1998.

Jean Wolf, Randall Guensler, and William Bachman. Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data. *Transportation Research Record*, (1768):125–134, 2001.

Jean Wolf, Shauna Hallmark, Marcelo Oliveira, Randall Guensler, and Wayne Sarasua. Accuracy issues with route choice data collection by using global positioning system. *Transportation Research Record*, (1660):66–74, 1999.

Lalit Yalamanchili, Ram M. Pendyala, N. Prabaharan, and Pramodh Chakravarthy. Analysis of global positioning system-based data collection methods for capturing multistop trip-chaining behavior. *Transportation Research Record*, (1660):58–65, 1999.