TKK Dissertations in Information and Computer Science
Espoo 2008 TKK-ICS-D1

DATA ANALYSIS METHODS FOR CELLULAR NETWORK PERFORMANCE OPTIMIZATION

Pasi Lehtimäki

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Faculty of Information and Natural Sciences for public examination and debate in Auditorium TU2 at Helsinki University of Technology (Espoo, Finland) on the 3rd of April, 2008, at 12 noon.

Helsinki University of Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science

Teknillinen korkeakoulu
Informaatio- ja luonnontieteiden tiedekunta
Tietojenkäsittelytieteen laitos

Distribution:
Helsinki University of Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science
P.O. Box 5400
FI-02015 TKK
FINLAND
URL: http://ics.tkk.fi
Tel. +358-9-451 3272
Fax +358-9-451 3277
E-mail: series@ics.tkk.fi

ISBN 978-951-22-9282-0 (Print)
ISBN 978-951-22-9283-7 (Online)
ISSN 1797-5050 (Print)
ISSN 1797-5069 (Online)
URL: http://lib.tkk.fi/Diss/2008/isbn9789512292837/

Multiprint Oy
Espoo 2008

Lehtimäki, P. (2008): **Data-analysis methods for cellular network performance optimization**. Doctoral thesis, Helsinki University of Technology, Dissertations in Computer and Information Science, TKK-ICS-D1, Espoo, Finland.

Keywords: cellular network, radio network, radio resource optimization, information visualization, regression, clustering, segmentation, optimization

ABSTRACT

Modern cellular networks including GSM/GPRS and UMTS networks offer faster and more versatile communication services for the network subscribers. As a result, it becomes more and more challenging for the cellular network operators to enhance the usage of available radio resources in order to meet the expectations of the customers.

Cellular networks collect vast amounts of measurement information that can be used to monitor and analyze the network performance as well as the quality of service. In this thesis, the application of various data-analysis methods for the processing of the available measurement information is studied in order to provide more efficient methods for performance optimization.

In this thesis, expert-based methods have been presented for the monitoring and analysis of multivariate cellular network performance data. These methods allow the analysis of performance bottlenecks having an effect in multiple performance indicators.

In addition, methods for more advanced failure diagnosis have been presented aiming in identification of the causes of the performance bottlenecks. This is important in the analysis of failures having effect on multiple performance indicators in several network elements.

Finally, the use of measurement information in selection of most useful optimization action have been studied. In order to obtain good network performance efficiently, the expected performance of the alternative optimization actions must be possible to evaluate. In this thesis, methods to combine measurement information and application domain models are presented in order to build predictive regression models that can be used to select the optimization actions providing the best network performance.

Lehtimäki, P. (2008): **Data-analyysimenetelmiä matkapuhelinverkkojen suorituskyvyn optimointiin**. Tohtorin väitöskirja, Teknillinen korkeakoulu, Dissertations in Computer and Information Science, TKK-ICS-D1, Espoo, Suomi.

Avainsanat: matkapuhelinverkot, radioverkko, radioresurssien optimointi, informaation visualisointi, regressio, klusterointi, segmentointi, optimointi

TIIVISTELMÄ

Nykyaikaiset matkapuhelinverkot kuten GSM/GPRS ja UMTS tarjoavat yhä nopeampia ja monipuolisempia palveluita käyttäjilleen. Tämän seurauksena verkko-operaattorit joutuvat yhä haasteellisempien tehtävien eteen pyrkiessään tehostamaan rajallisten radioresurssiensa käyttöä asiakastyytyväisyyden takaamiseksi.

Matkapuhelinverkot keräävät jatkuvasti runsaasti mittausinformaatiota, jota voidaan käyttää verkon suorituskyvyn ja palvelun laadun analysointiin ja parantamiseen. Tässä väitöskirjassa tutkitaan erilaisten data-analyysimentelmien soveltamista tämän mittausinformaation käsittelyyn siten, että matkapuhelinverkon suorituskyvyn analysointi ja palvelun laadun parantaminen tehostuu.

Tässä väitöskirjassa on kehitetty käyttäjäkeskeisiä menetelmiä, jotka mahdollistavat usean matkapuhelinverkon suorituskykyä kuvaavan indikaattorin yhtäaikaisen seurannan ja analysoinnin. Tämä mahdollistaa sellaisten suorituskyvyn pullonkaulojen tunnistamisen, joilla on vaikutuksia useaan suorituskykyindikaattoriin.

Tässä väitöskirjassa on kehitty menetelmiä myös suorituskykyongelmien aiheuttajien tarkempaan selvittämiseen. Tämä on ensisijaisen tärkeää sellaisten vikatilanteiden tutkimisessa, joissa suorituskykyongelman aiheuttajalla on suora vaikutus useisiin eri indikaattoreihin ja verkkoelementteihin.

Väitöskirjan loppuosassa on tutkittu mittausinformaation tehokasta hyödyntämistä varsinaisten optimointitoimenpiteiden valitsemisessa. Jotta parhaaseen suorituskykyyn päästäisiin, on vaihtoehtoisten toimenpiteiden vaikutukset suorituskykyyn oltava ennakoitavissa. Tässä väitöskirjassa on esitetty menetelmiä, joiden avulla aiemmin kerättyä mittausinformaatiota ja sovellusalan teoreettisia malleja voidaan käyttää ennustavien regressiomallien muodostamiseen ja optimaalisten optimointitoimenpiteiden valitsemiseen.

Preface

This work has been done in the Laboratory of Computer and Information Science at the Helsinki University of Technology. I wish to thank my supervisor Prof. Olli Simula and my instructor Dr. Kimmo Raivio, for their support during my work at the laboratory. Also, I would like to thank Dr. Jaana Laiho, M.Sc Mikko Kylväjä and M.Sc Kimmo Hätönen at Nokia Corporation as well as Dr. Pekko Vehviläinen and M.Sc Pekka Kumpulainen at Tampere University of Technology for their cooperation during our work with the cellular network performance analysis.

I am also grateful to my parents for their continuous support during my studies at the HUT.

Pasi Lehtimäki
Otaniemi, March 10, 2008

Contents

Publications of the Thesis

Here is the list of the publications:

1. Pasi Lehtimäki, Kimmo Raivio, and Olli Simula. Mobile Radio Access Network Monitoring Using the Self-Organizing Map. In *Proceedings of the European Symposium on Artificial Neural Networks* (ESANN), pages 231-236, Bruges, April 24-26, 2002.

2. Jaana Laiho, Kimmo Raivio, Pasi Lehtimäki, Kimmo Hätönen, and Olli Simula. Advanced Analysis Methods for 3G Cellular Networks. *IEEE Transactions on Wireless Communications*, Vol. 4, No. 3, pages 930-942, May 2005.

3. Pasi Lehtimäki, Kimmo Raivio, and Olli Simula. Self-Organizing Operator Maps in Complex System Analysis. In *Proceedings of the Joint 13th International Conference on Artificial Neural Networks and 10th International Conference on Neural Information Processing* (ICANN/ICONIP), pages 622-629, Istanbul, Turkey, June 26-29, 2003.

4. Pasi Lehtimäki and Kimmo Raivio. A SOM Based Approach for Visualization of GSM Network Performance Data. In *Proceedings of the 18th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems* (IEA/AIE), pages 588 - 598, Bari, Italy, June 22-25, 2005.

5. Pasi Lehtimäki and Kimmo Raivio. A Knowledge-Based Model for Analyzing GSM Network Performance. In *Proceedings of the 6th International Symposium on Intelligent Data Analysis* (IDA), pages 204 - 215, Madrid, Spain, September 8-10, 2005.

6. Pasi Lehtimäki and Kimmo Raivio. Combining Measurement Data and Erlang-B Formula for Blocking Prediction in GSM Networks. In *Proceedings of the 10th Scandinavian Conference on Artificial Intelligence* (SCAI), Stockholm, Sverige, May 26-28, 2008 (accepted).

7. Pasi Lehtimäki. A Model for Optimisation of Signal Level Thresholds in GSM Networks. *International Journal of Mobile Network Design and Innovation* (accepted).

Author's Contributions

Here, the author's contributions in the publications of this thesis are outlined. In **Publication 1**, the author developed an extension to the previously existing cell classification method in order to make it more suitable for cell monitoring purposes, and applied the new method in the analysis of new data set. All computational work carried out in Publication 1 was performed by the author.

Publication 2 includes the application of the method presented in Publication 1 for a new data set. In Publication 2, the presented method is compared to a more traditional method to analyze cell performance. The author was responsible for performing the computational work and interpretation of the results associated with the approach presented in Publication 1.

In **Publication 3**, the cell monitoring approach presented in Publications 1 and 2 were modified in order to take the dynamic nature of the data into account when distinguishing between different states of the cells. The author was responsible for developing the new approach and implementing the software used in the analysis.

Publications 1–3 used only a limited amount of a priori knowledge about the problem domain. **Publication 4** presents a visualization process suitable for the analysis of the GSM network performance degradations. In this work, prior knowledge about the most common performance degradations is used to focus on the most interesting parts of the measurement data to be visualized for the user. The author was responsible for developing the method, running all the technical computing and interpreting the results.

In **Publication 5**, a knowledge-based model is constructed in order to take the prior knowledge into account more efficiently. The available raw data was used to estimate the free parameters of the knowledge-based model. Finally, the estimated model is visualized as a hierarchical cause-effect chain representing the development of the failures in the cellular network. The author was responsible for developing the knowledge-based model and estimation of the model parameters. The visualization of the cause-effect chains and interpretation of the results were carried out by the author.

In **Publication 6**, the gap between the GSM network measurements and theoretical calculations associated with cell capacity is discussed. A method that uses both the Erlang-B formula as well as the network measurements is developed in order to predict the amount of blocking in SDCCH and TCH channels at differ-

ent amounts of demand. The author was responsible for developing the method, performing all of the computations and analyzing the results.

In **Publication 7**, the prediction method developed in Publication 6 is applied in automated parameter optimization. In addition, a model describing the effects of parameter adjustments to the performance data is defined. The author was responsible for developing the model and running the required computations.

Abbreviations

3GPP	3rd Generation Partnership Project
ANFIS	Adaptive Neuro-Fuzzy Inference System
AuC	Authentication Centre
BMU	Best-Matching Unit
BR	Blocking Rate
BS	Base Station
BSC	Base Station Controller
BSS	Base Station Subsystem
BTS	Base Transceiver Station
CDMA	Code Division Multiple Access
CN	Core Network
CSSR	Call Setup Success Ratio
DCR	Dropped Call Rate
FDMA	Frequency Division Multiple Access
FER	Frame Error Rate
GPRS	General Packet Radio System
GSM	Global System for Mobile
HLR	Home Location Register
HO	HandOver
HOSR	HandOver Success Ratio
IMSI	International Mobile Subscriber Identity
KPI	Key Performance Indicator
LMS	Least Mean Square
LU	Location Update
MDS	Multidimensional Scaling
MLP	Multilayer Perceptron
MS	Mobile Station
MSC	Mobile Switching Center
NMS	Network Management System
NRM	Network Reference Model
NSS	Network SubSystem
OMC	Operation and Maintenance Center
OSS	Operations Support System
PCA	Principal Component Analysis
RAA	Resource Allocation Algorithm
RNS	Radio Network Subsystem
RRM	Radio Resource Management
RMSE	Root Mean Square Error

SDCCH	Standalone Dedicated Control CHannel
SMS	Short Message Service
SOM	Self-Organizing Map
TCH	Traffic CHannel
TCP	Transmission Control Protocol
TDMA	Time Division Multiple Access
TMF	TeleManagement Forum
TMN	Telecommunications Management Network
TOM	Telecom Operations Map
TRX	Transceiver/Receiver
UE	User Equipment
UMTS	Universal Mobile Telecommunications System
UTRAN	Universal Terrestrial Radio Access Network
VLR	Visitor Location Register
WCDMA	Wideband Code Division Multiple Access

Chapter 1

Introduction

1.1 Motivation and overview

The number of mobile network subscribers increases constantly. At the same time, more efficient network technologies are developed in order to provide faster and more advanced data communication services for the subscribers. As a result, the current and new network technologies operate in parallel, making cost-effective network management more and more challenging. The network operator should be able to manage the radio resources to meet the current as well as the future demand without expensive investments to infrastructure.

This thesis presents approaches in which data-analysis methods, cellular network measurement data and application domain knowledge are combined in order to solve practical radio resource management problems. In practice, this involves the development of methods suitable for the detection of abnormal failures and performance bottlenecks from multivariate measurement data. In addition to finding bottlenecks in network performance, it is necessary to identify the cause or the limiting factor for the performance, and to select a management action in order to remove the causes of the failures and performance degradations. The first portion of methods focus on visualization of performance data for human optimizers. Also, methods to predict network performance under different conditions are presented. Finally, an automated method to select the optimal configuration adjustment for the network is presented.

1.2 Contributions of the thesis

The main contributions of this thesis are:

- the development of data visualization methods for expert-based optimization of UMTS network plans.

- the development of data visualization methods for expert-based optimization of operative GSM networks,
- the development of knowledge-based predictive models for optimization of operative GSM networks.

1.3 Outline of the thesis

The outline of this thesis is as follows. In chapter 2, the problem domain is introduced in more detail. That is, the GSM and UMTS network architectures as well as the management of the networks are shortly outlined. The focus of chapter 2 is on the wide range of methods developed for the optimization of radio resource usage. In chapter 3, the process monitoring problem is discussed, and the variety of research fields providing tools for process management is shortly reviewed. The emphasis is on various types of data-analysis methods and their usage in process monitoring. In chapter 4, the results of applying advanced data analysis methods to improve mobile network performance are presented. Finally, in chapter 5, the conclusions are made.

Chapter 2

Radio Resource Management in Cellular Networks

In this chapter of the thesis, the domain of application, that is, the cellular networks and their management are discussed. Firstly, the cellular network architectures including GSM and UMTS are described. In the following section, a four-layer model for the management of telecommunications networks is described, the focus being on network management layer of the model.

Then, the focus is turned on the network management functions associated with the radio network part of the system, and especially, the management of the radio resources. The radio resource management techniques are discussed in several sections. Firstly, different definitions for radio resource management are given. Then, a framework for using various radio resource management techniques in a systematic performance optimization process is presented.

In the remaining part of this chapter, the non-real time (offline) optimization loops are discussed. Especially, the focus is on data-driven approaches in which the network performance is defined by the BTS level measurements over relatively long time periods and the optimization is strongly based on intelligent processing of available measurement data collected from the network elements. A comprehensive literature study of most widely used approaches for data-driven non-real time optimization in both operative GSM networks and UMTS network simulations is conducted.

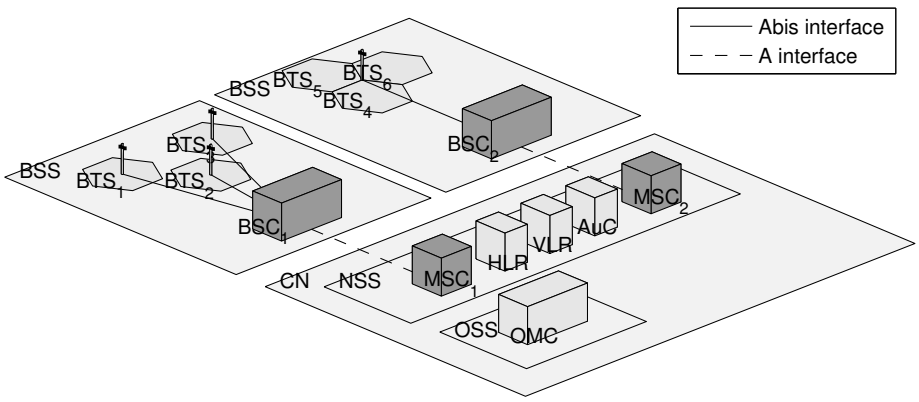


Figure 2.1: GSM network architecture.

2.1 Cellular Network Architectures

2.1.1 GSM Network

A Global System for Mobile communications (GSM) network consists of Network and Switching Subsystem (NSS), Base Station Subsystem (BSS) and Operations SubSystem (OSS). In Figure 2.1, the architecture of GSM network is depicted.

The BSS contains all the radio-related capabilities of the GSM network, being responsible for establishing connections between the NSS and the mobile stations (MSs) over the air interface. For this purpose, the BSS consist of several Base Station Controllers (BSCs) that can manage the operation of several Base Transceiver Stations (BTSs) through the Abis interface. Up to three BTSs can be installed on the same site, and usually, the BTSs are placed to cover separate sectors around the site. Each BTS is responsible for serving the users in its own coverage area, also called the cell, over the air interface. Depending on the user density in the cell served by a BTS, one or more Transceiver/Receiver pairs (TRXs) operating on separate radio frequencies can be installed to a BTS in order to obtain the required number of communication channels. In (Kyriazakos and Karetsos, 2004), the architecture of GSM network is described in more detail.

In GSM, the available radio frequency band is divided between different subscribers using Frequency Division Multiple Access (FDMA) and Time Division Multiple Access (TDMA) techniques. In practice, this means that up to 8 subscribers may operate on a single physical frequency, and the 8 users using the same physical frequency are separated by allocating different time slots for each of the users.

On a single physical channel, several logical channels operate in parallel in order to establish connections over the air interface. The most important logical channels used to implement the basic services such as voice calls, short message service (SMS) messages and location updates (LUs) include Standalone Dedicated

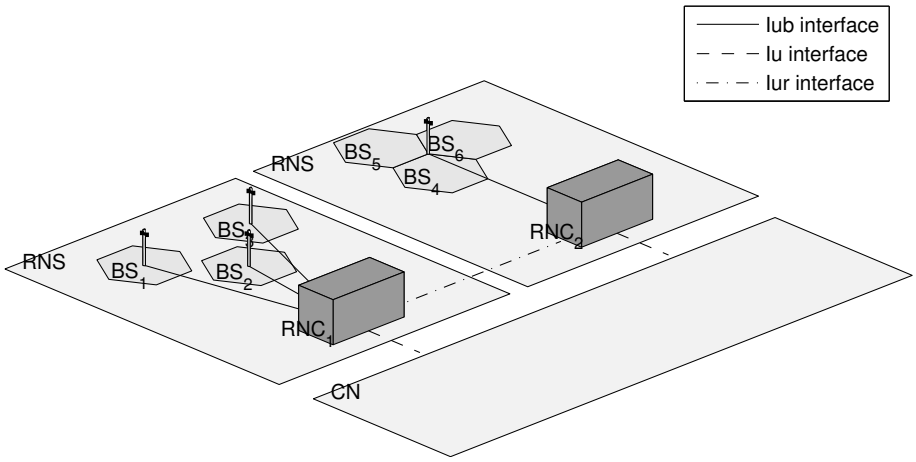


Figure 2.2: UMTS network architecture.

Control CHannel (SDCCH) and Traffic CHannel (TCH). For example, in voice call establishment the SDCCH is occupied during the negotiation phase in which the actual TCH channel carrying the voice data is allocated. The SMS messages and the LUs are usually transmitted in SDCCH, but TCH may be used for that purpose during congestion situations.

The role of the NSS is to operate as a gateway between the fixed network and the radio network. It consists of Mobile Switching Centre (MSC), Home Location Register (HLR), Visitor Location Register (VLR) and Authentication Centre (AuC). The MSC acts as a switching node, being responsible for performing all the required signaling for establishing, maintaining and releasing the connections between the fixed network and a mobile user. The Home Location Register (HLR) is a database that includes permanent information of the subscribers. This information includes International Mobile Subscriber Identity (IMSI) and for example, the identity of the currently serving VLR needed in routing the mobile-terminated calls. The VLR contains temporary information concerning the mobile subscribers that are currently located in the serving area of the MSC, but whose HLR is elsewhere. The AuC is responsible for authenticating the mobile users that try to connect to the GSM network. Also, a mechanism used to encrypt all the data transmitted between the mobile user and the GSM network are provided by the AuC.

The OSS consist of Operation and Maintenance Center (OMC) that is responsible for monitoring and controlling the other network elements in order to provide adequate quality of service for the mobile users. In other words, it measures the performance of the network and manages the network configuration parameters and their adjustments. Therefore, most of the methods and techniques discussed in this thesis are mostly implemented in the OSS.

2.1.2 UMTS Network

The Universal Mobile Telecommunications System (UMTS) consist of Universal Terrestrial Radio Access Network (UTRAN) and the Core Network (CN) connected via the Iu interface. In Figure 2.2, the UMTS network architecture is depicted.

The UTRAN consist of several Radio Network Subsystems (RNSs) that are responsible for connecting the User Equipment (UE) to the network. The RNS consist of Radio Network Controllers (RNCs) and Base Stations (BSs). RNC is the switching and controlling element of the UTRAN, located between the Iub and Iu interface. The RNC controls the logical resources of its BSs and is responsible, for example, to make handover decisions. The RNC and the BSs are connected through the Iub interface, while the RNCs within the same UTRAN are connected via the Iur interface. For more information about UMTS architecture, see (Kaaranen et al., 2005).

The main tasks of BS include radio signal receiving and transmitting over the Uu interface (air interface), signal filtering and amplifying, modulation/spreading aspects as well as channel coding and functionalities for soft handover. The BS includes transceiver/receiver equipment to establish radio connections between UEs and the network.

In UMTS, the available frequency band is divided between the users on the basis of Wideband Code Division Multiple Access (WCDMA) technique. In (W)CDMA, the data for each user is transmitted in the whole frequency band, and no separation in frequencies nor time is present. Instead, the user data is multiplied by a code sequence unique for each user (code chip-rate is higher than the bit-rate of the data). After multiplying the user data with the corresponding codes, a single spread spectrum signal is obtained and transmitted through the air interface. At the receiver, the spread spectrum signal is multiplied by the same, user specific codes which decodes the original data for each user.

The use of (W)CDMA technique causes the capacity of the UMTS network to be a more difficult issue to handle, and no clear separation between network capacity and coverage can be made. Also, the UMTS radio network becomes interference limited rather than frequency limited as is the case with GSM networks.

2.1.3 Telecommunications Management Network

The TeleManagement Forum (TMF) is an international organization consisting of service providers and suppliers from the communications industry. In order to improve and accelerate the availability of network management products and compatibility between products from different vendors, the TMF provides highly authoritative standards and frameworks for the management of telecommunication business operations.

The Telecommunications Management Network (TMN) model, as proposed by the TMF, gives a general framework for the processes involved in telecommunication

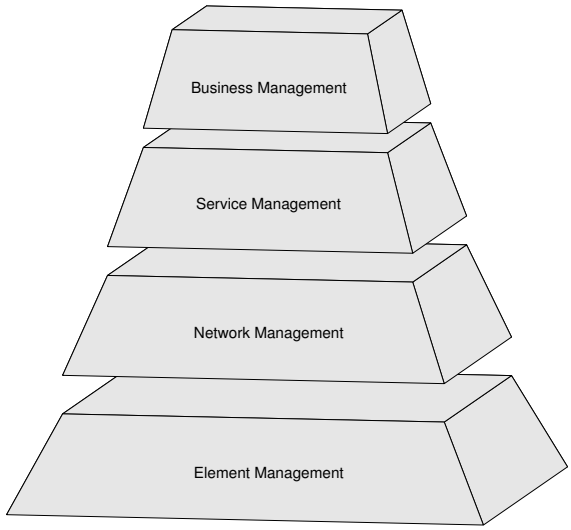


Figure 2.3: The TMN model.

business management. The same framework is adopted by 3rd Generation Partnership Project (3GPP) in order to create a globally applicable 3G generation cellular system known as the UMTS. According to Laiho et al. (2002c), the layers of TMN (see Figure 2.3) consist of

- business management layer,
- service management layer,
- network management layer, and
- element management layer.

The business management layer can be seen as goal setting rather than goal achieving layer, in which high-level planning, budgeting, goal setting, executive decisions and business-level agreements take place. For this reason, the business management layer can be seen as strategical and tactical management unit, instead of operational management like the other layers of the TMN model. The service management layer is concerned with tasks including subscriber data management, service and subscriber provisioning, accounting and billing of services as well as development and monitoring of services. The network management layer manages individual network elements and coordinates all network activities and supports the demands of the service management layer. Network planning, data collection and data analysis, as well as optimization of network capacity and quality are the main tasks of this layer. The element management layer monitors the functioning of the equipment and collects the raw data.

In addition to the TMN, the TMF has defined a Telecom Operations Map (TOM) in which the processes of the TMN layers are defined in more detail. The TOM links each of the high-level processes into a set of component functions and identifies the relationships and information flows between the component functions.

The above mentioned frameworks and guidelines help the service providers to define the organization of the human resources and the tasks related to different parts of the organization. In practice, the tasks adopted from the TOM require the use of software tools, which are implemented by the Network Management Systems (NMS). The NMS consist of all the necessary tools, applications and devices that assist the human network managers to maintain operational networks. The NMS tools are based on open interfaces in order to establish long-term supportability for the tools, compatibility between tools from different vendors, but also to enable rapid development of high quality tools and technologies. For this reason, the 3GPP has defined a Network Resource Model (NRM). The NRM defines object classes, their associations, attributes and operations as well as defines the object structure which is used in, for example, management of configuration and performance data.

In this thesis, the focus is on the network management layer of the TMN model, and especially, the activities focusing on radio network part of the GSM and UMTS networks. This is discussed in the next section.

2.2 Radio Resource Management

2.2.1 Control Loop Hierarchy

The objective of the radio resource management (RRM) is to utilize the limited radio spectrum and radio network infrastructure as efficiently as possible. The RRM involves strategies and algorithms for controlling parameters related to transmission power, channel allocation, handover criteria, modulation scheme, error coding scheme, etc. Most of the RRM algorithms operate in a loop, constantly monitoring the current state of the system, and if necessary, control actions are triggered in order order to improve radio resource usage.

In (Laiho et al., 2002c), a general hierarchy for different RRM techniques is presented in which the RRM loops are classified into three layers according to the response time of the algorithm (length of a single iteration) as well as the amount of information needed by the algorithm (see Figure 2.4). In the bottom layer, the fast real-time RRM loops ensure the adequate quality of the currently active radio links. These techniques are also called as the resource allocation algorithms (RAA) and examples include serving cell selection and transmission power control. In (Zander, 2001), a wide range of RRM techniques that belong to the fast real-time loops are presented. The fast real-time RRM algorithms for power control, channel allocation and handover control focus on maximizing operators revenue, that is, the incomes of the operator. The maximization of the incomes is closely connected with the concept of service quality, since only the services that meet the quality of service requirements contribute to the income of the operator. The quality of service are defined for each service, user and link separately, and therefore the fast real-time RRM loops are designed to meet these quality requirements for each link separately. Therefore, the maximization of the incomes implies that the number and duration of the communication links filling the quality requirements must be maximized. The quality of each link is optimized or controlled

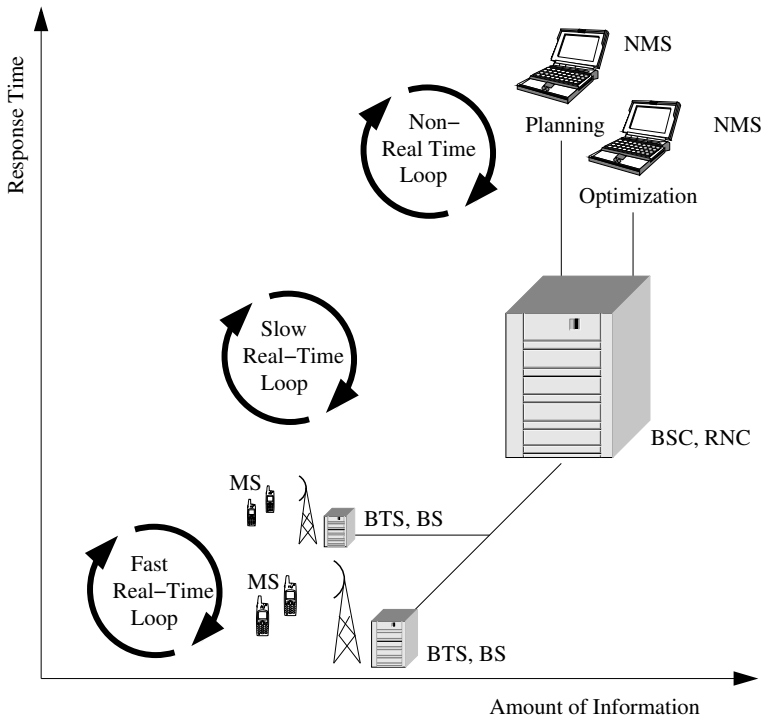


Figure 2.4: The hierarchy for RRM control loops.

separately, based on short number of measurements, typically averaged over short time intervals. These RRM loops are implemented in the MSs and BTSs.

The middle layer of RRM algorithms consist of slow real-time RRM loops implemented in the BSS or RNC of the network. Admission control and handover control algorithms are typically implemented at this layer. The slow real-time control loops perform RRM actions that are needed to maintain link-level performance, such as triggering a BSC initiated handovers in order to support seamless mobility. In (Kyriazakos and Karetsos, 2004), a wide range of adaptive dynamic RRM techniques are presented that belong to the second layer of the control layer hierarchy. These are fully automated control methods, but they are more closely related to improving the average performance of the network, that is, their operation affects on all links currently active in the cell. The emphasis in these dynamic reconfiguration methods is in congestion control, that is, making dynamic reconfigurations to the system when the system becomes highly loaded for relatively short time periods. These techniques are usually triggered several times during one day, and the length of congestion period typically lasts no longer than minutes. Examples of such methods for GSM networks include halfrate/fullrate tradeoff, forced handovers, dynamic cell resizing and RX-level adjustment.

The top layer of the hierarchy consist of statistical non-real time control loops implemented in the NMS. These loops are initiated and iterated offline and they are used to improve radio resource usage in both operative networks but also in network simulations taking place in the network planning phase. In this the-

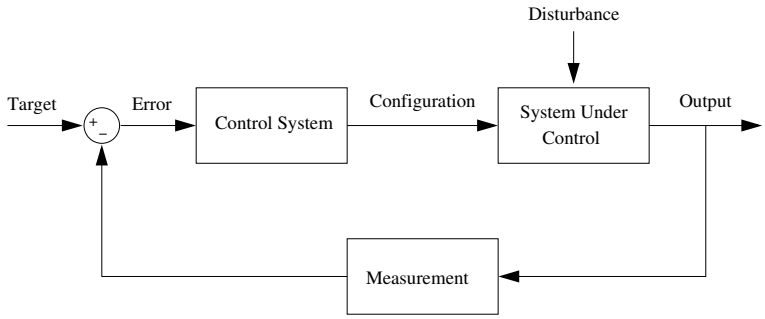


Figure 2.5: A Framework for control loop design based on control engineering.

sis, the non-real time (offline) control loops based on statistical measurements are called performance optimization techniques rather than control loops. These performance optimization techniques have the longest response time but more information sources (variables, network elements) than the control loops at the two bottom layers. The aim of these algorithms is to find the optimal configurations (in the long run) without adapting to the natural, daily (short-term) variations in the traffic patterns. The performance is measured in terms of Key Performance Indicators (KPIs) that describe, for example, various failure rates over long time periods, and are averaged or summed over all active users. The number of essential KPIs that need to be analyzed is typically between 10 and 30, and the number of raw network measurements related to the most important KPIs is hundreds or thousands. For more information about RRM techniques for wireless network planning and optimization, see (Laiho et al., 2002c; Kyriazakos and Karetsos, 2004; Lempiäinen and Manninen, 2003).

2.2.2 A Framework for RRM Control Loop

In (Halonen et al., 2002), a control engineering framework for RRM control loops aiming in enhancements in radio system is presented. The aim of the control loops is to adapt the wireless network configuration parameters so that the performance of the network is repeatedly improved. In other words, it is a process in which the radio resource management algorithms and techniques are systematically applied in a loop in order to improve system performance.

In Figure 2.5, a block diagram illustrating the control loop framework is depicted. The control engineering approach regards the

- configuration parameters as system input,
- the user generated traffic is interpreted as unpredictable disturbance for the system under control,
- the performance of the network in terms of statistical counters or KPIs is the output of the system under optimization, and
- the control system is responsible for generating the improved configuration

parameters given the deviation between the current performance of the system and the target performance.

The performance optimization proceeds by measuring the performance of the system with the current traffic load, and comparing it to the target performance. The error or deviation between the current and target performance is fed to the control module, which is responsible for producing a new system configuration in which the gap between measured and target performance is decreased. This loop can be iterated until the target performance is met.

Separate optimization loops can be developed for different subsystems of the mobile network so that the optimization of the performance of one subsystem has a minimal impact on other subsystems.

In this thesis, the focus is on the top level of RRM methods, that is, on the statistical non-real time performance optimization. Especially, the focus is on data-driven approaches for optimization and planning. The control engineering framework is used to distinguish the major functional blocks of the performance optimization approaches and to analyze the implementation of the individual blocks and the relationships between the functional blocks. This is important in order to fully understand the benefits of data-driven approaches when implementing RRM control loops based on information extracted from massive data records.

2.2.3 Non-Real Time Performance Optimization

The performance optimization approaches, that is, the statistical non-real time control loops, can be divided into:

- expert-based,
- adaptive autotuning,
- measurement-based, and
- predictive methods.

The most straightforward approach presented in the literature is based on performance data visualization and active role of human expert in analyzing the data. In this expert-based approach, the user is responsible for detecting the performance degradations from the presented graphical figures. Then, the user should be able to analyze the cause of the performance degradations. Finally, the user is responsible for deciding the optimal configuration among the alternatives based on his/her understanding of the performance bottleneck. In other words, the mapping from the alternative configurations to their expected performances takes advantage of reasoning that need not be represented explicitly as a software algorithm. Therefore, the tasks of the control system are actually performed by human resources. The expert based approach is focused on fault detection and diagnosis and representing related information in graphical form. The user is then responsible for analyzing the figures and making the control action decisions.

In the adaptive autotuning approach, the performance of the network under the current configuration is measured by collecting data from the output of the system under control. Then, the control system is responsible for intelligent decision-making in order to update the configuration (parameters) towards better ones. Finally, the new configuration is installed and new performance data is gathered. This loop is repeated until convergence of configuration parameters is obtained. The system under control must be a real operating network or a simulator and the performance of the current configuration should be possible to measure efficiently. The control system is usually equipped with expert-defined control rules that aim in selecting improved configuration by exploiting prior knowledge about the performance bottlenecks. The difference between adaptive autotuning and slow real-time control loops like congestion relief algorithms is that the slow real-time control loops are continuously active, and they can be triggered at any time. The adaptive autotuning methods have a clear starting point and duration, and the configurations achieved during the adaptation are fixed after the adaptation process.

The third approach is based on the use of network measurements. The measurement data allows the determination of the mapping between the alternative configuration parameters and the system performance explicitly. In this approach, there is no clear feedback from output variables to the updated configuration, but instead, the improved (optimal) configuration is directly computed from the target variables.

The fourth approach is based on developing predictive regression models using past measurements extracted from the network. The estimated models allow the prediction of network performance under unseen configurations and therefore, such models are useful in automated performance optimization. In this approach, a model for the system under control is obtained from past measurement data. The system model enables the computation of the performance with different configuration adjustments directly and the system model remains unchanged during the optimization process. No feedback loop is needed to test configurations during the decision making about the new configuration.

It should be mentioned here, that some of the autotuning methods developed for performance optimization can be implemented as fully automated slow real-time loops. Also, some of the autotuning methods developed for planning purposes may be directly applied in optimization of operational networks as a non-real time control loop or a slow real-time loop.

In the following sections, examples of above mentioned approaches for parameter optimization in operative GSM networks and UMTS network simulations performed during network planning are presented. In particular, the focus is on data-driven techniques in which real or simulated network data is used as a source of information in decision making regarding the optimal control action.

2.3 Cellular Network Performance

In cellular network planning phase, for example, no performance measurements like KPIs from live network are available and therefore, predictions of performance must be used in decision making. Also, early testing of new optimization algorithms in operative cellular networks may not be desirable in order to avoid unnecessary risks of confusing the current network configuration. For these reasons, most of the algorithms and methods developed for optimization of GSM and UMTS network performance are developed and tested with network simulators. From the optimization algorithm point of view, there is no significant difference in whether the performance of live network or simulated network is optimized. Therefore, it is possible to use most of the presented methods in both optimization of live network, but also in final phases of the network planning process which is strongly based on simulators. Firstly in this section, some basic theoretical models used for network performance predictions are reviewed. Then, the performance measurements of a real mobile network are introduced. Finally in Section 2.4, the performance optimization approaches presented in the literature are outlined.

2.3.1 Performance Prediction

Path Loss Models

The most frequently used models associated with network planning and simulation include various path loss models. The purpose of the path loss models is to compute the amount of attenuation in the radio signal on the propagation path. A model based on pure theoretical derivations is the ideal path loss model where link gain $G(R)$ at distance R is defined by

$$G(R) = \frac{C}{R^{\alpha}} \tag{2.1}$$

where C is an antenna parameter and α is a parameter describing the propagation environment (Zander, 2001). In decibel scale, the amount of path loss at distance R is

$$L(R) = 10 \log G(R) = 10 \log C - 10\alpha \log R. \tag{2.2}$$

The value $\alpha = 2$ is used for free space and values from 3 to 4 are used in urban environments.

Another widely used path loss model for the urban environments is the Okumura-Hata model

$$\begin{aligned} L(R) &= 26.16 \log f + (44.9 - 6.55 \log h_{BTS}) \log R \\ &\quad -13.85 \log h_{BTS} - a(h_{MS}) + 69.55, \end{aligned} \tag{2.3}$$

where f is the carrier frequency, R is the distance between BTS and MS antennas, h_{BTS} is the height of the BTS antenna and h_{MS} is the height of the MS antenna (Hata, 1980). The function $a(h_{MS})$ can be selected from three alternatives depending on the carrier frequency and the type of the operating environment (large, medium or small city).

The above mentioned path loss models are used, for example, in initial network planning phase (dimensioning) in order to compute the maximum operating range of 3G network base stations for given maximum transmission powers (Holma and Toskala, 2004). In addition, path loss models are used to predict the relationship between the original signal and interference, having direct impact on signal quality in the radio links.

Capacity of GSM network

One of the most important performance criteria in mobile networks is the availability of resources (communication channels) with variating traffic load. In order to predict the amount of traffic that can be supported, the blocking probability is calculated. Traditionally, the Erlang-B formula is used as a model when computing the amount of blocking with different number of channels and demand (Cooper, 1981). For example, consider the case when the incoming transactions follow the Poisson arrival process with arrival rate λ, transaction length is exponentially distributed with mean $1/\mu$ and the number of channels N_c is finite. The probability that n channels are busy at random point of time can be computed using the Erlang-distribution

$$p(n|\lambda, \mu, N_c) = \frac{(\lambda/\mu)^n/n!}{\sum_{k=0}^{N_c}(\lambda/\mu)^k/k!}. \tag{2.4}$$

Using this formula, it is possible to calculate the amount of traffic that is supported with given blocking probability. If the Erlang-B formula is applied in network planning (dimensioning), the number of communication channels that are needed is computed in order to meet the traffic and blocking probability requirements.

Capacity of UMTS network

Since the UMTS network supports several bit-rates and the capacity in networks using WCDMA multiplexing is interference limited, the estimation of the capacity is based on calculations in which the transmission powers and path losses for each active radio link must be known. The capacity of a WCDMA base station is measured, for example, in terms of uplink loading

$$\eta_{ul} = (1+i)\sum_{j=1}^{N}\frac{1}{1+W/\left[(E_b/N_0)_j R_j v_j\right]} \tag{2.5}$$

where $(E_b/N_0)_j$ is the signal to interference ratio of radio signal for user j, W is the chip-rate, R_j is the bit-rate of user j and v_j is the activity of the user j. It should be noted here, that the consumption of wireless network resources caused by a single user depends on the bit-rate of the service used, the speed of the user, and the path loss (distance) influencing the radio signal. Also, the number of users in the adjacent cells affect on cell capacity due to the interference originating from the surrounding cells. The higher the bit-rate of the used service, the greater the load factor for single user becomes. The larger the load factors of individual active users are, the less new users can be allocated to the system.

This load factor can be directly used to estimate the amount of interference (noise rise) in addition to the basic noise floor caused by thermal noise.

In networks based on WCDMA , the estimation of the blocking probability must take the occurrences of soft handover into account. The so-called soft capacity indicates the amount of traffic that can be supported by a WCDMA cell with prespecified blocking probability. A computational procedure for evaluating the soft capacity based on Erlang-B formula is described in (Holma and Toskala, 2004).

2.3.2 Performance Measurements

The operation of the cellular network can be interpreted to consist of a sequence of events. From the network operation point of view, certain events are closely associated with bad performance, lack of resources or failures. The number of undesired events during a measurement period (typically one hour) are stored by a set of corresponding counters. In this thesis, the performance of operative cellular networks is determined by the number of undesired events, such as blocked channel requests and dropped calls. The optimization of operative networks aims in minimizing the number of occurrences of such events.

Since the raw data consisting of the values of the counters at different time periods is impractical to analyze as such, a wide range of KPIs are defined that aim in more intuitive performance analysis (Halonen et al., 2002; Kyriazakos and Karetsos, 2004). The most important KPIs include the SDCCH and TCH Blocking Rates (SDCCH/TCH BR), Dropped Call Rate (DCR), Call Setup Success Rate (CSSR) and Handover Success Rate (HOSR). These KPIs can be computed in different ways depending on the network vendor and the operator, but in general, they are computed by dividing the number of undesired events with the total number of attempts. For example, the DCR can be computed by dividing the number of dropped calls due to inadequate radio link quality and other similar reasons in a measurement period with the total number of calls in the measurement period.

However, the KPIs are mostly useful in fault detection rather than studying the actual cause of undesired events. For example, the dropped calls can be caused by failures in the A, Abis or air interfaces or any other related network element. Observing a certain value of DCR does not indicate which of the network element of interface caused the calls to be dropped. In order to isolate the cause, the counter data must be studied. However, the use of counter data not always gives the actual cause for the undesired events. For example, the cause for bad radio link quality can be shadow fading or multipath fading, but also, interference originating from other cells operating on the same frequency has an effect on signal quality. There are no measurements available that could be used to distinguish between these different causes for bad signal quality.

Another difficulty with the use of KPIs in performance analysis is caused by strong interactions between close-by BTSs. For example, the HOSR can be on unacceptable level, but further analysis might reveal that the problem occurs mostly during the outgoing handovers into a certain close-by BTS. A possible explanation for failed incoming HOs may rely in lack of TCHs in the target cell. Therefore, the

capacity problems in a close-by BTS may be visible in HOSR of a BTS, and it is necessary to simultaneously analyze several KPIs from close-by BTSs in order to fully recognize the location of the bottleneck in network performance. Similar dependencies between KPIs may exist between BTSs on the same physical carrier, or between the BTSs sharing some other physical resources.

2.4 Data-Driven Performance Optimization

2.4.1 Expert-Based Approach

In the literature, a wide range of radio network optimization methods exploiting visualization and expert decision making are proposed. The expert-based approach has been studied using performance data from network simulators and live networks.

In (Zhu et al., 2002), a set of indicators are proposed for the detection of overloaded cells. The method is used to optimize pilot power settings in an UMTS network in order to obtain better network performance. A dynamic network simulator is used to demonstrate the benefits of the proposed indicators. The above method is based on designing good indicators that can be visualized in very simple form, such as time-series data or histogram. However, the visualization of performance data of a wireless network has been tackled also with advanced data analysis methods such as neural networks. For example, the works by Raivio et al. (2001) and Raivio et al. (2003) demonstrate the use of clustering and neural networks in the visualization of operational states computed from multivariate uplink performance data of a WCDMA network. Also, the problem of finding similar base stations according to uplink performance is tackled, enabling the simplification of autotuning of key configuration parameters.

The work presented in Publication 1 is a modification to the above mentioned method. In Publication 1, the downlink performance degradations in WCDMA network simulation are detected during continuous monitoring of the state of the network. The current states of the BSs are classified according to the shape of the distribution of the related performance variables over short time periods. The end-user is provided a simplified description of the possible states of the BSs. Then, the user is able to find out which of the obtained states are inappropriate for the BSs. By using a digital map roughly describing the radio signal propagation conditions in the network area, the end-user is responsible for deciding what is the limiting factor for the network performance. Also, the end-user is responsible for deciding how the configuration should be adjusted in order obtain better performance for the planned network.

In Publication 2, the same methodology has been applied for the analysis of uplink performance of a microcellular network scenario and comparisons to performance analysis based on WCDMA loading equations are presented. The presented method can also be used in cell grouping, aiming in more efficient optimization of large amount of BSs since similar BSs may share the same configuration parameters. In (Laiho et al., 2002b), the same methodology has been applied for

the analysis of uplink performance in a microcellular network scenario, but also, the flexibility of the presented methodology is demonstrated by using the same method in the analysis of uplink and downlink performance simultaneously both in micro- and macrocellular network scenarios. The general use of cell grouping in the network optimization process is discussed in (Laiho et al., 2002a) and (Laiho et al., 2002b).

In Publication 3, the problem of continuously monitoring the states of the cells is approached from a new perspective. The definition of the BS state used in performance monitoring and state classification is based on dynamics of the link performance. A linguistic description of the dynamics of the alternative BS states is provided. Using them, the user is responsible for deciding which states are inappropriate and how the BSs in such states should be adjusted.

Vehviläinen (2004) and Vehviläinen et al. (2003) give a comprehensive study for exploiting data mining and knowledge discovery methods in performance analysis of operative GSM networks. The use of soft computing techniques like rough sets, classification trees and Self-Organizing Maps for the easy analysis of important features of the performance data is discussed. In addition, methods to use a priori knowledge, that is, the application domain experience in the analysis process is greatly emphasized.

In (Multanen et al., 2006), a method to use KPI data from live GSM network to find city-sized low performing subnetworks from a very large network areas is presented. This method applies well for determining locations of performance degradations in which optimization should take place.

In Publications 4-5, the analysis of performance degradations in city-sized GSM networks is studied. In Publication 4, a method to analyze the real KPI data of an operating GSM network with neural network based visualization process is described. Several different kinds of visualizations are provided in order to help user's task to analyze alternative causes for the performance degradations. The user is responsible for deciding how the configuration should be adjusted in order to prevent the same performance degradations to appear in the future. The main benefit of this approach is that the same methods can be applied in optimization of many different configuration parameters and network subsystems with low costs as long as required expertise is at hand. For example, in Publication 4, the same method is used to analyze TCH and SDCCH capacity problems without any major modifications to the method. Also, the same methods are available for the analysis of operative networks as well as for the analysis of simulated data being the output of, for example, network planning activities.

However, the use of the expert-based methods requires extensive knowledge about the problem domain and the optimization actions proposed by different experts may not be consistent. Also, the main disadvantages of these methods include the inability to observe how close-by cells interact during faulty situations. Furthermore, the visual analysis of KPI data may be misleading, since the averaging performed during KPI computation lose essential information about the true source of the performance degradation.

In order to cope with these difficulties, a data-driven approache using the counter

data instead of KPI data may be used. In Publication 5, an explicit description of the network performance development is presented in order to study the cause-effect chains in which the bad performance is developing. Possible cause-effect relationships between the most important counters are searched from the data and presented for the end user in a tree-structured cause-effect chains. Also, the main objective was to study the bad performance situations in which the cause is in fact located in close-by BTSs.

Ricciato et al. (2005) have presented methods to discover bottlenecks in performance of live UMTS network. Several indicators of bottlenecks in TCP (Transmission Control Protocol) packet data transmissions are proposed. The visualization used in this work is based on plotting the proposed indicators in the form of time-series in which the presence of the bottlenecks are easily captured.

2.4.2 Adaptive Autotuning Approach

One of the most widely adopted approaches for performance optimization is the adaptive autotuning approach in which the initial configuration parameter values are repeatedly updated with better ones until convergence is obtained. The methods following this approach repeatedly change the configuration and measure the improvement in real network or apply a network simulator.

In (Olofsson et al., 1996) and (Magnusson and Olofsson, 1997), the design of optimal neighbor lists used by handover algorithms in GSM networks is discussed. The aim was to design an automatic procedure in order to avoid manual adjustment of neighbor lists for each cell. The presented method was based on simulations in which the potential new neighborhood relations were tested, and if the new relation proved out to be useful in the long run, it was finally included in the updated cell list.

Barco et al. (2001) have studied the optimization of frequency plans based on interference matrices. The interference matrices are derived from the measurement reports sent by the mobiles. The presented technique is tested under GSM/GPRS simulator, but it is mentioned, that field trials have provided good results also under live network environments.

For the performance optimization of UMTS networks, very similar approaches have been presented. Especially, the use of heuristic rules for deriving the improved configuration has been proposed frequently. Nearly all of the optimization techniques are developed under simulator based experiments. For example, Valkealahti et al. (2002b) suggest a rule-based control strategy in order to optimize common pilot power settings in an UMTS network. The work by Love et al. (1999) also proposes a rule-based approach for the optimization of pilot powers in a CDMA cellular system. In (Höglund and Valkealahti, 2002), a similar method has been presented for the optimization of downlink load level target and downlink power maxima. In (Höglund et al., 2003), the uplink load level target has been optimized with similar, rule-based approach. In (Valkealahti and Höglund, 2003), several parameters are optimized simultaneously with similar approach.

Another strategy in the autotuning approach is based on minimization of formally

defined cost functions rather than the use of heuristic control rules. The optimization of common pilot power by minimization of the formal cost function with a gradient method has been proposed in (Valkealahti et al., 2002a). In (Flanagan and Novosad, 2002a) and (Flanagan and Novosad, 2002b), a technique for finding soft handover parameters that provide minimal blocking in the network have been presented. Flanagan and Novosad (2003) suggest a cost function based approach for optimization of multiple parameters simultaneously, including soft handover parameters, uplink and downlink power maxima, as well as uplink and downlink load targets. Hämäläinen et al. (2002) have presented a cost function based autotuning method for the determination of planned, service specific E_b/N_o targets. Zhu and Buot (2004) discuss the dependencies between different KPIs and their sensitivities with respect to the optimized parameters. A sensitivity matrix is computed and the autotuning approach is based on the computed sensitivity matrix. Even tough the above mentioned approaches for UMTS network optimization are all tested with radio network simulators, they can be used also to optimize operating WCDMA networks without any major modifications to the control action decision making.

The adaptive autotuning approach has been applied in optimization of live GSM network by Magnusson and Oom (2002). The signal strength thresholds used in cell selection strongly affects the size of the cells, and therefore, they are tuned in order to obtain an optimal traffic balancing between cell layers. Simple intuitive rules are used to decide how the current configuration should be adjusted based on the performance measurements.

Toril et al. (2003) have proposed an algorithm for automatic offline optimization of handover margins in a live GSM network. The presented method is based on updating the current handover margin with a simple heuristic update rule depending on current amount of traffic and blocking.

The main characteristic of the above mentioned methods is that the mapping from the alternative configuration settings to the performance of the different configurations is determined by testing each of the configurations for certain amount of time in the live network or simulator. Human-defined heuristics or gradients of the objective function are used to select the direction and magnitude of the search in an intelligent manner in order to obtain faster convergence. Still, testing a large number of feasible configurations is a very time-consuming task and therefore, optimization of large number of network elements and parameters simultaneously may not be practical. For this purpose, the base stations could share the same optimized value of the parameters or they could be assigned into groups of similar BSs, and the BSs in the same group could use the same parameter values, thus decreasing the dimension of the parameter space. In Publications 1-3, possible methods to obtain this cell grouping have been presented.

2.4.3 Measurement-Based Approach

The third approach for the non-real time performance optimization is based on the use of network measurements directly in decision making. That is, the available data can be directly used to construct the configuration to performance mapping

without the need of advanced data-driven inference. In (Toril et al., 2002), level-quality data generated by a GSM/GPRS network simulator was gathered, and a mapping between the received signal strength and the network performance was obtained. Then, the selection of the updated signal level threshold was based on operator requirements for the signal quality with certain confidence level.

In (Chandra et al., 1997), handover related parameters were selected based on similar data records. During the operation, the network produced a data set that allowed the construction of mapping between the handover parameters and the amount of traffic carried by the cell. Nonlinear optimization was used to find the optimal parameter value given the previously mentioned mapping.

In both of these studies, the used measurement data allowed the determination of the mapping between the alternative configuration parameters and the system performance explicitly. The drawback of this approach is that reliable measurement data allowing the determination of the mapping between the configuration and the system performance is not available for most of the essential network parameters.

2.4.4 Predictive Approach

Once the bottlenecks of the performance are found, it is necessary to adjust the configuration in order to maximize the performance according to the operators needs. In the previous section, different approaches to decide the optimal control action were discussed. However, the decision about the new configuration may be very difficult to make without knowledge of how the network will behave in the new, unseen configuration. Predictive models can be applied in order to help and to automate the decision-making procedure.

In the study by Steuer and Jobmann (2002), traffic balancing through optimization of cell sizes is discussed. The cell sizes were modified by adjusting the signal strength thresholds, handover hysteresis settings and sector shapes of the smart antennas in order to avoid blocking. The presented approach is based on measurements including the locations of the mobiles. The method makes predictions about the performance of the system with the new, unseen configurations that are used to select the optimum setting, therefore being based on predictive modeling. The available data including the mobile locations were the driving force for deciding the optimal traffic balancing. The benefits of this study were demonstrated with GSM simulations.

The use of predictive modeling approach requires the use of common application domain models in order to make the necessary predictions about the performance of the adjusted configuration. However, the theoretical predictions and observations made from a real network are not always directly similar or comparable. In Publication 6, the significant differences between theoretical predictions and true measurements are highlighted. Also, a method to combine the use of common theories and measurement data in order to provide more accurate predictions about blocking in GSM networks is presented.

In Publication 7, a predictive modeling based approach is proposed for the optimization of signal strength thresholds in operative GSM networks. The model

includes a data-driven component exploiting the past measurements for the determination of the mapping between current configuration parameters and network performance and a knowledge-based component based on common theoretical models allowing the prediction of network performance under unseen configurations. These results are strongly based on the results provided in Publication 6.

Chapter 3

Data Analysis Methods

In this chapter, the data analysis methods used in this thesis are described. Firstly, the process monitoring tasks in data-rich production or manufacturing processes are discussed. Then, a wide range of methods and their usage to solve process monitoring tasks are described.

3.1 Tasks of Process Monitoring

In the process and manufacturing industries, there is a strong tendency to produce end-products of higher quality, to satisfy environmental and safety regulations and to reduce manufacturing costs. In mobile communication industry, there is a push to provide mobile communication services meeting the quality of service agreements for constantly increasing number of subscribers. However, the improvement of the product manufacturing or service provision is complicated by faults occurring in the processes. According to Chiang et al. (2001), "a fault is defined as an unpermitted deviation of at least one characteristic property or variable of the system" and in order to satisfy the performance requirements, the faults need to be

- detected,
- identified,
- diagnosed, and
- removed.

These tasks can be tackled by process monitoring methods. Fault detection is defined as determination of whether a fault has occurred or not. Fault identification involves selection of variables most relevant for the diagnosis of the fault. In fault diagnosis, the actual cause of the fault, but also, the type, location, magnitude, and time of the fault are determined. Process recovery involves removing the effects of the fault.

3.2 Survey of Research Fields

In engineering literature, a wide range of computational, data-driven methods that can be efficiently used in different process monitoring tasks have been presented. It turns out that methods useful for various process monitoring tasks are developed under very different research fields.

In traditional *statistics*, data analysis focuses on careful experiment design, hypothesis definition, data gathering and hypothesis testing. The emphasis is on confirmatory data analysis, that is, hypotheses about phenomena are made, and statistical tests are used to reject or confirm the hypotheses. For more information about statistical hypothesis testing, see (Meyer, 1975) and (Milton and Arnold, 1995). The tests usually involve estimation of models, for example, linear regression models and testing the significance of the dependency. Also, a wide range of mathematical tools like sample mean, variance and median are available to summarize the information of experimental data. In addition, horizontal bar charts, pie charts, line charts and scatter plots are often used to depict information about one variable, to hunt for correlations between variables and to graph multivariate data. The traditional statistical techniques are frequently used in quality control of industrial systems. For more information about traditional quality control, see (Mitra, 1998) and (Chiang et al., 2001).

Due to the rapid development of computer aided systems, the amount of available data has truly exploded and traditional hypothesis testing is no longer an efficient approach for many cases. One of the most recent and rapidly growing research fields related to inference in data-rich environments is *data mining*. According to Hand et al. (2001), "data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful for the data owner". In other words, data mining focuses on methods that can be used to rapidly increase the amount of knowledge of a system from which data is available. The emphasis is on hypothesis generation rather than testing well defined hypothesis. The main tasks of data mining include exploratory data analysis, descriptive modeling, predictive modeling, pattern and rule discovery and retrieval by content. For the analysis of unknown systems, the explorative data analysis task is the most important one in order to find out the basic structure of the data, see Hoaglin et al. (2000). Another useful set of methods developed under data mining discipline focus on descriptive modeling, in which structure in (multivariate) data is typically searched. Descriptive modeling consists of clustering and segmentation methods that apply well for fault detection problems in many industrial applications. The data mining methods for predictive modeling typically consists of classification and regression techniques. They are most useful in fault detection, identification and diagnosis of faults.

The science of graphical representation of data sets is also studied by an own research field, *data visualization*, that is strongly rooted in the exploratory data analysis. However, it has similar aims as statistics and basic scientific visualizations. According to Spence (2007), visualization means "forming a mental model or mental image of something". Another frequently quoted justification for data visualization states that "solving a problem simply means representing it so as

to make the solution transparent". Advanced visualization techniques have been developed due to the rapid increase in amount of data to be represented. Modern visualization techniques often rely on projection methods, in which the data is first projected into lower dimension, and the projected data is visualized with basic graphs. Therefore, the method for summarizing or data reduction differ from more traditional statistics. Visualization can be effectively used in different process monitoring tasks.

In *artificial intelligence*, the aim is to create machines to automate tasks requiring intelligent behavior. *Machine learning* is a subfield of artificial intelligence that concerns algorithms and methods that allow computers to learn from examples. For this reason, machine learning techniques are frequently applied in tasks in which process data are used to learn useful relationships in the process. The learning problems are typically very closely related to clustering, regression and classification methods (Cherkassky and Mulier, 1998). However, the focus of machine learning is on the computational properties of the methods such as computational complexity.

A sub-field of machine learning more focused on practical applications of learning methods is *pattern recognition*. It includes a wide range of information processing problems of great practical importance. For example, speech recognition, classification of handwritten characters, fault detection in machinery and medical diagnosis are important topics in pattern recognition (Bishop, 1995). Most of the recognition problems take the form of clustering, regression or classification, preceeded by careful data preprocessing and feature extraction.

System control is a field of engineering in which the aim is the control the operation of a system so that it would function as intended, for example, in a production process (see Åström and Wittenmark (1997)). The focus is not in management of unsuspected faults, but instead, the maintenance and optimization of the normal operational modes. An important part of system control is the system identification step, in which a mathematical model for system behavior is estimated from measurement data. System identification techniques are outlined in (Ljung and Glad, 1994). The system identification consists of similar methods and procedures as the predictive modeling techniques studied in the data mining, pattern recognition and machine learning communities. However, the identification of systems aiming in system control typically involves estimation of dynamical models from the data.

Next, the basics of above approaches and how they can be used in different process monitoring procedures are discussed. Then, different techniques such as neural networks, fuzzy systems etc. are described and how they can be used in above mentioned process monitoring problems.

3.2.1 Exploring and Visualizing Data

In (Hoaglin et al., 2000), a wide range of tools for explorative data analysis are described. The most simple examples of such tools include stem-and-leaf plots, letter-value displays and N-number summaries. In these techniques, the experi-

mental data or a batch is presented by a set of numbers describing the location and spread of the observations.

Hand et al. (2001), Fayyad et al. (2002) and Spence (2007) give a good summary of basic tools to visualize univariate and bivariate data. Univariate data is often displayed as histogram or box plots, and bivariate data is typically displayed as scatterplots. If the second variable is time, a time-series plot is often used.

The visualization of multivariate data can be done in at least two ways. The multivariate data can be visualized with, for example, Chernoff's faces (Chernoff, 1973) or parallel coordinate techniques (Inselberg and Dimsdale, 1990). Also, projecting the data into two dimensions and the use of basic bivariate data plots is of common practice. The projection is typically based on Principal Component Analysis (PCA) (Hotelling, 1933), Multidimensional Scaling (MDS) (Torgerson, 1952; Young, 1985) or Self-Organizing Maps (SOM) (Kohonen, 2001).

By using exploratory data analysis and data visualization techniques, a human analyst can efficiently study the information content of the data, and then, important conclusions and problem refinement emerge. For example, the analyst may be able to find out what are the most typical failure types of the processes, and it helps the researchers to focus on certain subproblems more closely related to the fault of interest. These methods are useful especially in fault detection applications in which all the possible fault types are not known in advance, but instead, new types of faults may occur. In this kind of applications, previously defined fault types and detection based on them do not necessarily provide adequate solutions. This is the case in mobile communication networks in particular, since new hardware and radio resource algorithms may be installed and integrated to the existing system, and the compatibility between different algorithms and equipment and possible side effects are not necessarily known.

3.2.2 Clustering and Segmentation

Another useful data analysis problem type is clustering. In clustering, the data is separated into groups or clusters so that the similarity between samples in the same cluster is maximized and the similarity between samples in different clusters is minimized.

According to Hand et al. (2001), clustering algorithms can be divided into three classes: those based on finding the optimal partitioning of the data into a specified number of clusters, those aiming in finding the hierarchical structure of the data, and those based on probabilistic models searching for the underlying cluster structure. The clustering algorithms aiming in dividing the data into specified number of clusters are referred as partitive algorithms. The hierarchical clustering algorithms that search for the structure of the data can be divided into top-down (divisive) and bottom-up (agglomerative) algorithms. For more information about algorithms, see Everitt (1993).

As the definition of clustering implies, the characterization of distances between samples and clusters is of great importance in clustering of data. For example, the selection of the measure for the within-cluster and between-clusters distances

greatly influences the solution returned by the algorithm. The within-cluster and between-clusters distances can be measured in several ways. For example, the within-cluster distance can be evaluated with average distance between each sample pair in the same cluster, or with the average distance of each sample from the centroid of the cluster. The between-clusters distances can be based on for example, the distances between the cluster centroids.

One of the main problems in data clustering is the availability of wide range of algorithms that tend to search for very different types of clusters from the data. Even the use of same clustering method with different number of target clusters raises the question of which clustering serves the problem solving process adequately. In the literature, several clustering validity indices have been proposed in order to select the optimal clustering, for more information, see Bezdek and Pal (1998).

A data analysis problem closely related to clustering is the segmentation of time-series data. In time-series segmentation, a sequence of N consecutive (multivariate) data samples is partitioned into groups so that each segment is as homogenous as possible (Bellman, 1961; Terzi, 2006). The homogeneity can be defined in several ways. Typically, each segment is represented by a model, for example, a distribution or a time-series model. Similarly to the clustering algorithms, top-down and bottom-up approaches for time-series segmentation have been proposed. A good review of segmentation approaches have been presented in (Terzi, 2006).

Clustering and segmentation techniques are useful in fault identification and diagnosis. For example, clustering can be used to divide the process data into a discrete set of states, and some of the states may represent undesired process conditions. By studying the properties of such clusters, the causes of the faults may be analyzed. In addition, the properties of the found fault clusters can be stored and used later in fault detection. The segmentation algorithms can be used in similar fashion. Especially, when the changes of the states occur more slowly than the sampling rate of the measurements, it is more natural to apply segmentation algorithms than clustering of rapidly changing data samples. In addition, segmentation can be used for separating between different operational modes of the process.

3.2.3 Classification and Regression

Classification and regression belong to the so-called supervised learning tasks. In classification, each sample $\mathbf{x}$ is assigned to one of the several classes C. The goal is to find a classification function $C = f(\mathbf{x})$ that is able to predict the classes of unseen samples $\mathbf{x}$ with minimal classification error. The classification function is estimated from labeled data, that is, using a data set consisting of N examples of $(\mathbf{x}_i, C_i)$ pairs, $i = 1, 2, \ldots, N$. In regression, the model output is continuous and may consist of several variables. In other words, the learning task consist of estimating a regression function $\mathbf{y} = \mathbf{f}(\mathbf{x})$ using examples of $\mathbf{x}_i$ and $\mathbf{y}_i$ in order to predict new samples with minimal prediction error.

For classification and regression, a wide range of methods is available. In (Milton

and Arnold, 1995) and (Jørgensen, 1993), the basic linear regression for single and multivariate cases are described. The principles of linear regression models can also be extended to fit to many nonlinear modeling problems. In (McCullagh and Nelder, 1983), a wide range of such extensions are described. One of the most interesting approach is based on estimating black-box models using neural networks. The neural network based classification and regression methods are discussed in detail, for example, in (Haykin, 1999), (Cherkassky and Mulier, 1998) and (Bishop, 1995). The estimation of dynamical regression models from process data is explained in (Ljung and Glad, 1994). Practical advises for building models in the presence of missing data, redundancy in data and colinearity between variables are given in (Hyötyniemi, 2001).

Classification and regression techniques can be used efficiently in many tasks related to process monitoring. For example, fault detection can be based on classifying measurement data into previously specified classes of normal and abnormal behavior. Fault identification and diagnosis can be based on regression models estimated between variables, and statistical tests can be used to find out the possible causes of the faults. Also, regression techniques can be used to build a model for the process under different operating regimes. The regression models can later be used to predict system behavior after certain adjustments to the process.

3.2.4 Control and Optimization

In system control, the behavior of the system under different situations is very well known and the system behavior can be affected through control signals. Especially, the system model is able to predict the outcome of the process to different control signals. Then, the problem is to decide how the system is supposed to function and how to select the control signals continuously so that the system operates as desired. In other words, control and optimization are most frequently used in fault recovery procedures.

Traditional approaches for system control are based on single variable control loops. The systems are described by linear, time-invariant differential or difference equations that are usually solved in frequency domain using Laplace, Fourier or Z-transforms (Lewis and Chang, 1997; Åström and Wittenmark, 1997).

Modern control theory is based on state-space representations of systems and they are solved in time-domain (Hakkala and Ylinen, 1978). These methods are available also for nonlinear and time-variant systems. Modern system theory exploits mainly the matrix algebra techniques.

The theory of optimal control is discussed in (Kirk, 1970) and (Åström and Wittenmark, 1997). The optimal control problem has a clearly defined cost function that is used to select the optimal controller from the alternative ones. In addition, some physical constraints can be included to the controller design. Examples of cost functions are the minimum resource and the minimum time problems.

The above approaches for system control are based on differential equation representation of the system and its use to design the controller. The neural networks can be used to model system behavior, but also to learn suitable controllers from

experimental data without explicit knowledge about the system equations. The use of neural networks in system control is discussed in (Nørgaard et al., 2000). The use of other multivariate regression techniques in system control are also discussed in (Hyötyniemi et al., 1997).

In some cases, the system can be influenced with static (constant) control signals rather than time-variable control signals. In these cases, various optimization approaches can be used to design adjustments to the system or process behavior. In (Bazaraa et al., 1993), several nonlinear optimization algorithms useful for designing optimal process adjustments are described.

3.3 Traditional Methods for Regression

As stated in the previous section, regression techniques can be used in different parts of process monitoring. Next, the basics of linear regression are described.

3.3.1 Linear Regression

Let us assume that we have observations of single output variable $y(k)$ and N predictor variables $x_i(k), i = 1, 2, ..., N$ from time instants $k = 1, 2, ..., K$. The parameters $\theta_i, i = 0, 1, 2, ..., N$ of the linear regression model

$$y(k) = \theta_0 + \theta_1 x_1(k) + \theta_2 x_2(k) + \ldots + \theta_n x_N(k) + e(k) \tag{3.1}$$

can be estimated by minimizing the error $e(k) = y(k) - \hat{y}(k)$ between the true observations $y(k)$ and predictions $\hat{y}(k) = \theta_0 + \sum_i^N \theta_i x_i(k)$. Also, let us denote $\mathbf{Y} = [y(1)\ y(2)\ \ldots\ y(K)]^T$, $\mathbf{E} = [e(1)\ e(2)\ \ldots\ e(K)]^T$ and

$$\mathbf{X} = \begin{bmatrix} 1 & x_1(1) & x_2(1) & \ldots & x_N(1) \\ 1 & x_1(2) & x_2(2) & \ldots & x_N(2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_1(K) & x_2(K) & \ldots & x_N(K) \end{bmatrix}. \tag{3.2}$$

Now, the multiple linear regression model can be written in form $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{E}$. An unbiased estimate $\hat{\boldsymbol{\theta}}$ for the parameter vector $\boldsymbol{\theta} = [\theta_1\ \theta_2\ \ldots\ \theta_N]^T$ can be obtained by minimizing the sum of squared error $\mathbf{E}^T\mathbf{E}$ between the true observation of the output variable and the prediction given by the model, leading to

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{3.3}$$

The ability of the model to explain the output data given the input data can be analyzed in several ways. One of the most frequently adopted practices is to compute the so-called coefficient of determination

$$R^2 = 1 - \frac{\sum_{k=1}^K \left[y(k) - \hat{y}(k)\right]^2}{\sum_{k=1}^K \left[y(k) - \bar{y}\right]^2} \tag{3.4}$$

where $\bar{y}$ denotes the mean value of the observations $y(k), k = 1, 2, \ldots, K$. If R^2 takes values close to 1, the data contains evidence about a strong linear trend between the input and the output variables and R^2 values close to zero indicate that no linear trend between input and the output variables is present in the observed data set.

3.3.2 Linear and Quadratic Programming

In many cases, the estimation of models of form $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta}+\mathbf{E}$ is based on minimizing the sum of squared error $\mathbf{E}^T\mathbf{E}$, but there are additional constraints that restrict the set of possible parameter vector $\boldsymbol{\theta}$ values. The error measure $\frac{1}{2}\mathbf{E}^T\mathbf{E} = \frac{1}{2}\mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\boldsymbol{\theta} + \frac{1}{2}\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta}$ can be rewritten in form $\frac{1}{2}\boldsymbol{\theta}^T\mathbf{H}\boldsymbol{\theta} + \mathbf{f}^T\boldsymbol{\theta}$ by omitting the constant term not depending on $\boldsymbol{\theta}$ and setting $\mathbf{H} = \mathbf{X}^T\mathbf{X}$ and $\mathbf{f} = -\mathbf{X}^T\mathbf{Y}$. A typical case where the sum of squared errors must be minimized and the values of parameter vector $\boldsymbol{\theta}$ are restricted is the so-called quadratic programming problem

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & \left\{ \frac{1}{2}\boldsymbol{\theta}^T\mathbf{H}\boldsymbol{\theta} + \mathbf{f}^T\boldsymbol{\theta} \right\} \\ w.r.t \quad & \mathbf{A}\boldsymbol{\theta} \leq \mathbf{b}, \\ & \mathbf{A}_{eq}\boldsymbol{\theta} = \mathbf{b}_{eq}, \\ & \boldsymbol{\theta}_{min} \leq \boldsymbol{\theta} \leq \boldsymbol{\theta}_{max}, \end{aligned} \tag{3.5}$$

where the matrix $\mathbf{A}$ and vector $\mathbf{b}$ define the inequality constraints of the problem, matrix $\mathbf{A}_{eq}$ and vector $\mathbf{b}_{eq}$ define the equality constraints and vectors $\boldsymbol{\theta}_{min}$ and $\boldsymbol{\theta}_{max}$ define the box constraints for the parameter vector $\boldsymbol{\theta}$.

Since the rows of $\mathbf{Y}$ may consist of samples with varying importance or variables measured in very different scales, it may be necessary to add a weighting scheme to the above model. The use of weighted error measure

$$\mathbf{E}_w = \begin{bmatrix} w_1 & 0 & \ldots & 0 \\ 0 & w_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & w_n \end{bmatrix} \left[\mathbf{Y} - \hat{\mathbf{Y}} \right] = \mathbf{W}\mathbf{E} \tag{3.6}$$

leads to the quadratic programming problem with $\mathbf{H} = \mathbf{X}^T\mathbf{W}^T\mathbf{W}\mathbf{X}$ and $\mathbf{f} = -\mathbf{X}^T\mathbf{W}^T\mathbf{W}\mathbf{Y}$.

Another example of constrained optimization is the linear programming problem taking the form

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & \left\{ \mathbf{f}^T\boldsymbol{\theta} \right\} \\ w.r.t \quad & \mathbf{A}\boldsymbol{\theta} \leq \mathbf{b}, \\ & \mathbf{A}_{eq}\boldsymbol{\theta} = \mathbf{b}_{eq}, \\ & \boldsymbol{\theta}_{min} \leq \boldsymbol{\theta} \leq \boldsymbol{\theta}_{max}. \end{aligned} \tag{3.7}$$

In other words, we are looking for the parameter vector $\boldsymbol{\theta}$ for which the value of $\mathbf{f}^T\boldsymbol{\theta}$ is as small as possible, but it is also required that the solution vector $\boldsymbol{\theta}$ must

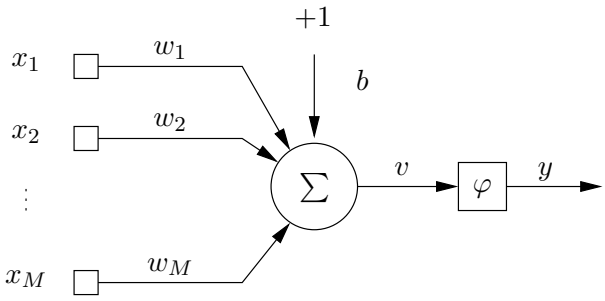


Figure 3.1: The basic elements of a neuron.

satisfy the inequality constraint $\mathbf{A}\boldsymbol{\theta} \leq \mathbf{b}$, the equality constraint $\mathbf{A}_{eq}\boldsymbol{\theta} = \mathbf{b}_{eq}$ and the box constraint $\boldsymbol{\theta}_{min} \leq \boldsymbol{\theta} \leq \boldsymbol{\theta}_{max}$. There are several methods that can be used to find the solution vector $\boldsymbol{\theta}$ for the problems stated above. For more information, see Bazaraa et al. (1993).

3.4 Neural Networks

Artificial neural networks are efficient computational systems in which the learning is inspired by neuro-biological systems. Due to the good learning properties, neural networks have found many applications in industry. Especially, neural networks are highly useful in many problems faced during process monitoring. These include visual data mining for systems not very well known, but also, they provide simple methods for nonlinear regression and classification as well as control of nonlinear systems. Next, the neural network structures used in this thesis are shortly reviewed.

3.4.1 Neuron Models

Neural networks aim to solve engineering problems by following the signal processing principles of neuro-biological systems. In general, neural networks are constructions of simple units, called neurons, and they are adaptively combined in order to solve various practical problems. In the literature, a variety of different models for a neuron have been proposed (Haykin, 1999), (Demuth and Beale, 1998). In Figure 3.1, a typical neuron model is depicted. The basic elements of a neuron are a set of *synapses*, a *summing junction* and an *activation function*. The synapses receive the input signal x_i for the neuron. In each synapse, a weight w_i is associated describing a strength of the corresponding synaptic connection. A summing junction or a linear combiner adds the weighted input signals and a *bias* b to form an *activation potential* v

$$v = \sum_{i=1}^{M} w_i x_i + b. \tag{3.8}$$

Finally, the activation potential is fed through an activation function φ limiting the output y of the neuron into an appropriate range

$$y = \varphi(v). \tag{3.9}$$

The most common choices for activation functions for regression problems are the logistic sigmoidal function

$$\varphi(v) = \frac{1}{1 + \exp(-av)} \tag{3.10}$$

and the hyperbolic tangent function

$$\varphi(v) = \tanh(v). \tag{3.11}$$

In a *linear* model of a neuron, the output signal y is exactly the same as the activation potential

$$y = v = \sum_{i=1}^{M} w_i x_i. \tag{3.12}$$

A neuron model in which a short-term memory is used to save the previous input samples in order to represent dynamical dependencies between the neuron input and output is called a *spatio-temporal* model of neuron.

3.4.2 Adaptive Filters

Most of the neural network algorithms focus on nonlinear regression tasks. By combining a high number of simple computational units (neurons), complex nonlinear dependencies can be modelled. In other words, the goal is to find a model that is able to approximate the behavior of a system given the input data $x(n)$ and the output of the system $y(n)$ from the time instants $1 \leq n \leq N$. When the dependencies between input and output variables of the unknown dynamical multi-input single-output system are assumed to be linear, the model reduces to the linear spatio-temporal model of a neuron, or to the so-called adaptive filter. The mapping carried out by the adaptive filter is of the form

$$y(n) = \sum_{i=1}^{M} w_i x_i(n), \tag{3.13}$$

where $x_i(n)$ denotes the ith input signal and w_i is the corresponding weight of the neuron. Note, that the multivariate input signal $x(n)$ may consist of measurements from many different variables but also to contain past measurements of the same variables. The adaptive filter structure is depicted in Figure 3.2.

The learning problem, that is, the determination of values for the weights of the adaptive filters can be solved by minimizing the sum of squared prediction errors of the model over the whole observation session $1 \leq n \leq N$, thus obtaining the so-called linear (multi-variable) least-squares filter (Haykin, 1999). However, it also possible to obtain an adaptive method for the solution of the same problem by

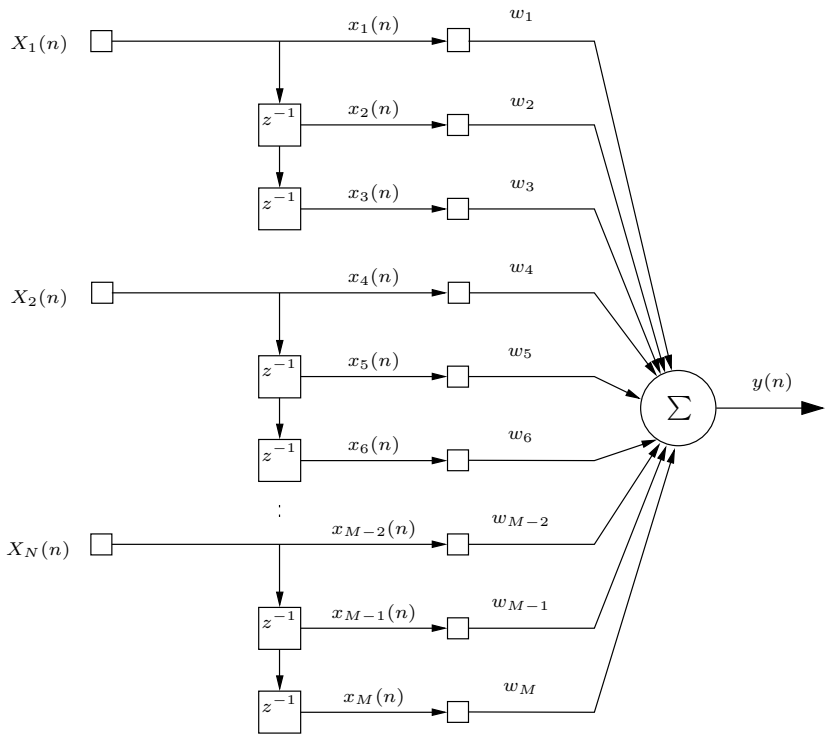


Figure 3.2: Adaptive filter model used in linear system modeling. For each input variable, two delay units are used to save the past observations.

minimizing the instantaneous squared prediction errors. In the literature, such a method to produce an instantaneous estimate for the model parameters of a linear model is called as the LMS (Least Meas Square) algorithm (Haykin, 1999). In the LMS algorithm, the parameters of the network, that is, the weights of the neuron are updated iteratively using

$$w_i(n+1) = w_i(n) + \eta x_i(n) e(n), \tag{3.14}$$

where $e(n) = y(n) - \hat{y}(n)$ is the error signal obtained using the output $\hat{y}(n)$ of the model and the true output $y(n)$ and $\eta < 1$ is a learning factor.

3.4.3 Multilayer Perceptrons

One of the most popular neural network architectures suitable for solving numerous nonlinear function approximation problems is the multilayer perceptron (MLP) (Haykin, 1999). It consists of multiple layers of nonlinear neurons. The

output of the neuron i in layer l of an MLP network is defined as

$$y_i^{(l)}(n) = \varphi\left(\sum_{j=0}^{M_{l-1}} w_{ij}^{(l)}(n) y_j^{(l-1)}(n)\right) \tag{3.15}$$

$$= \varphi\left(v_i^{(l)}(n)\right) \tag{3.16}$$

where $w_{ij}^{(l)}$ and $v_i^{(l)}(n)$ are the jth synaptic weight and the induced local field of the neuron i at layer l respectively and φ is the nonlinear activation function. It should be noted that the synaptic weight $w_{i0}^{(l)}$ of neuron i in layer l is also called the *bias*, that is, the synaptic weight corresponding to the additional fixed input signal $x_0(n) = 1, \forall n$.

The synaptic weights and the bias terms of the network are obtained by training the network using the back-propagation algorithm

$$w_{ij}^{(l)}(n+1) = w_{ij}^{(l)}(n) + \eta \delta_i^{(l)}(n) y_j^{(l-1)}(n) \tag{3.17}$$

where $\delta_i^{(l)}(n)$ is the local gradient

$$\delta_i^{(l)}(n) = \begin{cases} \varphi_i'(v_i(n)) e_i(n) & \text{, neuron } i \text{ in output layer} \\ \varphi_i'(v_i(n)) \sum_k \delta_k(n) w_{ki}(n) & \text{, neuron } i \text{ in hidden layer} \end{cases} \tag{3.18}$$

and the summation over k is performed over neurons in layer $l+1$ in which the neuron i at layer l is connected and $\eta < 1$ is the learning factor.

In order to process temporal patterns with MLP, a short-term memory can be included in the static MLP model. In Figure 3.3, a single-output MLP with 5 neurons in the first layer and a single neuron in the output layer of the network is depicted.

The most interesting property of an MLP is that it is able to approximate any nonlinear function with arbitrary accuracy given enough hidden layer neurons, making it very useful for the nonlinear system identification.

3.4.4 Self-Organizing Map

One of the most famous neural network algorithms is the Self-Organizing Map (SOM) (Kohonen, 2001). It consists of neurons or map units, each having a location in both a continuous multi-dimensional measurement space and in a discrete low-dimensional grid. Within this architecture, the SOM is able to produce a mapping from high-dimensional measurement space into low-dimensional grid of map units making it especially suitable for data visualization. In Figure 3.4, a SOM with 16 neurons arranged in a $[4 \times 4]$ regular grid for the processing of 3-dimensional input vectors $x = [x_1\ x_2\ x_3]^T$ is depicted.

During the training phase of the SOM, the map unit locations in the measurement space are repeatedly adjusted, while the locations of the map units in the discrete grid remain unchanged. The training of SOM consist of two phases: the winner

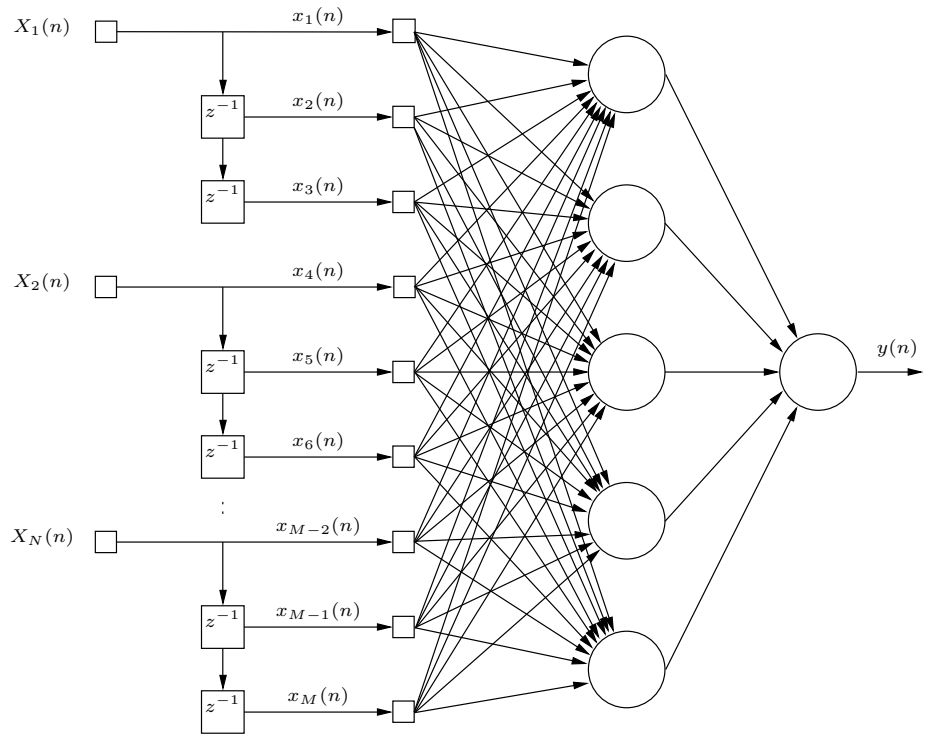


Figure 3.3: A single-output MLP with 5 neurons in the hidden layer. A tapped delay line with two delay units for each input variable $X_i(n)$ is used to model the short-term dependencies between the consecutive data samples.

map unit search and application of an update rule used to adjust the locations of the map units in the measurement space. The winner search is defined by:

$$c = \arg\min_i ||x - m^{(i)}||, \tag{3.19}$$

in which $m^{(i)}$ is the location of the ith map unit in the measurement space and c is the index of the winner map unit in the SOM grid. In winner search, an input sample x is picked randomly from the multidimensional measurement space and the map unit c closest to the input sample is declared as the winner map unit or the best-matching map unit (BMU). Then, the locations of all the map units in the measurement space are updated according to the rule:

$$m^{(i)}(t+1) = m^{(i)}(t) + \alpha(t)h_{ci}(t)[x(t) - m^{(i)}(t)], \tag{3.20}$$

in which the $h_{ci}(t)$ is, for example, the Gaussian neighborhood function

$$h_{ci}(t) = \exp\left(-\frac{||r_c - r_i||}{2\sigma^2(t)}\right) \tag{3.21}$$

and $0 < \alpha(t) < 1$ is the learning rate factor. In $h_{ci}(t)$, r_c is the location of the winner unit and r_i is location of the ith map unit in the discrete output lattice of

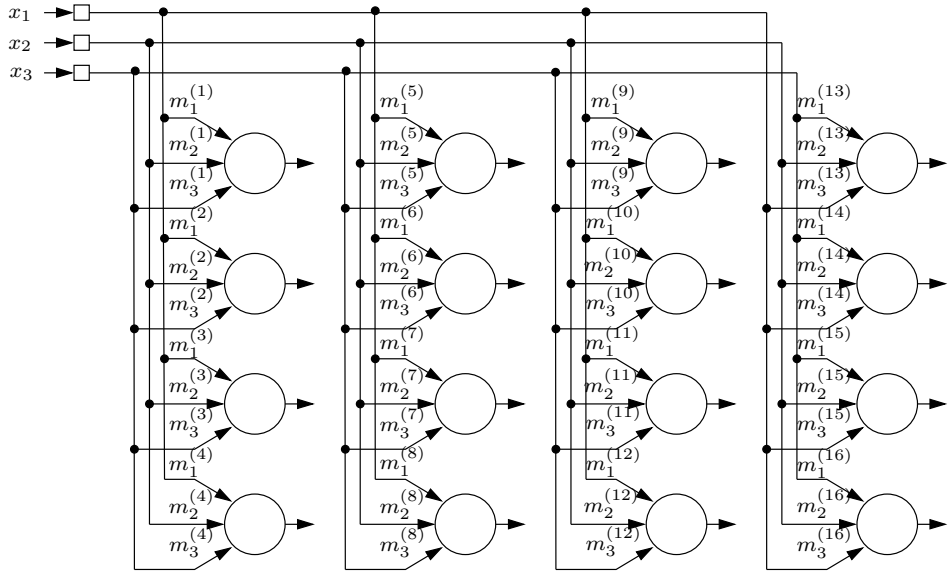


Figure 3.4: A $[4 \times 4]$ SOM in a 3-dimensional measurement space.

SOM. The learning rate factor $\alpha(t)$ and $\sigma(t)$ are monotonically decreasing functions of time t.

The SOM is an efficient visualization tool and it has been widely used in exploratory data analysis tasks involving multivariate data. The SOM can be interpreted to perform a nonlinear dimensionality reduction from the multivariate measurement space to the low-dimensional grid. For this reason, the SOM can be used in visualization of multivariate data.

3.5 Fuzzy Systems

In this section, the basics of fuzzy logic are presented. Fuzzy systems can be used in knowledge-based process monitoring in systems for which mathematical models are not available, but qualitative information is present. Also, fuzzy systems can be used to estimate nonlinear regression models whose behavior can be analyzed in terms of qualitative rules.

First, the concept of fuzzy sets is shortly introduced and the logical operations to form statements with the fuzzy sets are described. Then, the concept of fuzzy if-then rules for making logical reasoning using the fuzzy logic is presented.

3.5.1 Fuzzy Sets, Logical Operations and Inference

In traditional set theory, an object can either belong or not to belong to a set. In the fuzzy set theory, an object may also have partial memberships (Zadeh, 1965). The membership of an object x to a fuzzy set A is defined by the membership function $\mu_A(x)$ of a fuzzy set A. It maps all the values of x between 0 and 1 describing how well x fulfills the property specified by the set A.

As in the case of the Boolean logic, a set of logical operations can be used to form statements of multiple fuzzy sets. A fuzzy intersection (or T-norm) operator corresponds to the AND operation in Boolean logic. A simple example of a T-norm filling the so-called *boundary, monotonicity, commutativity* and *associativity* requirements is the multiplication of $\mu_A(x)$ and $\mu_B(x)$.

A fuzzy union (or S-norm) operator corresponds to the Boolean OR operation, and an example of a S-norm operator filling the same requirements as the multiplication T-norm is the *probabilistic or* is $S_{probor}[\mu_A(x), \mu_B(x)] = \mu_A(x) + \mu_B(x) - \mu_A(x)\mu_B(x)$.

The fuzzy sets can be used to construct fuzzy rules describing the dependencies between quantities. A fuzzy rule or if-then rule is a statement of the form: if x is A then y is B, in which A and B are linguistic labels of fuzzy sets defined by the corresponding membership functions. The if-part of the rule is called the antecedent (premise) and the then-part is called the consequence of the fuzzy rule. The interpretation of a fuzzy rule consists of two parts: evaluation of antecedent part value by applying the desired fuzzy operators for the membership function values of the given input x and the application of an implication method. For example, the evaluation of the antecedent part of a fuzzy rule "if x is *low* AND y is *high*, then z is *medium*" consists of application of the T-norm operator for $\mu_{low}(x)$ and $\mu_{high}(y)$ in order to obtain $\mu_{low \cap high}(x, y)$.

A *fuzzy implication* is a mapping that associates the values of the antecedent part to the value of the consequence part of the fuzzy rules. That is, the greater is the value of the membership function of the antecedent part, the closer the output of the implication is to the consequent part of the fuzzy rule. In *fuzzy aggregation* the result of implication of several fuzzy rules are combined into a single fuzzy output. Thus, it is needed when fuzzy logic is applied in reasoning with more than one fuzzy rule.

3.5.2 Fuzzy Inference Systems

Fuzzy inference systems are parametric models in which the output of the system as a response to system input is computed using fuzzy logic. The functional blocks of a typical fuzzy inference system are:

- *knowledge base* consisting of the fuzzy rules and the membership function parameters in the model;
- *decision making unit* performing the inference operations to the system input using the fuzzy rules in the knowledge base;

- *fuzzification interface* converting the numerical input signal into different degrees of matching of linguistic labels according to the membership functions stored in the knowledge base;
- *defuzzification interface* converts the fuzzy output of the decision making unit back to numerical in order to produce the system output.

In the literature, several types of fuzzy inference models have been proposed. A good review is presented in (Lee, 1990a) and (Lee, 1990b). The main differences in fuzzy inference systems are in the type of defuzzification method and the form of fuzzy rules used in the system. In this thesis, a fuzzy model called the *zeroth-order Sugeno fuzzy model* is used in system modeling. In this model, the fuzzy rules are of the form "if x is A and y is B, then z is r" where A and B are linguistic labels for two fuzzy sets defined by the corresponding membership functions $\mu_A(x)$ and $\mu_B(y)$ respectively and r is a crisply defined constant. The main differences in Sugeno fuzzy model with respect to other fuzzy models is that the membership functions of the consequent part of the fuzzy rules are singleton spikes and the implication and aggregation methods are fixed.

3.5.3 Adaptive Neuro-Fuzzy Inference System

As with the learning tasks associated with MLP networks, learning was used to find the model parameters in order to find good representation of the training data. Similarly in the case of fuzzy inference systems, the parameters of such a model, that is, the antecedent and consequent parts of the fuzzy rules may be estimated using a set of training data. This corresponds to learning rules from the data, instead of producing the output of the fuzzy inference system given the fuzzy rules and the input for the system.

Adaptive neuro-fuzzy inference system (ANFIS) is a multilayer network architecture similar to multilayer perceptrons (Jang, 1993). The only distinction is that the node functions perform fuzzy logic operations based on the Sugeno fuzzy inference model for the input data. ANFIS model structure consists of the following layers:

Fuzzification of system input: In this layer, the values of the input variables in the membership functions defined by the fuzzy rules are computed. Thus, the output of this layer is:

$$y_i^{(1)} = \mu_{A_i}(x_i), \tag{3.22}$$

where A_i is the linguistic label for ith fuzzy set. For each input variable component x_i, multiple fuzzy sets may be associated.

Application of the fuzzy AND-operation: In this layer, the strengths of the fuzzy rules are computed. The output of this layer is:

$$y_i^{(2)} = w_i = \prod_{j \in R_i} \mu_j(x), \tag{3.23}$$

where w_i denotes the strength of the ith rule and $\mu_j(x)$ denotes the fuzzy sets that belong to the premise part of the ith rule. The set R_i of rules is obtained

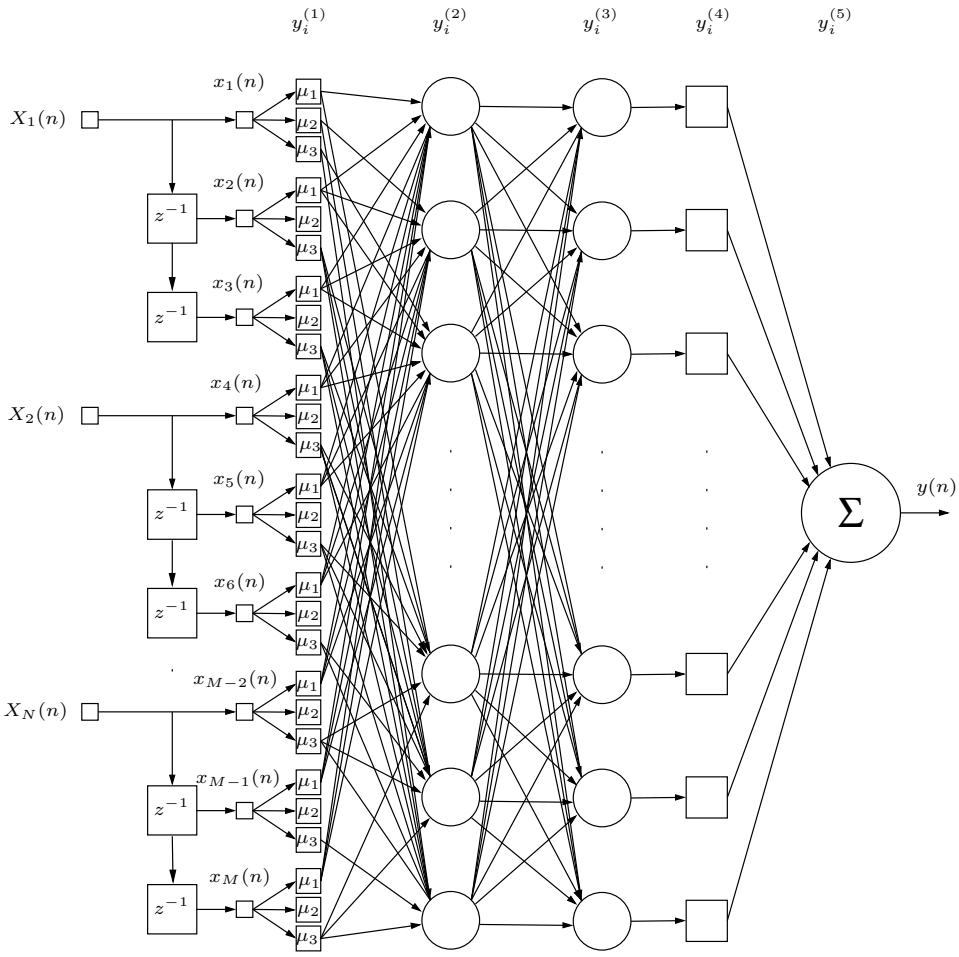


Figure 3.5: ANFIS model structure.

by forming all the possible fuzzy set combinations in which a single fuzzy set is included for each input sample component.

Normalization of the firing strengths of the rules: In this layer, the strengths of the rules of the inference system are normalized:

$$y_i^{(3)} = \overline{w}_i = \frac{w_i}{\sum_j w_j}. \tag{3.24}$$

Application of implication method: The application of the implication method corresponds to weighting the consequent part of the fuzzy rules according to their firing strengths. Thus, the fuzzy rule with high firing strength will have more contribution to the system output than a fuzzy rule with small firing strength. Thus, the output of this layer is:

$$y_i^{(4)} = \overline{w}_i r_i = \frac{w_i r_i}{\sum_j w_j}, \tag{3.25}$$

where the term $w_i r_i$ corresponds to the implication method.

Application of aggregation method: In this layer, the total output of the system $y(n) = y^{(5)}(n)$ is computed as the average of the fuzzy rule outputs weighted by the firing strengths of the rules. First, the consequent parts of the fuzzy rules are aggregated to form a single fuzzy set by summing the singleton spikes. Then, the result of aggregation is defuzzified to form the numerical system output by dividing the sum of weighted rule outputs by the sum of the firing strengths computed in the normalization layer:

$$y^{(5)} = \sum_i \overline{w}_i r_i = \sum_i \frac{w_i r_i}{\sum_j w_j} \tag{3.26}$$

Thus, the concurrent application of firing strength normalization, fuzzy implication and aggregation corresponds to the weighted average defuzzification method.

3.6 Clustering

Clustering of data means partitioning the data set into a set of clusters. Next, the k-means clustering algorithm and a validity index to choose the optimal clustering is discussed. In addition, methods to allow easy analysis of cluster properties are outlined.

3.6.1 k-means

One of the most widely used clustering method is the k-means algorithm (Everitt, 1993; Hand et al., 2001; Bezdek and Pal, 1998). It belongs to the partitive clustering algorithms, that is, it divides the data into specified number of clusters (k) by minimizing

$$E = \sum_{i=1}^{k} \sum_{x \in Q_i} ||x - c_i||^2, \tag{3.27}$$

where Q_i denotes the ith cluster and c_i denotes the center of the ith cluster. The definition of distances used in clustering makes implicit assumptions about the shape of the clusters. The k-means searches for spherical clusters.

The basic version of k-means operates as follows: first, k cluster centers are picked up in random fashion. Then, each sample in the data are assigned to the cluster whose center is closest to the sample (typically, in the Euclidean sense). Then, the mean vector of the samples in the same cluster is computed and defined as the new cluster centroid. These steps are iterated until memberships of the samples in the clusters remain unchanged.

3.6.2 Davies-Bouldin Index

When using partitive clustering methods such as the k-means, it becomes necessary to select the optimal number of clusters in which the data is divided. Several va-

lidity indices have been proposed. In this thesis, the Davies-Bouldin index (Davies and Bouldin, 1979) minimizing

$$\frac{1}{C}\sum_{k=1}^{C}\max_{l\neq k}\left\{\frac{S_c(Q_k)+S_c(Q_l)}{d_{ce}(Q_k,Q_l)}\right\} \tag{3.28}$$

is used. In above, S_c denotes within cluster distance and d_{ce} denotes between cluster distance. In the literature, other validity indices have been proposed. For more information, see Bezdek and Pal (1998).

3.6.3 Cluster Description

In data analysis associated with clustering, it becomes useful to know what properties are common in certain cluster and how the samples in a cluster differ from the samples in the other clusters. In this thesis, a cluster description algorithm presented in (Siponen et al., 2001) is used.

Each rule produced by the algorithm describe clusters in terms of value ranges for each measured variable. Single descriptive rules and their combinations for each cluster can be computed by maximizing the significance measure $S_r(i,r)$ of a rule r for cluster i:

$$S_r(i,r) = P(i|r)P(r|i) = \frac{n_{r\&i}^2}{n_r n_i}, \tag{3.29}$$

where n_r is the number of data samples in the data set for which the rule r is true, n_i is the number of data samples in cluster i, and $n_{r\&i}$ is the number of data samples in the cluster i for which the rule r is true.

3.6.4 Clustering of SOM

In (Vesanto and Alhoniemi, 2000), a two-phase method for clustering is proposed. Firstly, a SOM is trained for a given (multivariate) data set. Then, a clustering algorithm is applied for the reference vectors (map unit locations in the measurement space) of SOM, instead of the original data set. In this approach, benefits of SOM in multivariate data visualization are combined with the clustering method, providing an easy way to analyze the properties of the clusters. In addition, this approach provides computational efficiency for very large data sets, for which the direct clustering would be very time consuming.

Several clusterings are performed, for example, with k-means algorithm and different values for k and the best is selected according to the Davies-Bouldin index.

3.7 Segmentation

For the process monitoring purposes it is important to detect changes in the statistics of the measured data. In order to track such variations, a segmentation for the

time-series data can be obtained, and the properties of the segments can be studied, for example, using different kind of visualizations. Next, the use of histogram maps and operator maps in time-series segmentation is described.

3.7.1 Histogram maps

In (Raivio et al., 2001, 2003), a method for segmentation of time-series data was proposed. In this method, the model for the segments is characterized by a distribution of the samples in a multivariate data space. In other words, the samples of the sequences in the same segment are well described by the same distribution. For simplicity, the distributions are approximated by a histogram describing the number of samples in different clusters obtained with basic clustering algorithms.

In this method, the SOM is used to visualize the variety of histograms observed in the data. This is established by obtaining the initial clustering and then, performing histogram computation for each sequence of time-series data. Then, the histogram data is visualized with the basic SOM algorithm.

Typically, the histogram data tends to form clusters of similar histograms. The k-means algorithm can be used to link similar histograms and therefore, clusters of histograms can be obtained. The sequences of time-series data whose histograms belong the same cluster are declared to belong to the same segment.

Raivio et al. (2001, 2003) applied the segmentation method for the classification of mobile cells into similarly behaving groups in order to obtain more efficient network configuration optimization. In Publications 1 and 2, the histogram maps are applied in mobile network performance monitoring problem.

3.7.2 Operator Maps

The basic SOM algorithm has been applied in many real-world data analysis problems (Kohonen, 2001). However, the basic SOM algorithm is mostly useful in cases where the data samples in the data set are statistically independent. In many process monitoring problems, the systems may be dynamical, that is, the past measurements made from the system do have an influence on current system output.

There are several generalizations of the basic SOM algorithm in which the map units of the SOM are considered as *parametric models* instead of static data points in the measurement space, resulting in so-called *operator maps* (Kohonen, 2001). The operators may be, for example, parametric probabilistic models (Hollmen et al., 1999) or dynamical models (Barreto and Araujo, 2000; Hyötyniemi et al., 1997; Lampinen and Oja, 1989). Such extensions of SOM are more useful in the analysis of dynamical process data, and especially, in segmentation of process data in the form of time-series.

Usually, the SOM algorithms are based on gradient algorithms of the form

$$\theta(t+1) = \theta(t) - \alpha(t)\frac{\partial E(t)}{\partial \theta(t)}, \tag{3.30}$$

where $E(t)$ is the cost to be minimized (Kohonen, 1993). At each iteration t (epoch), the whole measurement data $1 \leq n \leq N$ is used to compute the updated model parameter values, therefore, the learning occurs in batch mode. When the map units are considered to represent parameters of dynamical models, the error measure $E(t)$ can be defined as the average prediction error of the model given the system input and true output $y(n)$ for the system weighted over the neighborhood $h_{c(n),i}(n,t)$ of the lattice of map units

$$E(t) = \frac{1}{2N}\sum_{m\in L}\sum_{n=1}^{N} h_{c(n),m}(t,n)[y(n) - \hat{y}^{(m)}(t,n)]^2, \tag{3.31}$$

where $\hat{y}^{(m)}(t,n)$ is the prediction of operator m for time point t for sequence n.

Note, that the winner search procedure in which the sequence $c(n)$ for the neighborhood function is computed can be used to adjust the length of the subprocess segments that are assumed stationary. For example, the map unit with smallest mean squared error over a process segment of length L and starting from time step n_0 can be declared as the winner unit for observations $\{x_i(n), y(n)\}, n = n_0, n_0+1, ..., n_0+L-1$.

In this thesis, several kind of parameterizations for the operators are used. The adaptive filters are used as operators in order to obtain a segmentation based on differences in linear dynamics in the process. The MLP networks are used as operators in order to model nonlinear dynamics within the segments. Finally, ANFIS models are used for the same purpose as MLP operators. However, the ANFIS operators enable the analysis of the nonlinear dynamical models in terms of fuzzy rules, being more useful for process monitoring purposes.

For operators consisting of adaptive filters, the output of the operator m is defined by:

$$y^{(m)}(n) = \sum_{i=1}^{M} w_i^{(m)} x_i(n), \tag{3.32}$$

where M is the dimension of the input space, $w_i^{(m)}$ is the ith weight of the mth adaptive filter corresponding to the ith input variable $x_i(n)$. The advantages of this map unit parameterization are the easy interpretation of the model parameters as correlation between variables and the efficient optimization of the parameters. Since the model is linear in parameters, a solution corresponding to the global minima can be obtained (for a single operational mode of the process).

In order to model any temporally local nonlinear dependence between the past and current observations, the parameterization of the map units can be generalized according to the MLP. In such a case, the output of the ith neuron in layer l of a single operator m is

$$y^{(m,l)}(n) = \varphi\left(\sum_{j=0}^{M_{l-1}} w_{ij}^{(m,l)} y_j^{(m,l-1)}(n)\right), \tag{3.33}$$

where $y_j^{(m,l-1)}(n)$ is the jth input component for the neurons at layer l, $w_{ij}^{(l)}$ is the jth synaptic weight of the neuron i at layer l, M_{l-1} is the number of inputs for layer l and φ is a nonlinear activation function. The main disadvantage of the MLP operators is the lack of interpretation for the estimated mapping.

In the analysis of spatio-temporal data, a human-comprehensible description about the process is needed. An adaptive model suitable for this purpose is the ANFIS model based on the zeroth-order Sugeno fuzzy inference system. In ANFIS, the dependence between the model input and output can be represented by a set of human-friendly fuzzy rules of the form: if x_1 is A and x_2 is B then y is r, where A and B are linguistic labels and r is a crisp consequent part parameter of the fuzzy rule. The output of an operator m in this case is:

$$y^{(m)}(n) = \frac{\sum_k \prod_{\forall i,j \in R_k^m} \mu_{ij}^{(m)}\left(x_i(n)\right) r_k^{(m)}}{\sum_l \prod_{\forall i,j \in R_l^m} \mu_{ij}^{(m)}\left(x_i(n)\right)}, \tag{3.34}$$

where $\mu_{ij}^{(m)}(x_i(n))$ is the value of the jth premise part membership function for the ith input component, $r_k^{(m)}$ is the crisp consequent part parameter of kth rule of the map unit m and R_k^m is the set of premise part membership functions that belong to the kth rule of the operator m.

Due to use of above mentioned nonlinear operators of different complexity, it would be appropriate to include a mechanism to prevent the over-fitting of the map units into the data in order to achieve better generalization capability. Thus, according to the regularization framework, an additional penalty term is added to the error measure to be minimized by the SOM

$$E_r(t) = \frac{1}{2N} \sum_{m \in L} \sum_{n=1}^{N} \left[h_{c(n),m}(t,n) \left[y(n) - \hat{y}^{(m)}(t,n) \right]^2 + \lambda \eta[\theta^{(m)}] \right] \tag{3.35}$$

where $\eta(\theta^{(m)})$ is the value of the parametric penalty function for the parameters of the mth map unit in the SOM and λ is the regularization parameter. Regularization can also be defined implicitly by using double resampling. In Publication 3, the operator maps were used in mobile network performance analysis.

3.8 Knowledge Engineering

According to Ljung and Glad (1994), there are two sources of knowledge for system properties. Firstly, the collected experiences of application domain experts and literature contain basic laws for the system operation. Secondly, the system itself, and especially, the observations made from the system is an important source of information. The methods explained in the previous sections are heavily based on using the measurement data from the system of interest. In this section, approaches are presented in which prior knowledge outside the data records play an important role in addition to the measurement data. In (Ljung and Glad, 1994), this is called knowledge engineering and it refers to tasks in which application domain knowledge and mathematical model construction techniques are combined in order to obtain knowledge-based models.

Examples of how prior knowledge can be used to focus on phenomena of interest include:

- selecting the most important variables from large amount of measurement data depending on the problem at hand,
- selecting a subset of time points (samples) in which the variables meet certain static requirements,
- including constraints for dependencies between measured variables in order to prune the amount of possible solutions, and
- deriving new mappings based on mathematical models in order to fill gaps between quantities of interest.

The use of application domain knowledge with the data-driven methods is strongly dependent on the problem at hand. In this chapter, only the principles of using prior knowledge are outlined. The more detailed description of using prior knowledge is given in the next chapter of this thesis, in which the results of applying data driven methods to radio resource optimization problems are described.

3.8.1 Variable and Sample Selection

In modeling tasks, it is essentially important to use only the relevant variables in modeling. In many cases, the variable selection problem can be attacked with methods not requiring prior knowledge about the data (Guyon and Elisseeff, 2003). In many applications such as telecommunications, the variables tend to be very colinear and correlate with the total amount of traffic in the system. In such applications, it is essentially important to know the meaning of the variables and to select the most relevant ones based on prior knowledge.

The simplest form of applying a priori information about the system in model construction is the selection of the important variables for the problem at hand. This step is nearly always executed when data analysis methods such as neural networks are used to process measurement data. In the analysis of data produced by networked systems, the topology and dependencies between individual data generating systems may be important to include in the system modeling. In the networked systems like wireless networks, this is especially important since geographically close-by BTSs are usually connected into same controllers and also, many of the cellular network performance problems tend to be local in nature. For these reasons, many modeling problems require that variables from close-by cells are included in the same model. Similarly, it may be necessary to include variables from other cells operating on the same frequency when building a model for a single BTS (see Publication 5).

In addition to the selection of important variables for the analysis, it is often useful to focus on certain value ranges of the selected variables. Especially in telecommunications, the KPIs have generally accepted thresholds that define the region of acceptable performance. When analyzing process faults, it is natural to focus

on measurements in which acceptable performance is not met. In (Vehviläinen, 2004), an example of data analysis method is given in which certain thresholds for important variables have been set and the data samples in which variables have values on the range of interest are taken into more careful analysis. In Publication 4, a similar approach to select a subset of data for more careful analysis is presented.

3.8.2 Constraining Dependencies between Variables

Sometimes, the prior information enables the simplification of the model estimation process by restricting the number of possible solutions. When building models with the counter data of the mobile networks, for example, it may become clear that the regression model must not have any parameters having negative values nor values greater to one. In addition, it is very common that the sum of regression model parameters must be equal to one. In Publication 5, such cases arise. In regression problems with such restrictions, the solution can be obtained with quadratic programming presented earlier in this thesis.

3.8.3 Importing Mathematical Models

In some cases, it is necessary to derive mappings between existing variables. The derivation of the mappings is typically based on mathematical models describing application domain phenomena.

In (Oussar and Dreyfus, 2001), a semi-physical approach is presented in which the prior knowledge is in the form of continuous differential equations. Firstly, the continuous differential equations are discretized and then, the selected components of the discrete model are implemented by a neural network. The results of increasing the number and complexity of the neural network based components indicate that a model of increasing accuracy for the analyzed process is obtained.

Another approach of including prior knowledge is the case in which human expert is able to give simple, linguistic rules about how the system operates/should operate in some important operating regimes. These rules can be converted to fuzzy rules and a fuzzy system can be obtained to perform the modeling or control task. In this case, the prior knowledge takes the form of rules, and no mathematical characterization of the knowledge is needed.

In telecommunications, the use of mathematical models in combination with measurement data is a common practice. The most typical examples of mathematical models in the telecommunication applications are path loss models and the Erlang-B formula describing the capacity of mobile networks. In Publication 6, a method for blocking prediction in GSM networks based on Erlang-B formula is presented. In Publication 7, the path loss models are used in prediction of signal quality and call dropping in the context of cell size optimization. The use of path loss models allows the calculation of cell radius with different signal strength requirements.

Chapter 4

Data-Driven Radio Resource Management

In this chapter, the results of applying data analysis methods and knowledge engineering techniques for the optimization of wireless network performance are presented. Firstly, the results of visualization of multivariate performance data of a 3G network simulations are discussed. Then, the results of performance analysis using visualization are presented for operating GSM networks. Finally, the predictive modeling approach for optimization of configuration parameters of an operating GSM network is discussed.

4.1 Expert-Based UMTS Network Optimization

In this section, the results of applying data visualization to cellular network data are presented. The purpose of the methods is to visualize useful information for human expert responsible for optimizing network performance. Since the task is usually associated with multiple performance measures and reasonable overall performance is of interest, this step reduces to visualization of multivariate data.

4.1.1 Network Scenarios

The UMTS network analysis is performed for two network plans. Firstly, a microcellular network scenario consisting of 46 base stations with omnidirectional antennas was analyzed. The network scenario is depicted in Figure 4.1. The data from this network scenario was analyzed in Publication 1 with the histogram map, the focus being on the analysis of downlink (BS to MS) performance of this network. In Publication 2, the analysis of this network scenario focused on the uplink (MS to BS) performance monitoring.

The second network plan consisted of 31 base stations with sector antennas and one

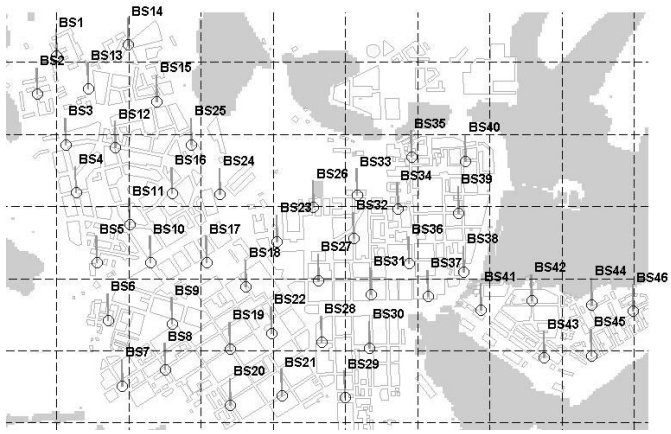


Figure 4.1: Microcellular network scenario with 46 base stations with omnidirectional antennas.

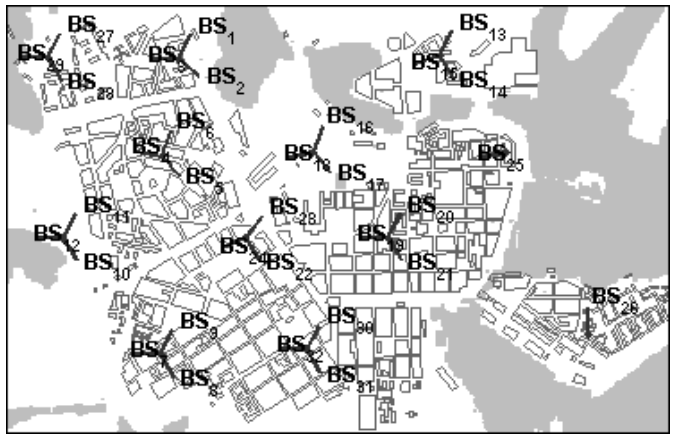


Figure 4.2: Macrocellular network scenario with 31 base stations with sector antennas and one omnidirectional antenna.

base station with omnidirectional antenna. This macrocellular network scenario is depicted in Figure 4.2. This network scenario was analyzed in Publication 3 with the operator maps. In these analysis, the focus was on the downlink performance.

4.1.2 Cell Monitoring

When a large data record describing the operation of wireless network is at hand, it is natural to question how it should be used in order to analyze how the network is operating. For this purpose, a some kind of description about process states and their variations should be obtained, for example, in graphical form in order to provide this information for human expert. Next, three approaches to define a process state and to estimate the process states in different time segments are presented (see Section 3.7).

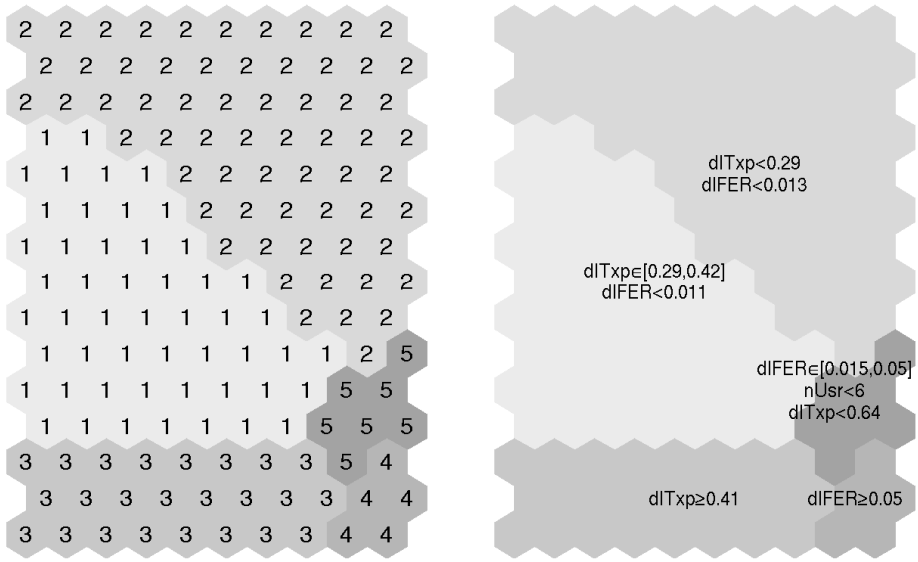


Figure 4.3: (a) Visualization of data clusters and (b) their properties using SOM.

Histogram Map

In Publication 1, the problem of monitoring the state of the cells in a microcellular UMTS network scenario was studied with the use of histogram map. The focus on downlink performance of the network. Firstly, the most important variables from all the 46 cells were used to train a SOM. These include the number of users (nUsr), downlink transmission power (dlTxp) and downlink frame error rate (dlFER). Then, the codebook of the SOM was clustered into five clusters using k-means clustering algorithm. Figure 4.3(a) shows the trained SOM and how the SOM grid is separated into clusters obtained with the k-means algorithm. The rules describing the clusters in the measurement space are shown in Fig. 4.3(b). From the rules it easy for application domain expert to assess the properties of each data cluster. For example, cluster 4 represents data samples with unacceptable high downlink frame error rate (dlFER > 0.05).

In order analyze slow variations in the behavior of the mobile cells, segmentation for the time-series data was performed. The histogram map was used in segmentation by first dividing the data into sequences of equal length, and computing the number of samples in previously computed clusters for each of the data sequence. The histograms of the sequences were used to train the histogram map.

Similarly to the SOM trained with the mobile network data, the histogram map can be clustered in order to obtain the segmentation of the time-series data. The trained histogram map was clustered into seven clusters using the k-means clustering algorithm in order to separate dissimilar areas on the map. Figure 4.4(a) shows the clustered histogram map.

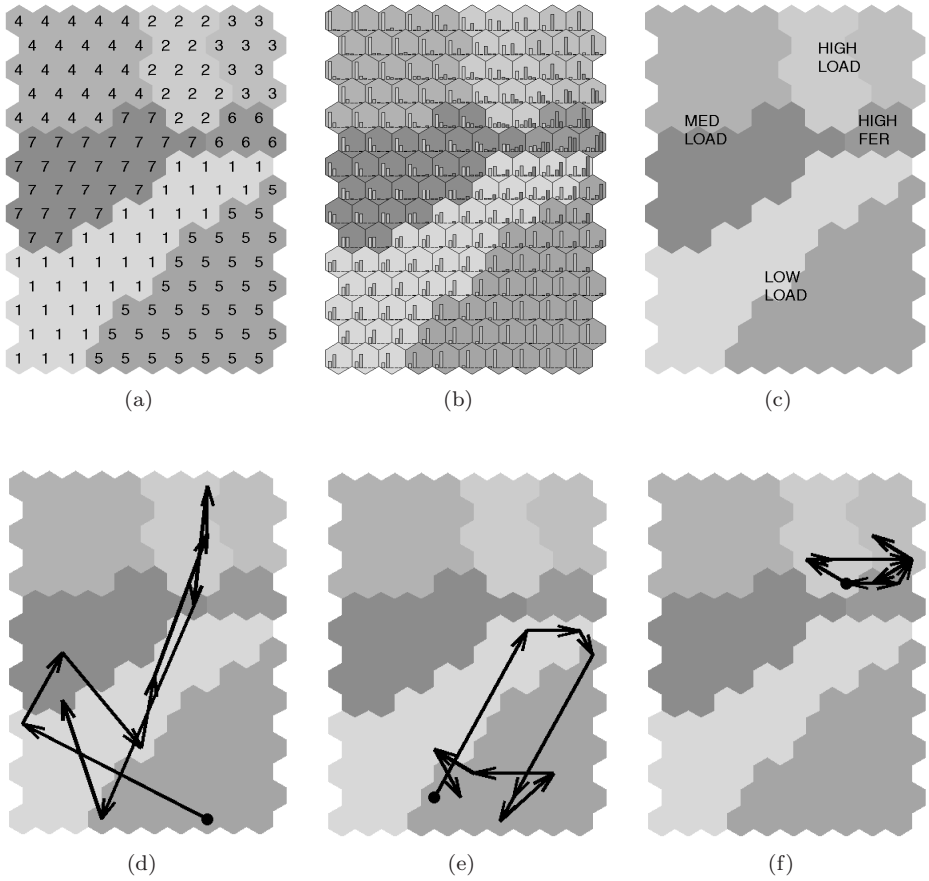


Figure 4.4: (a) Clusters on the histogram map and (b) their properties visualized by histograms. (c) Properties of clusters are manually labeled on top of SOM. Trajectories of mobile cells (d) 8, (e) 14 and (f) 44 of the mobile network on the histogram map.

In Figure 4.4(b) the properties of each cluster on the histogram map are shown. Each cluster represents differently distributed data samples in the measurement space. These distributions are visualized as a data histogram in which each bar represents the density (occurrence probability) of the corresponding data cluster in the measurement space. For example, cluster 6 on the histogram map represents density structures in which most of the data samples are located in data cluster 4 as indicated by the height of the 4th bar of the histograms. As mentioned earlier, the data cluster 4 represents data samples with unacceptable high downlink FER, indicating that cluster 6 on the histogram map describes undesired data distribution and thus it should be avoided by making corrections to the process configuration parameters. In Figure 4.4(c), the histogram map in which the cluster properties are manually labeled on the map is shown.

In order to monitor the behavior of the mobile cells, the trajectory of the map units on the histogram map that best describe the state of the consecutive sequences

of the cell data can be visualized. By following the trajectory, the changes in cell behavior are easily detected. Figures 4.4(d)-(f) illustrates this monitoring process for mobile cells 8, 14 and 44. For example, cell 8 starts from low load state, visiting medium load and high load areas before settling back to medium load area. Mobile cell 14 operates on low load, except a small peak in dlFER in the beginning of the data set. Cell 44 operates exclusively on high load and dlFER area.

In Publication 2, similar analysis was performed for the uplink data of the same network scenario. In addition, the presented results were compared with traditional network analysis based on uplink loading equations.

Linear Operators

A second approach to cell monitoring is based on tracking the changes in linear dynamics in the cells. This was based on the use of operator maps in which the operators were adaptive linear filters (see Publication 3). In Figure 4.5(a) the parameters of the adaptive filters of the $[4 \times 4]$ operator map are shown. Each bar of the histogram indicates the value of a single parameter of the corresponding model. The first bar corresponds to the coefficient of the non-delayed sample of the number of users and the second bar corresponds to the coefficient of the first delayed sample of the number of users. The third bar corresponds to the non-delayed sample of downlink average transmission power, and the fourth bar corresponds to the delayed sample of the downlink average transmission power. The last bar indicates the bias term of the model, that is, the mean of the output process (downlink frame error rate).

From the figure it can be seen that the parameters of the map are topologically ordered since the neighboring map units do have similar values for the parameters. Three clearly different areas can be found on this map: the map units in the upper left corner have the highest values for the coefficients corresponding to the current and past values of the number of users in the network. That is, the more users the mobile cell has, the worse is the quality of the service provided by the cell since there is a strong positive correlation between the number of users and the downlink frame error rate. In contrast, the coefficients corresponding to the downlink average transmission power tend to have small negative values. In other words, high values of the downlink average transmission power tend to have a weak lowering effect on the downlink frame error rate.

The upper right corner of the map represents similar behavior, except that the frame error rate is not well explained by the changes in the downlink average transmission power. This is due to the fact that the values of the corresponding parameters are nearly zero. Thus, the only conclusion that can be made is that the higher the number of users in the cell is, the higher is the frame error rate.

The lower half of the map consists of models with small positive values for all parameters, stating that both the number of users and the downlink average transmission power correlate positively with the frame error rate. Thus, the bad quality can up to some level be explained by the number of users, but an increase in transmission power does not seem to produce better performance.

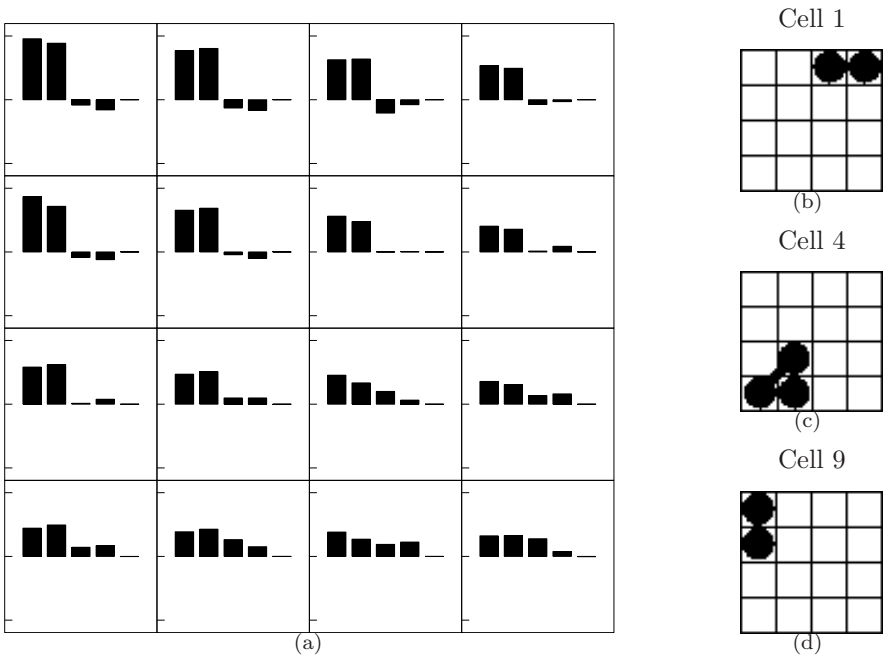


Figure 4.5: (a) Parameters of the adaptive filters of the operator map. The trajectories of the cells (b) 1, (c), 4 and (d) 9 of the mobile network.

The visualization of the parameters of the adaptive filters as shown in Figure 4.5(a) can be used to examine the spectrum of the data generating models in a piecewise stationary process. Similarly to the histogram map based visualization, the trajectories of the mobile cells on the operator map can be used to analyze long-term variations in the cells. In Figures 4.5(b)-(d), examples of such trajectories have been shown. In cell 1, the variation in performance can be explained only in terms of the high number of users, since they visit the upper right corner of the map. Cell 4 is an example of base station in which the bad performance occurs when the number of users is high, but the increase of the downlink transmission power does not seem to help in reaching the network users. Cell 9 suffers from a decrease in performance when the number of users is high and the average transmission power too low since they visit the upper left corner of the operator map.

Non-Linear Operators

The third approach for cell monitoring is based on operator maps in which the operators are simple ANFIS models (see Publication 3). For purposes of network planning, it is of interest to know when the cells generate high values for the frame error rate. Thus, the analysis of the rule plane shown in Figure 4.6(a) can be used to track the conditions for various quality problems. For example, the map units {6, 7, 11} tend to produce high FER values when the number of users in the cell is low. At the same time, the number of users in the past were high, indicating that the bad performance occurs when the number of users drops. In map units {7, 11},

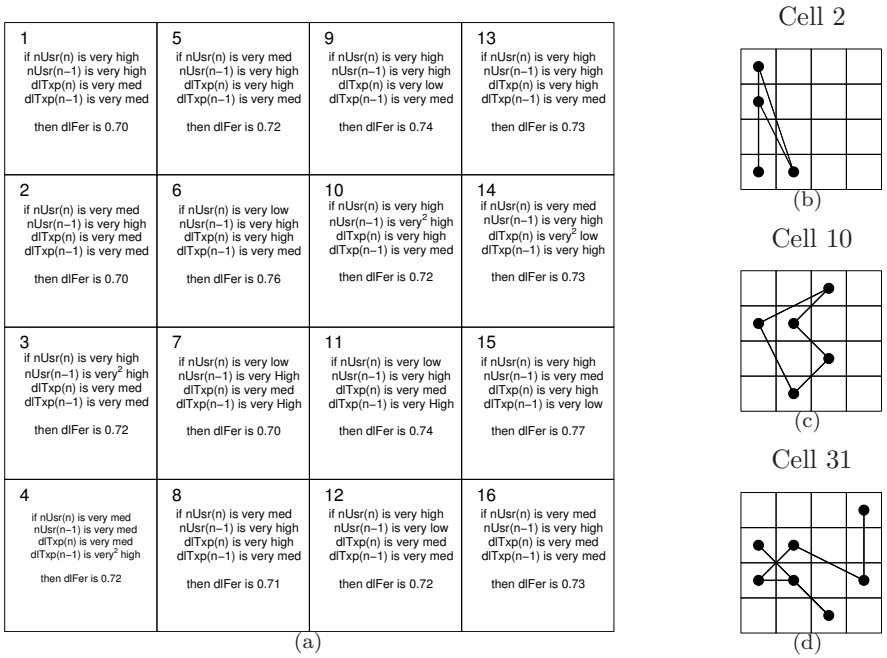


Figure 4.6: (a) Rule plane 3 of the best ANFIS operator map. The centers of the Gaussian premise part membership functions are described by the user-selected label set {LOW, Low, low, med, high, High, HIGH}. The shape of the membership functions are described by $\mu_{very^n A}(x) = \mu_A(x)^{n+1}$. The trajectories of the mobile cells (b) 2, (c), 10 and (d) 31 of the mobile network on the map of ANFIS models.

the transmission power drops from high to medium, in contrast with map unit {6} in which the increase in transmission power from medium to high produces the bad quality. The map unit {12} seems to generate bad performance when there is a rapid growth in number of users with transmission power being constantly medium.

According to the figure, cells 10 and 31 visit the map units representing undesirable behavior and the cell 2 represents normal operation.

4.1.3 Cell Grouping

In cell monitoring, the interest was in the analysis of variations in the cell performance during the network operation. This was accomplished by dividing the measurement period into sequences, and performing a segmentation among the sequences and monitoring the changes in the currently active segment.

Such a segmentation can be performed for long data sequences also, thus establishing a grouping of cell behavior according to long-term characteristics of operation. Such a grouping is especially useful when the static configuration parameters of the base stations need to optimized. In such cases, the variations in short term behavior is not as useful information as the long-term behavior. Since the opti-

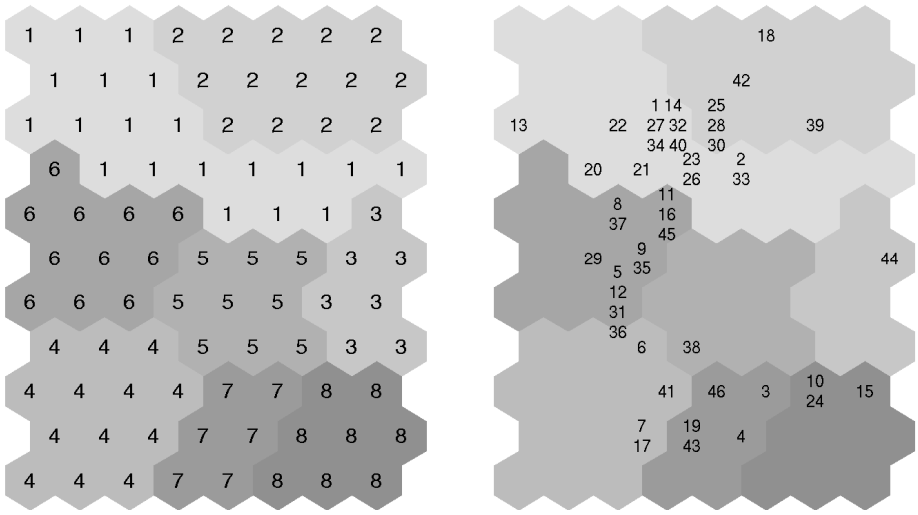


Figure 4.7: (a) Clusters, (b) classification of mobile cells.

mization of the mobile network parameters cell by cell can be very time consuming, the cell grouping can be efficiently used to simplify the optimization task. In other words, the mobile network parameter optimization reduces to optimization of the parameters of the clusters of cells instead of cell by cell (Laiho et al., 2002c) so that the cells in the same cluster have the same configuration parameters. The histogram map and the operator maps described in the previous section can be used to obtain this cell classification.

In Publication 2, the histogram map was used to obtain the cell grouping. In Figure 4.7(a), the clustered histogram map used for cell grouping is depicted. This classification is based on uplink data of the microcellular network scenario. The histogram map consists of 8 behavioral clusters, each indicated by a different gray-level. In Figure 4.7(b), the map units best describing the data from different mobile cells are shown.

4.2 Expert-Based GSM Network Optimization

4.2.1 A SOM Based Visualization Process

In Publication 4, a SOM based analysis process to track performance bottlenecks in operating GSM network was presented. The proposed SOM based analysis process is illustrated in Figure 4.8. In this study, the KPI data from the GSM network was studied. The network consisted of 41 base stations from which the most important KPIs over 10-week time period were measured.

When projection methods such as the SOM are used in data visualization, all

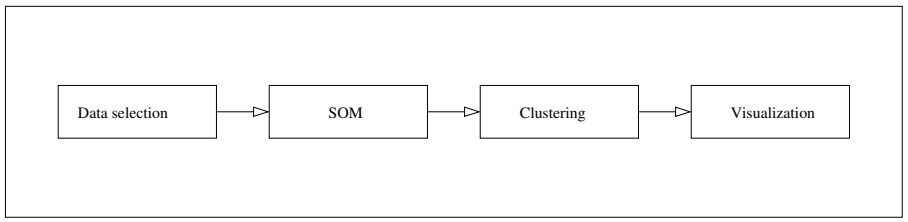


Figure 4.8: A block diagram illustrating the phases of the proposed analysis process.

samples usually have equal priority when determining the projection (dimensionality reduction) function. In order to focus on certain network problem types, the samples representing normal operation can be removed before the more detailed analysis. When analyzing the performance degradations of a GSM network, the most important samples can be selected by choosing the KPIs of interest and retaining only the samples for which the operator defined objective values are not met. For example, if an accurate visualization of traffic channel problems are desired, it would be justified to use only the samples in which traffic channel or signaling channel blocking occurs, or traffic channel drop rate exceeds 2%.

After the subset of data of interest is selected, the data is normalized in order to make all variables equally important independently on the measurement unit. Then, the normalized data is used as the input data in SOM training. The trained SOM is used to visualize the multi-dimensional input data using the component plane representation of SOM.

As mentioned earlier in the thesis, clustering can be used to divide performance data into different clusters representing different types of failures. In this study, the approach presented in Section 3.6.4 was adopted in which the clustering is performed for trained SOM in order to provide simple visualization of cluster properties. After the SOM training and clustering, a visualization of the selected multi-dimensional input data is obtained. This information helps the application domain expert to make inferences about the possible problem scenarios present in the data. The cluster analysis based on SOM component planes reveals the variety of failures faced by the network. It is relatively easy task for an expert to select the most important variables (KPIs) for each failure type. By analyzing the amount of samples in different fault clusters originating from each cell of the GSM network, the locations of the different failure types are efficiently obtained. Finally, the visualization of the times of occurrence of different fault types reveals additional temporal information about the faults. These three types of simple visualizations allows the selection of variables, cells and time periods that are taken into further analysis using conventional methods.

In the analysis of SDCCH problems in the network, the samples meeting the requirement SDCCH Blocking > 0 % was used in order to filter the uninformative (normal operation) samples from the analysis. Then, histogram equalization based normalization method were applied for the selected data set in order to obtain invariance w.r.t the scales of the variables. Then, a SOM was trained in which the map units were organized in a $[15 \times 10]$ hexagonal grid by applying 500 epochs of

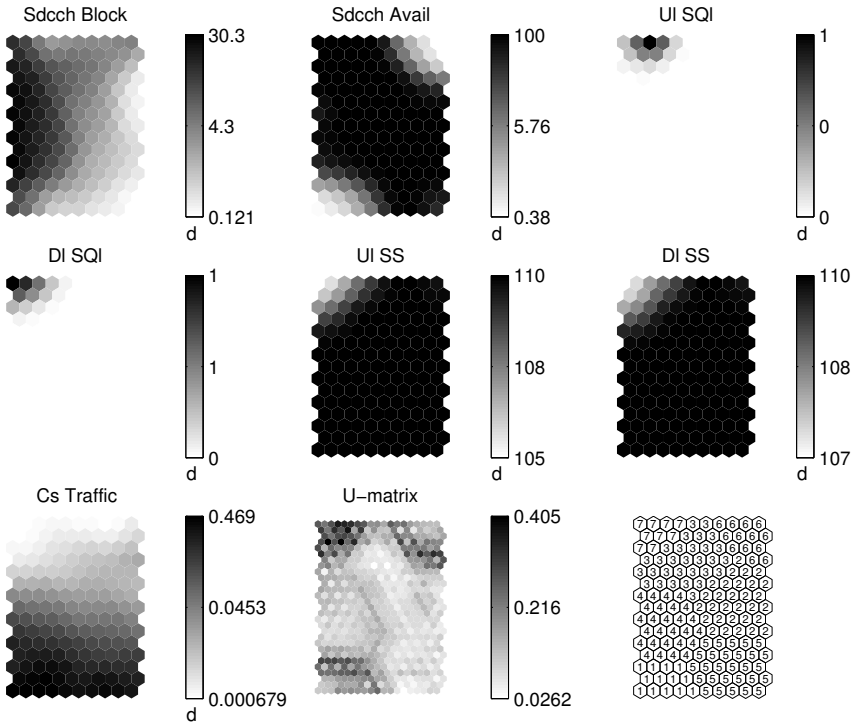


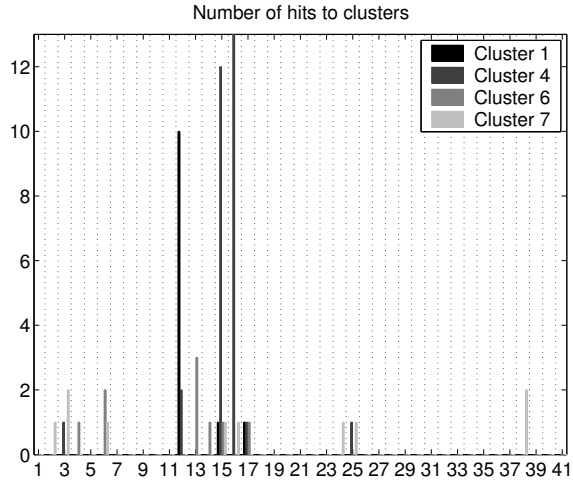
Figure 4.9: SOM of data samples representing possible signaling channel capacity problems. Clusters 1, 4, 6 and 7 represent possible signaling channel capacity problems.

batch training and 500 epochs of sequential training.

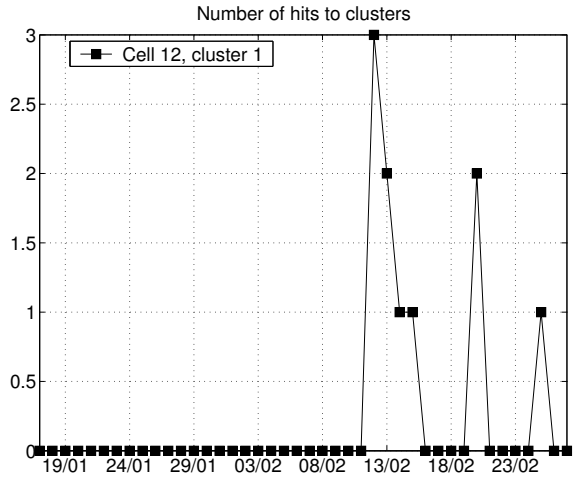
In Figure 4.9, the so-called component planes of the SOM are shown (Vesanto, 1999). In addition to reveal the cluster structure, the cluster indices of the map units obtained with the k-means are shown along with the visual U-matrix representation of SOM (Ultsch and Siemon, 1990).

By analyzing the properties of each cluster using the component planes, four clusters that represent possible signaling channel capacity problems can be identified: cluster 4 contains high values for signaling channel blocking, with moderate amount of traffic. Clusters 1 and 6 represent behavior in which a drop in channel availability is likely to cause the high blocking values. Cluster 7 represents channel blockings that are likely to be a result of bad signal quality, i.e. the connection is refused because the required channel quality could not be provided. The U-matrix reveals that the clusters 1, 6 and 7 are located further apart from the other clusters.

The locations of the various fault types can be analyzed by computing the number of hits into different fault clusters (see Figure 4.10(a)). From the figure, it becomes evident that nearly all of the samples in the fault clusters were generated by only three cells of the network. Hits from other cells can be viewed short-term failure situations that do not give reasons to major configuration adjustments and



(a)



(b)

Figure 4.10: (a) Cells 12, 15 and 16 contribute the most to the fault clusters. (b) In cell 12, the failures appear mostly during a 4-day period.

therefore can be ignored.

When plotting the times of occurrences of the hits to the fault clusters from these three cells, it was found that the cell 12 had a 4-day period when most of the samples into fault cluster 1 were generated (see Figure 4.10(b)). This suggests that the signaling channel availability were temporarily reduced (i.e. the amount of available channels dropped) and therefore, some of the channel requests were blocked. In order to verify this assumption, it was easy to plot the variable describing the number of available SDCCH channels against the number of blocked SDCCH requests during that 4-day time period. After this simple plotting, it was even more clear that it is the drops in availability that causes the requests to be

blocked. This procedure is a good demonstration that failures related to multiple variables can be easily found, the base stations and the time periods in which they occur is easy to obtain and verifying plots are straightforward to accomplish with this kind of analysis procedure.

In Publication 4, the same analysis procedure was followed for the analysis of TCH capacity problems in the same network. Several interesting failure situations were found from the data. Firstly, the drop in channel availability in close-by cells was detected. In addition, blocking problems in two cells due to the lack of resources were detected. In addition, interesting relationships between cell size and call dropping were found, stating that high call dropping rate is often observed in small cells and during time periods with low amount of traffic.

4.2.2 A Knowledge-Based Visualization Process

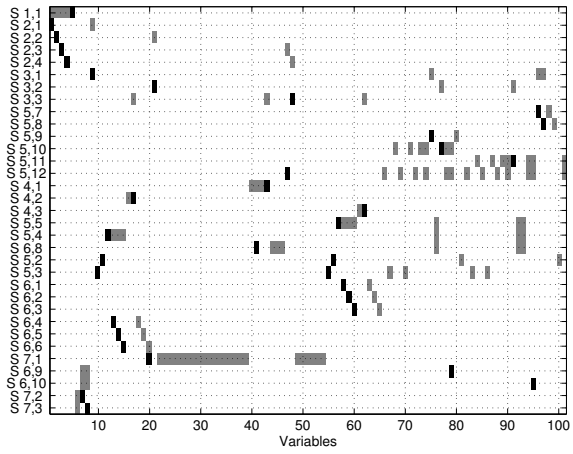
In previous section, the analysis of GSM network performance was based on KPI data. Also, the main objective have been in detecting the type and geographical location of the performance bottlenecks. The network analyzer have been responsible for analyzing the cause of the performance problems. In this section, the performance is analyzed using the raw counter data. By using the counter data, the possible causes for the performance problems can be tackled more automatically based on observations, which is not possible on the basis of KPI data.

In Publication 5, a multi-layer hierarchical structure of cause-effect pairings were developed. The set of available measurements were assigned into variable groups based on a priori knowledge about the semantics of the variables. Within a variable group, the variables are clearly connected while the variables from different groups are independent on each other. Therefore, the modeling problem is more easily solved by identifying a separate subsystem for each variable group separately.

In Figure 4.11(a), the memberships of the variables in different groups (subsystems) are shown. This plot shows the set of input variables (gray) and output variables (black) that belong to each of the subsystems (y-axis). The subsystems tend to form a hierarchical structure (see Figure 4.11(b)), i.e. the outputs of a subsystem describing some low-level phenomena can be an input to a higher-level subsystem. In the figure, the solid lines indicate that the input of the upper level system contains outputs of the lower level system from the same BTS only. The dashed line indicates, that the input of the upper level system contains input signals from lower level systems of the other BTSs also.

For each variable group, a simple constrained optimization problem with suitable constraints were defined. In Table 4.1, the different types of models and the corresponding constrained optimization problem for the subsystems are shown. For different model types, a priori knowledge about the associated variables was used to determine the constraint equations and, for example, the presence of bias in the model.

Once the model parameters were estimated from the data, the cause-effect chains related to performance degradations in the network can be analyzed using the model parameters. In order to visualize the most interesting and useful cause-



(a)

(b)

Figure 4.11: (a) Input and output variables (x-axis) of different subsystems (y-axis). (b) The subsystem hierarchy.

effect chains, a list of strong dependencies was obtained. An item to a dependency list is generated per each input-output variable pair, including the strength of the dependency between the input and output variable (the value of the parameter a), a measure of model accuracy (root mean square prediction error (RMSE) of the model), and a measure of models importance in overall network performance analysis (the average number of failures stored in the output variable of the input-output variable pair).

After all the models have been estimated and the properties of each input-output variable pairs are stored in the dependency list, a tree-shaped graph is constructed in order to analyze the cause-effect chains generating the major performance degradations of the network. Since the number of theoretically possible dependencies is extremely large, only the most important dependencies are included to the depen-

Table 4.1: The different types of models for the subsystems.

Type	Systems	Model	Parameter estimation
I	$S_{1,1}$, $S_{2,1}$, $S_{2,2}$, $S_{2,3}$, $S_{2,4}$, $S_{3,1}$, $S_{3,2}$, $S_{3,3}$	$y(t) = \sum_i x_i(t)$	$a_i = \sum_t x_i(t) / \sum_t y(t)$
II	$S_{5,7}$, $S_{5,8}$, $S_{5,9}$	$y(t) = ax(t) + b$ $= \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} x(t) \\ 1 \end{bmatrix}$ $= \tilde{\mathbf{a}}^T \tilde{\mathbf{x}}(t)$	$\min_{\tilde{\mathbf{a}}} \frac{1}{2}\tilde{\mathbf{a}}^T \tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \tilde{\mathbf{a}} - \mathbf{y}^T \tilde{\mathbf{X}} \tilde{\mathbf{a}}$ $\tilde{a}_1 \in [0, 1]$ $\tilde{a}_2 \geq 0$
III	$S_{5,10}$, $S_{5,11}$, $S_{5,12}$, $S_{6,9}$, $S_{6,10}$, $S_{7,1}$	$y(t) = \mathbf{a}^T \mathbf{x}(t) + b$ $= \begin{bmatrix} \mathbf{a} & b \end{bmatrix} \begin{bmatrix} \mathbf{x}(t) \\ 1 \end{bmatrix}$ $= \tilde{\mathbf{a}}^T \tilde{\mathbf{x}}(t),$	$\min_{\tilde{\mathbf{a}}} \frac{1}{2}\tilde{\mathbf{a}}^T \tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \tilde{\mathbf{a}} - \mathbf{y}^T \tilde{\mathbf{X}} \tilde{\mathbf{a}}$ $\sum_{i \neq N} \tilde{a}_i = 1,$ $\tilde{a}_{i \neq N} \in [0, 1]$ $\tilde{a}_N \geq 0$
IV	$S_{4,1}$, $S_{4,2}$ $S_{4,3}$, $S_{5,2}$, $S_{5,3}$, $S_{6,1}$, $S_{6,2}$, $S_{6,3}$, $S_{6,4}$, $S_{6,5}$, $S_{6,6}$	$\mathbf{y}(t) = \mathbf{x}(t)\mathbf{A},$	$\min_{\mathbf{A}} \sum_i \frac{1}{2}\mathbf{A}_i^T \mathbf{X}^T \mathbf{X} \mathbf{A}_i - \mathbf{y}_i^T \mathbf{X} \mathbf{A}_i$ $\sum_i a_{ij} = 1,$ $a_{ij} \in [0, 1]$
V	$S_{5,4}$, $S_{5,5}$, $S_{6,8}$	$y(t) = a_1(c_1(t) + c_2(t) + c_3(t))x_1(t)$ $+a_2(c_1(t) + c_2(t) + c_3(t))x_2(t)$ $+a_3(c_1(t) + c_2(t) + c_3(t))x_3(t),$	$\min_{\mathbf{a}} \frac{1}{2}\mathbf{a}^T \mathbf{X}^T \mathbf{X} \mathbf{a} - \mathbf{y}^T \mathbf{X} \mathbf{a}$ $\sum_i a_i = 1,$ $a_i \in [0, 1]$

dency tree.

Three criteria are used to prune uninteresting dependencies from the tree. Firstly, the model accuracy from which the dependency originates must be at a reasonable level. Otherwise, the analysis might be misled by very inaccurate models having large values for parameter a (which is forced in several models due to the equality constraints for the parameter vector). Secondly, the output variable of the dependency must be interesting enough (i.e relatively large number of failures must be observed in the output variable). Finally, only the dependencies that belong to the cause-effect chains contributing most to the overall network performance degradations are included into the dependency tree. For each subsystem, different minimum and maximum values for strength of dependency, model accuracy and model interestingness are defined.

Figures 4.12(a)-(d) show the pruned dependency trees in four separate cases. In Figure 4.12(a), the cause-effect chains of the most significant blocking problems are shown. Clearly, there are 4 BTSs (6,11,18,85) that suffer from lack of resources. BTSs (6,11,18) suffer from lack of half rate traffic channels and BTS 85 suffers from lack of full rate traffic channels. Only BTS 6, the causes for blocking can be said to result regularly from congestion.

In Figure 4.12(b), the results of the corresponding analysis for the call setup failures are shown. Here, four BTSs (17,52,66,74) seem to suffer from call setup failures regularly. In all these four BTSs, the failures tend to originate during SDCCH signaling and fail due to radio link problems. In BTSs 17 and 74 the radio link failures can be said to result from bad downlink signal quality and in BTSs 52 and 66 they are due to bad uplink signal quality.

Figure 4.12(c) shows the corresponding results for call dropping problems. Again, the cause-effect chains for describing the reasons for dropped calls in four BTSs

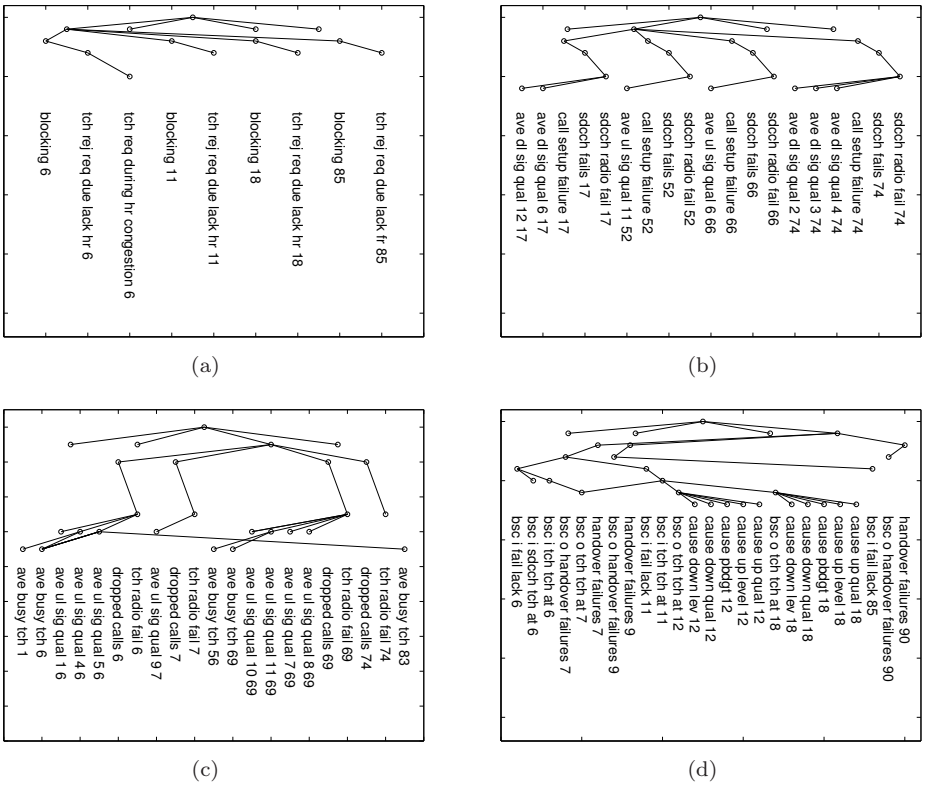


Figure 4.12: The analysis of the four main components of the user perceived quality. (a) Cause-effect chains for blocking of services, (b) call setup failures, (c) call dropping and (d) handover failures.

(6,7,69,74) are shown. The reasons for call dropping seem to be radio link failures. In BTSs 6, 7 and 69 the radio failures are likely due to bad signal quality in uplink. In two TRXs of BTS 6 and in one TRX of BTS 69 the number of bit errors seem to correlate with the amount of traffic in both the own BTS as well as in interfering TRXs on the same radio frequency.

Finally, in Figure 4.12(d) the analysis of the handover problem sources are shown. The results for three BTSs (7,9,90) having the worst handover performance are shown, indicating that the problems are in BSC controlled outgoing handovers. In BTSs 7 and 9, the problems seem to be in lack of resources in the target BTSs (6,11,85). In target BTS 6 suffering from lack of resources, there seems to be high amount of incoming TCH-TCH handover attempts from BTS 7. This same BTS tend to cause problems also for target BTS 11 suffering from lack of resources during handover attempts. For target BTS 11, two other BTSs (12,18) are found that can be said to generate high number of handover attempts. The handover attempts of these two BTSs are due to very similar reasons: the quality of the uplink and downlink radio connection and the uplink and downlink signal strength are not reasonable in these two BTSs, and some of the users are switched into more appropriate BTSs. Also, significant number of users are switched to another BTS in order to minimize the energy consumption of the MSs (power budget).

This study is a good indication of how a priori knowledge can be used to ease the model estimation by pruning variables. Only the theoretically sound dependencies need to be estimated. Then, regression techniques can be used to estimate the strength of the theoretically possible dependencies, and only the strongest ones are taken into more careful analysis. The possible dependencies between cells can be analyzed by including variables from other cells to the models. The presented method allows efficient fault detection, identification of strongest input variables, and diagnosis of the causes of the faults.

4.3 Predictive GSM Network Optimization

In previous sections, methods for fault detection, identification and fault diagnosis have been discussed. In this section, the focus is on the final step of the process monitoring procedure, that is, on the process recovery. In the context of mobile network performance monitoring, the process recovery is associated with making adjustments to configuration parameters so that the amount of predictable, regularly occurring performance degradations or faults is minimized. In order to automate the configuration parameter optimization, a computational method to evaluate the performance of alternative configurations must be available. In data-rich environments like cellular networks, such predictive models are most efficiently obtained with the use of past data records. Next, the data analysis methods and suitable knowledge engineering approaches are used to build predictive models that can be used in automated configuration parameter optimization.

4.3.1 Prediction of Blocking

In blocking prediction, the interest is to compute the number of blocked requests at different conditions. This can be based on the use of well known Erlang-B formula. The expected value for the number of blocked requests is obtained by multiplying the number of arriving requests with the blocking probability given by Equation 2.4, leading to $B = \lambda p(N_c|\lambda, \mu, N_c)$. The expected value for the congestion time is $C = p(N_c|\lambda, \mu, N_c)$ and the expected value for the number of channels in use is $M = \sum_{n=0}^{N_c} np(n|\ \lambda, \mu, N_c)$.

In Publication 6, it was shown that the Erlang-B formula does not provide accurate predictions for blocking in GSM networks if low sampling rate measurements of arrival process are used in the model. More traditional regression methods can be used for the same purpose with the assist of knowledge engineering approach in which Erlang-B formula and regression methods are combined. With the use of Erlang-B formula, the dependencies between B, C and M that remain the same in each base station system need not be estimated from data alone. The data can be used to estimate other relevant and additional parameters that are required in prediction. In Publications 6 and 7, a method to use Erlang-B formula and measurement data to predict blocking is presented. The regression techniques are used to estimate the arrival rate distribution describing the arrival process during short time periods. The Erlang-B formula is used to compute the amount

of blocking during the short time periods.

Suppose that the time period is divided into N_s segments of equal length. Also, assume that we have a vector $\boldsymbol{\lambda} = [0 \;\; 1\Delta_\lambda \;\; 2\Delta_\lambda \;\; \ldots \;\; (N_\lambda - 1)\Delta_\lambda]$ of N_λ possible arrival rates per segment with discretization step Δ_λ. Let us denote the number of blocked requests during a segment with arrival rate λ_i with $B_i = \lambda_i p(N_c|\lambda_i, \mu, N_c)$, where $p(N_c|\lambda_i, \mu, N_c)$ is the blocking probability given by the Erlang distribution. Also, the congestion time and the average number of busy channels during a segment with arrival rate λ_i are denoted with $C_i = p(N_c|\lambda_i, \mu, N_c)$ and $M_i = \sum_{n=0}^{N_c} np(n|\lambda_i, \mu, N_c)$. In other words, the segment-wise values for blocked requests, congestion time and average number of busy channels are based on the Erlang-B formula.

Now, assume that the number of segments with arrival rate λ_i is θ_i and $\sum_i \theta_i = N_s$. Then, the cumulative values over one hour for the number of requests T, blocked requests B, congestion time C and average number of busy channels M can be computed with

$$\begin{bmatrix} \lambda_1 & \lambda_2 & \ldots & \lambda_{N_\lambda} \\ B_1 & B_2 & \ldots & B_{N_\lambda} \\ \frac{C_1}{N_s} & \frac{C_2}{N_s} & \ldots & \frac{C_{N_\lambda}}{N_s} \\ \frac{M_1}{N_s} & \frac{M_2}{N_s} & \ldots & \frac{M_{N_\lambda}}{N_s} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_{N_\lambda} \end{bmatrix} = \begin{bmatrix} T \\ B \\ C \\ M \end{bmatrix} \tag{4.1}$$

or in matrix notation $\mathbf{X}\boldsymbol{\theta} = \mathbf{Y}$.

Now, the problem is that the vector $\boldsymbol{\theta}$ is unknown and it must be estimated from the data using the observations of $\mathbf{Y}$ and matrix $\mathbf{X}$ which are known a priori. Since the output vector $\mathbf{Y}$ includes variables that are measured in different scales, it is necessary to include weighting of variables into the cost function. By selecting variable weights according to their variances estimated from the data, the quadratic programming problem

$$\min_{\boldsymbol{\theta}} \quad \left\{ \frac{1}{2}\boldsymbol{\theta}^T \mathbf{H}\boldsymbol{\theta} + \mathbf{f}^T\boldsymbol{\theta} \right\} \tag{4.2}$$

$$w.r.t \quad 0 \leq \theta_i \leq N_s, \;\; i = 1, 2, ..., N_\lambda, \tag{4.3}$$

$$\sum_{i=1}^{N_\lambda} \theta_i = N_s \tag{4.4}$$

is obtained where $\mathbf{f} = -\mathbf{X}^T\mathbf{W}^T\mathbf{W}\mathbf{Y}$ and $\mathbf{H} = \mathbf{X}^T\mathbf{W}^T\mathbf{W}\mathbf{X}$ include the weighting matrix $\mathbf{W}$. In other words, the goal is to find the vector $\boldsymbol{\theta}$ that provides the smallest prediction errors for variables T, B, C and M. For many BTSs, the number of blocked requests B and the congestion time C are regularly zero and the use of the corresponding rows in $\mathbf{X}$ and $\mathbf{Y}$ is not possible due to the numerical problems occurring in the optimization. In such cases, the solution vector $\boldsymbol{\theta}$ can be solved using only the two remaining rows of $\mathbf{X}$ and $\mathbf{Y}$.

Now, suppose that we have N_d observations for variables T, B, C and M, all measured at hour h during different days. The above optimization problem could be solved for each of the N_d observation vectors separately, leading to N_d solution vectors $\boldsymbol{\theta}$ for hour h. Since we are interested in long-term prediction of blocking,

we should somehow combine the solution vectors so that behavior common to all solution vectors are retained and non-regular properties of the demand are given less attention.

Let us denote the ith solution vector for hour h with $\boldsymbol{\theta}_h^{(i)}$ and the jth element of the corresponding solution vector with $\theta_{jh}^{(i)}$. Since $\theta_{jh}^{(i)}$ described the number of segments with arrival rate $\lambda = \lambda_j$ during ith observation vector at hour h, the probability for a random segment during ith observation period to have an arrival rate $\lambda = \lambda_j$ can be computed from $\theta_{jh}^{(i)}$ with $p_{jh}^{(i)} = \theta_{jh}^{(i)}/N_s$, where N_s is the number of segments in a period.

Now, if we are interested in occurrences of λ_j at hour h in the long run, it would be straightforward to sum the occurrences of the λ_j during the N_d observations. In other words, the probability for observing a segment with arrival rate $\lambda = \lambda_j$ at hour h would become

$$p_{jh} = \frac{1}{N_d N_s} \sum_{i=1}^{N_d} \theta_{jh}^{(i)}. \tag{4.5}$$

Now, the arrival rates λ_j and their probabilities p_{jh} for hour h form a probabilistic model. Let us define a column vector

$$\underset{seg \mapsto hour}{\boldsymbol{\theta}_h} = \mathbf{p}_h N_s \tag{4.6}$$

that maps the segment-wise candidate arrival rates λ_j to the total number of arrived requests T in a single one hour time period with

$$T = \boldsymbol{\lambda} \underset{seg \mapsto hour}{\boldsymbol{\theta}_h}. \tag{4.7}$$

Note that the parameter vector $\boldsymbol{\theta}_{h,seg \mapsto hour}$ can also be used to map the vector $\mathbf{B} = [B_1 \ B_2 \ \ldots \ B_{N_\lambda}]$ of segment-wise blocking candidates to the total number of occurrences of blocked requests during one period. Similarly, the cumulative values for the average number of busy channels and the congestion time can be computed.

4.3.2 Prediction of Signal Quality and Dropped Calls

In the previous section, a method to predict the amount of input traffic and blocking was established. Next, the focus is turned on the prediction of call dropping and bad voice quality caused by low signal quality observed at the receiver. The signal quality and the number of bit errors are mostly TRX level rather than BTS level measurements and therefore, it is necessary to extend the BTS-wise model to contain also TRX level details. The signal quality observed in a TRX is mostly dependent on the distance between BTS and MS, but the amount of traffic clearly has an influence on signal quality. This is most likely caused by interference that is more severe during high load time periods.

In order to predict the signal quality in a TRX, a measure of amount of traffic per TRX is needed. Since the loads in the TRXs of a BTS are strongly correlated

with the total amount of traffic in the BTS, a linear model

$$c_i(k) = \underset{T \mapsto c_i}{\theta_0} + \underset{T \mapsto c_i}{\theta_1} T(k) + e(k) \tag{4.8}$$

predicting the number of calls $c_i(k)$ in TRX i of a BTS is estimated. In the model, $T(k)$ is the number of TCH requests in the BTS during period k that is used as a measure of amount of traffic or load in the BTS. The parameters are estimated with the simple linear regression method described earlier in this thesis.

As stated in the previous section, the TRX level signal quality is mostly affected by the amount of load in the TRX as well as the distance between the MSs and the BTS. In addition, the signal quality is measured separately on the uplink (MS to BTS) and downlink (BTS to MS). Therefore, the prediction of signal quality is based on linear regression model

$$u_i(k) = \underset{\bar{r}, c_i \mapsto u_i}{\theta_0} + \underset{\bar{r}, c_i \mapsto u_i}{\theta_1} \bar{r}(k) + \underset{\bar{r}, c_i \mapsto u_i}{\theta_2} c_i(k) + e(k) \tag{4.9}$$

where $u_i(k)$ is a measure for number of bit errors in uplink of TRX i of a BTS during time period k, $c_i(k)$ is the measure for the amount of traffic in TRX i at time period k and $\bar{r}(k)$ is the average distance between the users and the BTS during the corresponding time period. The same model structure can be used for the prediction of downlink signal quality $d_i(k)$. The parameters for the models are estimated in the same way as in the model for predicting the TRX-wise traffic load. Note that BTSs having more than one TRX need separate models for each TRX. Now, the signal or voice quality in a BTS can be measured by combining the TRX-wise measures from the uplink and downlink with

$$Q(k) = \sum_{i=1}^{N_{TRX}} u_i(k) + \sum_{i=1}^{N_{TRX}} d_i(k). \tag{4.10}$$

Intuitively, call dropping is mostly caused by bad signal quality between the MS and the BTS. Let $D(k)$ be the sum of dropped calls due to the radio failures in SDCCH and TCH channels during the kth measurement period. A linear model

$$D(k) = \underset{Q \mapsto D}{\theta_0} + \underset{Q \mapsto D}{\theta_1} Q(k) + e(k) \tag{4.11}$$

between the number of dropped calls and the signal quality measure $Q(k)$ can be estimated from the data using basic linear regression techniques.

In these regression models, knowledge engineering is implemented in careful selection of input and output variables of the models.

4.3.3 Optimization of Signal Strength Thresholds

In Publication 7, the aim was to determine the influence of minimum signal strength threshold adjustment ΔP to the most important performance measures, that is, the signal quality Q, number of dropped calls D, blocked SDCCH requests B_s and blocked TCH requests B_t. In order derive this dependency, several parameters had to be derived from well known theories in addition to the large number

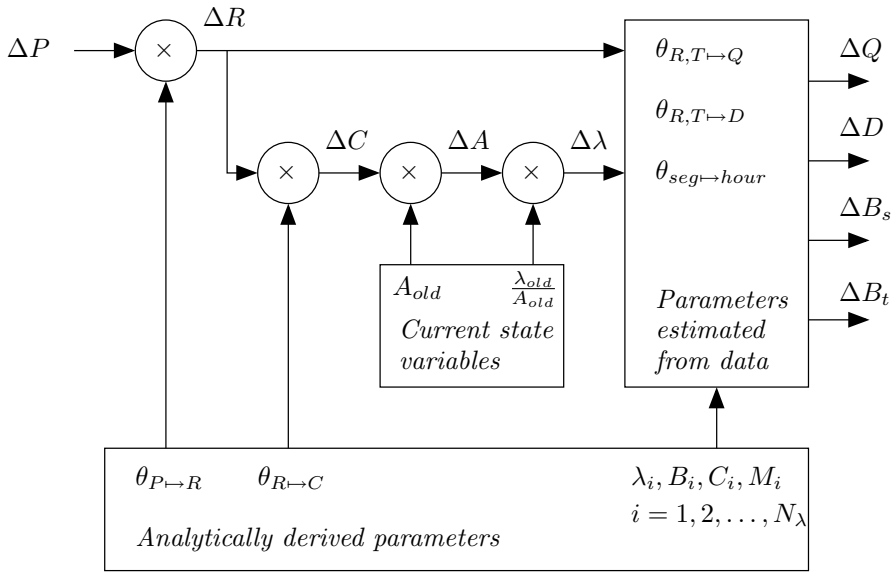


Figure 4.13: A block diagram describing the information flow from the input ΔP to the performance of the network.

of parameters that was estimated from the data. In Figure 4.13, a block diagram describing the use of analytically derived parameters, current state variables of the BTSs and the parameters that are estimated from the data is shown.

The input variable, ΔP, indicates the change in minimum signal strength that is required from a MS in order to allow access to the cell. This variable can be mapped to the change in operating radius of the cell by the use of path loss model. That is, the change in minimum signal strength is directly mapped to the change in maximum allowed path loss, which in turn can be used to compute the change in operating radius of the cell in meters. In other words, it is possible to derive the value of the parameter $\theta_{P\mapsto R}$ by the use of simple path loss modeling. Since the operating radius prior to adjustment can be estimated from the data, the new operating radius of a cell can be computed given the adjustment ΔP.

Now, the operating radius adjustment ΔR and the old operating radius R_{old} can be used to compute the new operating radius R_{new}. For the prediction of the signal quality and the number of dropped calls with the new configuration, the new operating radius is in very important role. However, it is also necessary to derive a measure of amount of traffic in the cell after the configuration adjustment. Assuming the cell coverage area is approximated by a round disk, the area change ΔA of the disk can be easily computed. Assuming the users are homogeneously distributed in the cell, the change in number of incoming transactions $\Delta\lambda$ can be computed given the ΔA and the measure of incoming transactions λ_{old} prior to adjustment.

In Publication 6, a method to estimate a discrete arrival rate distribution was presented. The presented method can be applied in prediction of SDCCH blocking

Table 4.2: The results of the experiments.

Number of optimized cells	Performance improvement in optimized cells	Performance improvement in whole network
5	13.1 %	1.3 %
10	11.2 %	2.0 %
15	12.1 %	2.8 %
20	13.6 %	3.3 %
25	12.1 %	3.7 %
30	9.8 %	3.7 %
40	8.7 %	3.8 %
50	7.4 %	4.0 %
60	6.9 %	4.2 %
80	5.8 %	4.2 %
100	5.5 %	4.3 %
120	4.3 %	4.3 %

B_s and TCH blocking B_t, but also, to predict the amount of traffic T after the configuration change required in prediction of Q and D.

Clearly, the performance of a single cell in terms of signal quality, blocking and dropping is improved if the size of the cell is dramatically decreased. Therefore, cell shrinking can not be performed without increasing the sizes of some other cells if introduction of coverage gaps must be avoided. Therefore, the cell shrinking and expanding must be in balance.

Now, the optimization problem including the balance requirement can be represented in form

$$\min_{\Delta \mathbf{P}} \quad \left\{ \mathbf{f}^T \Delta \mathbf{P} \right\} \tag{4.12}$$

$$w.r.t \quad -\mathbf{A}\Delta \mathbf{P} \leq \mathbf{0}, \tag{4.13}$$

$$\Delta \mathbf{P}_{min} \leq \Delta \mathbf{P} \leq \Delta \mathbf{P}_{max}, \tag{4.14}$$

where ΔP_{min} and ΔP_{max} denote the maximum allowed change, $\mathbf{A}$ is a matrix implementing the coverage balance equation, and $\mathbf{f}$ is a vector combining the effects of ΔP adjustment to call dropping, blocking and signal quality. Solving the above problem yields to the optimal configuration change minimizing the occurrences of SDCCH and TCH blocking, dropped calls and voice quality problems, while maintaining the current coverage.

This linear programming problem can be solved to find the cell size adjustments leading to the optimal overall performance. However, it may not be necessary to optimize all the cells in the network simultaneously, but instead, the focus may be in a small set of low performance cells. For these cells, the performance is optimized while causing minimal performance degradations to the other parts of the network.

The low performing cells to be optimized should have a relatively high number of failures, but also, their performance should be possible to influence through ΔP adjustments.

In Table 4.2, the obtained performance gains in the experiments in which different

number of cells were selected for optimization are presented. Clearly, the results indicate that selecting only a subset of low performance cells and maximizing their performance leads to improvement percentage of 10 % in these cells. Naturally, the boundary cells of the optimized cells suffer from performance degradation and therefore, the overall network performance improves only about 3 %. The more cells are taken into optimization, the less improvement in the optimized cells is obtained. On the other hand, the overall network performance gradually improves if more cells are taken under performance optimization.

Chapter 5

Conclusions

In this thesis, a wide range of data analysis methods were applied in order to analyze and optimize the usage of radio resources of a cellular network. It was demonstrated that a large collection network performance data is an important source of information and with the use of suitable data analysis methods, measurement data can be used to enhance performance optimization process. Especially, the potential of multivariate data analysis methods in detection and diagnosis of failures and performance degradations associated to multiple KPIs and mobile cells were emphasized.

The use of multivariate data visualization methods has several advantages in the analysis of cellular network data. Firstly, the methods are fairly simple to implement. In addition, the same methods can be used to analyze very different problem types without major changes to the used technique. Also, such approach is not sensitive with respect to algorithm and software updates made to the analyzed system and therefore they need not be updated simultaneusly with the other system.

In this thesis, approaches to include prior knowledge to the data analysis process was presented. The use of such techniques is useful in mobile communication networks in particular, since lots of useful models are available, and the measurement data is typically averaged in order to save disk space. Therefore, it may be necessary to combine these two sources of information when advanced decisions must be made.

It was also shown in this thesis, that the adjustment process can also be formalized and efficiently solved with data-driven methods. This enables the making of very difficult optimization decisions in which consequences of actions are very difficult to predict without quantitative methods. Especially, the use of past measurement data in model construction allows the possible performance gains to be evaluated prior to configuration adjustment. As a result, the operator has the opportunity to ignore the configuration adjustment if too moderate performance gains would be obtained.

It is also common that the cellular network data contains different types of anoma-

lies, making certain measurements uninformative. In the presented approach, separate preprocessing step can be easily performed prior to model estimation and performance prediction. Many of the autotuning approaches developed for performance optimization tend to require that no anomalies occur during parameter adaptation, making it more difficult to apply them in operative systems.

Implementation of such methods in a true environment is tractable, since the performance data can be analyzed offline and the predictions for the consequences of the selected actions can be based on models also estimated offline. In this case, real-time communication with the existing network elements is not necessarily needed during the optimization process.

Bibliography

Barco, R., Canete, F., Diez, L., Ferrer, R., and Wille, V. (2001). Analysis of mobile measurement-based interference matrices in GSM networks. In *Proceedings of Vehicular Technology Conference (VTC)*.

Barreto, G. d. A. and Araujo, A. F. R. (2000). Time in self-organizing maps: An overview of models. *International Journal of Computer Research*, 10(2):235–259.

Bazaraa, S. M., Sherali, D. H., and Shetty, C. M. (1993). *Nonlinear Programming: theory and algorithms*. John Wiley & Sons, Inc.

Bellman, R. (1961). On the approximation of curves by line segments using dynamic programming. *Communications of the ACM*, 4(6).

Bezdek, J. C. and Pal, N. R. (1998). Some new indexes of cluster validity. *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics*, 28(3):301–315.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford Univeristy Press, Inc.

Chandra, C., Jeanes, T., and Leung, W. (1997). Determination of optimal handover boundaries in a cellular network based on traffic distribution analysis of mobile measurement reports. In *Proceedings of 47th IEEE Vehicular Technology Conference (VTC)*.

Cherkassky, V. and Mulier, F. (1998). *Learning from Data*. John Wiley & Sons.

Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, 68(342):361–368.

Chiang, L. H., Russel, E. L., and Braatz, R. D. (2001). *Fault Detection and Diagnosis in Industrial Systems*. Springer.

Cooper, R. B. (1981). *Introduction to Queueing Theory*. Elsevier, 2nd edition.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):224–227.

Demuth, H. and Beale, M. (1998). *Neural network toolbox user's guide*. The MathWorks Inc.

Everitt, B. (1993). *Cluster Analysis*. Arnold.

Fayyad, U., Grinstein, G. G., and Wierse, A., editors (2002). *Information Visualization in Data Mining and Knowledge Discovery*. Morgan Kaufmann.

Flanagan, A. and Novosad, T. (2002a). Automatic selection of window add in WCDMA radio network based on cost function minimization. In *Proceedings of the 7th IEEE International Symposium on Spread-Spectrum Techniques and Applications*.

Flanagan, A. and Novosad, T. (2002b). WCDMA network cost function minimization for soft handover optimization with variable user load. In *Proceedings of 56th IEEE Vehicular Technology Conference (VTC)*.

Flanagan, A. and Novosad, T. (2003). Maximizing WCDMA network packet traffic performance: Multi-parameter optimization by gradient descent minimization of a cost function. In *Proceedings of the 14th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Reserach*, 3:1157–1182.

Hakkala, L. and Ylinen, R. (1978). *Johdatus moderniin systeemi- ja säätöteoriaan*. Otakustantamo.

Halonen, T., Romero, J., and Melero, J., editors (2002). *GSM, GPRS and EDGE Performance - Evolution Towards 3G/UMTS*. John Wiley & Sons.

Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*. MIT Press.

Hata, M. (1980). Empirical formula for propagation loss in land mobile radio services. *IEEE Transactions on Vehicular Technology*, VT-29(3):317 – 325.

Haykin, S. (1999). *Neural Networks: a comprehensive foundation, 2nd edition*. Prentice-Hall, Inc.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (2000). *Understanding Robust and Exploratory Data Analysis*. John Wiley & Sons.

Hollmen, J., Tresp, V., and Simula, O. (1999). A self-organizing map algorithm for clustering probabilistic models. In *Proceedings of Ninth International Conference on Artificial Neural Networks*, volume 2, pages 946–951.

Holma, H. and Toskala, A., editors (2004). *WCDMA for UMTS*. John Wiley & Sons, 3rd edition.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441,498–520.

Hyötyniemi, H. (2001). Multivariate regression: Techniques and tools. Technical Report Report 125, Helsinki University of Technology, Control Engineering Laboratory.

Hyötyniemi, H., Nissinen, A. S., and Koivo, H. N. (1997). Evolution based self-organization of structures in linear time-series modeling. In *Proceedings of 3rd Nordic Workshop on Genetic Algorithms*, pages 135–152, Helsinki, Finland.

Hämäläinen, A., Valkealahti, K., Höglund, A., and Laakso, J. (2002). Auto-tuning of service-specific requirement of received ebno in WCDMA. In *Proceedings of 56th IEEE Vehicular Technology Conference (VTC)*.

Höglund, A., Pöllänen, J., Valkealahti, K., and Laiho, J. (2003). Quality-based auto-tuning of cell uplink load level targets in WCDMA. In *Proceedings of 57th IEEE Vehicular Technology Conference (VTC)*.

Höglund, A. and Valkealahti, K. (2002). Quality-based tuning of cell downlink load target and link power maxima in WCDMA. In *Proceedings of 56th IEEE Vehicular Technology Conference (VTC)*.

Inselberg, A. and Dimsdale, B. (1990). Parallel coordinates: A tool for visualizing multidimensional geometry. In *Proceedings of the First IEEE Conference on Visualization (Visualization' 90)*, pages 361–378.

Jang, J.-S. R. (1993). ANFIS: Adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems, Man, and Cybernetics*, 23(3):665–685.

Jørgensen, B. (1993). *The Theory of Linear Models*. Chapmap & Hall, Inc.

Kaaranen, H., Ahtiainen, A., Laitinen, L., Siamäk, N., and Niemi, V., editors (2005). *UMTS Networks*. John Wiley & Sons, 2nd edition.

Kirk, D. E. (1970). *Optimal Control Theory*. Prentice Hall.

Kohonen, T. (1993). Thing's you haven't heard about self-organizing maps. In *Proceedings of International Conference on Neural Networks*, pages 1147–1156.

Kohonen, T. (2001). *Self-Organizing Maps, 3rd edition*. Springer.

Kyriazakos, S. and Karetsos, G. (2004). *Practical Radio Resource Management in Wireless Systems*. Artech House, Inc.

Laiho, J., Kylväjä, M., and Höglund, A. (2002a). Utilization of advanced analysis methods in UMTS networks. In *Proceedings of IEEE Vehicular Technology Conference (VTC)*.

Laiho, J., Raivio, K., Lehtimäki, P., Hätönen, K., and Simula, O. (2002b). Advanced analysis methods for 3G cellular networks. Technical Report A65, Helsinki University of Technology.

Laiho, J., Wacker, A., and Novosad, T., editors (2002c). *Radio Network Planning and Optimisation for UMTS*. John Wiley & Sons, Ltd, 2nd edition.

Lampinen, J. and Oja, E. (1989). Self-organizing maps for spatial and temporal AR models. In *Proceedings of Scandinavian Conference on Image Analysis (SCIA)*, pages 120–127, Helsinki, Finland.

Lee, C. C. (1990a). Fuzzy logic in control systems: fuzzy logic controller, part I. *IEEE Trans. Syst., Man, Cybern.*, 20(2):404–418.

Lee, C. C. (1990b). Fuzzy logic in control systems: fuzzy logic controller, part II. *IEEE Trans. Syst., Man, Cybern.*, 20(2):419–435.

Lempiäinen, J. and Manninen, M., editors (2003). *UMTS Radio Network Planning, Optimization and QoS Management*. Kluwer Academic Publishers.

Lewis, P. H. and Chang, Y. (1997). *Basic Control Systems Engineering*. Prentice-Hall, Inc.

Ljung, L. and Glad, T. (1994). *Modeling of Dynamic Systems*. Prentice-Hall, Inc.

Love, R. T., Beshir, K. A., Schaeffer, D., and Nikides, R. S. (1999). A pilot optimization technique for CDMA cellular systems. In *Proceedings of 50th IEEE Vehicular Technology Conference (VTC)*.

Magnusson, P. and Oom, J. (2002). An architecture for self-tuning cellular systems. *Journal of Network and System Management*, 10(2):211–223.

Magnusson, S. and Olofsson, H. (1997). Dynamic neighbor cell list planning in a micro cellular network. In *Proceedings of IEEE 6th International Conference on Universal Personal Communications (ICUPC)*, pages 223–227.

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models*. Chapmap & Hall, Inc.

Meyer, S. L. (1975). *Data Analysis for Scientists and Engineers*. John Wiley & Sons.

Milton, J. S. and Arnold, J. C. (1995). *Introduction to Probability and Statistics*. McGraw-Hill, Inc., 3rd edition.

Mitra, A. (1998). *Fundamentals of Quality Control and Improvement*. MacMillan, 2nd edition.

Multanen, M., Raivio, K., and Lehtimäki, P. (2006). Hierarchical analysis of GSM network performance data. In *Proceedings of the 14th European Symposium on Artificial Neural Networks (ESANN)*.

Nørgaard, M., Ravn, O., Poulsen, N. K., and Hansen, L. K. (2000). *Neural Networks for Modelling and Control of Dynamic Systems*. Springer-Verlag.

Olofsson, H., Magnusson, S., and Almgren, M. (1996). A concept for dynamic neighbor cell list planning in a cellular system. In *Proceedings of 6th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pages 138–142.

Oussar, Y. and Dreyfus, G. (2001). How to be a gray box: Dynamic semi-physical modeling. *Neural Networks*, 14(9):1161 – 1172.

Raivio, K., Simula, O., and Laiho, J. (2001). Neural analysis of mobile radio access network. In *Proceedings of the International Conference on Data Mining (ICDM)*.

Raivio, K., Simula, O., Laiho, J., and Lehtimäki, P. (2003). Analysis of mobile radio access network using the self-organizing map. In *Proceedings of the International Symposium on Integrated Network Management (ISINM)*.

Ricciato, F., Vacirca, F., and Karner, M. (2005). Bottleneck detection in UMTS via TCP passive monitoring: a real case. In *Proceedings of the ACM Conference on Emerging Network Experiment and Technology*, pages 211–219.

Siponen, M., Vesanto, J., Simula, O., and Vasara, P. (2001). An approach to automated interpretation of SOM. In *Advances in Self-Organizing Maps*, pages 89–94. Springer.

Spence, R. (2007). *Information Visualization: Desing for Interaction*. Prentice-Hall, Inc.

Steuer, J. and Jobmann, K. (2002). The use of mobile positioning supported traffic density measurements to assist load balancing methods based on adaptive cell sizing. In *Proceedings of IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pages 339–343.

Terzi, E. (2006). *Problems and Algorithms for Sequence Segmentations*. PhD thesis, University of Helsinki.

Torgerson, W. S. (1952). Multidimensional scaling: I. theory and method. *Psychometrika*, pages 17:401–419.

Toril, M., Pedraza, S., Ferrer, R., and Wille, V. (2002). Optimization of signal level thresholds in mobile networks. In *Proceedings of 55th IEEE Vehicular Technology Conference (VTC)*.

Toril, M., Pedraza, S., Ferrer, R., and Wille, V. (2003). Optimization of handover margins in GSM/GPRS networks. In *Proceedings of 57th IEEE Vehicular Technology Conference (VTC)*.

Ultsch, A. and Siemon, H. P. (1990). Kohonen's self-organizing feature maps for exploratory data analysis. In *Proceedings of the International Neural Network Conference (INNC)*.

Valkealahti, K. and Höglund, A. (2003). UMTS radio network multiparameter control. In *Proceedings of the 14th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*.

Valkealahti, K., Höglund, A., Parkkinen, J., and Flanagan, A. (2002a). WCDMA common pilot power control with cost function minimization. In *Proceedings of 56th IEEE Vehicular Technology Conference (VTC)*.

Valkealahti, K., Höglund, A., Parkkinen, J., and Hämäläinen, A. (2002b). WCDMA common pilot power control for load and coverage balancing. In *Proceedings of the 13th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*.

Vehviläinen, P. (2004). *Data Mining for Managing Intrinsic Quality of Service in Digital Mobile Telecommunications Networks*. PhD thesis, Tampere University of Technology.

Vehviläinen, P., Hätönen, K., and Kumpulainen, P. (2003). Data mining in quality analysis of digital mobile telecommunications network. In *Proceedings of XVII IMEKO World Congress*.

Vesanto, J. (1999). SOM-based visualization methods. *Intelligent Data Analysis*, 3(2):111–126.

Vesanto, J. and Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, 11(3):586–600.

Young, F. W. (1985). Multidimensional scaling. In *Encyclopedia of Statistical Sciences*, volume 5, pages 649–659.

Zadeh, L. A. (1965). Fuzzy sets. *Inform. Control*, 8:338–353.

Zander, J. (2001). *Radio Resource Management for Wireless Networks*. Artech House, Inc.

Zhu, H. and Buot, T. (2004). Multi-parameter optimization in WCDMA radio networks. In *Proceedings of 59th IEEE Vehicular Technology Conference (VTC)*.

Zhu, H., Buot, T., Nagaike, R., and Harmen, S. (2002). Load balancing in WCDMA systems by adjusting pilot power. In *Proceedings of 5th International Symposium on Wireless Personal Multimedia Communications (WPMC)*.

Åström, K. J. and Wittenmark, B. (1997). *Computer-Controlled Systems*. Prentice-Hall, 3rd edition.