Dynamic Location Management and Activity-based Mobility Modelling for Cellular Networks

by

John Scourias

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 1997

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

0-612-21539-3

Canada

The University of Waterloo requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Abstract

The challenge of supporting rapidly growing numbers of mobile subscribers, while constrained by limited radio spectrum, is being faced by cellular network operators worldwide. The standard technique for cellular networks is to decrease radio cell size, thereby maintaining a supportable subscriber density. However, smaller cell sizes result in increased signalling for location management procedures, which reduces the bandwidth available for user traffic.

Location management is an essential function in cellular networks that allows the network to maintain the position of subscribers, in terms of location areas. Location areas in current systems, such as GSM, consist of static and arbitrarily-defined collections of cells, which do not take into account individual subscriber mobility patterns, either in space or time. Consequently, performance suffers significantly in terms of signalling requirements.

A location management algorithm is proposed which uses the mobility history of individual subscribers to dynamically create individualized location areas, based on previous movements from cell to cell. The average visit duration in each cell is also maintained and is used to define paging areas, or subsets of the location area which are most likely to contain the subscriber.

To test the proposed algorithm, a stochastic mobility model was developed based on daily sequences of activities and their associated durations and locations. The mobility model outputs the cells traversed by a subscriber over time. A call arrival distribution, together with the mobility model, provided a simulation framework used to compare the proposed algorithm with variations of the static location management scheme.

Overall, the dynamic algorithm generated significantly lower location management costs, in terms of signalling messages generated, for all parameters examined. The proposed dynamic algorithm requires slightly increased memory and processing capabilities in the mobile station and network, but can significantly reduce location management signalling compared to current techniques.

Acknowledgements

First of all, I would like to thank God, my parents Pericles and Pavlina, and the Department of Computer Science at the University of Waterloo, for giving me the opportunity to continue learning and to complete this programme.

I am grateful to my supervisor, Dr. Thomas Kunz, for his support, constructive criticism, and valuable ideas.

I would like to thank the reviewers of the thesis, Dr. Jay Black and Dr. Grant Weddell. I would also like to thank Jane Prime, Debbie Mustin, and the other administrative staff in the Department of Computer Science for smoothing the administrative hurdles.

Thank you to Nicole Roslin, for her patience and encouragement, and for all her help, especially in sorting out the statistical analysis... Perl scripts are great, but SAS is better.

The inter-disciplinary nature of this thesis, among other things, required me to ask many questions, and I am very grateful to all who took the time to answer them, especially the following people:

- Shilling Yip, of the Region of Waterloo Planning Office, who provided me with the trip survey needed by the mobility model, traffic zone maps of the region, and answers to many questions.

- Dr. Bruce Hutchinson, of the Civil Engineering Department, who despite the vagueness of my early ideas, provided direction, encouragement, and answers to my many traffic engineering questions.

- Stuart Wachsberg, of the Computer Science Department, for letting me bounce ideas off him, and for his programming help.

- David Evans, of the Computer Science Department, for all his help, including help with the simulation code for the mobility model.

- **Ihab El-Khodari and Chris Hoff, of the Civil Engineering Department, for all their help and advice, including the monumental task of properly setting up the TransCAD software.**

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 General concepts in location management

Over the last several years, the worldwide cellular communications market has undergone explosive growth. This can be attributed to several factors, including decreasing prices, improved radio coverage, and compact, lightweight terminals. During 1996, for example, European countries had an average cellular subscriber growth rate of 33.9%, and many Asian countries had annual growth rates of over 50% [32]. In general, as the number of subscribers increases given a fixed radio spectrum allocation, the size of radio coverage cells must decrease, in order to accommodate the higher subscriber densities. This creates many challenges, especially concerning location management. Personal communication systems (PCS), which are higher-capacity and lower-power digital cellular networks with smaller cells, face similar location management issues.

As cellular operators deploy even smaller cells, creating microcellular and picocellular networks to increase capacity, the problem of efficient location management will become more severe due to the additional signalling created by more subscribers and more cells. The additional signalling consumes scarce radio bandwidth, as well as fixed bandwidth and processing capacity, and thus must be kept to a minimum. Handovers, which are an important aspect of wireless communications, and have received much attention in recent years, are not normally considered part of location management and are not addressed here.

Location management is concerned with the procedures required to enable the network to maintain location information for each subscriber, or more specifically, for each active mobile terminal with a registered subscriber, and to efficiently handle the establishment of incoming calls. One of the fundamental properties of cellular networks is that, because subscribers can roam throughout the network and possibly in other networks, there can be no permanent connection to a subscriber as is the case in fixed telephony. In order to be able to route incoming calls to mobile subscribers, the network needs the current location of each subscriber, in terms of the particular cell that the subscriber is currently *camped* on. Mobile stations camp on the base station which provides the best signal quality, and listen for paging messages addressed to the mobile subscriber indicating that an incoming call has arrived.

The location management procedures are required in cellular networks to manage subscriber mobility. In general, mobility in telecommunications networks may be categorized into terminal mobility and personal mobility, with the latter being the goal of modern cellular systems. Terminal mobility allows a terminal to physically move, or roam, throughout the coverage area of a network. This is an inherent feature in all cellular networks. Personal mobility goes a step further and allows a subscriber to roam throughout a network irrespective of a particular terminal. This is accomplished by allocating an identity number not only to the terminal, but also to the subscriber, and keeping an association between the subscriber and the terminal. In practice, this involves the use of a personalized smart card which can be inserted into any compatible terminal, and which allows calls to be made and received at that particular terminal, with proper billing and call routing. The implementation of personal mobility requires that it is the mobile subscriber, not the mobile terminal, that must be tracked, either directly or indirectly (by tracking the mobile terminal, but providing a mapping function between terminal and subscriber identifiers).

The two fundamental procedures which comprise the basis of location management are location updating and paging. Location updating is initiated by the mobile station, and informs the network of the subscriber's current location area. Paging is initiated by the network when an incoming call arrives. Paging messages are broadcast in one or more paging areas, contained within the current location area, and inform the target user of the incoming call. These concepts will be defined in more detail below.

Except for certain geographical positioning applications, the smallest unit of location information that cellular networks are concerned about is the *radio cell.* A cell is the geographical area over which the average signal strength from the radio transmitter that defines the cell is over some threshold, as compared to neighbouring cells and frequencies. Cells are ideally represented as hexagons, although in reality, interference and path losses due to irregular terrain and neighbouring cells, as well as power control algorithms, smear cell boundaries into fuzzy, irregular shapes.

For location management purposes, cells are usually grouped together into *location areas* and *paging areas.* A location area is a set of cells, normally (but not necessarily) contiguous, over which a mobile station may roam without needing any further location updates. In effect, a location area is the smallest geographical scale at which the location of the mobile station is known. A paging area is the set of cells over which a paging message is sent to inform a subscriber of an incoming call. Obviously, location areas and paging areas must be related to be useful. In most operational systems, location areas and paging areas are identical. For this reason, any grouping of cells for location management purposes is usually called a location area.

The relationship between location updates, pagings, location areas, and paging areas, can best be illustrated by the two extreme approaches to location management. The first approach maintains the location of the mobile subscriber at the individual cell level. In other words, the size of the location area in this case is exactly one cell. As mentioned above, this is the finest granularity of location information possible. Whenever the mobile station detects that it has moved to a new cell, which may happen very frequently in the case of an automobile-mounted mobile station, a location update is performed. This is clearly very inefficient. However, paging messages need only be sent to one cell, since the exact location of the mobile station is known.

At the other extreme, the location area of the mobile is comprised of all cells in the network. Knowledge of location information is at the coarsest level possible, with the network knowing only whether the mobile station is in the network or not. This means that location updates are not required at all. However, when an incoming call arrives, the network must page every cell in the network. Since this must be done for every incoming call to every mobile subscriber, this is clearly an unsatisfactory approach.

1.2 Problem definition

As described in Chapter 2, location management as currently performed is suitable for a relatively small number of subscribers, in a cellular environment that uses relatively large cells. The definition of location areas, which can obviously greatly influence the location updating and paging traffic, is performed statically, using aggregate statistics and traffic patterns. While the overall number of subscribers is relatively low, and the location areas are moderately large, the signalling traffic volume remains acceptable.

In recent years, cellular subscriber growth has increased dramatically worldwide. To provide the additional capacity, while constrained by the fixed available radio spectrum, operators have been using progressively smaller cells, through cell splitting and sectorization. To accommodate this trend, either the number of cells and subscribers per location area must increase (to maintain the same geographical size for the location areas), or the geographical size of location areas must decrease (to maintain the same number of cells per location area). Assuming the same subscriber mobility patterns and location management algorithms, the number of paging messages will increase in the former case, and the number of location updates will increase in the latter case. In the former case, the number of paging channels will need to be increased in all cells, thus reducing the number of channels available for voice or data traffic. In the latter case, the number of channel access attempts will increase due to location updates, thus decreasing the probability of successful call attempts.

The motivation behind this thesis is the realization that the above situation arises because location areas are geographically fixed, and apply to all subscribers, regardless of the individual subscriber mobility patterns. This can be illustrated by considering an example with a fast subscriber (for example, driving in a car) and a slow subscriber (for example, a pedestrian). The fast subscriber may cross several location areas during a trip, while the pedestrian may not even cross one. The fast subscriber can be handled efficiently by creating large location areas, to avoid numerous location updates. However, if these large location areas would also apply to the pedestrian subscribers, significant signalling would be wasted paging over the entire location area, when paging over a much smaller location area (or smaller paging areas within the location area) would suffice.

One approach to location management is to use relatively large location areas, which are subdivided into several paging areas. By using additional information about subscriber mobility patterns, the network is able to determine the current paging area with a high probability of success on the first attempt. In addition to a general velocity-based mobility class, there are several other parameters which affect subscriber mobility, and thus location management procedures [40].

One of these parameters is time of day. Regular full-time workers tend to follow the same mobility pattern with respect to time. The most obvious manifestation of this phenomenon is the morning and afternoon rush hour. Accurate knowledge of such phenomena may allow, for example, the use of smaller paging areas during low mobility time periods, thus minimizing signalling bandwidth, and the use of larger paging areas during high mobility time periods (e.g., morning rush hour).

Even if paging areas could be allocated taking into account certain mobility classes and the time of day, the procedures would still be far from optimal. The reason for this is that the location areas and paging areas would still apply to all subscribers equally, regardless of individual mobility patterns in time or space. Dimensioning location and paging areas for the average user does not take into account individual mobility patterns, and thus cannot be as efficient as an algorithm based on individual subscriber movements.

1.3 Basic proposal and thesis outline

The problem of statically-defined location areas and paging areas, which apply to all subscribers equally regardless of individual spatial or temporal mobility patterns, is addressed in this thesis. The first contribution is a proposal for a location management algorithm based on the use of individual mobility patterns to dynamically create personal location and paging areas. The second contribution is a stochastic mobility model used to compare different location management algorithms. The model is based on activity selections by individuals, and traces the path to the destination as a sequence of traversed radio cells.

Chapter 2 discusses the procedures involved in location management, both in general terms, and as implemented in the GSM cellular network standard. In Chapter 3, previous work on location management is reviewed. The advantages and disadvantages of various approaches are discussed.

The proposed mobility model is explained in Chapter 4, together with some background on transportation planning and activity modelling theory. The mobility model formed the basis of the simulation study, also described in this chapter. The proposed location management algorithm is discussed in Chapter 5, along with the requirements that it attempts to fulfill. In Chapter 6, the results from the simulation study are presented and analysed. Conclusions from the simulation and possible future work are discussed in Chapter 7.

Chapter 2

Location management procedures

2.1 General procedures

Location management procedures can be divided into the two broad categories of call-related and non-call-related procedures. The basic difference is that call-related procedures are invoked only when an incoming call arrives, while non-call-related procedures can be invoked at any time while a mobile station is active, or powered on.

2.1.1 Call-related location management functions

Call-related functions are initiated by the mobile network whenever it receives an incoming call for a mobile subscriber. The incoming call could originate from another user in the mobile network, or from a fixed network user. The purpose of call-related location management functions is to locate the user for whom the incoming call is intended. Locating the subscriber is a two-step process, which begins with an interrogation phase, where a location database is queried for the paging area of the subscriber. The next step is a paging phase, where each cell in the paging area broadcasts a paging message indicating the intended subscriber for the incoming call.

Interrogation

In any cellular network, the location information of a subscriber, as well as other subscription information, is stored in a database in the home network of the subscriber. The precise contents of this database (i.e., the granularity of location information) and the structure of the database (i.e., centralized or distributed) differ between cellular standards, and even between different implementations of the same standard. These issues are the focus of current research, some of which will be covered in Chapter 3.

During the interrogation phase, an intermediate switch in the mobile or fixed network, which has access to the location database, will query the database for the current location of the subscriber associated with the destination dialled number. One possible reply from the database is an address identifying another database which contains the required information (or which has pointers to the required information). Another reply may be a local telephone number (regular dialled number, compliant with ITU-T E.164) that enables a direct connection to the mobile subscriber.

There are a number of factors that should be taken into consideration when comparing different database structures and interrogation schemes. First, access delays translate directly into end-user connection delays, and thus adversely affect the perceived quality of service. From an implementation efficiency standpoint, the amount of storage space required is also a consideration. For distributed database implementations, providing transparency and internal consistency are important points.

Paging

Paging involves messages sent over the radio interface, informing the mobile subscriber that an incoming call is pending. When the mobile station replies, the exact base station to which it is attached will be known to the network, and the call setup can proceed.

Recall that the network knows the position of the mobile station only at the location area level.

Since radio spectrum is scarce, these messages must be kept to a minimum, by paging a minimum number of cells. The tradeoff, as mentioned above, is that in order to minimize the number of cells that must be paged, location updates must be more frequent. A fact that should also be taken into account is that because of the unpredictable nature of radio communications, paging messages may not arrive at the mobile with the first attempt, and there is usually some number of repetitions. Since the arrival of paging messages obviously cannot be predicted, a mobile station must listen to the paging channel continuously (or almost continuously, as will be explained for GSM). The constant observation of the paging channel consumes battery power, a limited resource for the mobile station.

2.1.2 Non-call-related location management functions

Unlike call-related location management functions, which are only initiated when an incoming call arrives for a mobile subscriber, non-call-related functions may be executed at any time while the mobile station is powered on. The purpose of this group of functions is to transmit location information from the mobile station to the network. Updated location information could be sent to the network when an active mobile station detects a change in its location area, or when a mobile station powers on or off.

Location Update

The main non-call-related function is the location update. A cellular network is comprised of a number of cells, which are normally grouped together into location areas, as described above. A base station will transmit a location area identifier as well as a cell identifier. A mobile station continuously monitors signal strength and/or quality from surrounding cells. If a mobile station decides that reception is better in another cell, and the location identifier of the new cell is different from the location area identifier of the previous cell, a location update is triggered.

Location updates require the setup of a signalling channel to convey the information, and should thus be minimized in a limited channel and limited spectrum environment. As explained above, however, there is a trade-off between paging and location updating.

Location registration

Many authors do not distinguish between location updates and location registrations. The difference concerns the transfer of database information between location registers, and may depend on the location update procedures and network topology. Generally speaking, a location update modifies subscriber data which already exists at a location register. During a location registration, this data is not present at the location register, and must be retrieved from another location register.

Attach and Detach

The attach and detach procedures are complementary to each other, and are used as a way of reducing the number of overall paging messages. The detach procedure is initiated when an active mobile station powers off. Since a powered-off mobile station is obviously unreachable for incoming calls, attempting to page that subscriber would result in wasted signalling. To prevent this, when a mobile station powers off, the network is notified and stores the information. Any subsequent incoming calls will be rejected at the network level, avoiding radio signalling and conserving radio resources.

The attach procedure is initiated when the mobile station powers on again. If the mobile station is still in the same location area as when it powered off, it will send an attach message notifying the network that the subscriber is again available to receive incoming calls. If the mobile station discovers that it is in a new location area, a normal location update will be performed.

It is also possible for the network itself to perform a detach operation on a subscriber [31, 39]. For example, a mobile station may be detached if it could not be reached after a certain number of paging attempts. Another possibility is related to periodic updating. A timer could be maintained by both the network and the mobile terminal, which would be reset each time there was contact between the two entities. If the mobile timer expired, the mobile station would re-attach. If the (longer duration) network timer expired without the terminal having re-attached, the mobile station would be considered unreachable, and would be automatically detached.

2.2 GSM location management fundamentals

The location management concepts and procedures will be explained in greater detail by looking at their implementation in the GSM (Global System for Mobile communications) digital cellular standard. For many location management proposals, a scheme similar to the one used in GSM is used as the standard of comparison. GSM is a second-generation cellular communications system which has rapidly gained market share worldwide since its debut in 1991. Unlike analog first-generation systems such as AMPS (Advanced Mobile Phone Service), GSM transmits digital signals over the radio interface as well as internally between different network elements. It provides ISDN compatibility in terms of user services offered and general protocol structure, although not in terms of data rates, since the allocated radio spectrum is limited.

In this section, a brief overview of GSM will be given, including the general system architecture, the parameters used for identifying the entities involved in location management, and the procedures used to select the serving network and cell. These concepts will be needed to understand the subsequent section, describing the location management procedures used in GSM.

2.2.1 Architecture of GSM

The functional architecture of a GSM network can be divided into the mobile station, the base station subsystem, and the network switching subsystem, as shown in Figure 2.1. Each subsystem is comprised of functional entities which communicate through various interfaces using specified protocols.

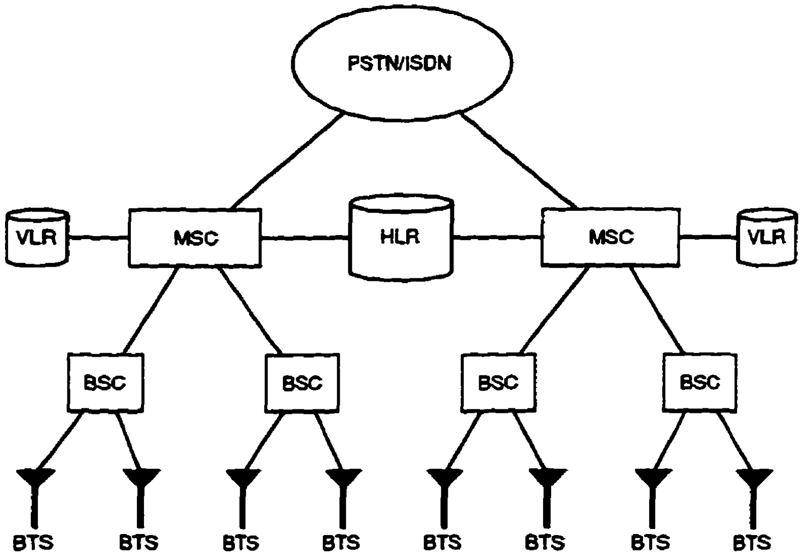

Figure 2.1: Architecture of a GSM network

Mobile Station

A GSM mobile station consists of two entities: the mobile equipment itself, and a smart card called the Subscriber Identity Module (SIM). The SIM card is an integral part of the GSM design, which allows the implementation of personal mobility. A subscriber can plug the SIM card into any GSM terminal and is able to place and receive calls, with proper routing and billing. The SIM card is also important for the mobility management functions since it stores subscriber identity and location information. The SIM card is also involved in the terminal side of security functions such as authentication and encryption.

Base Station Subsystem

All cellular networks overlay a geographical area with a grid of cells, where each cell is defined by a single Base Transceiver Station (BTS). A BTS is assigned a set of frequencies in such a way that interference from other cells in the network reusing the same frequencies is kept to a minimum, while maintaining enough radio channels in a cell to handle the required subscriber density.

In order to keep infrastructure costs low and to provide sufficient coverage, the management of the radio resources for several BTSs is handled by a Base Station Controller (BSC), including the setup and release of radio channels, frequency hopping, and inter-BTS handovers. The BTS and BSC together comprise the base station subsystem, and handle almost all radio resource aspects of GSM. Signalling between the mobile station and the network uses a layered protocol stack, the top layer of which is specified in GSM specification 04.08 [17]. This specification defines the type and size of information that passes over the radio interface for location management procedures.

Mobile services Switching Center

The Mobile services Switching Center (MSC) is the central component of the network switching subsystem. The MSC has functionality similar to a Public Switched Telephone Network (PSTN) switch, and in addition contains the functionality needed for handling subscriber mobility, such as location management, call routing and call establishment, and security. An MSC is connected to one or more BSCs. The functionality of the MSC is provided in conjunction with the following four intelligent databases defined in the

GSM architecture, which together with the MSC form the network switching subsystem:

- Home Location Register (HLR)
- Visitor Location Register (VLR)
- Authentication Centre (AuC)
- Equipment Identity Register (EIR)

The MSC and the databases communicate using the Mobile Application Part (MAP) protocol, which uses the services of Signalling System No. 7 (SS7). SS7 is a telecommunication signalling protocol standardized by the International Telecommunication Union's Telecommunication Sector (ITU-T), formerly known as the CCITT.

Home Location Register

In terms of location management and call routing, the relevant databases are the Home Location Register (HLR) and the Visitor Location Register (VLR). The HLR of each GSM operator contains administrative information for all subscribers, along with their current location. The location information is necessary for routing incoming calls to the mobile station, and is typically the Signalling System No. 7 address of the visited VLR. There is logically one HLR per GSM network, although it may be implemented as a distributed database.

Visitor Location Register

The Visitor Location Register (VLR) is associated with a particular geographical area, usually (but not necessarily) associated with the area covered by one MSC. It contains a subset of the administrative information from the HLR, necessary for call control and provision of subscribed services, for each mobile subscriber currently located in the geographical area controlled by that VLR. Although the VLR can be implemented as an independent entity, all manufacturers of switching equipment to date implement the VLR together with an MSC. The combination of the VLR and MSC speeds up access times to information that the MSC requires during call processing.

2.2.2 Identification parameters

International Mobile Subscriber Identity (IMSI): The IMSI is used to uniquely identify a subscriber. It consists of a maximum of 15 digits, which are divided up into a Mobile Country Code (3 digits), a Mobile Network Code (2 digits), and a Mobile Subscriber Identification Number (up to 10 digits).

Temporary Mobile Subscriber Identity (TMSI): The TMSI is allocated by a VLR to a subscriber, and is valid only while the subscriber is registered at that VLR. The purpose of the TMSI is to save paging bandwidth, since it is shorter than the IMSI, and also for privacy, since a TMSI alone does not uniquely identify a subscriber. A TMSI is allocated to a subscriber only in encrypted form.

Mobile Station ISDN number (MSISDN): The MSISDN is the number that is dialled by a caller to access the mobile subscriber. It follows the ITU-T E.164 numbering plan, and the format consists of a Country Code, a National Destination Code, and a Subscriber Number. The Country Code and National Destination Number should provide enough information such that it may be translated into an SS7 global title address for the HLR of the subscriber. The first few digits of the Subscriber Number may be used for additional routing information, if necessary.

Mobile Station Roaming Number (MSRN): The MSRN is a local telephone number, following the local numbering plan, which is used by an MSC and VLR to route an incoming call to a mobile station. The VLR may assign an MSRN from its local pool either on demand for an incoming call, or, less likely, when a mobile subscriber first registers with the VLR. The latter is a less efficient technique, since the mobile might not receive any calls while assigned a number.

Location Area Identifier (LAI): The LAI is broadcast by each base station, uniquely identifying the location area to which it belongs. The LAI is comprised of three parts: a Mobile Country Code, a Mobile Network Code, and a 2-octet Location Area Code.

Cell Identity: A particular cell within a location area is identified with the Cell Identity (CI), which has a fixed length of 2 octets. The CI identifies the cell locally, within its location area; globally, the CI is appended to the Location Area Identifier.

Base Station Identity Code (BSIC): Sometimes confused with the Cell Identifier, the BSIC serves a different purpose. The BSIC is comprised of two 3-bit fields, the Public Land Mobile Network (PLMN) Colour Code, and the Base Station Colour Code. The purpose of the BSIC is to resolve any ambiguities that may arise when a mobile station is able to receive two different base station broadcast channels, which use the same beacon frequency. Details of their use and allocation are beyond the scope of this paper, but it should be noted that the BSIC is not the same as the Cell Identity.

2.2.3 Network and cell selection

A location update is triggered, always by the mobile station, in the following cases:

- When better service, in terms of radio reception quality, can be obtained from a cell in another location area
- When the mobile station is switched on in a location area different than the last recorded location area
- When the mobile station attempts to get normal service from a different GSM network

The task of selecting the most appropriate network and cell is thus central to location updating, and is not trivial considering the instability of the radio environment. In addition, there are certain administrative constraints which must be taken into account. The network and cell selection criteria and algorithms employed by GSM will be discussed below in some detail.

Many countries which have GSM networks have more than one operational GSM network, and the coverage areas of the different networks often overlap. A mobile station is served by exactly one network at any point in time. An idle mobile station camped on a cell is able to listen to the Broadcast Control Channel (BCCH) to obtain information about the cell. The BCCH continuously broadcasts the cell identity, location area identity, network identity, power level parameters, and control channel structure. In addition, the beacon frequencies of the surrounding cells are broadcast on the BCCH of the serving cell, allowing the mobile station to easily monitor surrounding cells for possible cell reselection.

The home network is normally selected, but when more than one network is available, or when the home network does not provide adequate coverage, changing the serving network is possible. If the user is not subscribed to any available networks, only emergency calls are possible. In that case, the mobile station automatically selects the cell (and corresponding network) with the strongest signal from up to thirty monitored signals.

If normal service is possible, the user may select a serving network manually, or may let the mobile station select one automatically, based on three lists of networks stored on the SIM card: the preferred list, the found list, and the forbidden list.

Once a serving network is selected, the serving cell on which the mobile station will camp is selected. Cell selection in a mobile network under real operating conditions is far from trivial, due to the complicated nature of radio wave propagation. Momentary Rayleigh fades, for example, should not cause new cell selection. In addition, cell selection should be fairly stable along cell borders, which are quite fuzzy in practice, to avoid 'thrashing' between location areas, a situation which creates much unnecessary signalling, and may result in lost paging requests.

When the mobile station first switches on, the cell selection procedure uses the C1 criterion to determine the 'best' cell, between all suitable cells (i.e., cells belonging to the selected network, which are not barred from access). The cell with the highest positive value of C1 is selected, where C1 is calculated by the mobile station as follows:

$$C1 = averageReceiveLevel - p1 - \max(p2 - maxMobileStationPowerLevel, 0)$$

The values $p1$ and $p2$ are power level parameters that are broadcast on the BCCH of a cell. Once a cell has been selected, subsequent cell selections use the C2 parameter, which is the sum of the C1 parameter and some additional parameters. The C1 and C2 parameters effectively determine cell boundaries, and should be defined (using appropriate parameters in C2) to prevent thrashing by delaying the selection of cells that would needlessly trigger location updates, as well as to force cell reselection when appropriate, such as when power levels are excessively high.

2.3 Location management procedures in GSM

For mobility management, GSM takes the intermediate approach between full-system paging and per-cell location updating, using location areas and paging areas (which are equivalent in GSM) made up of one or more cells. The actual assignment of cells to location areas is operator-defined, and is constrained only by the fact that a location area cannot contain cells belonging to more than one MSC.

The sections below describe the GSM procedures used for location management. These include location updating (normal, periodic, and attach), and interrogation and paging for incoming calls. The descriptions are for the general, and successful, case. There are many variations possible due to unsuccessful attempts or operator-configurable system options.

2.3.1 Location update and registration

A location update is always initiated by the mobile station. It is normally initiated when the mobile station camps on a cell which broadcasts a location area identifier that is different from the one stored in the SIM card (if a previous value exists). In addition, location area updates may be triggered periodically, to let the network know that the mobile is still accessible, or to allow the network to update its location registers in case of failure. A similar procedure is the IMSI attach, which involves a smaller set of operations at the network side. As mentioned previously, no location updates are required while roaming within the same location area.

Location updates and location registrations are very similar, and differ only in additional data transferred on the network side for location registrations. If the mobile station registers at a VLR for the first time, no subscriber information exists there yet, and it must be requested from the HLR. For location updates between location areas under the same VLR, the subscriber information already exists.

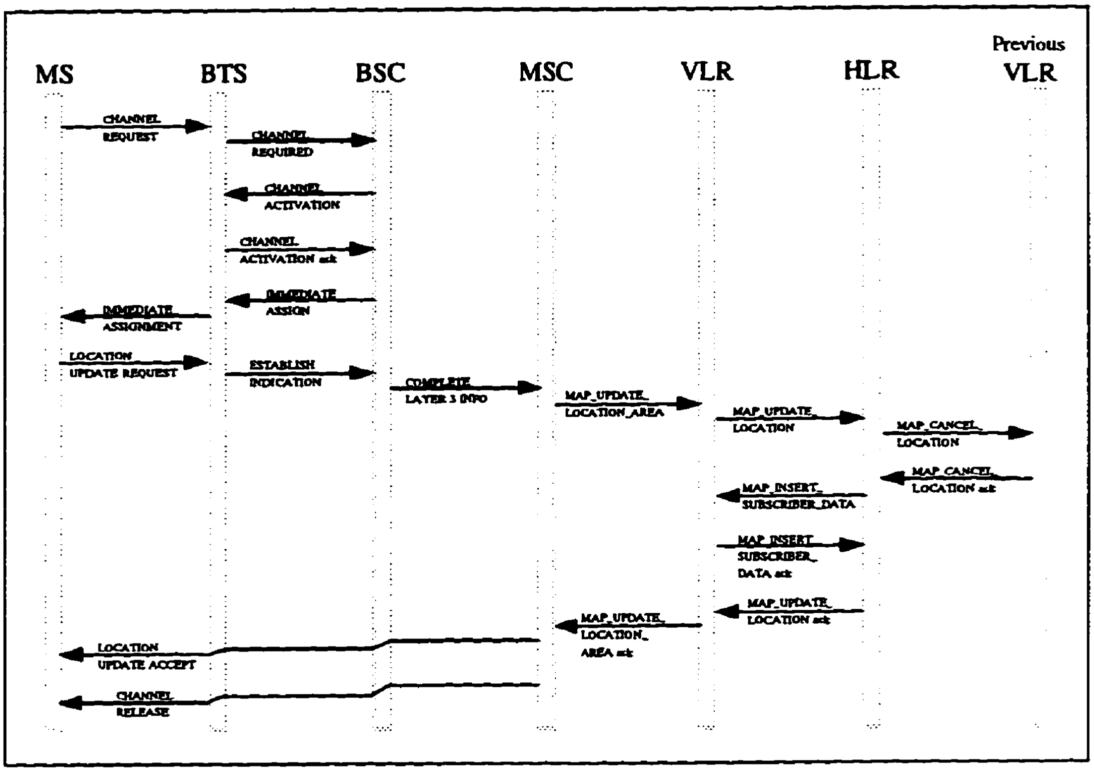

Figure 2.2: Message exchange for GSM location updates

The steps involved, and messages exchanged, for location updating are illustrated in Figure 2.2, and explained below.

1. The mobile station must first be synchronized to the base station, to be able to receive and decode the Broadcast Control Channel (BCCH). The BCCH broadcasts information necessary for the mobile station to be able to properly access the network, including the identity of the network, location area, and cell, information for handover and cell selection, information describing the current control channel

structure, parameters for properly transmitting on the random access channel, and other cell options.

2. The mobile station transmits a CHANNEL REQUEST message on the Random Access Channel (RACH). This message contains only 8 information bits in total, and includes the reason for the mobile access (3 to 6 bits) and a random discriminator used to prevent duplicate channel assignments. In this case, the reason for access is location updating.

3. If there is no collision of the access burst, and the mobile station is not barred from accessing the location area or cell, the Base Transceiver Station (BTS) measures the transmission delay, and sends a CHANNEL REQUIRED message to the Base Station Controller (BSC).

4. The BSC selects a signalling channel (possibly full rate) and sends its description in a CHANNEL ACTIVATION to the BTS. If the BTS is able to allocate that channel, it responds with a positive CHANNEL ACTIVATION acknowledgment.

5. If the BTS channel activation was successful, the BSC will forward an IMMEDIATE ASSIGN command to the BTS, which forwards it to the mobile station as an IMMEDIATE ASSIGNMENT message, describing the assigned radio channel, on the Access Grant Channel (AGCH).

6. The mobile station accesses the assigned channel, using the given timing advance and power control parameters, and transmits an initial message indicating its identity (IMSI or TMSI), the mobile equipment classmark (indicating its capabilities), and the type of location update in a LOCATION UPDATE REQUEST message. (This layer 3 message is actually piggy-backed on a layer 2 frame, whose acknowledgment establishes the layer 2 connection and resolves any access contention issues

that may exist.)

7. The BTS sends an ESTABLISH INDICATION message to the BSC confirming the setup of the channel, and carrying the layer 3 location update request information.
8. The BSC sets up an SCCP (Signalling System No. 7) connection with the Mobile services Switching Center (MSC). At this point, the mobile station can communicate with the MSC. Communication between the MSC and the mobile station is forwarded transparently by the base station subsystem. The BSC forwards a COMPLETE LAYER 3 INFO message to the MSC, which contains the location update request information from the mobile station.
9. The MSC uses the Mobile Application Protocol (MAP) [16] to communicate with the appropriate location registers to complete the location update. The MSC sends a MAP_UPDATE_LOCATION_AREA message (service request, technically) to its VLR, which includes the information received from the mobile station, such as the IMSI or TMSI, the location update type, and the new location area identifier.
10. At this point, there are a few optional procedures that may be invoked. These include authentication of the mobile, activation of the ciphering mode, and verification of the International Mobile Equipment Identity (IMEI) status to ensure that the mobile station itself is not listed as being stolen or incompatible. These procedures will not be described here, but their presence, and additional signalling load, should be noted.
11. If the mobile station is allowed to register with the VLR, the subscriber's HLR is notified of the change in VLR area through the MAP_UPDATE_LOCATION service request. Some variations are possible here, such as if the mobile station included its TMSI in the location update request the previous VLR would be queried to obtain

the subscriber's IMSI (since TMSI is only valid in conjunction with a Location Area Identifier). In that case, subsequent procedures, as described here, would be slightly modified, since a TMSI reallocation may be invoked. The service indication to the HLR includes the IMSI and VLR address.

12. Upon receiving the MAP_UPDATE_LOCATION service indication, the HLR sends a MAP_CANCEL_LOCATION request to the subscriber's previous VLR. This deletes any subscriber information from the previous VLR, which then sends a MAP_CANCEL_LOCATION acknowledgment to the HLR.

13. The HLR then sends a MAP_INSERT_SUBSCRIBER_DATA request to the new VLR. Clearly, if the subscriber registers for the first time in the new VLR, no prior subscription information will exist. This request sends information such as the subscriber's MSISDN, the type of subscribed services, details on subscribed supplementary services (such as call forwarding address), and any restrictions on the subscriber (such as roaming, or operator determined barring). This information is necessary for service provision by the new VLR and MSC. Upon receipt of this information, the request will be acknowledged by the VLR.

14. Once the subscription information is successfully acknowledged by the new VLR, the HLR will have completed the location updating, and will acknowledge the MAP_UPDATE_LOCATION request previously sent by the new VLR, which in turn will acknowledge the previous MAP_UPDATE_LOCATION_AREA request that the VLR had received from the MSC.

15. Upon completion of the network side location updating functions, indicated by the MAP_UPDATE_LOCATION_AREA acknowledgment, the MSC will send a LOCATION UPDATE ACCEPT message to the mobile station. This message is transpar-

ently forwarded by the BSC and BTS, using the Direct Transfer Application Part (DTAP) over the existing dedicated signalling connection to the mobile station.

16. The location update is now complete. Unless the mobile station had previously indicated that it wants to maintain the existing signalling link, or the network requires it for any additional signalling, the link is released when the mobile station receives a CHANNEL RELEASE request from the network (or the mobile station's timer for receiving this message expires).

2.3.2 Interrogation and paging

By definition, a mobile subscriber is not directly associated with a fixed telephone link. When an incoming call from the fixed network (PSTN) is directed at the mobile network, the currently serving MSC for the target subscriber must be found, and a connection must be established to that switch for the duration of the call. This is the general call routing procedure, and interrogation is the mobile network function that determines what MSC the incoming call should be routed to.

As described previously, the mobile network maintains the current location of the subscriber in the HLR. Typically, because of legal, administrative, or technical reasons, only certain switches are designated as Gateway MSCs (GMSC) which are able to directly query the HLR to obtain routing information for mobile terminating calls. GMSCs, usually in the destination subscriber's home mobile network, receive incoming calls from the PSTN, query the appropriate HLR for the intended subscriber's current location, and forward the call to the proper destination MSC. Although not technically impossible, PSTN switches are generally not able to directly query GSM network databases, such as the HLR.

As for the description of location updates, the mobile terminating call setup procedure

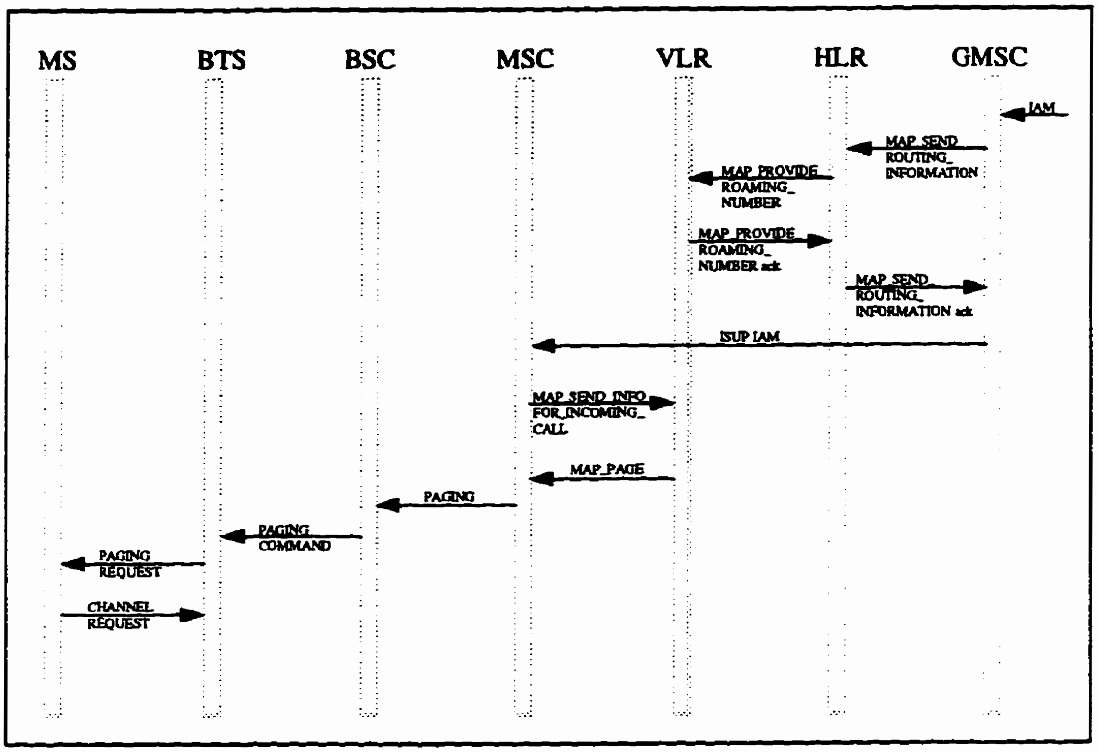

Figure 2.3: Message exchange for GSM interrogation and paging

described below and shown in Figure 2.3, is for the most general case and assumes a successful outcome. Although it does not affect the general procedure, it is assumed that the ISDN User Part (ISUP) is used for PSTN signalling. In certain cases, some of the steps may be unnecessary, such as when a mobile-to-mobile call is set up between two subscribers under the same MSC. Note also that the call setup procedure after the mobile subscriber has been successfully paged is not described below, since it is beyond the scope of location management.

1. The originating PSTN switch will be able to derive the address of the GMSC from the dialled digits (the intended mobile subscriber's MSISDN), and will send an ISUP Initial Address Message (IAM) to the GMSC indicating an incoming call.

2. The GMSC is able to derive the destination subscriber's HLR from the first few digits of the MSISDN. The GMSC will send a MAP_SEND_ROUTING_INFORMATION request, with the destination MSISDN, to the HLR. The GMSC ultimately requires a roaming number (MSRN) which it can use to route the call to the destination MSC.

3. The HLR maps the MSISDN to the IMSI, and looks up the current location of the destination subscriber. This address is normally in the form of the address of the subscriber's current VLR. The HLR sends a MAP_PROVIDE_ROAMING_NUMBER request to the current VLR, giving the IMSI of the subscriber and the current MSC (which is generally ignored, since there is in practice a one-to-one relationship between VLR and MSC).

4. The VLR allocates a roaming number (MSRN) from its assigned pool, and maps the IMSI to the MSRN. A response, with the allocated roaming number, to the MAP_PROVIDE_ROAMING_NUMBER request is sent to the HLR.

5. The HLR forwards the MSRN to the GMSC in response to the MAP_SEND_ROUTING_INFORMATION request. The GMSC is then able to continue with the call routing, and sends an IAM to the destination MSC, with the new MSRN.

6. Upon receipt of the IAM, the MSC sends a MAP_SEND_INFO_FOR_INCOMING_CALL request to the VLR, which indicates the MSRN and any other information that may have been included in the original IAM message, such as bearer service type. A response to this request is not forwarded to the MSC until the subscriber

has successfully responded to the paging.

7. Receipt of the MAP_SEND_INFO_FOR_INCOMING_CALL request triggers the VLR to issue a MAP_PAGE request back to the MSC. This request to page the mobile station includes the IMSI of the destination subscriber and the location area identifier that is stored in the VLR.

8. The MSC sends a PAGING message to the appropriate BSCs, which includes the IMSI, TMSI (if one has been allocated), a list of the cells to be paged, and, if applicable and known, the type of channel required for the incoming call (e.g., speech or data).

9. The BSC sends PAGING COMMAND messages to each BTS specified in the previous PAGING message. The PAGING COMMAND message includes the identity of the mobile subscriber (IMSI or TMSI) to be used, the *paging group* of the subscriber, and the type of channel required, if applicable. Paging groups are sections of the complete paging channel, in which all paging messages for particular groups of subscribers are normally sent. Since mobile stations, in general, need to listen only to their own paging group, and a paging group is much shorter (in terms of the number of transmitted timeslots) than the complete paging channel, less time is spent listening to the paging channel, thus conserving battery power. This concept is known as discontinuous reception.

10. There are three types of PAGING REQUEST messages that may be transmitted over the radio interface. These differ in the type of identifier (IMSI or TMSI) used and the number of paging requests that may be sent in one message. All messages are 23 octets (184 bits) long, and since a TMSI is shorter than an IMSI, more TMSI paging requests can be sent per message (up to 4 with a PAGING REQUEST TYPE

3 message). Due to the unpredictability of the radio environment, paging messages are normally repeated an operator-configurable number of times, to decrease the chance of lost pages (for example, if the subscriber was temporarily in a shadowed area).

11. When the mobile station successfully receives a paging message, it initiates a channel assignment procedure, identical to the first few steps of the location update procedure, with the exception that the establishment cause is 'response to paging' instead of 'location update'. A context is maintained on the network side to complete the circuit for the call once the mobile station has responded to the page and a radio signalling link has been established.

2.3.3 Conclusions

The detailed operation of the location management procedures in a practical mobile network, specifically a GSM network, serves to show the considerations that must be taken into account when developing new location management algorithms. For example, many proposals do not consider that at present, and likely for several years, mobile networks and fixed networks are disjoint. They do not have integrated switching and signalling functions, due not only to technical constraints, but also legal and administrative ones.

Other factors which should be clearer after this review include the fact that a mobile network must store certain subscription information in a database, regardless of the location management algorithm used. Even in an Intelligent Network architecture, a switch requires that information in order to properly provide service. Certain location management proposals ignore this detail when deciding to implement distributed databases with heavy replication.

Some of these practical issues will be discussed in the literature review in Chapter 3.

At this point, it should be noted that location updating, interrogation, and paging are relatively complex in terms of signalling on both the radio interface and within the fixed network, and that there are several factors that should be considered for an efficient location management scheme.

Chapter 3

Previous work on location management

Location management has received a fair amount of attention in recent research literature, as higher capacity cellular systems are being developed. Some of the recent proposals are reviewed in this chapter. A survey of the different location management algorithms is given, together with the mobility models used for their evaluation.

The proposals are evaluated in terms of the overall goals of location management:

- Minimize signalling, both over the radio interface and within the network, to reduce bandwidth and increase capacity
- Minimize call setup delays (especially due to paging procedures) in order to maintain acceptable quality of service
- Minimize the amount of data stored for each subscriber, to reduce equipment costs
- Minimize the computational complexity of algorithms, to reduce hardware costs, conserve battery power, and improve response times

- Maximize the applicability of the algorithms, by avoiding reliance on estimated parameters (such as user classes)
- Distribute location update requests evenly among cells, rather than concentrating them in cells on the border of location areas

In general, these are conflicting goals, and trade-offs must be made. Typically, proposals concentrate on improving one particular aspect by compromising another. For practical systems, a balance must be found, such as, for example, increasing the computational complexity and data storage requirements to decrease signalling requirements, and thus increase system capacity.

The overall location management functionality is comprised of the three inter-related aspects of location updating, interrogation, and paging. Different researchers have concentrated on different aspects, and they have assigned different priorities to different system resources (such as database storage capacity) and quality of service parameters (such as call setup delay).

Location management proposals fall into two broad categories: those that concentrate on the interrogation aspect of location management, and those that focus on location updating and paging algorithms. The two categories are not necessarily mutually exclusive, and may be combined in a comprehensive strategy.

3.1 Interrogation-based proposals

Several proposals have attempted to improve the signalling and querying efficiency of the network-side database querying required by location management, with a secondary goal of reducing radio interface signalling in terms of paging. The driving requirement is to remove the bottleneck that may arise as a result of a centralized Home Location Register (HLR), which must be queried for every incoming call and every location update associated with a subscriber. Replacing the two-tier HLR/VLR (Visitor Location Register) scheme used in current systems like GSM, a hierarchical location register structure is proposed in most of these proposals. These location registers are, in general, closely coupled with the fixed network, an assumption which does not hold in current networks like GSM, in which an intermediary Gateway MSC (GMSC), which is part of the mobile network, is accessed by the fixed network in order to query the HLR.

Jain et al. In [20], the authors propose a per-user caching scheme, given an Intelligent Network (IN) architecture of linked switches and databases. In general terms, an IN architecture separates the switching and the logic in the network. The intelligence is located in Service Control Points (SCPs) which the switches, or Service Switching Points (SSPs), can query for additional control information as required by a call. The 800 toll-free service, for example, is implemented using IN principles. A pointer to the called subscriber's last known VLR is stored in the call originating switch. The goal is to avoid querying the HLR, which in turn queries the current VLR to obtain the roaming number, and just query the VLR directly. If the subscriber has moved out of the cached VLR area, a normal HLR query will have to be performed. The fixed and mobile network architectures are assumed to be sufficiently integrated to support the direct VLR querying.

Although it may work well in certain cases, this approach is not generally applicable. A high Local Call to Mobility Ratio (LCMR) is required, meaning a subscriber receives many calls from a specific switch relative to the number of location updates performed. If the LCMR is below some threshold, database storage at the switch is wasted, and total setup delay increases due to the 'cache miss' when an out-of-date VLR is queried.

Wang In [48], a hierarchical Intelligent Network (IN) architecture is assumed, with each node in the hierarchy being associated with a location database. The highest level nodes may represent countries, possibly under a common root, followed by regions, area codes and cities, with the leaves of the hierarchy representing location areas.

Subscribers belong to a particular *class*, which is the number of the layer in the hierarchy that contains a node which subtends *all* of a subscriber's service areas. A service area is a particular geographical area, corresponding to one of the nodes, in which the subscriber elected to obtain service. Service areas do not have to be from the same layer.

When a subscriber registers in a location area, a series of database entries is created up to the node corresponding to the subscriber's class. A parent node will contain a record for the subscriber with a pointer to the child node under which the subscriber is located. When a subscriber moves to a new node, a message is sent to the new node, which creates a new series of location pointers up to the node which subtends the tree containing the old and new location areas. Another message is sent to the old node, which marks the previous location entries to indicate that the new location node of the subscriber is not subtended by that particular node. Other details are included in the algorithm to remove unnecessary nodes, and to deregister the mobile station when it moves out of the selected service areas.

This proposal is one of several possible interrogation strategies outlined in [10] which are under consideration for the Universal Mobile Telecommunication system (UMTS) [41], the next-generation European wireless system. Using a hierarchical location register architecture, there are three general strategies:

1. Queries could be directed to an HLR, which would contain the current location register address.

2. Queries could be sent directly to the current VLR, which means pointers would always need to be propagated in the tree.

3. Queries could be initially directed to the HLR, but if a reference to the current VLR is found along the way, that information is used to access the VLR directly, without querying the HLR.

There are different trade-offs in terms of database storage and querying delay between the different strategies, and in addition, different caching and replication schemes may be used. Wang's proposal was one implementation of the last strategy. However, the transfer of subscription information to the current node, a critical task, was not considered in Wang's proposal, which concentrated only on the location information. As a result, the required database storage calculations were not completely accurate.

Badrinath, Imielinski, and Virmani The proposal in [3] is a relatively short description of work in progress. The algorithm assumes a Personal Communications Network, which is basically a hierarchical, integrated Intelligent Network architecture. It describes in general the concept of *partitions*, which are subtrees containing location servers (corresponding to location areas) most often visited by subscribers, as specified by their user profile. Three different general interrogation algorithms are described, and the effect of partitions on them is investigated. The interrogation algorithms included flat (search the entire network hierarchy), expanding (search the home location server, and proceed up the hierarchy if unsuccessful), and hybrid (search the home location server, then the parent, then the entire network) schemes.

The algorithm relies on partitions to determine the location of the subscriber, thus avoiding location pointers and their associated maintenance in the location register hierarchy. The trade-off is that if the subscriber roams over a large area, paging volume may be excessive, and the extra paging steps could create long call setup delays. However, this paper was quite brief, and details on the creation and maintenance of the partitions were not presented.

Shivakumar and Widom The proposal by [43] discusses replication of the user profile, containing the subscriber's current location, in several location registers. Assuming an integrated Intelligent Network architecture, outgoing calls from a particular zone (area controlled by a VLR) would first check the local VLR, in an attempt to avoid querying a distant centralized database, i.e., the HLR, for the current location of a specific mobile subscriber. The decision of where to place the replicated profiles, given the costs for updating the replicated records, is based on a minimum-cost, maximum-flow algorithm described in the paper. The HLR still maintains the current location of its mobile subscribers, and manages the updating of all their replicated records.

The algorithm requires the construction of a flow network, which is a graph containing nodes for each subscriber and each zone in the network. The edges between the subscriber nodes and the zone nodes have attributes which require estimates for the number of calls from a particular zone to a particular subscriber, as well as the average number of location updates performed by a particular subscriber. These estimates, together with estimates on the replication costs and database capacities, are used by the minimum-cost maximum-flow algorithm to determine for which subscribers, and in which nodes, replication of the user profile would be cost-efficient. The possibility of varying mobility and call arrival patterns is considered, and four algorithms are presented for modifying the flow network and minimum-cost, maximum-flow algorithm to avoid having to recompute the entire algorithm, which is exponentially bound.

3.2 Location updating and paging proposals

Another approach to minimizing the total location management cost focuses on reducing the overall number of location updating and paging messages sent over the air interface. The location register architecture is generally of no concern, and the assumption of a

GSM-like HLR/VLR mechanism would be sufficient.

Xie, Tabbane, and Goodman In [49], the authors propose an algorithm which uses individual location area sizes based on subscriber characteristics. Using a somewhat unrealistic square grid representing the cellular network, the algorithm is based on a variable k, which represents the length, in cells, of the sides of a square location area. They assume that the cells can be described using an (x_n, y_n) ordered pair, and that an initial location of (x_i, y_i) and initial value of k are stored by both the mobile station and the network. If the mobile station enters a new cell (x_j, y_j) such that $\lfloor(x_j/k)\rfloor \neq \lfloor(x_i/k)\rfloor$ or $\lfloor(y_j/k)\rfloor \neq \lfloor(y_i/k)\rfloor$, then the mobile station performs a location update, storing the new values of (x_j, y_j) and also transmitting them to the network. When an incoming call arrives, the network pages successively in cells (a, b) where $a = (x_j, x_j + 1, ..., x_j + k - 1)$ and $b = (y_j, y_j + 1, ..., y_j + k - 1)$. As described, successively paging, on average, $\frac{k}{2}^2$ cells can introduce a long call setup delay, especially since paging messages are typically repeated.

Details were not given on how to practically calculate the mobility parameter required for the optimization algorithm. In addition, upon updating, the subscriber is apparently at a corner of the new location area; a more appropriate location would be closer to the center, in case the subscriber reverses direction.

Okasaka et al. The proposal in [35] is described as two separate but related algorithms which address the problem of location management radio signalling being concentrated only on border cells of location areas, as well as the problem of thrashing between location areas along location area borders. The first algorithm, multi-grouping, arranges cells into several conceptual layers of location areas, with location area boundaries of one layer staggered in relation to other layers. Mobile stations are also divided into several groups, and each group is assigned to one or more layers. On its own, this proposal means that mobile stations belonging to different groups observe different location areas, and thus perform location updates in different cells. This distributes location updating traffic relatively evenly amongst different cells.

The second algorithm, multi-switching, improves upon the multi-grouping algorithm by making the mobile station switch to another location area *layer* (within the set of several layers allocated to its group) for every location update. For example, assume mobile station M belongs in a group to which location area layers L_1, L_2, and L_3 are assigned. If M is currently in layer L_2 and performs a location update, M will shift to the corresponding location area in layer L_3. If the layers are staggered in relation to each other, then after the location update, the mobile station will effectively be moved from the edge to the middle of a new location area. With careful staggering of layers, and assignment of cells to location areas, this algorithm reduces the problem of thrashing at location area boundaries, in addition to evenly distributing location updating traffic.

Since subscribers frequently change short-term mobility behaviour or calling patterns, a mechanism, which was not discussed, must be provided to transfer subscribers from one group to another, or the problems of improperly matched mobility patterns and location area sizes, discussed above, would arise. A related issue is that cells are still statically assigned to location areas, and location areas are statically and heuristically grouped into layers, with the associated inefficiencies discussed earlier.

Bar-Noy, Kessler, and Sidi The authors in [4] compare three different location updating strategies, using two different mobility models. The updating strategies include a time-based updating strategy, in which the mobile station sends a location update after a time interval T. Another strategy has the mobile send a location update after it detects M cell crossings. Finally, a distance-based strategy is considered, where the mobile station sends a location update after it detects that the distance (in terms of cells) between its current cell and the cell in which the last location update was sent, exceeds some parameter D.

In their comparison of the different update strategies, using two somewhat similar mobility models, the authors claim that the distance-based method is better than the other two, in terms of the expected search cost for equal expected update costs. The search (or paging) algorithm consists of searching concentrically until the subscriber is found. If the value of D or M is set relatively high (for efficient updating), the paging cost would be prohibitive. Finally, no mention is made of how a distance-based strategy would be implemented.

Plassmann The article by Plassmann [37] is an overview article discussing the topic of location management, and describes a number of different approaches, from the most basic to some of the more recent proposals discussed in this chapter. A general cost analysis formula was given, but exact numbers were not substituted from any of the proposals. A new proposal was very briefly described, which was a combination of two approaches described in the article. The mobile keeps a record of the number of times each cell is visited, and the number of incoming calls at each cell. The product of these two values, p, results in a home area of high values of p, which is divided into several paging areas, and an outside area of relatively low values of p, in which the mobile terminal performs normal location updates. The details of algorithm are not well-defined, however.

Tabbane Tabbane has written several papers describing different versions of his general scheme [33, 38, 44, 45]. The most comprehensive [46] description is reviewed here. In the proposal, user profiles are maintained for each subscriber. For each time period T, an ordered set of tuples (a_j, p_j) is stored, where a_j is a paging area j, and p_j is the probability of locating the user in paging area j. The paging area can be comprised of one or more cells. The paging areas are ordered by decreasing probability p_j. If the subscriber is called during some time period T, the subscriber will be paged sequentially, or in parallel, in the paging areas specified in the user profile for time period T. The set of paging areas specified for a particular time period in the user profile can be thought of as a location area. If the subscriber roams outside of this location area, a normal location update will be performed.

The user profile may be built up by monitoring the user over some period of time. This long-term user profile seldom changes, although the actual updating procedures are not specified. The location accuracy may be improved by reordering the paging areas in the long-term profile according to short-term or medium-term algorithms, or by paging concentrically around the last known location area. The former approach attempts to reorder the paging areas so that areas associated with a higher probability of containing the subscriber are paged first, reducing signalling and delay. Call arrival and origination rates from particular paging areas may be used for a medium-term reordering of the user profile. Alternatively, estimated distance travelled since the last known location may be used at the time of an incoming call to determine which areas to page first.

The contents of the user profile are outlined in the paper, but by ranking likely location areas for each arbitrarily-defined time period, more data and complexity are introduced than may be necessary. In addition, the methods for initializing the user profile, as well as long-term updates, are not clearly specified. Some of the proposed methods for short-term updates to the user profile, such as paging based on estimated distance travelled since the last known location, may be difficult to implement without additional topological knowledge in the network. Despite some unresolved issues, the general approach seems efficient and practical, and the proposal described in Chapter 4 is most closely related to this type of approach.

Madhow, Honig, and Steiglitz The discussion in [29, 30] is quite theoretical and mathematically intensive, although the initial assumptions are fairly simple. On a per-user basis, the expectation of the total paging and updating costs, from the time the last call arrived (and hence the last known location) until the next call arrival, is derived given the updating cost, the updating policy (whether to update or not as a function of the relative position since the last known location), and a paging cost function, which depends on the position at the time of the next call arrival. The solution to the dynamic programming equations given took from hundreds to tens of thousands of iterations to converge — such computationally-intensive algorithms are not particularly well-suited to current mobile stations.

Cho and Marshall Location management also applies to the area of mobile computing and IP routing. The proposal in [7] uses the concept of a *local region*, which includes the mobile agents and mobile routers most often visited by a given mobile host, and possibly source hosts which frequently interact with the mobile host. In addition, frequent source hosts outside the local region are maintained in a patron list. The mobile agents in a local region are hierarchically arranged in a redirection tree, with a redirection agent at the root which maintains the current location of the served mobile hosts. Packets to a served mobile host are intercepted by the redirection agent, and forwarded to the current location of the mobile host. When a mobile host leaves the local region, or re-enters it, messages are sent to the members of the patron list, notifying them of the mobile host's new location. This is done only for the first outside location area — forwarding pointers are used from that point on, until the mobile station re-enters the local region. More than one local region per mobile host is possible. The procedure for creating the local region is described vaguely in terms of monitoring host movements and accesses, or by getting it from the subscriber in advance.

3.3 Proposal evaluation and mobility models

It is evident that in order to quantitatively and realistically assess any location management scheme, an accurate and plausible mobility model must be used. A mobility model is an analytical or statistical representation of the daily movements of mobile subscribers. There have been several different approaches to mobility modelling for location management, and some of the methods used in the literature are reviewed.

Location update rates, call to mobility ratios, and other metrics used in the proposals depend on some description or model of subscriber movements between location areas. In many cases, some type of random mobility model is used. In certain papers, where the mobility model is explicitly required in mathematical derivations, a one-dimensional model is used to simplify the calculations. Average velocity and other global parameters are sometimes used.

Jain et al. The analysis in [20] depends on the Local Call to Mobility Ratio (LCMR), defined as the number of calls from a particular switch to a particular subscriber, relative to the number of location updates performed by that subscriber. An analytical approach assuming Poisson call arrivals and exponential roaming times in VLR areas was used to determine a threshold for LCMR. Use of this approach over normal HLR querying was justified only when the value of LCMR crossed the threshold.

Xie, Tabbane, and Goodman The evaluation of [49] uses a general cost function for paging and location updating based on the call arrival rate a and the location updating rate u, respectively, and the square location areas with sides of length k cells. The location updating rate u can be reduced to $\frac{b}{k}$, where b incorporates the expected velocity for random mobility. Using the normalized cost function of the call arrival rate a, and the variable b, an optimum value for k can be calculated. The evaluation of the algorithm uses different call arrival probability functions and plots normalized cost functions comparing the proposed scheme with a fixed scheme.

Okasaka et al. The mobility model used to evaluate the proposed algorithms in [35] was very simple, consisting of a square grid of cells and one subscriber moving at constant velocity (speed and direction). The main purpose of the analysis was to evaluate the performance of the protocol with respect to avoiding thrashing between cells and location areas due to radio propagation effects.

Wang The analysis in [48] does not use an actual mobility model, but makes assumptions on the probabilities that a subscriber will be roaming outside of a particular layer of their service area (the probabilities get smaller for higher layers, as the subtended area increases). The number of database entries that would be required with this scheme was then calculated, and the author concluded that for a given number of subscribers, and given the estimated probabilities described above, the total number of database entries is significantly less than for a centralized system. However, the calculation is flawed because the author does not take into account the storage requirements for subscription information. A further point of comparison is the updating and querying cost, calculated as the product of the mean distance travelled by signalling messages between the different nodes and the average number of database entries for different categories of subscribers. This last metric is fairly insignificant, since databases are addressed directly (and electrical signals travel close to the speed of light).

Badrinath, Imielinski, and Virmani In the evaluation of [3], the metric used is total cost in terms of the subscriber's call to mobility ratio (incoming calls to location updates). The mean times between successive incoming calls and successive location updates are given by exponentially distributed random variables. Total cost is comprised of update cost, in terms of the number of links traversed, and paging cost, in terms of the number of location servers paged. The (briefly described) simulation results compare the total cost for the three interrogation algorithms, with and without the use of partitions, for a number of different call to mobility ratios.

Shivakumar and Widom The algorithm in [43] is analytically compared to [20] and another similar approach, based on the network query size, and the location lookup delay. An actual mobility model is not used to obtain any quantitative results, but rather variables are used to represent the number of location updates and the number of incoming calls from a particular zone, for generic subscribers.

Madhow, Honig, and Steiglitz The mobility model used in [29, 30] is a simple, discrete-time, one-dimensional random walk. During each time period, there is a probability p that the mobile station moves one cell to the right or one cell to the left, and thus a probability of $q = (1 - 2p)$ that the mobile remains in the current cell.

One of the two mobility models used in [4] is similar to the one used in [29, 30]. The model considers so-called independent and identically distributed (IID) movements. Time is slotted, and during any slot there is a certain probability p that the mobile station will move to the right or to the left, and thus a probability $(1-2p)$ that the mobile station will remain in the current cell, independently of previous movements. The second mobility model considered in [4] is a three-state Markovian model. The three states are $right$ (the mobile station moves one cell to the right of the current cell), $stay$ (the mobile station stays in its current cell), and $left$ (the mobile station moves one cell to the left of the current cell). Different probabilities are assigned to the nine possible state transitions.

Cho and Marshall The analysis of the scheme in [7] uses assumed values for parameters such as bandwidth, latency, and packet size, as well as a call to mobility ratio and other parameters related to the calling patterns of patron hosts and sources in the local region. The scheme is compared to a basic scheme in terms of packet transmission time relative to given call to mobility ratios.

Tabbane An analytical evaluation of Tabbane's proposal [46] is given which assumes typical values for certain parameters, such as cell size, average subscriber velocity, and average number of call arrivals and call originations. The size and number of signalling messages generated over each interface per unit time is the chosen comparison metric. A simulation is also used for comparison. The mobility model is a simple one, with a subscriber moving with an average velocity and random direction, having a certain probability of remaining in a certain paging area. The call arrival and origination rates are uniformly distributed. The Mobility Predictability Level (MPL) is a key parameter used in the comparisons to give an estimate of the randomness of the mobility patterns, and is given by $MPL = \sum_{i=1}^{k} p_i$, where k is the number of paging areas for the current time period, and p_i is the probability of locating the subscriber in area i. Results were plotted comparing the number of messages across various network interfaces, and the ratio of total cost of the proposed algorithm versus the standard GSM algorithm, as a function of the MPL.

Seskar et al. In [42], the authors propose a traffic model which simulates vehicle movement based on the relationships between vehicle speed, vehicle density per street length, and volume of vehicles. By defining the street layout, and probabilities of turning at intersections, the number of vehicles crossing at particular points can be estimated by a simulation. Even if overall car crossings can be accurately determined, however, the movements of cellular subscribers are not obvious. Subscribers may stop somewhere for an extended period of time, even if traffic does not. In addition, as cellular systems grow, increasing numbers of mobile stations are no longer used exclusively in vehicles, but are being used by pedestrians. Pedestrians are not confined to street grids, and their average velocities are not easily derived.

MONET Group A comprehensive discussion of mobility modelling is presented in [40], where alternative approaches to updating, paging, and interrogation are discussed and contrasted. Although the report places emphasis on UMTS procedures and concepts (such as Customer Premises Networks), generally applicable ideas are presented concerning mobility management. In defining their mobility model, detailed actual measurements on parameters related to pedestrian, private automobile, and public transportation movements were gathered. The measurements were intended as inputs to simulations for various types of users. The measurements are used in one example comparison of traffic flow, and results between theory and practice were compared, with moderate agreement.

3.4 Conclusions

At best, the proposals above used some version of a random mobility model, typically one-dimensional. Clearly, such simplifications for as critical a component of location management as subscriber mobility, are not satisfactory. A random mobility model fails to take into account certain aspects of travel such as trip purpose, and so-called attraction zones, where people spend a large proportion of their time. Random mobility models may have provided a rough model for certain early users of cellular telephones, such as travelling salespersons, whose movements were fairly random. However, as wireless technology increasingly becomes a consumer item, such mobility patterns will no longer reflect reality.

In Chapter 4, some of the problems associated with random mobility models are described, and a new mobility model is proposed. Mobility of people, in general, is a topic which is of concern not only to cellular network planners, but to city planners, transportation engineers, and others. Transportation theory is a complex area, one where there are competing theories, and which is still developing. However, certain techniques and ideas can be drawn from work in travel behaviour and traffic planning. A new location management algorithm will be presented in Chapter 5, in which the mobility patterns of average subscribers form the basis of the proposed updating and paging strategy.

Chapter 4

Mobility model description and simulation design

A mobility model, in the context of location management, is a model of the daily movements of mobile subscribers, or more precisely, the daily movements of registered mobile stations. Such a model is of paramount importance in mobility management studies. The number of paging and location update messages, required for comparisons of the efficiency of different location management schemes, depends fundamentally on user mobility patterns.

There have been several proposals for location management algorithms in the recent past, to improve upon the current method of using statically-defined location areas. In attempting to quantify the performance of their proposals, most authors have used mobility models which are overly simplified and unrealistic, completely random, or dependent upon many assumed 'typical' values for key variables.

Different location management proposals react differently to the randomness of user behaviour. A mobility model is needed which is more realistic than a random mobility model. Random mobility models have been frequently used in the past due to their simplicity, and the fact that early users of cellular communications had typically been very mobile subscribers whose behaviour could be relatively well-modelled as random. With the tremendous growth of cellular communications, random mobility models are no longer accurate for the majority of subscribers, whose behaviour is more routine. A random mobility model would cause very biased results for proposals which attempt to utilize the behavioural aspects of subscriber mobility — for example, the importance of the home and workplace as trip origins and destinations.

An overview of travel demand models and transportation theory from the field of transportation planning is presented. The proposed activity-based mobility model uses some of the ideas from transportation theory, and is described in the latter part of the chapter.

4.1 Transportation planning models

In the areas of transportation planning and traffic engineering, transport models are used to describe and forecast *travel demand*, usually in terms of trips taken by households, and typically concentrating on urban areas. There are many different models in use — the particular model chosen depends on many variables, of which an important one is the availability of appropriate data, as well as the particular application of the model. The main purpose of most transport models is to analyze observed urban travel behaviour and its relationship to certain socioeconomic variables, in an effort to forecast future travel behaviour in the face of population growth and upon implementation of different transport policies or services. These forecasts form the basis for decisions involving transportation system investments, urban planning and zoning, and public transit policies.

A basic element of transportation modelling is the *trip*. A trip is usually defined as a one-way movement of one person from an origin to a destination, utilizing one mode of transportation. Trips are often categorized by the purpose behind the trip, such as work, shopping, and return home.

Other important concepts in transportation modelling include the concept of *traffic zones*. Traffic zones represent the spatial structure of a geographical area, such as a municipality. They aggregate the location and other factors of individuals or households into usually no more than a few hundred traffic zones, which can range in size from a few hundred square meters, to several square kilometers. Trips are assumed to originate and terminate at a single fixed point within a traffic zone called the *centroid*. Various socioeconomic variables such as employment or retail space can be measured per traffic zone. The definition of zones depends on the particular goals of a transportation study, and is usually a compromise between representational accuracy, analytical complexity, and availability of data.

4.1.1 Elements of travel demand

There are several different approaches to travel demand modelling. All integrated models, however, attempt to describe and forecast the four basic components of travel demand [5, 8, 14]. Some methods view and model each component separately, while others combine two or more steps.

Trip generation The first step consists of estimating the total number of trips T_i originating from zone i and/or the total number of trips T_j terminating at zone j. In practice, the trips are classified by trip purpose, such as work, shopping, or social/recreation, and are often grouped into home-based or non-home-based trips, depending on whether one end of the trip is at the home. Trip generation may be based on zonal averages, or it may be based on type of household.

Trip distribution Given the total number of trips originating in a given zone T_i, this stage calculates T_{ij}, the number of trips from each zone i to every other zone j. The number of trips entering a particular zone can similarly be distributed to all other origin zones.

Modal split Mode refers to the different types of transportation mode, such as car or transit. Modal split determines the relative ratio of the different modes used for T_{ij}. Although an important component of urban transport modelling, mode was not taken into account in the mobility model described here. The mode has major implications for handover, due to the different speeds of different modes, but it does not affect location management to the same degree. The path taken is the main factor for location management, and though it may be influenced by the mode taken (for example, a bus may take a different route from point A to point B than a car or pedestrian would have taken), the extra complexity required was not justified. This is particularly true when modelling a medium-sized area such as the Region of Waterloo, which does not have a subway system or underground pedestrian pathways, which would be examples of modes of transport using unique paths.

Trip assignment Given the number of trips T_{ijm} from a zone i to a zone j using a mode m, this step maps the trip route onto the existing or proposed transportation infrastructure. The selected route is typically the one with the smallest generalized cost, usually some combination of time, distance, speed, and convenience.

4.1.2 Aggregate models

Aggregate models are the earliest of the travel demand models, and are still being used in practice. They attempt to model the behaviour of aggregations of the population, based on average socioeconomic indicators, such as age and income, on a per traffic zone or per household type basis. The classical transport model is the most basic aggregate model, with many modifications having been made in different studies.

The classical transport model

The classical transport model is an aggregate model which has been used since the 1960s to forecast travel demand. Various techniques and submodels exist for determining the variables for each of the four elements of travel demand. Some of the submodels are described below [8, 14, 36].

Trip generation One method used in trip generation is (usually linear) regression analysis, applied on survey and census data, where the dependent variable is T_i or T_j, the number of trips originating or ending in a particular zone. The independent variables are characteristics of the zone or of the households within the zone. The general form of these equations is [36]:

$$T_i = a_i + \sum_k b_i X_i^k$$

It is also possible to use regression analysis at the household level [8], which avoids the problem of variation between households in a zone. The number of trips per household can be estimated using independent variables at the household level. Zonal totals can be simply derived (in linear models) by knowing the number of households in each zone.

Another method, known as category analysis or cross-classification [36], classifies households into groups and estimates trip rates per household type, and per trip purpose. The household type is a tuple based on certain socioeconomic variables, such as income level, car ownership level, and household structure. The number of trips T_i^p from zone i

for purpose p can be represented by:

$$T_i^p = \sum_h n_i(h) t^p(h)$$

where h is the household type, $n_i(h)$ is the number of households of type h in a zone i, and $t^p(h)$ is the trip rate for purpose p.

Trip distribution The most commonly used trip distribution model is the gravity model, which is similar in form to Newton's law of gravitation. Its general form is [8]:

$$T_{ij} = \alpha O_i D_j f(c_{ij})$$

where O_i and D_j are the number of trips originating from zone i, and terminating in zone j, respectively, α is a proportionality constant, and $f(c_{ij})$ is a decreasing function (typically negative exponential or inverse power) of the generalized cost of travel between zones i and j.

Modal split Modal split may be performed before or after the trip distribution step. The basic modal split models involve a diversion curve. A binary modal split (for example, between private car and transit) is an S-shaped binary logit curve of the form [36]:

$$P_c = \frac{1}{1 + e^{c_c - c_t}}$$

where P_c is the probability of choosing the private car, and c_c and c_t are the generalized costs of the car and transit, respectively.

Trip assignment Trip assignment assigns the trips T_{ijm} to one or more routes, resulting in the number of trips T_{ijmr} from zone i to zone j using mode m and route r.

The all-or-nothing approach assigns all trips to the path of minimum generalized cost (usually minimum travel time), but without considering congestion. The capacity of the actual path is considered in the capacity-constrained trip assignment algorithms, which iteratively calculates the cost and minimum paths until equilibrium is reached [14].

Direct demand models

Treating each of the four elements of travel demand independently has no compelling theoretical support, and direct demand models attempt to estimate trip generation, trip distribution and modal split simultaneously [8], possibly together with trip assignment. The most general form of these equations is [36]:

$$T_{ijmr} = \alpha_{ijmr} \prod_{k} \beta^{k}_{ijmr} P^{k}_{ijmr}$$

where T_{ijmr} is the number of trips from i to j using mode m over route r, α and β are calibration parameters, and P^k is the kth dependent variable. Although calibration of direct demand models is simpler and more intuitive than calibrating the individual submodels of the classical model, there are both theoretical and practical disadvantages, such as the large number of parameters that are needed.

4.1.3 Disaggregate models

The classical travel demand model has a number of shortcomings that have been recognized for some time. One of the main criticisms is that it does not represent a coherent theory of travel behaviour [5]. Aggregate models describe or replicate aggregate travel demand, without explaining it. Other criticisms are that policy is generally not taken into account, meaning that predictive abilities are limited [9]. Since aggregate models are basically correlative, large quantities of input data are required for accuracy, so such studies tend to be expensive. Disaggregate models [5, 8, 34, 36] have attempted to improve travel demand analysis by modelling the behaviour of individuals, rather than socioeconomic groups.

Discrete choice models

One of the more common approaches to disaggregate travel demand is the discrete choice model, which is based on random utility theory borrowed from microeconomics. The underlying principle of this theory is that individuals will attempt to maximize their utility, or minimize their disutility, when making travel-related choices. The utility functions U_{ad} for different alternatives a in a given choice set C_d for each decision-maker d, are of the form [8, 28]:

$$U_{ad} = V_{ad} + \epsilon_{ad}$$

where

$$V_{ad} = \sum_k \beta_{ka} X_{kad}$$

X is a vector of variables which determine a measurable and systematic value of the utility of a given alternative, and include both socioeconomic variables for the individual and variables describing the given alternative; β are the coefficients for the different variables, which differ among alternatives but are assumed to be constant for all individuals. To incorporate individual tastes and measurement error, a random component ϵ is used. Depending on the properties and distribution of ϵ, different models can be created. The most popular discrete choice model is the multinomial logit model, which assumes that ϵ is independent and identically distributed (IID) with a Gumbel distribution. This simplifying assumption gives the probability P_{ad} for decision-maker d to select alternative

a from the choice set C_d as [8, 28]:

$$P_{ad} = \frac{\exp V_{ad}}{\sum\limits_{k \in C_d} \exp V_{kd}}$$

The β coefficients can be estimated using maximum likelihood techniques with relatively small data sets.

Discrete choice models offer a more realistic and behaviour-based approach to travel demand, but they also suffer from a number of drawbacks. The definition of a choice set poses difficulties when dealing in practice with many discrete options such as trip destinations. Given more than a few alternatives in a choice set, the mathematics quickly become overly complicated. Discrete choice models have been best used for modal split choice. Aggregation of the results, which is necessary for practical forecasting, becomes difficult in cases where the models are not linear. It has also been claimed [5] that discrete choice models do not adequately account for policy considerations.

Activity-based approaches

Travel demand is generally considered a *derived demand*, since there is ultimately some other purpose for which travel is undertaken. Travel is therefore a means to an end, rather than an end in itself, except for relatively rare cases such as pleasure drives. A description of the daily (or longer term) activity patterns of individuals and households will therefore provide a deeper understanding of travel behaviour, and thus better descriptions and predictions of travel, given certain spatial and temporal constraints.

There has been a fair amount of work in recent years into activity-based approaches to travel demand, although it tends to be fragmented and without a dominant methodology. The different approaches do share some common features, including a focus on *patterns of behaviour*, rather than on statistics of individual trips; consideration of the timing and duration of activities and travel, instead of just peak traffic; and consideration of interpersonal constraints between household members [22].

Time budget studies have been carried out for a number of years. These studies describe and attempt to explain observations of time spent on various activities, or of participant counts in different activities during the day [21]. Somewhat related has been research on activity pattern studies, which incorporate both temporal and spatial aspects of individual and household activities, and attempt to explain and model these activity patterns.

Chapin [6, 11] concentrates on the time allocation patterns of individual and household activities. Activities, or episodes, are defined as intervals of time devoted to a dominant purpose. Associated with each activity are a duration, a location or path, and a date and time, as well as individual socioeconomic characteristics and other factors. These activities are linked with preceding and succeeding episodes, which together form an activity pattern, over a period of time. Two classes of activity are defined, although the categorization of activities is not absolute for all individuals. *Obligatory* activities are caused by basic needs (such as eating and sleeping) or by employment, and they tend to structure the day. *Discretionary* activities are not central elements of the daily routine, and may be substituted for other activities.

Hägerstrand [47], at the University of Lund, has used the approach of time geography, where time and space are intricately connected. This approach is 'physicalist' in the sense that it views time and space as resources which constrain human activities. Emphasis is placed on the following groups of physical constraints:

- Capability constraints, which arise from the physical and biological limitations of people (and machines)
- Coupling constraints, which are caused by the need to have people and equipment

bound together at given points of space-time (e.g. for work)

- Authority constraints, which are the result of restrictions to access

Between the above two approaches of focusing on choices or on constraints lie a number of conceptual and functional models [2, 22]. There remain several challenges in the activity approach, such as the handling of complexity involved in travel behaviour, the gathering of relevant and detailed survey data, and the aggregation of results to allow forecasts. Nevertheless, the approach has much potential for providing a deeper understanding of travel behaviour, and hence for better forecasting techniques.

4.2 Proposed activity-based mobility model

The proposed mobility model is presented in this section, starting with the general requirements of the model. The survey which provided the raw data used in the simulation is also described, as every transportation model is to a certain extent limited by the available data.

4.2.1 Model requirements and data availability

The goal of the mobility model is to provide, at the individual subscriber level, a realistic set of paths traversed on a daily basis. Since the model will be used to evaluate a location management algorithm which relies on a long-term user profile, a period of several days will be simulated to allow for the creation of a user profile. The model will be applied to a cellular communications system, so the routes traversed can be superimposed on a geographical framework of radio cell boundaries. The final output from the model will therefore be a list of cells traversed by an individual subscriber over a period of several days, together with the amount of time spent in each cell.

A variety of mobility information is collected through surveys periodically conducted by metropolitan area traffic planning offices. These surveys have many uses, such as traffic planning and public transportation policy development. This data normally includes trip information for a sample of households for a single weekday, although the exact data collected varies from region to region and from survey to survey.

The input data for this mobility model was derived from the trip survey conducted by the Regional Municipality of Waterloo in 1987. The trip survey captured information on trips taken by residents during a specific weekday. A travel diary was completed by each household member over 5 years of age, in which details on all trips taken during the survey day were recorded. Another section of the survey contained information about the household and individuals, some of which was inaccessible for privacy reasons. Data from the survey was aggregated into one file containing items relevant for this simulation. The following data was included for each recorded trip:

- Person identifier
- Trip start time
- Trip end time
- Trip purpose (or activity) at the origin and destination
- Home traffic zone of trip-maker
- Employment status (full-time or part-time, in or out of the home) of trip-maker
- Fixed work location (true or false) for trip-maker

The trip survey could not be used in its original format primarily because it only captured trip information for one weekday. Several mobility management schemes, including the one proposed here, involve some type of user movement profile built up over time. The model uses aggregated input data extracted from the trip survey to simulate subscriber mobility over a period of several weekdays. The input data to the simulation and the simulation procedure are described below.

4.2.2 Overview of the mobility model

The eventual goal of the mobility model is to simulate the daily movement of cellular subscribers over a period of several days, in terms of cells crossed and time spent in each cell. Using this information, location management algorithms can be compared according to the relative number of location updates and paging messages they generate. After reviewing the transportation theory literature (Section 4.1), it was decided that the best approach for the required mobility model, in terms of solid and intuitive theoretical basis, simplicity, and implementability given the available data, was a modified activity-based approach.

The driving parameter in the simulation is the notion of trip purpose, which is equivalent to an activity in this proposal. This is an important parameter collected in the original trip survey, which classifies the reason for the trip into nine categories, namely (1) work, (2) work related, (3) school, (4) serve passenger, (5) shopping, (6) social/recreation, (7) personal business, (8) return home, and (9) other.

In the simulation, the 24-hour day is divided into a user-defined number of equal segments, referred to as *time periods*. Time periods are used to aggregate data from the trip survey by time of day. The simulated day for a particular subscriber consists of a sequence of activities which have an associated duration and location. The paths taken to travel from one activity location to another, together with the time spent at each activity, form the spatial and temporal output data required from the simulation.

The concept of activity, or trip purpose, is central to the simulation. Each activity has an associated time of day, duration, and location (at the level of a cell). An activity is selected based on the previous activity and the current time period. The probability of transition from one activity to another uses the activity transition matrix described below. Once the next activity is selected, its duration is determined using the activity duration matrix. Finally, the location of the activity is selected, based on the type of activity, and some heuristics. Since the current location of the subscriber is already known, once the location for the next activity is selected, the intermediate route (in terms of cells crossed) and the total distance are determined from a lookup table. Using a user-defined system-wide average speed, the total time and the time in each intermediate cell are calculated. The subscriber stays in the destination cell for the duration of the activity, and the sequence is repeated.

Each simulated subscriber has certain characteristics, one of the more important of which is person type. The distributions for the activity transition matrix and for the activity duration are indexed by person type. Four categories of person type were defined, similarly to [26]. This categorization attempted to create groups of somewhat similar subscribers, with similar mobility patterns, using the information that was available from the trip survey. The four categories were:

- Full-time employed outside the home
- Part-time employed outside the home, but not a student
- Student, secondary or post-secondary, possibly employed part-time outside the home
- Not employed outside the home, and not a student

Another important characteristic is whether the work location is fixed from day to day, since work is a major activity. Other characteristics of the subscriber include the home, work, and school locations at the cell level. This information is used to select the destination cell when the corresponding activity (e.g., home, work, or school) is selected.

4.2.3 Input data

In determining the next activity, its duration, location, and the geographical path to take, entries are looked up in several tables, sometimes using a random index. The derivation and use of the tables are specified below.

Activity transition matrix

After validation, the trip survey data was manipulated by an external program to obtain data that was directly applicable to this simulation. The activity transition matrix was one of the data tables extracted from the trip survey data.

The activity transition matrix is an empirical distribution of transitions from one activity (or trip purpose in the trip survey) to another, recorded as a four-dimensional array. The distribution is indexed by the subscriber's person type, the current time period, and the previous activity. Associated with each entry is an empirical cumulative distribution of the transition probability to the next activity, as shown in Table 4.1. Implicit in this model is the assumption that activity linkages depend on individual characteristics and on the time of day (i.e., time period). This assumption has been supported by other studies [13].

Activity duration matrix

The activity duration matrix (Table 4.2), which was also created by an external program, is a table indexed by person type, time period, and current activity, giving a cumulative probability distribution of the different activity durations that were observed. The duration of a particular activity can be derived from chronologically sorted trip survey

person type	*time period*	*previous activity*	*next activity*	*cumulative probability*
1	4	8	1	0.351724
1	4	8	2	0.393103
1	4	8	3	0.420690
1	4	8	4	0.475862
1	4	8	5	0.696552
1	4	8	6	0.793103
1	4	8	7	0.986207
1	4	8	8	1.000000
1	4	8	9	1.000000
1	4	9	1	0.000000

Table 4.1: Sample data from the activity transition matrix

records by calculating the difference between the arrival time of one trip record, and the departure time of the following trip record. Of course, the previous record's destination (arrival) purpose and the subsequent record's origin (departure) purpose must be equal, otherwise the pair of records is rejected. The relative frequency of the durations (aggregated into 5 minute intervals) was recorded, together with the corresponding person type, time period, and trip purpose, to create the empirical cumulative distribution.

person type	*time period*	*activity*	*duration*	*cumulative probability*
0	7	6	330	0.948718
0	7	6	360	0.961538
0	7	6	400	0.974359
0	7	6	460	0.987179
0	7	6	540	1.000000
0	7	7	0	0.125654
0	7	7	5	0.235602
0	7	7	10	0.413613
0	7	7	15	0.539267
0	7	7	20	0.596859

Table 4.2: Sample data from the activity duration matrix

In addition, when the origin purpose of the earliest recorded trip matched the destination purpose of the last recorded trip for a particular subscriber, the duration of that purpose (i.e., overnight duration) was assumed to be valid and was recorded. In almost every case, the first recorded trip started at the home, and the last recorded trip ended at the home (i.e., the purpose or activity was *home*).

Geographical path

In a realistic mobility model, not all cells are visited for the same duration. Some cells are the location of extended activities, while other cells are crossed briefly as the subscriber travels to the final destination. This concept of intermediate cells is critical to the mobility model, in order to reflect the distribution of time spent in various cells. Given an origin and destination cell, a table was created to look up the path taken, at the cell level. Another table contained the total distance from origin to destination.

Routing information was derived using a geographical information system (GIS) package called TransCAD. The street grid of the Kitchener-Waterloo-Cambridge area was entered into TransCAD, onto which was superimposed a layer of *cells*. Each cell is a geographical area roughly corresponding to a cellular radio coverage cell. Cells are groups of adjacent *traffic zones* as defined by the Region of Waterloo Planning Office. Traffic zones are important because they represent the basic geographical unit for which survey and census data is collected, such as retail employment or land use, which affect trip attractions. The assignment of traffic zones to cells was somewhat arbitrary, but followed some basic rules, such as making cells as convex as possible, and making cell size roughly inversely proportional to road and population density. Each cell was represented by a *centroid*, a point roughly in the center of a cell, and connected to the street grid. The path taken from one cell to another was in practice from one centroid to another.

A total of 45 cells was defined for the Kitchener-Waterloo-Cambridge area, out of the

235 corresponding traffic zones. Using the minimum path function of TransCAD, the minimum distance between each pair of centroids was calculated, and the cells crossed by this minimum path were recorded manually. In the cases where cell borders coincided with the minimum path, the rule followed was to minimize the number of cell changes by remaining in a cell as long as possible. In a few cases where a corner of another cell was 'clipped' before returning to the same cell, no cell change was recorded, since radio cell boundaries are fuzzy in practice, and a good cell selection algorithm (see Section 2.2.3) should avoid such changes. Using the calculated minimum path length, the travel time was calculated by dividing the path length by a user-defined average speed, assumed to be constant for the entire region. This average speed should be relatively low to account for time spent waiting at stop signs, traffic lights, and congested roads. Although this assumes motorized transport, this assumption is justified since over 85% of all trips in the Region of Waterloo trip survey had some form of motor vehicle as the mode of transport [1].

Subscriber information

The mobility model describes individual subscriber movement, and so certain information must be maintained for each subscriber. The output derived from the model should reflect the individuality of a subscriber, otherwise it would be indistinguishable from a random mobility model. In transportation theory, individuals (or groups) can be categorized by many different parameters, such as income, car ownership, etc. To avoid unnecessary complexity, yet still differentiate to some extent between different groups, a person type variable was defined for each subscriber. This attribute, as defined earlier, had one of four values, based primarily on employment status.

Certain locations (cells) which were associated with an individual were also included, such as home, work, and school locations. Work location may or may not be fixed over time, and a variable maintains this information. The home and school locations are assumed fixed, however. This information is used to determine the destination cell whenever the corresponding activity is selected. Thus, the proposed mobility model incorporates the structure of daily travel patterns, which is completely ignored by random mobility models. The current activity, cell, and location area are also maintained for each subscriber in the simulation.

4.2.4 Simulation procedure

Using the information in the matrices defined above, a discrete-event simulation was developed. Briefly, a discrete event simulation schedules events in a queue to occur at certain times based on a global clock, and these events are dequeued in chronological order and acted upon based on the event type.

The three event types in the mobility simulation are the *trip-generation* event, the *cell-change* event, and the *call-arrival* event. The occurrence of a particular type of event triggers a corresponding function. The simulation is started with an initial trip-generation event, whose parameters are entered by the user. The trip-generation function creates and queues a set of cell-change events. These cell-change events correspond to the cells crossed as the subscriber selects a new activity and travels to its new location. A cell-change function performs any location updating functions which may be triggered when a change of cell occurs. The trip-generation function also creates and queues a second trip-generation event, corresponding to the end of the duration of the destination activity.

Call arrivals are not modelled using the typical Poisson arrival process, since that model is not very realistic for individual subscribers receiving calls. A subscriber is much more likely to receive a call at 2 pm than at 2 am, a fact not accounted for by a Poisson model. A key parameter used in the simulation is the number of call arrivals, c_a, over 24 hours, which is defined at run-time. A call-arrival function generates and queues c_a call arrivals for the next 24 hours based on a heuristic probability distribution for call arrival time, described below. The call-arrival function also queues an event which will trigger the generation of the next set of call arrivals.

Four independent variables are used as inputs to the simulation. Person type and work location variability parameterize the mobility aspects, while the number of daily call arrivals and the maximum size of the dynamic location area (described in Chapter 5) affect the location management algorithms. For each combination of the independent variables, fifteen repetitions of the simulation are executed, each with a randomly selected set of internal parameters (such as home location).

The simulation continues running until the master clock crosses a user-defined threshold, nominally set at 15 simulated days. Output detailing location update and paging messages generated by the different location management algorithms is logged to external files, which are analyzed by separate software.

Trip-generation event

Trip-generation events cause all of the subscriber movement during the simulation. Inputs to the function are the previous activity, subscriber type, and the simulation clock time (which is converted to the corresponding time period). The output is a set of cell-change events. After the cell-change events are queued, another trip-generation event is generated and queued, representing the next activity and destination, as described below. The cell-change events correspond to the cells crossed as the subscriber moves from the current location to the location of the next activity.

Determination of the subscriber's next activity drives the daily mobility pattern generation, as shown in Figure 4.1. The next activity is selected using the activity transition matrix. Given the current activity, the current time period, and the person type, the next activity is randomly selected from the corresponding entries in the activity transition matrix. Together with the subscriber type and the current time period, the selected next activity is used to determine the duration of the activity, using the duration matrix. The next destination cell also depends on the selected activity. If the activity is work (for fixed work locations), school, or home, constant locations are used from the stored subscriber information. The location returned for the other activities (except for shopping) is a uniformly distributed random number representing cells from 1 to 45.

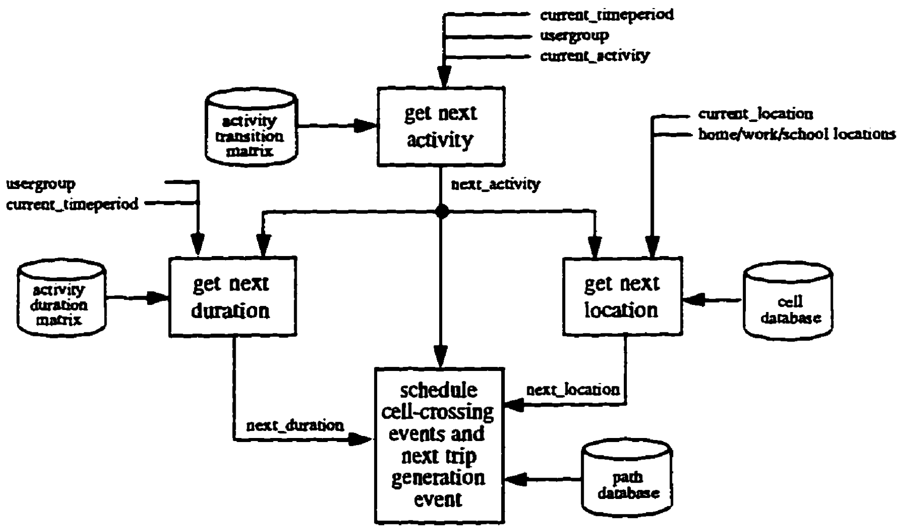

Figure 4.1: Data flow schematic for trip generation in the proposed mobility model

For shopping, the possible destinations are limited due to shopping location constraints, as well as habit and convenience among individuals. Therefore, shopping may be more realistically modelled in terms of zonal retail employment [23], or distance to shopping location [14]. The method used in this simulation is a combination of these two approaches. The zonal retail employment (aggregated to form cell-level retail employment) is divided by the distance from the current location to the target cell. The destination cell is randomly selected from the top five target cells. Intuitively, this method prefers cells which are closer to the current location of the subscriber, and is more likely to select shopping districts.

Using the current cell and destination cell, the path and total distance are looked up in the path and distance tables. The total distance is divided by a user-defined system-wide average speed to calculate the total time, which is divided by the number of crossed cells to derive the cell crossing time. An assumption was made that cell crossing times are equal for all the intermediate cells; endpoint cells have half of the cell crossing time each. For each intermediate cell, a cell-change event is created and queued, with an event time equal to the current time plus an appropriate multiple of the cell crossing time. Finally, a second trip-generation event is created and queued, with an event time equal to the current time plus the total travel time to the destination plus the activity duration time (i.e., once the selected activity is completed, a new activity should be derived).

Cell-change and call-arrival events

When a cell-change event occurs, the location management function is called. The exact location update actions depend on the particular location management algorithm implemented. Details of location updates, such as the time and cell in which they occurred and the cells which comprise the new location area, are logged to an external file for later processing.

Independent of any movement, calls may arrive for the subscriber, represented by call-arrival events. The simulation will randomly create c_a call arrival times using a probability distribution. The day is divided up into three periods, each with a corresponding percentage of daily call arrivals. This approximately represents the observed busy peri-

Time of day	*Probability of incoming calls*
00:00 to 08:00	0.2
08:00 to 18:00	0.5
18:00 to 24:00	0.3

Table 4.3: Probability distribution of incoming calls during the day

ods during the morning and mid-afternoon, and slow periods during late night and early morning. The heuristic distribution used here is shown in Table 4.3.

Similarly to the cell-change event, when a call-arrival event occurs a location management function is called. This function handles the paging actions depending on the particular location management algorithm implemented. Details on the number of paging messages required are logged to an external file for later processing.

Chapter 5

Dynamic location management algorithm proposal

Location management algorithms, as implemented in current cellular systems and as presented in most proposals, assume as fundamental the existence of spatially static location areas. The mapping of cells to location areas is performed infrequently, using heuristics and aggregate statistics, and the resulting location areas apply to all subscribers. Clearly, not all subscribers have the same mobility patterns, spatially or temporally, and thus static and globally applicable location areas are far from optimal.

The proposed location management algorithm attempts to utilize the mobility history of the subscriber to dynamically create location areas for individual subscribers, and to dynamically determine the most probable paging area. The motivation behind the proposal, the information available as input to the algorithm, the goals and constraints, and a detailed outline of the algorithm are presented below.

5.1 Motivation

As was seen in Chapter 3, many proposals, given fixed location areas, attempt to improve on the current system by using hierarchical location registers or databases, usually integrated in an Intelligent Network type of architecture. These proposals attempt to find an efficient way of searching through the hierarchy to locate the subscriber, using pointers created by location updates. Although this type of proposal removes the potential bottleneck created with a centralized location register architecture, it does not address the problem of efficiently performing the location updates which are used to build the location register hierarchy. No attempts are made to reduce the frequency of location updates, which directly affects the efficiency of many of these proposals.

Several proposals suggest the use of user profiles to individualize the location areas. However, the detailed contents of the user profiles, as well as procedures for their creation and maintenance, were not well defined. In addition, proposed algorithms to dynamically determine individual location areas generally suffer from excessive complexity, which is not supportable on mobile stations. Other location management proposals are based on grouping subscribers with similar behaviour. Although plausible, such grouping is fairly arbitrary since subscriber mobility characteristics may change. The key issue is that algorithms must remain flexible, not overly complicated, and efficient, even in the face of mobility behaviour changes.

In developing the following proposal, ideas were borrowed from transportation theory and activity pattern theories discussed in Chapter 4. In the vast majority of cases, a trip is undertaken for a particular purpose occurring at a particular destination. The purpose behind a trip is generally some activity, such as work or shopping. These activities can take place at a limited number of locations. For example, most people work at a fixed location, and schools and shopping areas are located at specific sites. The home is normally the hub from which all other trips are taken, and is generally a place where subscribers spend a substantial amount of time during the day.

Location management techniques can make use of the fact that the average subscriber has a limited number of frequently visited locations, which form part of the daily mobility pattern. Of course, there will be deviations from the standard mobility pattern, but the more fundamental activity pattern seldom changes. The proposed algorithm exploits the spatial and temporal characteristics of trips. For example, during a typical trip, a number of cells may be crossed between the origin and destination of a trip; the subscriber will spend some amount of time at the destination (depending on the activity), which will likely be longer than the time spent in the intermediate cells. The intermediate cells should be included in the location area, but only the destination cell (or some other cell in which a relatively long period of time was spent) should be paged on the first try. The subscriber is less likely to be found in the intermediate cells.

Similar logic is used in determining location areas. In general, larger location areas are preferable, since location updates will be reduced. The drawback is a correspondingly large number of paging messages. This can be mitigated by using intelligent paging, as outlined above, to page the most likely cells first. Even in this case, however, a location area must still be of limited size, to bound the number of pages both in first-round paging, and in worst case situations. If an entire city, for example, was one location area, a subscriber could have dozens of cells in which to be paged in the first round.

If fixed location areas are used, the daily movements of particular subscribers might still cross several location areas on a single trip undertaken daily, generating unnecessary location updates. A personal, or individualized, location area, anchored around the cell in which the location update occurs, would consist of the most likely cells that the subscriber could traverse. The information to create the personal location area is derived from the user profile, which contains a record of previous transitions from cell to cell, together with the length of time spent in each cell.

5.2 Mobility history and user profiles

The user profile maintains a record of the mobility history of the subscriber. The mobility history consists of the number of transitions the subscriber has made from cell to cell, and the average duration of visits to each cell. Specifically, a counter $N_{a,b}$ is kept for each cell a of the number of times the subscriber has moved from cell a to each neighbouring cell b. Also associated with each visited cell a is the average time T_a that the subscriber has spent in that cell.

As a mobile station roams, it selects and camps on the best cell, in terms of signal reception. If the mobile station moves from cell a to cell b, the counter $N_{a,b}$ is incremented by 1, and a timer t_b is started, which is stopped either when the mobile station moves to a new cell, or the mobile station is turned off. The timer may be incremented in minutes or deci-hours (6 minute intervals), or some other appropriate increment. When t_b is stopped, the average value of T_b is updated, since the total number of visits to cell b can be derived from the table storing the counters N. The counters and timers are stored in non-volatile memory on the mobile station (such as the SIM card in a GSM system).

The information stored in $N_{a,b}$ and T_a for all visited cells a and b can be represented as a directed graph, where the nodes represent visited cells, and the links represent transitions between cells. The weight of a link (a, b) is the value of $N_{a,b}$, and the weight of a node a is the value of T_a.

When a subscriber enters a new cell which is not in the current location area, a location update is performed in which the mobile station dynamically determines the new location area. The new location area is created based on the most likely cells to be visited by the subscriber from the new cell, using information from the user profile.

5.3 Goals and constraints of the algorithm

In Chapter 3, different proposals were compared against the general goals of location management. These same goals form the basic requirements for the development of the proposed algorithm. Certain trade-offs are necessary since many of the goals are contradictory in practice. Priority was given to minimizing the bandwidth required for location updates and paging. Performance should be better than or equivalent to current proposals for all mobility patterns. Providing adequate quality of service in terms of call setup delay (due to paging delay), as well as preserving battery life, were also given high priority. The cost was slightly more computational complexity in the mobile station than currently required, and additional memory requirements at the mobile station and the network databases.

5.4 Algorithm definition

Given the above requirements and constraints, a dynamic location management algorithm was designed using as input the mobility history (or user profile) of the subscriber. The user profile can be represented as the directed graph previously described. A set of cells are selected for the new location area, such that the probability of the subscriber subsequently roaming within those cells is sufficiently high, as defined below. For an incoming call, the subset of the location area where the subscriber is most likely to be found is paged first, based on the subscriber's average visit duration in each cell.

5.4.1 Location updating strategy

When a subscriber enters a new cell which is not part of the previous location area, when the mobile station is first powered on in a cell, or periodically as required by the network, a location update is triggered. Upon completion of the location update, a new location area will have been generated for the subscriber, comprised of cells which the subscriber is most likely to go to from the current cell.

The user profile, conceptually represented as a mobility graph, is used to derive the most likely cells. The overall decision of which cells to include in the new location area depends on several considerations:

- Selecting a subset (possibly null) of the outgoing links at each node
- Selecting a metric used to rank the possible links
- Selecting the order of inclusion of new nodes into the new location area (i.e., breadth-first or depth-first order)
- Deciding when to stop selection of nodes along a particular path

The goals and constraints previously mentioned were used in deciding among the alternatives. The decisions taken are explained in the overall sequence of steps for the location updating procedure, outlined below.

1. Whenever a new cell is entered (for example, from cell a to cell b), the counter $N_{a,b}$ is incremented, and the average duration T_a is updated. This occurs regardless of whether a location update is performed, or whether a location update was dynamic or static.

2. An attach procedure, similar to GSM, is possible when the current cell at power-on is within the last location area recorded before the previous detach procedure at power-off. In this case, a simple message would be sent to the HLR, notifying it that the mobile station is again reachable.

3. Upon entering a new cell and having decided a location update is necessary, the mobile station looks up the new cell in the user profile. If the new cell is not found

in the user profile (i.e., if the user has never visited that cell or if the cell has been purged from memory), a classical location update is performed. For this purpose, the network is overlayed with relatively large static location areas, similar to current implementations.

4. If the new cell is found in the user profile, the list of neighbouring cells previously visited is read, together with the number of times the subscriber has moved to those cells from the new cell ($N_{a,b}$ for all cells b, or the link weight in the mobility graph). Due to the physical layout of cells in cellular networks, there can be only a limited number of neighbouring cells, normally six. The average weight $\bar{N}$ of the links to each neighbouring cell is calculated.

5. The cells corresponding to the links whose weights are greater than or equal to the average weight $\bar{N}$ are added to the new location area, in order according to the link weight. The average weight is used to discriminate between paths which the subscriber often traverses, and paths which are seldom traversed. The mean is a relatively simple way of determining the break between frequently and infrequently used cells, in a non-arbitrary manner. Adding all qualified cells from the neighbouring cells implies a breadth-first method, which is used to limit the possibility of additional location updates if the subscriber subsequently traversed a directly neighbouring, but unselected, cell. A depth-first method would normally select a single, extended, and frequently used path, but would ignore paths through less frequently traversed neighbouring cells.

6. Once selected cells from the first *ring* of neighbouring cells have been added to the personal location area, the above two steps are repeated for the cells in the personal location area, in order by largest link weight. The average weight of the cell's outgoing links is calculated, and any nodes whose associated link weight is

greater than or equal to the average are added to the personal location area, such that all cells descended from the first ring of cells are ordered by link weight.

7. It is possible that one of the selected links leads back to a previous node. Obviously, a particular cell needs to be listed in the new location area only once, so duplicated cells are not added to the personal location area.

8. The above steps are repeated until no other cells are left for inclusion, or the maximum dynamic location area size has been reached. This limit is defined externally; maximum dynamic location area size is one of the independent variables used in the simulation analysis. A larger limit implies larger location areas, as well as additional memory requirements at the mobile station and the network. The effect of larger dynamic location areas is discussed in Chapter 6.

5.4.2 Paging strategy

The paging strategy is closely tied with the location updating strategy. A large location area is beneficial in terms of efficient location updating, but not for paging unless some intelligent paging strategy is used to reduce the number of cells that are paged.

In this proposal, one of the parameters maintained by the mobile station is the average time spent in each visited cell a, represented by T_a. Cells with a low value of T_a are ones which the subscriber tends to cross quickly during a trip to some other destination. An example would be an intermediate cell along the path from home to work. It may be frequently crossed, but the subscriber spends little time within the cell, and will seldom be found there. On the other hand, cells with a high value of T_a indicate that the subscriber is quite likely to be found there. Examples of such cells are work or home cells.

As the mobile station roams, it maintains a table of the values of T_a for each cell. During a location update, the set of T_a values for the cells comprising the new location area are transmitted to the network. Up-to-date T_a values enable the network to perform more efficient paging. However, this data should be sent as infrequently as possible to minimize bandwidth usage, and should use an existing radio connection where possible (e.g., using the signalling channels during a conversation).

The paging strategy involves the use of the parameter T_a to limit the paging cost. Cells within the current location area with a high value of T_a, and thus more likely to contain the subscriber, form the paging area which is paged first. Cells with lower values of T_a are subsequently paged if there is no reply from the subscriber.

The problem of determining which specific cells to page first, given a corresponding list of T_a values, is similar to the problem of determining which links to include in the location updating strategy. The average value of T_a among the cells in the current location area is calculated. Cells where T_a is greater than or equal to the average form the paging area used in the first round of paging. If this paging attempt is unsuccessful, all cells in the location area are paged in the second round of paging (in case the subscriber moved to the first paging area just after the first round of paging). Note that a location area may be broken up into more than two paging areas, for example by finding the average of the first group of cells and using that to divide the first group into two segments. Different implementations of the general paging scheme are possible depending on the number of cells per location area, and the number of cells comprising a paging area. The number of paging areas, however, should be kept low to avoid long paging-related call setup delays which adversely affect quality of service.

5.5 Summary

A dynamic location management scheme was proposed, based on the mobility patterns of individual subscribers. The scheme attempts to improve upon the performance of current location management schemes in terms of bandwidth efficiency. The mobile station maintains a mobility history consisting of the number of transitions between visited cells, and the average time spent in visited cells. The mobility history is used for the dynamic determination of efficient and individualized location areas and paging areas.

Chapter 6

Analysis

The proposed location management algorithm is compared to existing protocols in this chapter, primarily in terms of radio bandwidth efficiency. As discussed in previous chapters, one of the most important goals of location management algorithms is efficient utilization of radio spectrum, which is a fixed resource. This requirement can be satisfied by minimizing the number of location updating and paging messages that must be transmitted across the radio interface.

6.1 Description of data and procedures

The analysis involved comparisons of the dynamic model with the standard fixed location management algorithm, as typically implemented in today's cellular networks. The 45 radio cells in the simulation environment were grouped in four different fixed location area layouts, in an ad hoc manner based on geographical proximity. These location areas are equivalent to paging areas, and the terms are used interchangably in the discussion below. Fixed Group 0 has one cell per location area. Fixed Group 1 has thirteen location areas of three or four cells each. Fixed Group 2 is similar, with three to five cells, generally different from Group 1, in each of eleven location areas. Finally, Fixed Group 3 has five large location areas, with eight to ten cells per location area. These fixed groupings, or algorithms, were compared with the proposed dynamic algorithm, whose maximum location area size is parameterized.

The original output from the mobility model provided details about paging and location update messages generated over time by a simulated subscriber, for each of the fixed algorithms and for the proposed dynamic algorithm. The movements and messages generated by each subscriber were logged for 15 simulated days, an interval assumed at the outset to be long enough for the model to reach a steady state. For each set of controlled variables (person type, fixed work location, size of dynamic location area, and number of daily call arrivals), fifteen repetitions were run, each with a different set of internal parameters (initial start time, location of home, work, and school). The number of location updates and pages were tabulated, per simulated day, for each different algorithm and for each different set of controlled variables. Overall, there were 160 parameter combinations (4 person types × 2 fixed work location states × 5 daily call arrival levels × 4 dynamic location area sizes) with 15 repetitions for each combination, for a total of 2400 individual runs.

Due to the number of independent and dependent variables and the complexity of the data, statistical analysis on the data was deliberately kept simple. Various means, standard deviations, and test statistics were calculated using the statistical package SAS and in certain cases, Perl scripts. The assumption of normal distribution of the data, necessary for t-tests and other techniques, was verified through SAS.

6.2 Comparison of location management-related signalling

The average number of location update and paging messages generated by the different fixed algorithms were compared with the proposed dynamic algorithm. The total location management cost, a rather abstract but very significant concept, was also calculated and used to compare the overall performance of the different algorithms. A global comparison is given first, followed by comparisons between different subscriber types, and between employeed subscribers with fixed and non-fixed work locations. Finally, the effect of different maximum sizes for the dynamic location area is analyzed.

6.2.1 Overall average performance

Location updating

The graph in Figure 6.1 shows the number of location update messages for the different algorithms, averaged over all relevant independent variables (person type, fixed work location, and maximum size of the dynamic location area). It shows the overall relationship between the different fixed algorithms and the proposed dynamic algorithm. As expected, Fixed Group 0, with one cell per location area, performed the worst. Fixed Groups 1 and 2, with about four cells per location area, performed better than Fixed Group 0, but worse than Fixed Group 3, which had about 9 cells per location area.

The 'learning curve' associated with the dynamic algorithm is clearly visible on the graph. As time progresses, the user profile is able to provide better information for cell selection. After the first day, the dynamic algorithm was comparable to Fixed Groups 1 and 2. The dynamic algorithm outperformed (statistically significant with P-value = 0.0020) Fixed Group 3 after four elapsed days, and the performance gap increased as time increased.

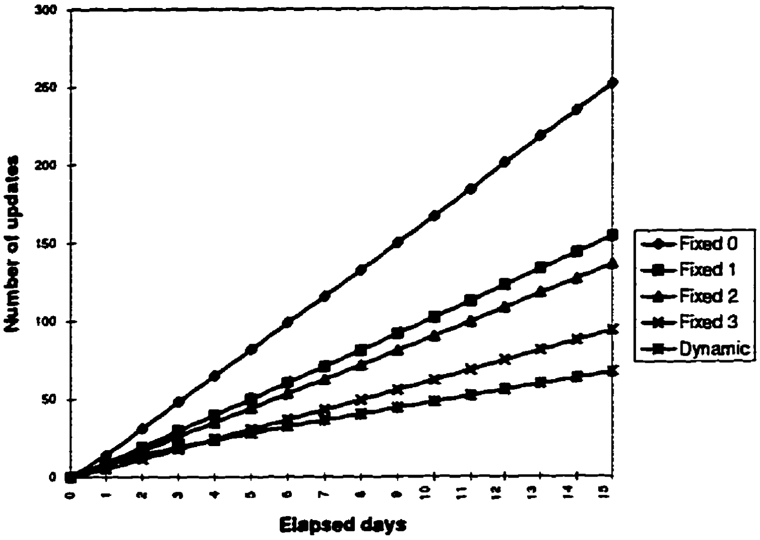

Figure 6.1: Global average number of location updates versus elapsed time

Paging

The overall average number of cells paged is shown in Figures 6.2, 6.3, 6.4, and 6.5 for levels of 3, 6, 9 and 12 incoming calls per day, respectively. The results scale almost identically for the different incoming call levels, indicating that the dynamic paging algorithm performs consistently relative to the fixed algorithms regardless of paging load.

The paging cost for Fixed Group 0 is optimal, since the current cell is always known, and consequently exactly one cell gets paged. However, this optimal paging cost is in general far outweighed by the associated large location updating cost. Conversely, the paging cost for Fixed Group 3 is comparatively the worst, due to the large paging (location) areas. As expected, the paging cost for Fixed Groups 1 and 2 falls in between the two extremes, with Fixed Group 1 being slightly better due to the smaller average number of cells per location area.

Despite the typically large location area for the dynamic algorithm (from 5 up to 20

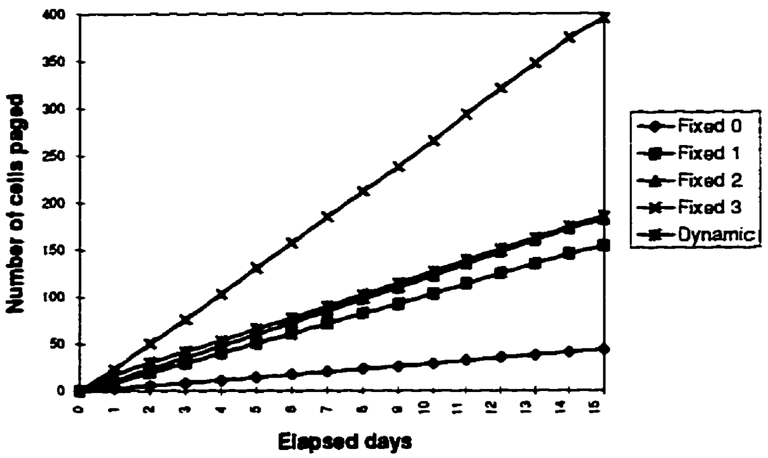

Figure 6.2: Global average number of cells paged for 3 incoming calls, versus elapsed time

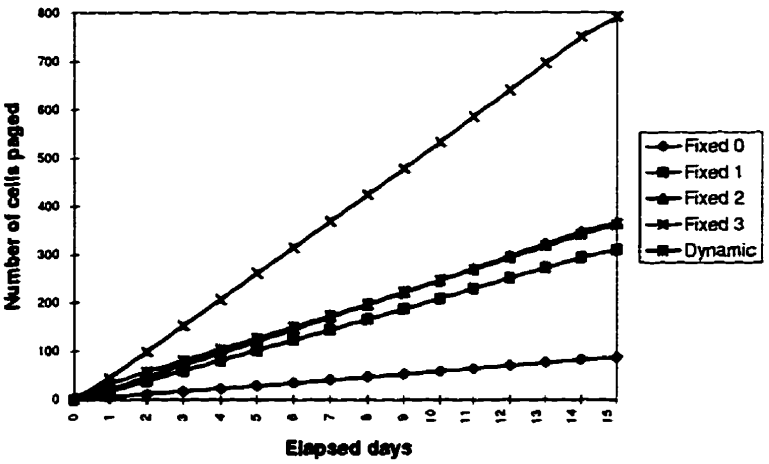

Figure 6.3: Global average number of cells paged for 6 incoming calls, versus elapsed time

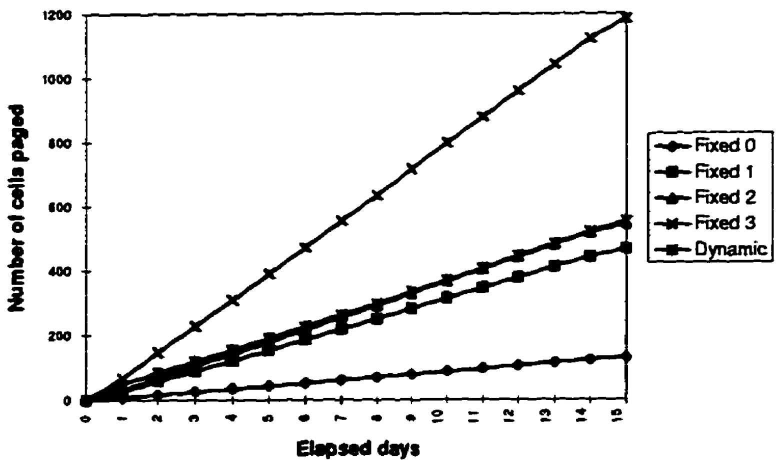

Figure 6.4: Global average number of cells paged for 9 incoming calls, versus elapsed time

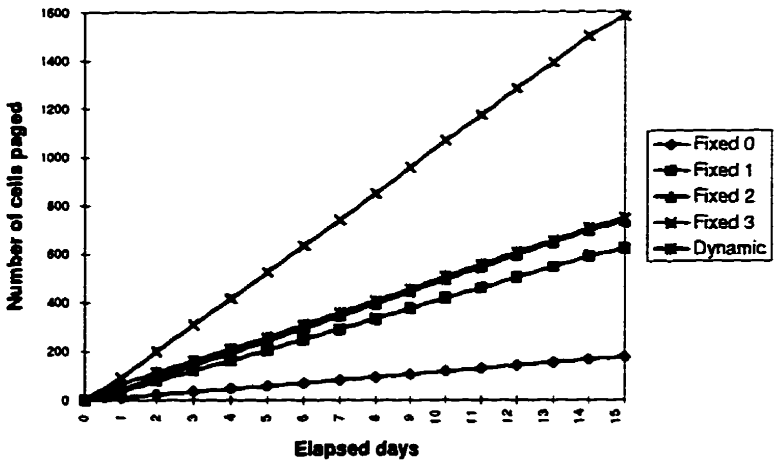

Figure 6.5: Global average number of cells paged for 12 incoming calls, versus elapsed time

cells, compared to 10 for Fixed Group 3), the paging cost for the dynamic algorithm is almost identical to Fixed Group 2, which has only three to five cells per location area. During the first few days, while the user profile is being compiled, the performance of the dynamic algorithm was somewhat worse than either Fixed Group 1 or 2. This gap, however, typically disappeared after 5 or 6 elapsed days, after which the performance of the dynamic algorithm and Fixed Group 2 were almost identical.

6.2.2 Total location management cost

The dynamic scheme performs best for location updates, but is not optimal for pages, due to the trade-off between efficient location updating and efficient paging. The total location management cost, defined as a linear combination of the number of location updating and paging messages, is a way to observe the overall effects and behaviour of the different algorithms. Total cost is a relatively abstract concept in the scope of location management. It can be defined in several ways, such as the total number of signalling messages crossing a specific interface, or the total size in bytes of all exchanged signalling messages. Some authors simply represent cost as an abstract constant. In this analysis, an intermediate approach has been taken, which is quantitative, yet not limited to a specific technology. Total location management cost is given by:

$$cost = c\,N_{lu} + N_p$$

where N_{lu} is the number of location update messages generated, N_p is the number of cells paged as a result of an incoming call, and c is a constant representing the relative cost of updates to pages. Since a location update is a much more expensive procedure in terms of signalling than pages, the two representative values of c are taken as 5 and 10. These values are somewhat arbitrary, but represent the approximate size and number of signalling messages required by a location update compared to a paging message, as discussed in Chapter 2.

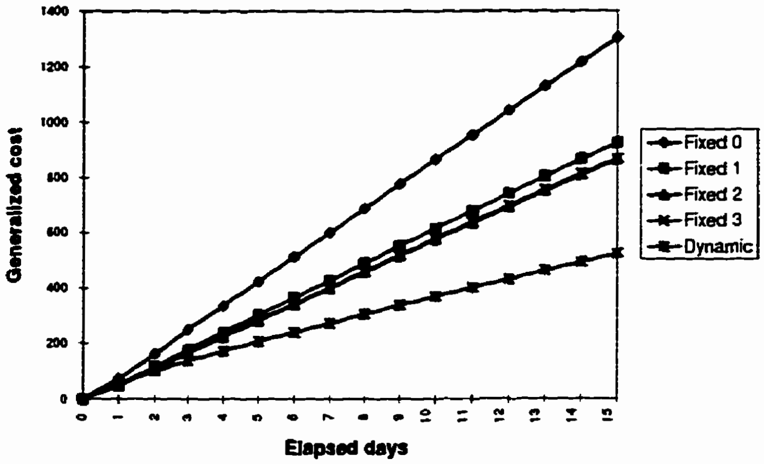

Figure 6.6: Global total cost (c = 5) for 3 incoming calls, versus elapsed time

Significant differences in the relative cost of the different algorithms arise as the number of incoming calls increases. For relative cost $c = 5$, and low levels of 3 incoming calls per day (Figure 6.6) and 6 incoming calls per day (Figure 6.7), the Fixed Group 0 algorithm incurs the highest cost. This is due to the excessive updating cost required by this algorithm, which is not offset by its inherently low cost of paging since there are few incoming calls. Fixed Groups 1 and 2 have very similar performance, with Fixed Group 1 performing slightly better, although the gap decreases as more incoming calls arrive. As the number of incoming calls increases (Figures 6.8 and 6.9), Groups 0, 1 and 2 approach a similar total cost. The performance of Fixed Group 3, which is initially quite good, declines rapidly as the number of incoming calls increases, so that for 9 and 12 incoming calls per day, its performance is significantly worse than all other algorithms. This is expected, since it performs the worst in terms of paging cost, which forms an important

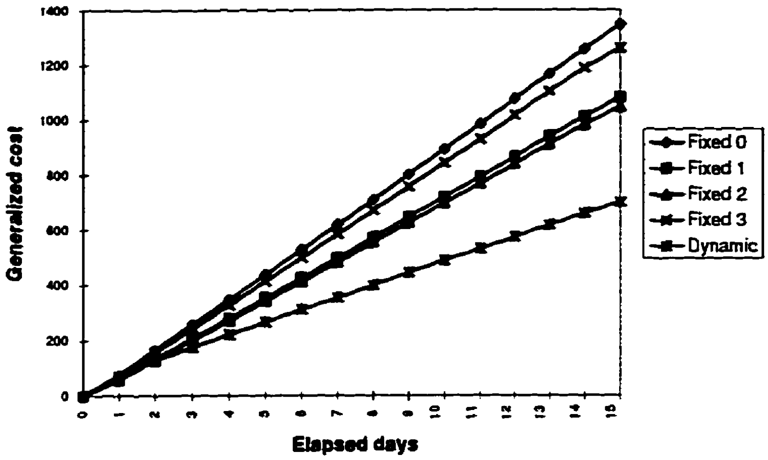

Figure 6.7: Global total cost (c = 5) for 6 incoming calls, versus elapsed time

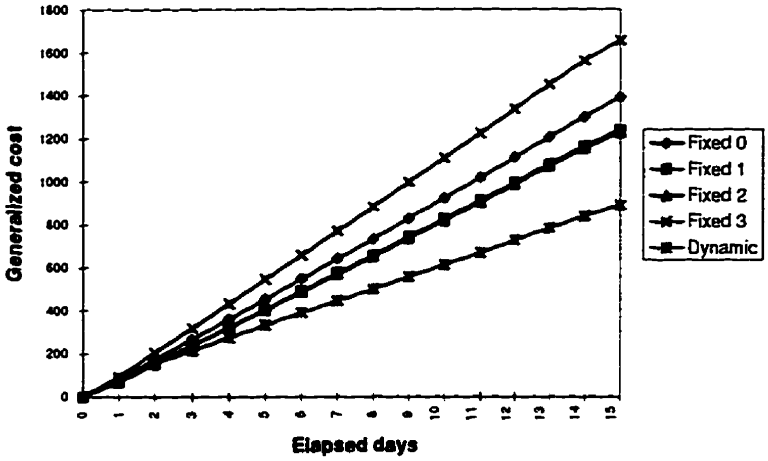

Figure 6.8: Global total cost (c = 5) for 9 incoming calls, versus elapsed time

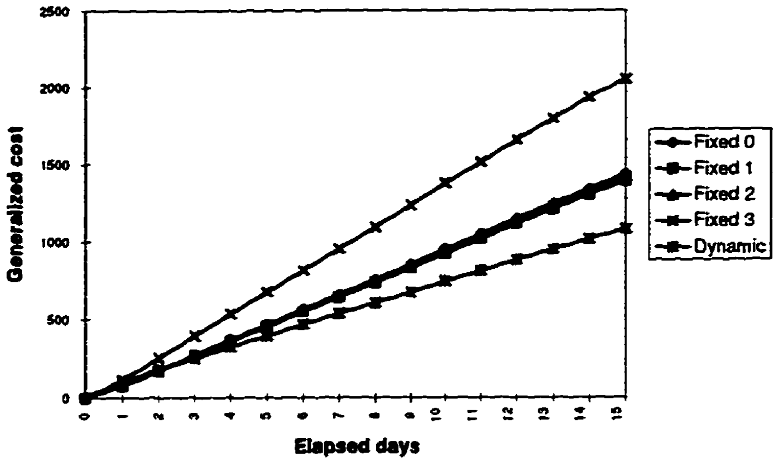

Figure 6.9: Global total cost (c = 5) for 12 incoming calls, versus elapsed time

proportion of the total cost for a value of 5 for relative cost c.

The dynamic algorithm outperforms all fixed algorithms after a brief 'learning curve' of no more than three elapsed days. Although the paging cost of the dynamic algorithm is comparable with Fixed Groups 1 and 2, as discussed above, the improvement in location updating cost is large enough to offset the paging cost, even though the cost of an update relative to a page is not very large.

For a larger relative cost of updates ($c = 10$), the worst performer is Fixed Group 0 for all levels of incoming calls (Figures 6.10, 6.11, 6.12, and 6.13). This was not surprising, since updates are relatively more expensive, and this algorithm generates by far the largest number of updates. As before, Fixed Groups 1 and 2 behave very similarly, with Fixed Group 1 again slightly ahead. Fixed Group 3, with the large location areas and corresponding low rates of updates, starts out the best of the fixed algorithms, but its large paging cost causes total cost to rise significantly with higher levels of incoming calls. For 9 incoming calls per day, Fixed Group 3 has a higher cost than Fixed Groups 1 and

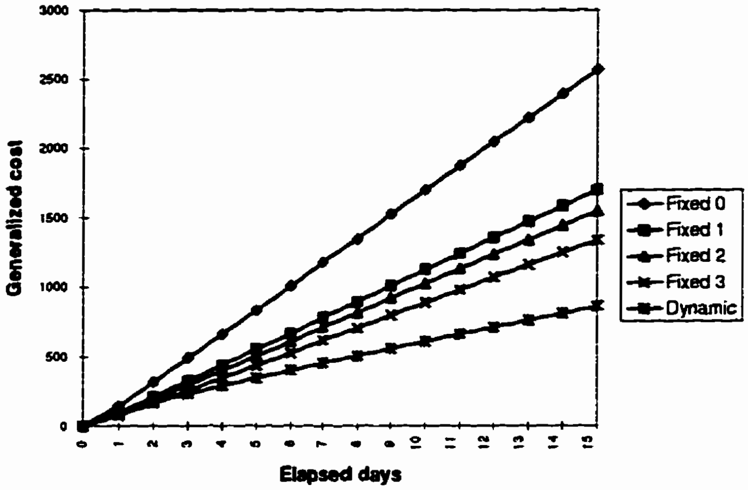

Figure 6.10: Global total cost (c = 10) for 3 incoming calls, versus elapsed time

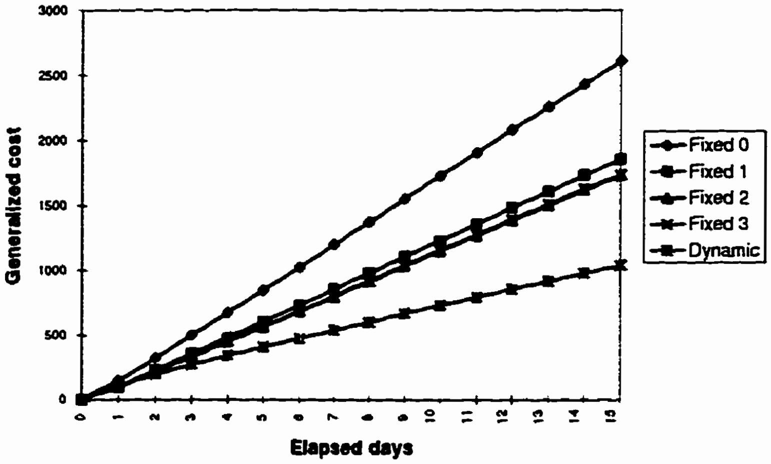

Figure 6.11: Global total cost (c = 10) for 6 incoming calls, versus elapsed time

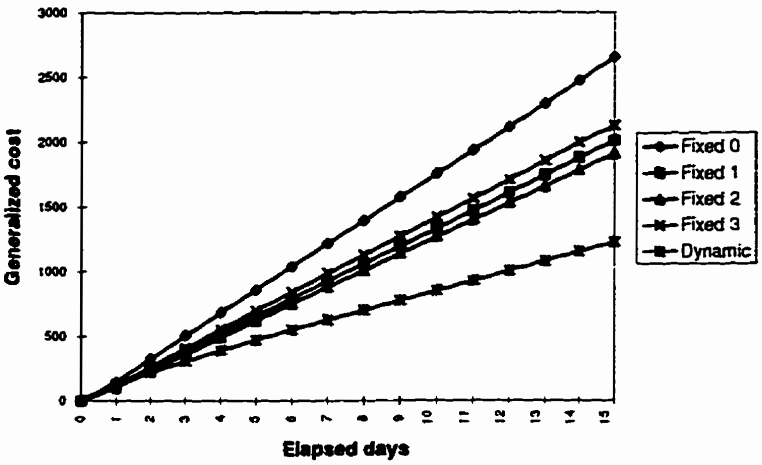

Figure 6.12: Global total cost (c = 10) for 9 incoming calls, versus elapsed time

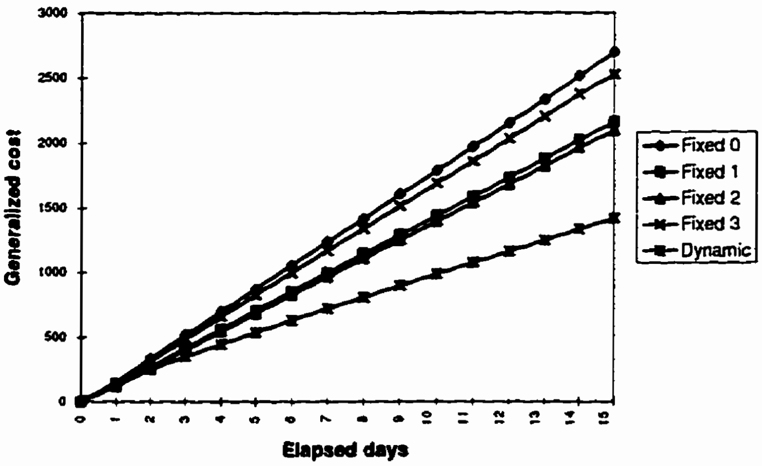

Figure 6.13: Global total cost (c = 10) for 12 incoming calls, versus elapsed time

2, and it approaches the cost of Fixed Group 0 for 12 incoming calls per day.

The dynamic algorithm again outperforms the fixed algorithms, after a 'learning curve' of 2 to 3 days. The gap between the total cost of the dynamic and fixed algorithms is even greater for $c = 10$, since the updating efficiency of the dynamic algorithm becomes even more apparent.

6.2.3 Average performance relative to subscriber type

The effect of different subscriber types on the performance of the various location management algorithms is discussed in this section. Recall that the four person types were full-time employed (Type 0), part-time employed (Type 1), unemployed student (Type 2), and unemployed, non-student (Type 3). In the simulation, the different types were equally represented. It should be noted that the following analysis used the cumulative results after 15 simulated days, both for analytical simplicity, and to allow the dynamic algorithm and the mobility model to stabilize their behaviour after the initial transient effects.

Location updating

The relative proportion of location updates for different person types remained almost identical for the fixed algorithms. Overall, the dynamic algorithm generated the fewest location updates for all person types (Figure 6.14). Relative to person type, there appear to be more location updates for part-time employed subscribers, followed by the full-time employed and student types, with the non-employed, non-student type exhibiting the lowest number of updates. Since the fixed algorithms make no distinction among the different trip characteristics of different person types, this relationship is interpreted as the relative level of trips undertaken by each type. Although it has not been further investigated, it does not seem counter-intuitive that part-time employed subscribers would travel more than subscribers who were neither employed nor students.

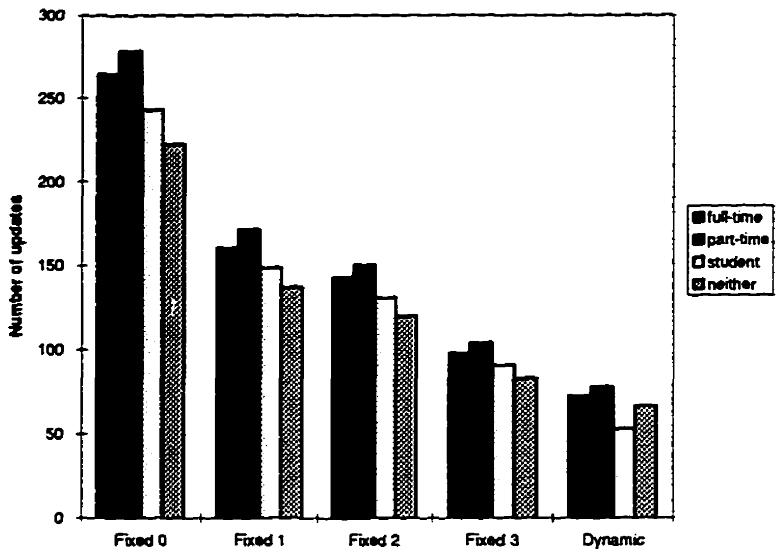

Figure 6.14: Average number of location updates per user type for different algorithms, after 15 days

The results for the dynamic algorithm were somewhat less intuitive, since the travel habits of a subscriber can affect the performance of the dynamic algorithm (for example, if the vast majority of trips by the subscriber were to activities located in fixed locations). Several observations can be made in comparing the absolute and relative number of location updates generated by the different subscriber types. First, the number of updates are significantly lower for all person types (P-values < 0.0001), indicating that all person types benefit from the implementation of the dynamic algorithm. Second, the relative number of updates generated by the student group is noticably smaller compared to the fixed algorithms, indicating that the relative performance of the dynamic algorithm is greater for the student group. The reason for this has not been further investigated, but a possibility could be that students travel to a more limited number of different cells, enabling the dynamic algorithm to work more efficiently by keeping almost all visited cells within the personal location area. Third, the relative number of updates for non-employed, non-students is higher compared to the fixed algorithms. Again, a possible reason could be that this person type travels to a larger set of different cells, and thus reducing the relative efficiency of the dynamic algorithm.

Paging

As with location updating, the fixed algorithms exhibit the expected paging behaviour, with the larger location areas incurring a higher cost. Note that since person type does not affect the number of cells paged for the fixed algorithms, and that equal numbers of each person type were input to the simulation, any differences in the number of cells paged across different person types for each fixed algorithm is due to the randomness inherent in the simulation. The number of cells paged for the different algorithms scales linearly with respect to the number of incoming calls, so for brevity, only the graph for 6 incoming calls is shown in Figure 6.15.

Subscriber person type did, however, play a role in the performance of the dynamic algorithm. Overall, the number of paging messages generated by the dynamic algorithm was roughly comparable with the number generated by fixed algorithm Groups 1 and 2. However, employed subscribers generated more paging messages than non-employed subscribers. The student type performed best with the dynamic algorithm, followed by the non-employed, non-student type. The full-time employed type generated the most paging messages, followed by the part-time employed type. The reason for this behaviour, as before, might be that students spend a large amount of time in few cells, thus forming a small paging area that can be accurately determined from the user profile. Conversely, the employed types may distribute their time over more activities (and cells), thus making the accurate derivation of the 'primary' paging area more difficult.

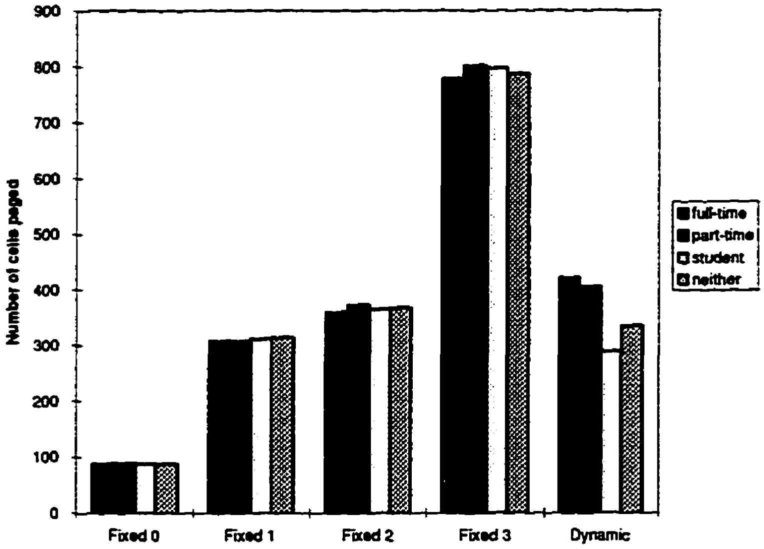

Figure 6.15: Average number of cells paged for 6 incoming calls, by user type, versus the different algorithms, after 15 days

Total location management cost

For a relative cost $c = 5$ and 3 incoming calls per day (Figure 6.16), Fixed Group 0 is the worst performer, with the rest of the fixed algorithms having similar costs. As the number of incoming calls increases (Figures 6.17, 6.18, and 6.19), the relative cost of Fixed Group 0 decreases due to its optimal paging cost, until total cost for Fixed Groups 0, 1 and 2 are almost equal at the level of 12 incoming calls. The performance of Fixed Group 3 deteriorates, however, as more incoming calls arrive, due to the high paging cost of Fixed Group 3, and for 9 and 12 incoming calls, it is the most expensive algorithm.

With c equal to 10, Fixed Group 0 is relatively disadvantaged, due to its frequent updates, while Fixed Group 3 is relatively advantaged. For fewer incoming calls (Figures 6.20 and 6.21) Fixed Group 3 is the least expensive of the fixed algorithms, although the dynamic algorithm is significantly less expensive for all user types. For 9 and 12

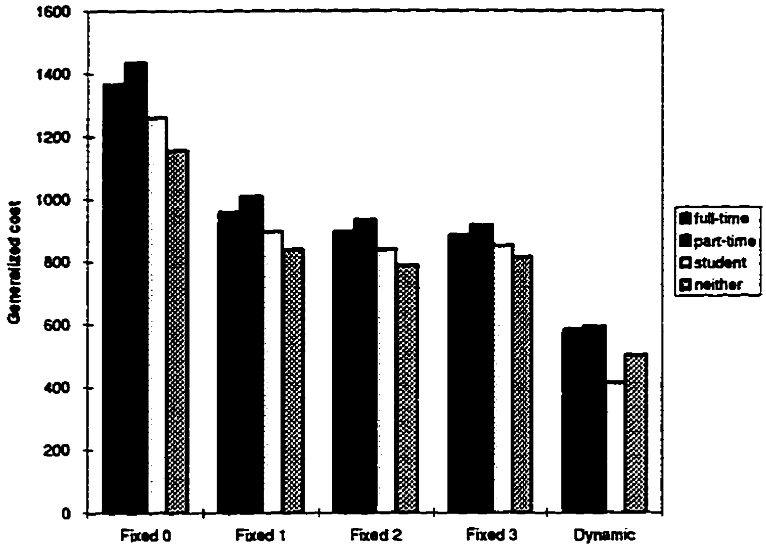

Figure 6.16: Total location management cost (c = 5) for 3 incoming calls, by user type, versus the different algorithms, after 15 days

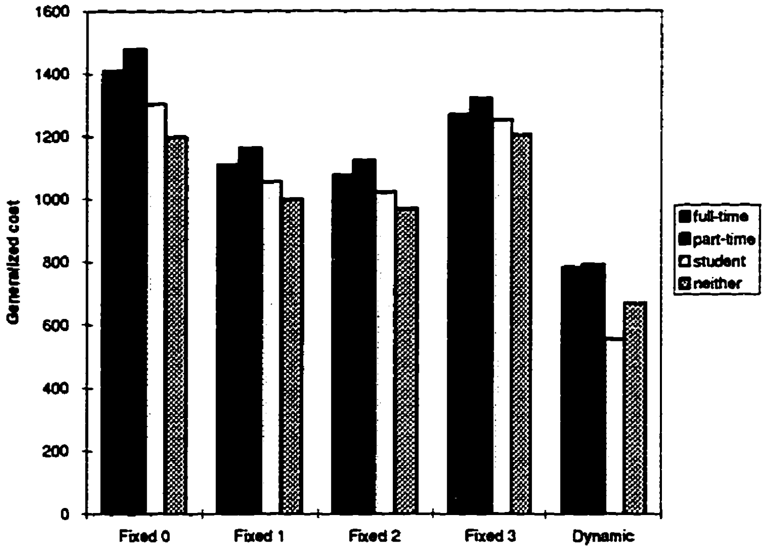

Figure 6.17: Total location management cost (c = 5) for 6 incoming calls, by user type, versus the different algorithms, after 15 days

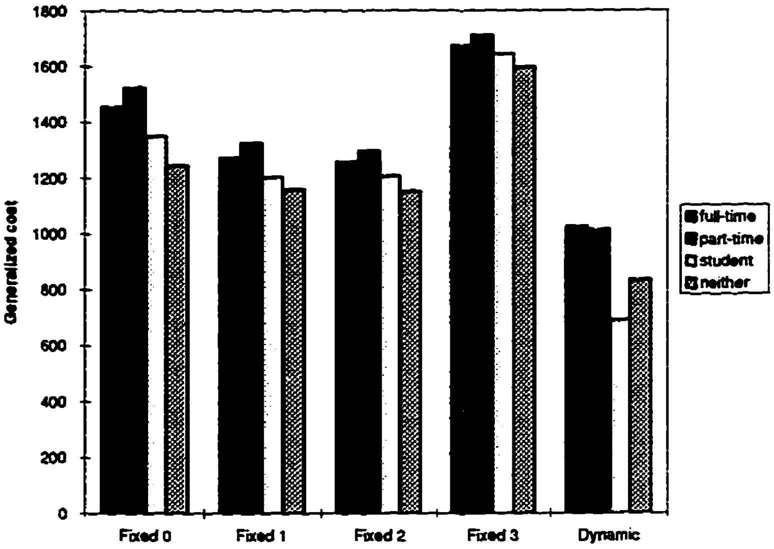

Figure 6.18: Total location management cost (c = 5) for 9 incoming calls, by user type, versus the different algorithms, after 15 days

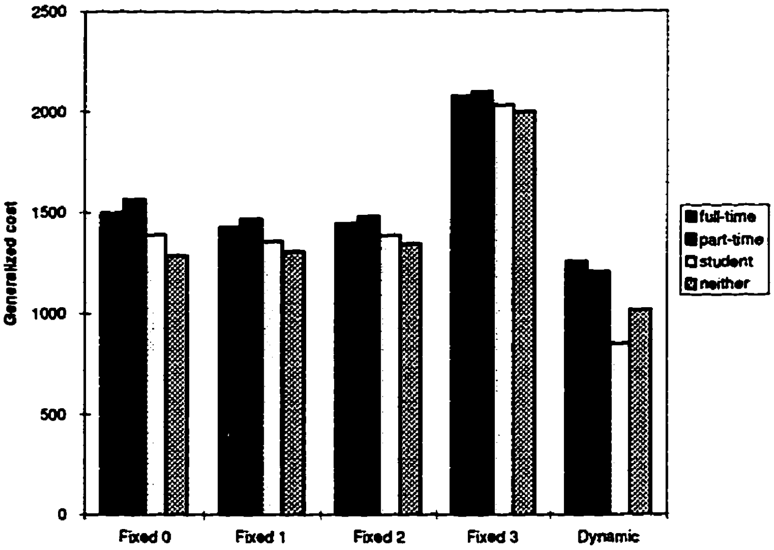

Figure 6.19: Total location management cost (c = 5) for 12 incoming calls, by user type, versus the different algorithms, after 15 days

incoming calls per day (Figures 6.22 and 6.23), Fixed Group 3 is no longer the least expensive fixed algorithm due to its higher paging cost, and Fixed Groups 1 and 2 incur lower costs. Again, the dynamic algorithm is the least expensive and Fixed Group 0 is the most expensive.

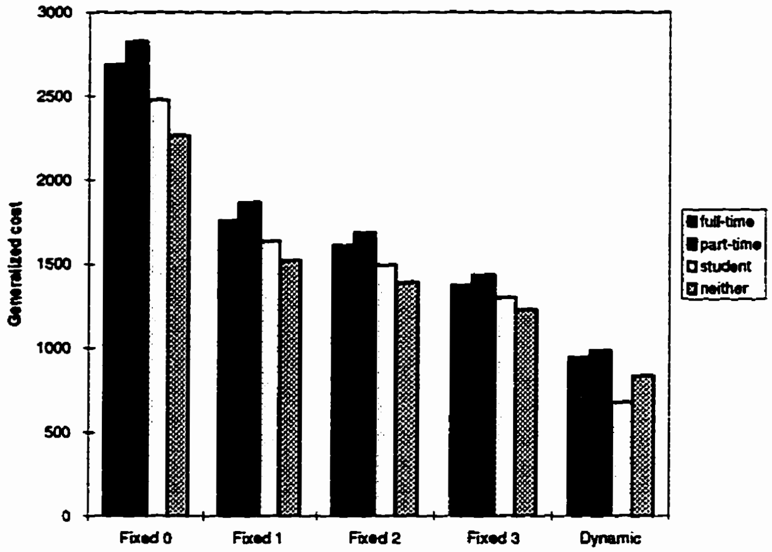

Figure 6.20: Total location management cost (c = 10) for 3 incoming calls, by user type, versus the different algorithms, after 15 days

Note that the relative difference among the different user types is smaller for Fixed Group 3 than for Fixed Group 0. This is because the cost for Fixed Group 3 is mainly influenced by the number of paging messages, which were nearly equal across the different user types. Due to this effect, the total cost in the case of 12 incoming calls, and c = 10, for the non-employed, non-student type with Fixed Group 0 is about 1% less than the cost with Fixed Group 3, even though the cost for the part-time employed type with Fixed Group 0 is actually 13% *more* than with Fixed Group 3.

For all levels of incoming calls, and all person types, the dynamic algorithm had the

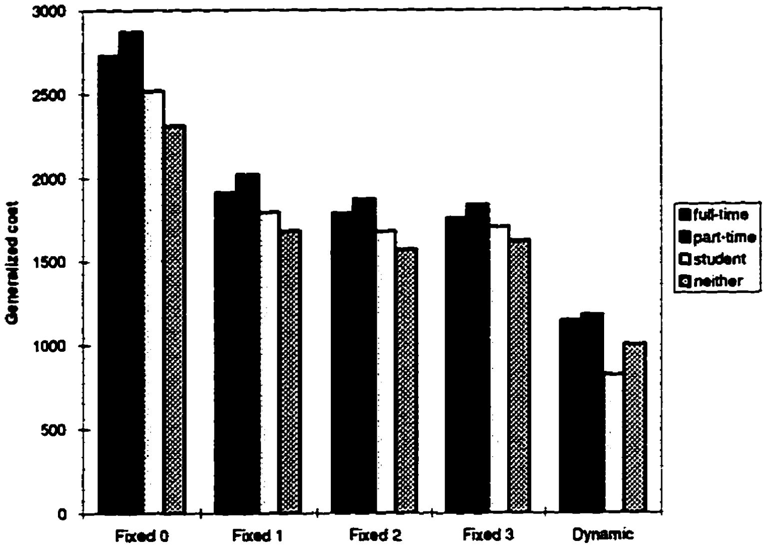

Figure 6.21: Total location management cost (c = 10) for 6 incoming calls, by user type, versus the different algorithms, after 15 days

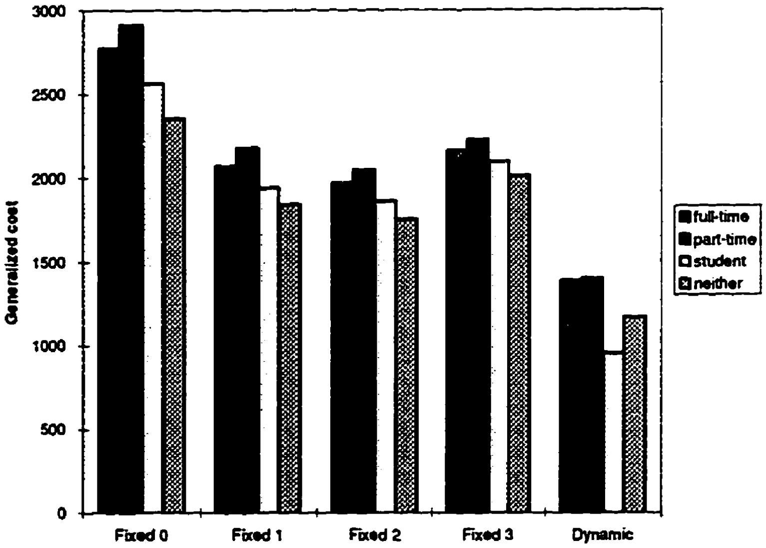

Figure 6.22: Total location management cost (c = 10) for 9 incoming calls, by user type, versus the different algorithms, after 15 days

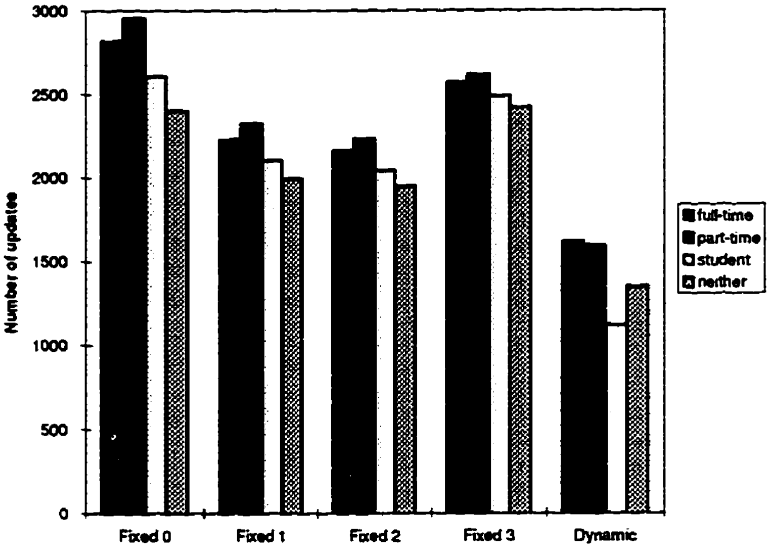

Figure 6.23: Total location management cost (c = 10) for 12 incoming calls, by user type, versus the different algorithms, after 15 days

lowest cost, especially for the student type. The relative behaviour of the different person types was similar for both updating and paging, so it remains similar for total cost. The employed person types had higher total cost; a possible, but untested, reason for this may be a higher trip rate associated with the employed types.

6.2.4 Average performance relative to fixed work location

The performance of the location management algorithms for employed subscribers who work at fixed or variable work locations is discussed in this section. Since the non-employed person types did not make any significant number of work trips, they will not be considered here. Also, note that the cumulative results after 15 simulated days were used, as in the previous person type discussion.

Location updating

As with subscriber type, the relative proportion of updates among work location categories for the fixed algorithms remained nearly constant. Overall, as shown in Figure 6.24, the dynamic algorithm generated the fewest updates for all work location categories (statistically significant with P-values < 0.0001). This shows that all work location categories benefit from the implementation of the dynamic algorithm.

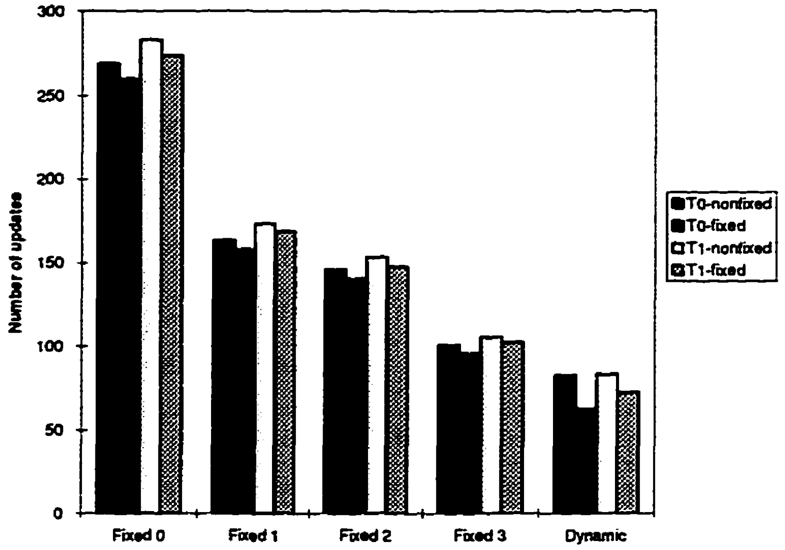

Figure 6.24: Average number of location updates by fixed work location category, for full-time (T0) and part-time (T1) users, versus the different algorithms, after 15 days

Compared to the fixed algorithms, the dynamic algorithm appeared to generate a lower number of updates for the fixed work location categories, particularly for the full-time employed, but also for the part-time employed. This observation supports the expected behaviour of the dynamic algorithm, which performs better when there is less randomness in the activity destinations. Since work is a major activity for these person types, a fixed work location should significantly improve the performance of the dynamic algorithm.

Paging

The fixed algorithms were, as expected, not affected by the different work location categories. The number of cells paged also scaled linearly with the number of incoming calls, so for brevity only the graph for 6 incoming calls is shown, in Figure 6.25. The dynamic algorithm, however, displayed relatively large variations in performance depending on the work location category. Subscribers with non-fixed work locations, both full-time employed and part-time employed, generated substantially more paging messages than the fixed work location groups. For fixed work location groups, the dynamic paging algorithm performed as well as, or better than, the Fixed Group 1 and 2 algorithms.

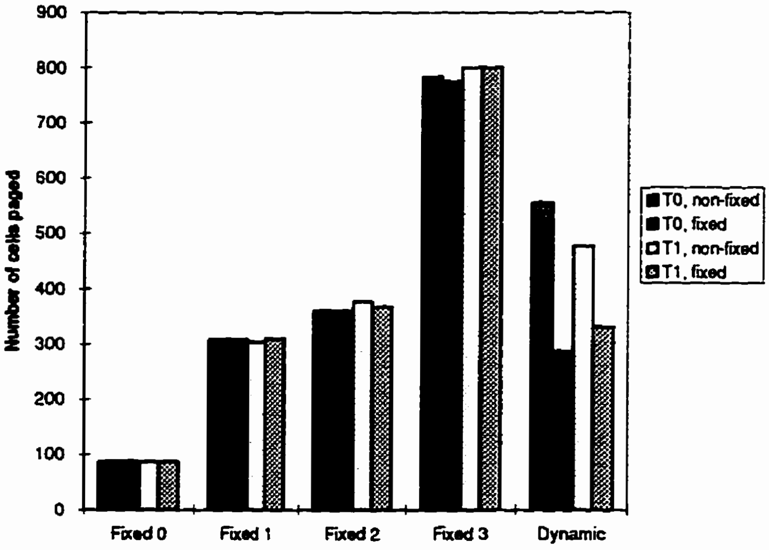

Figure 6.25: Average number of cells paged for 6 incoming calls, by fixed work location category, for full-time (T0) and part-time (T1) users, versus the different algorithms, after 15 days

Total location management cost

Among the fixed algorithms, a general pattern arises where the fixed work location full-time employed subscribers generate a lower total cost, and variable work location part-time employed subscribers generate a higher total cost, than the other two categories. The differences are relatively small, but the pattern is generally present.

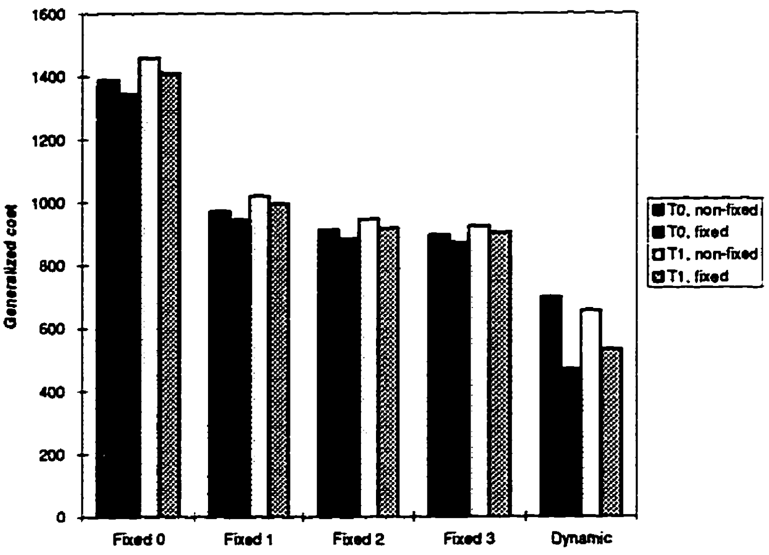

Figure 6.26: Total location management cost (c = 5) for 3 incoming calls by fixed work location category, for full-time (T0) and part-time (T1) users, versus the different algorithms, after 15 days

For relatively less expensive updates (c = 5), Fixed Group 0 is the most expensive algorithm for 3 and 6 incoming calls (Figures 6.26, 6.27), but is overtaken by Fixed Group 3 as the number of incoming calls increases (Figures 6.28, 6.29). Fixed Groups 1 and 2 have similar costs, which are lower than the other two fixed groups for larger incoming call levels. For more expensive updates (c = 10), Fixed Group 3 benefits from its fewer generated updates. It remains the least expensive for low numbers of incoming calls (3

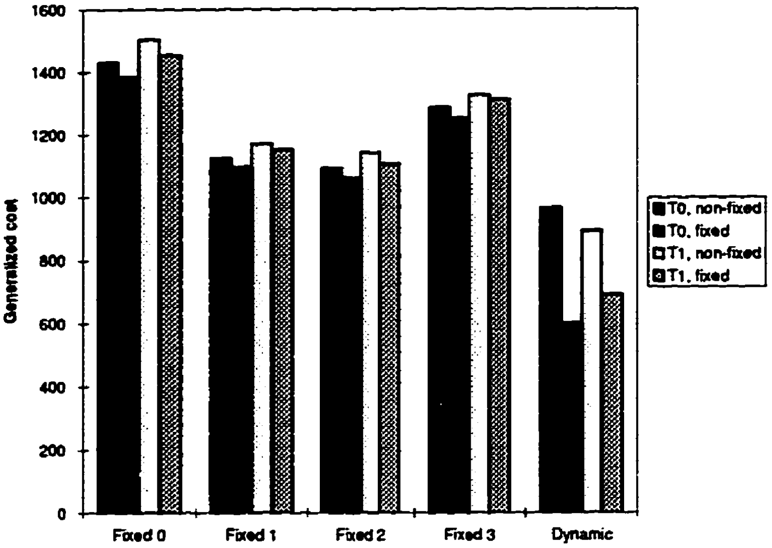

Figure 6.27: Total location management cost (c = 5) for 6 incoming calls by fixed work location category, for full-time (T0) and part-time (T1) users, versus the different algorithms, after 15 days

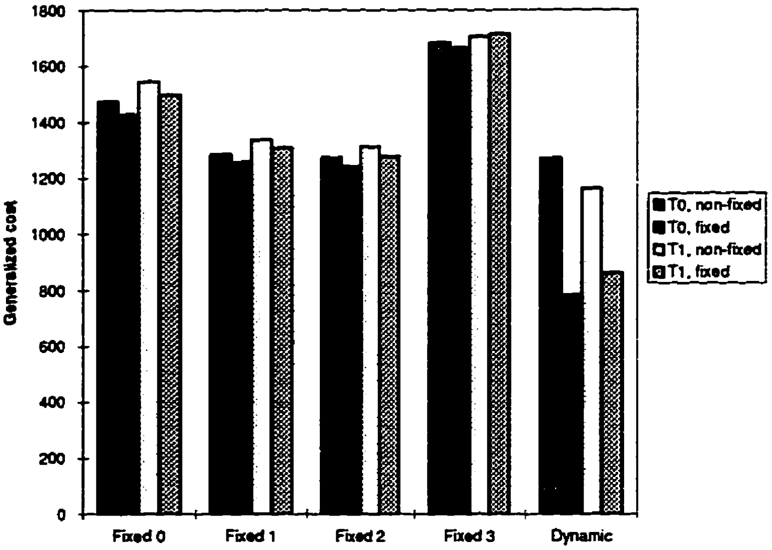

Figure 6.28: Total location management cost (c = 5) for 9 incoming calls by fixed work location category, for full-time (T0) and part-time (T1) users, versus the different algorithms, after 15 days

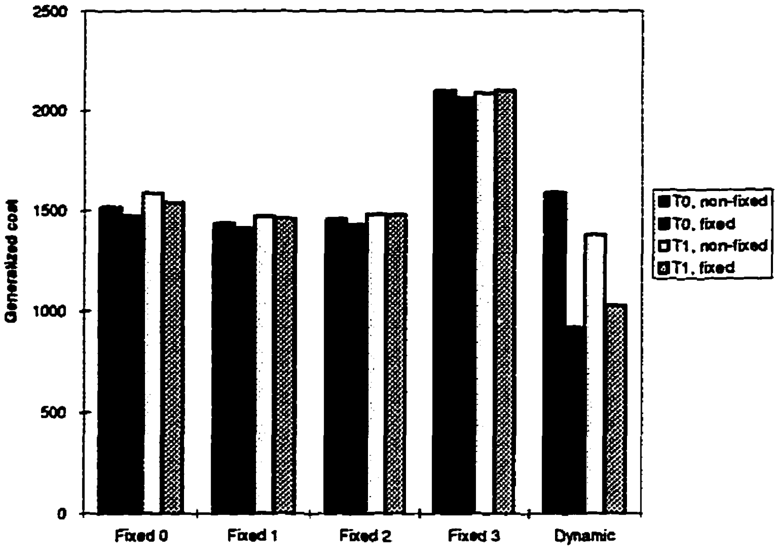

Figure 6.29: Total location management cost (c = 5) for 12 incoming calls by fixed work location category, for full-time (T0) and part-time (T1) users, versus the different algorithms, after 15 days

and 6 per day, as shown in Figures 6.30 and 6.31), and even though its total cost increases relative to other algorithms as the number of incoming calls increases (Figures 6.32 and 6.33), it remains below the total cost of Fixed Group 0.

Subscribers with a fixed work location, especially full-time employees, benefit significantly from the dynamic algorithm. For subscribers with a variable work location, especially full-time employees, the high degree of randomness does not enable the dynamic paging algorithm to consistently and accurately identify a small, initial paging area. As discussed above, the dynamic location updating is not as impacted by randomness as the dynamic paging. Due to this, for relatively more expensive updates (c = 10), the dynamic algorithm outperforms the fixed algorithms for all categories of fixed work location. For relatively more expensive paging (c = 5), the full-time employed group with variable work locations has a slightly higher total cost for 9 and 12 incoming calls than

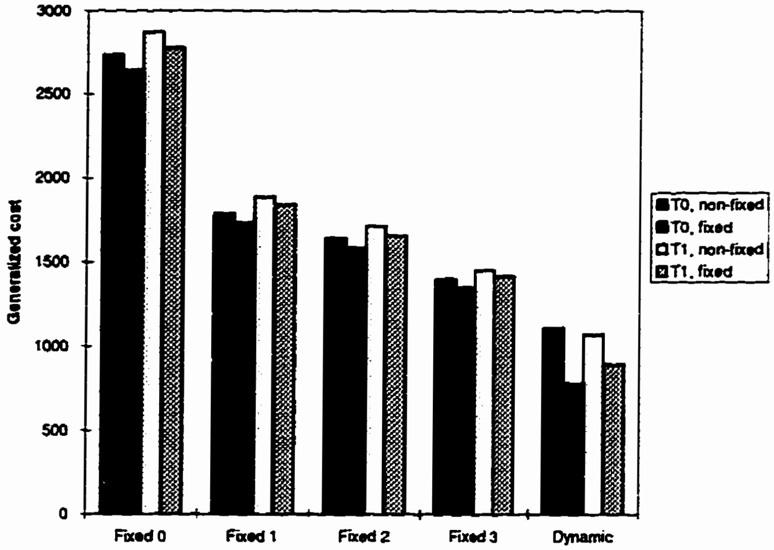

Figure 6.30: Total location management cost (c = 10) for 3 incoming calls by fixed work location category, for full-time (T0) and part-time (T1) users, versus the different algorithms, after 15 days

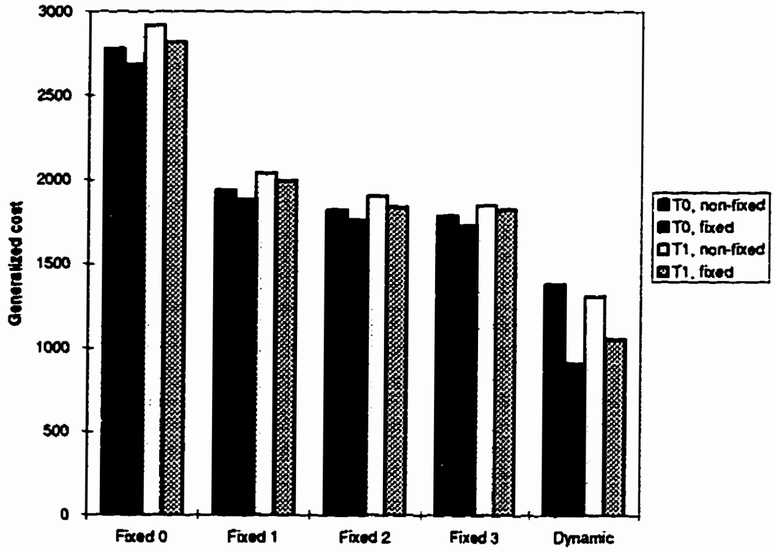

Figure 6.31: Total location management cost (c = 10) for 6 incoming calls by fixed work location category, for full-time (T0) and part-time (T1) users, versus the different algorithms, after 15 days

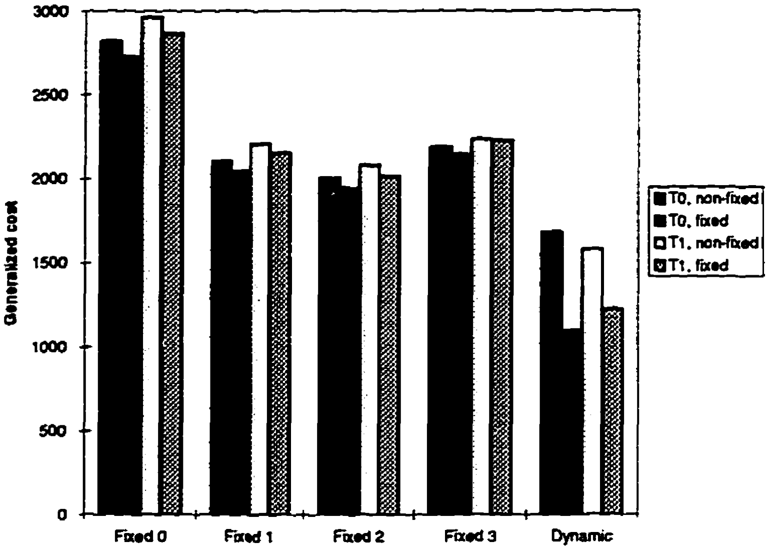

Figure 6.32: Total location management cost (c = 10) for 9 incoming calls by fixed work location category, for full-time (T0) and part-time (T1) users, versus the different algorithms, after 15 days

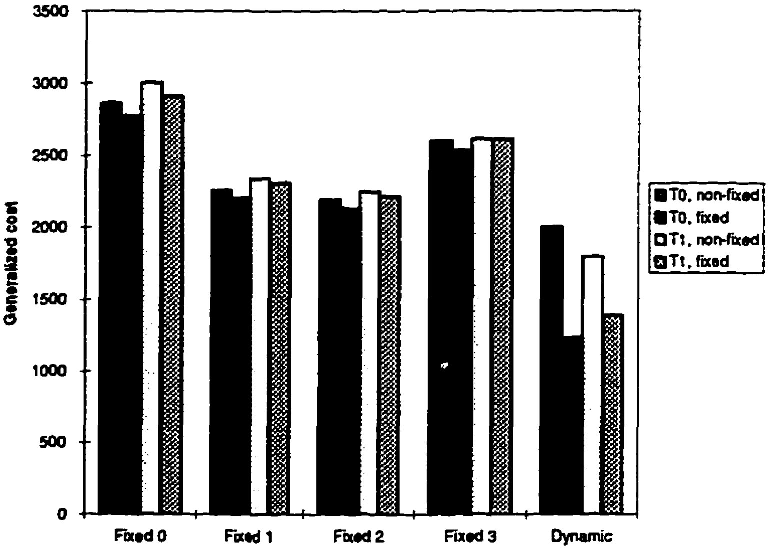

Figure 6.33: Total location management cost (c = 10) for 12 incoming calls by fixed work location category, for full-time (T0) and part-time (T1) users, versus the different algorithms, after 15 days

Fixed Groups 0, 1 and 2, although still significantly less than the cost for Fixed Group 3.

6.2.5 Comparison of different dynamic location area sizes

As described in Chapter 5, the dynamic algorithm generates a 'personal' location area during a location update, based on the user profile. The maximum size of this personal location area was one of the independent variables of the simulation. The four different location areas had a maximum of 5, 10, 15, and 20 cells. Recall that Fixed Groups 1 and 2 had roughly 4 cells per location area, and Fixed Group 3 had about 9 cells per location area. Intuitively, a larger location area, combined with the dynamic paging algorithm, would provide better performance than a smaller location area, since the number of location updates would be reduced. Obviously, the size of the dynamic location area does not affect the fixed algorithms in any way, and any differences shown on the graph arise from the randomness of the simulation.

Location updating

The general trend observed in Figure 6.34 was that the number of location updates decreased, but at a slower rate, as the dynamic location area size increased. With a location area of 5 cells, the dynamic algorithm generated slightly more updates than Fixed Group 3, although significantly fewer updates than Fixed Groups 1 or 2 (P-values < 0.0001). For all other dynamic location area sizes, the dynamic algorithm outperformed all fixed algorithms (P-values < 0.0001).

Paging

The performance of the algorithms scaled almost linearly as the number of incoming calls increased, so for brevity only the graph for 6 incoming calls is shown in Figure 6.35. The number of paging messages generated by the dynamic algorithm increased as the dynamic

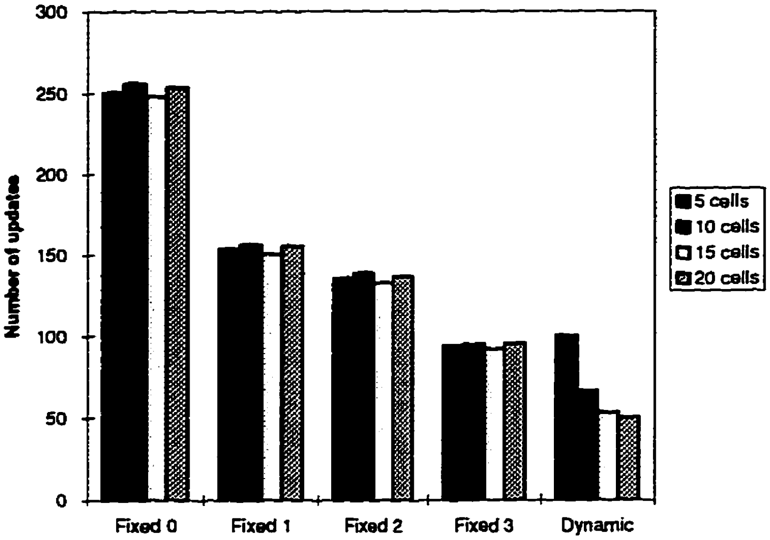

Figure 6.34: Average number of location updates by dynamic location area maximum size, versus the different algorithms, after 15 days

location area size increased. This was not unexpected, since a larger dynamic location area, assuming a similar rate of failure for finding the subscriber in the first paging area, would result in more cells being paged. However, the dynamic paging algorithm is quite efficient, and did not scale linearly. For example, doubling the dynamic location area size from 10 cells to 20 cells caused roughly a 40% increase in paging messages. The fixed algorithms were obviously not affected by the dynamic location area size, and any differences between the dynamic location area categories for fixed algorithms are due to random fluctuations in the simulation.

The paging cost for a dynamic location area of size 5 was somewhat lower than an extrapolation from the other levels would suggest. A possible reason for this is that the penalty for a 'cache miss', when the first paging area tried does not contain the subscriber, is relatively small compared to the other location area sizes. This is because

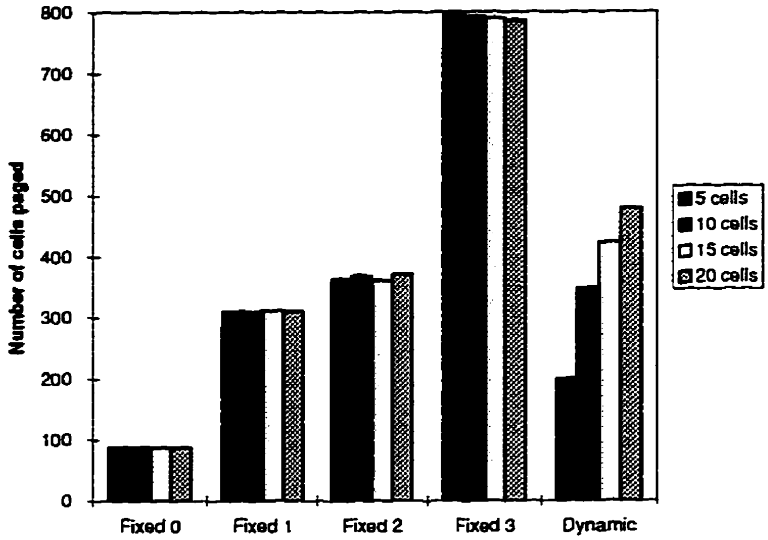

Figure 6.35: Average number of cells paged for 6 incoming calls, by dynamic location area maximum size, versus the different algorithms, after 15 days

the first paging area, from preliminary observations, is usually quite small (on the order of two or three cells).

Total location management cost

In terms of total location management cost, the dynamic algorithm was more efficient than all fixed algorithms, for both $c = 5$ and $c = 10$. For low levels of incoming calls, the relatively high updating cost of the 5-cell dynamic location area made that configuration the most expensive. As the number of incoming calls increased, the more expensive paging cost of the larger location area sizes was more dominant, especially for lower cost updates ($c = 5$), as shown in Figures 6.36, 6.37, 6.38, and 6.39.

For relatively less expensive paging ($c = 10$) shown in Figures 6.40, 6.41, 6.42, and 6.43, the almost linear relationship between larger dynamic location areas and higher paging cost was less dominant, and the smaller location areas were more expensive for lower levels of incoming calls. However, the high paging cost of the larger location areas was still significant, and for 12 incoming calls, the 20-cell dynamic location area was the most expensive among the different location area sizes.

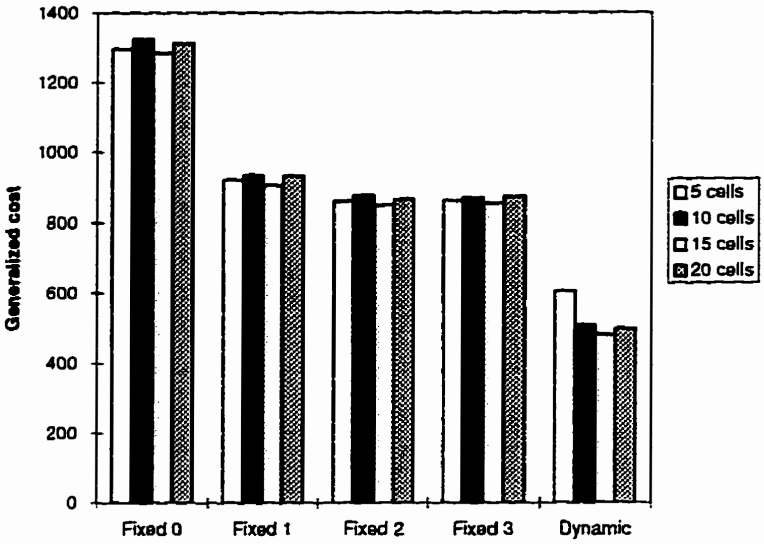

Figure 6.36: Total location management cost ($c = 5$) for 3 incoming calls by dynamic location area size versus the different algorithms, after 15 days

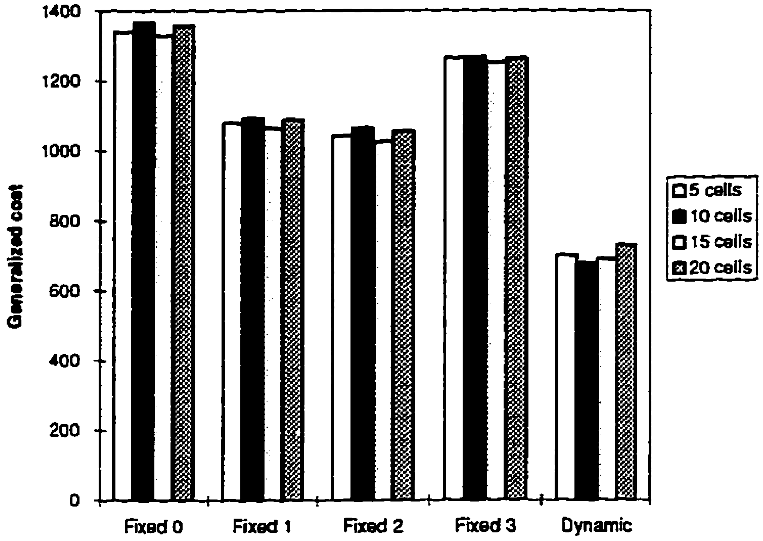

Figure 6.37: Total location management cost (c = 5) for 6 incoming calls by dynamic location area size versus the different algorithms, after 15 days

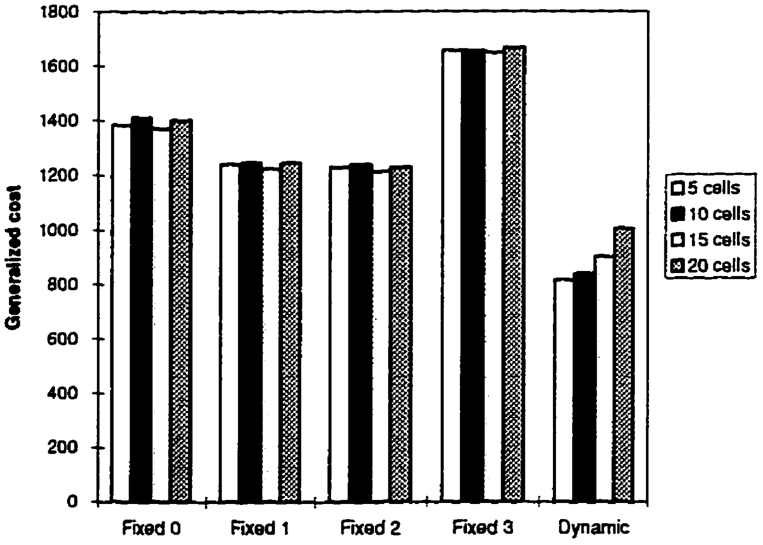

Figure 6.38: Total location management cost (c = 5) for 9 incoming calls by dynamic location area size versus the different algorithms, after 15 days

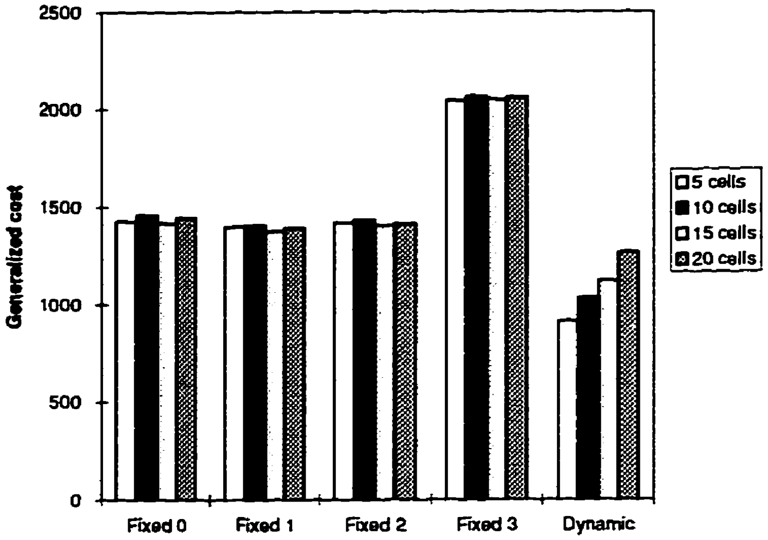

Figure 6.39: Total location management cost (c = 5) for 12 incoming calls by dynamic location area size versus the different algorithms, after 15 days

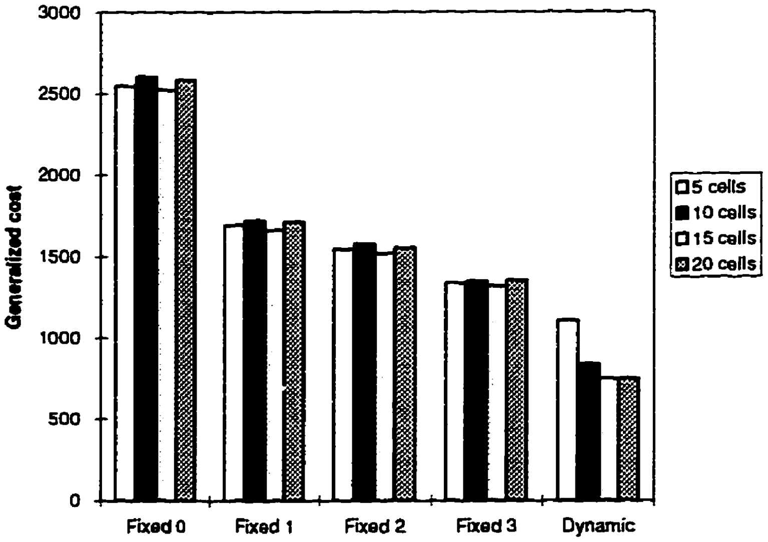

Figure 6.40: Total location management cost (c = 10) for 3 incoming calls by dynamic location area size versus the different algorithms, after 15 days

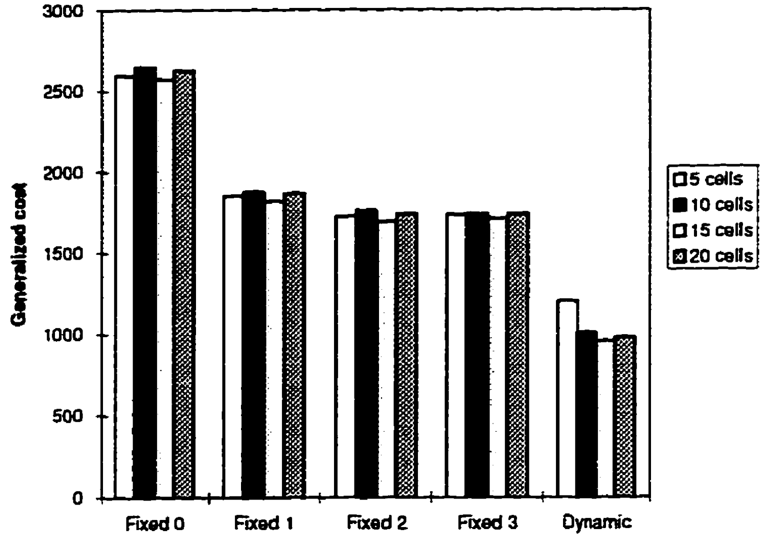

Figure 6.41: Total location management cost (c = 10) for 6 incoming calls by dynamic location area size versus the different algorithms, after 15 days

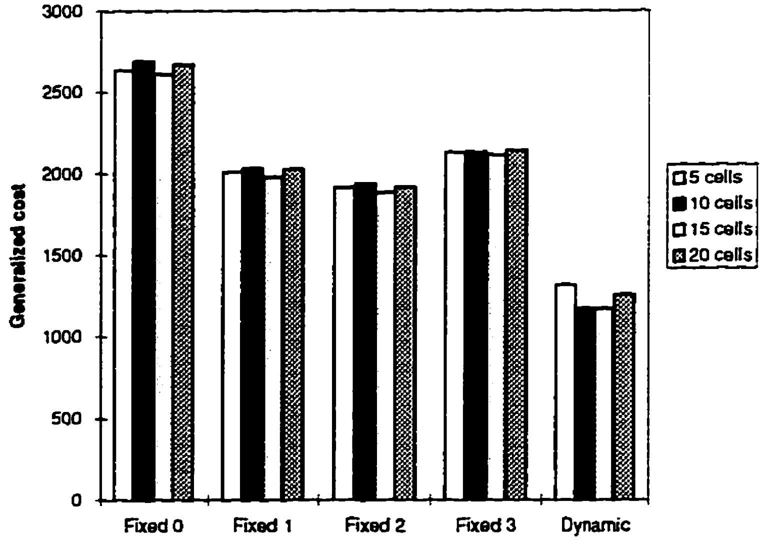

Figure 6.42: Total location management cost (c = 10) for 9 incoming calls by dynamic location area size versus the different algorithms, after 15 days

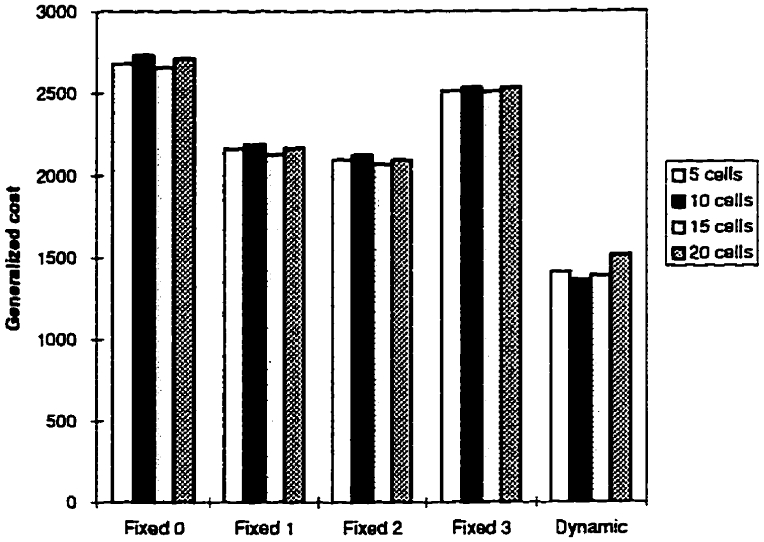

Figure 6.43: Total location management cost (c = 10) for 12 incoming calls by dynamic location area size versus the different algorithms, after 15 days

Chapter 7

Conclusions

Two distinct but related topics were addressed in the previous chapters. The first was a mobility model designed to simulate realistic mobility behaviour patterns, based on stochastically derived activities using data from a regional trip survey. A particular location and duration are associated with each activity. A trip thus consists of the path between the locations of adjacent activities. The mobility model was adapted for cellular network subscribers, by defining locations and paths in terms of radio cells.

The model was used as a tool for the second topic of the thesis, a location management algorithm utilizing user profiles for the dynamic creation of location and paging areas. Based on the number of times a subscriber had moved from the current cell to neighbouring cells, a personal location area is dynamically defined during a location update. A history of the average time spent in each cell is used dynamically to determine a subset of cells within the personal location area most likely to contain the subscriber, and which is paged first.

7.1 Proposal evaluation

During the evaluation of previous work in location management in Chapter 3, as well as when developing the proposed algorithms, the following overall goals were set. They represent an ideal situation, since in most cases, there are certain inherent trade-offs. The behaviour of the proposed location management algorithm will be discussed in this section.

Minimize signalling, both over the radio interface and within the network, to reduce signalling bandwidth and increase capacity

As discussed extensively in Chapter 6, the dynamic algorithm outperforms the standard fixed algorithms, in terms of total signalling cost, for nearly all different subscriber types defined, levels of incoming calls, work location variability, and dynamic location area size. The proposed dynamic algorithm would significantly reduce location management cost compared to currently used techniques. Due to the variable nature and number of the parameters involved, an exact measurement of the improvement would be rather meaningless, but significant improvement can be seen in almost all analyses performed in Chapter 6. As discussed below, the dynamic algorithm passes marginally more data during a location update, since the location information passed is a set of cell identifiers, as opposed to one location area identifier, although there are many fewer updates generated.

Minimize call setup delays (especially due to paging procedures) in order to maintain acceptable quality of service

Some of the proposed paging algorithms (see Chapter 3) paged successive paging areas, within some large location area, in an attempt to locate the subscriber. Because of the delays involved in paging (waiting for a response, possible repeated paging due to the harshness of the radio environment, and eventual call setup), repeating the entire process for more than one or two paging areas would lead to unacceptably high delays, in terms of quality of service.

The proposed paging algorithm pages at most two paging areas, and although this may in some cases result in higher paging costs, a short overall call setup time is preferable. If necessary, the algorithm is easily modifiable to page more than two paging areas.

Maximize the applicability of the location management algorithms

Some location management algorithms suggest that subscribers be classified into certain user classes, such as a highly mobile class, or a pedestrian class. Unless some method can be used to update this status, this is an inefficient mechanism since user class changes for each subscriber throughout the day (e.g., highly mobile while in a car, stationary at work or home, or pedestrian while shopping).

The proposed algorithm does not make any assumptions on the mobility status of a subscriber. The user profile associated with each subscriber provides a relatively efficient, and remarkably effective, method for storing location management information.

Minimize the computational complexity of algorithms and the amount of data stored for each subscriber, to reduce costs

In providing its significant location management cost savings, the dynamic algorithm requires some additional logic and memory on the mobile station and network (base station and/or switch). In terms of computation, the proposed algorithm requires the mobile station to record the average time spent in each cell, and the number of transitions from one cell to another. This requires only a timer (and a couple of elementary operations to update the previous average duration) and a set of counters, of which there are several in modern digital cellular protocols. In addition, the comparison process when entering a new cell is more involved, requiring the comparison of a cell identifier against a list of cell identifiers (the personal location area), as opposed to the comparison of a Location Area Identifier (LAI) against the broadcast LAI. The number of comparisons, however, is on the order of ten or twenty.

Another additional cost of the dynamic algorithm is the transmission of the user profile to the network. Only the information related to paging (i.e., the average visit duration for cells in the current location area) needs to be transferred. This transfer would generally be done during a location update, making use of the existing signalling channel. The amount of data transferred is directly related to the size of the dynamic location area, which itself is on the order of tens of cells.

The size of the user profile, which determines how much cell transition and stay duration information can be used by the dynamic algorithm, was assumed in the simulation to be unlimited. In practice, it would have been approximately 45 cells × [cell identifier (1 integer) + 6 labelled neighbouring cell transition counters (2 integers each) + average stay duration (1 integer)], or roughly 2.5 kbytes assuming 4-byte integers. The latest GSM (Phase 2) SIM cards have 8 kbytes of user memory available. A larger user profile contains the associated location management information of more cells, and thus allows the algorithm to function more effectively. When mobile station memory is limited, the problem of maintaining current and usable location management information becomes similar to a cache replacement problem. Although this has not been considered in detail, one possible strategy would be to replace cells which have been visited the fewest times, since these would be less likely to be visited again (care should be taken not to replace newly 'popular' cells).

On the network side, the HLR would need additional memory to store the list of cells forming the current dynamic location area, as opposed to storing only the current LAI. The associated paging-related information (average stay duration for the cells in the current dynamic location area) would also need to be stored. Assuming 20 cells per location area, this would be approximately 20 cells × [cell identifier (1 integer) + average stay duration (1 integer)], or 160 bytes, assuming 4-byte integers. Assuming 500,000 subscribers per HLR, the dynamic algorithm would require an additional 80 Mbytes of storage. Although it is a small additional cost, the addition of more storage to an HLR would be a favourable trade-off for the additional radio interface capacity provided by the dynamic algorithm, since radio spectrum is a fixed resource.

Distribute location update requests evenly among cells, rather than concentrating them in cells on the border of location areas.

Although this issue was not examined in detail, it is fairly intuitive that the dynamic algorithm would indeed distribute location updates evenly. Initially, an empty user profile would use the overlaid fixed location areas, and would thus generate updates only along location area borders. However, as a user profile is built, a subscriber will form individualized location area boundaries, which also vary with time. Since a user profile is unique, aggregating the resulting individual location area boundaries, especially since they vary in time, would statistically result in randomly distributed boundaries and location updates.

7.2 Future work

The scope of both the mobility model and the dynamic algorithm is quite large, and several possibilities and alternatives were covered only partially or not at all.

The concept of aging the user profile data would likely need to be considered in an actual implementation, in order to reduce the time needed by the algorithm to adapt to drastic changes in subscriber behaviour. For example, consider a subscriber moving to a new home (in a different cell). The old home cell would have a large average stay duration, and would very likely be paged first if included in the location area. Similarly, the number of transitions from neighbouring cells into the home cell would be very high, and thus the cell would very likely be selected in nearby dynamically generated location areas. In both cases, the behaviour of the algorithm may be different from that desired for effective location management. Aging the user profile data would gradually reduce the influence of previously active cells which are no longer as important. This would require minimal computational power to periodically multiply the user profile values by a reasonable depreciation factor.

Another area for further experimentation and possible improvement is the selection algorithm for the dynamic location area, given the user profile. The algorithm used here, as described in Chapter 5, selected neighbouring cells with link weights (the number of transitions from one cell to a neighbouring cell) above the average link weight for the source cell. The procedure was then repeated using the selected cells, in order of their link weight, until no other cells were found in the user profile, or until the personal location area was filled up. It would be interesting to compare the relative performance of several other possible alternatives. As a simple example, a certain fixed number of neighbouring cells may be selected, according to link weight, eliminating the need for calculating an average. As discussed in Chapter 5, there were intuitive reasons for using the average link weight in the cell selection process.

If it was decided to use a larger location area, allowing the addition of cells from several levels away (or *rings*, assuming a regular cell layout), it would be useful to further limit the selection of cells for the dynamic location area. At each node (cell), a decision is made on whether to consider any of its outgoing links, or to stop and move on to another cell at the same level. The number of cells to be added should be small, and only cells with a high possibility of being traversed should be added to the dynamic location area.

One possible decision metric is based on the ratio of spread to peak of the link weight distribution. One of the requirements for the algorithm is to keep the computational cost and complexity low, so the metric selected is the ratio of the difference between the maximum and minimum link weights, over the mean link weight value. The rationale behind this metric deals with whether a subset of the link weights has large enough values compared to the rest, as to increase greatly the probability of that subset of links being traversed by the subscriber. If the difference between the maximum and minimum link weights is small relative to the mean, it suggests that the link weight values are all relatively close together. In this situation, there is no subset of link weights which is much more likely to be selected than others, and therefore the links should not be considered. On the other hand, if the difference is large compared to the mean, it indicates that the link weight values are relatively spread apart; some will have low values, while others will have high values. In this case, the nodes on links with the relatively high values (selected through comparison with the mean link weight) deserve inclusion in the new location area. If this difference-to-mean ratio is greater than or equal to one, the node will be considered for addition to the new location area. Alternatively, a fixed number of nodes with the highest values of the ratio could be selected.

Another future experiment would be to implement a random mobility model, and compare the performance of the different location management algorithms using both the proposed and the random mobility models. It would be expected that the dynamic algorithm would not perform as well when using a completely random mobility model, since the randomness of cells visited would not allow useful user profiles to be constructed.

Finally, the only algorithm used for comparison here was the one currently implemented in systems like GSM. A comparison with some of the proposed models would be interesting, although for some of them the underlying assumptions and architecture used here might not apply.

7.3 Summary

An intelligent location management algorithm was developed, which dynamically generates individualized location areas and paging areas based on a user profile. The user profile stores the average duration spent in each cell, and the number of transitions from each cell to each of its neighbours.

A mobility model was also developed, which simulates the daily movements of mobile subscribers (in terms of a sequence of radio cells), incorporating realistic, individualized activity patterns and geographical focal points. The mobility model was used to compare the performance of the dynamic location management algorithm with several variations of the fixed algorithm used in current cellular networks. The dynamic algorithm significantly outperformed the fixed algorithms in terms of total location management cost, at the cost of additional logic and memory in the mobile station and network.

Bibliography

[1] Tranplan Associates. *Waterloo Region Travel Survey 1987: An Overview of the Survey Findings.* Regional Municipality of Waterloo, Department of Planning and Development, October 1989.

[2] Kay W. Axhausen and Tommy Gärling. Activity-based approaches to travel analysis: conceptual frameworks, models, and research problems. *Transport Reviews*, 12(4), 1992.

[3] B. R. Badrinath, T. Imielinski, and A. Virmani. Locating strategies for personal communication networks. In *Workshop on Networking of Personal Communications Applications*, December 1992.

[4] Amotz Bar-Noy, Ilan Kessler, and Moshe Sidi. Mobile users: To update or not to update? In *IEEE INFOCOM*, volume 2, 1994.

[5] Michael J. Bruton. *Introduction to Transportation Planning.* UCL Press, 3rd edition, 1985.

[6] F. Stuart Chapin and Henry C. Hightower. Household activity systems — a pilot investigation. Monograph, Center for Urban and Regional Studies University of North Carolina, Chapel Hill, May 1966.

[7] Gihwan Cho and Lindsay F. Marshall. An efficient location and routing scheme for mobile computing environments. *IEEE Journal on Selected Areas in Communications*, 13(5), June 1995.

[8] Juan de Dios Ortuzar and Luis G. Willumsen. *Modelling Transport*. John Wiley and Sons, second edition, 1994.

[9] Thomas A. Domencich and Daniel McFadden. *Urban Travel Demand: A Behavioral Analysis*. North-Holland, 1975.

[10] C. Eynard et al. A methodology for the performance evaluation of data query strategies in universal mobile telecommunication systems (UMTS). *IEEE Journal on Selected Areas in Communications*, 13(5), June 1995.

[11] Philip G. Hammer and F. Stuart Chapin. Human Time Allocation: A Case Study of Washington, D.C. Technical report, Center for Urban and Regional Studies, University of North Caroline, Chapel Hill, 1972.

[12] Susan Hanson and K. Patricia Burnett. Understanding complex travel behavior: Measurement issues. In Peter R. Stopher, Arnim H. Meyburg, and Werner Brog, editors, *New Horizons in Travel-Behavior Research*, chapter 11. Lexington Books, 1981.

[13] George C. Hemmens. Analysis and simulation of urban activity patterns. *Socio-Economic Planning Sciences*, 4, 1970.

[14] Bruce G. Hutchinson. *Principles of Urban Transport Systems Planning*. Scripta, 1974.

[15] European Telecommunications Standards Institute. Base Station Controller - Base Transceiver Station (BSC - BTS) Interface Layer 3 Specification. GSM 08.58, September 1994.

[16] European Telecommunications Standards Institute. Mobile Application Part (MAP) Specification. GSM 09.02, November 1994.

[17] European Telecommunications Standards Institute. Mobile Radio Interface Layer 3 Specification. GSM 04.08, November 1994.

[18] European Telecommunications Standards Institute. Mobile-services Switching Centre - Base Station System (MSC - BSS) Interface Layer 3 Specification. GSM 08.08, September 1994.

[19] Bijan Jabbari et al. Network issues for wireless communications. *IEEE Communications Magazine*, January 1995.

[20] Ravi Jain et al. A caching strategy to reduce network impacts of PCS. *IEEE Journal on Selected Areas in Communications*, 12(8), November 1994.

[21] P. M. Jones, M. C. Dix ans M. I. Clarke, and I. G. Heggie. *Understanding Travel Behaviour*. Oxford Series in Transport. Gower, 1983.

[22] Peter Jones, Frank Koppelman, and Jean-Pierre Orfeuil. Activity analysis: State-of-the-art and future directions. In Peter Jones, editor, *Developments in Dynamic and Activity-Based Approaches to Travel Analysis*. Avebury, 1990.

[23] Peter M. Jones. Destination choice and travel attributes. In David A. Hensher and Quasim Dalvi, editors, *Determinants of Travel Choice*, chapter 13. Saxon House, 1978.

[24] Peter M. Jones. Activity approaches to understanding travel behavior. In Peter R. Stopher, Arnim H. Meyburg, and Werner Brog, editors, *New Horizons in Travel-Behavior Research*, chapter 13. Lexington Books, 1981.

[25] Frank S. Koppelman. The application of disaggregate choice models to travel demand forecasting: Issues and methods. In Michael Florian, editor, *Transportation Planning Models*. North Holland, 1984.

[26] Eckhard Kutter. A model for individual travel behaviour. *Urban Studies*, 10, 1973.

[27] Eckhard Kutter. Some remarks on activity-pattern analysis in transportation planning. In Peter R. Stopher, Arnim H. Meyburg, and Werner Brog, editors, *New Horizons in Travel-Behavior Research*, chapter 12. Lexington Books, 1981.

[28] Steven R. Lerman. Recent advances in disaggregate demand modelling. In Michael Florian, editor, *Transportation Planning Models*. North Holland, 1984.

[29] U. Madhow, M.L. Honig, and K. Steiglitz. Optimization of wireless resources for personal communications mobility tracking. In *IEEE INFOCOM*, volume 2, 1994.

[30] U. Madhow, M.L. Honig, and K. Steiglitz. Optimization of wireless resources for personal communications mobility tracking. *IEEE/ACM Transactions on Networking*, 3(6), December 1995.

[31] Michel Mouly and Marie-Bernadette Pautet. *The GSM System for Mobile Communications*. published by the authors, 1992.

[32] MTA-EMCI. European Cellular Subscribers, Analogue & Digital (PCN/GSM). *Mobile Communications International*, (37), December 1996/January 1997.

[33] Rola Nevoux and Sami Tabbane. An intelligent location tracking method for personal and terminal FPLMTS/UMTS communications. In *International Switching Symposium 1995*, 1995.

[34] Walter Y. Oi and Paul W. Shuldiner. *An Analysis of Urban Travel Demands*. Northwestern University Press, 1962.

[35] Sadaatsu Okasaka et al. A new location updating method for digital cellular systems. In *IEEE 41st Vehicular Technology Conference*, 1991.

[36] Norbert Oppenheim. *Urban Travel Demand Modeling*. John Wiley and Sons, 1995.

[37] Dieter Plassmann. Location management strategies for mobile cellular networks of 3rd generation. In *IEEE 44th Vehicular Technology Conference*, 1994.

[38] Gregory Pollini and Sami Tabbane. The intelligent network signalling and switching costs of an alternate location strategy using memory. In *IEEE 43rd Vehicular Technology Conference*, 1993.

[39] MONET Project. Stage 2 (draft) specification of location management. Deliverable R2066/PTTNL/MF1/DS/P/006/b1, RACE, September 1992.

[40] MONET Project. Mobility model for UMTS (final). Deliverable R2066/SESA/GA2/DS/P/015/b2, RACE, November 1993.

[41] MONET Project. UMTS System Structure Document. Deliverable R2066/BT/PM2/DS/P/070/b2, RACE, December 1994.

[42] Ivan Seskar et al. Rate of location area updates in cellular systems. Technical Report WINLAB-TR-29, WINLAB, Rutgers University, April 1992.

[43] N. Shivakumar and J. Widom. User profile replication for faster location lookup in mobile environments. In *ACM MOBICOM*, 1995.

[44] Sami Tabbane. Comparison between the Alternative Location Strategy (AS) and the Classical Location Strategy (CS). Technical Report WINLAB-TR-37, WINLAB, Rutgers University, August 1992.

[45] Sami Tabbane. Evaluation of an alternative location strategy for future high density wireless communications systems. Technical Report WINLAB-TR-51, WINLAB, Rutgers University, January 1993.

[46] Sami Tabbane. An alternative strategy for location tracking. *IEEE Journal on Selected Areas in Communications*, 13(5), June 1995.

[47] Nigel Thrift. *An Introduction to Time Geography.* Geo Abstracts, 1977.

[48] John Zhonghe Wang. A fully distributed location registration strategy for UPCS. *IEEE Journal on Selected Areas in Communications*, August 1993.

[49] Hai Xie, Sami Tabbane, and David J. Goodman. Dynamic location area management and performance analysis. In *IEEE 43rd Vehicular Technology Conference*, 1993.