NETWORK CONFIGURATION IMPROVEMENT AND DESIGN AID USING ARTIFICIAL INTELLIGENCE

by

Sebastiaan Jan van Graan

Submitted in partial fulfilment of the requirements for the degree

Master of Engineering (Computer Engineering)

in the

Faculty of Engineering, the Built Environment and Information Technology

in the Department of Electrical, Electronic and Computer Engineering

UNIVERSITY OF PRETORIA

August 2007

Ms M Snyman
Prof. E Barnard

ACKNOWLEDGEMENTS

A special word of thanks to my supervisors, ms Magdaleen Snyman and prof. Etienne Barnard for their support and guidance throughout this research.

I would also like to thank MTN South Africa for granting me access to their systems and for supplying me with all the data required to conduct the proposed research in this dissertation.

An extra special thanks to my family, who supported and motivated me throughout the research process.

Lastly I would like to thank God for giving me the knowledge and the strength to start and complete this work.

ABSTRACT

This dissertation investigates the development of new Global system for mobile communications (GSM) improvement algorithms used to solve the nondeterministic polynomial-time hard (NP-hard) problem of assigning cells to switches. The departure of this project from previous projects is in the area of the GSM network being optimised. Most previous projects tried minimising the signalling load on the network. The main aim in this project is to reduce the operational expenditure as much as possible while still adhering to network element constraints. This is achieved by generating new network configurations with a reduced transmission cost. Since assigning cells to switches in cellular mobile networks is a NP-hard problem, exact methods cannot be used to solve it for real-size networks. In this context, heuristic approaches, evolutionary search algorithms and clustering techniques can, however, be used.

This dissertation presents a comprehensive and comparative study of the above-mentioned categories of search techniques adopted specifically for GSM network improvement. The evolutionary search technique evaluated is a genetic algorithm (GA) while the unsupervised learning technique is a Gaussian mixture model (GMM). A number of custom-developed heuristic search techniques with differing goals were also experimented with.

The implementation of these algorithms was tested in order to measure the quality of the solutions. Results obtained confirmed the ability of the search techniques to produce network configurations with a reduced operational expenditure while still adhering to network element constraints. The best results found were using the Gaussian mixture model where savings of up to 17% were achieved. The heuristic searches produced promising results in the form of the characteristics they portray, for example, load-balancing. Due to the massive problem space and a suboptimal chromosome representation, the genetic algorithm struggled to find high quality viable solutions.

The objective of reducing network cost was achieved by performing cell-to-switch optimisation taking traffic distributions, transmission costs and network element constraints into account. These criteria cannot be divorced from each other since they are all interdependent, omitting any one of them will lead to inefficient and infeasible configurations. Results obtained further indicated that the search space consists out of two components namely, traffic and transmission cost. When optimising, it is very important to consider both components simultaneously, if not, infeasible or suboptimum solutions are generated. It was also found that pre-processing has a major impact on the cluster-forming ability of the GMM. Depending on how the pre-processing technique is set up, it is possible to bias the cluster-formation process in such a way that either transmission cost savings or a reduction in inter base station controller/switching centre traffic volume is given preference.

Two of the difficult questions to answer when performing network capacity expansions are where to install the remote base station controllers (BSCs) and how to alter the existing

BSC boundaries to accommodate the new BSCs being introduced. Using the techniques developed in this dissertation, these questions can now be answered with confidence.

Keywords: GSM network optimisation/improvement; GSM network planning; GSM network modelling; transmission cost optimisation/improvement; traffic optimisation/improvement; traffic modelling; cell-to-switch association problem; Gaussian mixture models; genetic algorithms; constrained optimisation; heuristic based optimisation.

OPSOMMING

Hierdie verhandeling ondersoek die ontwikkeling van nuwe verbeteringsalgoritmes vir GSM-netwerke wat gebruik kan word om die nie-deterministiese veelterm moeilike (NP-moeilike) probleem van sel na sentrale optimering op te los. Die aspek wat hierdie projek van vorige projekte onderskei, lê op die gebied van die optimering van die GSM-netwerk. Die meeste vorige projekte het gepoog om die beheerseinlas op die netwerk te minimeer. Die hoofdoel van hierdie projek is egter om die bedryfskoste van netwerke so ver moontlik te verminder met inagneming van beperkings op die netwerkelemente. Hierdie doelwit is bereik deur nuwe netwerkkonfigurasies met verlaagde transmissiekostes te genereer. Aangesien die toekenning van selle aan sentrales in sellulêre mobiele netwerke 'n NP-moeilike probleem is, kan presiese metodes nie gebruik word om dit op regte wêreldgrootte-netwerke toe te pas nie. Heuristiese benaderings, evolusionêre soek-algoritmes en nie-toesighoudende soektegnieke kan egter in hierdie konteks gebruik word om die probleem op te los.

Hierdie verhandeling bied 'n omvattende en vergelykende studie van bogenoemde kategorieë van soektegnieke wat spesifiek vir GSM-netwerkverbetering aangepas is. Die evolusionêre soektegniek wat beoordeel word, is 'n genetiese algoritme terwyl die nie-toesighoudende soektegniek 'n Gaussian-mengselmodel is. 'n Aantal doelgemaakte heuristiese soektegnieke, elk met unieke doelwitte, is ontwikkel.

Die toepassing van dié algoritmes is getoets om die kwaliteit van die oplossings te meet. Die resultate het bevestig dat die soektegnieke wel oor die vermoë beskik om netwerkkonfigurasies teen laer koste, wat steeds aan die beperkings van die netwerkelemente voldoen, te lewer. Die beste resultate is gelewer met die gebruik van die Gaussian-mengselmodel waarmee 'n besparing van tot 17% behaal is. Belowende resultate is deur die heuristiese soektegnieke opgelewer in die vorm van die eienskappe wat dit uitbeeld. 'n Voorbeeld hiervan is las-balansering. Die reuse soekruimte, asook 'n sub-optimale chromosoomverteenwoordiging, het veroorsaak dat die genetiese algoritme moeilik oplossings van 'n bevredigende kwaliteit kon lewer.

Die doelwit van die vermindering van netwerk-bedryfskoste is bereik deur sel-na-sentrale optimering uit te voer met inagneming van verkeersdistribusies, transmissiekostes en netwerk-elementbeperkings. Hierdie kriteria kan nie van mekaar geskei word nie, aangesien hulle interafhanklik is. Die weglating van een sal tot ondoeltreffende en ongeldige oplossings lei. Die resultate van hierdie studie dui verder aan dat die soekruimte uit twee komponente bestaan, naamlik verkeer en transmissiekoste. Tydens optimering moet albei komponente gelyktydig in ag geneem word, anders word ongeldige, sub-optimale oplossings gegenereer. Die groot impak van die voorafverwerking van die datastel op die Gaussian-mengselmodel is duidelik waargeneem. Na gelang van die tipe voorafverwerking, is dit moontlik om die selgroeperingsproses op so 'n manier te beïnvloed dat óf vermindering van transmissiekoste óf inter-BSC/sentrale verkeersvolumevermindering voorkeur kan geniet.

Tydens netwerkkapasiteitsopgradering moet twee moeilike vrae beantwoord word, naamlik waar moet die BSC's geïnstalleer word en hoe die grense van die huidige BSC-areas aangepas moet word om die nuwe BSC's te akkommodeer. Deur gebruik te maak van die tegnieke wat in die verhandeling ontwikkel is, kan hierdie vrae nou met gemak beantwoord word.

Sleutelwoorde: GSM-netwerkoptimering/verbetering; GSM-netwerkbeplanning; GSM-netwerkmodellering; transmissiekoste-optimering/verbetering; verkeersoptimering/verkeersverbetering; verkeersmodellering; sel-na-sentrale assosiasieprobleem; Gaussian-mengselmodel; genetiese algoritmes; beperkte vryheidsoptimering; heuristiese optimering.

LIST OF ABBREVIATIONS

AI	artificial intelligence
AUC	authentication centre
BSC	base station controller
BSS	base station subsystem
BTS	base transceiver station
CDR	call data record
CPU	central processing unit
CSA	cell-to-switch association
DXX	digital cross connect
E1	2Mbps link divided into 32 64kbps time slots
EIR	equipment identity register
EM	expectation maximisation
GA	genetic algorithm
GMM	Gaussian mixture model
GMSC	gateway mobile services switching centre
GoS	grade of service
GSM	global system for mobile communications
HLR	home location register
IMEI	international mobile equipment identity
IMSI	international mobile subscriber identity
IVR	interactive voice response
LAPD	link access protocol D
MA	memetic algorithms
MS	mobile station
MSC	mobile services switching centre
NCR	network call reference number
NP-hard	nondeterministic polynomial-time hard
NSS	network and switching subsystem
OPEX	operational expenditure
OSS	operation and support system
PABX	private automatic branch exchange
PCM	pulse code modulation
PCN	personal communication network
PCS	personal communication system
PLMN	public land mobile network
PSTN	public switched telephone network
SA	simulated annealing
SC	switching centre
SIM	subscriber identity module
STM1	synchronous transmission module 1
TRAU	transcoder and rate adaption unit
TRC	transcoder
TRX	transceiver
VLR	visitor location register

TABLE OF CONTENTS

1 INTRODUCTION

More than a decade has passed since the initial rollout of GSM in Africa, which by now is a mature and established technology in most markets. Due to shifting traffic patterns, decisions made a few months ago regarding the network topology might not be the optimal solution any more. This leads to areas in the network being either over- or under-dimensioned; the profitability of these areas is not as high as can be. The reduction in profitability is due to a loss of income caused by congestion or due to unnecessary operational costs spent on the hiring of underutilised transmission links. By optimising/improving the network topology, not only is grade of service improved, but revenue is also saved.

Improving a GSM network with the goal of reducing cost is a difficult and time-consuming task. The shear size and complexity of today's networks discourage manual planning. Changes to the network topology have to be done based on an in-depth knowledge and understanding of the underlying traffic distributions. By only considering radio, transmission or core network parameters in isolation, non-optimal network topologies with unnecessarily high costs will be generated.

Transmission cost is one of the major expenditures in personal communication networks (PCN) such as GSM. The task of efficiently optimising network configurations of large GSM networks has consumed vast amounts of time and resources from trained and dedicated engineers over the years. The aim of this dissertation is to develop new network improvement algorithms which produce improved network configurations with a reduced cost while still adhering to network-element constraints.

When improving and re-optimising existing networks, one has the advantage of large amounts of data being available. One of the most important data sources is call data records (CDRs) which are generated by the mobile switching centres (MSCs). By evaluating the CDRs detailed traffic distributions can be generated. The problem of efficient network planning and optimisation can be solved by combining access transmission network improvement and core network improvement as well as detailed traffic distributions.

Due to constantly changing subscriber demands and network growth, cut-overs and network upgrades are a common occurrence. Network improvement should be performed in conjunction with the above-mentioned tasks to ensure the delivery of optimal/improved network configurations. An automated improvement process is recommended due to the complexity and frequent occurrence of these tasks.

The departure of this project from previous projects is in the area of the GSM network being improved. In most of the previous projects, optimisation was done to minimise the signalling load on the network. The main aim of this project is to reduce the operational expenditure as much as possible while still adhering to network element constraints.

Operational expenditure is reduced by generating new network configurations with a reduced transmission cost. New network configurations are generated by cutting sites over to new/different parents thus altering the network topology. In the literature, this problem is known as the cell-to-switch association (CSA) problem.

Assigning cells to switches in cellular mobile networks is a NP-hard problem. This problem can thus not be solved using exact methods for real-size mobile networks. In this context, heuristic approaches, evolutionary techniques and clustering techniques can, however, be used.

This dissertation presents a comprehensive and comparative study of the above-mentioned categories of search techniques adapted specifically for GSM network improvement. The evolutionary search technique evaluated was a genetic algorithm while the unsupervised learning technique used was a Gaussian mixture model. A number of custom-developed heuristic search techniques with differing goals were also experimented with.

The implementation of these algorithms was tested in order to measure the quality of the solutions. Results obtained confirmed the ability of the search techniques to produce network configurations with a reduced operational expenditure while still adhering to network element constraints. The best results found were using the Gaussian mixture model where savings of up to 17% were achieved (while keeping the MSC and MSC locations fixed). The heuristic searches produced promising results in the form of the characteristics they portray, for example, load-balancing. Due to the massive problem space, the genetic algorithm struggled to find high quality viable solutions.

New optimisation criteria were introduced in the dissertation which distinguishes it from previous works done in the field, namely.

- Cells were reassigned to switches with the goal of reducing network cost while adhering to network element constraints.
- All results presented in this dissertation were based on actual measurements taken from a GSM network with over 5 000 sites. This serves as evidence that the technique developed is capable of producing feasible, implementable results in large real-world networks.
- One of the most important requirements when addressing the CSA problem for GSM networks is cluster consistency. This requirement was not enforced by the literature surveyed, but will be a deciding factor when selecting network improvement/optimisation techniques for real-world networks.
- When performing network capacity expansions, two of the most difficult questions to answer are where to put the remote BSCs down and how to alter the existing BSC boundaries to accommodate the new BSCs being introduced. Using the techniques developed in this dissertation, new optimal BSC area boundaries can

easily be generated. The new BSC area's centre of mass should then be used as a starting point when searching for a suitable remote BSC site. Transmission cost savings of up to 36% were achieved when placing remote BSCs at the cluster's centre of mass.

Throughout the dissertation references are made to optimisation and improvement techniques. When referring to optimisation in this work it should be seen in the context of improvement rather than the theoretical optimum solution.

The dissertation is structured in six chapters. Chapter 2 provides background information on the GSM architecture, GSM principles and teletraffic theory. It further presents the current state of knowledge regarding the cell-to-switch optimisation/improvement problem. Chapter 3 describes the modelled GSM network as well as the traffic distribution creation process. Each of Chapter 4 and Chapter 5 is dedicated to a family of optimisation/improvement techniques. Every search technique is discussed and implemented with the results and discussion thereof presented at the end of each chapter. Chapter 4 explores heuristic search techniques with certain predefined goals, e.g. load-balancing. It also discusses an evolutionary search technique called a genetic algorithm. Chapter 5 is dedicated to a class of unsupervised learning techniques known as Gaussian mixture models. Finally, the dissertation is concluded in Chapter 6.

2 BACKGROUND

2.1 INTRODUCTION

The problem of assigning cells to switches is a complex and multifaceted one. It crosses into the domains of optimisation, teletraffic analysis and telecommunication network design. To fully understand and appreciate the problem, a basic knowledge of these three fields is required. Section 2.2 provides an overview of the GSM architecture and principles. Relevant teletraffic aspects are introduced in Section 2.3. Assigning cells to switches is not a new problem, a fair amount of work has already been done in the field. The current state of knowledge thereof is presented in Section 2.4.

2.2 GSM ARCHITECTURE AND PRINCIPLES

2.2.1 GSM building blocks

All the information mentioned in this section was obtained from [1]. The GSM network can be divided into four main areas as illustrated in Figure 2.1.

- The mobile station (MS)
- The base station subsystem (BSS)
- The network and switching subsystem (NSS)
- The operation and support subsystem (OSS)

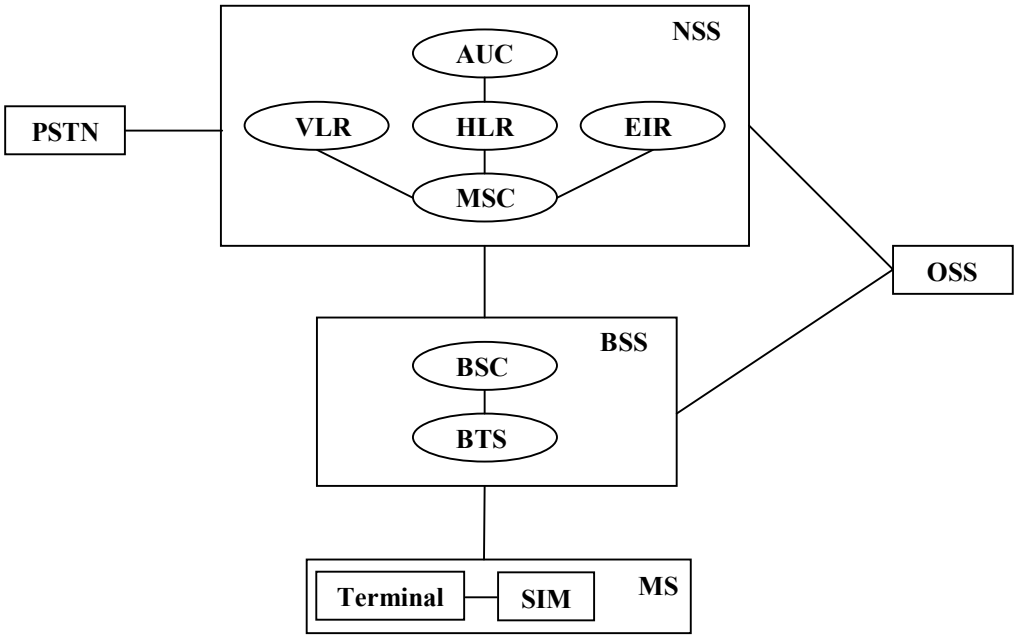

Figure 2.1 Generic GSM architecture

Each of these areas as well as all the elements contained within them is discussed next.

2.2.1.1 Mobile station (MS)

A Mobile station consists of two main elements.

- The mobile equipment or terminal
- The subscriber identity module (SIM)

2.2.1.1.1 The terminal

The terminal is the actual user-side device used for communication. This includes any type of mobile phone or GSM modem.

2.2.1.1.2 The SIM card

SIM cards are used to identify users. By inserting a SIM card into a terminal, the user gains access to all his/her subscribed services. Users can thus have access to all their subscribed services on any terminal by using their SIM card. Without a SIM card only emergency calls can be made from the terminal. SIM cards contain parameters used to identify the subscriber to the system, one such a parameter is the international mobile subscriber identity number (IMSI). Data stored on the SIM card supports the mobile mobility protocol in the network.

2.2.1.2 The base station subsystem (BSS)

The BSS connects to the mobile station and the network switching subsystem (NSS). It is in charge of radio transmission as well as reception. The BSS can be divided into two parts, namely.

- The base transceiver station (BTS) or radio base station
- The base station controller (BSC)

2.2.1.2.1 The base transceiver station (BTS)

A BTS is the hardware on which a number of cells are defined. The BTS houses all the necessary radio equipment such as transceivers and antennas needed to establish and maintain a communication channel with the mobile station. Its transmitting power defines the size of a cell. Each BTS has between one and 16 transceivers depending on user density.

2.2.1.2.2 The base station controller (BSC)

The BSC controls a group of BTS and manages their radio resources. A BSC is principally in charge of handovers, frequency hopping, exchange functions and controlling the radio frequency power levels of the BTSs.

2.2.1.3 The network and switching subsystem (NSS)

Its main role is to manage communications between mobile users and other users such as fixed-telephony users. It also includes databases for storing subscriber information and manages user mobility.

2.2.1.3.1 The mobile services switching centre (MSC)

This is the central component of the NSS. MSCs perform the switching functions of the network. They also provide connections to other networks. A number of databases are connected to the MSCs to help with locating and authenticating users on the network.

2.2.1.3.2 The gateway mobile services switching centre (GMSC)

The GMSC is the interface between the public land mobile network (PLMN) and other circuit-switched networks. It is in charge of routing calls from the interconnected networks towards GSM users.

2.2.1.3.3 Home location register (HLR)

The HLR is a very important database. It stores the current location (i.e. mobile switching centre/visitor location register (MSC/VLR) coverage area) and status of all subscribers as well as their subscribed services.

2.2.1.3.4 Visitor location register (VLR)

The VLR contains information from a subscriber's HLR. When a subscriber enters the coverage area of a new MSC, the VLR associated with the MSC requests information about the subscriber from the HLR. The VLR then has enough information to provide the user its subscribed services without enquiring the HLR each time a communication attempt is made. A VLR is always implemented together with an MSC. All the cells connected to a MSC constitute the MSC/VLR area.

2.2.1.3.5 The authentication centre (AUC)

The AUC register is used for security purposes. It provides the parameters required for subscriber authentication and encryption functions. These parameters help verify a subscriber’s identity.

2.2.1.3.6 The equipment identity register (EIR)

The EIR is also used for security purposes. It is a database containing information about mobile equipment. A terminal is uniquely identified by its international mobile equipment identity (IMEI). The EIR allows the blocking of calls from stolen or unauthorised terminals.

2.2.1.4 The operation and support system (OSS)

The OSS offers centralised support for maintenance and monitoring activities required by cellular networks. Components monitored by the OSS include MSCs, BSCs, VLRs, HLRs, EIR and the AUC.

2.2.2 Links and interfaces

Now that the major elements in a GSM network are known, it is necessary to discuss how they connect to one another. Figure 2.2 indicates the different interfaces used for interlinking nodes in a GSM network [2].

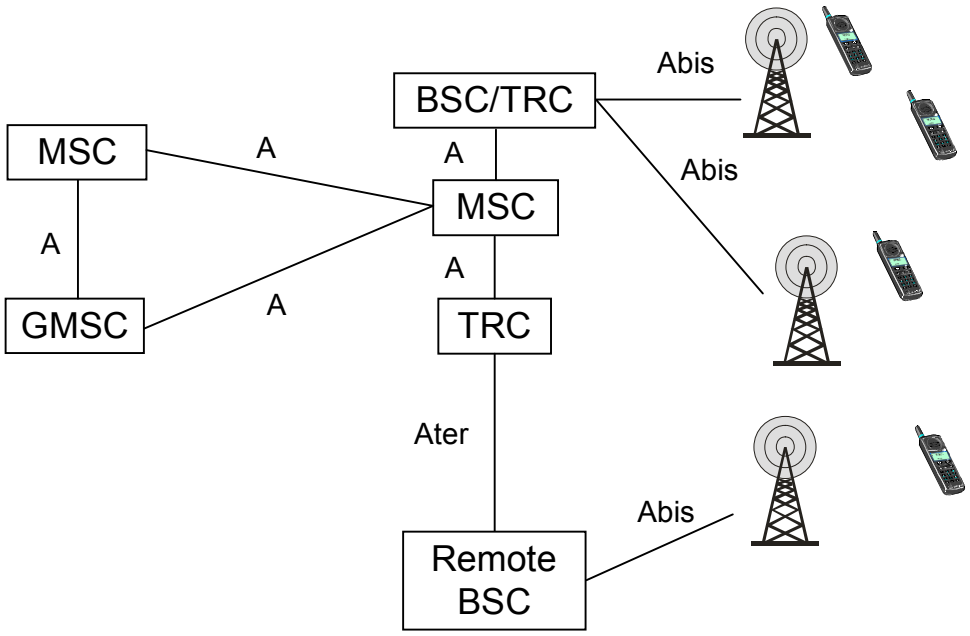

Figure 2.2 GSM interfaces

The GSM network modelled in this study uses E1 (2Mbps link divided into 32 64kbps channels) as well as 155Mbps synchronous transmission module 1 (STM1) links to connect network elements to each other. The structure of these links is discussed in the following paragraph. On top of these links different interfaces are used by network elements to communicate with each other. The interface used depends on the connected network element pair. All the main interfaces used to connect traffic carrying network elements are discussed in sections to follow.

E1 links are used to connect BTSs to BSCs as well as BSCs to each other. These 2Mbps links are divided into 32 64kbps time slots. Usually 30 of these time slots are available to carry traffic, the remaining time slots are used for signalling and synchronisation. MSCs and GMSCs are connected using high bandwidth STM1 links. These links have a bandwidth of 155Mbps (1953 64kbps time slots). Each interface listed in Figure 2.2 is discussed next.

2.2.2.1 Abis interface

The Abis interface is used between BTSs and BSCs. Each transceiver (TRX) in a BTS is capable of handling eight simultaneous connections. If no link access protocol D (LAPD) concentration is used, three 64kbps time slots per transceiver are required. Two 64kbps time slots are used to carry the user-plane traffic load of the eight time slots, at 16kbps per time slot, on the radio interface. The 16kps is reserved regardless of whether there is traffic on the particular air-interface time slot. Of the third time slot only 16kbps is used. This 16kbps is used for transceiver signalling traffic over the LAPD protocol. Without LAPD concentration, the Abis interface is able to handle 10 transceivers per E1 link.

When LAPD concentration is used, two or more transceivers share the same 64kbps time slot for signalling. A maximum of four transceivers' signalling can be concentrated together leading to 13 transceivers per E1 link. Refer to Figure 2.3 for a schematic representation [3].

2.2.2.2 Ater interface

The Ater interface is used between base station controllers / transcoders (BSC/TRCs) and remote BSCs. A remote BSC is a BSC that does not have a transcoder and rate adaption unit (TRAU). Such a BSC is only able to connect to a MSC via another BSC that does have a TRAU or a TRC. The LAPD signalling (from the BTSs) terminates at the remote BSC where the BSC performs a grooming function that multiplexes used time slots of all the incoming Abis interfaces together. The Ater interface requires only 16kbps to carry a single active voice call. This translates to a call capacity of 120 simultaneous voice calls per E1 link over the Ater interface.

One of the main reasons for using remote BSCs is the massive saving in transmission costs these BSCs introduce. Instead of having a large number of sites each working off a distant BSC/TRC, these sites now rather connect to a remote BSC located near the traffic centre of mass. The remote BSC then connects to a BSC/TRC parent using the Ater interface. Less links are required between a remote BSC and its BSC/TRC parent than the number of links terminating on the remote BSC from BTSs. This is due to the fact that no Abis signalling traffic needs to be carried over the Ater interface and because of the multiplexing ability of the remote BSC. Only timeslots carrying traffic need to be passed on to the BSC/TRC. Not all of the sites will operate at 100% capacity at exactly the same time. It is thus possible to dimension the route between a remote BSC and its BSC/TRC parent to such an extent that fewer links are required while still maintaining the required grade of service (GoS).

2.2.2.3 A interface

The A interface is used between BSC/TRCs and MSCs as well as between MSCs and GMSCs. Speech and circuit switched data are encoded at a rate of 13kbps in GSM. At each BSC/TRC, there is a TRAU. The TRAU is responsible for speech coding and data rate adaption from 16kbps (13kbps traffic and 3kbps signalling) to 64kbps. It thus converts the 16kbps data rate used in GSM to the 64kbps data rate used on pulse code modulation (PCM) links [4]. One complete 64kbps time slot is used on the A interface to carry the traffic load of a single voice call.

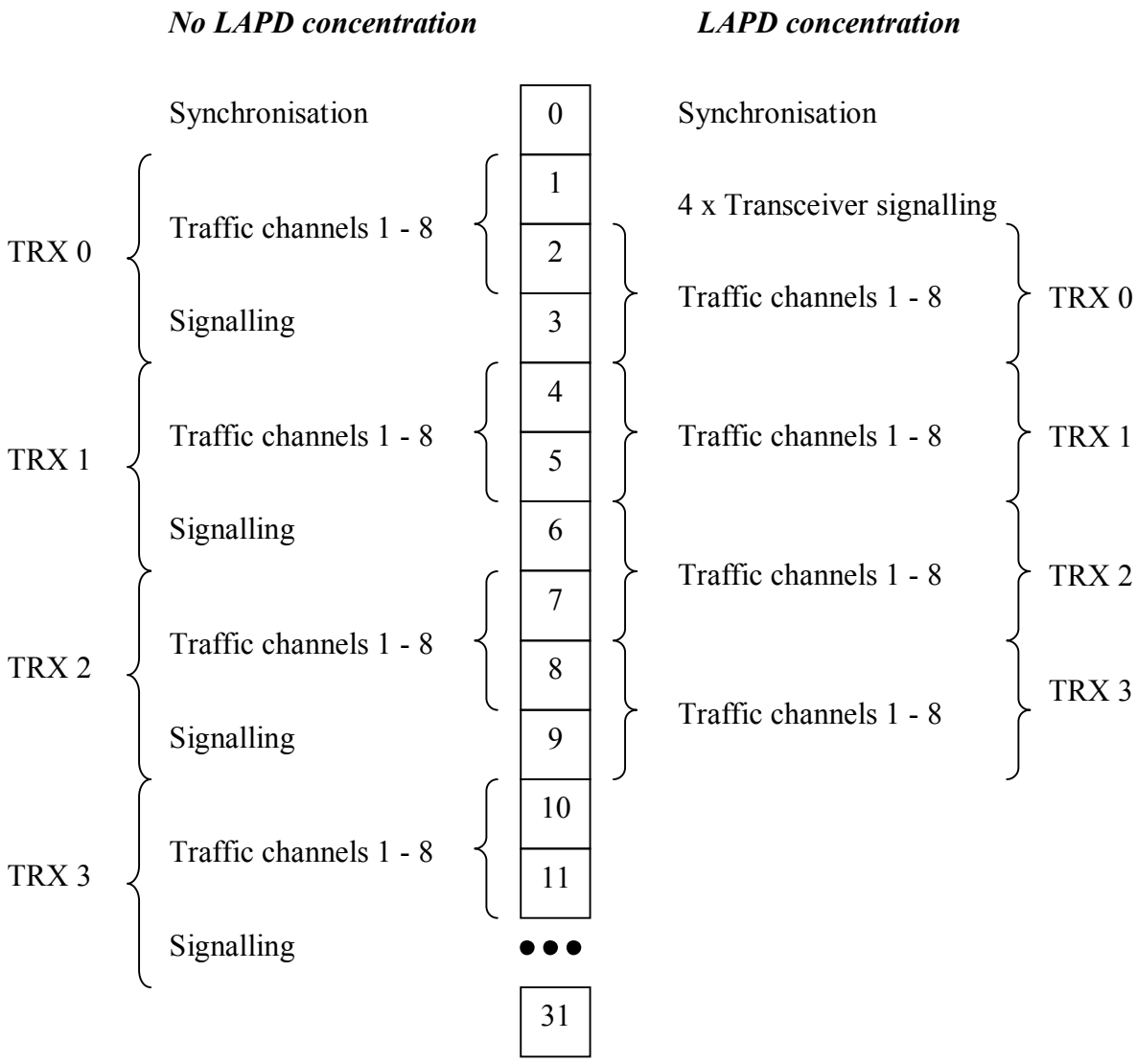

Figure 2.3 LAPD concentration compared with no LAPD concentration

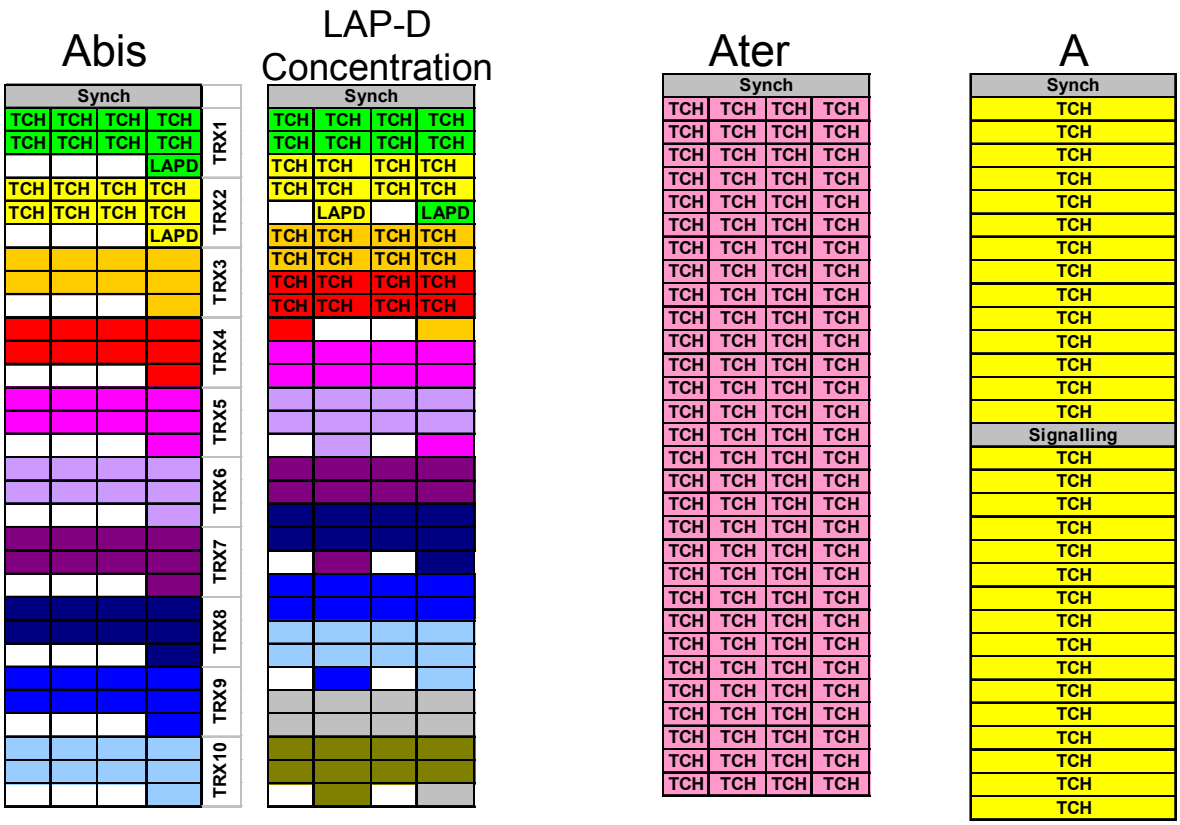

Figure 2.4 GSM interface protocol stacks

2.2.3 Link costs

Various models are used to model the cost of transmission links connecting different nodes in the network to one another. One factor that has a major impact on the link-costing model is whether the links are hired or self-owned. The network under investigation hired its entire transmission network from a public switched telephone network (PSTN). The main focus of this dissertation is thus reducing operational expenditure (OPEX) in networks that specifically hire their transmission infrastructure.

Link costs were calculated based on the PSTN's pricing rules. For STM1 as well as for E1 links, fixed rules are used to calculate the cost of a link based on link distance and availability. All links were modelled as having an availability of 99.95%. It should be noted that there is a correlation between the distance and cost of a link.

2.2.4 Switching regions

A switching region is a geographical area defined by all the cells connected to a switching centre. A switching centre (SC), in turn, is defined by all the switches located at the same premises. A GSM network can thus easily be broken down into its switching regions. It is desirable to keep as much traffic as possible local to a switching region. Keeping traffic local implies traffic originating from cells within a switching region also terminates at cells

within the same switching region. By achieving this, less OPEX is required for hiring expensive interswitching centre links used to carry traffic.
A schematic representation of how switching regions are defined is given in Figure 2.5. New switching regions are defined by changing the links between BTSs and BSCs. The existing links connecting BSCs to MSCs and MSCs to one another will remain unchanged. The reason for changing the links between the BTSs and BSCs is that this is the highest resolution at which it is financially and practically viable to make large-scale changes to the network configuration.

Figure 2.5 illustrates conceptual switching regions. Sites with a high amount of traffic between them are shaded. The solid line represents the initial switching area boundary. This boundary is not optimal because a group of sites with a large amount of traffic between them are situated in different switching regions. By moving BTS-X off BSC/TRC2 to BSC/TRC3, the inefficiency is removed. The new switching area boundary is indicated by the dashed line.

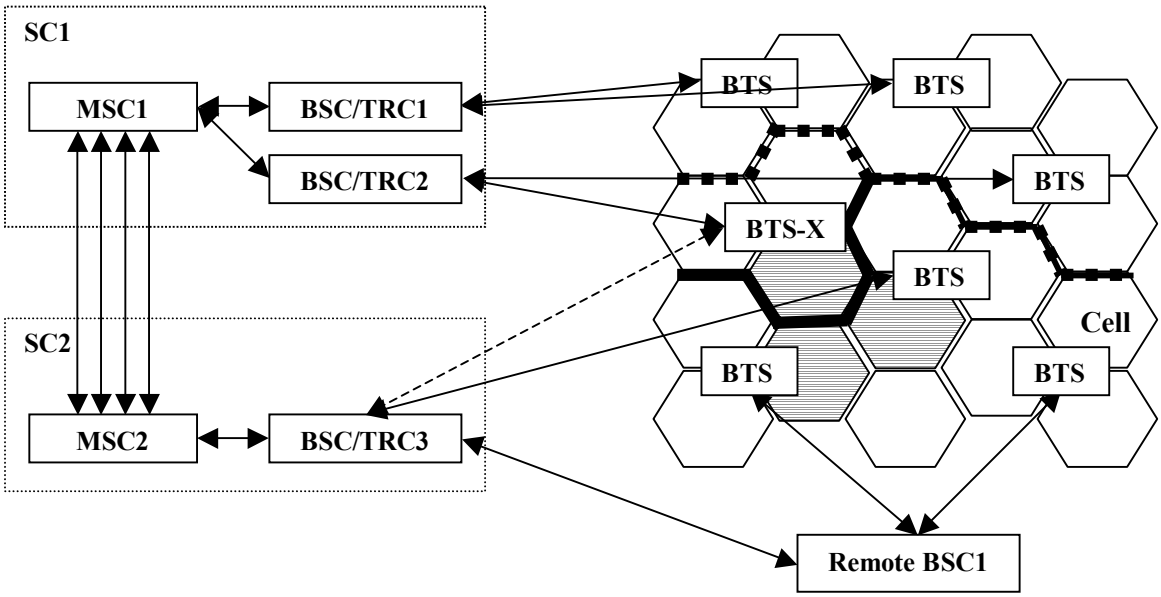

Figure 2.5 Switching regions

2.2.5 Mobility considerations in a cellular network

The fact that a mobile subscriber is free to move throughout the network while still enjoying coverage without doing anything on his or her side is a significant feature. This feature comes at a price. The price is that there is a certain amount of signalling traffic generated by the phone. This communication is crucial, since the network has to be kept informed of the current location area and status of all its subscribers. In GSM, updates are performed using two techniques.

- Location updates
- Periodic updates

The system has to know the location area of its users. Only this area is paged in the case of an incoming call. If subscribers could not be found using location areas, all the cells in the network would have to be paged to find the location of a subscriber.

There is a relationship between the frequency at which paging and location updates take place. If location updates are performed regularly, the network will have a good idea where each subscriber is. In the event of an incoming call, only a small number of cells have to be paged to determine where the subscriber is. A low location update rate, on the other hand, will mean that the network only has an approximate idea where each subscriber is. In the event of an incoming call, a large number of cells have to be paged to determine where the subscriber is. There is thus always a trade-off between the traffic generated by location updates and the traffic generated by paging requests. Two techniques are used for location updating.

- Periodic location updates
- Location updates due to location area border changes

Periodic location updates are exactly what the name implies. The mobile station is required to periodically send its identity. This happens whether or not the mobile station has changed position from the previous location update. The drawback of this technique is the amount of unnecessary traffic caused by the mobile station, especially if the frequency at which the periodic updates are sent is too high.

Location update on location area border changes, on the other hand, is only performed when the mobile station crosses location area boundaries. This requires the mobile station to listen to the current location area it is in and compare it with the last-known location area. If these two are the same, no action is taken. If they differ, a location update is performed. A stationary terminal will thus not create location updates until it starts to move. Terminals with a high mobility will constantly send location updates, specifically if the defined location areas are very small. The advantage of this technique is that unnecessary location updates are minimised.

Most of the time mobile operator companies implement a hybrid approach. A location update is performed each time the mobile station enters a new location area. In addition, if no communication between the mobile station and the network took place within a predefined amount of time, the mobile station automatically sends periodic update information.

Location updates can in turn be divided into two subclasses, namely.

- Simple location updates

- Complex location updates

Simple location area updates occur when the mobile terminal changes location areas but both location areas fall under the control of the same MSC. The strain imposed on the system due to simple location updates is minimal.

Complex location area updates occur when the mobile terminal changes between location areas that fall in different MSC/VLR areas. A significantly larger strain is placed on the system by complex location updates than by simple location updates. This is due to the more complex exchange of information between the mobile terminal and the BTS as well as the required updating of all the appropriate databases in the system.

2.2.6 Traffic case

An example of the high-level interaction between the GSM building blocks is presented next. This example shows what happens in a PLMN when a PSTN user phones a GSM user.

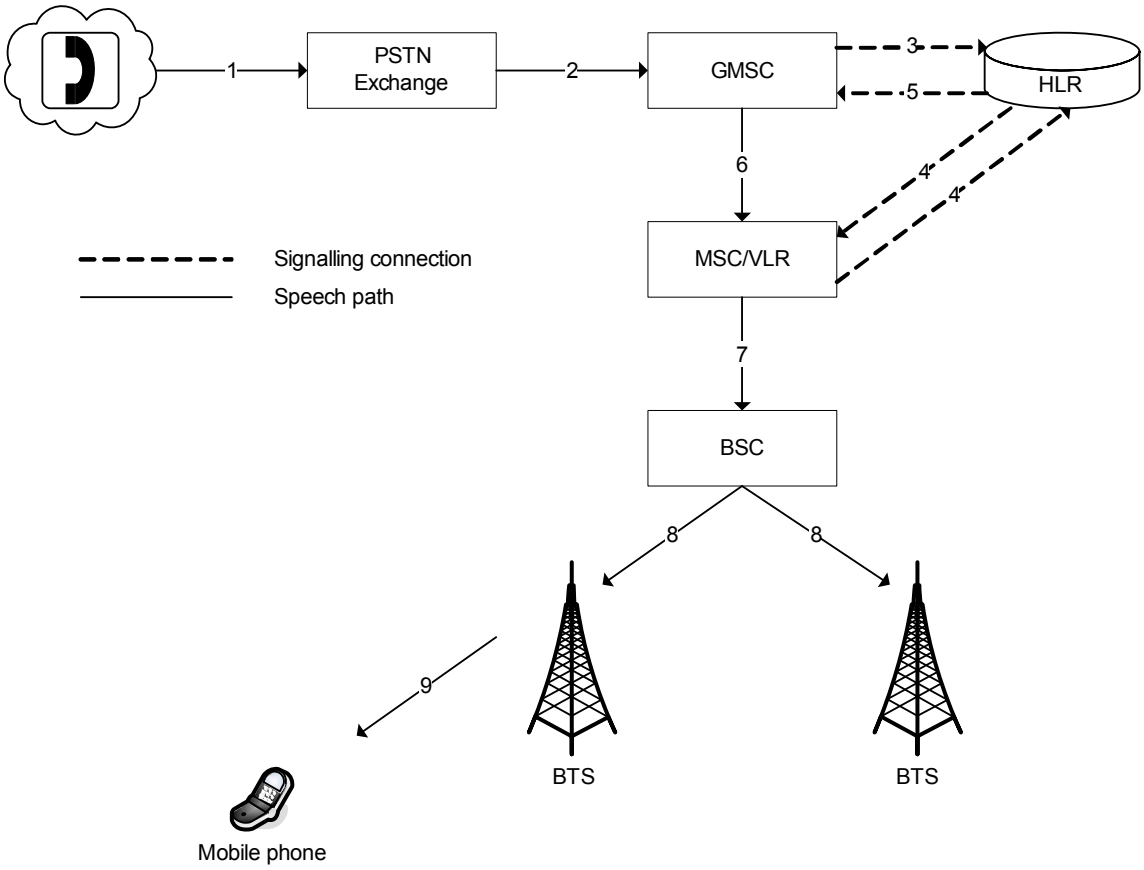

Figure 2.6 Call to MS from PSTN

1. PSTN user phones a GSM user. The PSTN exchange determines that the dialled number is destined for a PLMN.
2. The PSTN exchange routes the call to a point of interconnect with the specific PLMN.
3. The GMSC asks the HLR for information so that the call can be routed to the MSC/VLR where the MS is registered.
4. The HLR stores the address of each subscriber's current serving VLR. Using this address, the HLR contacts the appropriate VLR to obtain a roaming number. The roaming number serves as an address and is used to route the call to the appropriate MSC. The VLR sends a roaming number back to the HLR.
5. The HLR forwards the roaming number to the GMSC.
6. With the help of the roaming number, the GMSC routes the call to the appropriate MSC.
7. The MSC knows which location area the MS is located in and sends a paging message to the BSC handling the location area.
8. The BSC distributes the paging request to the BTSs in the location area.
9. The BTSs page the MS.

As soon as the MS detects its identity in the paging message, a series of request and acknowledge messages is exchanged between the MS and various network elements. When the subscriber answers the call, it causes a connect message to be sent from the MS. This message prompts the network to complete the connection path. Upon path establishment, a connection acknowledgement is sent to the MS indicating that a duplex path for traffic is open and ready for use.

2.3 TELETRAFFIC THEORY

The basic goal of traffic analysis is to determine the cost-effectiveness of different network configurations. The effectiveness of a network can be evaluated by determining how much traffic it carries during normal load conditions and how often the offered traffic exceeds the network's capacity.

Traffic analysis techniques can be divided into two main categories, namely.

- Delay systems
- Loss systems

GSM network operators model their systems using the loss model [5]. Only this model and a few important aspects around it are discussed in this section. Refer to Appendix A for an overview of teletraffic theory.

A useful measure of traffic is the traffic intensity that the network is capable of carrying. One measure of traffic intensity is called the Erlang. Traffic intensity can be calculated

assuming that the average call arrival rate and mean call holding time are known using the formula

$$A = \lambda t_m \tag{2.1}$$

where λ is the average arrival rate, t_m is the mean holding time and A is the traffic intensity in Erlang. It should be noted that the traffic intensity calculated above is the average utilisation during a period and not the maximum load experienced by the system.

In a telephone system, it is assumed that calls arrive in a random fashion. The call arrival rate is the rate at which calls arrive at the serving switch. It is always assumed that the arrival of one call has absolutely no correlation with the arrival of any other call. All call arrivals are thus assumed to arrive independently.

The exponential holding time distribution is one of the most commonly used distributions for regular telephone calls and is modelled by the negative exponential distribution. The negative exponential distribution has the property that the probability of the holding time is independent of how long the call has been in progress up to that point. Absolutely no form of history of the call is thus taken into consideration.

In a lost calls cleared system, call requests are denied during overload conditions. Losing calls is highly undesirable. Every lost call is a potential loss of revenue for the operator. Performance in a lost calls cleared system is measured in terms of the probability of rejecting a call.

Erlang's B formula is used to calculate the blocking probability [6]. This formula assumes a random arrival rate and an arbitrary holding time distribution. The blocking probability is simply the probability of an unsuccessful call attempt due to congestion. The blocking probability is based on the offered traffic load as well as the number of servers available and is given by

$$B = \frac{A^N}{N!\sum_{i=0}^{N}\left(\frac{A^i}{i!}\right)} \tag{2.2}$$

where N is the number of servers or channels and A is the offered traffic in Erlang.

The relation between traffic volume, number of devices and blocking probability as stated in Equation 2.2 requires tedious numerical computations. To ease this problem, tables were compiled. These tables are commonly referred to as the "Erlang B Tables". Two common

versions are available, the first version indicates the maximum traffic load that can be carried given the number of servers and blocking probability. The second version indicates the blocking probability given the traffic load and number of servers. Refer to Appendix A for an example of these tables.

2.4 CURRENT STATE OF KNOWLEDGE

Merchant *et al.* [7, 8] did the first work in the field of cell-to-switch assignment. They not only defined the problem, but also proposed several heuristic techniques to solve it. By publishing the techniques used for data generation, they enabled others to compare results in a subjective and scientific manner. To the present day, newly developed techniques are compared with those original heuristics proposed in [7].

Bhattacharjee *et al.* [9, 10] followed in the footsteps of Merchant and Sengupta. They published articles comparing the original heuristics by Merchant and Sengupta with their own heuristics. Bhattacharjee, Saha and Mukherjee, were also the first to introduce the idea of adding balanced load-sharing between the MSCs as a constraint.

Pierre *et al.* [11, 12, 13, 14, 15, 16, 17] and Din *et al.* [18] made their contribution to the field by using artificial intelligence search techniques. This is an important contribution since it broke away from the heuristic solutions tried by others. By implementing artificial intelligence (AI) techniques, access to a variety of global optimisation algorithms was gained.

Of all the articles encountered, the article by Demirkol *et al.* [19] is the closest to the problem addressed in this project. The major similarities are.

- The use of captured measurements from a GSM operator
- The use of an artificial intelligence search technique
- The level of detail at which the GSM network is modelled

2.4.1 Overview of the current body of knowledge

In the literature, the problem of assigning cells to switches in PCS networks is known as the cell-to-switch assignment problem. The following optimisation goals were identified from the literature studied.

- Designing a location area layout in such a manner that the signalling load on the air interface due to paging and location updates is minimised [7, 11, 12, 21, 25, 26, 27, 41, 49]
- Designing a location area layout in such a manner that uniform load-balancing is achieved between the MSCs [9, 10]
- Designing a location area layout in such a manner that the cabling cost between the cells and switches is minimised [7, 11, 12, 25, 41]

- Designing a location area layout in such a manner that the switch's call-handling capacity is not exceeded [12, 19, 27]

All of the papers oversimplified the cabling cost. In determining the cable cost, it was assumed that a cell would connect directly to one of the available switches. The network configurations used were also always oversimplified. In most of the cases, the entire network existed of switches and cells only. The exception is that Demirkol *et al.* [19] modelled the network to include BSCs. [20, 21, 22] modelled the network by only considering the cells and their interaction with one another.

These simplifications are justified by the fact that the problems were defined as performing optimisation on personal communication system (PCS) networks. PCS networks include any form of network that fits into the personal communication system category. GSM is only one of a large number of technologies that fit into this category. Different systems will have different architectures but the fundamental problem stays the same. Simplifying to a level such that the solution applies to any PCS network can be justified in this sense.

Looking at the body of knowledge, the following four aspects kept on reappearing in all of the optimisation problems.

- Determining the flow of traffic in a system [8, 19, 21]
- Calculating the cost associated with a network configuration [7, 10, 11, 12, 19, 21, 25, 26, 27, 41, 49]
- Generating new network configurations [19, 22, 23, 24, 25]
- Adhering to network element constraints [12, 19, 27]

Each of the above-mentioned topics will be discussed next in greater detail.

2.4.2 Determining the flow of traffic in a system

Depending on the type of optimisation being performed, *the flow of traffic* will have different interpretations. The majority of papers read were concerned with the minimisation of the signalling load on the air interface due to paging and location updates. "*The flow of traffic*" in these systems was mostly concerned with the mobility patterns of mobile subscribers. The main aspects here were the mobility rate and mobility pattern of subscribers as well as their density.

Different approaches were used to obtain mobility data. Mobility data is required to determine the interaction between adjacent cells in a network. Using this data, it is thus possible to model the handovers that occur between the cells in a network. The approaches used to obtain mobility data are discussed next.

- Demirkol *et al.* [19] used handover statistics from a network operator in Turkey. It was assumed that the mobility pattern of users not engaged in conversations on their mobile phones would follow the same trend as that of the handover statistics. Handover statistics were used because it is the only recorded measurement an operator has regarding the movement of its subscribers. Handover statistics are collected when a mobile engaged in a voice conversation crosses a cell area boundary. In idle mode, no records are kept regarding cell area boundary crossings.
- Cayirci *et al.* [21] used geographical information to predict the traffic load between cells. They assumed that all cells are hexagons of equal size. A grid of cells was placed over a digitised map. Roads on the map were used as geographical features from which to derive inter-cell handovers. Each type of road was classified according to its anticipated traffic density. The number of predicted handovers between two cells was computed as being a function of the number of roads as well as the type of road between the two cells.
- Other researchers generated all the data needed for the problem. This included the initial network configuration as well as the traffic load and mobility pattern of the subscribers. The first set of data generated was by Merchant *et al.* [8]. A number of researchers followed using the same technique to create their test networks. This enabled them to compare their findings with the original work done by Merchant *et al.* [8].

These methods, used to obtain data, all mainly dealt with the mobility of users in the network while their handsets are in idle mode. One of the main departures of this project in comparison with the previous works is in the measure of traffic used.

The main source of data used was call data records (CDRs). Each time a call is made CDRs are created at all the switches the call was routed through. From the CDRs, it is possible to determine the cell in which the subscriber was when the call was initiated. It is thus possible to model the entire path of each call through the network. Cells with a strong connectivity based on actual voice traffic can easily be identified.

The reason for using CDR data instead of the mobility patterns is that the amount of traffic generated by voice calls is exponentially more than the amount of traffic generated by signalling. Voice traffic currently accounts for more than 80% of the total traffic carried on the network. Optimising the transmission network based on voice traffic will yield much larger savings than optimising the transmission network based on signalling traffic.

2.4.3 Calculating the cost of a network configuration

Calculating the cost of a network configuration is one of the most important aspects of the problem at hand. The cost of a network configuration is used to compare its performance with other possible network configurations. Setting up a costing function that accurately models the network's performance is thus crucial.

Depending on the aspect of the network being optimised, different cost measures were used. Below is a list of the most popular costing considerations encountered, namely.

- Cost of location updates
- Cost of paging
- Cabling cost

Aspects the costing function should contain greatly depends on the following.

- The detail at which the network is modelled
- The search algorithm employed
- Optimisation criteria

The detail at which the network is modelled may include aspects such as whether it is assumed that cells connect directly to switches or whether a model more closely representing the actual architecture is used. Each network component can in turn also be modelled. Demirkol *et al.* [19] modelled the MSCs to have a finite central processor unit (CPU) capacity and the BSCs to have a finite transceiver (TRX) capacity. Most other researchers only modelled the CPU capacity of the MSCs.

Depending on the search algorithm, penalty terms may or may not be applicable. In most situations where heuristics were used to produce new network configurations, a penalty term was not necessary. This is because heuristic methods use fixed rules to determine new configurations. These rules are often chosen in such a way that component constraints are not violated. It is, however, possible to write heuristics that can generate infeasible network configurations. In these situations, penalty terms are applicable. Global optimisation techniques using artificial intelligence such as genetic algorithms and simulated annealing introduce a certain amount of randomness into the search. It is quite possible for these search techniques to create infeasible network configurations. A network configuration will typically be deemed infeasible because of one or more violated constraints. Such configurations are then penalised. This serves the purpose of identifying suboptimal configurations.

Optimisation criteria have a large impact on the network cost. In [10], network optimisation was performed to ensure maximum network scalability. This was achieved by assigning cells to switches in such a manner that all the MSCs carried nearly identical

traffic loads. The cost of the network rose by between 1% – 14% compared with the same network being optimised but without load-balancing.

2.4.4 Generating new network configurations

The cell-to-switch assignment problem is known to be NP-hard [7]. The selection of reference material chosen was greatly based on the variety of techniques used to solve this problem. Popular techniques used to solve the problem by other researchers are.

- Linear programming
- Graph partitioning and clustering
- Genetic algorithms
- Simulated annealing
- Heuristics
- Greedy searches
- Tabu search
- Scatter search
- Memetic algorithms
- Ant colony optimisation

The problem size (search space) plays a very important role in the selection of an appropriate search technique. For small systems up to 35 cells, linear programming can be used to solve the problem. These solutions are guaranteed to be optimal. Unfortunately, real-world systems seldom consist of merely 35 cells. The GSM network modelled in this thesis consisted of more than 5 000 sites. The most popular search techniques used will be discussed next.

Genetic algorithms (GAs) were used by [22, 23, 24]. Genetic algorithms are efficient global optimisation search techniques. They tend not to get stuck in suboptimal solutions as often as most heuristics do. They are ideal to use in situations where the search space is extremely large. Populations of individuals are created. Each individual represents a possible network configuration. Through the genetic operators of mutation and crossover, offspring is created from selected parents. The newly created offspring explores the search space. Single as well as multi-objective GAs were implemented. Multi-objective GAs have the advantage that different design options can be played off against each other.

Simulated annealing (SA) was used in [19, 25]. This global optimisation technique is ideal to use in situations where the initial configuration is known to be of high quality. Initially, the probability of selecting a solution worse than the current one is relatively high. As time goes by, this probability, also known as the temperature, decreases. A cooling schedule is responsible for defining the rate at which the temperature decreases. The ability of SA to effectively explore the search space is greatly influenced by the temperature variable. The one drawback that SA has is its speed. SA takes a long time to reach a good solution. This

problem was addressed by [25]. Menon *et al.* [25] introduced a heuristic in their implementation of SA. The heuristic was a pricing mechanism that compared possible network solutions with each other. The heuristic chose the best (based on network price) of the possible solutions and presented it to the SA as the next possible move. This approach helped guide the SA algorithm in a direction of decreasing cost which led to reduced simulation times.

A large number of different heuristics have been developed in [7, 9, 10, 19, 20, 21, 22, 25, 26, 27, 49]. Mixed results were obtained. A problem with heuristics is their inability to escape local minima. Bhattacharjee *et al.* [10] reported that when introducing the load-balancing constraint, the network cost rose above that which was achieved when ignoring load-balancing. Demirkol *et al.* [19] performed a comparison between two heuristics and SA. SA constantly outperformed the heuristics.

Greedy searches can be seen as a form of local optimisation. Demirkol *et al.* [19] implemented a greedy search technique by setting the temperature variable in simulated annealing to zero. This has the consequence that only solutions that improve on the current solution are chosen. The starting point of the greedy search algorithm will greatly influence the quality of the obtained solution.

Memetic algorithms (MAs) are population-based heuristic search approaches, which can be used to solve combinatorial optimisation problems based on cultural evolution [11]. The memetic algorithm implemented in [11] made use of global as well as local optimisation. The local optimisation was done via Tabu search while the global optimisation was performed using a GA. The MA works by creating an initial population, each individual in the population then explores its neighbourhood using Tabu search. Once a certain level of development has been achieved, the individuals interact with each other via the GA. Parents are selected and crossover is performed to produce new offspring. Each child is sent on a Tabu search trip in order to find the local minima in its surroundings. Mutation is performed probabilistically on all the individuals. If mutation does occur, the individual first has to perform a Tabu search before being added back into the population. The best individuals of the population are selected and the process continuous. Results obtained were compared with results from [7, 12]. Improvements of 1.9% - 2.3% were achieved.

2.4.5 Adhering to network element constraints

It is of vital importance to introduce network element constraints at each level of the network hierarchy being optimised. The reason for introducing constraints is to identify infeasible solutions. Consider the following example based on the CSA problem. If the switch call-handling constraint was removed, it is quite possible that the trivial infeasible solution of connecting all the cells to a single switch would be obtained. Demirkol *et al.* [19] modelled the BSCs and MSCs in the GSM network. For each equipment type, a certain set of constraints, which had to be adhered to, was defined.

When a constraint is broken, one of three possible actions can be taken, namely.

- The proposed solution can be corrected. This is done by a repair algorithm. If the solution is found to be irreparable, it can either be discarded or penalised.
- A penalty term may be added to the cost of the configuration. The configuration is thus left unchanged.
- The configuration can be discarded and a new configuration generated.

The following constraints were modelled in the literature studied.

- Switch level: Call handling capacity and load balancing
- BSC level: Call handling capacity, TRX capacity and paging capacity

3 IMPLEMENTED GSM MODEL

3.1 INTRODUCTION

Chapter 3 describes the modelled GSM network as well as the process developed for compiling traffic distributions. The network model as well as the traffic distribution program was implemented using the Java programming language [28]. The chapter is concluded with a section describing the interpretation of the experimental results.

3.2 TRAFFIC DISTRIBUTIONS

One of the most significant departures of this project, in contrast to others, is in the use of traffic distributions, as discussed in detail in section 2.4.

Call data records (CDRs) were used as the source from which traffic distributions were generated from. Each time a call is made CDRs are created at all the switches the call was routed through. From the CDRs, it is possible to determine the cell in which a subscriber was when initiating a call. It is thus possible to model the entire path of each call through the network.

Figure 3.1 illustrates the typical path of a call through a GSM network. A subscriber in cell A phoned a subscriber in cell T, the subscriber in cell T successfully answered the call. The path of the call through the network is illustrated by the dashed line. Solid lines are used to indicate other routes in the network. CDRs are generated by MSCs as well as GMSCs. In this example, the following nodes would thus generate CDRs: MSC2, GMSC1, GMSC2 and MSC4.

Determining the path of a call using CDRs requires the identification of all the CDRs generated by the call. This is achieved using the network call reference number (NCR). The NCR is unique per call, all the CDRs forming part of a call will have the same NCR. Now that all the CDRs defining a call are identified they need to be arranged in the correct order. This is achieved by firstly identifying the originating as well as terminating CDRs of the call. From the originating and terminating CDRs, the MSC/GMSCs at which the call originated and terminated can be identified. Once this has been done, the remaining CDRs are used to find the path through the network from the originating MSC/GMSC to the terminating MSC/GMSC. This is achieved using the incoming route, outgoing route and exchange identification fields of each CDR.

The example below describes the on-net mobile originating to mobile terminating call scenario. The originating CDR will be a mobile originating CDR (created on MSC2) and the terminating CDR will be a mobile terminating CDR (generated on MSC4). It is possible for a MSC/GMSC to generate multiple CDRs per call. One of these CDRs will indicate the next hop of the call. By looking at the outgoing route field of the CDRs generated at the originating MSC/GMSC, the next hop in the call path can be identified. Next the CDRs generated at the “next hop” node specified in the previous step need to be

evaluated. This process continues until the CDRs generated by the terminating MSC/GMSC are found. The complete path of the call can now be traced through the network.

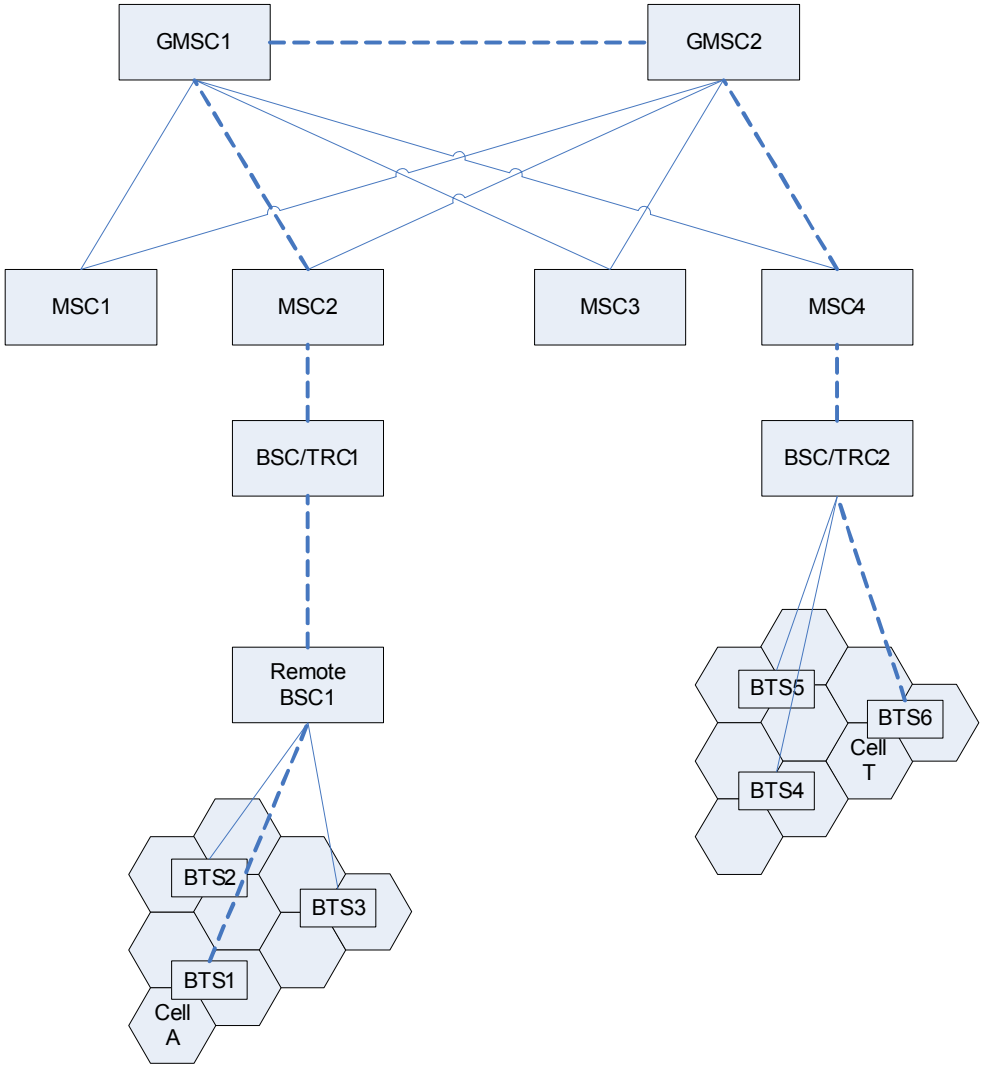

Figure 3.1 CDR generation

A conceptual example of the matched CDRs for the call scenario of Figure 3.1 is presented in Table 3.1. The path of the call through the network is found by looking at the exchange identification field of the matched CDRs in order, which in this case is MSC2 → GMSC1 → GMSC2 → MSC4.

Hop number	Exchange ID	Incoming route	Outgoing route	Cell ID	Call duration (sec)
1	MSC2	Remote BSC1	GMSC1	Cell A	96
2	GMSC1	MSC2	GMSC2		96
3	GMSC2	GMSC1	MSC4		96
4	MSC4	GMSC2	BSC/TRC2	Cell T	96

Table 3.1 Conceptually matched CDRs

The main traffic type, of interest in this project, is on-net mobile originating to mobile terminating traffic. A site-level traffic distribution is required for this traffic type. It was created by summarising a set of matched CDRs on a call-by-call basis. The summary is in the form of a traffic demand matrix as illustrated in Table 3.2. Only the originating and terminating entries per matched call are used to update the traffic demand matrix. It should be noted that a cell's site ID is derived directly from its cell ID.

Originating location	Terminating location			
	BTS 1	BTS 2	BTS 3	BTS 6
BTS 1	-	-	-	96
BTS 2	-	-	-	-
BTS 3	-	-	-	-
BTS 6	-	-	-	-

Table 3.2 Site level traffic distribution

In Table 3.2, rows are used to indicate the originating location while columns are used to indicate the terminating location. From the table, it can thus be seen that 96 seconds worth of traffic originated at BTS 1 and terminated at BTS 6.

3.3 NETWORK TOPOLOGY

Figure 3.2 illustrates the modelled GSM network. The network was modelled at different layers, these layers can be divided into two classes. The first class is used to indicate traffic statistics, while the second class is used to indicate transmission costs. Each of these classes as well as the units that make them up is discussed next.

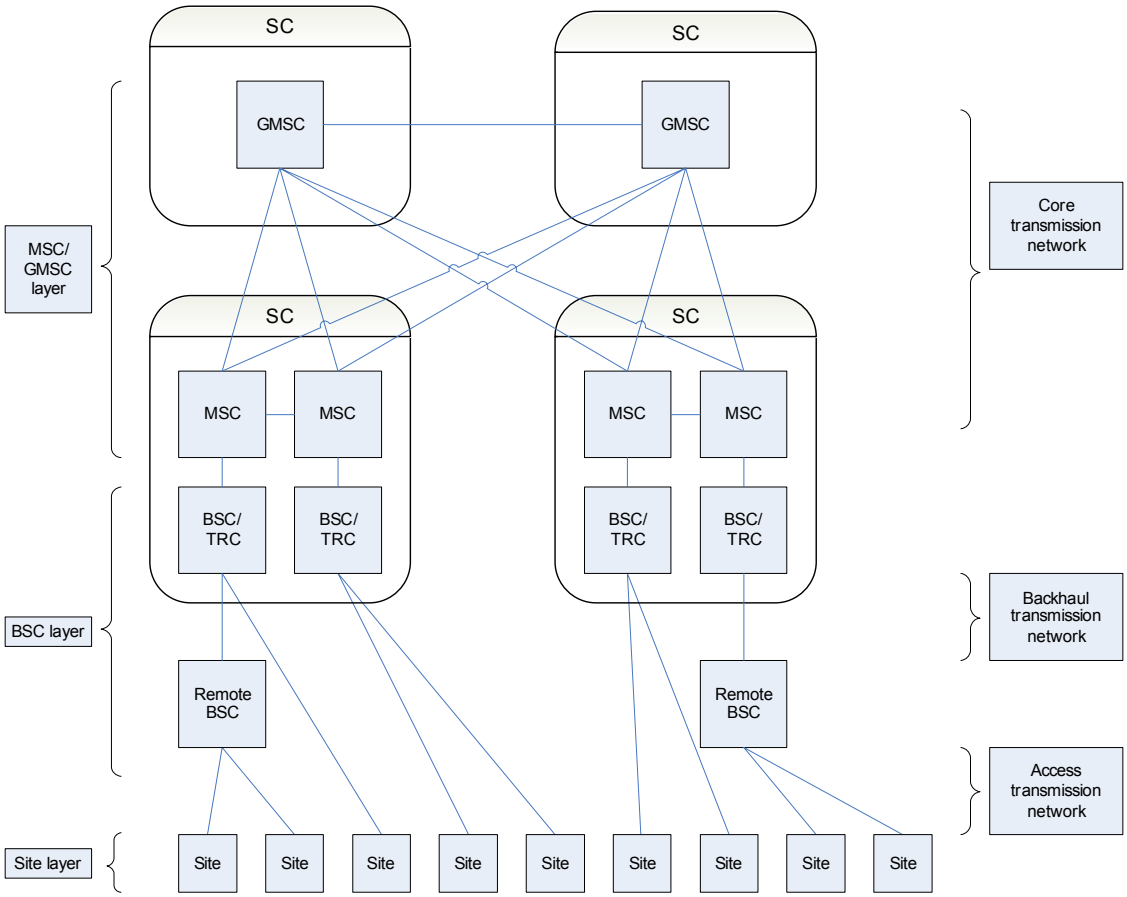

Figure 3.2 Modelled GSM network

3.3.1 Network traffic class

The network traffic class is used to measure traffic statistics at the site, BSC, MSC and SC levels. Traffic at each level is represented by a traffic matrix indicating the amount of traffic between originating and terminating pairs. A network with 5 000 sites will thus have a 5 000 by 5 000 site-level traffic matrix where rows indicate the originating site and columns the terminating site. Refer to Table 3.3 for an example.

Originating location	Terminating location			
	Site A	Site B	Site C	Site D
Site A	5 000	200	600	1 000
Site B	800	3 200	500	200
Site C	200	500	4 000	1 000
Site D	600	350	1 200	2 800

Table 3.3 Example of an intersite traffic distribution

Looking at Table 3.3 it can be seen that 5 000 seconds worth of traffic originated and terminated at site A while 200 seconds originated at site A and terminated at site B. These detailed site-level traffic statistics were calculated using raw CDRs. Site-level traffic distributions are fixed and cannot be altered.

BSC, MSC and SC traffic distributions are all derived from the site-level traffic distribution. From Figure 3.2, the following relationship for every site in the network can be derived.

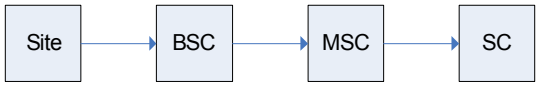

Figure 3.3 GSM element relationships

Using these relationships, the site-level traffic distribution can easily be summarised to BSC, MSC and SC level. It should be noted that these traffic distributions indicate the traffic demand between nodes and should be treated as traffic demand matrixes. The influence of moving a site to another BSC can thus easily be determined by summarising the site-level traffic distributions using the new network layout and comparing these results with the traffic distributions before the move.

The discussion above applies to the on-net mobile originating to mobile terminating call scenario. On-net mobile originating to mobile terminating traffic only accounts for a portion of the total traffic load carried in the core network. Even though this is a significant portion, it was felt necessary to model the traffic on the core as accurately as possible. More than 90% of the total traffic load on the core was modelled. This was achieved by modelling the traffic load generated at five traffic sources and eight traffic termination points. In total, 40 call scenarios were thus modelled. The five traffic sources were the operator's own BSCs as well as the four interconnect parties. The eight traffic termination points were the same entities mentioned under traffic sources as well as three on-net systems. These systems were the voicemail platform, interactive voice response systems (IVRs) and private automatic branch exchanges (PABXs). The CDR matching process used in these call scenarios is exactly the same as that used to match on-net mobile originating to mobile terminating calls. The only difference is that different CDRs are used as the originating and terminating points.

3.3.2 Transmission cost class

This class is used to calculate the transmission cost at different layers of the network. The transmission cost class can be divided into three layers, namely access, backhaul and core transmission. Each layer is discussed in greater detail below.

3.3.2.1 Access transmission cost

The access transmission cost represents the cost of all the links required to connect the sites to their BSC parents. The Abis interface is used over E1 links to connect sites to BSCs. The number of links required between a site and its parent BSC depends on the number of transceivers (TRXs) at the site. A maximum of 10 TRXs per E1 link are supported without LAPD concentration. With LAPD concentration, a maximum of 13 TRXs are supported per E1 link. LAPD concentration was not used in the model.

3.3.2.2 Backhaul transmission cost

Backhaul transmission cost represents the cost of all the links required to connect BSCs to their MSC parents. E1 links are used to connect a BSC to its parent. The Ater interface is used on links between remote BSCs and BSC/TRCs, while the A interface is used on links between BSC/TRCs and their MSC parents. An E1 link has 32 64kbps time slots, 30 of these time slots are used to carry traffic. On the A interface, a complete 64kbps time slot is required to carry the traffic of a single voice call. Thirty simultaneous voice calls can thus be carried per E1 link over the A interface (30 servers available). The Ater interface, on the other hand, only requires 16kbps to carry a voice call. A total of 120 simultaneous voice calls can be carried per E1 link over the Ater interface (120 servers available).

The traffic load (in Erlang) between a BSC and its parent is calculated by summing the traffic load of all the BSC's child nodes (sites). Knowing the traffic load and the GoS (0.5%), the Erlang B table is used to determine the number of servers necessary to carry the specified traffic load at the desired GoS. Knowing all of this, the number of links is calculated by dividing the number of servers required by the number of servers supported per E1 link for the interface used (A or Ater).

It should be noted that the modelled network was designed with the philosophy of always co-locating BSC/TRCs with their MSC parents. No E1 links are thus required between a BSC/TRC and its MSC parent, for they are normally in the same switching centre. Tie-cables are used to connect co-located nodes to each other, there is no cost associated with tie-tables. Remote BSCs are never co-located with their BSC/TRC parents, there is always a transmission cost associated with them.

3.3.2.3 Core transmission cost

Modelling traffic on the core network is not a trivial task. Two types of traffic are present in today's core networks, namely circuit switched and packet switched traffic. Circuit switched voice traffic still accounts for the bulk of traffic carried on the core and is the only type considered in this dissertation.

MSCs and GMSCs are at the heart of the core network. These nodes are responsible for the routing of traffic. None of these nodes are fully interconnected, routing cases are thus required to efficiently guide traffic from the originating switch to the terminating switch. The core transmission network can be seen as consisting of predefined routes along which all the traffic is routed. These routes are defined by routing cases and governed by the physical connections between nodes.

Before the cost of the core transmission network is calculated, it is necessary to determine the amount of traffic on each route. Once the traffic load is known, links can be assigned to the routes. The number of links assigned to a route is always the minimum number of links

capable of carrying the traffic load at the required GoS. Refer to the example below for a more detailed explanation.

Table 3.4 indicates a MSC level traffic distribution. For the purpose of this example, only the routes used to transport traffic from MSC A to MSC C is of concern. Looking at the table, one can see that 6 000 seconds needs to be routed from MSC A to MSC C.

Originating location	Terminating location			
	MSC A	MSC B	MSC C	MSC D
MSC A	50 000	2 000	6 000	10 000
MSC B	8 000	32 000	5 000	2 000
MSC C	2 000	5 000	40 000	10 000
MSC D	6 000	3 500	12 000	28 000

Table 3.4 Example of an inter-MSC traffic distribution

Node	Destination node	Next hop	Probability (Percent)
MSC A	MSC C	GMSC G	50
MSC A	MSC C	GMSC H	50
GMSC G	MSC C	MSC C	100
GMSC H	MSC C	MSC C	100

Table 3.5 Example of routing cases

Figure 3.4 illustrates the installed links between the nodes as well as the actual routes taken by calls from MSC A to MSC C. From Table 3.4 and Table 3.5, it can be seen that 50% of the calls follow the MSC A → GMSC G → MSC C (dashed) route while the other 50% of the calls follow the MSC A → GMSC H → MSC C (dotted) route.

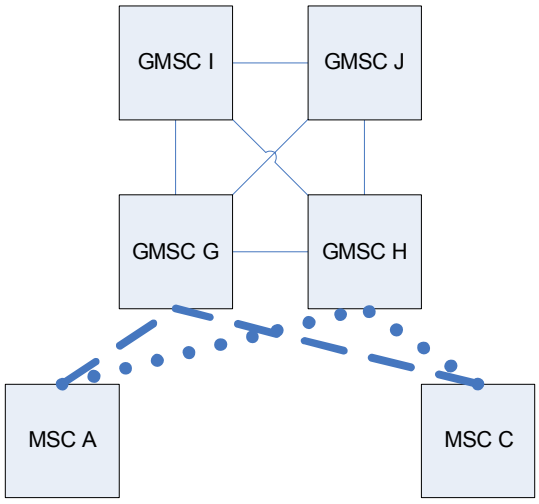

Figure 3.4 Routes between MSC A and MSC C

Once the amount of traffic on all the routes is known, it is possible to determine the total amount of traffic between any two directly connected nodes. The number of links between any two nodes is calculated as follows.

The first step is to calculate the total amount of traffic between the nodes. Using the Erlang B table, the number of servers required to carry the traffic load is then easily determined. Finally, the number of STM1 links necessary to provide the number of required servers is calculated.

Finally, all that is left is the task of calculating the transmission cost. The transmission cost per link is calculated based on its capacity and distance using the pricing rules. Once the cost of all the core transmission links is known, it is added together to form the total core transmission cost.

3.4 DESIGN SIMPLIFICATIONS AND ASSUMPTIONS

The following simplifications and assumptions were made.

- A GoS of 0.5% was used throughout the project.
- All signalling was ignored, because it only makes up a small percentage of the total traffic load.
- Due to the fact that signalling and packet switched data were ignored, only units processing circuit switched voice traffic were modelled.
- Inter-BSC as well as inter-MSC handovers were not modelled.

Sites were modelled as if always connecting directly to their BSC parents. In reality, the backhaul transmission layer consists of a digital cross connect (DXX) network. A number of sites will connect to a DXX and the DXX then connects to a BSC. The purpose of the DXX network is three-fold. Firstly, it introduces a timeslot grooming functionality which translates to a transmission cost saving. Secondly, it introduces redundancy. It is possible to define backup routes when using DXXs. If the primary route goes down, the DXX is capable of switching over to a backup route. The third major advantage of DXXs is that they ease network administration due to their remote monitoring functionality. The DXX network is an extremely complex and difficult network to model and was deemed outside the scope of this project.

The last design consideration to take into account is the backup/overflow routes defined between the MSCs and GMSCs. All the routes on the core network have a backup/overflow route associated with them. The purpose of these secondary routes is to provide a level of redundancy. If the primary route goes down, traffic is switched over to a secondary route. The secondary route is also used to carry overflow traffic in the event of congestion on the primary route. The backup routes were not considered part of the optimisation process and were thus not modelled.

Inter-BSC as well as inter-MSC handovers were not modelled in this project. The main reason for this is that only the originating and terminating locations of calls are captured in the traffic demand matrixes. It is thus not possible to derive the handover traffic from the

traffic matrixes. Handovers impact more than just the signalling traffic, it impacts the user-plane traffic as well. When a mobile terminal is handed over between MSCs during a phone call, the call is still routed through the original MSC the subscriber was on. It is common for calls that were handed over to have long inefficient routes that span over multiple MSCs and GMSCs. The reason for this inefficiency is due to limitations on the billing system. On a network-wide level, the extra traffic introduced by handovers is not significant. There will, however, be certain hotspot areas in the network where the extra user-plane traffic caused by handovers will introduce congestion if not specifically catered for.

3.5 GENERAL DESIGN CONSIDERATIONS

Before the results are presented, it is important to discuss three general aspects regarding the problem at hand.

Firstly, the network selected for this study is a mature GSM network that has been in existence for nearly 13 years. The network was selected specifically for this reason. GSM today is a mature technology that has been in existence for over a decade. There are a large number of GSM networks out there that has been in existence for a number of years. A large portion, if not most, of these networks are in a situation similar to the target network of this study where optimisation of the current network topology (switching area boundaries) will yield vast savings. All these networks thus stand to benefit from the findings made in this project.

Secondly, one should keep in mind that a large amount of time and effort has already been spent by the network operator optimising the network. Using the current network configuration thus provides a good starting point. In the light of this, using hill-climbing and heuristic algorithms to perform local optimisation is justified.

A third aspect that needs to be discussed is the concept of cluster consistency. Refer to Figure 3.5 for an example. The three clusters in the left image are consistent. This means that they have clear-cut boundaries separating them from each other. The boundaries in the image on the right-hand side are, however, not consistent. This is because there are no clear-cut boundaries separating the red and green clusters from each other.

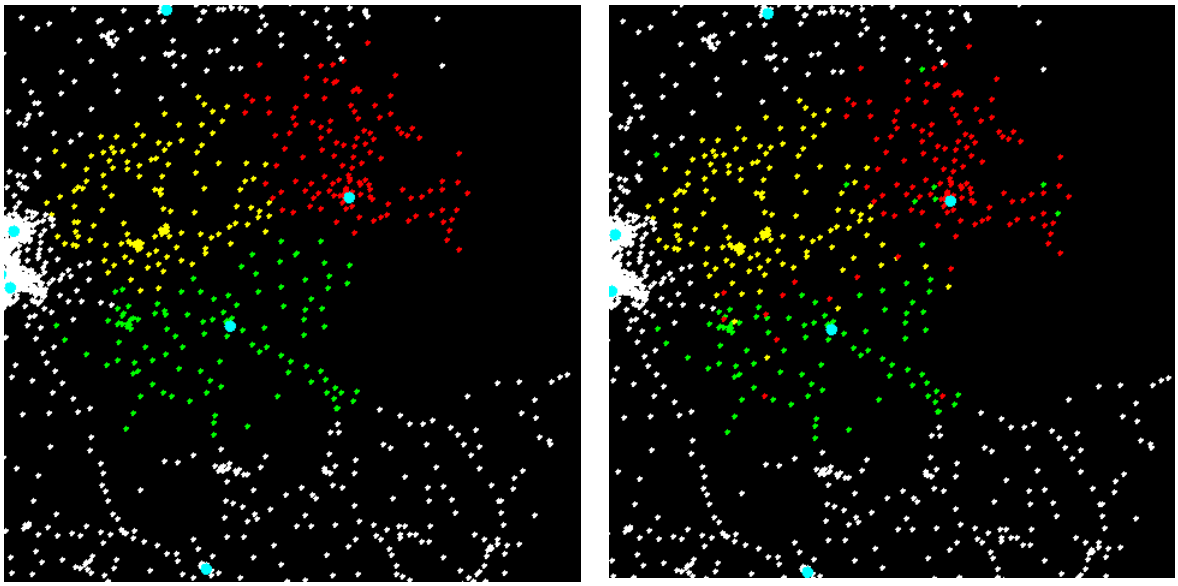

Figure 3.5 Consistent clusters (left) versus inconsistent clusters (right)

Each cluster represents the sites belonging to a specific BSC. The configuration in Figure 3.5 (left) can easily be implemented and is considered feasible. The configuration in Figure 3.5 (right), although possible to implement in theory, is deemed infeasible. One of the main reasons for not implementing configurations such as this is the massive signalling load that it introduces. When crossing multiple cell boundaries, multiple location updates have to be performed. Not only does this cause undesired traffic on the radio signalling channels, but it also places an unnecessary load on network elements.

3.6 INTERPRETATION OF EXPERIMENTAL RESULTS

All traffic as well as revenue measures are presented as percentages in the dissertation. This was done to protect the highly confidential and sensitive nature of the data without compromising the useful information that can the extracted from the results.

All measurements/comparisons were made relative to the operator's current network configuration (topology). If, for example, a technique states a core network transmission cost of 80% after the optimisation process, it should be interpreted as a 20% saving. This saving (reduction) is relative to the operator's current core network transmission cost. Likewise, if the core network traffic is stated as being 115%, it should be interpreted as a 15% increase in core network traffic (relative to the operator's current configuration). Lastly, if a measure is stated as being 100%, it means no change relative to the operator's current network configuration was achieved.

The experimental results of the investigated and developed heuristics is presented in the chapters to follow (Chapter 4 and Chapter 5) and should be interpreted as described above.

4 HEURISTIC SEARCHES

4.1 INTRODUCTION

Five heuristic search techniques were developed. This was done to satisfy two needs. Firstly, to explore the search space and obtain a feel of the complexity of the problem. Secondly, to determine the extent, if any, to which improvements on the current network configuration using heuristics can be made. Results generated by these experiments served as benchmarks to compare the results of the evolutionary and unsupervised search techniques with.

The same network region was used in all of the experiments. This region consisted of three BSCs and 385 sites. The colour of a site indicates the BSC it is working off. Figure 4.2 (left) illustrates the operator's current network configuration. BSCs are indicated with cyan-coloured dots. Sites coloured red work off the Nelspruit BSC located at the top right-hand corner of the image. Sites coloured green work off the Ermelo BSC situated at the bottom of the image. Yellow-coloured sites work off the Pretoria BSC situated at the top left-hand side of the image.

The Ermelo and Nelspruit BSCs are both remote BSCs, Pretoria BSC is a BSC/TRC unit. Nelspruit BSC and Pretoria BSC work off the same MSC situated in the Pretoria switching centre, while the Ermelo BSC works off an MSC located in the Germiston switching centre. There is no direct route between the Pretoria and Germiston switching centres, all traffic between them is routed via the Germiston gateway MSC.

4.2 SEARCH TECHNIQUE 1 (SEARCH BEST)

4.2.1 Overview

Search best is a hill-climbing algorithm with the goal of reducing the inter-switching centre traffic as much as possible. No network element constraints, load-balancing or transmission costs were taken into account. During this optimisation process, a single site was reparented at a time until no further improvement could be made. The operator's current network configuration was used as the starting point.

This experiment had two very important goals. First, the extent to which the core network traffic could be reduced while ignoring all constraints had to be established. Secondly, the impact this had on the BSC cluster consistency and feasibility had to be determined.

4.2.2 Implementation

A standard hill-climbing algorithm was implemented. The entire network was modelled but changes were only made to the optimisation region. Every site in the optimisation region had a chance of being reparented. This reparenting was performed one site at a time. The change introduced by the reparenting was saved and the site was moved back to its original BSC. Once all the sites had a chance of being reparented, the change introducing the largest saving was selected. Only this change was made permanent. This resulting

network configuration was then used as the new starting point and the process repeated itself. Optimisation continued until no further savings could be achieved. A flow diagram illustrating the process is presented in Figure 4.1.

4.2.3 Results and discussion

Figure 4.2 compares the original BSC boundaries (left-hand side) with the boundaries found by the heuristic (right-hand side). Unfortunately, the configuration found by the heuristic is not feasible. The main reason for this is the inconsistent cluster boundaries.

Interswitching centre traffic was reduced by nearly 3%. This was done at the expense of increasing inter-BSC traffic by 4%. When moving a site between BSCs working off different switching centres, it is possible to decrease the interswitching centre but increase the inter-BSC traffic as was found in this experiment.

The access transmission cost of the area under consideration rose by 8.7%. Most of the reparented sites were moved to BSCs further away than their original BSC parents. This explains the increase in access transmission cost.

Backhaul transmission cost within the optimisation area fell by 4%. Two of the BSCs in the optimisation are remote BSCs while one is a BSC/TRC co-located with its MSC parent. Enough sites were moved from the remote BSCs to the co-located BSC to cause a reduction in Ater transmission links. The rise of traffic on the co-located BSC does not cause a rise in the backhaul transmission cost since tie cables are used to connect the BSC/TRC to its MSC parent.

From Figure 4.2 (right) and Table 4.3, it can be seen that optimising purely based on traffic distributions creates infeasible network configurations. The generated configuration is not only infeasible due to the inconsistent clusters, but also more expensive transmission-wise and breaks network element constraints. These results emphasise the importance of taking all the constraints into consideration when performing optimisation. It also illustrates the high level of complexity of the problem at hand. By optimising a specific aspect, several others may possibly be affected negatively.

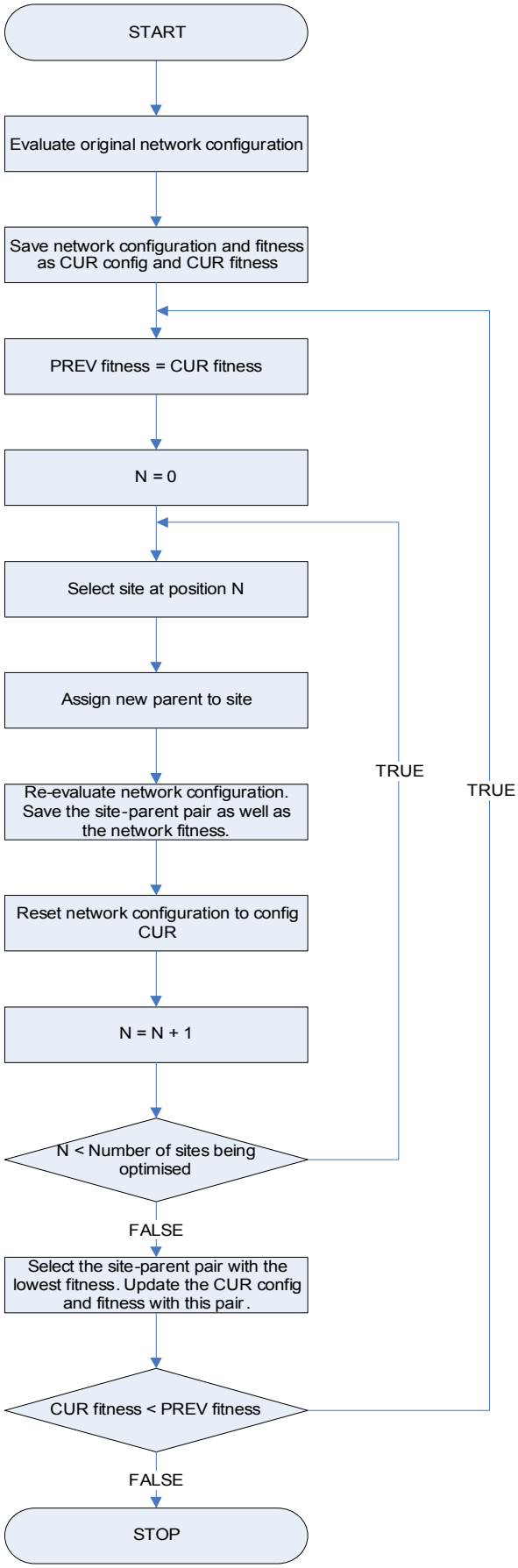

Figure 4.1 Search heuristic 1 flow diagram

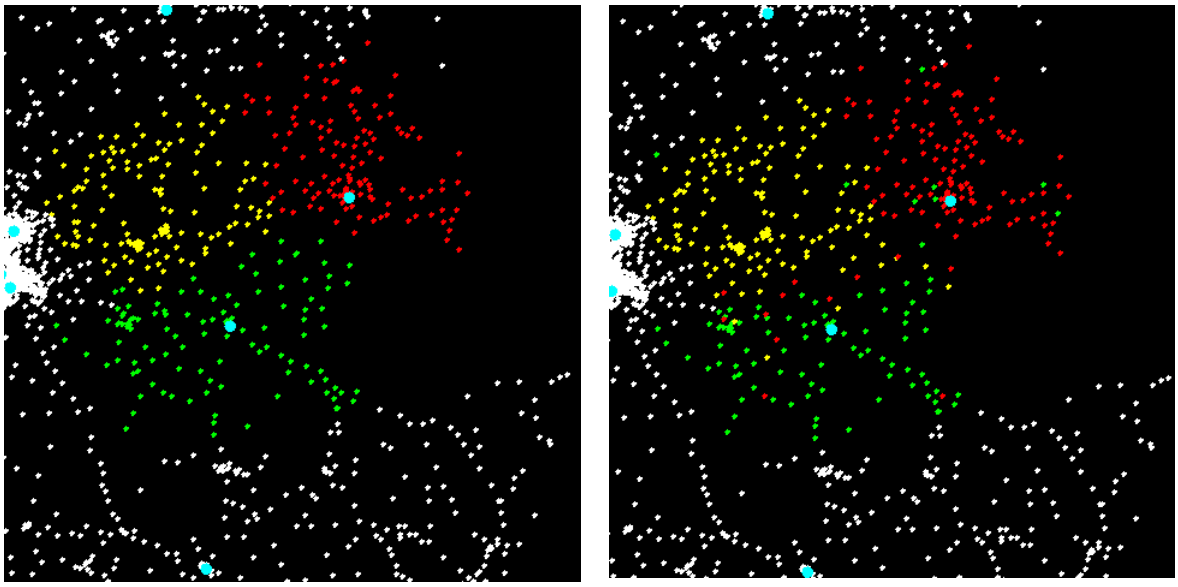

Figure 4.2 Original cluster formation (left), search technique 1 final cluster formation (right)

% Change in traffic	Inter-BSC traffic	Inter-SC traffic
Optimisation region	4.41%	-2.62%
Entire network	0.03%	-0.12%

Table 4.1 Traffic measurements

% Change in transmission cost	Access transmission	Backhaul transmission	Core transmission	Total transmission
Optimisation region	8.78%	-3.85%	-	-
Entire network	0.79%	-0.54%	0.47%	0.61%

Table 4.2 Transmission cost

ID	TRX utilisation original	TRX utilisation optimised	TRA utilisation original	TRA utilisation optimised	Unit utilisation original	Unit utilisation optimised
ERBSC1	57.29%	50.26%	-	-	57.29%	50.26%
NSBSC1	75.52%	77.86%	-	-	75.52%	77.86%
PTBSC3	58.66%	61.88%	87.88%	91.29%	73.27%	76.58%

Table 4.3 BSC utilisation

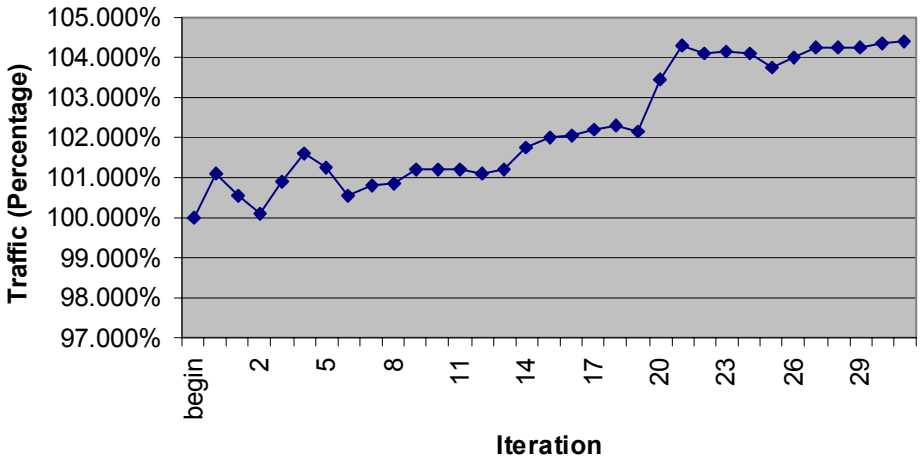

Figure 4.3 Inter-BSC traffic (optimisation region)

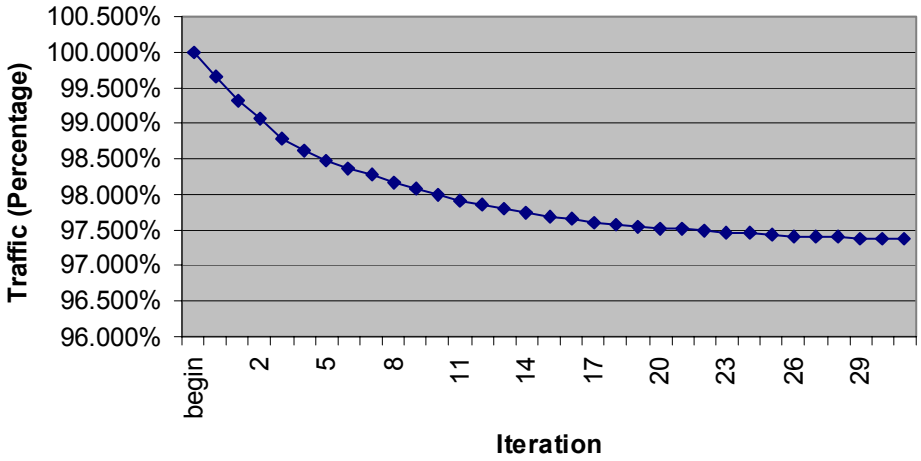

Figure 4.4 Inter-SC traffic (optimisation region)

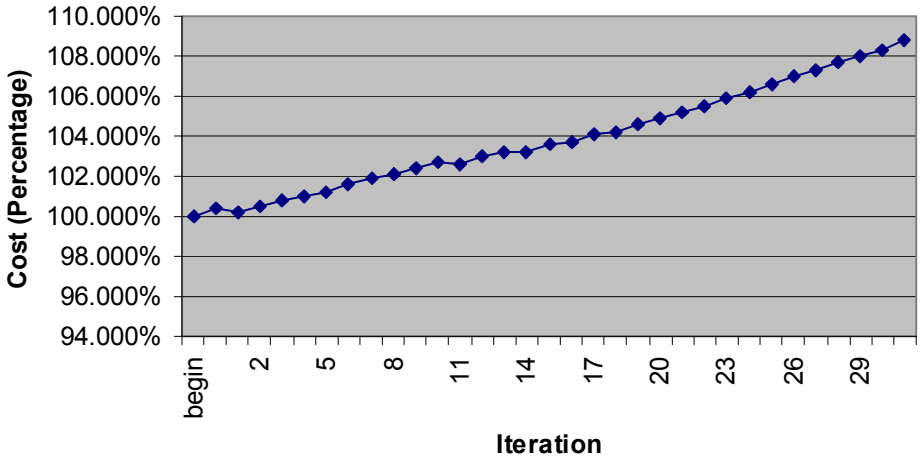

Figure 4.5 Access transmission cost (optimisation region)

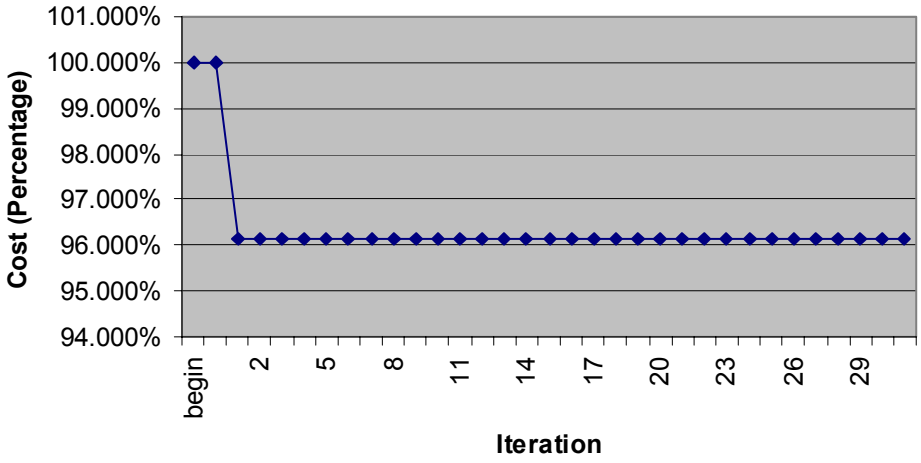

Figure 4.6 Backhaul transmission cost (optimisation region)

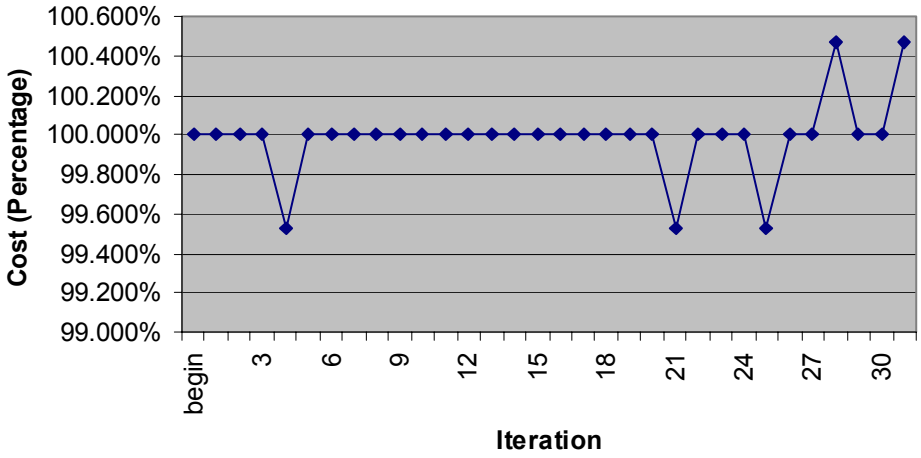

Figure 4.7 Core transmission cost (entire network)

4.3 SEARCH TECHNIQUE 2 (SEARCH CLOSEST)

4.3.1 Overview

The search closest algorithm tried minimising the access transmission cost as much as possible. This was achieved by connecting every site in the optimisation region to its closest BSC. There is a correlation between link distance and link cost. By connecting a site to its closest BSC parent the access cost is minimised. The original network configuration was used as a starting point. One site at a time was reparented, after each alteration the network was re-evaluated and the results saved. No constraints, load-balancing or interswitching centre traffic was used as part of the optimisation process. This experiment aimed at illustrating the effect on the network when connecting each site to its closest parent.

4.3.2 Implementation

The original MTN configuration was used as starting point. All the sites in the optimisation region were connected to the closest possible BSC. Sites were reparented one at a time, after each alteration the network was re-evaluated to obtain the new transmission costs as well as traffic measurements. A flow diagram of the algorithm used to perform the reparenting is presented below.

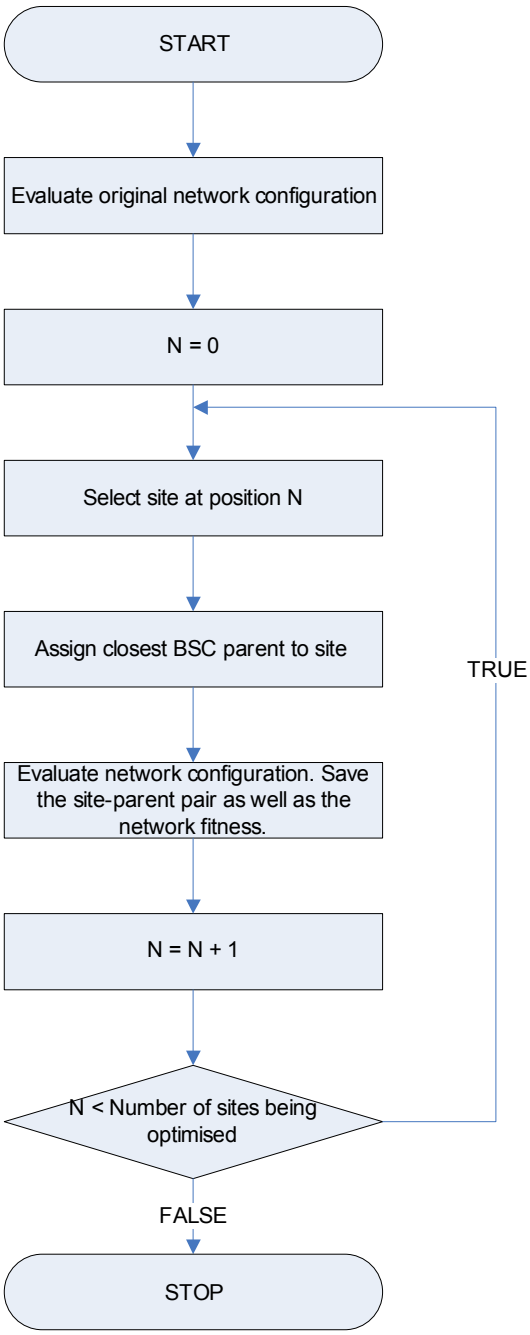

Figure 4.8 Search heuristic 2 flow diagram

4.3.3 Results and discussion

From Figure 4.9 it can be seen that the clusters formed were mostly consistent. Inter-BSC as well as inter-SC traffic increased significantly.

A massive reduction in access transmission costs was achieved. This was expected since there is a direct relationship between the length and cost of a link. The access transmission cost of the area under investigation was reduced by 13%. The transmission bill of the entire network taking all link types into consideration dropped by 0.3%. This represents a significant saving.

Backhaul transmission costs increased. The yellow cluster is the only one with a BSC that is co-located with its MSC parent. This cluster was reduced in size while the other two clusters (with remote BSC parents) grew. The consistent increase in backhaul transmission cost is thus justified. Although the newly formed clusters are consistent, the network configuration is infeasible, due to the 80% load limit constraint which was broken.

From these results, it can be seen that by reducing the access transmission costs significant savings can be achieved. These savings are, however, useless if they are not performed within the boundaries set by the constraints. On a network-wide level, the core transmission costs increased by 0.5%. This increase accounts for a substantial portion of the total transmission cost. By also taking the inter-SC traffic into consideration and minimising the core transmission as well as access network costs, even larger savings can be achieved.

The effect of the new cluster formation on the traffic and transmission cost is supplied below.

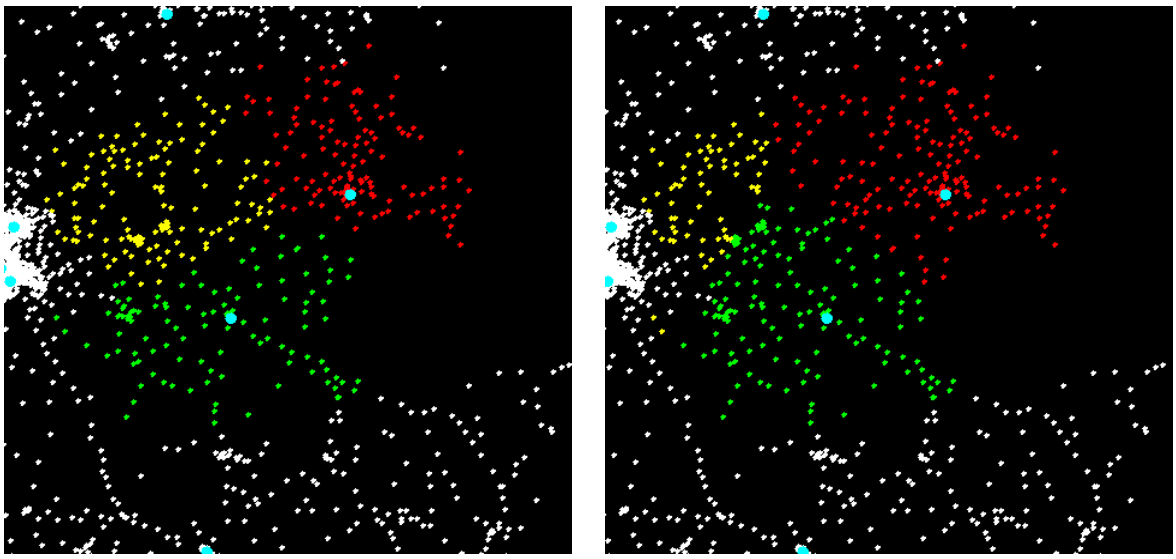

Figure 4.9 Original cluster formation (left), search technique 2 final cluster formation (right)

% Change in traffic	Inter-BSC traffic	Inter-SC traffic
Optimisation region	41.43%	11.15%
Entire network	0.26%	0.49%

Table 4.4 Traffic measurements

% Change in transmission cost	Access transmission	Backhaul transmission	Core transmission	Total transmission
Optimisation region	-13.14%	33.14%	-	-
Entire network	-1.18%	4.62%	0.47%	-0.32%

Table 4.5 Transmission cost

ID	TRX utilisation original	TRX utilisation optimised	TRA utilisation original	TRA utilisation optimised	Unit utilisation original	Unit utilisation optimised
ERBSC1	57.29%	77.21%	-	-	57.29%	77.21%
NSBSC1	75.52%	93.36%	-	-	75.52%	93.36%
PTBSC3	58.66%	32.77%	87.88%	76.52%	73.27%	54.64%

Table 4.6 BSC utilisation

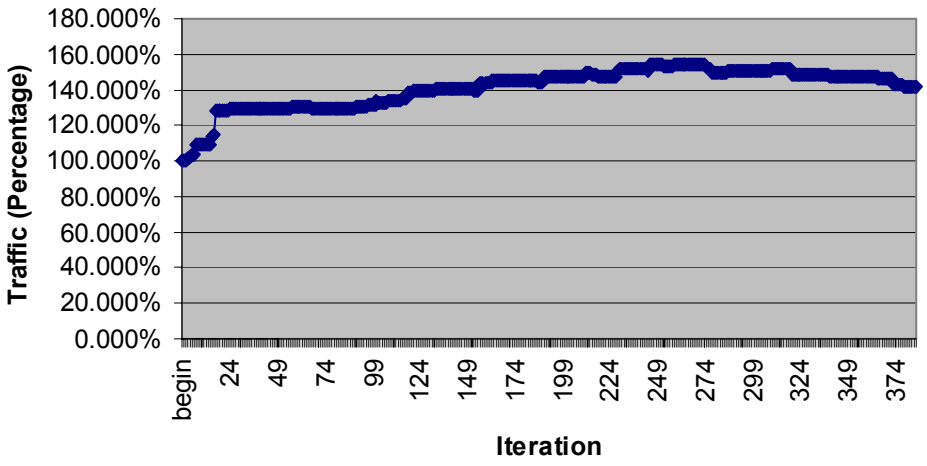

Figure 4.10 Inter-BSC traffic (optimisation region)

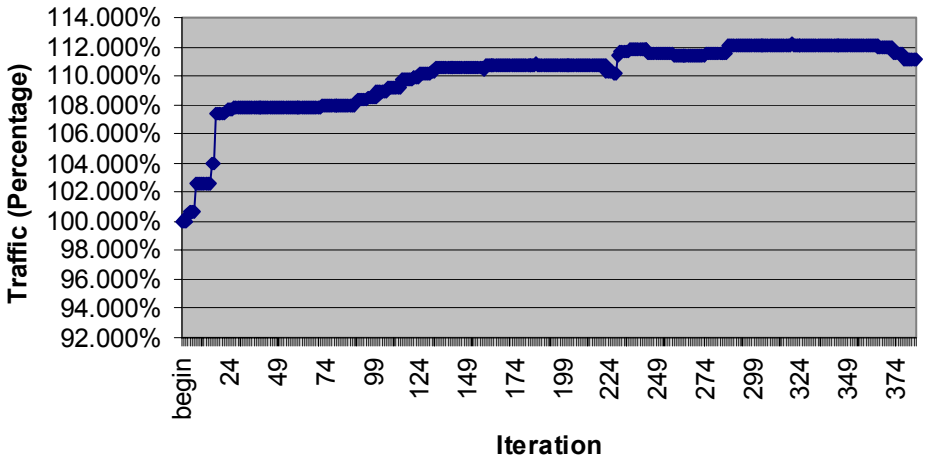

Figure 4.11 Inter-SC traffic (optimisation region)

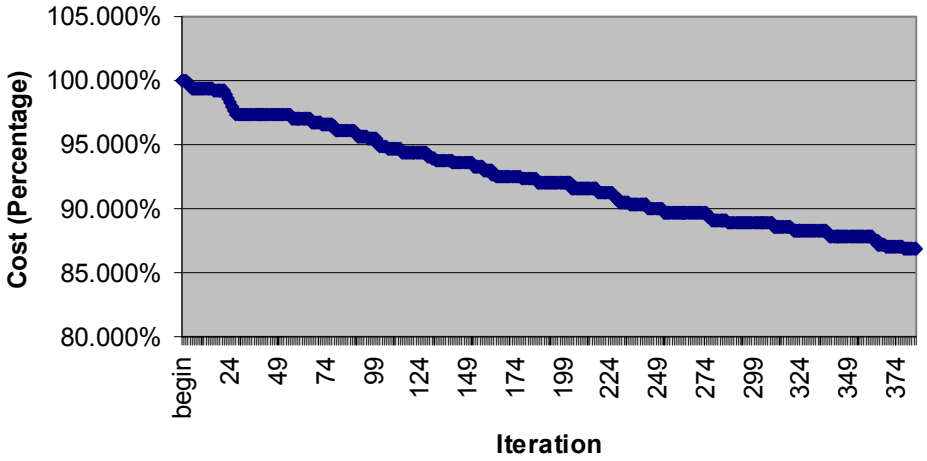

Figure 4.12 Access transmission cost (optimisation region)

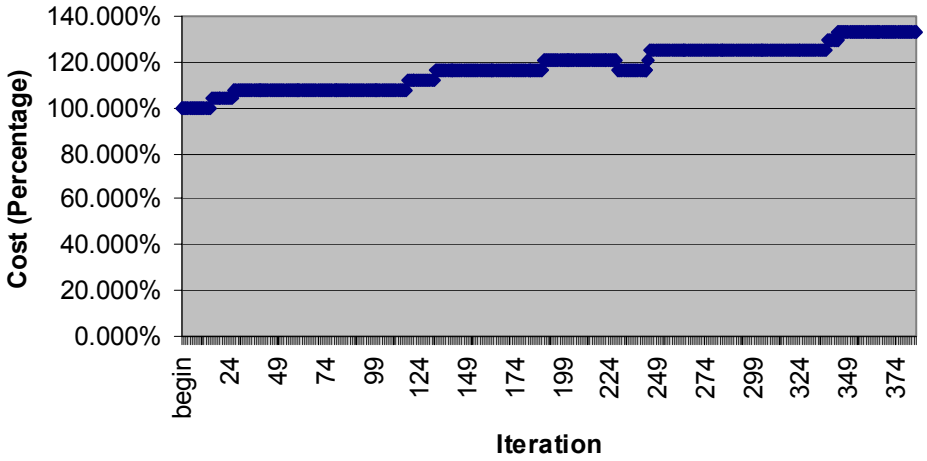

Figure 4.13 Backhaul transmission cost (optimisation region)

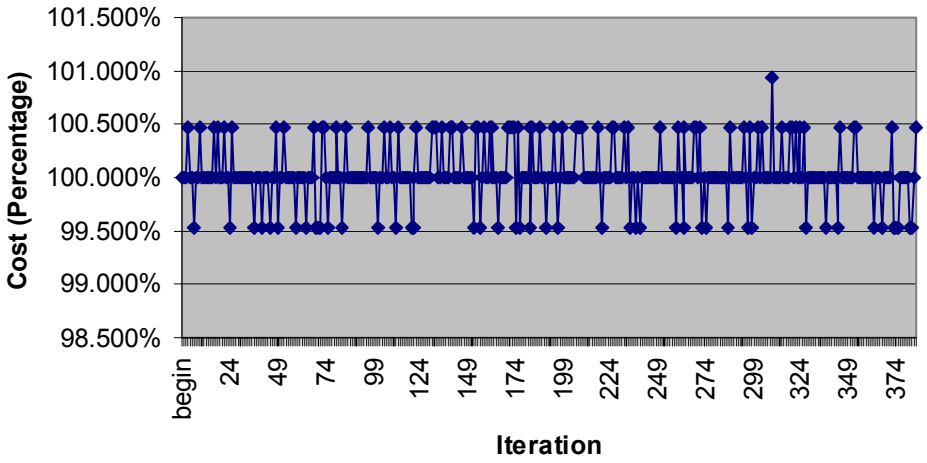

Figure 4.14 Core network cost (entire network)

4.4 SEARCH TECHNIQUE 3 (SEARCH BSC-BALANCED)

4.4.1 Overview

Search technique 3 tried balancing the load on the BSCs as far as possible. No network element constraints or traffic distributions were used as part of the decision criteria. Search technique 3 shares a common characteristic with search technique 2, namely taking access transmission cost into consideration. The major difference is that heuristic 3 does not blindly assign a site to its closest BSC parent as heuristic 2 did, but tries to keep all the BSCs under consideration nearly equally loaded. From this experiment the researcher wanted to see how a load-balanced network configuration compares traffic- and transmission-wise with other configurations where load-balanced BSCs were not of importance.

4.4.2 Implementation

All the sites being optimised start out as being parentless. The following process then repeats itself until there is no more unparented sites left. The BSC in the optimisation region with the lowest utilisation is selected. If there is a tie between two or more BSCs, one of them is selected at random. The closest unparented site to the selected BSC is identified and connected to it. Figure 4.15 illustrates the optimisation process.

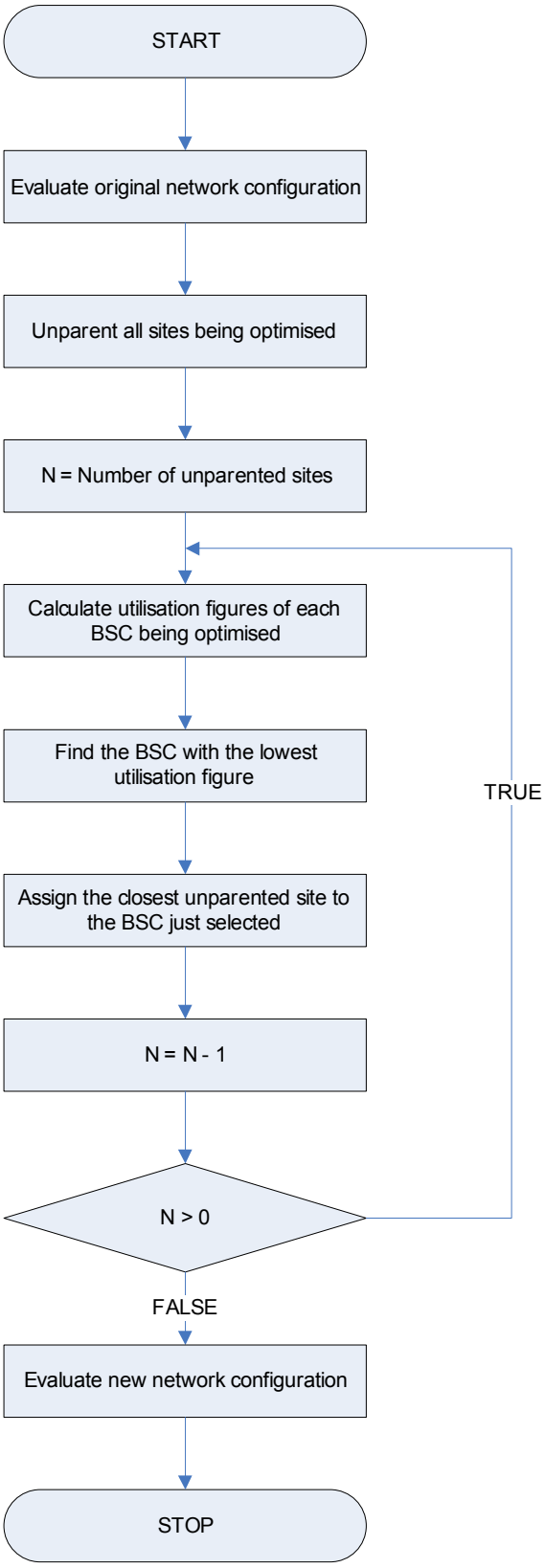

Figure 4.15 Search heuristic 3 flow diagram

4.4.3 Results and discussion

The newly formed clusters are mostly consistent. All of the inconsistencies are located near the cluster edges. These inconsistencies can easily be resolved by running the configuration through a cluster-correcting algorithm. Sites such as the three green ones in the right-hand site of the red cluster will be reparented to form part of the red cluster.

Surprisingly good results were obtained. Even though the BSCs are now nearly identically utilised, access and backhaul transmission costs were reduced by 3.91% and 1.19% respectively (optimisation region) with the core transmission cost remaining constant. No network element constraints were broken. These savings as well as the fact that the BSC loads are now balanced were achieved at the expense of increased inter-BSC and -SC traffic.

The total transmission saving on a network-wide level made by heuristic 3 is comparable with that of heuristic 2. Heuristic 3 adds the advantage of having load-balanced BSCs afterwards which eases future network expansion. The new BSC boundaries, although feasible, are not really practical. This impracticality is caused by the green strip separating the red and yellow clusters. In practice, this strip would generate a large amount of handovers and location updates putting an unnecessary high strain on the signalling channels, signalling network and BSC as well as MSC processing units.

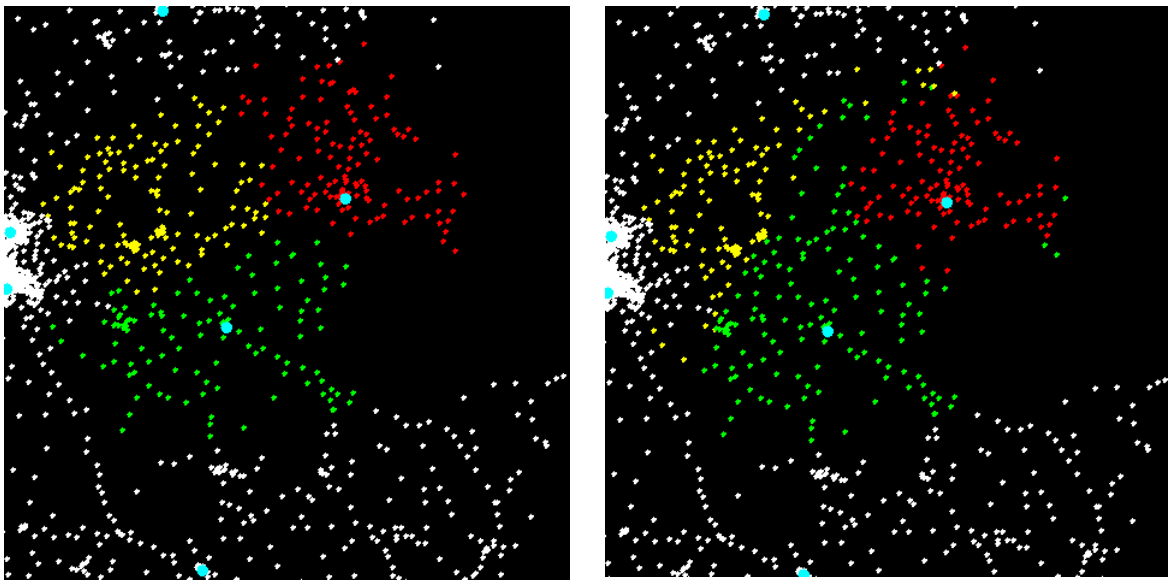

Figure 4.16 Original cluster formation (left), search technique 3 final cluster formation (right)

% Change in traffic	Inter-BSC traffic	Inter-SC traffic
Optimisation region	43.90%	11.22%
Entire network	0.28%	0.50%

Table 4.7 Traffic measurements

% Change in transmission cost	Access transmission	Backhaul transmission	Core transmission	Total transmission
Optimisation region	-3.91%	-1.19%	-	-
Entire network	-0.35%	-0.17%	0.00%	-0.22%

Table 4.8 Transmission cost

ID	TRX utilisation original	TRX utilisation optimised	TRA utilisation original	TRA utilisation optimised	Unit utilisation original	Unit utilisation optimised
ERBSC1	57.29%	68.62%	-	-	57.29%	68.62%
NSBSC1	75.52%	68.62%	-	-	75.52%	68.62%
PTBSC3	58.66%	55.63%	87.88%	81.44%	73.27%	68.53%

Table 4.9 BSC utilisation

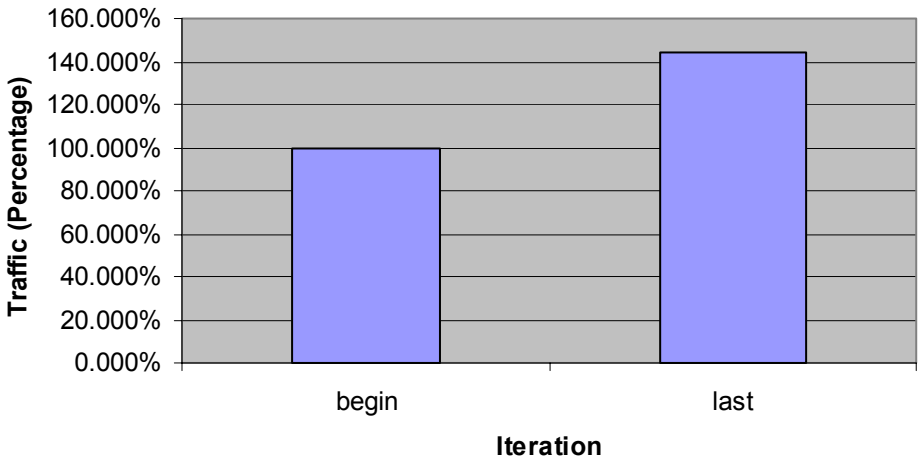

Figure 4.17 Inter-BSC traffic (optimisation region)

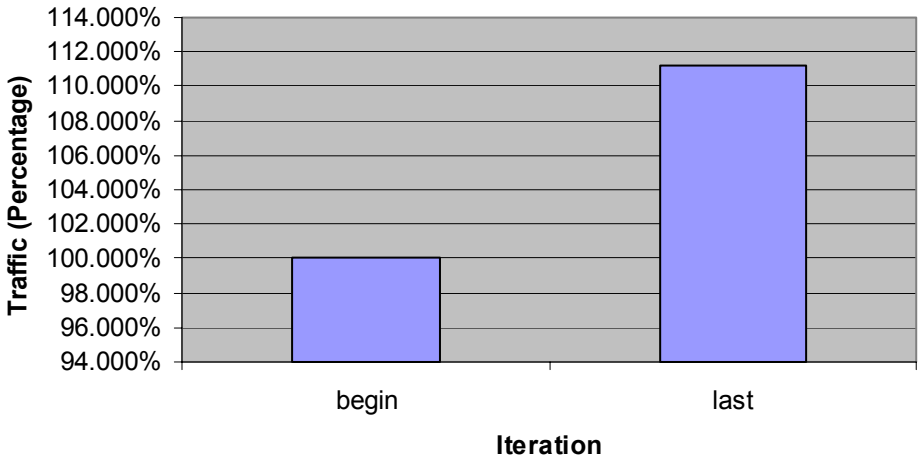

Figure 4.18 Inter-SC traffic (optimisation region)

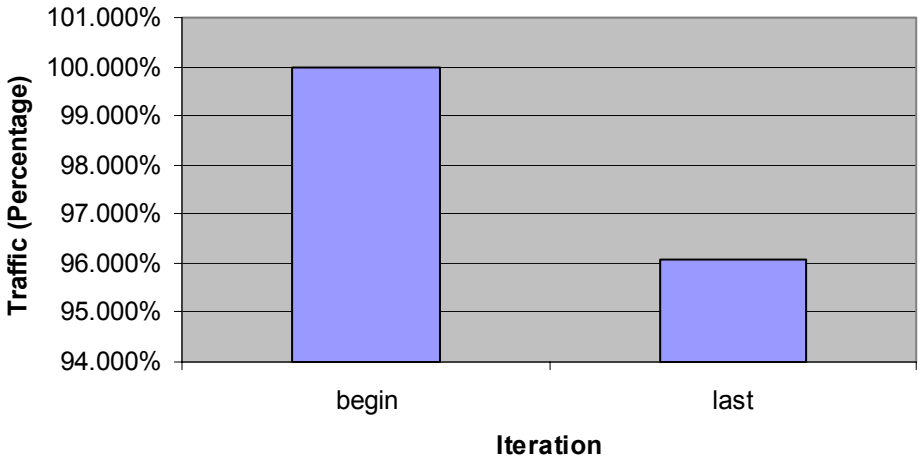

Figure 4.19 Access transmission cost (optimisation region)

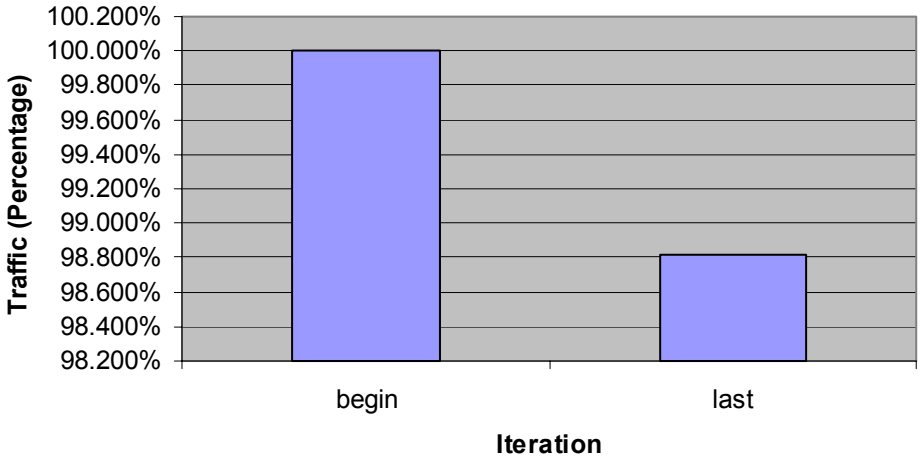

Figure 4.20 Backhaul transmission cost (optimisation region)

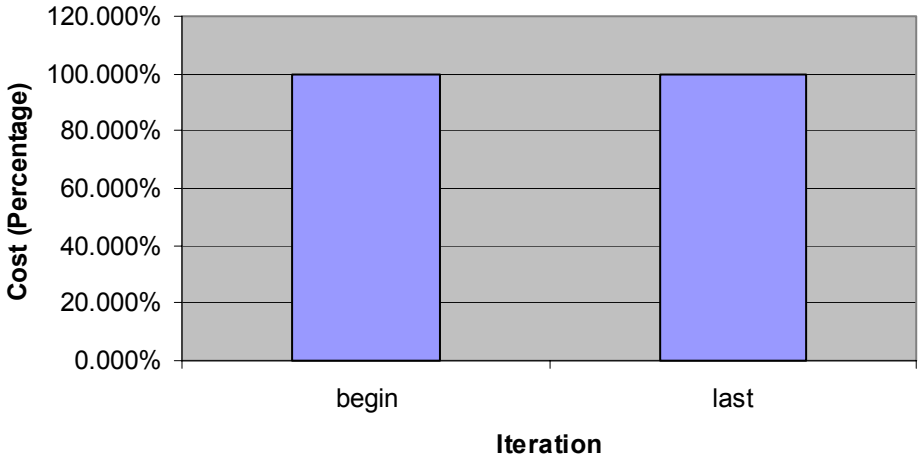

Figure 4.21 Core transmission cost (entire network)

4.5 SEARCH TECHNIQUE 4 (SEARCH BSC CONSTRAINTS GROUPED CLUSTERS)

4.5.1 Overview

Search technique 4 is an extension of search technique 2. In search technique 2, the cost of the access network was maximally reduced, no traffic information or unit constraints were taken into account. In this technique, the aim is again to minimise the access transmission cost but with the requirements that no BSC constraints are broken and that the formed clusters remain consistent (feasible). Using the results of this experiment, the researcher wanted to perform two comparisons. Firstly, to see how the new traffic distributions compared with the current configuration's traffic distributions. Secondly, to compare the savings made using heuristic 2 with the potential savings made by heuristic 4.

4.5.2 Implementation

The operator's original configuration was used as a starting point. Per iteration the following steps were taken. An individual was selected. The BSC parents of the four closest sites to the site being optimised were determined. The site was connected to the closest BSC parent that had enough spare capacity to accommodate it. This process continued until no further improvements were made.

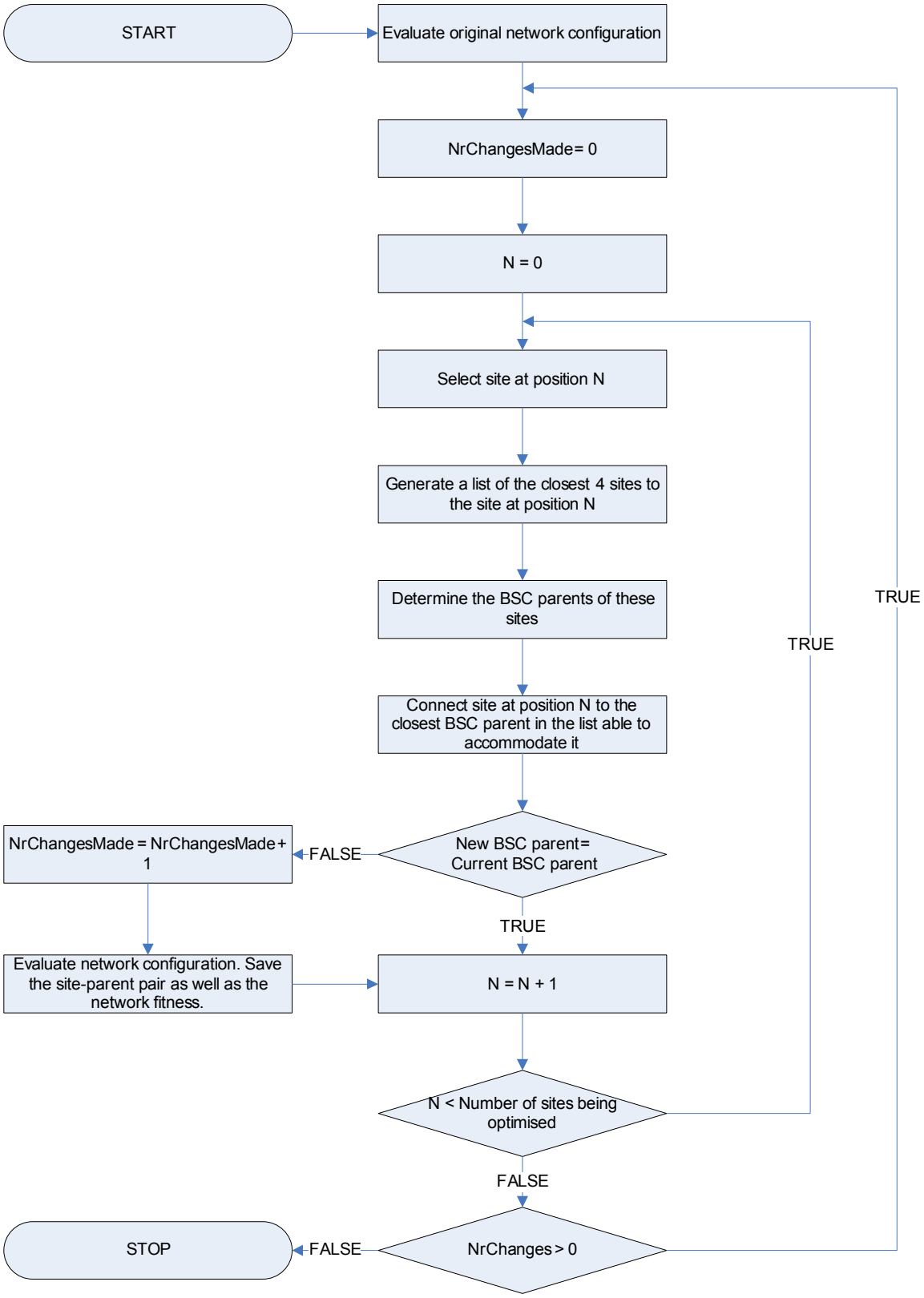

Figure 4.22 Search heuristic 4 flow diagram

4.5.3 Results and discussion

Results obtained from this heuristic were extremely positive. The network configuration created was feasible. The clusters formed were thus consistent and all network element constraints were adhered to. The total transmission cost was lower than that of the original configuration as well as that obtained by heuristic 2. Heuristic 2 produced a 0.2% larger access transmission network saving than heuristic 4, but it had a 2% higher backhaul and 0.94% higher core transmission costs associated with it (based on the entire network's transmission cost). When taking all the transmission costs and network element constraints on a network-wide level into account, heuristic 4 clearly outperformed heuristic 2.

The increased backhaul transmission cost in the solution found by heuristic 4 is expected. From Figure 4.23, it can be seen that the traffic load of the two remote BSCs increased while the traffic load of the co-located BSC decreased. This will naturally translate into an increased backhaul cost.

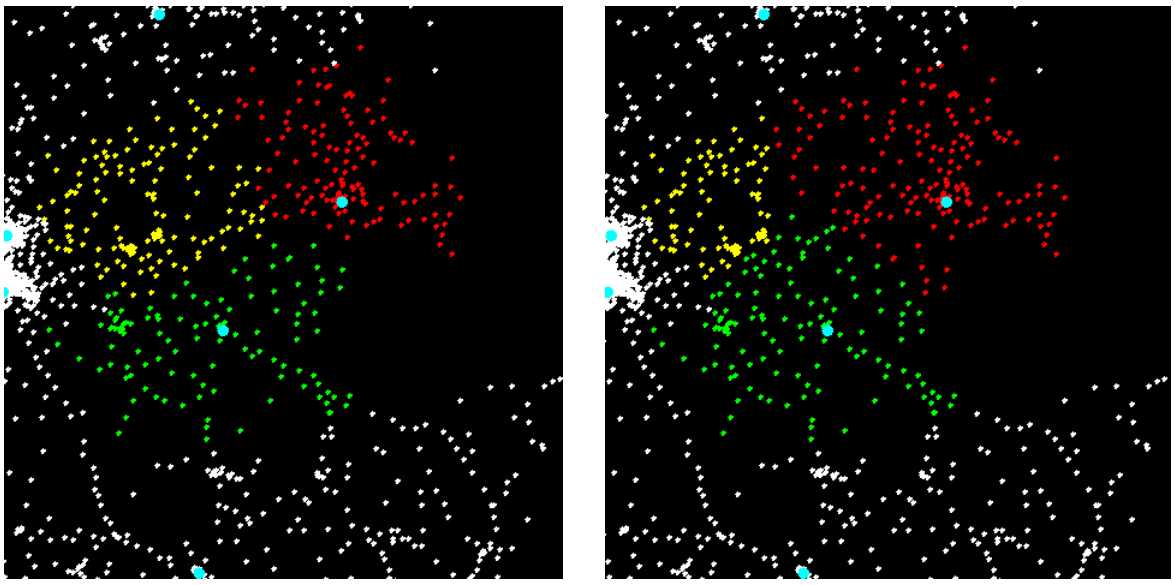

Figure 4.23 Original cluster formation (left), search technique 4 final cluster formation (right)

% Change in traffic	Inter-BSC traffic	Inter-SC traffic
Optimisation region	33.74%	3.95%
Entire network	0.21%	0.17%

Table 4.10 Traffic measurements

% Change in transmission cost	Access transmission	Backhaul transmission	Core transmission	Total transmission
Optimisation region	-10.94%	17.75%	-	-
Entire network	-0.99%	2.47%	-0.47%	-0.63%

Table 4.11 Transmission cost

ID	TRX utilisation original	TRX utilisation optimised	TRA utilisation original	TRA utilisation optimised	Unit utilisation original	Unit utilisation optimised
ERBSC1	57.29%	61.98%	-	-	57.29%	61.98%
NSBSC1	75.52%	93.36%	-	-	75.52%	93.36%
PTBSC3	58.66%	43.21%	87.88%	86.74%	73.27%	64.98%

Table 4.12 BSC utilisation

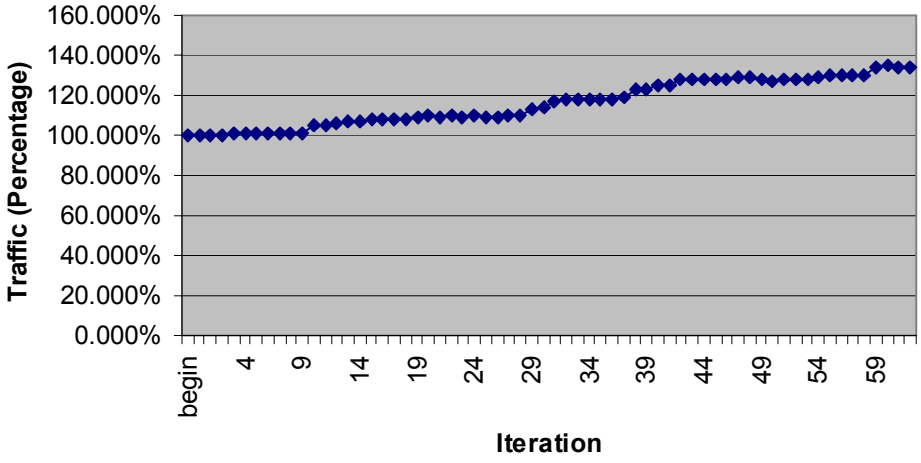

Figure 4.24 Inter-BSC traffic (optimisation region)

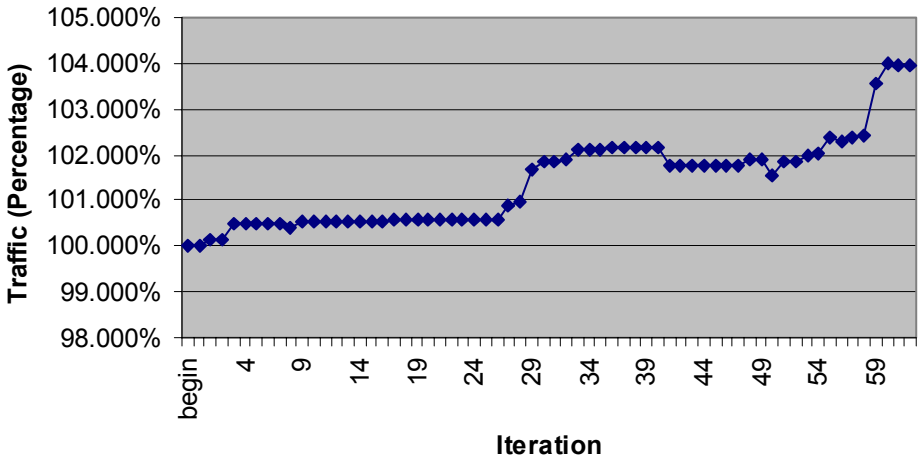

Figure 4.25 Inter-SC traffic (optimisation region)

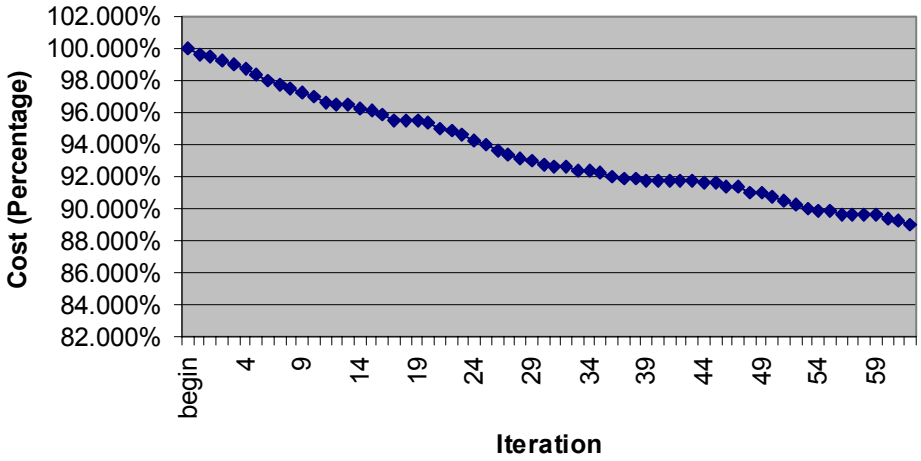

Figure 4.26 Access transmission cost (optimisation region)

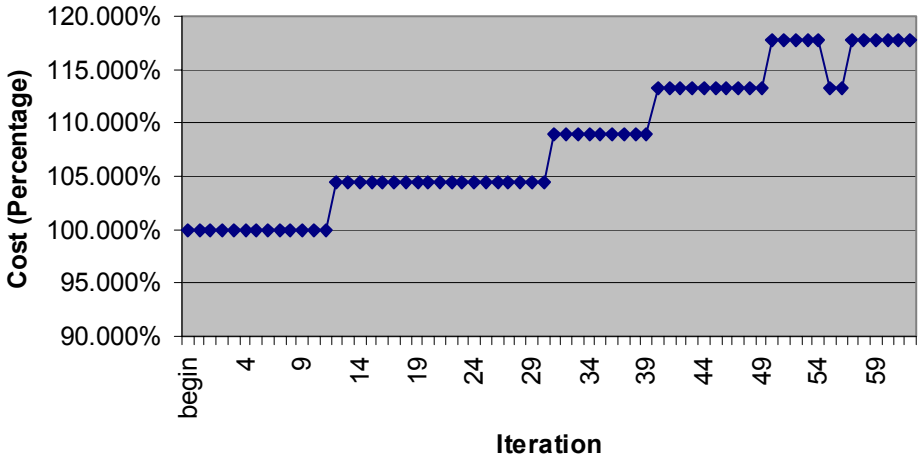

Figure 4.27 Backhaul transmission cost (optimisation region)

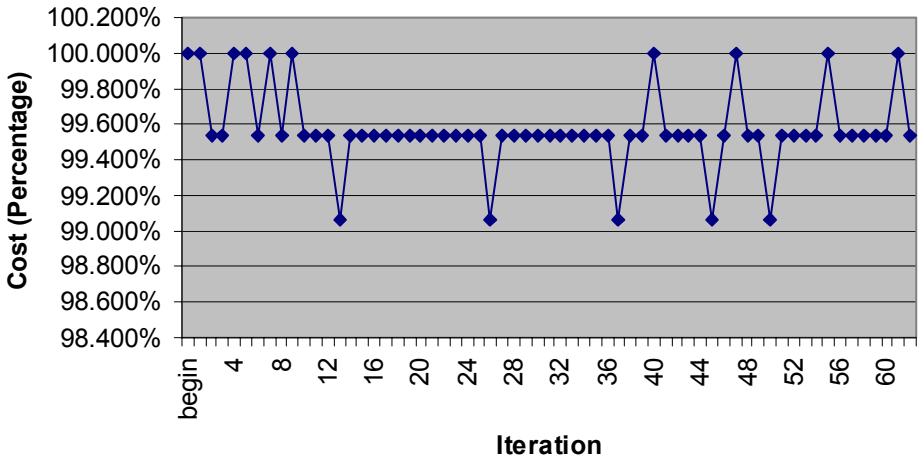

Figure 4.28 Core transmission cost (entire network)

4.6 SEARCH TECHNIQUE 5 (SEARCH BSC-BALANCED TABLE METHOD)

4.6.1 Overview

Search technique 5 is an extension of search technique 3. In search technique 3, the BSCs were load-balanced without taking network element constraints, transmission costs, cluster feasibility or traffic distributions into account. With search technique 5, load-balancing was performed within the constraints of the network elements (BSCs). Cluster feasibility as well as access transmission costs were also taken into account.

Load-balancing was achieved by cutting sites with weak connections off their current BSC to a new BSC with a lower load. The connection between a site and its parent is defined by either the distance to the parent, the cost of the link connecting the child to the parent or the traffic load between the child and the parent. In this simulation, connection strength was defined by distance. The further a site was away from its parent, the weaker the connection.

From this experiment, the researcher wanted to see the effect the load-balanced BSCs with node and cluster feasibility constraints taken into account had on the transmission cost and traffic distributions. Another interesting outcome will be to compare the level of load-balancing achieved by technique 5 with that of technique 3.

4.6.2 Implementation

The following approach was taken. The original network configuration was used as a starting point. The BSC in the optimisation region with the highest load was selected. Next the site with the weakest connection to the high-load BSC was selected. The BSC parents of the two closest sites (to the selected site) were determined. The site was cut over to the closest BSC found in the previous step capable of accepting it without breaking any

constraints. If the weakest connected site could not be cut over, the second-weakest connected site was tried. This continued until finally the site with the strongest connection was tried. After each successful cutover, the BSC loads were recalculated and the entire process repeated itself. This continued until no further changes were made. A flow diagram illustrating the entire process is presented on the next page.

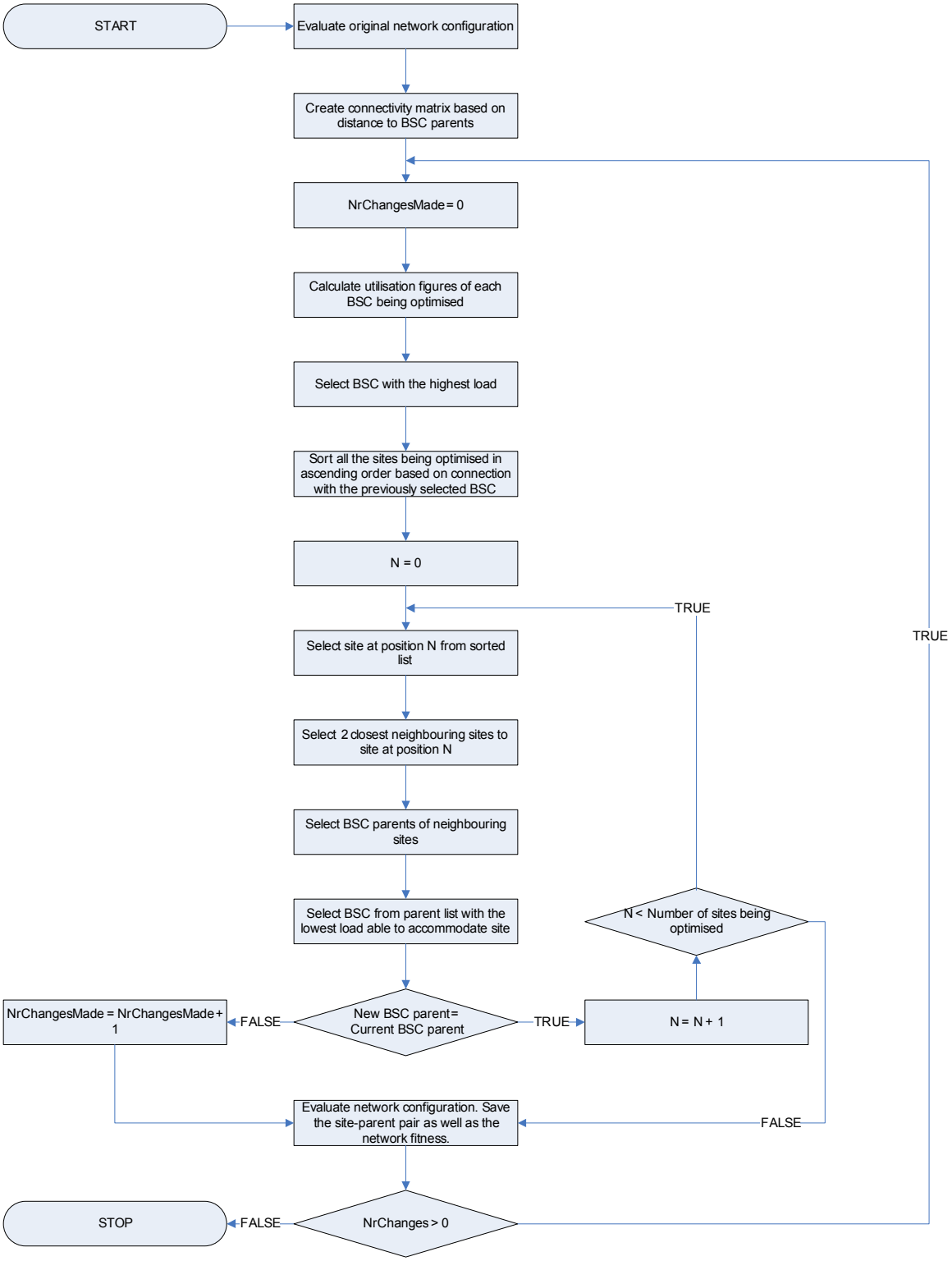

Figure 4.29 Search heuristic 5 flow diagram

4.6.3 Results and discussion

The newly formed clusters are for the most part consistent and bear a striking resemblance to the clusters formed by search heuristic 3. The level of load-balancing achieved by heuristic 5 is comparable with that of heuristic 3, unfortunately the saving achieved by heuristic 3 could not be matched by heuristic 5. A net saving of 0.68% on the access transmission cost in the optimisation region was achieved by heuristic 5 compared with the 3.91% saving achieved by heuristic 3. The difference in the core and backhaul transmission costs is neglectable.

Prior to optimisation, the load on PTBSC3 (yellow) and NSBSC1 (green) were nearly identical and considerably higher than that of ERBSC1 (red). Load-balancing was achieved by cutting sites over from PTBSC3 and NSBSC1 to ERBSC1. The reparented sites were the ones with the weakest connection to the PTBSC3 and NSBSC1. These sites were the ones on the boundary between PTBSC3 and NSBSC1. By moving them over, a strip of ERBSC1 sites separating PTBSC3 sites from NSBSC1 sites was defined.

Heuristic 3 started out with all the sites being parentless and then connecting unparented sites to the closest BSC while maintaining a balanced load between all the BSCs. Heuristic 5 started out with the network operator's current configuration and moved sites with weak connections off high-load BSCs to BSCs with a lower load. This was done with the goal of load-balancing all the BSCs in the optimisation region. These two different approaches achieved the same end goal but with different results. Even though the cluster formations have a similar shape, they are not identical. The clusters formed by heuristic 3 have the remote BSCs in the centre of the clusters, thus minimising transmission cost. Clusters formed by heuristic 5 do not have the remote BSC at the centre of the clusters, thus not minimising access transmission costs to the same extent as heuristic 3. This explains the larger saving achieved in the access transmission network by heuristic 3.

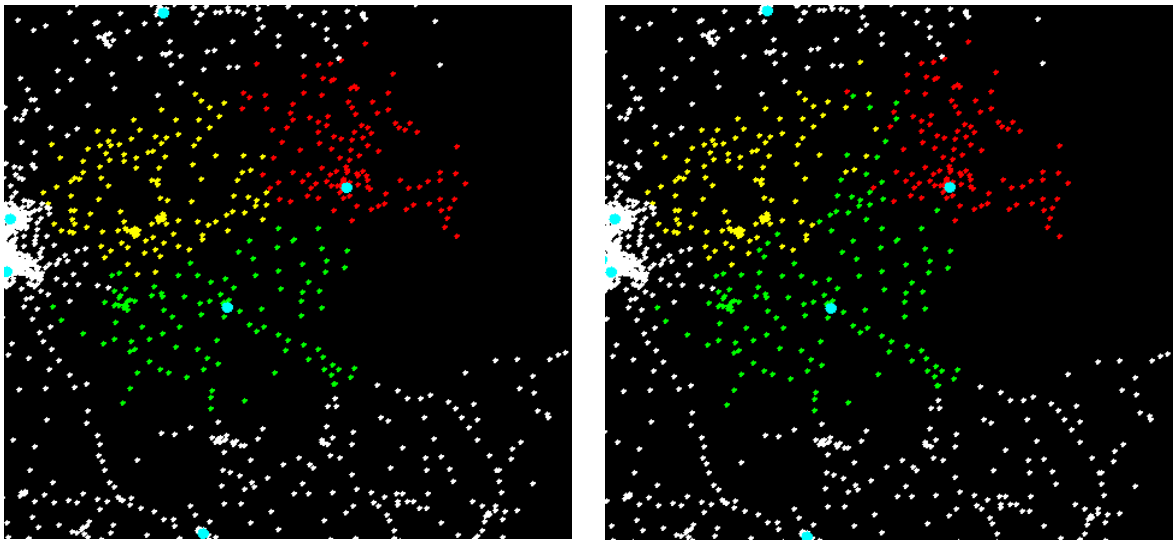

Figure 4.30 Original cluster formation (left), search technique 5 final cluster formation (right)

% Change in traffic	Inter-BSC traffic	Inter-SC traffic
Optimisation region	12.89%	4.74%
Entire network	0.08%	0.21%

Table 4.13 Traffic measurements

% Change in transmission cost	Access transmission	Backhaul transmission	Core transmission	Total transmission
Optimisation region	-0.68%	-1.19%	-	-
Entire network	-0.06%	-0.17%	0.00%	-0.05%

Table 4.14 Transmission cost

ID	TRX utilisation original	TRX utilisation optimised	TRA utilisation original	TRA utilisation optimised	Unit utilisation original	Unit utilisation optimised
ERBSC1	57.29%	69.01%	-	-	57.29%	69.01%
NSBSC1	75.52%	68.88%	-	-	75.52%	68.88%
PTBSC3	58.66%	55.18%	87.88%	82.58%	73.27%	68.88%

Table 4.15 BSC utilisation

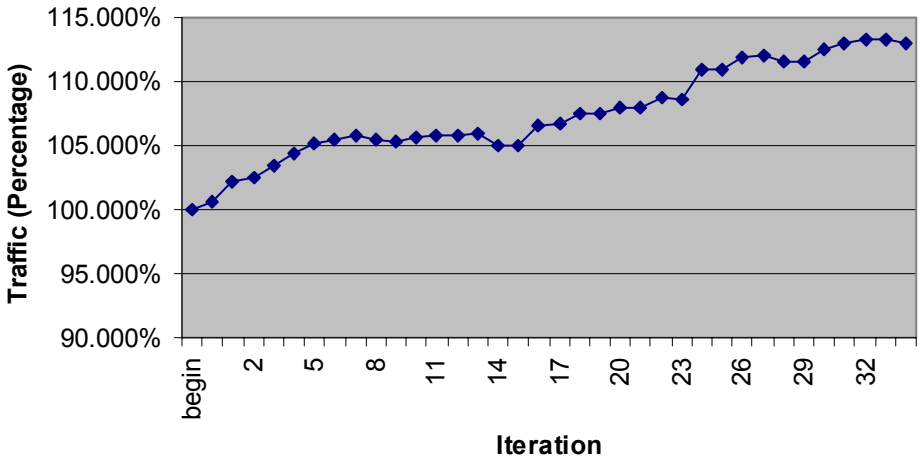

Figure 4.31 Inter-BSC traffic (optimisation region)

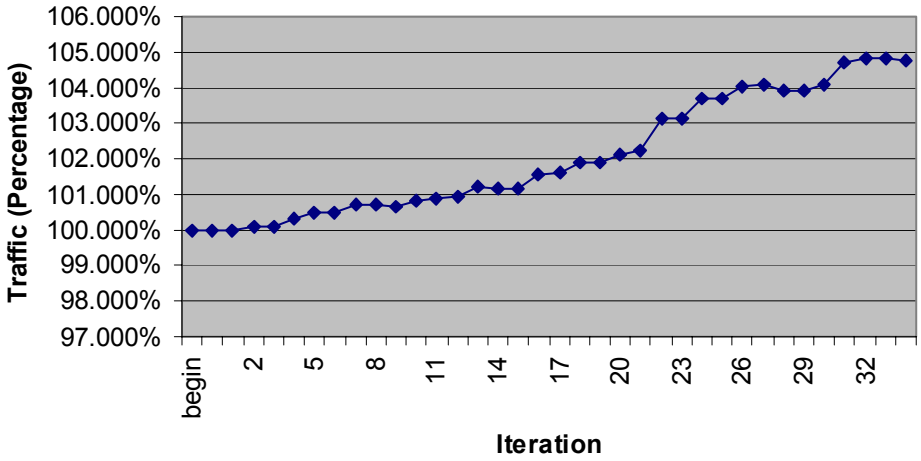

Figure 4.32 Inter-SC traffic (optimisation region)

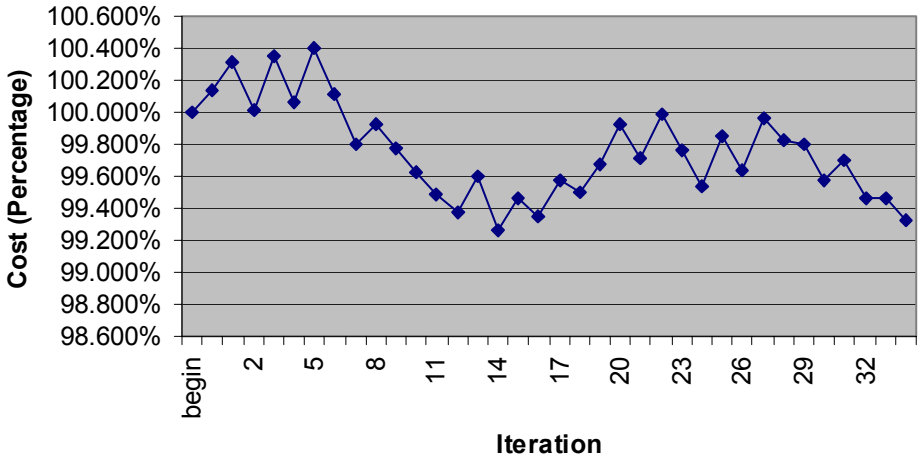

Figure 4.33 Access transmission cost (optimisation region)

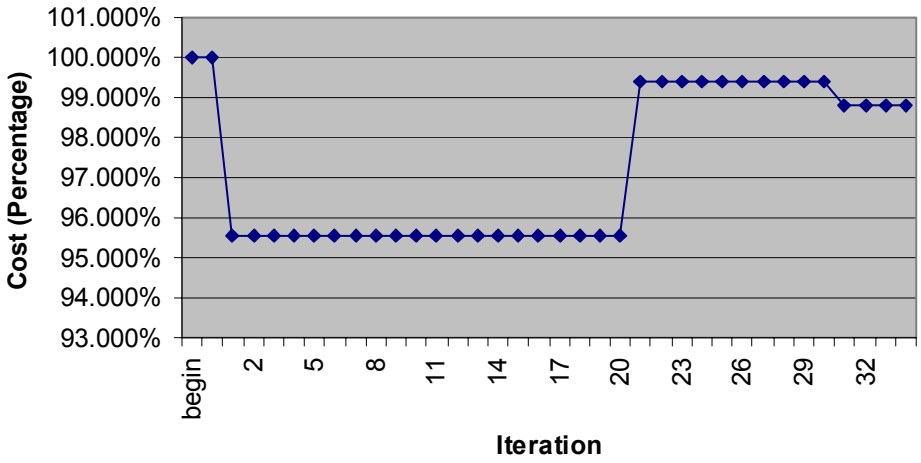

Figure 4.34 Backhaul transmission cost (optimisation region)

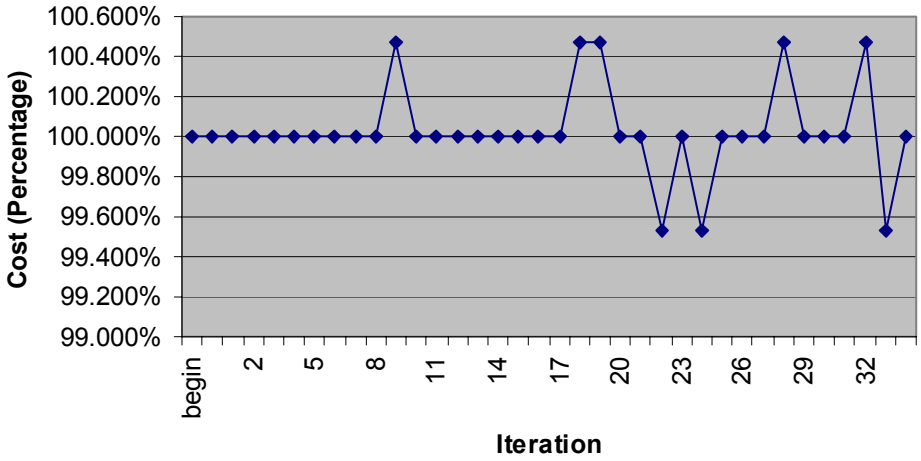

Figure 4.35 Core transmission cost (entire network)

4.7 GENETIC ALGORITHMS

4.7.1 Overview

Evolutionary computation simulates evolution on a computer. Through evolution, the quality of solutions is iteratively improved until an optimal, or at least feasible, solution is found. The goal of evolution is thus to generate a population of individuals with increasing fitness. The process of natural selection ensures that more fit individuals have the opportunity to mate most of the time, leading to the expectation that the offspring have a similar or better fitness. A genetic algorithm is used as the evolutionary computation strategy in this thesis.

In principle, a genetic algorithm starts out with a population of individuals (chromosomes) scattered around in the search space. Through the process of natural selection fitter

individuals will be selected more often to mate, hopefully creating even fitter offspring. After a number of iterations, the individuals will start to move towards the fittest individual in the population. This movement of individuals enables them to explore the search space and potentially find more optimum points. This process will continue until convergence is achieved. When convergence is achieved, all the individuals lie near the same point in the search space and no more effective exploration is thus possible.

Each chromosome represents a point in the search space. A chromosome consists of a number of genes, where the gene is the functional unit of inheritance. Each gene represents one characteristic of the individual. In terms of optimisation, a gene represents one parameter of the optimisation problem.

A genetic algorithm was chosen as the evolutionary computation strategy due to the flexibility that can be built into the fitness function. It is possible to define a custom fitness function taking all the optimisation criteria as well as constraints into consideration. This flexibility combined with the fact that genetic algorithms are global optimisation techniques made them a very attractive option.

Two approaches were followed. The first approach generated an initial population where each individual was initialised to a random network configuration. This meant that for each individual, each site in the optimisation region was assigned randomly to one of the BSCs serving the area. From this experiment, the researcher wanted to determine whether the GA is capable of finding a viable solution given this extremely tough starting point.

In the second approach, the initial population was generated to contain variations of the current network configuration. This was achieved by cloning the current network configuration to each individual and then reassigning 10% of the sites per individual at random to any one of the BSCs serving the optimisation area. With this approach, all the individuals are scattered around the current network configuration. The expectation is that the GA will be able to find an improved network configuration not differing dramatically from the current network configuration. Such a solution would represent a viable and realistic solution in the operator's eyes since it can be realised without making dramatic changes to the network's current topology.

4.7.2 Implementation

The genetic algorithm was implemented using two programs, a simulator and an evolutionary computational research system. The simulator, written in the Java programming language, was used for evaluating individuals (chromosomes) and for generating the initial population. The genetic algorithm framework was provided by a freely available Java-based evolutionary computational research system [29]. This framework provided the basic genetic algorithm functionality which was extended to meet the requirements of the problem at hand.

The logical functioning of a genetic algorithm can be broken down into the following major steps [30].

1 Represent the problem variable domain as a chromosome of fixed length
2 Define a fitness function to measure the performance of a chromosome
3 Randomly generate an initial population of chromosomes
4 Calculate the fitness of each chromosome
5 Select a pair of chromosomes for mating
6 Create a pair of offspring using crossover
7 Mutate the offspring with a fixed probability
8 Place the newly created offspring in a new population
9 Return to step 5 until the new population is equal in size to the initial population
10 Replace the initial population with the new population
11 Return to step 4 until the termination criteria are met

Figure 4.36 illustrates the flow of events as described above.

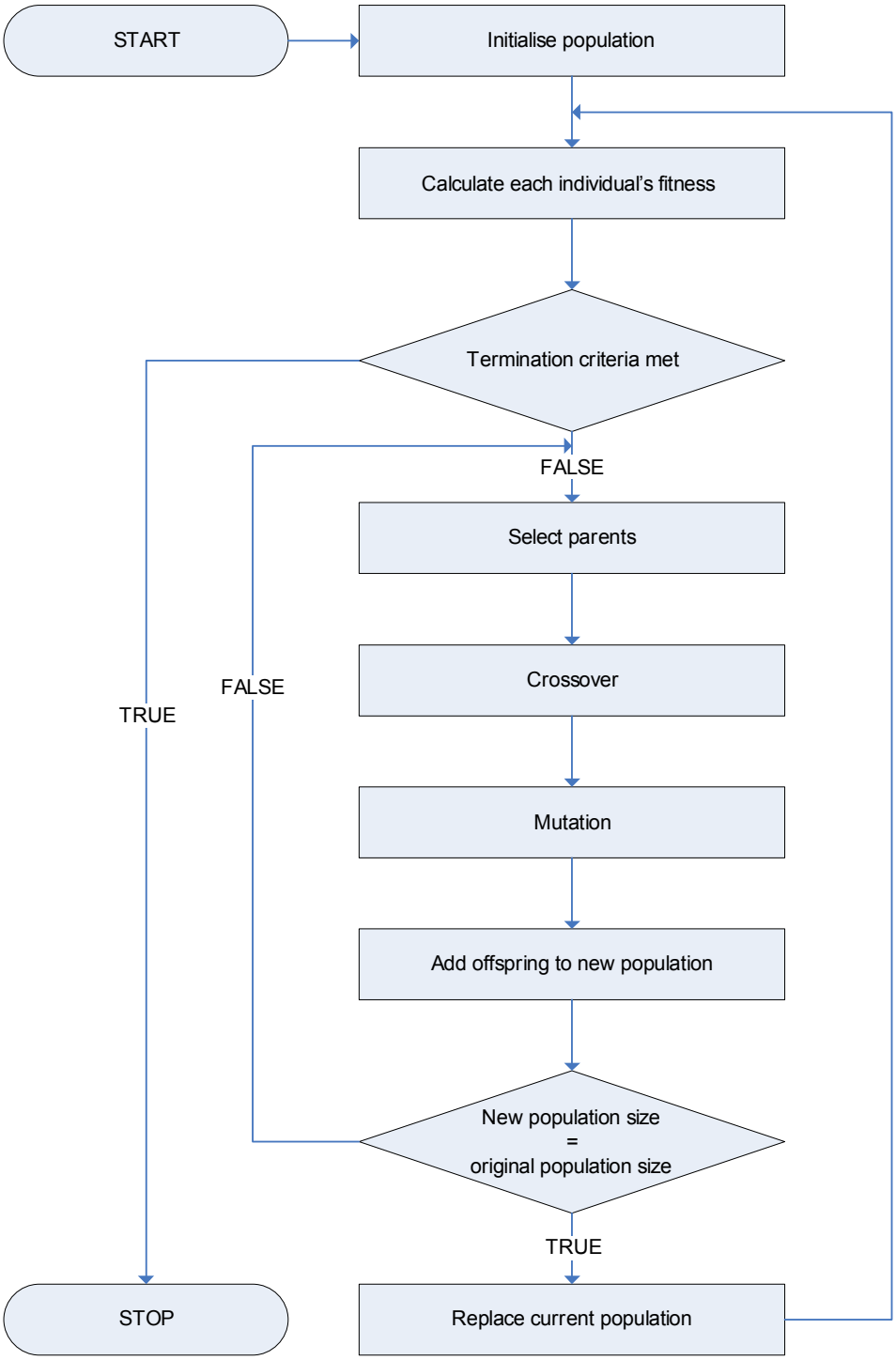

Figure 4.36 Genetic algorithm flow diagram

The most important aspects of the genetic algorithm and the adoptions made to suit the specific problem under investigation are discussed next.

4.7.2.1 Chromosome representation

A very important step in the design of a GA is to find an appropriate chromosome representation. The efficiency and complexity of a search algorithm greatly depend on the representation scheme. Chromosomes have a fixed length and are made up of a number of genes where each gene represents a single parameter to be optimised. The chromosome representation used in this thesis is depicted in Figure 4.37.

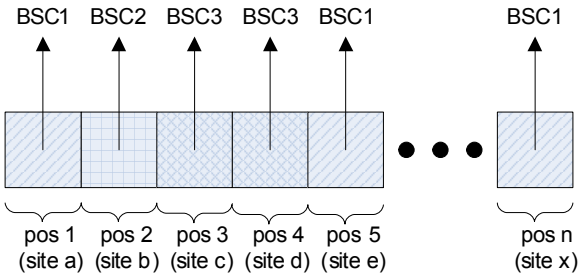

Figure 4.37 Chromosome representation

In the chromosome representation used, the position of the gene in the chromosome correlates to a specific site, while the value of a gene specifies the parent (BSC) of the site.

4.7.2.2 Initial population

Another very important aspect to consider when setting up a GA is the initial population. The standard approach of generating an initial population is to populate the chromosomes with gene values selected at random from the allowed set of values. This ensures that the initial population is a uniform representation of the entire search space. If prior knowledge of the search space and problem is available, heuristics can be used to bias the initial population towards potentially good solutions. There is a danger associated with this approach, namely that certain areas of the search space will not be explored. This might lead to premature convergence of the population to a local minimum.

Two different population initialisation approaches were experimented with. The first approach assigned gene values at random to the chromosomes. These gene values were selected from a set of allowable gene values. The allowable set of gene values contained a list of all the BSCs serving the optimisation area. The advantage of this initialisation technique is that the chromosomes are spread out uniformly in the search space which in theory is good. The major disadvantage is that these randomly initiated chromosomes suffer from the inconsistent cluster phenomena described in section 3.5. The GA thus starts out with a very tough starting point.

The second approach used prior knowledge available to bias the initial population to starting points in the vicinity of the operator's current network configuration. This was achieved by cloning the current network configuration to each individual and then

reassigning 10% of the sites per individual at random to any one of the BSCs serving the optimisation area. With this approach, all the individuals are thus scattered around the current network configuration. The expectation is that the GA will be able to find an improved network configuration not differing dramatically from the current network configuration. Such a solution would represent a viable and realistic solution in the operator's eyes since it can be realised without making dramatic changes to the network's current topology.

4.7.2.3 Fitness function

The purpose of the fitness function is to map a chromosome representation into a scalar value. This scalar value is known as the fitness of an individual (chromosome) and is used to judge the fitness of a solution relative to the other solutions in the population. The fitness of an individual thus defines the quality of its solution. The better a solution is, the higher the fitness associated with that solution will be. It is, therefore, very important that the fitness function accurately models the optimisation problem.

The fitness function should include all criteria being optimised and is also used to enforce constraints. In the event of a constraint being broken, a penalty term is added to the fitness value. Crossover, mutation, selection and elitism make use of the fitness associated with each chromosome. The fitness function is thus possibly the most important component of any evolutionary algorithm.

The fitness function implemented considered the following factors when evaluating individuals.

- Access network transmission cost
- Backhaul network transmission cost
- Core network transmission cost
- BSC constraints
- MSC constraints

An individual's fitness is calculated as being

$$fitness = \frac{1}{networkCost} \tag{4.1}$$

where *networkCost* is calculated as

$$networkCost = accessTX + backhaulTX + coreTX + BSCpenalty + MSCpenalty \tag{4.2}$$

where *accessTX, backhaulTX* and *coreTX* represent the access, backhaul and core transmission costs repectively. *BSCpenalty* and *MSCpenalty* represent the BSC and MSC penalty terms respectively.

4.7.2.4 Selection

Selection is the process used to select individuals from the parent population. These individuals are then used to create offspring using genetic operators such as crossover and mutation. Selection schemes use the fitness of individuals as a decision criterion. There are different types of selection schemes such as random selection, proportional selection, tournament selection and rank-based selection.

A good selection scheme is one that maintains a large enough diversity between the selected individuals. It is important to maintain diversity otherwise premature convergence to a suboptimal solution might occur. A good selection scheme is tournament selection [31]. It insures that the fittest individuals are not repeatedly selected and also that the most unfit individuals are not selected.

During tournament selection, a group of k individuals is selected at random each time an individual needs to be picked. From this group, an individual with a high fitness value is more likely to be selected than an individual with a low fitness value. For this reason, tournament selection was used as the selection scheme.

4.7.2.5 Crossover

During crossover, genetic material of the parent chromosomes is exchanged to create offspring. Parents selected for crossover do not always produce offspring. Crossover happens at a certain probability, called the crossover rate. A value of 0.8 is usually used as the crossover rate. During crossover, a certain crossover point is selected at random. This point defines the position in the chromosomes at which crossover will occur. The crossover process is illustrated in Figure 4.38.

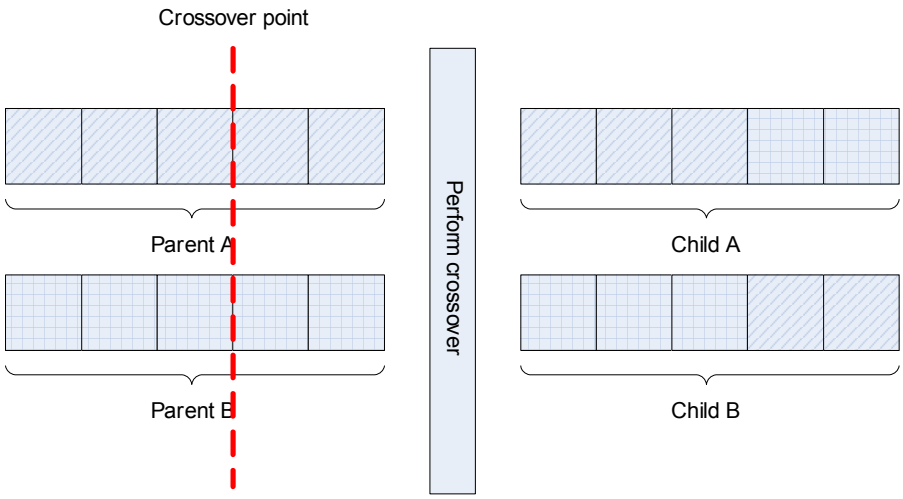

Figure 4.38 Crossover illustration

A crossover rate of 0.8 was used in this study. By implication a selected pair of parents thus only has an 80% chance of producing offspring. The crossover scheme explained above is known as single-point crossover and was used as the crossover operator in this study.

4.7.2.6 Mutation

Mutation adds new genetic material into an individual thus adding diversity to the population. Through the combination of mutation and crossover the entire search space is accessible. Mutation is usually performed on unfit individuals. As with crossover, not all of the offspring is exposed to mutation. Mutation happens at a certain probability called the mutation rate. Different mutation schemes exist and the choice of mutation scheme is highly problem-dependent. Two of the most popular schemes are random mutate and in-order mutate. In random mutate, each gene is considered separately for mutation. In in-order mutate, two random genes are selected and all the genes between then are mutated.

The mutation scheme implemented in this project worked as follows, only chromosomes with a low fitness were exposed to mutation. These chromosomes then had a 10% probability (mutation rate) of being mutated. Mutation was applied by randomly altering 5% of the chromosome's gene values. New gene values were selected from a list of allowable gene values. These allowable gene values were the BSCs serving the optimisation area.

4.7.3 Results and discussion

Section 4.7.3.1 and section 4.7.3.2 below detail the population fitness per generation of the different population initialisation schemes investigated. Results produced from both initialisation schemes highlights the complexity and size of the search space. Neither of the initialisation techniques was able to produce feasible network configurations. The network

configurations found were infeasible due to the lack of consistent clusters as can be seen in Figure 4.40 and Figure 4.42.

The GA was selected for the global optimisation characteristics it possesses. GAs are much less likely to get stuck in local minima than the heuristic search techniques explored earlier in the chapter. This global optimisation characteristic is one of the GA's biggest strengths.

For the GA to effectively explore the search space and generate high quality solutions, it is extremely important to use a chromosome representation that reflects the characteristics of the optimisation problem. Using an appropriate chromosome representation, thus plays a pivotal role in the effectiveness of the genetic algorithm as a global optimisation technique.

One very important characteristic of the cell-to-switch allocation problem for mobile networks is the cluster consistency requirement. This requirement is not emphasised enough in the current chromosome structure and hence inconsistent clusters were formed. This requirement can be dealt with by using a chromosome representation that specifically reflects the consistent cluster requirement. Alternatively, the current chromosome representation can be used but with an adopted fitness function that includes the cluster consistency requirement as one of the constraints.

4.7.3.1 Uniform population initialisation

The uniform population initialisation scheme initialised each gene of each chromosome to a value selected at random from a list of allowable gene values. These gene values represent the BSCs serving the optimisation area. The effect this had was that the individuals were distributed evenly throughout the search space, thus potentially making any region accessible.

From this experiment, the researcher wanted to see whether the GA is capable of increasing the average fitness of the initial population over a number of generations and whether the found clusters were consistent.

Figure 4.39 indicates the average population fitness as well as the fitness of the best individual in the population per generation. The fitness value used was the total network cost consisting of the access, backhaul and core transmission costs as well as the cost of penalties due to broken constraints. The population started out with a relatively high total network cost which the GA was able to improve on (reduce) with time. Convergence was achieved after 100 generations. The network cost (entire network) of the best-found individual was, however, 2.07% higher than the operator's current network cost. In addition to the increased cost, the clusters formed were not consistent thus making the solution infeasible as illustrated in Figure 4.40.

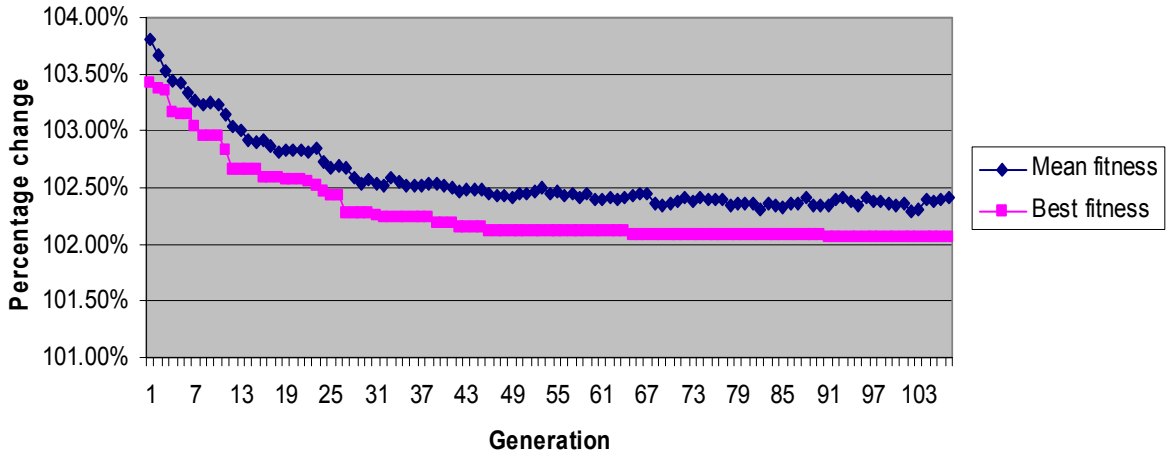

Figure 4.39 Population fitness (entire network)

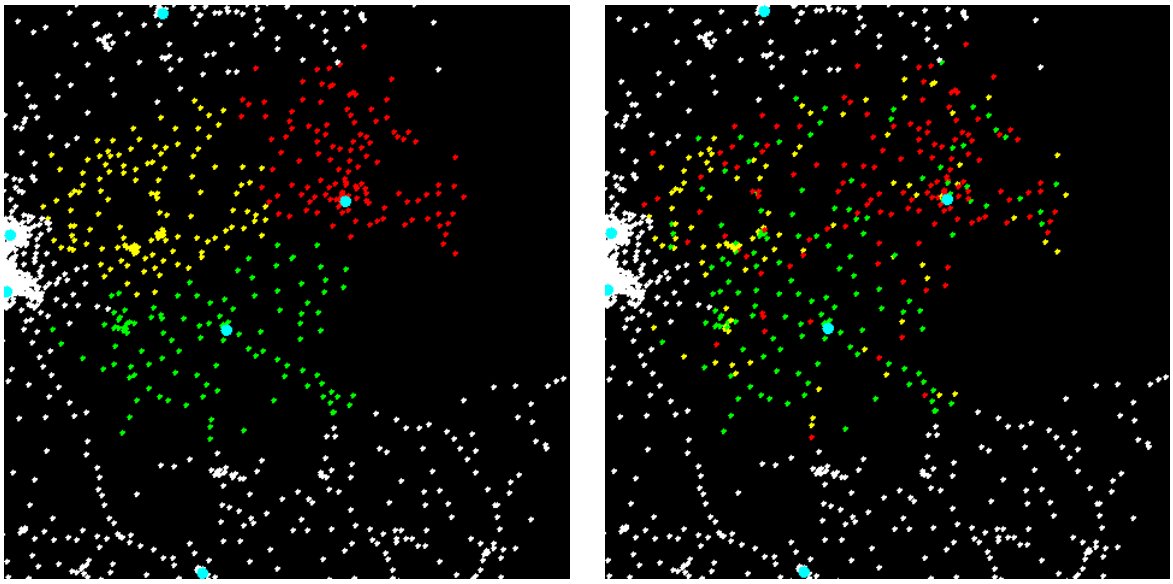

Figure 4.40 Operator's original cluster formation (left), final cluster formation found by GA (right)

% Change in traffic	Inter-BSC traffic	Inter-SC traffic
Optimisation region	162.63%	39.52%
Entire network	1.03%	1.09%

Table 4.16 Traffic measurements

% Change in transmission cost	Access transmission	Backhaul transmission	Core transmission	Total transmission
Optimisation region	20.30%	15.97%	-	-
Entire network	1.83%	2.23%	0.47%	1.38%

Table 4.17 Transmission cost

ID	TRX utilisation original	TRX utilisation optimised	TRA utilisation original	TRA utilisation optimised	Unit utilisation original	Unit utilisation optimised
ERBSC1	57.29%	70.96%	-	-	57.29%	70.96%
NSBSC1	75.52%	69.01%	-	-	75.52%	69.01%
PTBSC3	58.66%	53.75%	87.88%	80.30%	73.27%	67.03%

Table 4.18 BSC utilisation

4.7.3.2 Existing topology-based population initialisation

The second population initialisation scheme used the operator's current network configuration as a starting point and randomly altered 10% of the genes in the optimisation area of each individual. Altered gene values were again picked from a list containing only the BSCs serving the optimisation area. The effect this had was that the entire population was initialised to positions in the search space near the operator's current network configuration. The entire search space is thus not accessible any more.

From this experiment, the researcher wanted to determine whether the GA is capable of improving on the operator's current configuration without making massive changes to the network's current topology. Secondly, the researcher wanted to see whether the newly found solutions had feasible cluster formations.

The initial population started out with an average network cost of 0.8% higher than the operator's current network cost. The population's mean fitness rapidly decreased (cost of network increased) after the first few generations. This can be contributed to the following fact. Initially, only 10% of the genes per individual differed from the operator's current network configuration. This 10% change was limited to any randomly selected genes per individual within the optimisation area. Each individual thus had its own unique combination of genes that differed from that of the original configuration. The crossover operator creates offspring by exchanging the genetic material of two parents. If each parent thus had a unique combination of sites differing from the original configuration, the children of those two parents are expected to have even more sites differing from that of the original configuration. This causes the diversity of the population to increase after each generation. The increase in diversity goes hand in hand with the decreased mean population fitness as seen during the first 10 generations of Figure 4.41.

The GA was unable to improve the initial population's mean fitness using this initialisation technique. The best-found individual's network cost (entire network) was 0.65% higher than that of the operator's current network cost. Looking at Figure 4.42, it can be seen that the best-found individual also had inconsistent clusters causing the solution to be infeasible.

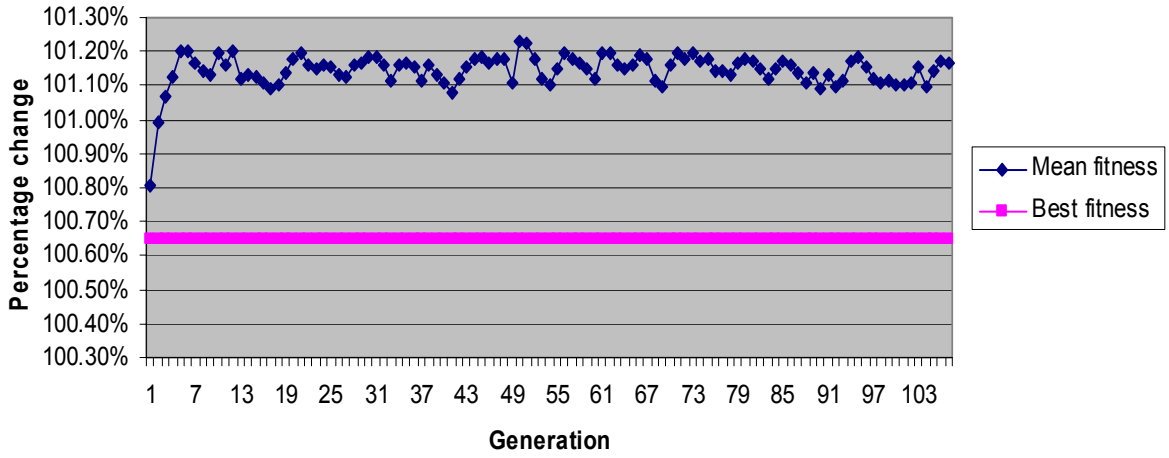

Figure 4.41 Population fitness (entire network)

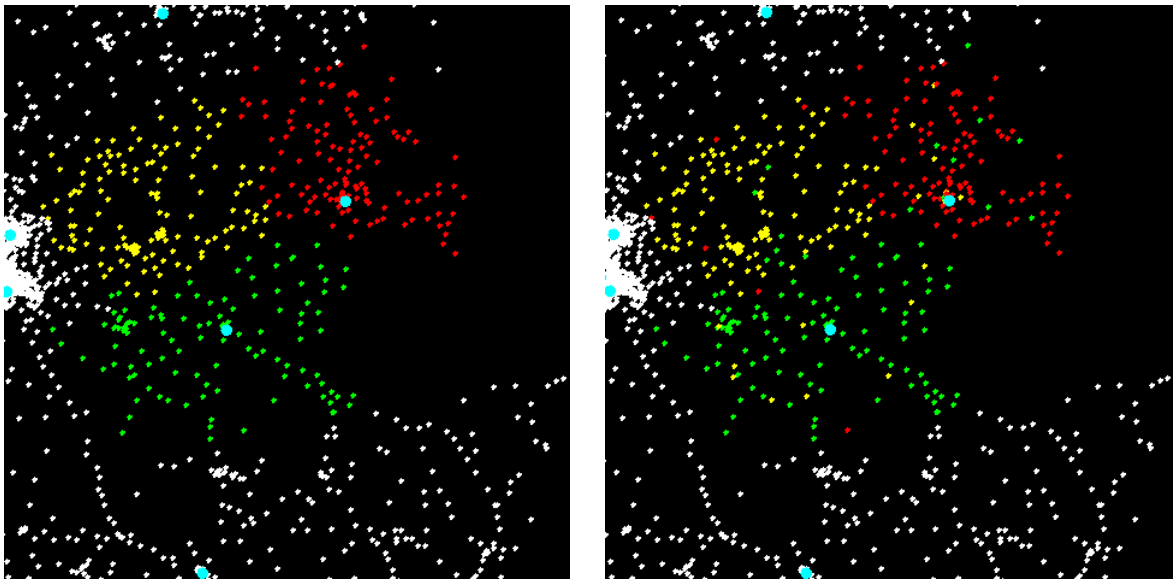

Figure 4.42 Operator's original cluster formation (left), final cluster formation found by GA (right)

% Change in traffic	Inter-BSC traffic	Inter-SC traffic
Optimisation region	41.83%	5.93%
Entire network	0.27%	0.26%

Table 4.19 Traffic measurements

% Change in transmission cost	Access transmission	Backhaul transmission	Core transmission	Total transmission
Optimisation region	10.08%	-8.28%	-	-
Entire network	0.91%	-1.15%	0.00%	0.49%

Table 4.20 Transmission cost

ID	TRX utilisation original	TRX utilisation optimised	TRA utilisation original	TRA utilisation optimised	Unit utilisation original	Unit utilisation optimised
ERBSC1	57.29%	56.51%	-	-	57.29%	56.51%
NSBSC1	75.52%	72.14%	-	-	75.52%	72.14%
PTBSC3	58.66%	61.52%	87.88%	89.02%	73.27%	75.27%

Table 4.21 BSC utilisation

4.8 Conclusion

The experiments conducted in this chapter had two imported goals. Firstly, to explore the complexity of the search space and, secondly, to determine the extent, if any, by which the current network configuration can be improved on.

Results obtained from the heuristics indicated that the search space is extremely complex. The search space can be thought of as consisting of two components, namely traffic and transmission cost. When optimising only one of the components, the other is affected negatively. It is thus very important to consider both components when trying to find optimal solutions.

The network operator's current configuration was used as a starting point in many of the heuristics. This was mainly done because in practice this is the most likely starting point any PCS network optimiser will have. Cutting a massive amount of sites over is not a viable solution for most operators. A more acceptable solution would be one where a smaller amount of changes are made to current network configuration.

Heuristic 4 showed that it is possible to achieve significant improvements on the current network configuration. A saving of 10.94% on the access network transmission cost was achieved in the optimisation region. This translated into a 0.6% saving made on the network-wide transmission cost. By optimising only 7.6% (385 out of the 5 000 sites) of the network, a total transmission cost saving of 0.6% was achieved. Heuristic 1 and 2 highlighted the importance of taking both the traffic and transmission cost factors into account when optimising.

Results obtained from the genetic algorithm highlighted the importance of using a chromosome representation that reflects the characteristics of the optimisation problem. The chromosome representation used did not place enough emphasis on the cluster consistency requirement and hence was unable to produce high quality feasible solutions.

5 UNSUPERVISED SEARCHES

5.1 INTRODUCTION

Unsupervised learning is the study of how systems can learn to represent input patterns in a way that reflects the statistical structure of the overall collection of input patterns. By contrast, with supervised or reinforcement learning, there is no explicit targets or environmental evaluations associated with each input; rather the unsupervised learner brings to bear prior biases as to what aspects of the structure of the input should be captured in the output [32].

With unsupervised learning, representations of the input data is constructed and used for decision making. Patterns in the data above and beyond what would be considered pure unstructured noise are thus discovered. Clustering and dimensionality reduction are two classic examples of unsupervised learning.

5.2 CLUSTERING (UNSUPERVISED LEARNING)

Clustering algorithms are referred to as unsupervised learning algorithms because they do not use class labels to separate data. Instead, they seek to segment the data based on structure inherent in the data.

Clustering is a technique that attempts to segment objects into groups with the goal of forming groups such that objects in a group are more similar to one another than to objects in other groups. One such a group can be regarded as a cluster.

It should be noted that there are nearly as many classifications of clustering algorithms as there are clustering algorithms. The classification of clustering algorithms presented below is from the density estimation point of view. Three approaches to density estimation are.

- Parametric methods
- Non-parametric methods
- Semi-parametric methods

With parametric methods, a specific functional form for the density model is assumed. The model is defined by a number of parameters optimised by fitting the model to the data. The main drawback with parametric methods is that the form of the density function is decided on before the optimisation process starts. In situations where the actual form of the density function is unknown, an inappropriate density form incapable of providing a good representation of the true density might be selected.

Non-parametric optimisation does not assume a predefined functional form, the form of the density function is determined entirely by the data. The main drawback of non-parametric techniques is that the number of parameters in the model depends on the size of the dataset, not the complexity of the dataset. Large datasets can thus easily lead to large unwieldy non-parametric models.

Semi-parametric models try to achieve the best of both worlds by allowing a very general class of functional forms where the number of parameters depends on the complexity of the density function being modelled and not the size of the dataset. The class of semi-parametric models used in this project for density estimation is known as mixture models.

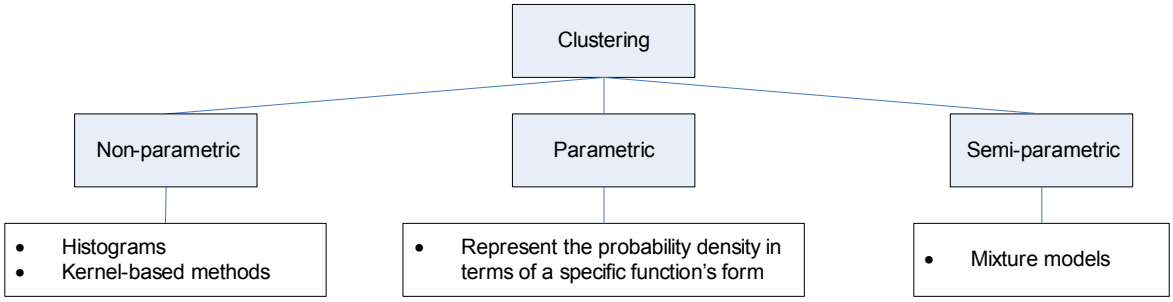

Figure 5.1 Probability density based clustering algorithm classification

The following aspects have to be thought through before selecting a specific clustering technique.

- What number of classes should be searched for
- Should all the attributes of a data point be used or only a subset
- What useful information can be extracted from the discovered clusters

As with most data mining and artificial intelligence techniques, domain knowledge is important and can be used to guide initial parameter values. In this project, the number of clusters required is governed by the number of BSC areas being optimised. It should be noted that when optimising a region covered by, for example, three BSCs, clustering should be performed based on four or five clusters. Superior cluster separation was achieved in this way. The extra clusters are manually merged afterwards until only the desired number of clusters is left.

The attributes to use are also to a large extent guided by domain knowledge. There is, however, a number of techniques available to assist in the parameter selection process, for example, determining the correlation between parameters or determining the measure of dependence between parameters.

When analysing the clusters formed, it is always very important to keep in mind the attributes used to perform clustering on, the pre-processing performed on the attributes, if applicable, as well as the distance function. For example, clustering based on the traffic between sites produced a set of clusters very different to the set produced when clustering was based on intersite distance.

5.2.1 Overview (GMMs)

This project used a Gaussian mixture model as the clustering algorithm. A short description of the main principles behind GMMs is presented below. A more detailed explanation can be found in [33].

In order to understand Gaussian mixture models, it helps to understand the problem they help to solve, namely density estimation. Density estimation can be described as follows.

Given a set of N points in D dimensions, $x_1, x_2, \ldots, x_n \in R^D$, and a family F of probability density functions on R^D, find the probability density $f(x) \in F$ that is most likely to have generated the given points. One way to define the family F is to give each of its members the same mathematical form and to distinguish different members by different values of a set of parameters θ [34]. In the case of Gaussian functions, θ would represent the set of parameters specifying the mean (m), standard deviation (σ) and mixing probability (p) of each function. Assuming there are K functions, F can thus be represented as

$$f(x;\Theta) = \sum_{k=1}^{K} \mathrm{p}_k g(x;m_k;\sigma_k) \tag{5.1}$$

where

$$g(x;m_k;\sigma_k) = \left(\frac{1}{\left(\sqrt{2\pi}\sigma_k\right)^D}\right) * \left(e^{-\frac{1}{2}\left(\frac{\|x-m_k\|}{\sigma_k}\right)^2}\right) \tag{5.2}$$

Mixtures of Gaussians are well suited to modelling clusters of points. When using Gaussians to model clusters, a Gaussian is typically assigned to a cluster with the mean of the Gaussian near the centre of the cluster. The spread of the cluster is modelled by the standard deviation of the Gaussian. The likelihood of a data point being generated by a specific process is indicated by the mixing probability (prior probability).

Density estimation using mixture models can thus be seen as the process of finding the vector of parameters θ that specifies the model from which the points were most likely drawn. In principle, any maximisation algorithm can be used to find θ. The expectation maximisation (EM) algorithm is an efficient and effective way of estimating θ. The EM algorithm operates within the maximum likelihood framework, a short outline of the EM algorithm is presented next.

EM is best explained with the assistance of an example. Suppose a set of data points is generated by two processes. The idea is then to use a mixture of Gaussians to represent the two processes and use EM to estimate the parameters of the GMM.

The negative log likelihood of a GMM for a specific data set serves as the error function that needs to be minimised. The values of the GMM's parameters (θ) can be found by iteratively performing the two steps of the EM algorithm (Figure 5.2) which guarantee convergence to a local minimum [33]. During step 1, the E-step, data points are assigned to the process that fits it best. During step 2, the M-step, the parameters of the processes (Gaussians) are updated using only points assigned to it. The negative log likelihood of the GMM (equation 5.1) can be expressed as equation 5.3 assuming there are N data points.

$$\lambda(X;\Theta) = -\sum_{n=1}^{N} \ln \sum_{k=1}^{K} \mathrm{p_k}\, g(x_n; m_k; \sigma_k) \tag{5.3}$$

The intuition behind EM is that each of these steps is easy to perform assuming the other step is solved. In other words, assuming we know the assignment of each datapoint, we can easily estimate the parameters of each process by only using the points assigned to the specific process. Likewise, if we know the parameters of each process (Gaussian) we can assign each point to the process (Gaussian) that most likely generated it.

The basic structure of the EM algorithm can be summarised as depicted in Figure 5.2.

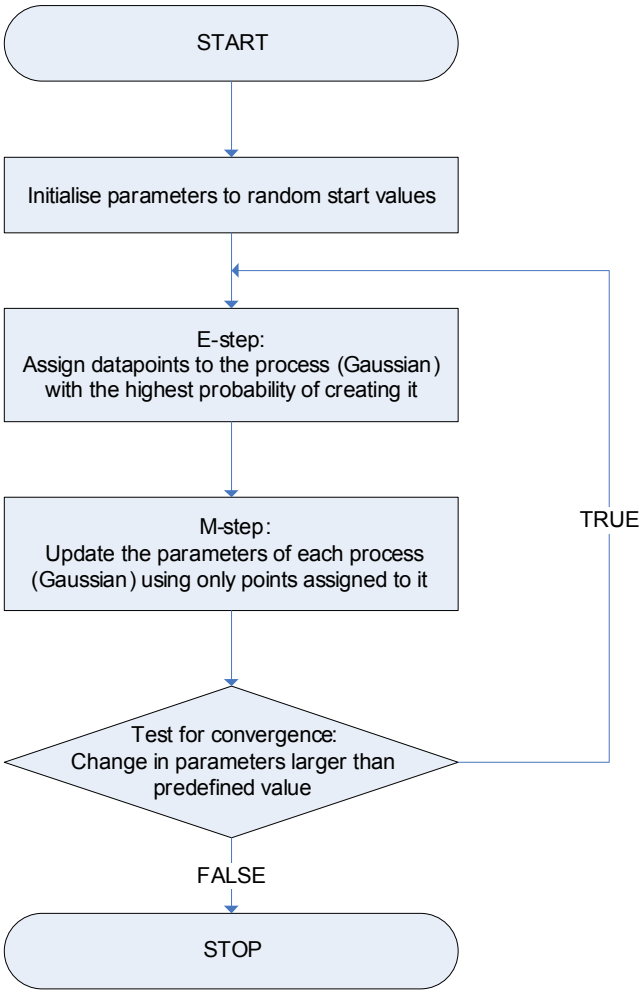

Figure 5.2 EM algorithm flow diagram

5.2.2 Implementation

Clustering, using a GMM, was implemented using two programs. Firstly, the GSM simulator written in the Java programming language was used for generating the distance and traffic matrixes used in the GMM, as well as evaluating the clusters found by the GMM. The GMM was implemented in Matlab using the freely available Netlab [35] toolbox. A detailed overview of the main areas of the clustering process is presented next.

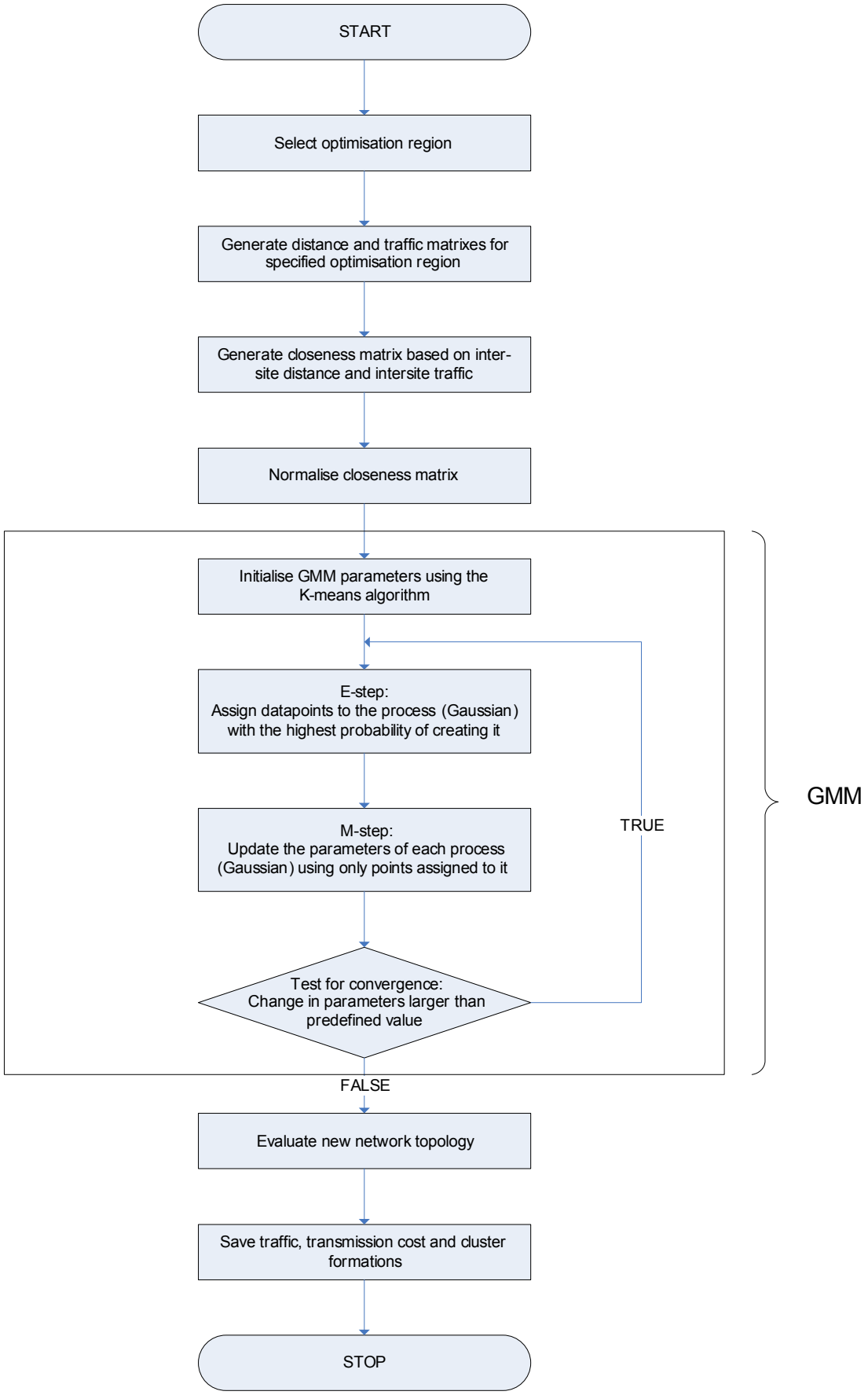

Figure 5.3 GMM clustering implementation

5.2.2.1 Distance measure

One of the most important aspects to consider in most clustering algorithms is the measure of similarity (closeness) between objects. The measure of similarity is a way of identifying objects that have several features in common. In this implementation, two objects (sites) that are very similar will have a high closeness measure while objects that do not have many features in common will have a low closeness measure.

When trying to generate clusters of sites with the end goal of reducing OPEX, two, often opposing, optimisation goals need to be solved. Firstly, the access cost of the network needs to be reduced as much as possible since this will introduce large savings in OPEX. In an extreme scenario, this can be achieved by connecting a site to its closest BSC. This, however, breaks the second optimisation criteria. Secondly, reducing the amount of traffic in the core network will free up expensive high bandwidth links which also translates into an OPEX saving. Optimising with the goal of only reducing the traffic in the core network is also not a viable option since infeasible configurations such as the ones discussed in Section 3.5 are generated. From an optimisation point of view, it is thus clear that the access transmission cost, as well as core network traffic criteria, needs to be optimised concurrently.

This was achieved by creating a closeness measure that takes both optimisation criteria into account. The GSM simulator was used to generate two tables specific to the sites in the optimisation region. Both tables had the same dimension. The first table indicated the amount of traffic in seconds from each site to every other site in the optimisation region, while the second table contained the intersite distances for all the sites in the optimisation region. The similarity of two objects was measured as follows.

$$closeness_{ij} = \alpha_{ij} + \beta_{ij} \tag{5.4}$$

$$closeness_{ij} = \frac{traffic_{ij}}{90thPrctileTraffic_i} + \frac{10thPrctileDistance_i}{distance_{ij}} \quad (5.5)$$

Equation 5.4 defines the closeness of sites *i* and *j*. Note that $closeness_{ij}$ is not necessarily equal to $closeness_{ji}$. The first term, alpha, is used to relate the connection strength of two sites based on the amount of traffic between them. Beta, the second term, defines the connection strength of two sites based on their geographical distance from each other. The combination of these two terms gives the closeness of two sites. Large values indicate "close" sites and small values indicate "distant" sites.

The idea behind the alpha term was to create a term that would be larger than one for sites with a substantial amount of traffic between them and less than one for sites with only a small amount of traffic between them. The alpha term between site *i* and *j* is calculated by dividing the total amount of traffic between the two sites by the 90th percentile traffic value of site ***i.*** This percentile value was calculated by generating a list of all the sites to/from which site *i* sent/received traffic. The 90th percentile was then calculated from this subset. The reason for doing it this way is that all the sites with zero traffic skewed the result. By ignoring the zero traffic sites, a more realistic value was obtained.

The beta term creates a term that is larger than one for sites that are geographically close to each other and less than one for sites that are far apart. The beta term between sites *i* and *j* is calculated by dividing the 10th percentile distance of site ***i*** by the distance between site *i* and ***j.*** This percentile value was calculated by generating a list of distances from site *i* to all the other sites being optimised. The 10th percentile was then calculated from this list.

5.2.2.2 Pre-processing

There was, however, a few extremely large outliers present in both the alpha and beta terms of equation 5.4. These outliers negatively affect the clustering ability of the GMM. To deal with the outliers, pre-processing was performed on the intersite closeness measures. This was achieved by taking the logarithm of each term in equation 5.5 before adding them together. It had the effect of bringing outliers in while still maintaining separation between the non-outlier measurements. Once the logarithm of all the alpha and beta terms were taken, they were rescaled to ensure that they were between zero and one. The scaling technique just mentioned was the outcome of several experiments. The experiments together with their results are presented next.

5.2.2.2.1 Alpha and Beta in the unscaled form

Figure 5.4 indicates the optimisation region under consideration as well as the current BSC boundaries. Three BSCs are included in this region namely, Umtata BSC1, Umtata BSC2

and East London BSC1. In total, there are 283 sites in the optimisation region. The final closeness table is a 283 x 283 matrix indicating the closeness between all the sites. There is thus a closeness measure from every site to every other site (including itself) in the optimisation region. These closeness measures are used as features by the GMM during the clustering process. This specific optimisation region represents a 283 dimensional search space.

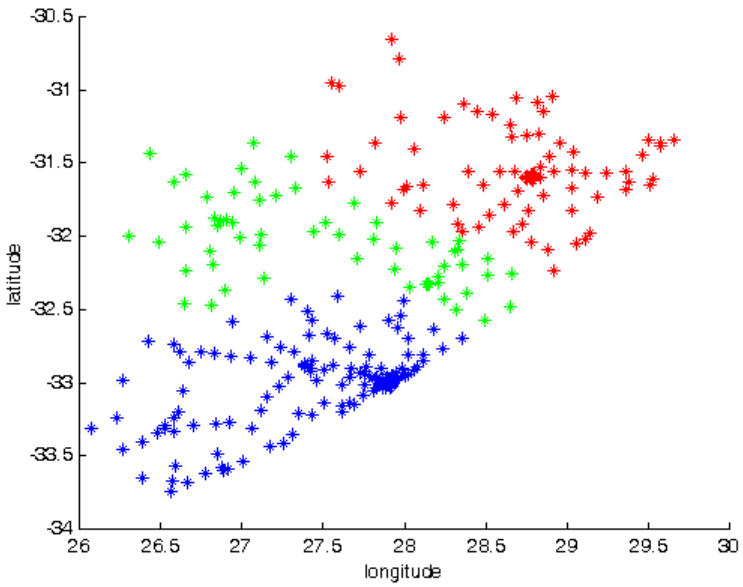

Figure 5.4 GMM optimisation region indicating the current BSC boundaries

The alpha and beta terms in their unscaled form as well as their distributions are presented below. The sections to follow are used to indicate the effect of different pre-processing techniques on the distributions of the alpha and beta terms and to illustrate the effect this has on the clusters formed by the GMM.

$$\alpha_{ij} = \frac{traffic_{ij}}{90thPrctileTraffic_i} \tag{5.6}$$

$$\beta_{ij} = \frac{10thPrctileDistance_i}{distance_{ij}} \tag{5.7}$$

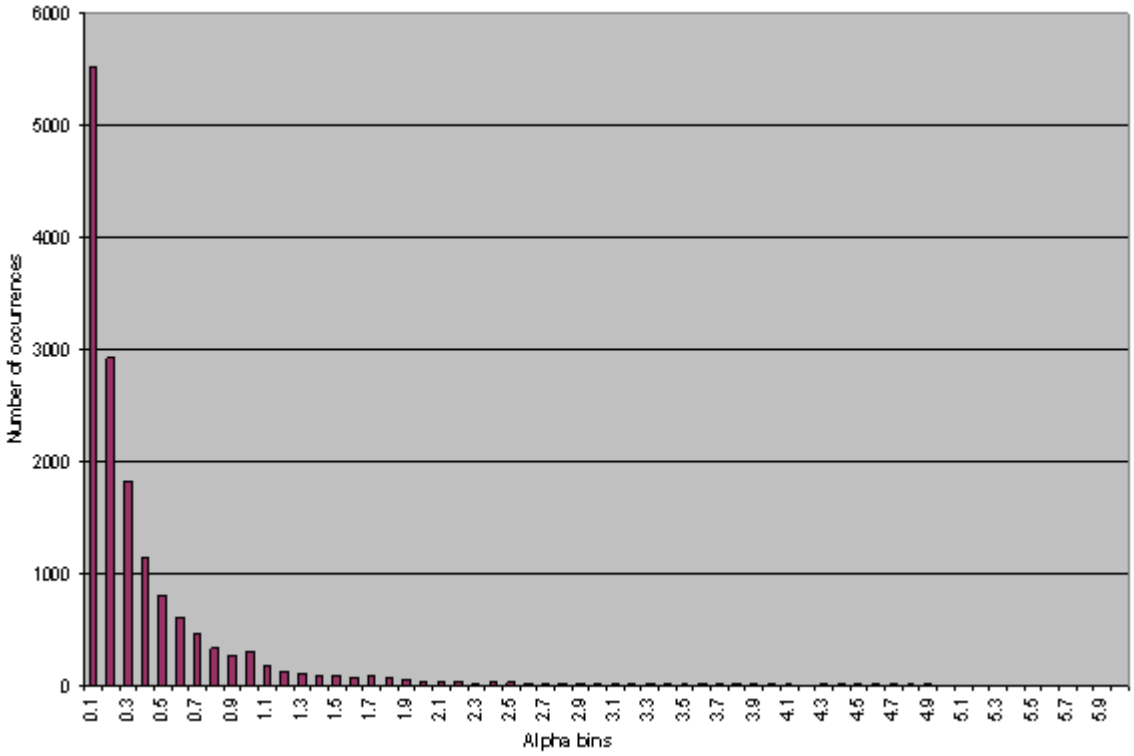

Figure 5.5 Unaltered alpha distribution

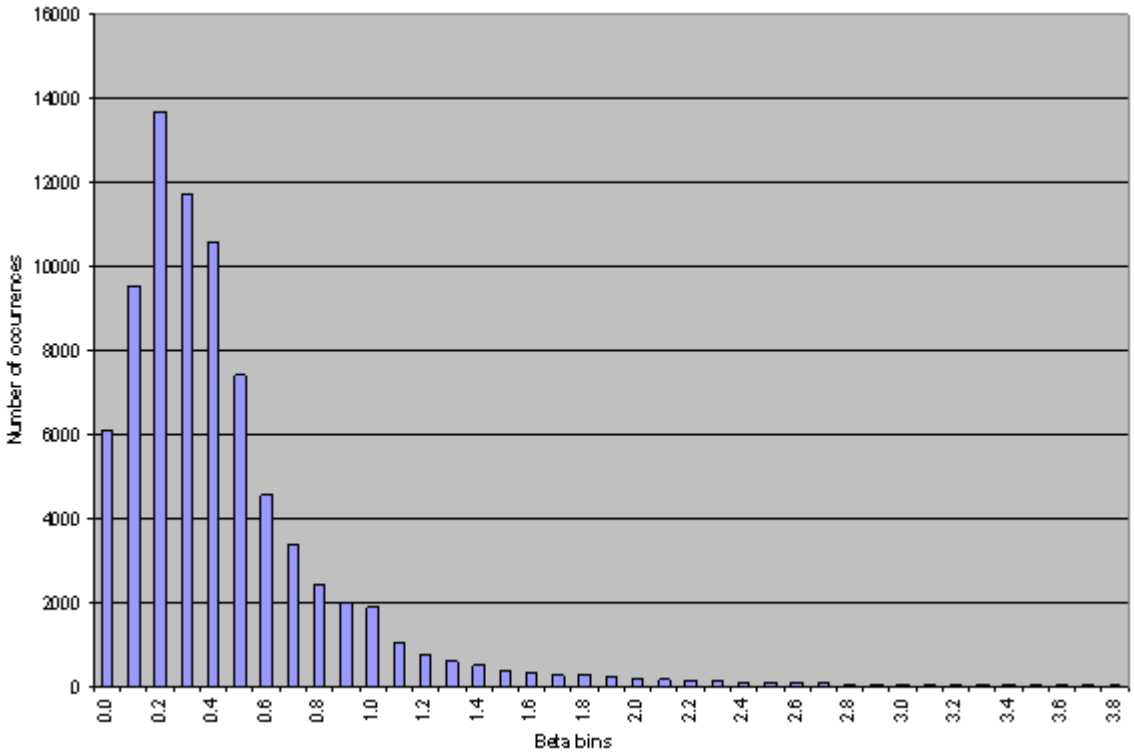

Figure 5.6 Unaltered beta distribution

The following cluster formations displayed were the two most often encountered when not altering the alpha and beta terms.

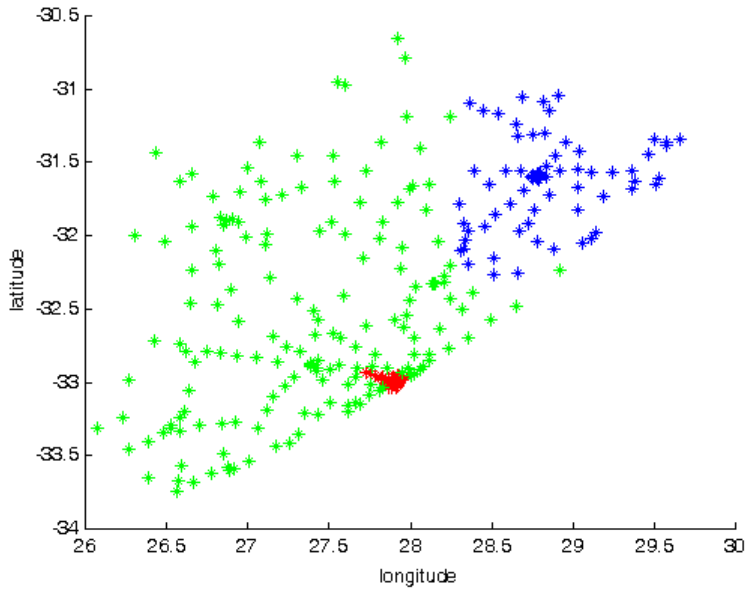

Figure 5.7 Unaltered alpha and beta cluster formation 1

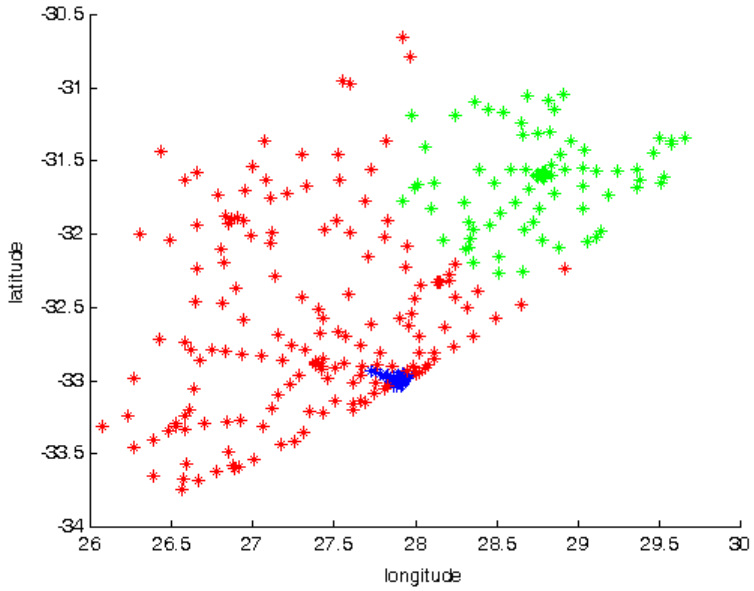

Figure 5.8 Unaltered alpha and beta cluster formation 2

5.2.2.2.2 Hard-limited scaling technique

The hard-limited scaling technique introduced the following two rules.

- If $\alpha_{ij} > 1$ then $\alpha_{ij} = 1$ else $\alpha_{ij} = \alpha_{ij}$
- If $\beta_{ij} > 1$ then $\beta_{ij} = 1$ else $\beta_{ij} = \beta_{ij}$

Figure 5.9 and Figure 5.10 illustrate the new alpha and beta distributions after applying the hard-limited technique.

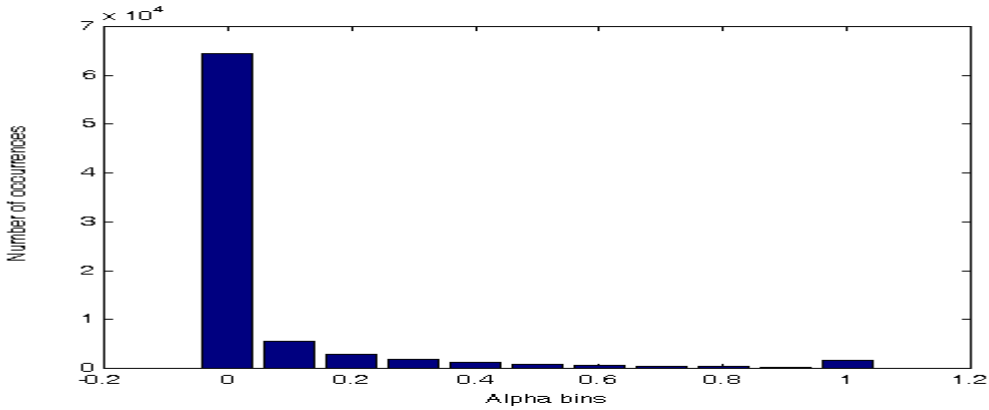

Figure 5.9 Hard-limited alpha distribution

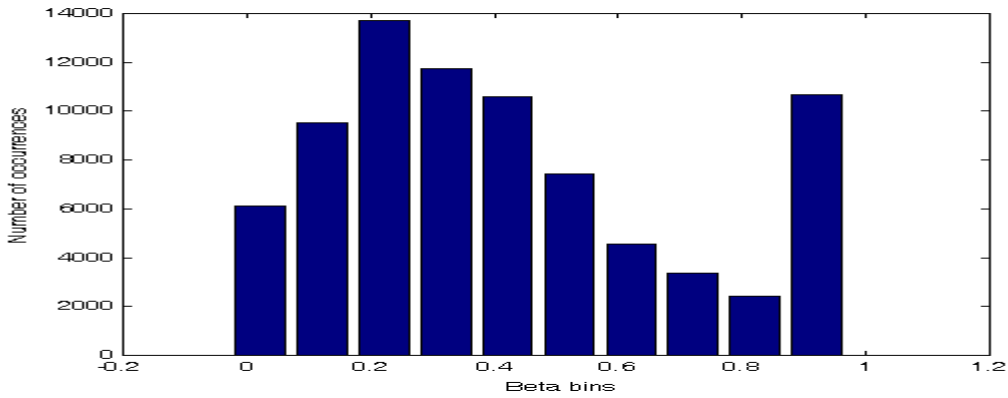

Figure 5.10 Hard-limited beta distribution

The following cluster formations were the ones most often encountered when using the hard-limited scaling technique on the alpha and beta terms.

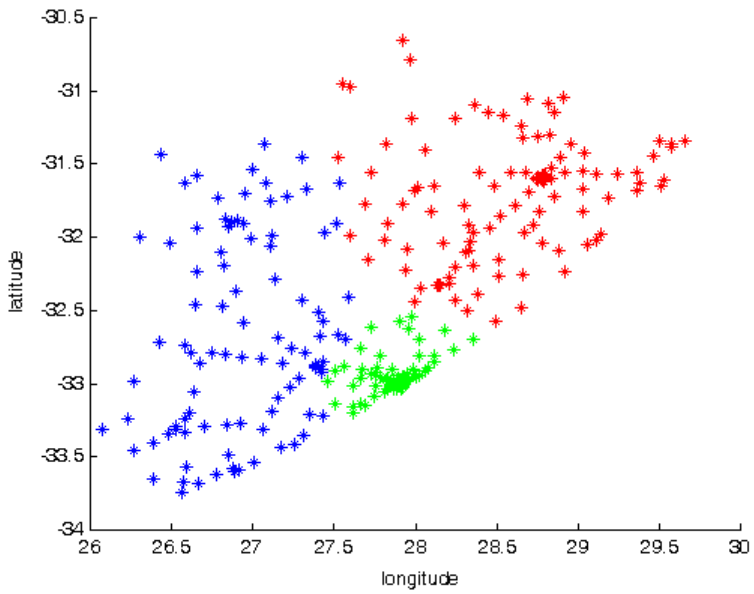

Figure 5.11 Hard-limited alpha and beta cluster formation 1

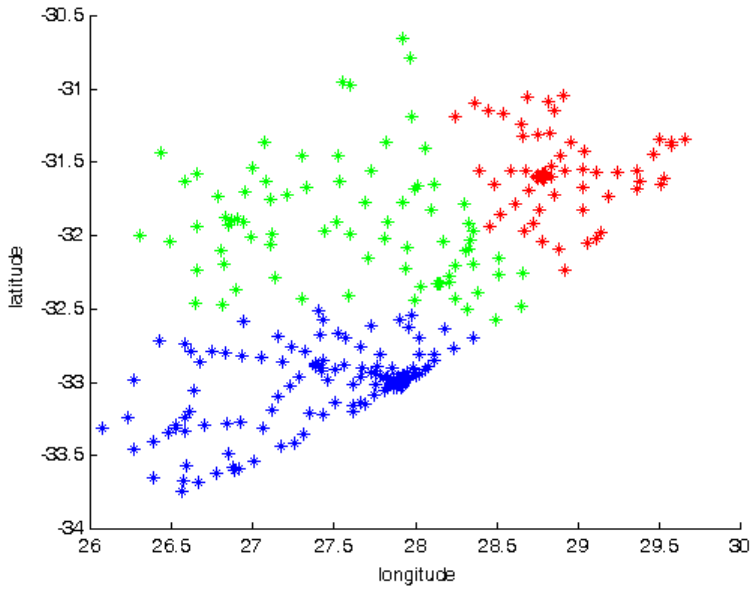

Figure 5.12 Hard-limited alpha and beta cluster formation 2

5.2.2.2.3 Logarithm-based scaling technique

This technique involved taking the logarithm of the alpha and beta terms as well as rescaling the result afterwards to ensure that the end-result is a value between 0 and 1.

To deal with the outliers, the logarithm of each term was taken before adding them together. This had the effect of bringing outliers in while still maintaining separation between these measurements. Once the logarithm of all the alpha and beta terms was taken, they were rescaled to ensure that only values between 0 and 1 were produced as output.

The alpha term was calculated as follows.

- Calculate the $\ln(10 * \alpha_{ij})$ value for all the alpha values.
- Rescale all the new alpha values to ensure that they are between 0 and 1. This was achieved by

$$\alpha_{ij} = \frac{\alpha_{ij} - \min(\alpha_i)}{\max(\alpha_i) - \min(\alpha_i)} \tag{5.8}$$

Similarly the beta term was calculated as.

- Calculate the $\log_{10}(\beta_{ij})$ value for all the beta values.
- Rescale all the new beta values to ensure that they are between 0 and 1. This was achieved by

$$\beta_{ij} = \frac{\beta_{ij} - \min(\beta_t)}{\max(\beta_t) - \min(\beta_t)} \tag{5.9}$$

Figure 5.13 and Figure 5.14 illustrate the effect logarithm-based scaling had on the alpha and beta terms.

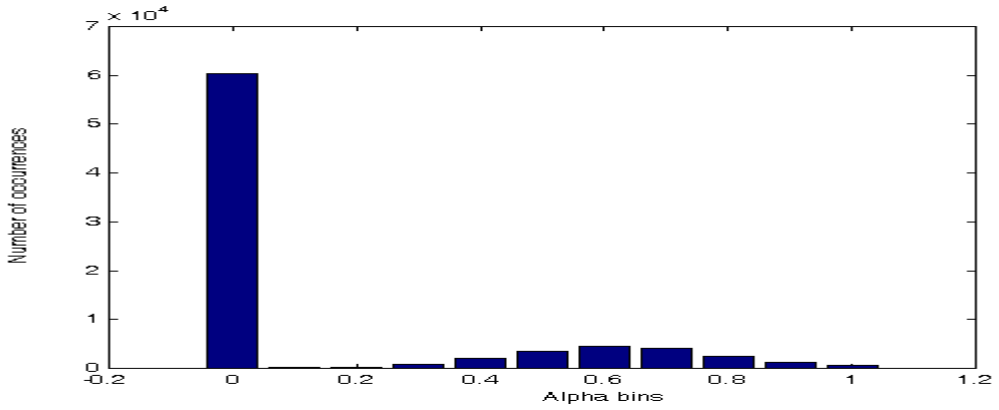

Figure 5.13 Logarithm-based alpha distribution

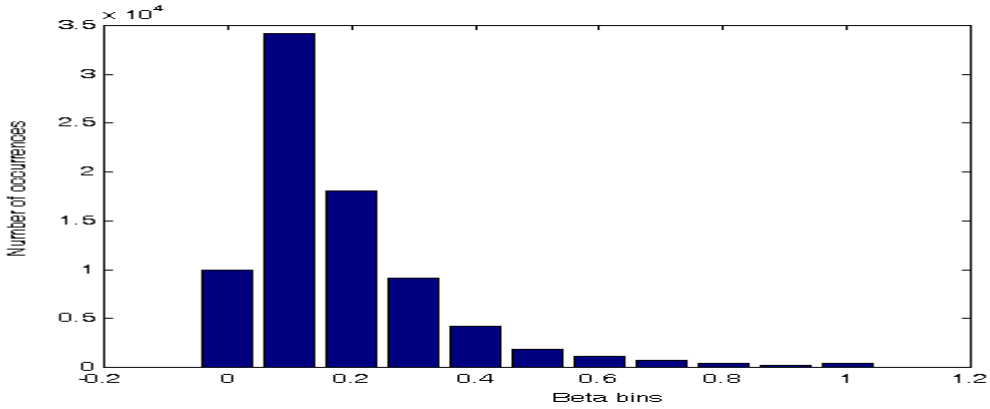

Figure 5.14 Logarithm-based beta distribution

The following cluster formations were most often encountered when using the logarithm-based approach to scale the alpha and beta terms.

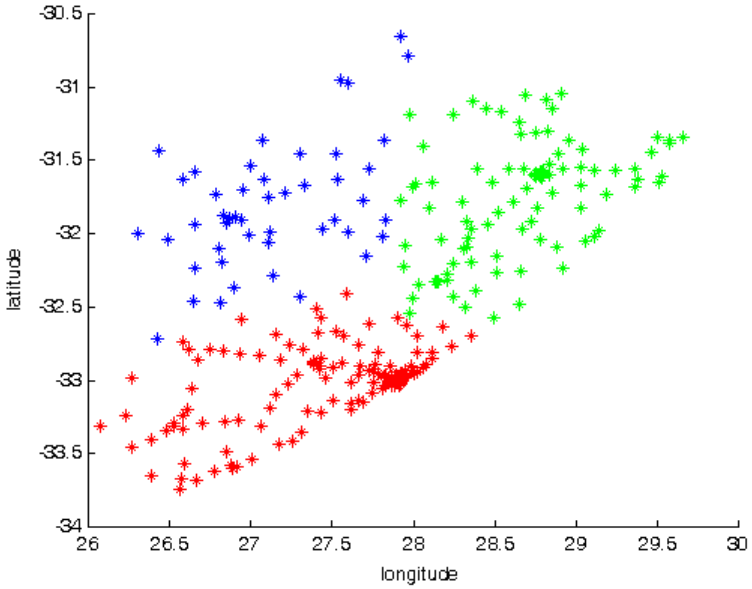

Figure 5.15 Logarithm-based alpha and beta cluster formation 1

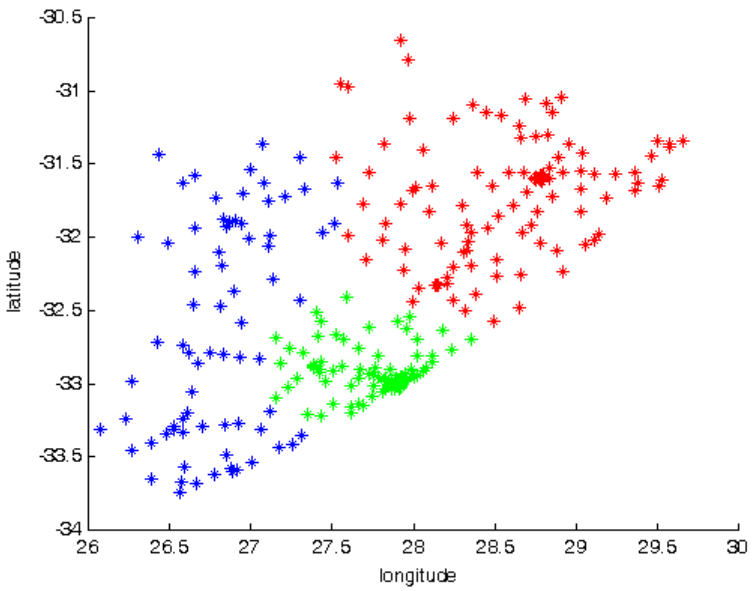

Figure 5.16 Logarithm-based alpha and beta cluster formation 2

5.2.2.3 Initialisation

Since a GMM is guaranteed to converge on a local minimum, the initial starting point has a large impact on the quality of solution (cluster formation) found. For this reason, the K-means algorithm was used to determine the initial cluster centres of the Gaussians. The mixing coefficients (prior probabilities) of each cluster were computed from the proportion of samples assigned to each cluster. A brief overview of the K-means algorithm is presented next.

The K-means algorithm seeks to partition a set of N data points into K disjoint subsets S_j containing N_j data points, in such a way as to minimise the sum-of-squares clustering function given by [36]

$$J = \sum_{j=1}^{K} \sum_{n \in S_j} \left\| x^n - \mu_j \right\|^2 \tag{5.10}$$

where μ_j is the mean of the data points in set S_j and is given by

$$\mu_j = \frac{1}{N_j} \sum_{n \in S_j} x^n \tag{5.11}$$

The algorithm starts off by assigning points at random to K sets and then computing the mean vectors of the points in each set. Next, each point is reassigned to the set with the closest mean vector. The means of the sets are now recalculated and the process repeats itself until there is no further change in the grouping of data points.

5.2.3 Results and discussion

Clustering was performed on two different geographical areas in the network. The first optimisation area was the same as the one used by the heuristic searches and evolutionary search algorithm (genetic algorithm). Promising results were obtained using this area. These results motivated further experimentation, a second area with significantly differed cluster formations was chosen. The best results achieved in this project were based on the second optimisation area. Optimisation region one and two are illustrated in Figure 5.17.

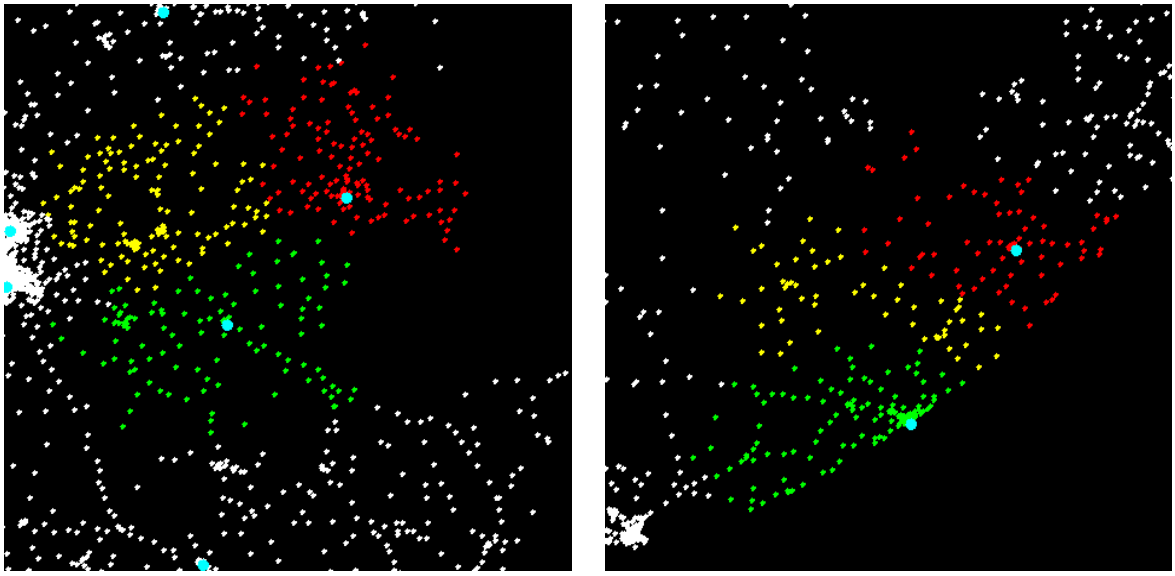

Figure 5.17 Optimisation area 1 cluster formation (left), optimisation area 2 cluster formation (right)

Optimisation region one consisted of three BSCs and 385 sites. The colour of a site indicates the BSC it is working off. Figure 5.17 (left) illustrates the operator's current network configuration. BSCs are indicated with cyan-coloured dots. Sites coloured red work off Nelspruit BSC 1 located at the top right-hand corner of the image. Sites coloured green work off Ermelo BSC 1 situated at the bottom of the image. Yellow-coloured sites work off Pretoria BSC 3 situated at the top left-hand side of the image.

The Ermelo and Nelspruit BSCs are both remote BSCs, Pretoria BSC is a BSC/TRC unit. Nelspruit BSC and Pretoria BSC work off the same MSC situated in the Pretoria switching centre while the Ermelo BSC works off a MSC located in the Germiston switching centre. There is no direct route between the Pretoria and Germiston MSCs, all traffic between them is routed via the Germiston gateway MSC.

Optimisation region two consisted of three BSCs and 289 sites. Figure 5.17 (right) illustrates the operator's current BSC boundaries. Red sites work off Umtata BSC 1 located at the right-hand side of the image while yellow-coloured sites work off Umtata BSC 2. Both Umtata BSC 1 and 2 are co-located at the remote switching centre indicated by the cyan dot at the right-hand side of the image. All green sites work off East London BSC 1 located near the bottom of Figure 5.17 (right).

All the BSCs in optimisation region two are remote BSCs. East London BSC 1 works off a Port Elizabeth BSC/TRC located at the Port Elizabeth switching centre. Umtata BSC 1 and 2 also work off BSC/TRCs located at the Port Elizabeth switching centre. The two Umtata BSCs work off the same MSC located at the Port Elizabeth switching centre while East London BSC 1 works off a different MSC also located at the Port Elizabeth switching centre.

All results presented below are supplied at two levels. Firstly, the optimisation region is evaluated in isolation. This enables one to see exactly what effect the new cluster topology (BSC boundaries) has on the traffic and transmission costs within the region under investigation. Secondly, the entire network is evaluated. By evaluating the entire network, the effect of the BSC boundary change can be observed on a network-wide level. Increases in interswitching centre traffic and transmission costs are thus easily spotted.

5.2.3.1 Optimisation region one

Optimisation region one is the same area as used by the heuristic searches and evolutionary search technique. A common optimisation area was used to enable a fair and unbiased comparison between the different optimisation techniques experimented with in this project.

The effect of the different pre-processing techniques discussed in section 5.2.2.2 is presented next. Figure 5.18 and Figure 5.19 compare the traffic and transmission costs associated with the new cluster boundaries with that of the original cluster boundaries. Looking at Figure 5.18, for example, it is possible to see that the cluster boundaries found when using the logarithm-based pre-processing technique introduced an access transmission cost saving of more than 4% in the region under investigation. The total saving introduced by these new boundaries (again for the logarithm-based approach) equates to a 0.4% saving in access transmission costs on a network-wide level.

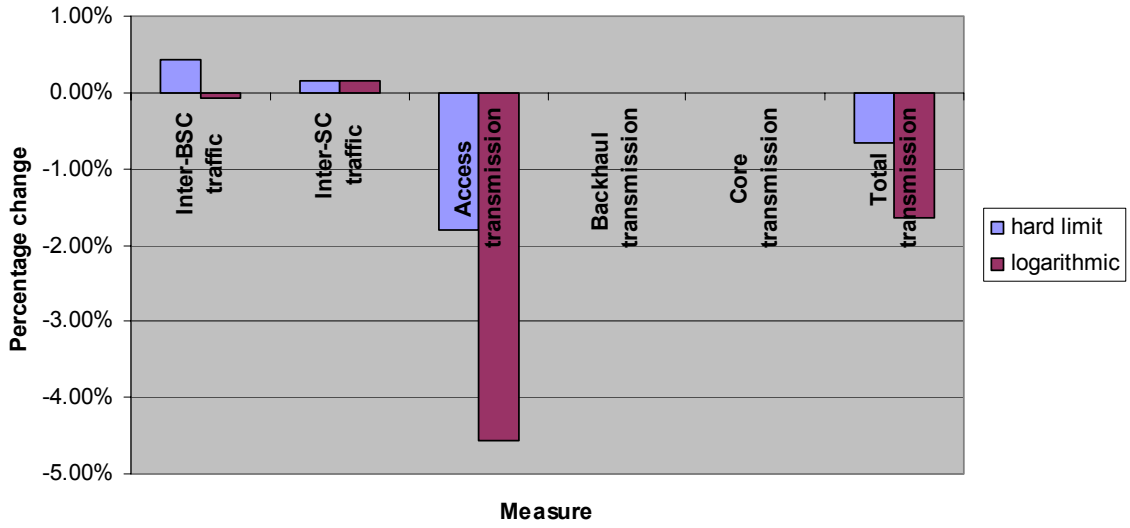

Figure 5.18 GMM optimisation results (optimisation region)

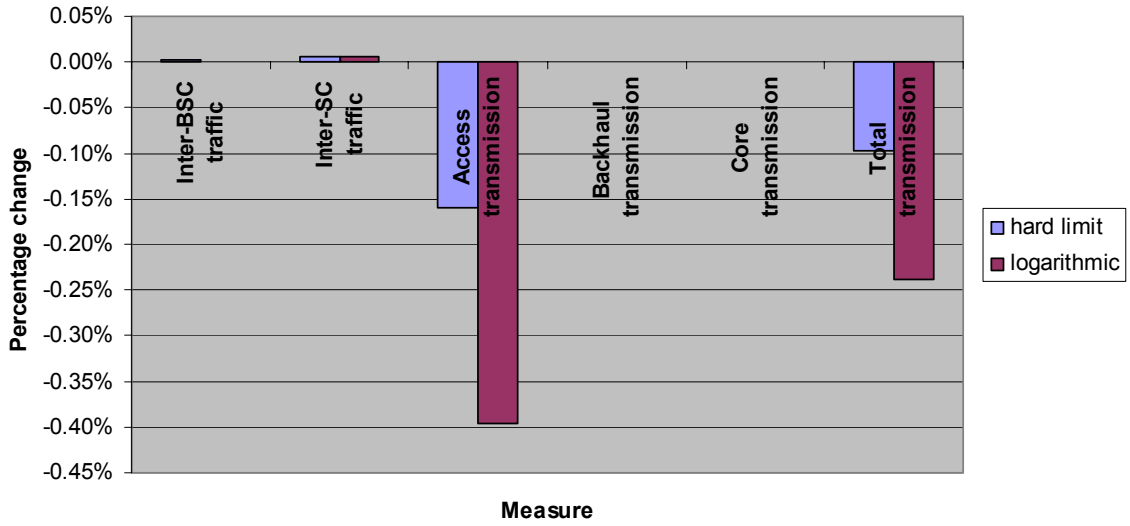

Figure 5.19 GMM optimisation results (entire network)

5.2.3.1.1 Hard-limited pre-processing

The clusters formed by the GMM were all consistent and can easily be implemented by the network operator. A saving of nearly 2% on the access transmission cost was achieved when using the hard-limited pre-processing technique. When comparing the newly formed clusters with the existing clusters, a striking resemblance is noticeable. The serving area of ERBSC1 as well as NSBSC1 increased, while the serving area of PTBSC3 decreased. The increased ERBSC1 serving area was achieved by cutting sites over from PTBSC3. Similarly, the increased serving area of NSBSC1 was achieved by the cutting over of sites from PTBSC3. All the reparented sites were closer to their new BSC parents than to their original parents. This explains the decrease in access transmission cost.

No significant change in the backhaul transmission cost is expected since the number of sites cut over from PTBSC3 to ERBSC1 and NSBSC1 is relatively small. Cutting sites over from PTBSC3 to ERBSC1 and NSBSC1 increases the amount of traffic being backhauled. PTBSC3 is a BSC/TRC unit with no backhaul transmission cost associated with it while ERBSC1 and NSBSC1 are remote BSCs with backhaul transmission costs associated with them. The spare capacity off the backhaul transmission links was sufficient to cater for the additional traffic load, no extra links were thus required.

A slight increase in interswitching centre traffic was observed. This increase was, however, not large enough to cause a change in the number of interswitching centre links.

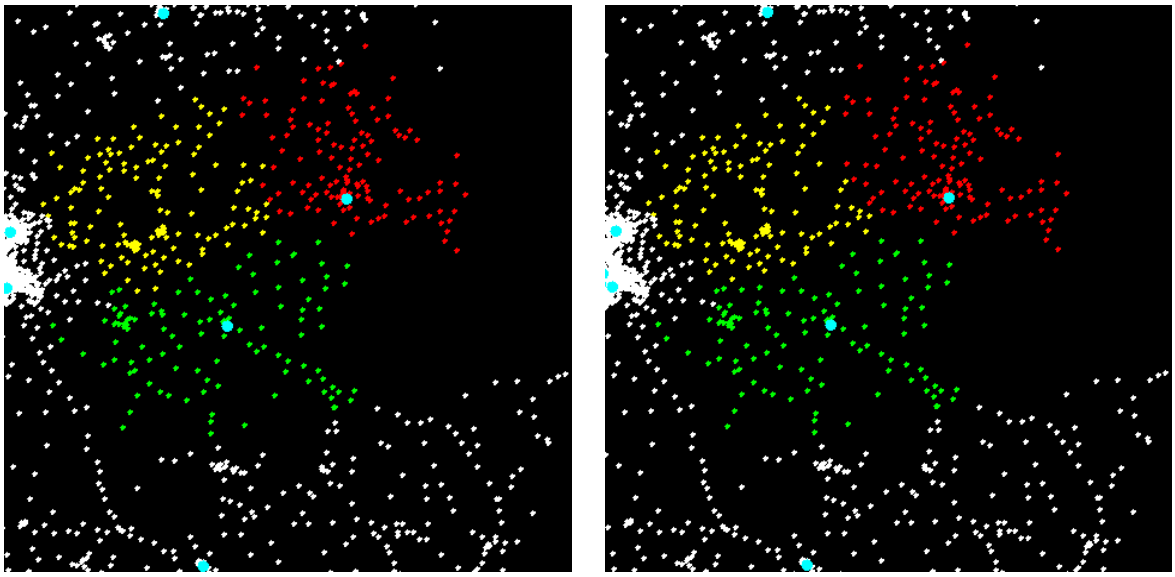

Figure 5.20 Original cluster formation (left), hard-limited technique (right)

% Change in traffic	Inter-BSC traffic	Inter-SC traffic
Optimisation region	0.44%	0.15%
Entire network	0.00%	0.01%

Table 5.1 Traffic measurements

% Change in transmission cost	Access transmission	Backhaul transmission	Core transmission	Total transmission
Optimisation region	-1.80%	0.00%	0.00%	-0.66%
Entire network	-0.16%	0.00%	0.00%	-0.10%

Table 5.2 Transmission cost

ID	TRX utilisation original	TRX utilisation optimised	TRA utilisation original	TRA utilisation optimised	Unit utilisation original	Unit utilisation optimised
ERBSC1	57.29%	58.98%	-	-	57.29%	58.98%
NSBSC1	75.52%	76.95%	-	-	75.52%	76.95%
PTBSC3	58.66%	56.52%	87.88%	87.12%	73.27%	71.82%

Table 5.3 BSC utilisation

5.2.3.1.2 Logarithm-based pre-processing

The clusters formed using logarithm-based pre-processing are similar to the clusters formed using the hard-limited approach. When comparing the newly formed clusters with the original clusters, the difference is an increase in NSBSC1 and ERBSC1's service areas while PTBSC3's service area decreased. The increased serving area of NSBSC1 was achieved by cutting sites over from ERBSC1 and PTBSC3. ERBSC1's service area was grown by cutting sites over from PTBSC3. All the sites that were cut over were closer to their new parents than to their old parents, thus a saving in access transmission costs. More sites were reparented using the logarithm-based pre-processing technique than the hard-limited pre-processing technique. This is why the access transmission cost saving using the logarithm-based pre-processing approach is larger than the hard-limited approach.

The NSBSC1's backhaul transmission cost did not increase compared with the original cluster formation's backhaul cost because the installed links had enough spare capacity to cater for the additional traffic. The same reasoning holds for ERBSC1's backhaul transmission cost. The change in the interswitching centre traffic is minuscule and therefore had no impact on the core network transmission cost.

Overall, the logarithm-based approach produced larger savings than the hard-limited approach. These savings can mostly be contributed to the fact that larger access network transmission cost savings were achieved by the logarithm-based approach due to more sites being reparented to closer BSCs without negatively impacting the core network traffic.

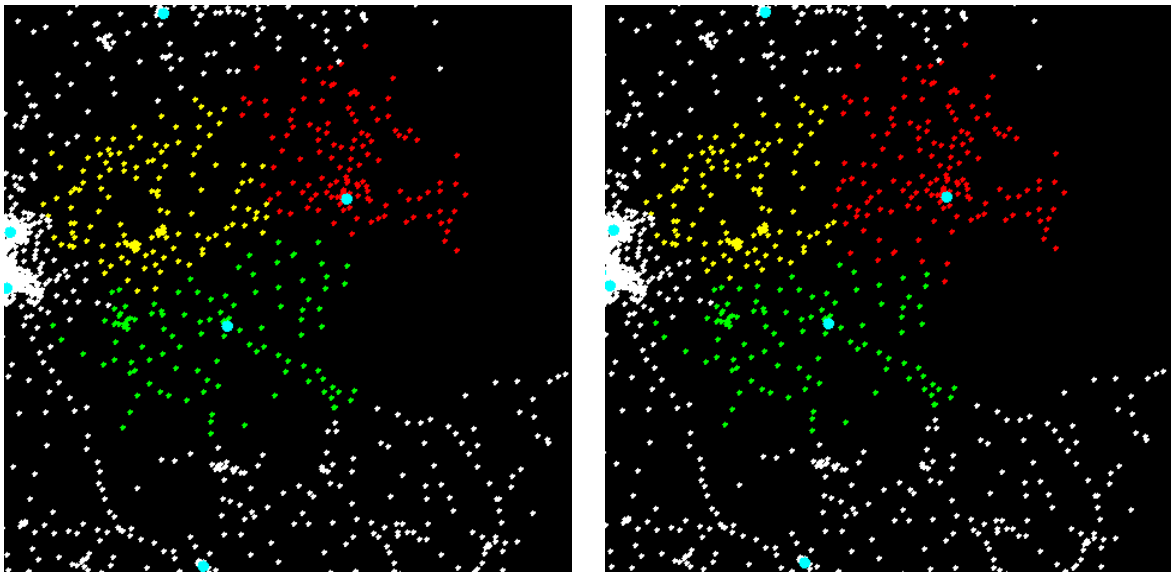

Figure 5.21 Original cluster formation (left), logarithm-based technique (right)

% Change in traffic	Inter-BSC traffic	Inter-SC traffic
Optimisation region	-0.08%	0.15%
Entire network	0.00%	0.01%

Table 5.4 Traffic measurements

% Change in transmission cost	Access transmission	Backhaul transmission	Core transmission	Total transmission
Optimisation region	-4.58%	0.00%	0.00%	-1.66%
Entire network	-0.40%	0.00%	0.00%	-0.24%

Table 5.5 Transmission cost

ID	TRX utilisation original	TRX utilisation optimised	TRA utilisation original	TRA utilisation optimised	Unit utilisation original	Unit utilisation optimised
ERBSC1	57.29%	58.98%	-	-	57.29%	58.98%
NSBSC1	75.52%	79.04%	-	-	75.52%	79.04%
PTBSC3	58.66%	55.09%	87.88%	87.12%	73.27%	71.11%

Table 5.6 BSC utilisation

5.2.3.2 Optimisation area 2

Optimisation area two was specifically chosen for its cluster shapes. Optimisation region two consists of two elongated clusters and one oval cluster, whereas optimisation region 1 consists of three nearly round clusters. Since all the BSCs in optimisation region two work off the same switching centre no change in interswitching centre traffic or core transmission cost is expected.

The results of each of the pre-processing techniques are presented next. The optimisation runs on area two generated cluster configurations very different from the original cluster configurations. The configuration found using the logarithm-based pre-processing technique produced a superior cluster formation compared with the original cluster formation and that formed using the hard-limited pre-processing technique.

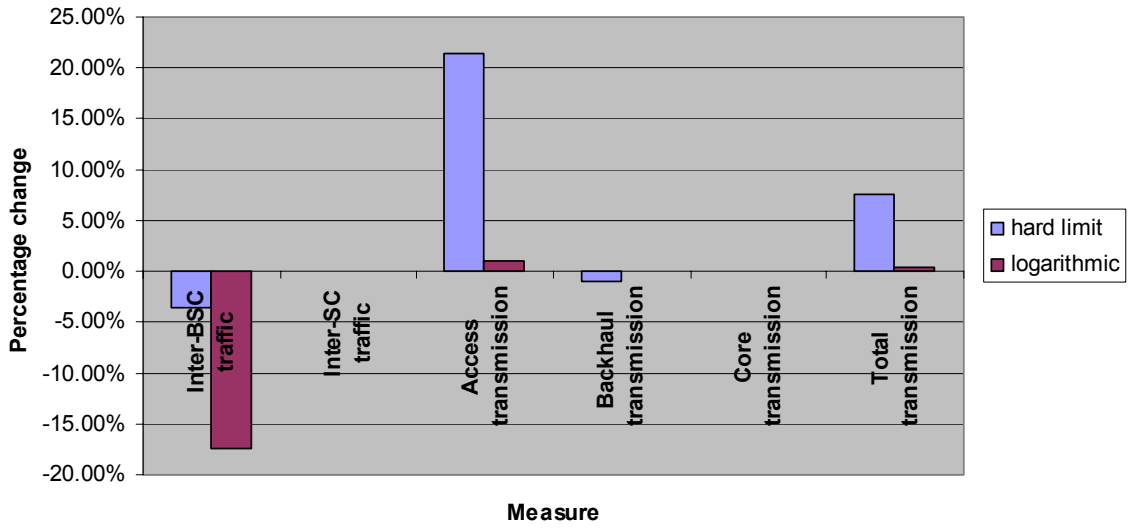

Figure 5.22 GMM optimisation results (optimisation region)

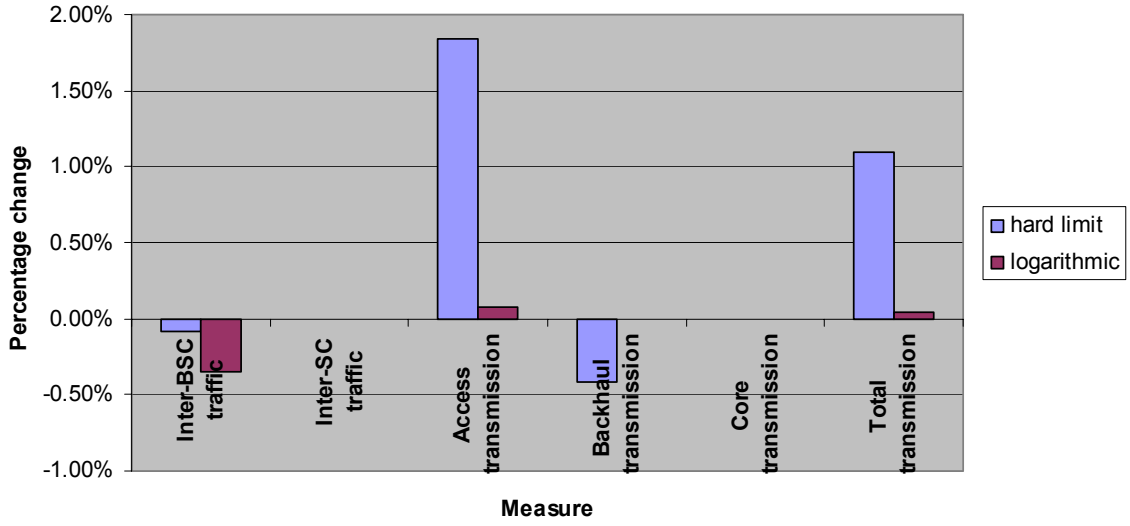

Figure 5.23 GMM optimisation region (entire network)

5.2.3.2.1 Hard-limited pre-processing

The clusters formed were all consistent and easy to implement. These newly formed clusters represent an entirely new cluster formation. A saving of nearly 4% was achieved on the inter-BSC traffic compared with the original configuration. Unfortunately, the saving achieved in inter-BSC traffic is overshadowed by the massive increase in the access

transmission cost which jumped by nearly 22%. A second problem associated with the new cluster configuration is that the TRX capacity constraint of UMBSC2 is broken.

The above-mentioned problems highlight the one deficiency of the GMM clustering approach, namely that the clustering technique is unaware of BSCs. During the cluster formation process, BSC constraints and locations are thus not taken into account. Clustering is performed based purely on the intersite distance and traffic load. Only once clusters have been formed, are they assigned to their closest BSC. The main consequence of the GMM not being aware of BSCs is that the clusters formed can easily break BSC constraints. Some sort of repair algorithm is thus required to reparent sites on the borders of BSC boundaries until all the BSC constraints are met. To minimise the access transmission cost, each of the formed clusters ideally has to be centred around its BSC parent. Without knowledge of the BSCs, the GMM forms clusters independently of where the BSCs are located. This can cause an increase in the access transmission cost as illustrated by this experiment.

The increase in access transmission cost is mainly caused by the UMBSC2 (yellow) sites. UMBSC2 is co-located with UMBSC1 at the cyan dot on the right-hand side of Figure 5.24 (right). Most of the sites working back to UMSC2 are now further away from UMBSC2 than in the original cluster formation. This increase in distance is the main contributing factor to the increase in access transmission cost. The service areas of UMBSC1 and UMBSC2 increased significantly while the service area of ELBSC1 decreased by nearly 35%.

The saving in backhaul transmission costs was achieved due to the fact that the number of backhaul links from the East London remote BSC site decreased while the number of backhaul links from the Umtata remote BSC site increased. Both the Umtata and East London remote BSC sites work back to the Port Elizabeth switching centre. The Umtata remote BSC site is closer to the Port Elizabeth switching centre than the East London remote BSC site. The number of long-distance backhaul transmission links was thus reduced, which translated into a backhaul transmission cost saving. The interswitching centre traffic and transmission cost remained constant as was expected.

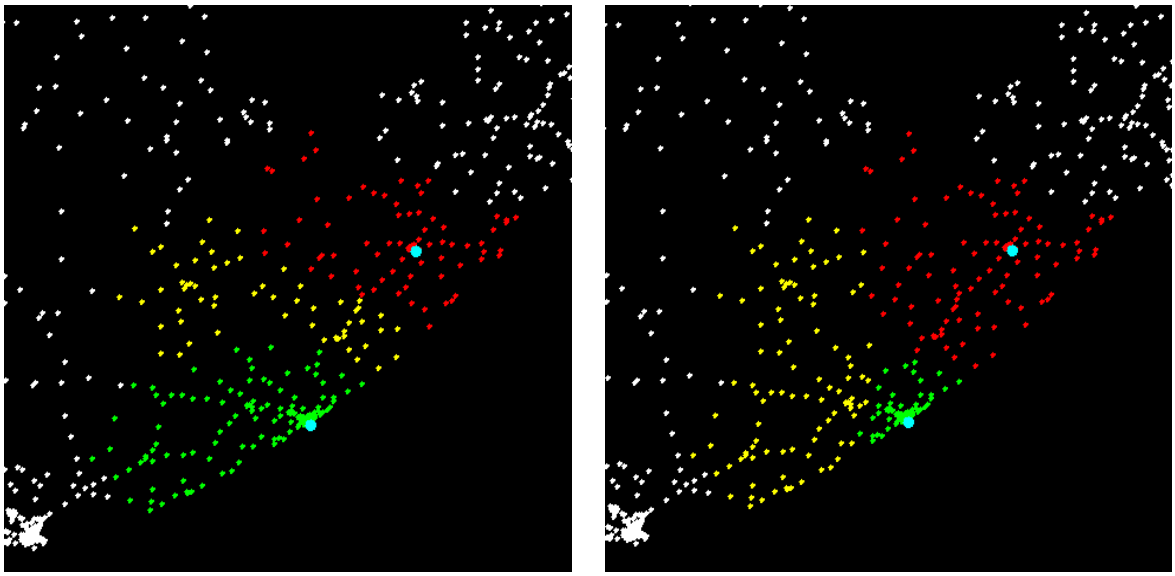

Figure 5.24 Original cluster formation (left), hard-limited cluster formation (right)

% Change in traffic	Inter-BSC traffic	Inter-SC traffic
Optimisation region	-3.59%	-
Entire network	-0.08%	-

Table 5.7 Traffic measurements

% Change in transmission cost	Access transmission	Backhaul transmission	Core transmission	Total transmission
Optimisation region	21.33%	-1.04%	0.00%	7.60%
Entire network	1.84%	-0.41%	0.00%	1.10%

Table 5.8 Transmission cost

ID	TRX utilisation original	TRX utilisation optimised	TRA utilisation original	TRA utilisation optimised	Unit utilisation original	Unit utilisation optimised
ELBSC1	83.69%	48.40%	-	-	83.69%	48.40%
UMBSC1	51.13%	66.76%	-	-	51.13%	66.76%
UMBSC2	86.13%	117.97%	-	-	86.135	117.97%

Table 5.9 BSC utilisation

5.2.3.2.2 Logarithm-based pre-processing

The formation presented in Figure 5.25 represents the most optimal cluster formation found in this project. All the clusters are consistent and thus represent a feasible solution. Inter-BSC traffic within the optimisation region was reduced by nearly 18%. This was achieved at the expense of the access transmission cost. A 1% increase in access transmission cost was recorded.

The backhaul transmission cost remained unchanged. The main reason for this is that the distribution of links between the two remote BSC sites (East London and Umtata) did not change. The service area of ELBSC1 (green sites) was slightly reduced in size. This reduction was not large enough to cause a reduction in the number of backhaul transmission links from the East London remote BSC site. The service area of UMBSC1 (red sites) increased significantly, the TRX count in the new area is 20% up compared with its original service area, there was thus an increase in backhaul transmission links. UMBSC2's (yellow sites) service area was reduced in size, this led to less backhaul transmission links being required. The total service area of the Umtata remote BSC site did not change significantly. This had the effect that the total number of backhaul links from the East London and Umtata remote BSC sites did not change, thus not affecting the backhaul transmission cost.

As with the previous experiment, the interswitching centre traffic and transmission cost remained unchanged as expected. None of the BSC constraints were broken, these BSC boundaries thus represent a feasible solution that can be implemented by the operator.

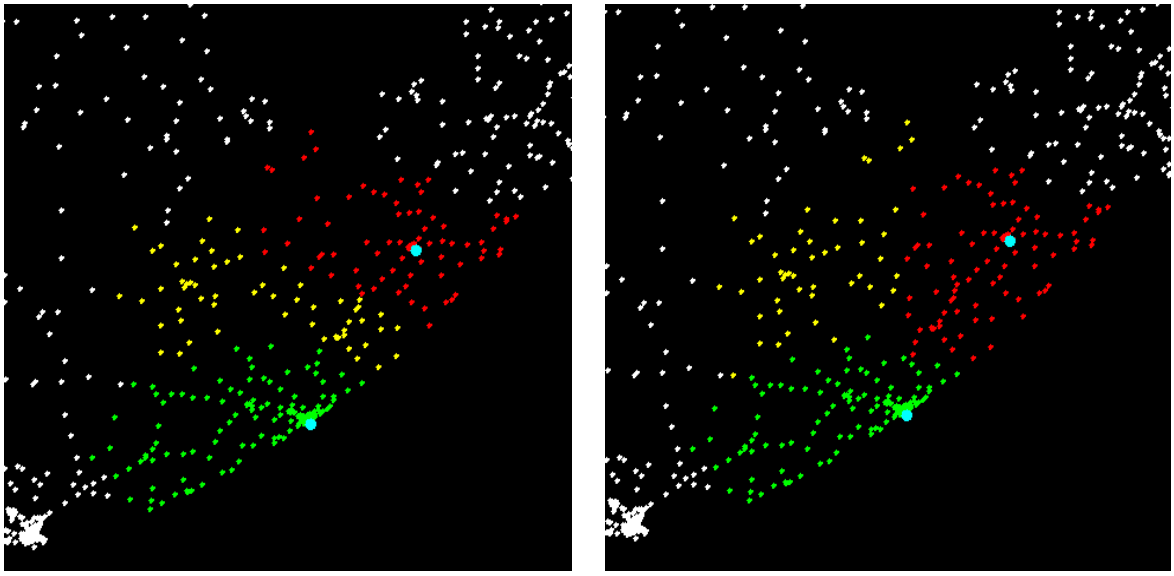

Figure 5.25 Original cluster formation (left), logarithm-based cluster formation (right)

% Change in traffic	Inter-BSC traffic	Inter-SC traffic
Optimisation region	-17.37%	0.00%
Entire network	-0.35%	0.00%

Table 5.10 Traffic measurements

% Change in transmission cost	Access transmission	Backhaul transmission	Core transmission	Total transmission
Optimisation region	1.08%	0.00%	0.00%	0.33%
Entire network	0.07%	0.00%	0.00%	0.05%

Table 5.11 Transmission cost

ID	TRX utilisation original	TRX utilisation optimised	TRA utilisation original	TRA utilisation optimised	Unit utilisation original	Unit utilisation optimised
ELBSC1	83.69%	82.80%	-	-	83.69%	82.80%
UMBSC1	51.13%	62.57%	-	-	51.13%	62.57%
UMBSC2	86.13%	54.49%	-	-	43.07%	27.25%

Table 5.12 BSC utilisation

5.2.3.2.3 Moved cluster centres

The final GMM experiment was performed to determine what the maximum achievable saving is, given the freedom to relocate BSCs. The BSC boundaries of optimisation area two found using the logarithm-based pre-processing technique were used as starting point. These boundaries were selected since they represent the best solution found in the dissertation. As Figure 5.26 and Figure 5.27 illustrate, massive savings were achieved.

The centre of mass of each cluster was determined. Next, the BSC parent of each cluster was relocated to the cluster's centre of mass. This reduced the access transmission cost of the optimisation region to its theoretical minimum, given the BSC boundaries under investigation. Note that the relocation of UMBSC2 made the largest contribution to the reduced access transmission cost.

An access transmission cost saving of 36.4% was achieved in the optimisation area while a network-wide total transmission cost saving of 1.88% was achieved.

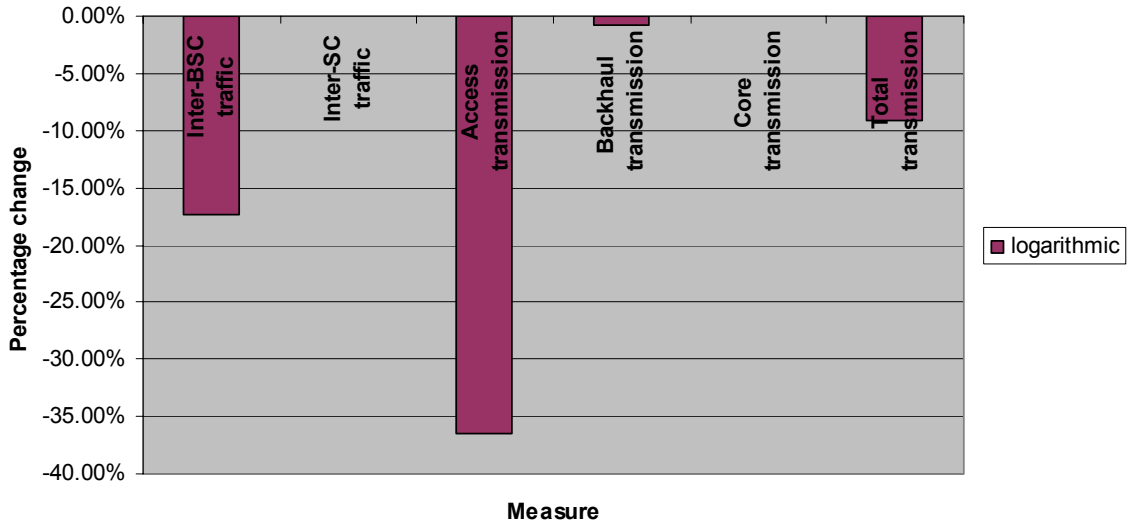

Figure 5.26 GMM optimisation results (optimisation region)

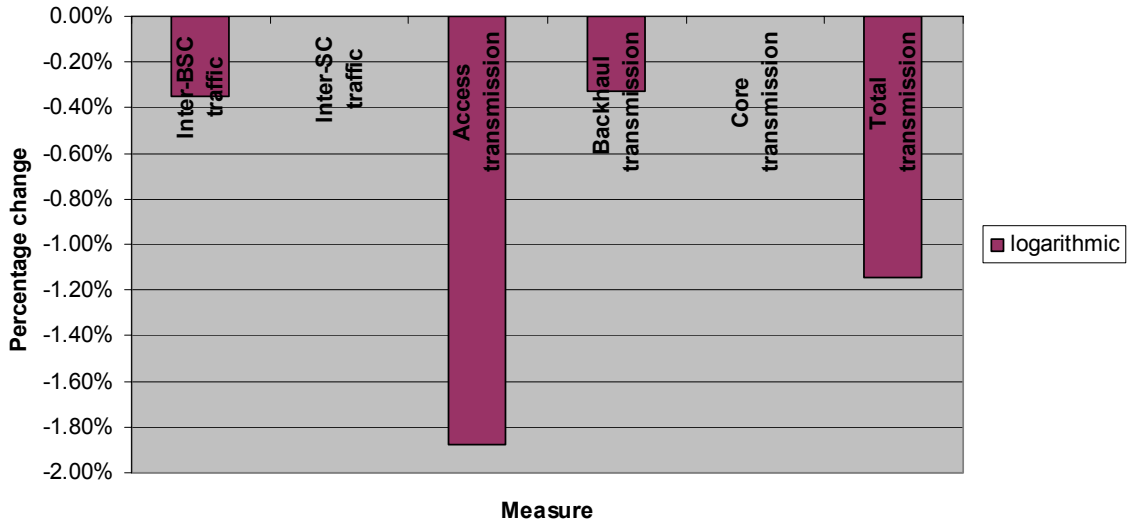

Figure 5.27 GMM optimisation results (entire network)

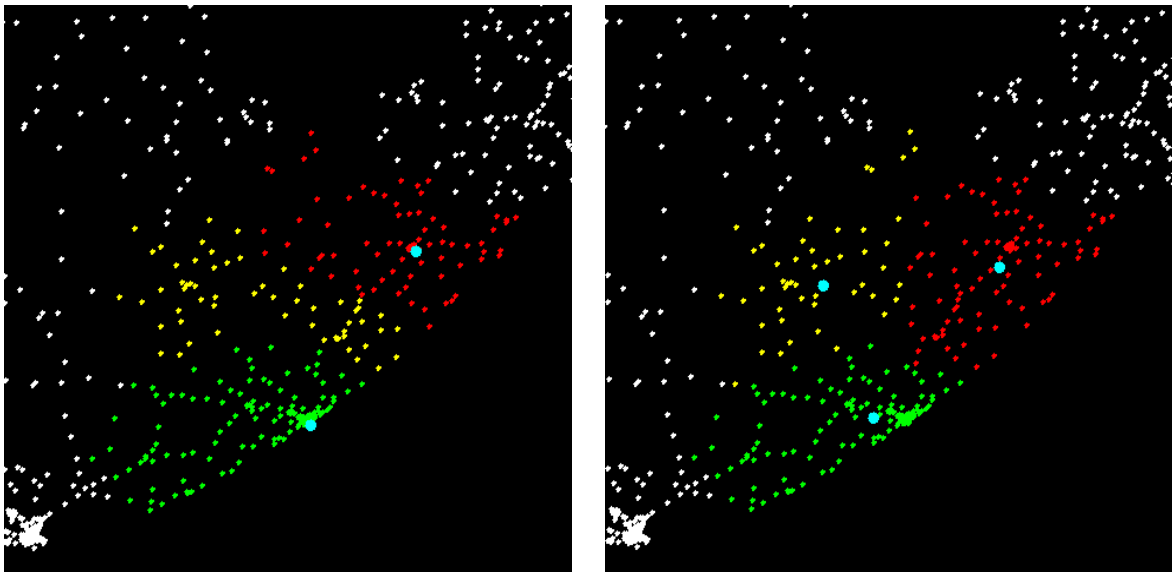

Figure 5.28 Original cluster formation (left), moved BSC cluster formation (right)

% Change in traffic	Inter-BSC traffic	Inter-SC traffic
Optimisation region	-17.37%	0.00%
Entire network	-0.35%	0.00%

Table 5.13 Traffic measurements

% Change in transmission cost	Access transmission	Backhaul transmission	Core transmission	Total transmission
Optimisation region	-36.44%	-0.83%	0.00%	-9.14%
Entire network	-1.88%	-0.33%	0.00%	-1.14%

Table 5.14 Transmission cost

ID	TRX utilisation original	TRX utilisation optimised	TRA utilisation original	TRA utilisation optimised	Unit utilisation original	Unit utilisation optimised
ELBSC1	83.69%	82.80%	-	-	83.69%	82.80%
UMBSC1	51.13%	62.57%	-	-	51.13%	62.57%
UMBSC2	86.13%	54.49%	-	-	43.07%	27.25%

Table 5.15 BSC utilisation

5.3 CONCLUSION

All the experiments conducted in this chapter centred on using an unsupervised learning technique to generate new BSC area boundaries (clusters) based on intersite traffic dispersions and distances.

Results obtained indicated the impact of different pre-processing techniques on the ability of the Gaussian mixture model to produce efficient clusters (BSC service areas) introducing either a transmission saving or reducing the interBSC/switching centre traffic volume. The major advantage of the GMM when forming clusters is that both the intersite distance and intersite traffic dispersion are used during the optimisation process. This results in clusters that are feasible to implement due to their consistency but it also has the advantage of reducing the interBSC/switching centre traffic load and transmission costs.

When performing network capacity expansions, two of the difficult questions to answer are where to put the remote BSCs down and how to alter the existing BSC boundaries to accommodate the new BSC being introduced. Using the approach developed in this chapter, new optimal BSC area boundaries can easily be generated. The new BSC area's centre of mass should then be used as a starting point when searching for a remote BSC site.

One problem is, however, that the GMM does not explicitly take the BSC constraints into consideration. Some of the solutions found by the GMM did break BSC constraints. A repair algorithm can be used to evaluate the clusters formed by the GMM. If BSC constraints are broken, sites on the borders of the BSC boundaries can be cut over until all BSC constraints are met. Future work could focus on the development of efficient repair algorithms.

6 CONCLUSION AND FURTHER WORK

6.1 CONCLUSION

The objective of reducing network cost was achieved by performing cell-to-switch optimisation taking traffic distributions, transmission costs and network element constraints into account. These criteria cannot be divorced from each other since they are all interdependent, omitting any one of them will lead to inefficient and infeasible configurations. Several approaches were presented in this dissertation to solve the problem of assigning cells to switches in cellular mobile networks. Experiments were conducted to measure the quality of solutions provided by the proposed algorithms. Experimental results from solving different instances of assignment problems show that significant savings can be achieved.

To aid in the understanding of the problem, an overview of the GSM architecture as well as teletraffic theory was presented. An in-depth study of the current body of knowledge concerning the cell-to-switch assignment problem and the leaders in the field is also presented. Lastly, the GSM model implemented in software as well as the traffic dispersion generation process was described in Chapter 3.

The experiments conducted in Chapter 4 had two important goals: firstly, to explore the complexity of the search space and, secondly, to determine the extent, if any, by which the current network configuration can be improved on. Results obtained from the heuristic searches indicated that the search space is extremely complex. The search space consists of two components, namely traffic and transmission cost. It was found that when optimising for only one of the components, the other is affected negatively. In order to generate high quality feasible network configurations, it is very important to consider both components simultaneously during the optimisation process.

Results obtained from the Gaussian mixture model experiments indicated the impact of different pre-processing techniques on the ability of the Gaussian mixture model to produce efficient clusters (BSC service areas). Depending on the pre-processing technique used, the solution found by the GMM could be biased towards either producing a transmission cost saving and/or reducing the inter-BSC/switching centre traffic volume. The major advantage of the GMM when forming clusters is that both the intersite distance and intersite traffic dispersion are used during the optimisation process. This not only results in clusters that are feasible to implement due to their consistency but also has the advantage of reducing the inter-BSC/switching centre traffic load and transmission costs.

The best results found were using the Gaussian mixture model where transmission cost savings of up to 17% were achieved (while keeping the MSC and MSC locations fixed). The heuristic searches produced promising results in the form of the characteristics they portray, for example, load-balancing. Due to the massive problem space and suboptimal

chromosome representation, the genetic algorithm struggled to find high quality viable solutions.

New optimisation criteria were introduced in the dissertation which distinguishes it from previous works done in the field, namely.

- Cells were reassigned to switches with the goal of reducing network cost while adhering to network element constraints.
- All results presented in this dissertation were based on actual measurements taken from a GSM network with over 5 000 sites. This serves as evidence that the technique developed is capable of producing feasible, implementable results in large real-world networks.
- One of the most important requirements when addressing the CSA problem for GSM networks is cluster consistency. This requirement was not enforced by the literature surveyed, but will be a deciding factor when selecting network improvement/optimisation techniques for real-world networks.
- When performing network capacity expansions, two of the most difficult questions to answer are where to put the remote BSCs down and how to alter the existing BSC boundaries to accommodate the new BSCs being introduced. Using the techniques developed in this dissertation, new optimal BSC area boundaries can easily be generated. The new BSC area's centre of mass should then be used as a starting point when searching for a suitable remote BSC site. Transmission cost savings of up to 36% were achieved when placing remote BSCs at the cluster's centre of mass.

The developed technique thus assists by generating new, feasible, efficient BSC boundaries when introducing a remote BSC in a region, as well as suggests an optimal location to place the BSC.

In conclusion, the investigation conducted in this dissertation is deemed successful based on the grounds that the NP-hard optimisation problem of assigning cells to switches (with the goal of reducing network cost while adhering to network element constraints) was efficiently solved with transmission cost savings of up to 17% being.

6.2 FURTHER WORK

Further research to be done include the following.

- Development of repair algorithms used specifically to alter network configurations found by the GMM, deemed efficient but infeasible due to broken network element constraints.
- Development of dimension reduction procedures to alleviate the curse of dimensionality. In the current setup, each site represents a dimension, it is worthwhile exploring if this limiting factor cannot be improved upon.

- Development of a chromosome structure that efficiently models the cluster consistency constraint.
- Combining the work done in this dissertation with a secondary program that verifies the feasibility of moving a site over to a new parent. In certain scenarios it is not always possible to reparent a site, for example due to topological limitations such as mountain ranges.

REFERENCES

[1] CME 20/CMS 40 System survey, Ericsson Radio Systems AB

[2] CME 20/CMS 40 Advanced system technique, Ericsson Radio Systems AB, pages 66 - 69

[3] CME 20/CMS 40 Advanced system technique, Ericsson Radio Systems AB, pages 177 - 182

[4] CME 20/CMS 40 System survey, Ericsson Radio Systems AB, pages 137 - 159

[5] T. Rappaport, "*Wireless Communications*", 2001, New York: Prentice Hall, pages 601 – 610

[6] J. Bellamy, "*Digital telephony*", New York: John Wiley & Sons, 1991, pages 459 – 507

[7] A. Merchant, B. Sengupta, "*Assignment of cells to switches in PCS networks*", IEEE/ACM Transactions on Networking, volume 3, issue 5, pages 521 – 526, October 1995

[8] A. Merchant, B. Sengupta, "*Multiway graph partitioning with applications to PCS networks*", 13th IEEE Proceedings on Networking for Global Communications, INFOCOM 1994, volume 2, pages 593 – 600, 12-16 June 1994

[9] P.S. Bhattacharjee, D. Saha, A. Mukherjee, "*Heuristics for assignment of cells to switches in a PCSN: a comparative study*", IEEE International Conference on Personal Wireless Communication, pages:331 – 334, 17-19 February 1999

[10] P.S. Bhattacharjee, D. Saha, A. Mukherjee, "*CALB: a new cell-to-switch assignment algorithm with load balancing in the design of a personal communication services network (PCSN)*", IEEE International Conference on Personal Wireless Communications, pages: 264 – 268, 17-20 December 2000

[11] A. Quintero, S. Pierre, "*A memetic algorithm for assigning cells to switches in cellular mobile networks*", IEEE Communications Letters, volume 6, issue 11, pages 484 – 486, November 2002

[12] S. Pierre, F. Houeto, "*Assigning cells to switches in cellular mobile networks using taboo search*", IEEE Transactions on Systems, Man and Cybernetics, part B, volume 32 , issue 3, pages 351 – 356, June 2002

[13] J.R.L. Fournier, S. Pierre, *"Assigning cells to switches in mobile networks using an ant colony optimization heuristic",* Computer Communications, 28 (1), pages 65 - 73, 2005

[14] A. Quintero, S. Pierre, "*Sequential and multi-population memetic algorithms for assigning cells to switches in mobile networks*", Computer Networks, volume 43, issue 3, pages 247 – 261, 22 October 2003

[15] A. Quintero; S. Pierre, "*Evolutionary approach to optimize the assignment of cells to switches in personal communication networks*", Computer Communications, volume 26, issue 9, pages 927 – 938, 2 June 2003

[16] A. Quintero, S. Pierre, "*Assigning cells to switches in cellular mobile networks: a comparative study*", Computer Communications, volume 26, issue 9, pages 950 – 960, 2 June 2003

[17] S. Pierre, F. Houéto, "*A tabu search approach for assigning cells to switches in cellular mobile networks*", Computer Communications, volume 25, issue 5, pages 464 – 477, 15 March 2002

[18] D.R. Din, S.S. Tseng, "*A genetic algorithm for solving dual-homing cell assignment problem of the two-level wireless ATM network*", Computer Communications, volume 25, issue 17, pages 1536 – 1547, 1 November 2002

[19] I. Demirkol, C. Ersoy, M.U. Caglayan, H. Delic, "*Location area planning and cell-to-switch assignment in cellular networks*", IEEE Transactions on Wireless Communications, volume 3, issue 3, pages:880 – 890, May 2004

[20] S. Zahid Ali, R.J. Read, "*Design and performance evaluation of an optimization methodology for optimal solution of Cellular Layout Design Problems*", International Symposium on Signals, Circuits and Systems, SCS 2003, volume 2 , pages:521 – 524, 10-11 July 2003

[21] E. Cayirci, I.F. Akyildiz, "*Optimal location area design to minimize registration signaling traffic in wireless systems*", IEEE Transactions on Mobile Computing, volume 2, issue 1, pages 76 – 85, January-March 2003

[22] R. Subrata, A.Y. Zomaya, "*A comparison of three artificial life techniques for reporting cell planning in mobile computing*", IEEE Transactions on Parallel and Distributed Systems, volume 14, issue 2, pages 142 – 153, February 2003

[23] T. Ozugur, "*Multiobjective hierarchical location and routing area optimization in GPRS and UMTS networks*", IEEE International Conference on Communications, ICC 2002, volume 1, pages 579 – 584, 28 April-2 May 2002

[24] T. Ozugur, A. Bellary, F. Sarkar, "*Multiobjective hierarchical 2G/3G mobility management optimization: niched Pareto genetic algorithm*", IEEE Global Telecommunications Conference, GLOBECOM 2001, volume 6, pages 3681–3685, 2001

[25] S. Menon, R.Gupta, "*Assigning cells to switches in cellular networks by incorporating a pricing mechanism into simulated annealing*", IEEE Transactions on Systems, Man and Cybernetics, part B, volume 34 , issue 1, pages:558 – 565, February 2004

[26] S. Mandal, D. Saha, A. Mahanti, "*A heuristic search for generalized cellular network planning*", IEEE International Conference on Personal Wireless Communications, pages 105 – 109, 15-17 December 2002

[27] M. Maitra, A. Mukherjee, D. Saha, S.S. Chowdhury, "*A heuristic-based approach for minimization of inter-switch handoff cost with load balancing for optimal location area planning in a PCSN*", IEEE International Conference on Personal Wireless Communications, pages 172 – 176, 15-17 Dec. 2002

[28] Java programming language [Online], Available: http://www.java.sun.com, Last accessed: 9 April 2007

[29] A Java-based Evolutionary Computation Research System [Online], Available: http://www.cs.gmu.edu/~eclab/projects/ecj/, Last accessed: 9 April 2007

[30] M. Negnevitsky, "*Artificial intelligence*", Harlow, England: Addison-Wesly, 2002, pages 220 – 221

[31] A.P. Engelbrecht, "*Computational intelligence*", London: Wiley, 2003, pages 126 – 130

[32] D. Peter, "*Unsupervised Learning*", The MIT Encyclopaedia of the Cognitive Sciences, 1999

[33] C.M. Bishop, "*Neural Networks for Pattern Recognition*", Oxford: Oxford University Press, 2003, pages 59 – 73

[34] C. Tomasi (2004). *Estimating Gaussian Mixture Densities with EM – A Tutorial* [Online]. Available: http://www.cs.duke.edu/courses/spring04/cps196.1/handouts/EM/tomasiEM.pdf, Last accessed: 9 April 2007

[35] Netlab pattern analysis toolbox for Matlab [Online], Available: http://www.ncrg.aston.ac.uk/netlab/, Last accessed: 9 April 2007

[36] C.M. Bishop, "*Neural Networks for Pattern Recognition*", Oxford: Oxford University Press, 2003, pages 187 – 190

[37] Erlang B tables [Online]. Available: http://www.itu.int/itudoc/itu-d/dept/psp/ssb/planitu/plandoc/erlangt.html, Last accessed: 9 April 2007

[38] I.F. Akyildiz, W. Wang, "*A dynamic location management scheme for next-generation multitier PCS systems*", IEEE Transactions on Wireless Communications, volume 1, issue 1, pages:178 – 189, January 2002

[39] B.L. Mark, Z.R. Zaidi,, "*Robust mobility tracking for cellular networks*", IEEE International Conference on Communications, ICC 2002, volume 1, pages:445 – 449, 28 April-2 May 2002

[40] S. Mandal; D. Saha, A. Mahanti, "*Heuristic search techniques for cell to switch assignment in location area planning for cellular networks*", IEEE International Conference on Communications, volume 7, pages:4307 – 4311, 20-24 June 2004

[41] W.N.N. Hung, Xiaoyu Song, "*On optimal cell assignments in PCS networks*", 21st International IEEE Performance, Computing, and Communications Conference, pages 225 – 232, 3-5 April 2002

[42] S. Zahid Ali; R.J. Read, "*Design and performance evaluation of an optimization methodology for optimal solution of Cellular Layout Design Problems*", International Symposium on Signals, Circuits and Systems, SCS 2003, volume 2, pages 521 – 524, 10-11 July 2003

[43] L. Du, J. Bigham, L. Cuthbert, C. Parini, P. Nahi, "*Cell size and shape adjustment depending on call traffic distribution*", IEEE Wireless Communications and Networking Conference, WCNC 2002, volume 2, pages 886 – 891, 17-21 March 2002

[44] Chou Hsinghua, G. Premkumar, Chu Chao-Hsien, "*Genetic algorithms for communications network design - an empirical study of the factors that influence performance*", IEEE Transactions on Evolutionary Computation, volume 5, issue 3, page 236 – 249, June 2001

[45] G. Celli, E. Costamagna, A. Fanni, "*Genetic algorithms for telecommunication network optimization*", IEEE International Conference on Systems, Man and Cybernetics, 'Intelligent Systems for the 21st Century', volume 2 , pages 1227 – 1232, 22-25 October 1995

[46] M. Negnevitsky, "*Artificial intelligence*", Harlow, England: Addison-Wesly, 2002, pages 217 – 257

[47] A.P. Engelbrecht, "*Computational intelligence*", London: Wiley, 2003, pages 123 – 132

[48] A.P. Engelbrecht, "*Computational intelligence*", London: Wiley, 2003, pages 133 – 145

[49] C.U. Saraydar, O.E. Kelly, C. Rose, "*One-dimensional location area design*", IEEE Transactions on Vehicular Technology, volume 49, issue 5, pages:1626 – 1632, September 2000

[50] V. Garg, J. Wilkes, "*Principles & Applications of GSM*", New York: Prentice Hall, 1998, pages 287 – 327

[51] W. Stallings, "*Wireless communications and networks*", New York: Pearson Education, 2002, pages 531 – 539

ADDENDUMS

A TELETRAFFIC THEORY

A.1 INTRODUCTION

Analysing the flow of traffic through any communication system is very important. There is a balance the communication provider always tries to achieve. The communication provider must supply enough resources to be able to serve its customers during the busy hour at the specified grade of service. Depending on the communications regulatory authority and country, this GoS may be specified in the licence agreement. Supplying resources is, however, terribly expensive. If the operator supplies twice the amount of required resources, it stands to lose a lot of money due to underutilisation.

By using traffic analysis techniques, it is possible to derive the peak traffic load the system has to be able to carry. All licence agreements can then be met without wasting exorbitant amounts on unnecessary equipment. The body of knowledge presented in section A.2 and section A.3 was obtained through a detailed study of [5, 6, 50, 51].

A.2 TELETRAFFIC ANALYSIS THEORY

The basic goal of traffic analysis is to determine the cost-effectiveness of different network configurations. The effectiveness of the network can be evaluated by determining how much traffic it carries during normal load conditions and how often the offered traffic exceeds the network's capacity.

Traffic analysis techniques can be divided into two main categories, namely.

- Delay systems
- Loss systems

The type of technique to use will be determined by the method the system uses to cope with overload traffic. In a delay system, calls arriving that cannot be processed at the current instant in time is pushed into a queue. In a lost calls cleared system, call requests are denied during overload conditions. It should be noted that even in delay systems, at some point calls could be dropped. This is due to the fact that the queues are of a finite length. Once a queue is full, any further calls being pushed into the queue will simply be dropped.

Dropping calls is highly undesirable. Every dropped call is a potential loss of revenue for the operator. Performance in a lost calls cleared system is measured in terms of the probability of rejecting a call. In delay systems, the average holding time is mostly used as a measure of performance.

In a telephone system it, is assumed that all calls arrive in a random fashion. Figure A.1 depicts exactly this assumption. The bottom half represents the traffic offered to the system. The top part shows the number of trunks (servers) necessary to carry the offered traffic load. The maximum number of servers necessary is three.

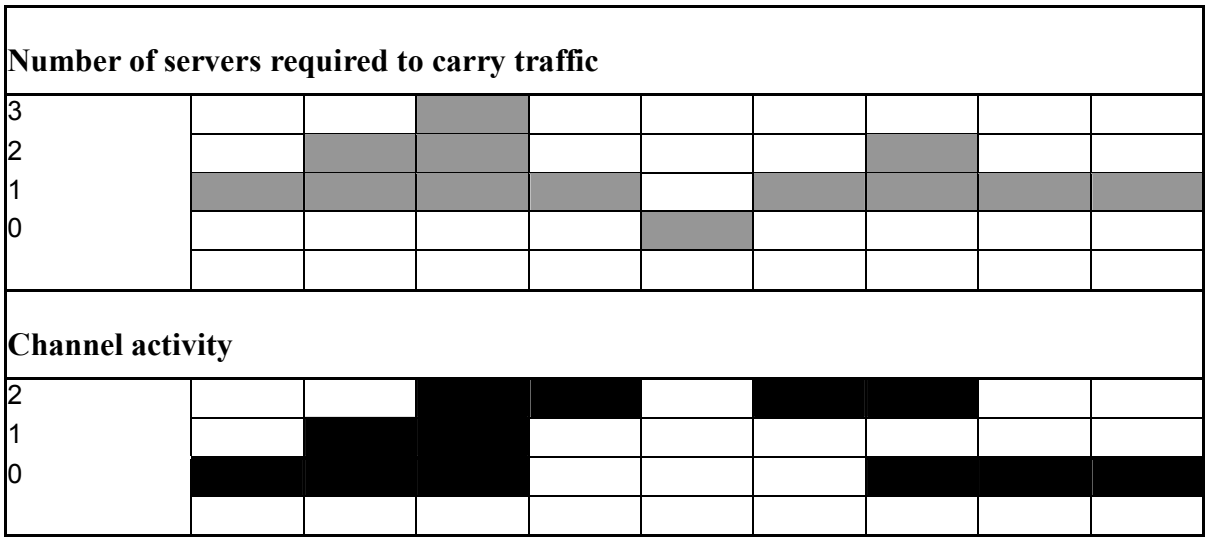

Figure A.1 Offered load and the number of servers required

A.2.1 The measure of traffic

Certainly, one of the best measures of a network's capacity is the amount or volume of traffic that the network is capable of successfully handling during a specified period. A useful measure of traffic is thus the traffic intensity that the network is capable of carrying. One measure of traffic intensity is called the Erlang. This measure is named after the pioneer traffic theorist A.K. Erlang. The traffic intensity is obtained by dividing the traffic volume (in seconds, for example) by the duration during which this traffic volume was measured (also in seconds). As can be seen, the traffic density is dimensionless. The unit, Erlang, is thus only used in honour of A.K. Erlang.

In the terminology used in traffic theory, a server is referred to as a single channel. An E1 link has 32 TDMA time slots. Each time slot can carry a single voice call at a time. It thus has a maximum capacity of 32 Erlang. This assumes that all the time slots are available to carry traffic and that each time slot can only carry one voice call at a time.

Traffic can also be calculated assuming that the average call arrival rate and mean call holding time are known using the formula

$$A = \lambda t_m \qquad \text{(A.1)}$$

where λ is the average arrival rate, t_m is the mean holding time and A is the traffic intensity in Erlang. It should be noted that the traffic intensity calculated above is the average utilisation during a period and not the maximum load experienced by the system.

Telephone networks are designed to carry the busy-hour traffic at a predefined GoS. This GoS may be assigned by the communications authority regulating telecommunication in that country. The busy-hour traffic, as the name implies, is the hour of the day during which the telephone network has to carry the largest amount of traffic. The busy hour is influenced by the geographical area as well as the type of subscriber.

Offered traffic and carried traffic; in an ideal world, these two terms would mean exactly the same. Due to the cost of equipment it is not viable for the operators to provide enough capacity in their networks to handle all the potential calls simultaneously. The GoS indicates the probability of an unsuccessful call due to congestion. The offered traffic is all the traffic that is presented to the network, whether the network can cope with it or not. The carried traffic is the traffic that was presented to the network and that the network was able to carry. When congestion occurs, the offered traffic would be more than the carried traffic. This is because certain calls will not be able to complete the call set-up process due to congestion on trunks, congestion on signalling links or finite CPU capacity at the switches.

A.2.2 Call arrivals

The call arrival rate is the rate at which calls arrive at the serving switch. It is always assumed that the arrival of one call has absolutely no correlation with the arrival of any other call. All call arrivals are assumed to arrive independently. This forms one of the most fundamental assumptions of classical traffic analysis. Although this assumption might not always hold, it is possible to adjust the random arrival assumption so that useful results can still be obtained.

When speaking of arrival distributions, it is important do distinguish between the following.

- Interarrival time
- Arrival distribution

Each of these topics will be covered next in more detail.

A.2.2.1 Interarrival times

The interarrival time is modelled by a negative exponential equation. The following assumptions are made.

- Only one call can arrive per time interval.
- The probability of a call arrival in a given time period is directly proportional to the length of the time period.
- The probability of a call arriving in a time period has no correlation to what happened in the previous time periods. All calls thus arrive independently.

The probability of zero calls arriving in a specified period is given by

$$P_0(\lambda t) = e^{-\lambda t} \tag{A.2}$$

where λ is the average call arrival rate and t is the specified time period.

A.2.2.2 Arrival distributions

A valuable piece of information will be to know how many calls are expected to arrive within a specified period of time. This cannot be directly computed from the previous equations since it only models the interarrival times. The probability of j arrivals within time period t can, however, be computed using the equation

$$P_j(\lambda t) = \frac{(\lambda t)^j}{j!} e^{-\lambda t} \tag{A.3}$$

A more useful equation would be one that is able to indicate the probability of j or more calls arriving within the specified period of time. This can be derived from equation 4 as follows

$$\begin{aligned} P_{\geq j}(\lambda t) &= \sum_{i=j}^{\infty} P_i(\lambda t) \\ &= 1 - \sum_{i=0}^{j-1} P_{<i}(\lambda t) \\ &= 1 - P_{<j}(\lambda t) \end{aligned} \tag{A.4}$$

A.2.3 Holding times

The first principle of traffic intensity has just been discussed, namely the arrival rate. The second factor considered is the holding time of the phone calls. The two most commonly used holding time distributions are.

- Constant holding time
- Exponential holding time

Each one of the above holding times will be discussed in more detail in the sections to follow.

A.2.3.1 Constant holding times

Constant holding times cannot be applied to all voice calls in general. There are, however, cases where a constant holding time is a valid assumption. One example of such a system is one where an automated response is played to the people that phoned in. Another example is in a data network where a constant packet size is used.

If a constant holding time is assumed, the probability of j channels being busy during any period of time is the same as the probability that j calls arrive during that same period of time.

Remember that

$$A = \lambda t_m \tag{A.5}$$

where λ is the average arrival rate, t_m is the mean holding time and A is the traffic intensity in Erlang.

Equation 3 can then be rewritten as

$$\begin{aligned} P_j(\lambda t_m) &= P_j(A) \\ &= \frac{(A)^j}{j!} e^{-A} \end{aligned} \tag{A.6}$$

where A is the traffic intensity in Erlang, λ is the arrival rate and t_m is the constant holding time.

P_j is thus the probability that j channels will be occupied and is only dependent on the traffic load in Erlang.

A.2.3.2 Exponential holding times

The exponential holding time distribution is one of the most commonly used distributions for regular telephone calls. The exponential holding time distribution is given below

$$P(>t) = e^{\frac{-t}{t_m}} \tag{A.7}$$

where t_m is the average holding time.

The above equation expresses the probability that a holding time exceeds the time of t.

The following two graphs illustrate how closely the negative exponential holding time distribution resembles the actual measured holding time distribution.

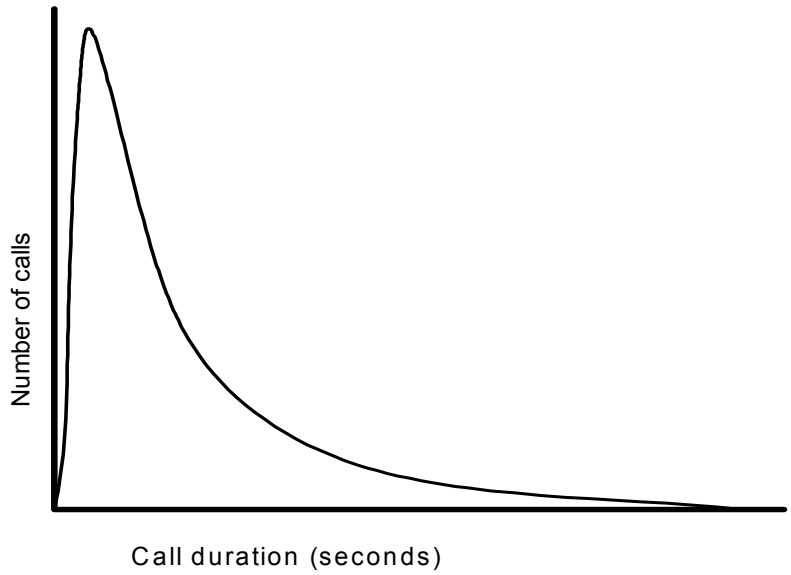

Figure A.2 Measured holding time distribution

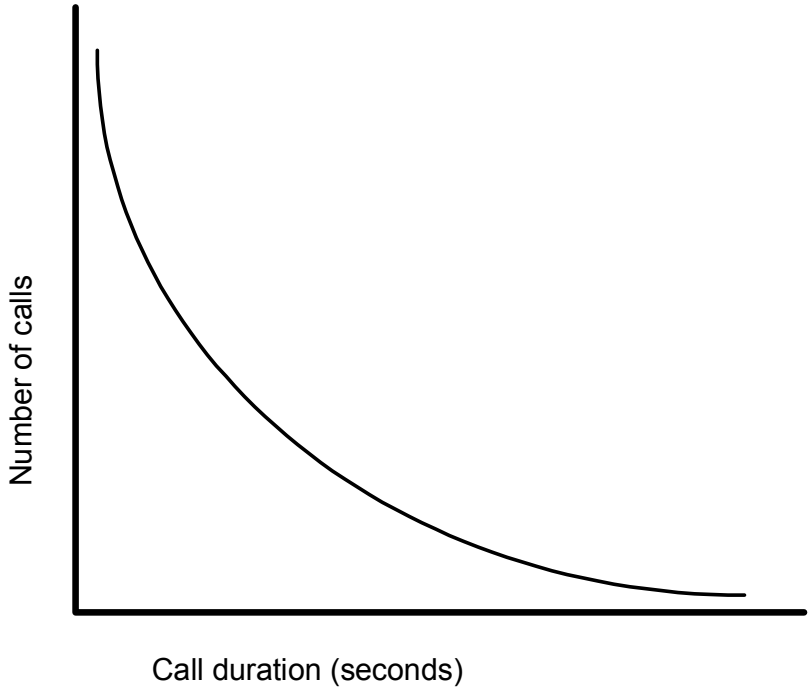

Figure A.3 Negative exponential holding time distribution

One observation that is worthy of being mentioned is that the negative exponential distribution has the property that the probability of the holding time is independent of how long the call has been in progress up to that point. Absolutely no form of history of the call is thus taken into consideration.

The probability distribution of the number of active circuits in a trunk can also be modelled by equation 5 in circumstances where the holding time of the calls is exponential. The probability of j circuits being used assuming exponential holding times can be represented by

$$P_j(A) = \frac{(A)^j}{j!} e^{-A} \tag{A.8}$$

where A is the traffic intensity in Erlang.

A.2.4 Loss systems

A.2.4.1 Lost calls cleared system

In a lost calls cleared system, Erlang's B formula is used. This formula is also referred to as Erlang's formula of the first kind or Erlang's loss formula. Erlang's B formula is based on the concept of statistical equilibrium.

Statistical equilibrium implies that the state of a system is independent of the time at which the system is evaluated. When a system is in equilibrium, it is as likely to have a new

arrival as it is to have a departure. In the case where the number of active circuits increases above the average, the probability of departures increases while the probability of arrivals decreases. In the event of the number of active circuits falling below the average, the probability of calls arriving increases while the probability of calls leaving the system decreases.

Erlang's B formula is given by

$$B = \frac{A^N}{N!\sum_{i=0}^{N}\left(\frac{A^i}{i!}\right)} \tag{A.9}$$

where N is the number of servers or channels and A is the offered traffic in Erlang.

Erlang's B formula specifies the blocking probability for a system with random arrivals and assumes an arbitrary holding time distribution. The blocking probability is simply the probability of an unsuccessful call attempt due to congestion. The blocking probability is based on the offered traffic load as well as the number of servers available.

The output utilisation represents the traffic carried by each server and is given below

$$p = \frac{(1-B)A}{N} \tag{A.10}$$

where B represents the blocking probability, A is the offered traffic in Erlang and N is the number of servers available to carry the traffic.

The relation between traffic volume, number of devices and blocking probability requires tedious numerical computations. To ease this problem, tables have been compiled. These tables are commonly referred to as the "Erlang B Tables". Two common versions are available, the first version indicates the maximum traffic load that can be carried given the number of devices and blocking probability. The second version indicates the blocking probability given the traffic load and number of servers. Refer to section A.4 for an example of these tables.

A.3 DIMENSIONING

What does dimensioning really mean? When network dimensioning is performed, appropriate strategies are followed to select the optimum equipment for the best network design. The best network design is one in which the user is provided with high quality services while network efficiency is maximised.

When trying to model a cellular system, there are three aspects to cover, namely.

- The call model
- The mobility model
- The topology model

Each of the above-mentioned aspects will be covered in more detail in the sections to follow.

A.3.1 The call model

It is used to generate traffic for each of the subscribers in the network. The call model, in turn, can be divided into two subcategories, namely.

- A call traffic model
- A caller distribution model

The call traffic model describes the usage of the offered services of the operator by the subscribers.

The caller distribution model describes the behaviour of individual customers. It takes into account the fact that any given subscriber is more likely to phone a certain group of people than he or she is to phone anyone at random.

A.3.2 The mobility model

The mobility model is the part of the mobile operator's dimensioning process that truly distinguishes it from the techniques followed by the fixed-line operators. The mobility model is used to describe the occurrence of events such as location updates and handovers. These two events are only found in mobile systems.

Modelling the movement of subscribers through the system is not a trivial problem. Several models have been developed to model the mobility of the users. Some of the most popular models are.

- The fluid model
- The Markovian model
- The gravity model
- The topology model

Each of there models will next be discussed in more detail.

A.3.2.1 The fluid model

The fluid model, as its name suggests, models the movement of subscribers to be the same as that of a fluid. According to assumptions made by the fluid model, the number of people leaving a region is proportional to the population density of that region, the velocity of the movement and the length of the boundary of the region. All these factors can be combined to calculate the average number of crossings per unit time. A crossing is the occurrence where a subscriber exits one region and enters another and can be modelled as

$$\lambda = \frac{vpL}{\pi} \tag{A.11}$$

where λ represents the number of crossings per unit time, p is the average population density in the region, v is the average movement velocity in the region and L is the perimeter of the region.

Due to the fact that the fluid model uses the average population density and average velocity, it is most accurate when used to describe the behaviour of a large population of people. This model is not intended to model the population on a user-by-user basis.

A.3.2.2 The Markovian model

This model was developed to model the behaviour of an individual. It is also known as the random walk model. According to this model, there is a certain probability that a user will stay within the current region (p), and another probability ($1 - p$), that the user will cross a boundary and enter an adjacent region. Various transition probability distributions can be used depending on the behaviour of the users of the specific region. There is, however, a downside to using this model and that is that there is no concept of trips. All transitions between regions are purely based on the transition probability distribution.

A.3.2.3 The gravity model

This is one of the more commonly used models in transportation research and is used to model human movement behaviour. This type of model has been successfully applied to regions with vastly differing sizes. It has been implemented from town level all the way up to the international level. The main downside of this approach is the large number of parameters required per region. To model a geographic area with many regions becomes unpractical.

A.3.3 The topology model

The topology model is used to describe the network area. It should thus contain all the information regarding possible/current signalling and traffic points. Any information that describes the region and that can have an influence on the placing of equipment is added to

this model. The optimisation process is thus dependent on the topology model for an accurate description of the region.

A.4 ERLANG B TABLES

Both Table A.1 and Table A.2 were obtained from [37]. Note that only partial extracts of both tables are presented due to space constraints.

n	Loss probability (E)									n
	0.00001	**0.00005**	**0.0001**	**0.0005**	**0.001**	**0.002**	**0.003**	**0.004**	**0.005**	
1	.00001	.00005	.00010	.00050	.00100	.00200	.00301	.00402	.00503	**1**
2	.00448	.01005	.01425	.03213	.04576	.06534	.08064	.09373	.10540	**2**
3	.03980	.06849	.08683	.15170	.19384	.24872	.28851	.32099	.34900	**3**
4	.12855	.19554	.23471	.36236	.43927	.53503	.60209	.65568	.70120	**4**
5	.27584	.38851	.45195	.64857	.76212	.89986	.99446	1.0692	1.1320	**5**
6	.47596	.63923	.72826	.99567	1.1459	1.3252	1.4468	1.5421	1.6218	**6**
7	.72378	.93919	1.0541	1.3922	1.5786	1.7984	1.9463	2.0614	2.1575	**7**
8	1.0133	1.2816	1.4219	1.8298	2.0513	2.3106	2.4837	2.6181	2.7299	**8**
9	1.3391	1.6595	1.8256	2.3016	2.5575	2.8549	3.0526	3.2057	3.3326	**9**
10	1.6970	2.0689	2.2601	2.8028	3.0920	3.4265	3.6480	3.8190	3.9607	**10**
11	2.0849	2.5059	2.7216	3.3294	3.6511	4.0215	4.2661	4.4545	4.6104	**11**
12	2.4958	2.9671	3.2072	3.8781	4.2314	4.6368	4.9038	5.1092	5.2789	**12**
13	2.9294	3.4500	3.7136	4.4465	4.8306	5.2700	5.5588	5.7807	5.9638	**13**
14	3.3834	3.9523	4.2388	5.0324	5.4464	5.9190	6.2291	6.4670	6.6632	**14**
15	3.8559	4.4721	4.7812	5.6339	6.0772	6.5822	6.9130	7.1665	7.3755	**15**
16	4.3453	5.0079	5.3390	6.2496	6.7215	7.2582	7.6091	7.8780	8.0995	**16**
17	4.8502	5.5583	5.9110	6.8782	7.3781	7.9457	8.3164	8.6003	8.8340	**17**
18	5.3693	6.1220	6.4959	7.5186	8.0459	8.6437	9.0339	9.3324	9.5780	**18**
19	5.9016	6.6980	7.0927	8.1698	8.7239	9.3515	9.7606	10.073	10.331	**19**
20	6.4460	7.2854	7.7005	8.8310	9.4115	10.068	10.496	10.823	11.092	**20**
21	7.0017	7.8834	8.3186	9.5014	10.108	10.793	11.239	11.580	11.860	**21**
22	7.5680	8.4926	8.9462	10.180	10.812	11.525	11.989	12.344	12.635	**22**
23	8.1443	9.1095	9.5826	10.868	11.524	12.265	12.746	13.114	13.416	**23**
24	8.7298	9.7351	10.227	11.562	12.243	13.011	13.510	13.891	14.204	**24**
25	9.3240	10.369	10.880	12.264	12.969	13.763	14.279	14.673	14.997	**25**
26	9.9265	11.010	11.540	12.972	13.701	14.522	15.054	15.461	15.795	**26**
27	10.537	11.659	12.207	13.686	14.439	15.285	15.835	16.254	16.598	**27**
28	11.154	12.314	12.880	14.406	15.182	16.054	16.620	17.051	17.406	**28**
29	11.779	12.976	13.560	15.132	15.930	16.828	17.410	17.853	18.218	**29**
30	12.417	13.644	14.246	15.863	16.684	17.606	18.204	18.660	19.034	**30**
31	13.054	14.318	14.937	16.599	17.442	18.389	19.002	19.470	19.854	**31**
32	13.697	14.998	15.633	17.340	18.205	19.176	19.805	20.284	20.678	**32**
33	14.346	15.682	16.335	18.085	18.972	19.966	20.611	21.102	21.505	**33**
34	15.001	16.372	17.041	18.835	19.743	20.761	21.421	21.923	22.336	**34**
35	15.660	17.067	17.752	19.589	20.517	21.559	22.234	22.748	23.169	**35**
36	16.325	17.766	18.468	20.347	21.296	22.361	23.050	23.575	24.006	**36**
37	16.995	18.470	19.188	21.108	22.078	23.166	23.870	24.406	24.846	**37**
38	17.669	19.178	19.911	21.873	22.864	23.974	24.692	25.240	25.689	**38**
39	18.348	19.890	20.640	22.642	23.652	24.785	25.518	26.076	26.534	**39**
40	19.031	20.606	21.372	23.414	24.444	25.599	26.346	26.915	27.382	**40**
41	19.718	21.326	22.107	24.189	25.239	26.416	27.177	27.756	28.232	**41**

Table A.1 Erlang B table indicating maximum traffic load given the loss probability (E) and number of devices (n)

A	Number of devices n										A
	1	2	3	4	5	6	7	8	9	10	
3.00	.750000	.529412	.346154	.206107	.110054	.052157	.021864	.008132	.002703	.000810	**3.00**
3.05	.753086	.534550	.352104	.211655	.114346	.054933	.023376	.008833	.002985	.000909	**3.05**
3.10	.756098	.539585	.357975	.217178	.118671	.057771	.024946	.009574	.003287	.001018	**3.10**
3.15	.759036	.544519	.363764	.222676	.123027	.060670	.026576	.010356	.003612	.001136	**3.15**
3.20	.761905	.549356	.369475	.228145	.127409	.063628	.028265	.011180	.003959	.001265	**3.20**
3.25	.764706	.554098	.375107	.233584	.131816	.066642	.030012	.012046	.004331	.001406	**3.25**
3.30	.767442	.558748	.380660	.238991	.136244	.069710	.031818	.012955	.004728	.001558	**3.30**
3.35	.770115	.563308	.386137	.244365	.140690	.072831	.033681	.013908	.005150	.001722	**3.35**
3.40	.772727	.567780	.391536	.249703	.145152	.076001	.035601	.014905	.005599	.001900	**3.40**
3.45	.775281	.572167	.396861	.255006	.149627	.079220	.037577	.015947	.006076	.002092	**3.45**
3.50	.777778	.576471	.402110	.260271	.154112	.082484	.039608	.017033	.006581	.002298	**3.50**
3.55	.780220	.580693	.407286	.265498	.158606	.085791	.041694	.018166	.007114	.002519	**3.55**
3.60	.782609	.584837	.412389	.270685	.163105	.089140	.043834	.019344	.007678	.002756	**3.60**
3.65	.784946	.588905	.417419	.275832	.167608	.092527	.046026	.020567	.008272	.003010	**3.65**
3.70	.787234	.592897	.422379	.280938	.172113	.095952	.048269	.021837	.008898	.003281	**3.70**
3.75	.789474	.596817	.427269	.286002	.176617	.099412	.050564	.023153	.009555	.003570	**3.75**
3.80	.791667	.600665	.432090	.291024	.181119	.102905	.052907	.024515	.010245	.003878	**3.80**
3.85	.793814	.604445	.436843	.296003	.185616	.106428	.055298	.025922	.010967	.004205	**3.85**
3.90	.795918	.608157	.441529	.300939	.190108	.109980	.057737	.027376	.011724	.004552	**3.90**
3.95	.797980	.611803	.446149	.305831	.194592	.113559	.060221	.028875	.012514	.004919	**3.95**
4.00	.800000	.615385	.450704	.310680	.199067	.117162	.062749	.030420	.013340	.005308	**4.00**
4.05	.801980	.618904	.455195	.315483	.203531	.120789	.065320	.032010	.014200	.005718	**4.05**
4.10	.803922	.622362	.459623	.320243	.207983	.124437	.067933	.033644	.015095	.006151	**4.10**
4.15	.805825	.625761	.463990	.324958	.212422	.128103	.070586	.035323	.016027	.006607	**4.15**
4.20	.807692	.629101	.468295	.329628	.216846	.131788	.073278	.037046	.016994	.007087	**4.20**
4.25	.809524	.632385	.472540	.334254	.221254	.135488	.076008	.038812	.017998	.007591	**4.25**
4.30	.811321	.635614	.476726	.338835	.225645	.139202	.078774	.040621	.019038	.008120	**4.30**
4.35	.813084	.638788	.480855	.343371	.230019	.142928	.081574	.042472	.020115	.008674	**4.35**
4.40	.814815	.641910	.484926	.347862	.234373	.146666	.084408	.044365	.021229	.009254	**4.40**
4.45	.816514	.644980	.488941	.352309	.238707	.150412	.087274	.046299	.022380	.009861	**4.45**
4.50	.818182	.648000	.492901	.356712	.243021	.154167	.090171	.048273	.023567	.010494	**4.50**
4.55	.819820	.650971	.496806	.361070	.247313	.157927	.093096	.050286	.024792	.011155	**4.55**
4.60	.821429	.653894	.500658	.365385	.251583	.161693	.096050	.052338	.026054	.011843	**4.60**
4.65	.823009	.656770	.504458	.369655	.255830	.165462	.099030	.054428	.027352	.012559	**4.65**
4.70	.824561	.659600	.508206	.373882	.260054	.169234	.102035	.056555	.028687	.013304	**4.70**
4.75	.826087	.662385	.511904	.378065	.264253	.173007	.105063	.058719	.030059	.014077	**4.75**
4.80	.827586	.665127	.515552	.382206	.268427	.176780	.108115	.060917	.031467	.014879	**4.80**
4.85	.829060	.667826	.519150	.386304	.272576	.180551	.111187	.063150	.032911	.015711	**4.85**
4.90	.830509	.670483	.522701	.390359	.276700	.184320	.114280	.065417	.034391	.016572	**4.90**
4.95	.831933	.673100	.526204	.394372	.280797	.188086	.117390	.067717	.035907	.017464	**4.95**
5.00	.833333	.675676	.529661	.398343	.284868	.191847	.120519	.070048	.037458	.018385	**5.00**

Table A.2 Erlang B table indicating the loss probability given the traffic load (A) and number of devices (n)