Fault Detection in Cable Modem Networks

by

Caedmon David Austen Somers
B.Eng, University of Victoria, 1997

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of

MASTER OF APPLIED SCIENCE

in the Department of Electrical and Computer Engineering

Supervisor: Dr. N.J. Dimopoulos

ABSTRACT

Cable television networks provide analog and digital television broadcasts as well as internet connectivity to subscribers. Exposure to the elements and gradual wear reduce the performance of the analog portion of the hybrid fiber and coaxial network increasing signal noise and disruption of service. Network monitoring capabilities are minimal and network failures require the rapid deployment of a cable technician. Advanced warning of failing components allows earlier and more targeted equipment servicing, which may yield better quality of service and higher internet subscriber capacity. This thesis shows that status information returned from cable modems can be used to give a measure of network health and provide a means for fault detection. Several techniques for detecting behavior deviations of individual modems have been developed and evaluated. The topology of the network provides constraints which are used to determine the part of the network where faults may have occurred that manifests itself as a behavior deviation of particular cable modems. Regions of the network with unusual modem behaviour are shown to relate to areas with more customer service requests.

Table of Contents

List of Tables

List of Figures

Glossary

"Bad" Modem A modem with a power signal that is considered significantly abnormal.

Cable plant A cable television and cable modem distribution network served by a single head end.

CRC Cyclic redundancy check. An digital error detection scheme.

"Good" Modem A modem with a power signal that is not considered significantly abnormal.

Modem Data Status signal data collected from cable modems in cable plants, including power control feedback and CRC signals. These come in hourly samples.

Segment A subtree of the cable modem network that is fed from the same head end modem, and typically contains several SHUBs.

SHUB Secondary Hub. A subtree of a cable network that is rooted where the coaxial network stems from the optical fibre loop.

SMT Status Monitoring Transponder. These are the signal measuring devices equipped on cable trunk amplifiers that provide the SMT status data. SMTs are often referred to instead of the amplifier itself.

SMT Data Status data collected from SMTs within a cable plant, including a temperature reading. Sampled approximately once every three minutes.

Stability Data Data describing a cable plant's performance at the segment level including error levels and customer work service requests (WSRs). Data values cover approximately one month.

WSR Work service request. A request by a subscriber which initiates a service call.

Trademarks

LANcity is a trademark of Nortel Networks, Inc.

Matlab is a trademark of MathWorks, Inc.

Terayon is a trademark of Terayon Communication Systems Corporation.

Acknowledgements

I would first like to thank my supervisor, Dr. Nikitas Dimopoulos, for his guidance in this research and giving me the opportunity. I would also like to acknowledge the contribution of Dr. Stephen Neville in starting the research that is the subject of this thesis. I am thankful for his detailed explanations of relevant material and answers to my many questions. Jon Kanie was helpful in maintaining the data and running some analyses. Erik Laxdal offered assistance in many technical matters including document preparation and grad studies advice. I would also like to thank Nicos Kourounakis for his practical views, Glenn Barr for convincing me to do a Masters, André Schoorl for paving the way, Dr. Kin Li for his advice, Kier Robins for his encouragement, and my Mum for her strength and advice over these past years.

Finally, I would like to thank Rogers Cable and the Canadian Cable Labs Fund for their support of this research.

Dedication

To my parents.

Chapter 1

Background

This chapter provides an overview of cable television networks and fault detection systems. Specific to this research is the cable modem network that provides internet connectivity over cable networks. Sufficient background information is given to support the description of the cable network fault detection system presented in this thesis.

1.1 Cable Television Network Architecture

Cable television is a service available to many homes through which subscribers receive television broadcasts. The *cable network* or *cable plant* is the physical communications network in a city which distributes the broadcast signals to each home. Recently, additional services have been made available through these cable networks, such as cable internet and digital television, introducing the need for reliable digital signal transmission.

A cable network is tree structured branching from the head end out to the residences within its scope [1, 3, 23]. At different levels in the network hierarchy the transmitted signal passes through devices which reproduce and possibly broadcast it along multiple paths. The purpose of the network is to deliver the signals to the paying customers with high reliability and low noise.

Broadcast signals for the variety of channels are collected at the cable network *head end* from a variety of sources. Some come remotely from satellite feeds, some are played from recordings, while others are filmed locally and broadcast live. Signals for all channels

are combined at the head end and frequency division multiplexed, each alloted a 6 MHz band in the spectrum ranging from 54 to 550 MHz for the analog channels and 550 to 860 MHz for the digital channels [5]. The combined signal is transmitted from the head end over the cable network where the individual channels may be selected and viewed with a digital or analog TV tuner by each subscriber.

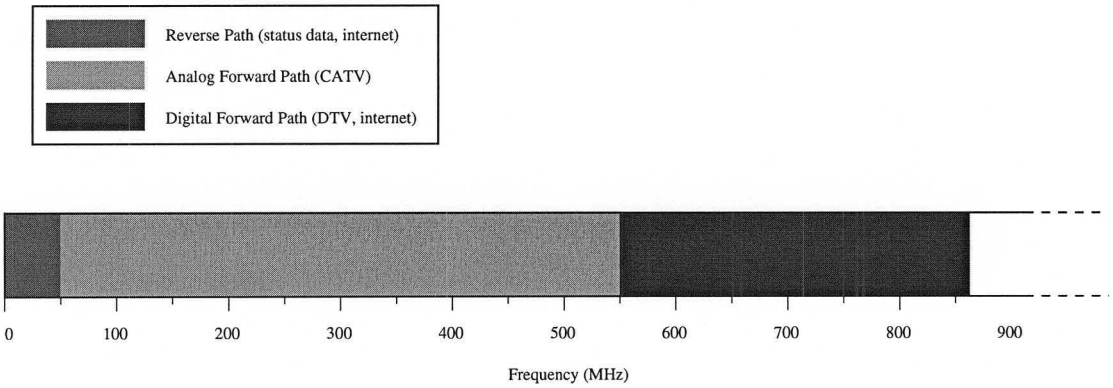

Figure 1.1. *Television Broadcast Spectrum*

From the head end, also called a primary hub (PHUB), the transmission medium is initially fiber optic cable. After this, the signals are converted to an electrical form to be transmitted over coaxial cable. The various subnetworks created by this conversion are called secondary hubs (SHUBs). These are electrically independent from one another but carry copies of the original signal. Each SHUB branches out into a tree of analog trunk amplifiers that boost the signal to make up for transmission losses. In some networks the trunk amplifiers are equipped with status monitoring transponders (SMTs) which provide status information about the amplifier operating conditions. The trunk amplifier tree can sometimes form a chain of over 20 amplifiers in cascade. The tree branches off to distribution amplifiers which are each capable of feeding a small number of subscribers. The resulting signal is fed into the subscriber residences along the coaxial cable which connects to television sets, set top boxes, and cable modems (see figure 1.2). Parts of the signal spectrum are filtered out at the tap to each residence to block services for which they did

not subscribe.

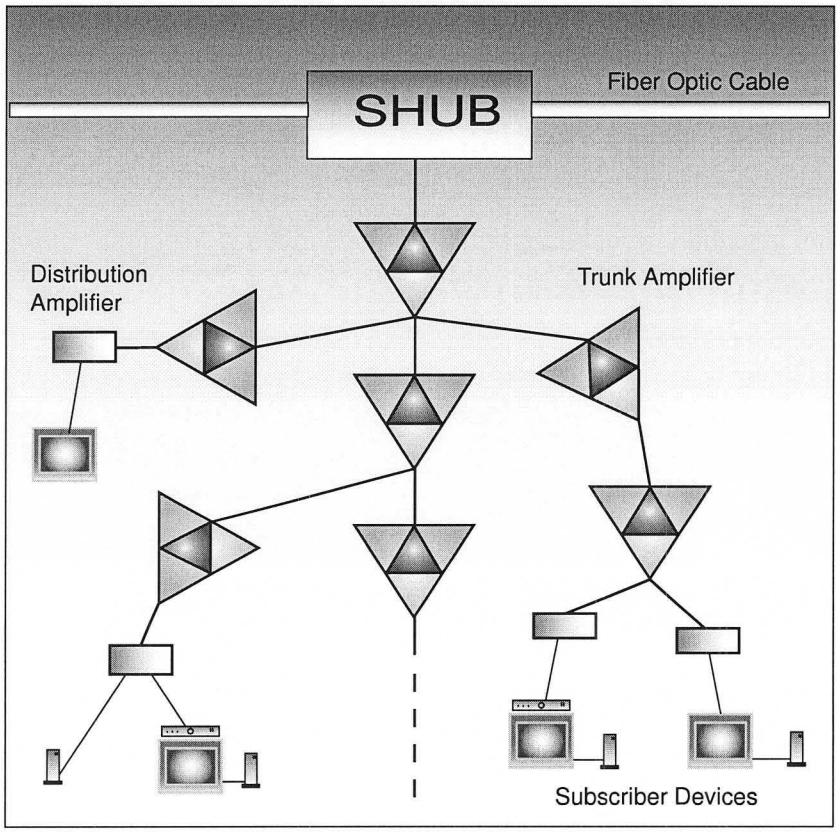

Figure 1.2. *Cable Network Structure*

The cable network is fairly static in that it is not reconstructed as individual subscribers request or cancel their service. The transmission equipment must be in place and capable of providing a service before it is available to the subscriber. From time to time the network is extended to reach a new geographic area or rebuilt to replace old hardware. The fibre optic portion of the network is occasionally extended to reach further down the tree, providing a cleaner signal and greater bandwidth as far as it spans.

Several services require a reverse transmission path to facilitate communication upstream to the head end [5]. These are the interactive services such as cable internet and digital television, as well as network status monitoring equipment. Although the cable networks were not designed for this, the transmission hardware has been augmented to

send an upstream signal in the 5-42 MHz frequency range which does not interfere with the downstream broadcast [20]. Originally unused because of the high noise level in this range, modern equipment has made it possible to utilize this part of the spectrum as well.

Cable modems modulate and demodulate digital data signals on the coaxial cable in the subscriber's home. It provides an ethernet connection to link the modem to the local area network within the home or office. The downstream traffic (from the internet to the subscriber) is broadcast on one of the traditional television channel frequency bands. The upstream traffic (from the subscriber to the internet) is sent along the upstream channel of the cable network to a head end cable modem and through it to the internet. These provide the link between the cable network and the rest of the internet. A head end modem serves several SHUBs, grouping them into *segments*.

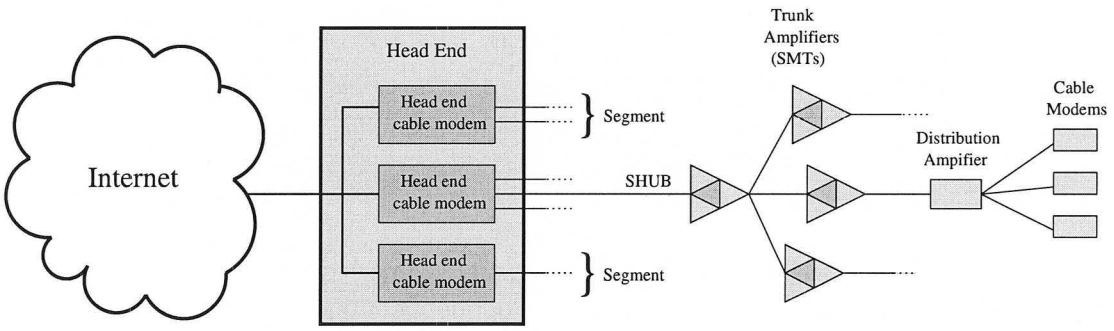

Figure 1.3. *Cable Modem Network Infrastructure: Each head end modem serves several SHUBs that form a single segment. Internet traffic is modulated onto the analog cable network and flows both downstream and upstream. From the head end, cable modem traffic is routed onto the public Internet.*

The data carrying network has additional status information to allow remote monitoring of the network transmission and stability. Status data is sent from the individual cable modems up to the head end modem where it is collected and stored for later analysis.

A typical plant contains dozens of segments, hundreds of SHUBs, several hundred trunk amplifiers, and thousands of cable modems.

1.2 Fault Detection in Large Scale Engineering Plants

Large scale engineering systems, or plants, require the correct operation of many subsystems and components for the system to function. Examples include automotive factories, the power grid, and communication systems. The failure or malfunction of a single component can have far reaching effects on the plant and the service it provides. For instance, an entire manufacturing pipeline can be stalled by the interruption of a single stage, or a downed power line can cut off thousands of residences. Plants therefore require monitoring and active maintenance. In most cases measurement and feedback of operational levels is essential to the operation of a system. Each system has its own model of what is considered normal behaviour. The combination of this model and status monitoring information provides system operators with a means to detect system malfunction.

In the worst case scenario, faults within the system become apparent only in the event of absolute failure. Yet it is possible that certain observable component behaviours are indicative of future problems, such as the gradual overheating of physical components, oscillations in systems with feedback, or the deterioration of electrical conductivity. A mechanism to detect abnormal behaviours may provide adequate warning of malfunction. What is most necessary is a means to detect behaviour outside an acceptable operational range. This raises the question of what is normal behaviour. Modelling specific faulty component behaviours is prohibitive due to the open ended spectrum of fault modalities. Identification or localization of a particular fault is less essential than mere awareness of anomalous behaviour.

A traditional approach to status monitoring is bounds checking, where levels are compared to upper and lower-bound thresholds, and alarms are generated when the operational level leaves the predefined range. This approach has several drawbacks. The thresholds must be set wide enough to account for all ranges of normal behaviour although the dynamic range may be small at any particular time. When the operational level is near a threshold, variations in the signal can generate a large number of alarms even if the level

does not wander far from the threshold itself. The behaviour of the signal within the accepted range is ignored. Variations within this range may clearly indicate faulty or abnormal behaviour, yet no alarms are generated at all.

In a complex system with many components a fault may be implied even if no subsystem is symptomatic of failure. Such a fault may only be apparent by recognizing an abnormality in the collective behaviour of multiple subsystems. For example, an above average operating temperature of one component may be quite normal, but a problem may be indicated when a number of related components are running hot. Only through a higher level analysis can such patterns be detected.

Advanced fault detection aims to give warning of pending component failure before the fault symptoms are obvious and detrimental to system operation. Adequate lead time to critical fault events gives service personnel the opportunity to identify, isolate, and rectify waning components. The result is higher system reliability, less wear on other components, and lower maintenance costs.

In plants without direct status monitoring capabilities it may still be possible to see the effects of system components through indirect measurements. Such opportunities may be wholly unanticipated and only discovered through analysis of existing status information sources.

Advanced fault detection require analysis of operational status data to reveal trends applicable to fault inference opportunities, yet each system is unique and likely requires specific analysis to reveal them. Fortunately, such analyses can be performed off-line, using data archives of status measurements. Any discoveries in the historical behaviour might then be applied in real time in an active fault detection system.

To verify any fault detection technique operational feedback is required to benchmark the accuracy of fault predictions. In offline analysis, a historical account of actual plant fault events serves this purpose.

The identification of observable behavioural trends leading to system malfunction may be approached in a number of ways. Techniques in statistics and machine learning can be

applied to relate system states to failure states. Data mining approaches can be attempted to automatically discover such relations, given appropriately structured data and sufficient computing resources. Success in any computer centric approach requires a guiding hand for appropriate direction and leads suggested by domain knowledge.

1.3 Fault Detection in Cable Networks

Cable networks are like standard engineering plants in that they require monitoring and continuous maintenance for reliable operation. Physical components age and their performance degrades over time until they fail or are replaced. Cable networks offer a particular challenge because of their size and continuous exposure to the elements. With the general lack of direct monitoring, cable operators are often made aware of network problems only when subscribers report a loss or deterioration of service. Component failure will affect cable television feeds as well as internet and digital television services. Levels of noise which only diminish analog broadcast quality can altogether prevent digital signal transmission. Timely repair is required to maintain quality of service and maintenance is costly and ongoing.

Previous efforts have shown that advanced fault detection is possible using status information provided by SMTs on the cable trunk amplifiers [19, 14] yielding increased capacity and improved reliability. Unfortunately not many networks have the SMT monitoring installed.

Modern services require a higher level of quality and reliability in the cable networks. Cable modems provide a new opportunity to monitor the state of the cable network. Status signals returned by subscriber cable modems provide a particular view of a cable network from a very large number of locations.

Cable modem signals are used to identify problems in the network around "bad" modems. This is not meant to imply that there is a problem with the modems themselves. It is the terminology adopted to refer to modems whose status signals are significantly abnormal

and possibly suggestive of network problems in vicinity of the modems.

This thesis describes an effort towards fault detection of cable networks using cable modem status information. The next chapter presents an overview of this work, including the analysis and software required to build such a system. Later chapters detail the specific tasks involved.

Chapter 2

Overview of Fault Detection Analysis

This chapter outlines the research performed for this thesis and the steps taken during its progression. It clarifies the scope of this work and distinguishes it from related research. The purpose of this work is to analyze operational status data supplied by Rogers Cable which was collected from cable modems in their cable networks, and to uncover trends that may lead to improved network fault detection.

There are several intended outcomes of this work. First, data processing algorithms and their implementation specifically targeted to cable modems. These algorithms are intended to de-noise, aggregate, store, and extract relevant data. Second, the creation of a software environment which may be used for future development of fault detection techniques fully qualified and tested. Third, algorithms that discover aberrant behaviours of the cable plants based on the status data collected from cable modems.

The above goals were arrived at through a progression of analysis tasks. Although they are outlined here in a serial fashion the actual work involved some backtracking and revisiting as goals changed and new data became available. This chapter is a summary of the major steps involved in the analysis. The chapters that follow provide specifics.

2.1 Modem Sweep Software

The software environment used for the implementation of the data processing techniques described in this thesis is Matlab^{TM}. This is a fairly high level programming platform that

is particularly well suited for data analysis and visualization. Beneath this software layer is a collection of programs used to extract the variety of data sources from different cable plants and present them in a form suitable for Matlab to interpret. The process of running a full analysis of the modem data is called a modem *sweep*. Each sweep targets one or more plants for a specified interval of time.

2.2 The Modem Sweep Fault Detection System

There are four steps in the cable modem sweep: data preparation, feature extraction, feature analysis, and fault determination. These steps correspond to the flow of data through the system from the input data sources through to the projected fault reports.

The goal of a fault detection system is to use the available data sources to detect or predict the failure of elements within the monitored system or plant. As is the case here, this must be done without the benefit of an existing model of normal plant behaviour. Discovering the normal system behaviour is part of the process of building the fault detection system. Detecting plant elements whose operation falls outside this model provides advanced fault detection in the form of generated alarms. With additional information describing actual plant faults the system can be extended to predict the kind of faults present or pending within the plant. Application of the system may reduce the impact of faulty equipment and improve plant stability as a whole.

2.2.1 Data Preparation

Data preparation takes the available data sources and prepares them for use within the system. In the modem sweep several raw data sources must be uncompressed, preprocessed, and integrated into cohesive data structures before being read into Matlab for processing in the following stages.

The primary source of data used by the modem sweep is cable modem data, collected from individual cable modems in each cable plant. It consists of several time series of

status signals from each modem, which may be missing data and have inconsistent time bases.

Chapter 3 describes the sources of data available for use in the cable modem fault detection system and the associated challenge of interpreting the data in the fault detection process.

2.2.2 Feature Extraction

Feature extraction takes the highly detailed input data sources and reduces the dimensionality into a set of features describing the most salient features of the various network elements. The choice of features and their generation changes based on the findings of the analysis so the feature extraction process is iterative. The goal of this stage is to reduce the complexity of searching for faults in the plant while maintaining those characteristics of data which allow the variety of fault behaviours to be detected.

In chapter 4 the features extracted from the various data sources are described along with the algorithms used to automate their extraction.

2.2.3 Feature Analysis

Given a set of high level features of the numerous plant elements at different hierarchical levels a model of normal plant behaviour must be formed. Patterns of features inspected in isolation and in conjunction with other features that fall outside the expected norm are searched for consistencies with known plant behaviours and faults. The analysis is open ended and gives insight to characteristic behaviours that tend to surround faulty network elements in the network structure and the time of fault events. This is a lengthy and iterative process which also involves the need for additional features generated from the feature extraction stage. Different methods of combining and summarizing feature sets are attempted with the goal of exposing clear patterns for predicting or detecting faults within the plant.

Chapter 5 describes the most useful analysis techniques used to find the patterns used

for predicting plant faults as well as the algorithms used towards this end.

2.2.4 Fault Determination

The final representation of the plant and methods for extracting the telling features along with the pattern detection mechanisms to expose plant elements which fall outside the learned plant model form the fault detection system. Chapter 5 covers how suspect regions of the cable networks are determined. The result of that analysis can be used to generate a fault report for the targeted cable plant.

Chapter 3

Data Sources

A fault detection system relies on measurement data from the plant under inspection. Archived data describing the system structure and dynamics can be analyzed off-line. In combination with knowledge of plant dynamics, these data sources may reveal trends that provide a model for normal and abnormal plant operation, and methods to perform fault detection on the plant.

Yet status data provides only a restricted view of the total system state, limiting the fault modalities within reach of detection. A fault detection system is only as good as the data accessible to it. Many plants are either not instrumented for fault detection, or the monitoring infrastructure is very rudimentary and retrofitting is expensive. Thus the only viable option is to attempt to extract as much information as possible from existing incomplete, sparse, coarse and noisy data sources.

Making the most of the data requires a concerted effort. Prior to application data sources must be analyzed and their limitations understood. Failure to properly assess the data sources could lead to mistaken conclusions about the system, including erroneous fault claims.

A number of data sources were provided by Rogers Cable for the purpose of fault detection in their cable networks. Each provides a different view of the network, contributing to potential analyses of the cable plants. The primary source of data is the modem data, which provides several status values from each modem sampled in the networks, as well as some topological information. The next most significant data source is the network stabil-

ity information. Although limited, it provides segment level view of network transmission quality. Another limited source of data used in previous cable network fault detection projects is the SMT (status monitoring transponder) data. This provides several regularly sampled status signals at the trunk amplifiers in the networks. In relation to this is the SMT network topology information which was used sparingly. Each of these data sources are described in the following sections. A summary of the data sources and the periods of time which they cover is given in section 3.6.

3.1 Modem Data

A wealth of data collected from individual modems in the cable modem network was provided by Rogers for analysis towards the goal of advanced fault detection. This data was bundled into daily files and transmitted to the university lab where it was stored and later processed.

The data comes from LANcity^{TM} modems, one of the two brands of cable modem used by internet subscribers. Although the LANcity data does not include every modem used in the Rogers networks, the majority of the cable plants and their subnetworks are represented to some extent within the data.

The data contains status and topological information. For each modem in each plant, a sequence of hourly samples containing two status signals, a modem power level and a cyclic redundancy check (CRC) level[1], are present. The topological information from each modem, although repetitious, gives a picture of the static network hierarchy relating modem to SMT, SMT to SHUB (secondary hub), and SHUB to plant. These attributes are discussed in the following sections.

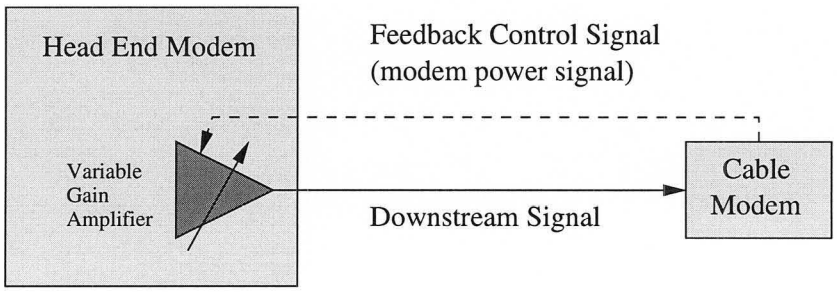

Figure 3.1. *Modem Power Feedback Signal*

3.1.1 Modem Power Signal

The modem power signal is a measurement in dBmV of the automatic gain control feedback signal the modem sends upstream to the headend (see figure 3.1). This level is intended to keep the downstream signal strength consistent. The fluctuations in this signal give a view of the impedance of the cable network from the head end to the cable modem. Given that electrical conductance varies with temperature and that the ambient temperature of the environment varies over time, it is expected that this signal will vary in accordance with the ambient temperature of the city. The power-temperature relationship is a primary feature of the modem data and it is discussed in further detail throughout this thesis. Some typical modem power signals are shown in figure 3.2.

Many atypical waveforms are observed as well, since additional factors influence the power reading. Their origin is not clear but they are important in that some observed behaviours could be indicative of plant faults. Several unusual modem power signals are shown in figure 3.3. The most common abnormal behaviours are power spikes, level shifts, and flat regions. These are described later in this section, while extraction of these signal features is discussed in section 4.1, and preprocessing of the power signal is presented in section 4.3.

To better understand the power signal a closer look at the measurement details is required. The digitally sampled signals impose several limitations on the representation of

[1]This is not the CRC value itself, but rather a measure of the CRC error rate for the modem

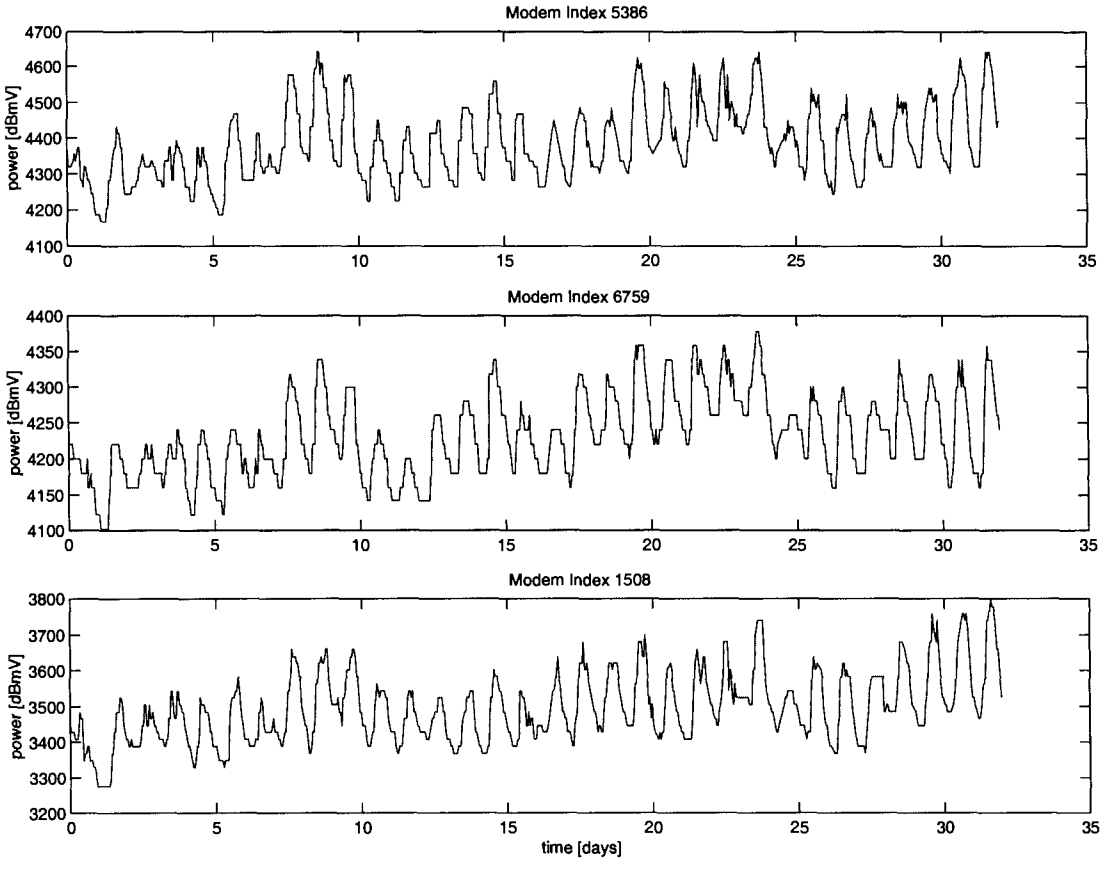

Figure 3.2. *Normal Modem Power Signals*

the underlying signal. In particular the modem data has aperiodic sample times, a bounded sampling range, and non-uniform quantization levels.

3.1.1.1 Sampling Interval

In the case of the power (and CRC) signal, the sampling period is about one hour. Application of traditional signal analysis techniques assumes this sampling frequency to be consistent. The frequency of different time intervals between successive samples for a single (randomly chosen) modem is shown in figure 3.4. The majority of sample intervals are 3600 or 3601 seconds, but there is some slight deviation from this. For this one modem's 1002 samples, the average sampling interval is 3601.1 and the standard deviation is 19.9.

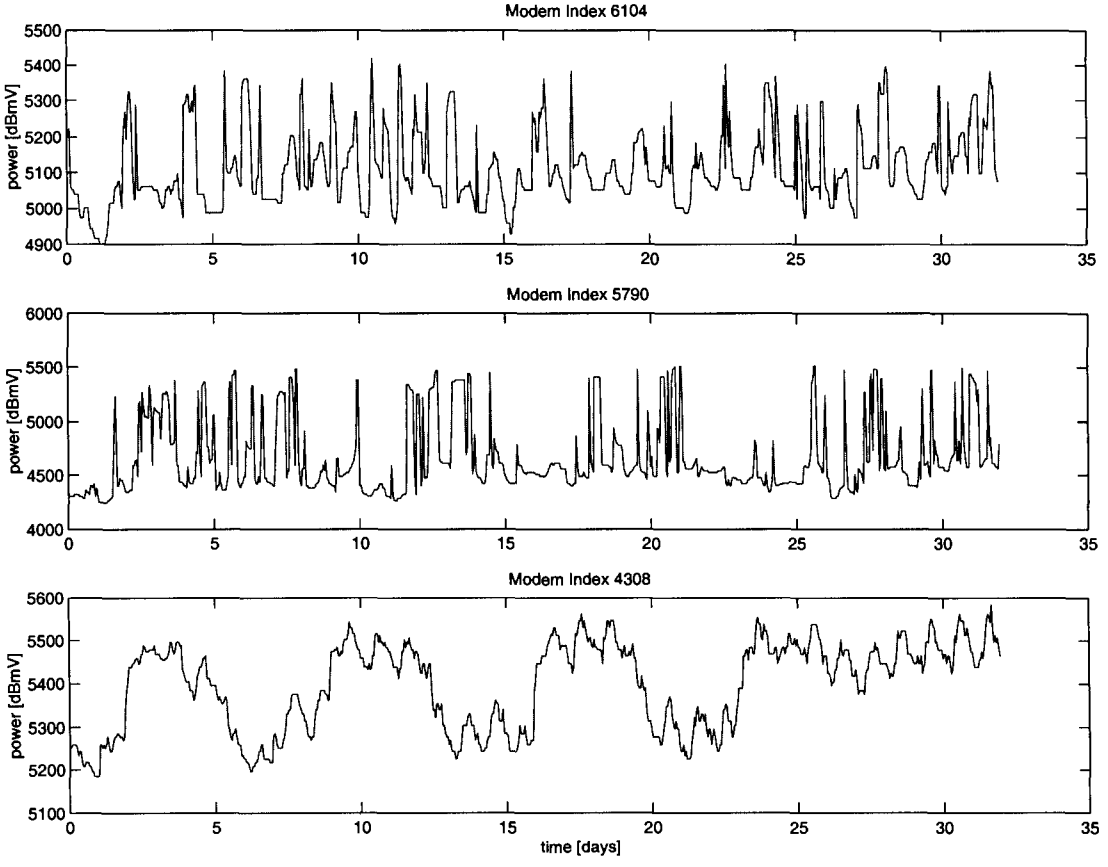

Figure 3.3. *Unusual Modem Power Signals*

Although the sampling rate is not perfectly consistent the error this introduces into the analysis of these signals should have minimal impact. The exact sample time stamps are used when appropriate and are not assumed to be one hour apart. The reason for the inconsistent sample times is not known, but it could have to do with rounding at some point of the data collection process.

3.1.1.2 Bounded Sampling Range

Sampled signals have an effective minimum and maximum measured value due to either the sensor capabilities or the encoding limitations. Measured values of modem power fall between approximately 2500 dBmV and 5600 dBmV. The histogram in figure 3.5 shows

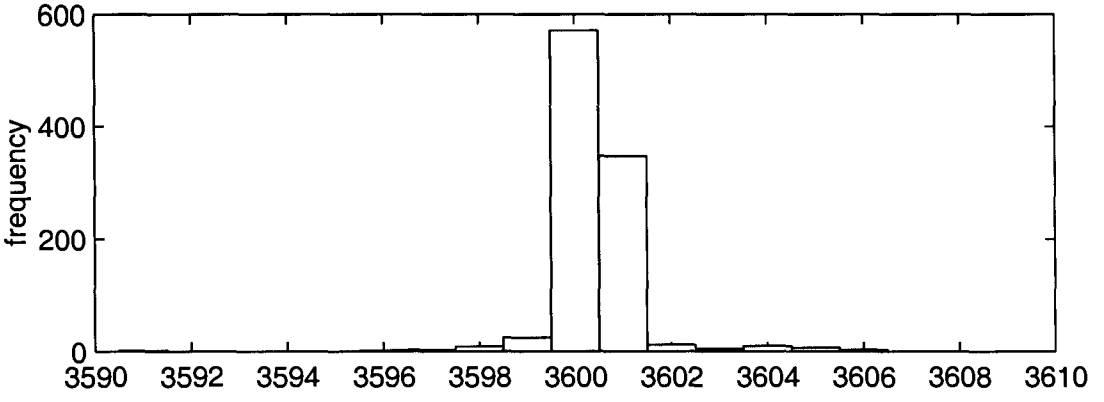

Figure 3.4. *Histogram of Sample Time Differences*

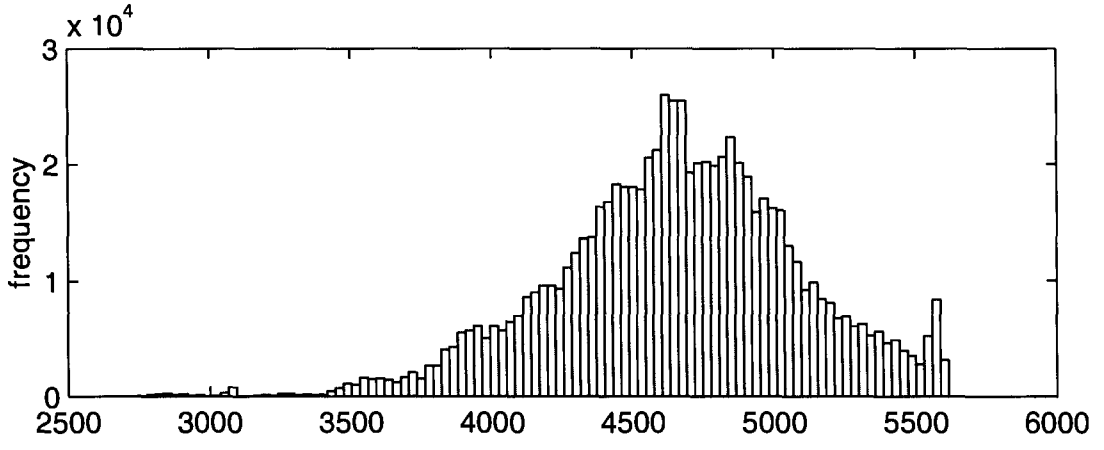

Figure 3.5. *Histogram of Power Signal Levels*

the relative frequency of different power levels for 800 randomly selected modems from two plants[2].

The distribution falls off on each side of the center but the upper range appears limited around 5600 dBmV. Inspection of individual power signals reveals what appears to be an upper limit to the sampling range. Figure 3.6 shows instances of clipped power signals. Of the 800 randomly selected modems, 59 exhibited clipping. Figure 3.7 shows the maximum power levels for all modems from one plant that have 50 or more identical maximum

[2]Three percent of the power values lie at zero but they are not shown in the histogram.

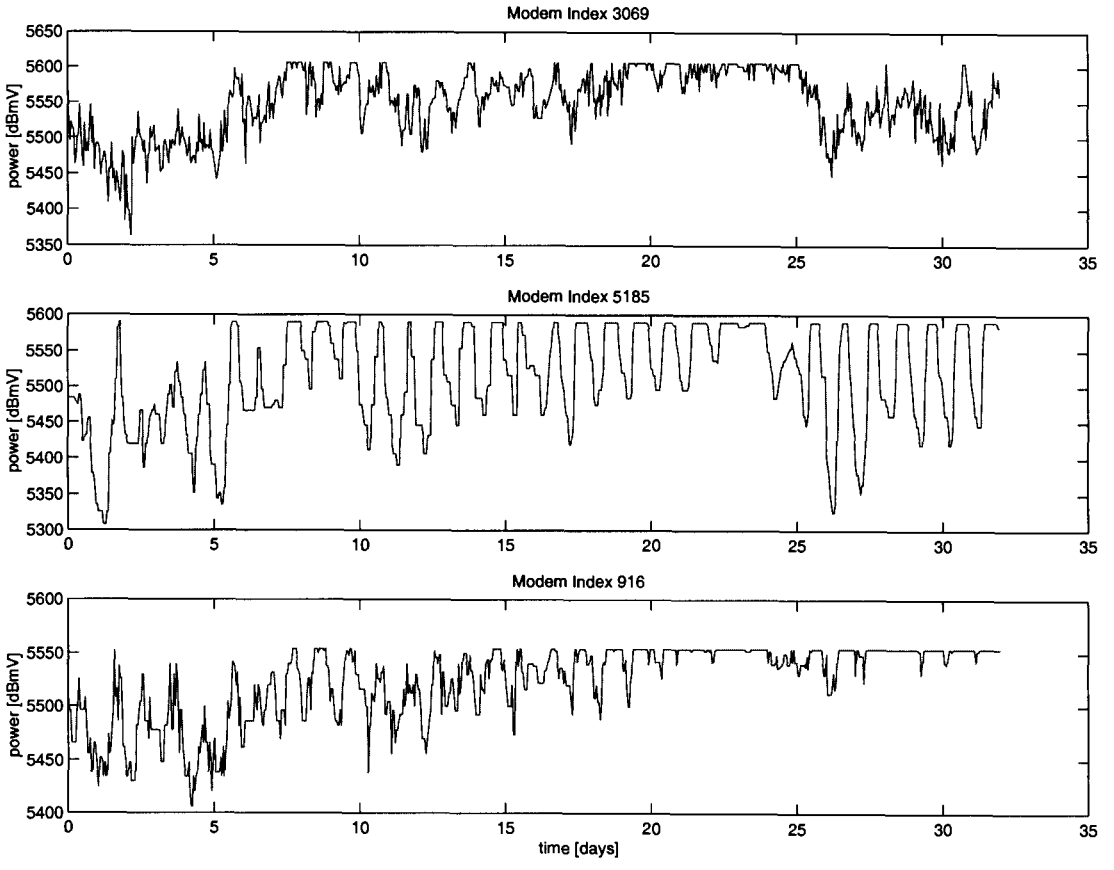

Figure 3.6. *Clipped Modem Power Signals*

samples (468 of 6811 modems). The maximum appears to vary from modem to modem but clipping consistently occurs somewhere around 5600 dBmV. The clipping is unlikely due to natural causes and is assumed to be a limitation of the data collection process. The consequence is an additional source of opacity to the real network and the introduction of misleading signal artifacts. Filtering of invalid data is discussed in section 4.3 and clipping in particular is treated in section 4.3.3.

The lower range of the power signal does not appear to be prematurely clipped.

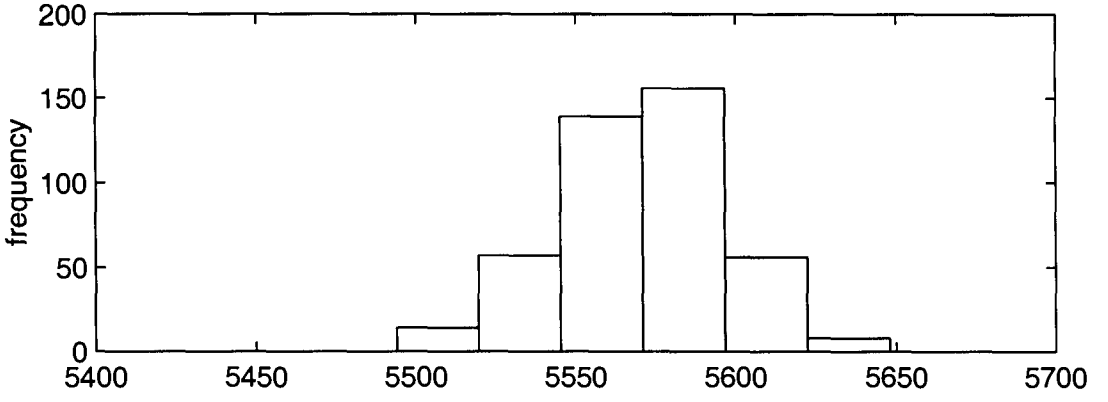

Figure 3.7. *Histogram of Maximum Power Signal Levels*

3.1.1.3 Quantization Levels

Digital signals are limited in the accuracy of their amplitude measurement. The modem power levels are provided in the data with integer dBmV values which imposes a lower limit on the effective quantization step size (1dBmV). However, only a subset of the possible dBmV values within the dynamic range are observed for any modem. The modem power levels appear to be sampled into a set of fixed quantization levels, and those power levels are not multiples of a single quantization step size. In addition, different modems have different quantization levels.

The two histograms in figure 3.8 show the number of modem power samples at each power level for two different modems from the same plant[3]. Quantization beyond the single dBmV accuracy is apparent given the obvious gaps between spikes in the histogram. The signals had plenty of opportunity to occupy the intermediate levels given the high repetition at the observed levels. The spikes appear at different dBmV values for the two modems, and as figure 3.9 shows, the difference between adjacent quantization levels is not constant.

These factors are perhaps caused by different representation accuracies at the various stages of the sampling, transmission, and data archiving process. Despite these measure-

[3]The dynamic ranges of these two modem signals are unusually small, exasperating the quantization effect.

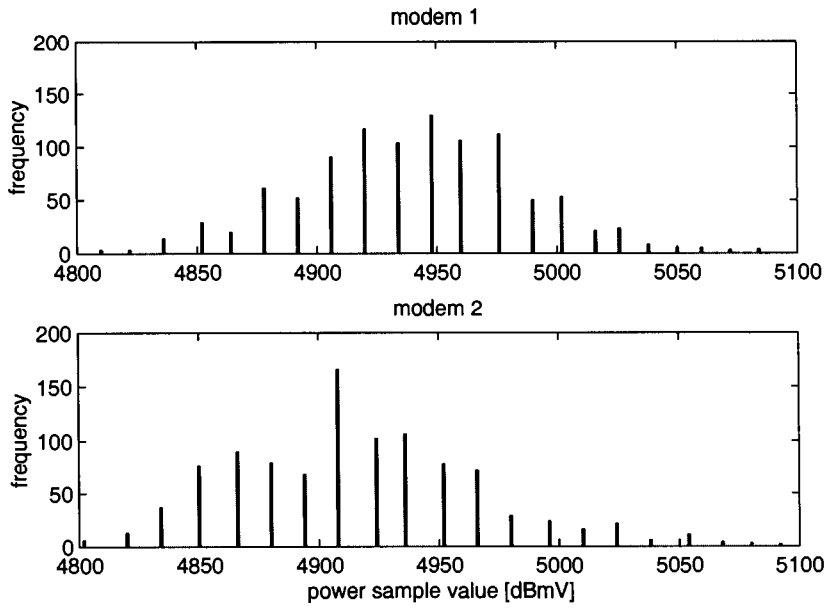

Figure 3.8. *Power Level Histograms for Two Modems*

ment peculiarities, the discretization of measurement is small in contrast to the dynamic range of the signals and the error introduced into the analysis is assumed to have little effect on the results.

3.1.1.4 Power Spikes

A commonly observed power signal event is a power spike. These are situations where the power level quickly rises and falls. Several examples are shown in figure 3.10. The origin and impact of power spikes is unknown.

3.1.1.5 Level Shifts

Inspection of modem power signals reveals a common event called a level shift. At a level shift the power signal appears to shoot up or down by a large amount and then continue on as it was prior to the shift but with a DC offset, as seen in figure 3.11. Natural power variation is very unlikely the cause of these events. They are possibly the influence of a

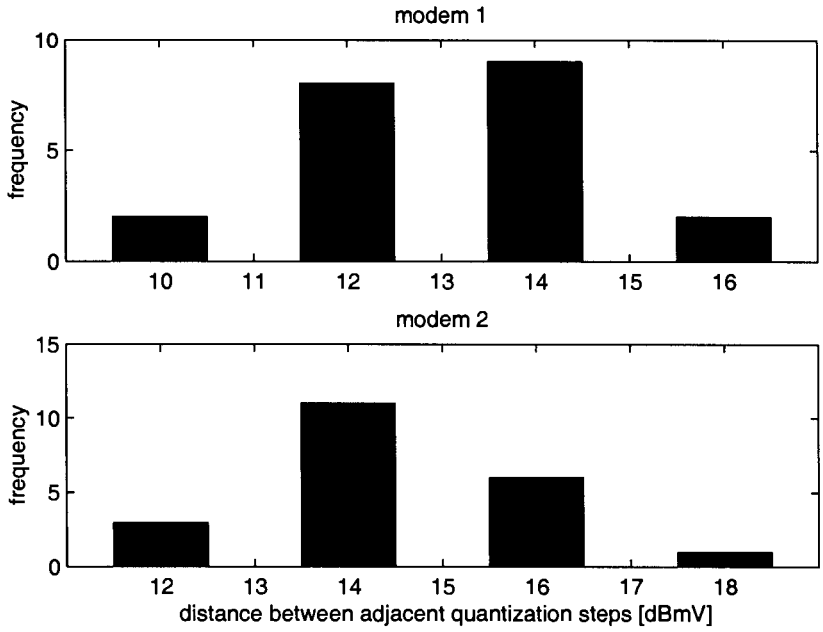

Figure 3.9. *Quantization Level Differences for Two Modems*

control system with very coarse feedback control or perhaps a data collection artifact.

3.1.1.6 Flat Regions

Flat regions are regions in signals that remain at a fixed value for many consecutive samples. They (flat regions) are a common signal artifact and often last for hours or days. The sampling is sensitive enough compared to the typical variation of a modem power signal that each sample should be different than the last. Flat regions must therefore be caused by something other than a truly constant power signal. It is uncertain at which point in the data collection process the flat regions arise, whether it is at the measurement sensors or a problem with the data collection system.

If the cause of flat regions is not within the cable network itself, it introduces an additional source of network opacity making the goal of fault detection more challenging. Any real signal fluctuations within the flat region interval cannot be seen by the fault detection system. It is not assumed that the absence of observed events during these periods implies

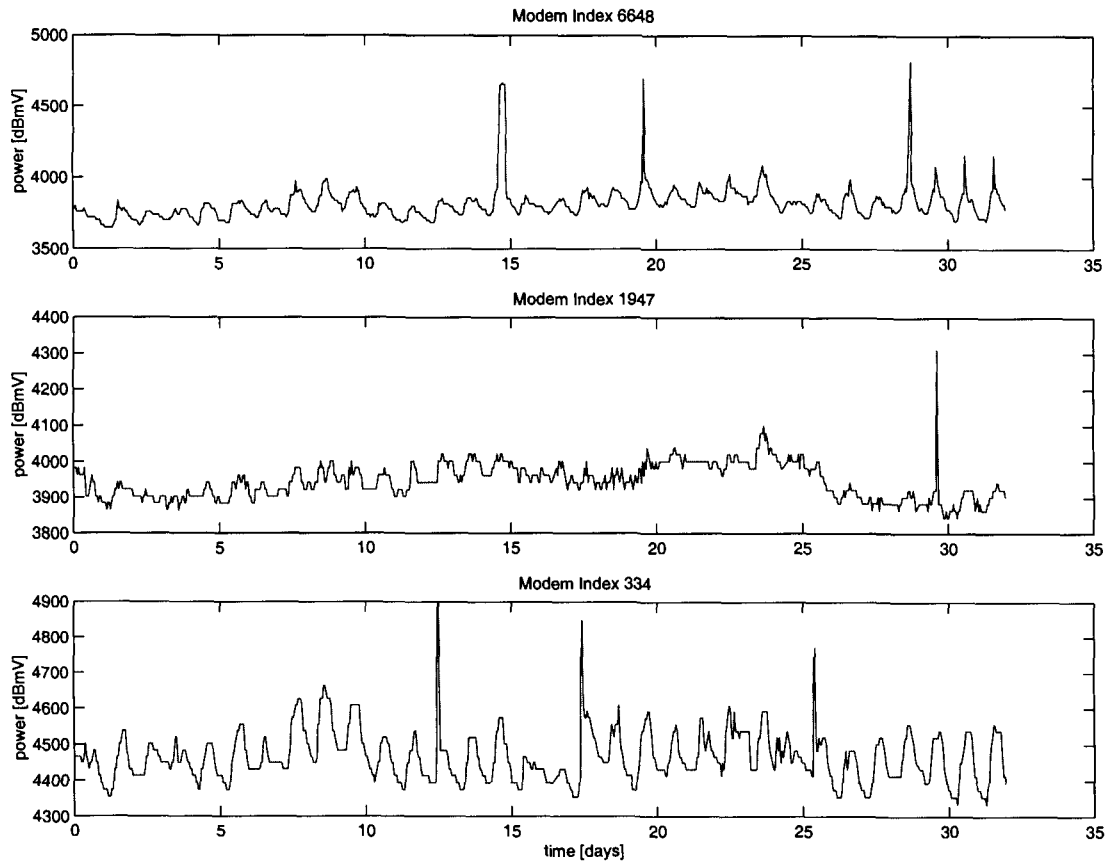

Figure 3.10. *Modem Power Signals with Power Spikes*

the absence of real events. Signals with flat regions instead are considered only partially present and leave open the possibility of otherwise observable signal artifacts during these times.

If the cause of the flat regions is from the cable network, the presence of flat regions in modem signals may be an indication of some kind of network fault. The identification of these regions serves to both improve interpretation of the data and provides a quantifiable feature of the data. Flat region identification is detailed in section 4.3.1.

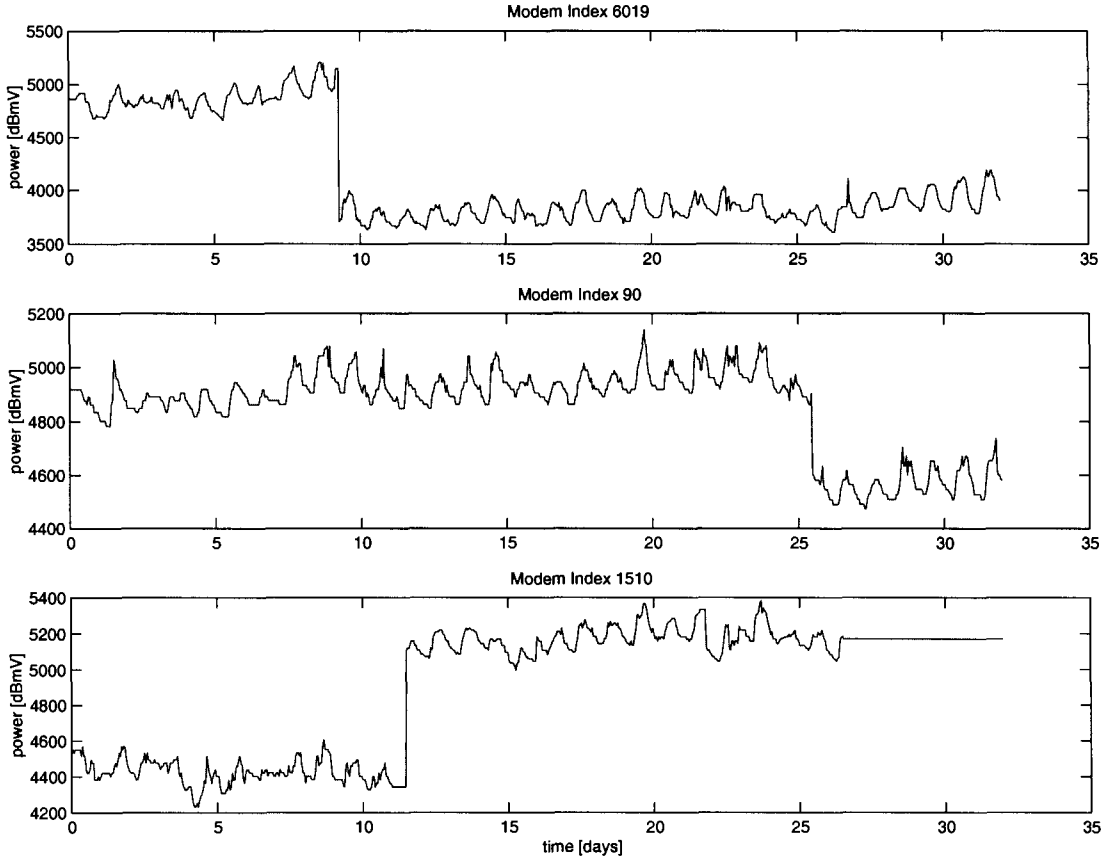

Figure 3.11. *Modem Power Signals with Level Shifts*

3.1.1.7 Zero Levels

Zero levels are an instance of one or more consecutive power samples at zero. A power level of zero is not considered valid because a dBmV value of zero implies an infinitely negative voltage. These events are thought to represent a data collection issue of some kind, such as failure to communicate with the modem. Section 4.3.2 describes the identification of zero levels.

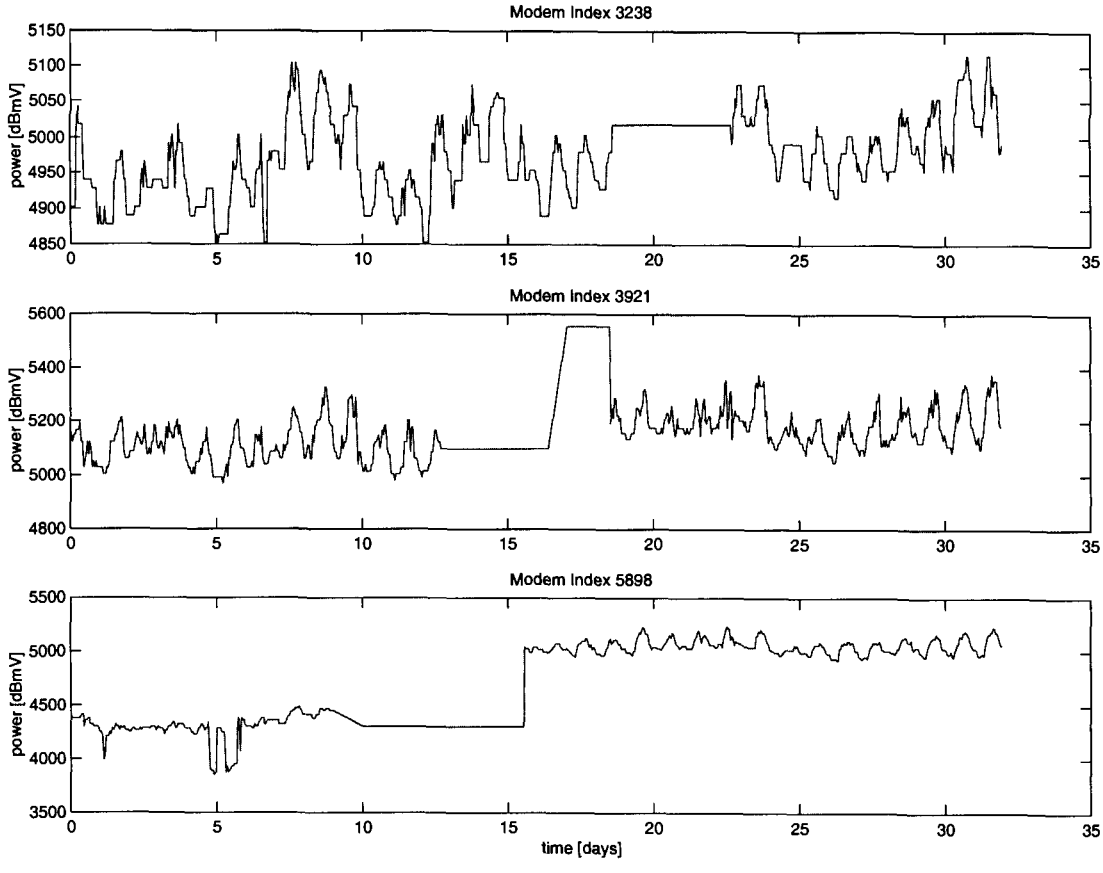

Figure 3.12. *Modem Power Signals with Flat Levels*

3.1.1.8 Time Gaps

Since each modem is sampled approximately every hour a gap of more than an hour between samples indicates that some number of samples are missing. Even one missing sample produces a blind period of two hours in which significant network events could transpire. The absence of this one sample may itself be an indication of a network problem thus time gaps of any duration are recorded as modem events. Time gaps are easily detected by applying a threshold to the difference between normalized timestamps of consecutive modem samples.

Despite the promise of fault detection using time gaps these events are very rare. One

would expect individual modems or related groups of modems to come in and out of contact with the sampling system as parts of the network malfunction or are disconnected.

One obvious source of the observed time gaps is missing modem data. This is a related issue and is discussed in section 3.5.1. These time gaps are plant wide and often last for days. Although time gaps of this origin can be anticipated from the available modem data, it is important within the sweep to keep a record of these gaps and make it apparent to the sweep operator that faults within these periods cannot be explicitly detected.

3.1.2 Modem CRC Signal

The modem CRC (cyclical redundancy check) signal is a measurement of the quality of the digital data transmission between the head end modem and the individual subscriber modems. The specifics of this measure were not disclosed, but it is understood that it reflects the amount of noise in the transmission medium. Noise interferes with the signal, causing the receiver to misinterpret the digital signal bits. To detect this problem, a digital communication system can transmit checksums along with the payload data, and the receiving end compares the transmitted checksum with a locally computed checksum on the payload. A mismatch indicates a corruption of the signal, and the frequency of checksum mismatches suggests the overall level of noise over that communications channel. The time varying frequency of CRC errors is recorded for each cable modem in the modem CRC signal. The units of the CRC measurement are not known, however, it is their relative values that are important. Several modem CRC signals are shown in figure 3.13.

3.1.2.1 Sampling Interval

The modem power and CRC samples are given within the same sample line in the modem data and thus the sampling interval for the CRC signal is the same as the power signal discussed in section 3.1.1.1.

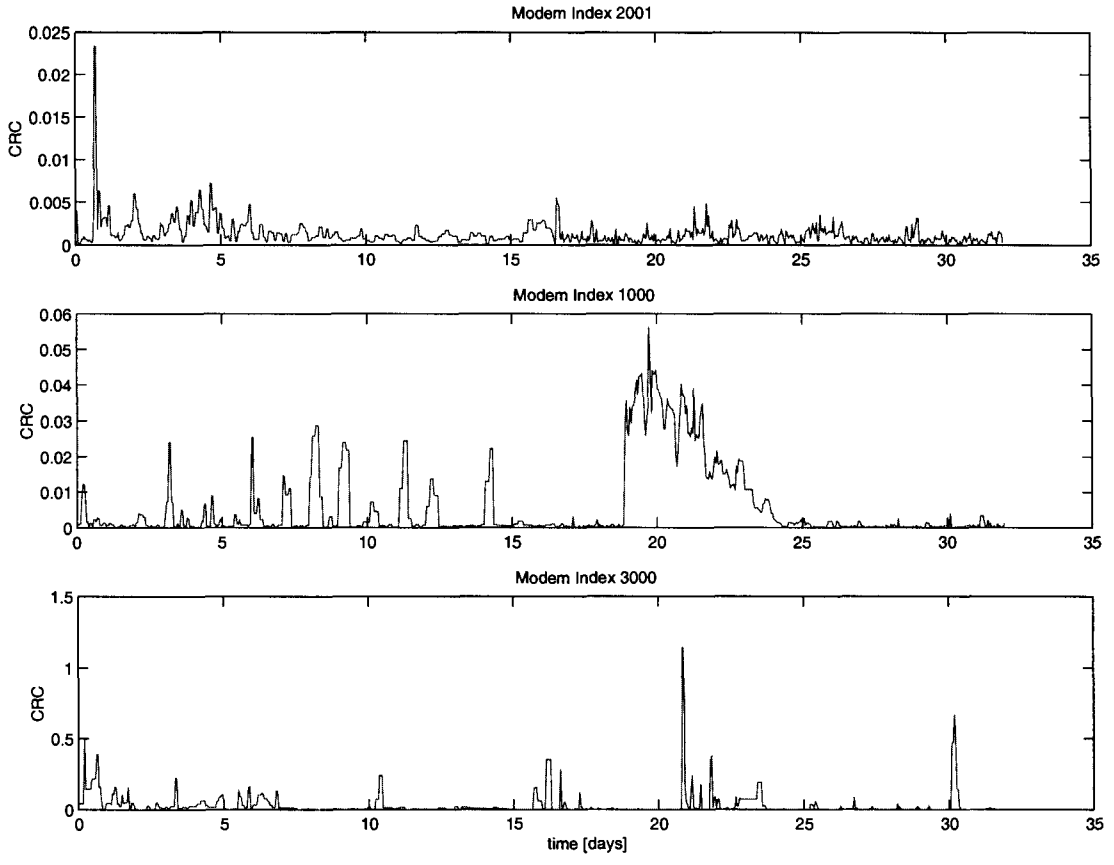

Figure 3.13. *Modem CRC Signals*

3.1.2.2 Sampling Range

A histogram of CRC values from 800 random modems in two plants is shown in figure 3.14. The minimum observed CRC level is zero, and the peak is just slightly above zero. There is no apparent clipping in the CRC signal as there is in the power signal. The upper tail of the distribution decays to zero, although values as high as 22 are observed. 6.4% of the CRC samples are not shown (they are above 0.014), and the mean CRC value is 0.017.

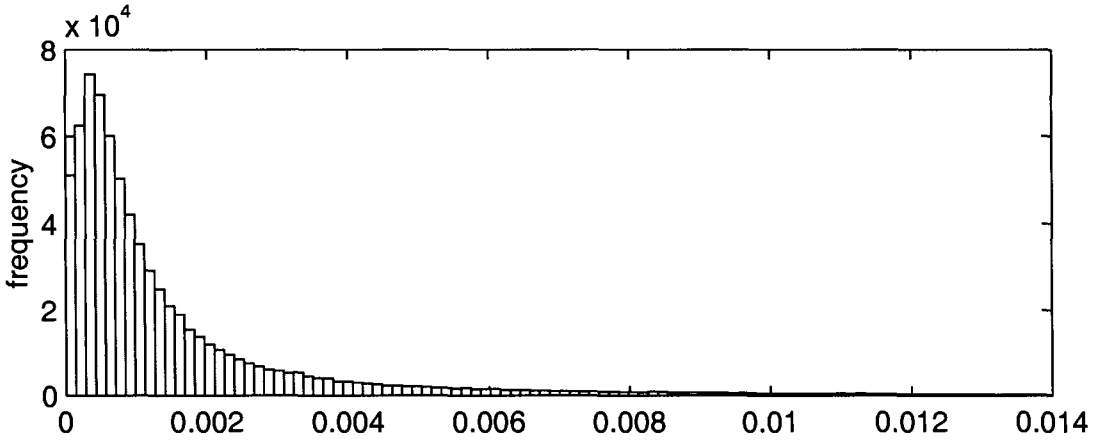

Figure 3.14. *CRC Level Histogram*

3.1.2.3 Quantization Levels

The modem CRC signals in the data are given with four decimal places of accuracy, and unlike the power signal, they appear to take on any representable value. Figure 3.15 shows CRC level histograms for two modems. The effective quantization is the same as the data representation of 10^{-4}.

3.1.3 Modem Data Topology

Each modem sample includes information on the location of the modem in the cable network. In particular the modem's SMT, SHUB, and plant are given. The location of the modem in the network is important to the fault detection system because it allows observations from the modem signals to be traced back to particular regions of the cable network. This information collectively reveals topological mapping from modem to SMT, SMT to SHUB, and SHUB to plant.

The tree structure of the network is not completely specified from this data however, since the parent/child relationships of the SMT trunk amplifiers is not inferable, the segments are not given, and neither are the distribution amplifiers. The topological picture is thus seen as a collection of subnetworks within subnetworks of modems without any

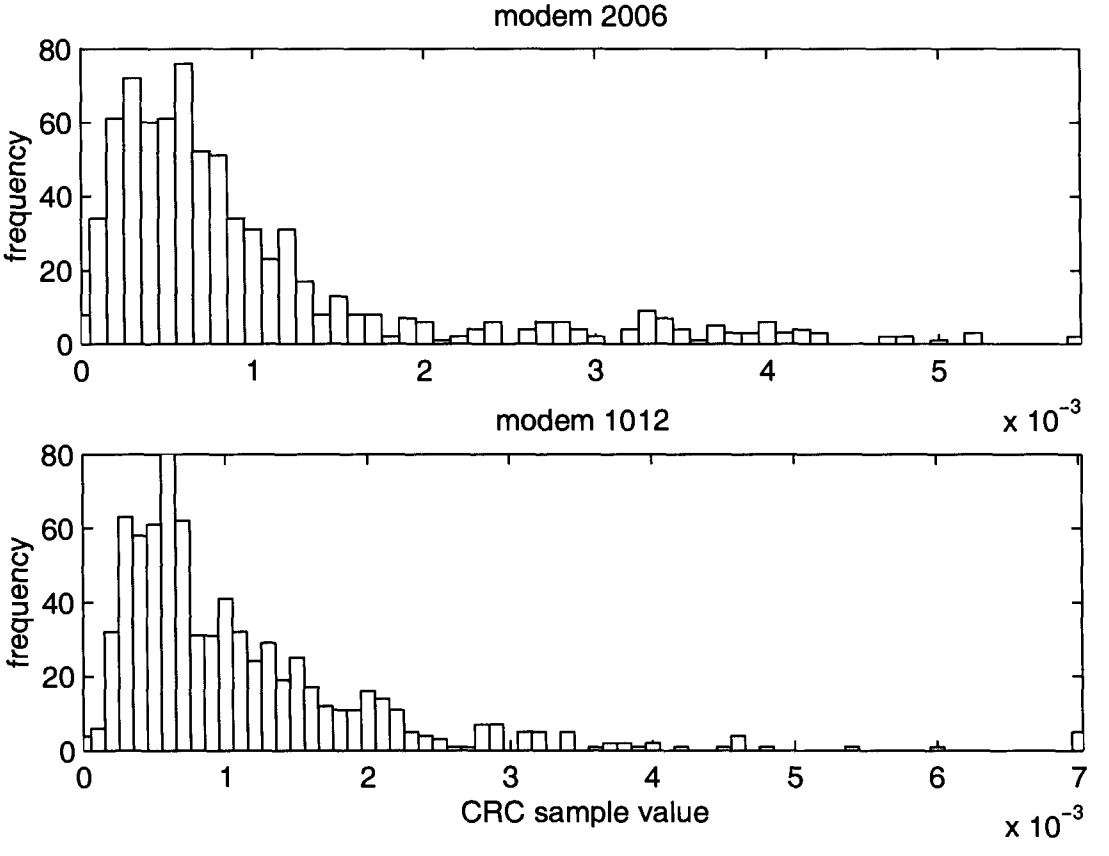

Figure 3.15. *CRC Quantization Level Histogram*

knowledge of structure within the subnetworks (see figure 3.16).

Although the topological picture is not complete there is potential for fault attribution at the modem, SMT, SHUB, segment, and plant level.

3.2 Segment Stability Reports

To assist in the identification of troubled regions in the cable networks, Rogers provided a set of spreadsheets containing stability information for each plant at the segment level. The stability information is given as a set of data transmission metrics over several time scales as measured by the head end cable modem. This data was a one time offering covering all

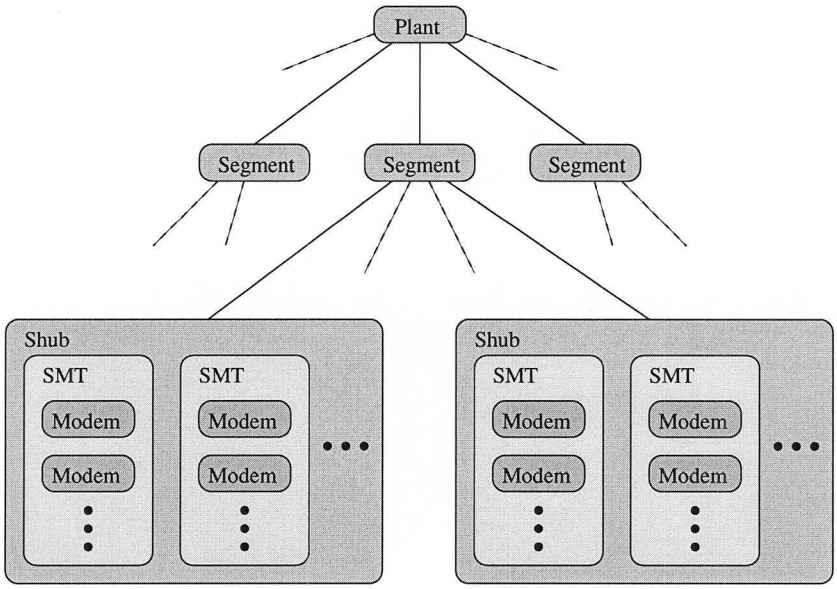

Figure 3.16. *Effective Cable Network Topological View Given in Modem Data*

plants for the one month period ending April 7 2002.

The stability data is important because it gives a top down view of the observed plant performance. This contrasts with the bottom up view provided by the modem signals. Taken as a measure of true plant performance, the stability data is used to relate observed plant behaviour through modem signals to observed plant transmission quality. Trends and correlations observed between these two data sets serve as a basis for automated fault detection based on the modem data alone. This also serves as an alternative to confirmation of suspected network difficulties through Rogers. A major advantage is that it is possible to automate the search for inferential fault detection patterns between the modem and stability data.

Modem data was available for the same time period which the stability data covers, making a comparative analysis possible.

There are two spreadsheets, one for LANcity cable modems and one for Terayon^{TM} cable modems. The metrics and time frames differ slightly between the two sets of data. Separate data records are available for each head end modem which serves a segment con-

taining one or more SHUBs.

For each metric each segment is given a percentage of error free days over each time scale. The SHUBs within the segment are also listed, giving a topological picture of the segment to SHUB mapping to complement the topological information provided by the modem data, although this topology differed slightly from that in the modem data. The stability metrics and timescales in the two spreadsheets are summarized in table 3.1.

Table 3.1. *Terayon Segment Stability Fields*

Feature	Description
IP Address	IP Address of head end modem
# Customers	Number of cable modems served in segment
SHUBs	List of SHUBs served by head end modem
# WSR	Work Service Requests - client requests for service
UCS (14,10,5,1 week)	Upstream Channel Status - % of error free days (4 different values)
SNR (14,10,5,1 week)	Signal to Noise Ratio - % of error free days (4 different values)
Combined UCS & SNR (14,10,5,1 week)	Error free days for both UCS and SNR

The meaning of the individual stability metrics was not disclosed but it can be inferred that they represent aspects of the quality of digital data transmission over the corresponding segment.

The number of modems in each segment compares with the numbers provided by the cable modem data quite well but not perfectly. The discrepancy might be due to the differing times the modems were tallied.

The UCS (upstream channel status) and SNR (signal to noise ratio) values are metrics tabulated by the head end modem that serves the segment. An *error free day* is probably a day where the measured values do not exceed a certain threshold that constitutes an error. The percentage of error free days are given over four time periods of different duration. The 14 week score is for the 14 weeks ending April 7th, the 10 week score is for the 10 weeks ending April 7th, etc. The combined stability measure is always at least as low as the lowest other measure for its corresponding time frame. This figure represents the proportion of days that are free of any kind of UCS or SNR error. These two measures are

mentioned in [17] (a classified document), however their exact nature is not critical to the analysis.

The CRC field is likely related to the modem CRC signal, only this measure is taken from the head end modem and reflects the entire segment.

Of particular interest is the *work service request* (WSR) measure. This counts the number of truck rolls directed at a segment for the one month period ending April 7th 2002. The WSR count for a particular segment reflects both the size and the quality of service in that area. Unlike the other features, the WSR factors strongly on human behaviour, since it is a reflection of the customer requested service activity in the region of the network.

It should be noted that the measurement values at the segment level do not imply a consistent behaviour over the entire scope of the measurement. It is quite possible that a poorly behaving segment is recorded as such due to just one SHUB or a few SMTs which bring down the stability metric for the entire segment. The particular subnetwork causing the poor stability cannot be seen from this level but it is hoped that the problematic region of the network will be apparent from the modem data signals.

The LANcity stability spreadsheet contains similar features including number of customers and WSRs, but instead of SNR and UCS it has fields for CRC, L2, and BS. It is expected that these fields serve purposes similar to the Terayon fields. In the subsequent analysis, the LANcity WSR is the primary stability metric used.

3.3 SMT Data

Another source of data used was the SMT data. This data was collected from cable trunk amplifier SMTs over a period of many years and was the basis for earlier cable network fault detection analyses [12, 14, 15, 16, 19]. The SMT data was archived in the UVic lab but transmission of new data was discontinued as of 2001. It was used initially in the cable modem analysis and provided several useful insights from the time frame when both SMT and modem data were available.

The SMT data consists of a stream of data samples for each amplifier in a cable network. The sampling period varies between plants but is typically around the 3 minute range. There are many sampled fields but most utilized were the forward pilot and the temperature signals [6, 16]. The reverse pilot and current signals were the basis for another analysis [12].

The SMT temperature signal provides significant information to the analysis of the cable modem power signals as plant wide temperature estimates. Appendix A details the derivation of the plant-wide temperature estimation.

3.4 SMT Topology

From previous cable network fault detection efforts [reference], a source of network SMT topology was used. This topology defines the cable trunk amplifier tree structure from the headend downwards, which the modem data does not do. Unfortunately this data source is largely incomplete and out of date. The provided topology data is absent for most plants. For those that had data present, the topology given represented the network at one specific time. This view becomes invalid over time as the cable network changes with the addition of new cable amplifiers or the replacement of trunk amplifiers with fibre optic nodes. Thus, at present, SMT topology is thus mostly inapplicable, although in the past it was used extensively for fault cluster inferences [15].

3.5 Data Issues

The data sources which make fault detection possible must be interpreted carefully. Data is prone to error and misinterpretation leading to invalid conclusions unless the limitations of the data sources are understood and accounted for. This is particularly challenging because the data collection system is itself prone to faulty operation. In a real world system the data sources are like random processes, potentially including any range of valid or invalid

sequences. A fault detection system should have robustness and minimize the production of misleading results when exposed to imperfect data. Towards this goal, the analysis should include an examination of the kinds of data defects that are present or possible, and attempts should be made to minimize their influence. A number of observed limitations in the data sources and their implications are discussed in the following sections. The significance of imperfect data sources should not be underestimated. It is natural to overlook these issues because many projects operate on synthetic or very reliable data. In the case of real world data analysis however, these issues should not be left unchecked.

3.5.1 Missing Data

The most obvious and prevalent data limitation is the lack of completeness. There are a number of reasons why the various data sources are incomplete.

3.5.1.1 Causes for Missing Data

Although the archiving process in the UVic lab was very reliable, there were frequent occasions when the daily transmission of data never arrived. Thus the data has many days which are entirely missing. The reason for this is unknown. It was likely an operational error closer to the data collection source.

Even when the expected daily transfer was conducted, data within the files transmitted was often incomplete. The reason for this is also unknown. Interestingly, the periods of missing data tended to come from plants that were geographically related, such as all the plants in Toronto. This implies that something further up the data collection chain was at fault.

The result of these two problems is the presence of many holes in the modem and SMT data. A chart of modem data availability is given in figure 3.17.

A significant reason for missing data was the exchange of several cable plants between Rogers and Shaw cable companies in November 2000. This altogether terminated the re-

ceipt of modem data from the BC area plants and introduced several plants in the Ontario area. From the data point of view there is missing data after that date for the BC plants and missing data before that data for the Ontario plants. Although this is accommodated by shifting focus between plants, it hampers the analysis because follow up or historical analysis cannot be conducted on these plants.

Another reason why data is unavailable, for specific modems or subnetworks in a plant, is because of gradual topological change. Internet subscribers come and go, thus the status information from their modem will only be available during their subscription period.

Perhaps the most useful source of missing data is equipment malfunction that prevents the data collection from within the scope of the failing elements. These occurrences, if they can be identified, may give valuable leads towards faulty behaviour.

3.5.1.2 Dealing with Missing Data

Regardless of the cause, missing data must be handled appropriately. A major problem with missing data is that it makes inferences about the behaviour of affected signals less reliable compared to signals whose data was available in its entirety. For example, suppose only one day of data is present for a particular modem in a month long sweep. Modem CRC signals vary widely over time and if the one available day of data for a particular modem happens have high CRC levels cannot be assumed that this level is representative of the entire month. Yet the CRC mean could easily be interpreted as such because it hides the quantity of data considered. More specifically, for a given probability distribution, the mean of a small set of samples is more likely to stray from the distribution mean than the mean of a large set of samples. Consequently, within a plant, those modems with the highest mean CRC levels are often simply at the extreme because they have less available data. This problem has many different faces and must be considered during analysis to avoid making invalid conclusions.

Another key consideration is that the absence of data does not imply the absence of faulty behaviour despite the fact that none was detected. The best way to deal with this

issue is to make the absence of data apparent in any results presented. For example, if a badly performing SHUB has no problems for a week, it must be reported if there was no data for that week so it is not assumed that the behaviour was normal during that time.

In temporal analysis, a missing chunk of data will usually introduce a large jump in the signal level between adjacent samples, which is not an actual sharp signal change. Such occurrences might trigger false events so events should be neutralized by time gap events if they occur concurrently.

On the visualization side, missing data in signal plots will either produce a sharp discontinuity or a large gap. This makes it difficult to interpret the signal visually, especially while scanning across the signal, as it is very distracting to the eye.

In this thesis, the focus is to use the available and valid data to attempt to extract information on the health of the plant from their behaviour.

3.5.2 Event Encoding

In some cases signal levels are not meant to be interpreted at face value. Information may be embedded into the data stream to signify special events or status. In the modem data, for example, zeros are found in the modem power signals. These conditions do not represent real power levels because the dB scale does not reach zero. The zeros are present in both the power and CRC signals at the same time. These encodings are risky because without due attention they will be interpreted normally. Not only are they invalid signal levels but they are usually at extreme values of the signal range. If left in the data stream they might severely influence the results of any statistic of, or any algorithm applied to that signal. These encoded values should at least be filtered from the signals before automatic analysis is conducted. Of course, the encoded information can be used advantageously by utilizing their intended meaning.

In practice, the presence of encoded event values within data streams may not be anticipated, especially if no formal specification for the data format is given, as is the case with the modem data. Over time, new events may be added to the system or very rare events

may be encountered. This is one of the lessons learned early when dealing with large, real world systems. Too often, the only way these events are discovered is when the analysis system produces bizarre results or fails altogether. The code is specifically adjusted to handle these occurrences. One preventative measure to deal with these events is to filter signal values that are considered outliers in the distribution of normally observed values. This eliminates the events encoded with extreme values without affecting the underlying signal. Statistics of the proportion of signals filtered in this way are collected so that unusally high concentrations of eliminated samples are not passed unnoticed.

3.5.3 Inconsistent Topological Information

In real systems, where data is collected from a variety of independent systems, it is possible that the data sources do not entirely agree with one another. In the various sources of cable network data the topological information is often inconsistent, and this ambiguity must be resolved because many analyses make the assumption of topological consistency.

3.5.3.1 Sources of Topological Inconsistency

There are a number of possible reasons for the discrepancy. They mostly arise because the actual structure of the network occasionally changes. The different data sources used in an analysis may have slightly different time frames or topological changes may be updated in different data sources at different times. It is especially complicated to track topological changes within a single analysis time frame, thus topological information is usually taken from a snapshot of the network at a specific time. Another issue occurs often when topological information is maintained by hand. This is often done because the topology is known to change infrequently and automating the topology generation may be difficult. These conditions are subject to human error and the provided topological information may be incorrect and even inconsistent with itself.

Incorrect structural information, such as topology, leads to a number of problems. The

most apparent problem is that analyses using the topological model will produce erroneous results. For example, a poorly behaving network element may imply a network issue in an area where there is none because it is incorrectly specified in the topology. A misrepresented network element will influence statistics calculated on that region of the network. A less likely but more severe problem is a structural loop implied by a circular topological specification. In some cases this might cause the analysis to loop infinitely while searching what is assumed to be a tree structure. It is easier to remove loops prior to analysis than to have each algorithm detect them, and this is the approach taken as described in the following section.

3.5.3.2 Dealing With Topological Inconsistency

Although it is impossible to guarantee correct and consistent topological information, a good practice is to make sure that any structural information used is at least self-consistent. This requires making sure parent-child associations are symmetrical and the overall structure is acyclic. These two requirements are achieved simultaneously with the following method.

The modem data provides topological information indirectly since each modem data sample is tagged with the plant, SHUB, SMT, and a modem identifier. A mapping from SHUB to SMT, for example, can then be formed by collecting the set of SMTs that appear in modem samples that also reference that SHUB. This basic strategy is subject to inconsistencies if the same SMT appears in two different SHUBS.

A robust strategy to form maps from each topological level to each other level is to first form associations between adjacent topological levels going up the network tree. For example, each modem is associated to one SMT. Contradictory associations are resolved in an unambiguous manner by taking the *last* observed association, effectively taking the topological view as it was at the *end* of the data period. Once this process is complete for each topological level, other upstream mappings are produced consistently by following the tree backwards to the top. Thus a modem's SHUB is its SMT's associated SHUB, etc.

The reverse mappings may then be formed by reversing the associations. For example, the modems underneath a particular SHUB is the set of modems which associate to that SHUB. This procedure generates a set of self-consistent mappings, although it does not guarantee to be the most accurate when subject to inconsistent data.

Once the topological information is guaranteed to be well formed, algorithms written to analyze these structures are simpler to write and safer to apply. Topological inconsistencies are reported so that analysis of the effected network elements can be observed with caution and the structural information can possibly be rectified at the source.

3.5.4 Unknown Issues

One of the biggest difficulties in dealing with data collected from real world operations is the unpredictability of the system. No amount of historical analysis can prepare the fault detection system for all future possibilities. New and unanticipated imperfections within the data sources should be expected on occasion. Consequently, the fault detection system may fail. Unfortunately this is often the case, especially in a prototype stage, as it is with many software systems that are exercised with newly encountered input conditions. What may be worse is the problems may go unnoticed, and the faulty results are taken without question.

One can never deal with this problem entirely. Ways to mitigate the effects of unexpected problems is to test the system over as large a set of real input data as possible, and to inspect the validity of results on occasion by drilling down from the high level results. Unknown issues are a large reason for the continued monitoring of any large system.

3.6 Data Sources Availability

Figure 3.17 shows the availability over time of the modem, SMT, and stability data[4]. Common periods of data are required for an effective analysis that integrates the data sources.

[4]SMT data is also available for many years prior to this.

The SMT data was being phased out and is scarce during this period, so it is fortunate that modem data is available at all in the same time window. There is no period where all three data sources overlap, although some modem data was available for the period that the stability data covers. Larger gaps of missing data are apparent from the diagram, although there are many one or two day gaps for the different plants that are not visible.

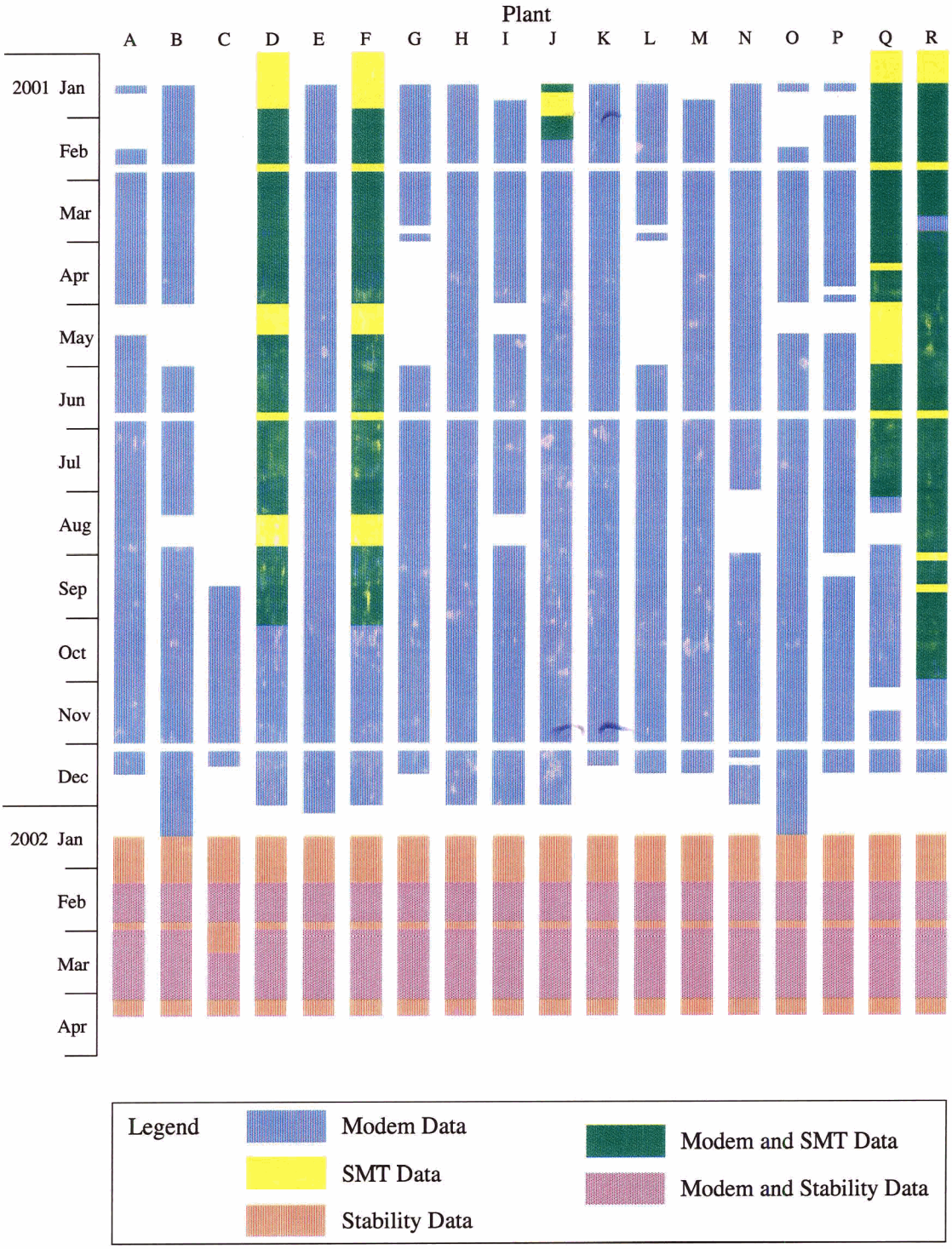

Figure 3.17. *Data Availability*

Chapter 4

Feature Generation

To facilitate the analysis of the data sources, high level features are extracted from the raw data. These are in the form of traditional statistics of the data signals as well as some custom feature detectors to elicit more specific quantitative measures of the signals. Effective feature extraction aims to reduce the dimensionality of the input data while maintaining the meaningful properties that may lead to fault detection. A network element is represented by a feature vector which is amenable to a wide selection of data analysis techniques. The various features derived from the data sources are described in the sections that follow.

For the sake of clarity when referring to the analysis of data sources it will be assumed that only the data values within the analysis time period are considered and not the entire historical archive.

To help distinguish features from other quantitative values, such as signal names, feature names will be identified using short descriptions rather than single symbols. For example while a modem power signal for modem m is represented as $p_m(t)$, a feature for modem m will have a more verbose name of the form $feature_m$. For consistency features are defined only in terms of the data signals and not from other features.

Features are first defined individually, then feature vectors representing network elements are summarized. Since elements within a network form a hierarchy, features for higher levels in the network are partially characterized by statistics of network elements at lower levels within their scope. Thus, low level modem features are not only used to characterize individual modems, but they are also used to generate aggregate features for

SMTs, SHUBS, etc.

Formulas are presented to give a precise specification for the derivation of features because literal descriptions are often ambiguous. The equations are also meant to provide enough information to help recreate the exact analysis methodologies described herein.

4.1 Modem Data Features

The modem power and CRC signals are processed to extract a set of features that are used in fault detection analysis. Symbolic representations for the modem data signals are presented here to help specify formulas involving the modem signals in mathematical form. This treatment aims to eliminate ambiguity in their presentation and provide sufficient detail to allow reproduction of these methods.

The modems in a plant for a specific analysis time period are each assigned a unique number m, the order of assignment has no meaning and only serves to individuate them. A modem m has a power signal $p_m(t)$ and CRC signal $c_m(t)$ as provided by the modem data. Each signal is defined for $t \in T_m$, the sample times unique to that modem which form an ordered sequence $T_m = \{t_{m_1}, t_{m_2}, ..., t_{m_{\|T_m\|}}\}$.

4.1.1 Number of Samples

The number of samples for both the power and CRC signals of a modem provides a feature that indicates how much data was recorded from that particular modem. Typically, this number will be close to the number of hours in the analysis time period. Without speculating why the amount of data received from each modem varies, the feature is simply extracted for later analysis. The measure for modem m is simply the size of the set of samples for the modem, which is equal to the number of elements t in the set T_m for modem m.

$$num\_samples_m = \|T_m\| \tag{4.1}$$

4.1.2 Mean Power

The mean power of a modem signal is a simple average of all the power signals present over the analysis period for a particular modem. Specifically,

$$mean\_power_m = \overline{p_m} = \frac{\sum_{t \in T_m} p_m(t)}{\|T_m\|} \tag{4.2}$$

This feature gives a simple idea of the operating level of the modem power. Although it disregards all information pertaining to the signal variation, the level alone could reveal issues that significantly influence the operational level of a modem.

4.1.3 Standard Deviation of Power

Ignoring the mean operational level of the modem power, the standard deviation [7] gives a measure of the signal variation that helps give an indication of the amount of power signal fluctuation.

$$std\_power_m = \sqrt{\frac{\sum_{t \in T_m} (p_m(t) - \overline{p_m})^2}{\|T_m\| - 1}} \tag{4.3}$$

4.1.4 Mean CRC

The mean of a modem CRC signal provides a feature that gives an indication of the overall CRC activity for the modem. This feature could help reveal trends that relate poor modem data transmission qualities with other features. Similar to the modem power, it is calculated as

$$mean\_crc_m = \overline{c_m} = \frac{\sum_{t \in T_m} c_m(t)}{\|T_m\|} \tag{4.4}$$

4.1.5 Standard Deviation of CRC

The modem CRC standard deviation gives a measure of variability of the modem transmission error rates. The CRC signals are quite bursty in nature and this feature helps differentiate between signals that have the same average but different spike magnitudes.

$$std\_crc_m = \sqrt{\frac{\sum_{t \in T_m}(c_m(t) - \overline{c_m})^2}{\|T_m\| - 1}} \tag{4.5}$$

4.1.6 Number of CRC Spikes

The modem CRC signal spikes, in contrast to the mean and standard deviation of the CRC signal, provide information about the signal that may correspond to significant events that cause transmission errors. The CRC spike count for a modem is generated from a feature detector algorithm, as opposed to the well established statistics used for the previous few features. The spikes are identified using the following algorithm which produces a set of spike events characterized by both their magnitude and time of occurrence. The number of modem CRC spikes feature is determined as the size of this set.

The spike detection method is fairly rudimentary and simply uses a threshold parameter h which defines the CRC level above which the level is considered to be within a spike. Consecutive samples above the threshold are considered part of the same spike, therefore the number of spikes is equal to the number of times the CRC signal crosses the threshold in the positive direction.

$$num\_crc\_spikes_m = \|\{i | 2 \leq i \leq \|T_m\| \text{ and } c_m(t_{i-1}) < h \text{ and } c_m(t_i) \geq h\}\| \tag{4.6}$$

This definition does not pose an upper or lower limit on the duration of the CRC spike nor does it require that the spike ever end (although there can be at most one spike that does not pass back below the threshold). An absolute threshold was chosen instead of a difference-based threshold so that CRC spikes from different modems have an equivalently interpretable meaning.

A threshold value of $h = 0.05$ was selected based on observation of many CRC signals, safely above the noise floor but not so high as to miss what appear to be significant spike events.

4.1.7 Power-Temperature Correlation

A correlation measure between a modem signal and the plant temperature was used initially as the primary behavioural indicator. Due to the importance of a general behavioural measure and a few shortcomings of this feature, a more effective measure was developed later in the analysis (4.5).

The correlation between two signals gives an indication of their similarity. In this case, the modem power signal is correlated with the SMT derived plant temperature estimate if it is available (if the SMT data is available for the particular analysis period), otherwise it is correlated with the modem power derived plant temperature estimate. Correlation compares two signals sample to sample, so they are resampled to a common time base using a method similar to that in section 4.3.7 prior to comparison. To produce a feature that is comparable between modems, the signals are also normalized to have a peak-to-peak amplitude of two units. This produces a correlation coefficient in the range $[-1, 1]$. The correlation of two identical signals yields a coefficient of 1 while a signal and one 180 degrees out of phase yields a coefficient of -1.

For a power and temperature signal, $p(t)$ and $m(t)$, both sharing sample times $t \in T$, the correlation coefficient is

$$r_{pm} = \frac{\sum_{t \in T}(p(t) - \overline{p})(m(t) - \overline{m})}{\sqrt{\sum_{t \in T}(p(t) - \overline{p})^2 \sum_{t \in T}(m(t) - \overline{m})^2}} \tag{4.7}$$

where $\overline{p}$ and $\overline{m}$ are the means of the signals $p(t)$ and $m(t)$, respectively.

This direct correlation has a serious drawback due to the frequent level shifts in the modem power signals. Since the correlation subtracts out the average of each signal, a two similar signals correlate poorly if one of them has an artifact such as a level shift in it. The two signals will not overlap closely because the level shift creates an offset that cannot be compensated for by the simple mean cancellation present in the correlation formula. Since level shifts are detected in a separate feature and the power-temperature correlation is intended to give a measure of similarity, this is an undesirable effect.

For this reason the power-temperature correlation feature is derived using a modified

technique. The two signals are broken into several shorter signal segments of 50 samples each, and are then compared piece-wise. A vector of correlation coefficients $\mathbf{R}_m$ is produced by applying equation 4.7 to each corresponding pair of 50 sample segments.

$$power\_temp\_corr_m = \overline{\mathbf{R}_m} \tag{4.8}$$

The power-temperature correlation feature is the average of these correlation coefficients, excluding those which are invalid due to zero variation in at least one of the signal segments.

4.1.8 Power-Temperature Correlation Standard Deviation

Another modem feature is generated simply by taking the standard deviation of the vector of correlation coefficients generated while determining the power-temperature correlation between a modem power signal and the plant temperature signal.

$$power\_temp\_corr\_std_m = std(\mathbf{R}_m) \tag{4.9}$$

4.2 Minimum Mean Squared Error

As a distance measure between two discrete time signals the minimum mean squared error (MMSE) technique is used. This technique shifts and scales one signal, y_1, to find the minimum error with respect to a reference signal, y_2, using per-sample differences. This is similar to the least squares line [4] but it scales a second signal instead of a line.

Selection of this metric is based on the desire to find similarity between signals, possibly from different amplitude dimensional scales, such as modem power and SMT temperature. It is also insensitive to small time shifts which may arise due to delayed temperature effects or unsynchronized timestamp clocks.

The modem power-temperature correlation from section 4.1.7 is a measure with similar intent, yet this method was devised because there was a need for very sensitive error measurements used to validate the temperature estimation methods in section 4.4.

To ensure the MMSE measure produces comparable values between pairs of signals the reference signal y_2 is normalized to have a standard deviation of 1. Consequently, two arbitrary signals (i.e. two modem power signals or a modem power signal and a SMT temperature signal) can be compared and the MMSE will provide a consistent measure of similarity.

The two signals are assumed to have been preprocessed to remove any invalid data and have been resampled to consistent times so that per-sample differences can be computed. The signals have equal length N.

For a given time shift of k, the mean squared error between y_1 and y_2 is

$$MSE_k(y_1, y_2) = \frac{1}{N}\sum_{n=1}^{N}(\alpha y_1(n-k) + \beta - y_2(n))^2 \tag{4.10}$$

where α is a vertical scaling factor and β is a vertical shifting factor. These are the coefficients that will be used to minimize the mean squared difference. Without loss of generality, the signals are assumed to have means of zero, since the MMSE can shift y_1, arbitrarily with respect to y_2, so the initial offset is irrelevant.

$$\sum_{n=1}^{N} y_1(n-k) = \sum_{n=1}^{N} y_2(n) = 0 \tag{4.11}$$

However, mean centering simplifies the derivation of parameters that minimize the error. As a quadratic function with positive second derivatives with respect to α and β, the minimum occurs when

$$\frac{\partial MSE_k(y_1, y_2)}{\partial \alpha} = \frac{\partial MSE_k(y_1, y_2)}{\partial \beta} = 0 \tag{4.12}$$

The partial derivatives evaluate to

$$\frac{\partial MSE_k(y_1, y_2)}{\partial \alpha} = \frac{2\alpha}{N}\sum_{n=1}^{N} y_1^2(n-k) + 2\beta\sum_{n=1}^{N} y_1(n-k) - \frac{2}{N}\sum_{n=1}^{N} y_1(n-k)y_2(n) = 0 \tag{4.13}$$

and

$$\frac{\partial MSE_k(y_1, y_2)}{\partial \beta} = \frac{2\alpha}{N}\sum_{n=1}^{N} y_1(n-k) + 2\beta - \frac{2}{N}\sum_{n=1}^{N} y_2(n) = 0 \tag{4.14}$$

Solving for the vertical shifting coefficient β from 4.11, 4.12 and 4.14, we find

$$\beta = 0 \tag{4.15}$$

implying that the minimum error is achieved when the means of the signals are equal. The α coefficient, from 4.11, 4.13, 4.12, and 4.15, is then

$$\alpha = \frac{\sum_{n=1}^{N} y_1(n-k)y_2(n)}{\sum_{n=1}^{N} y_1(n-k)^2} \tag{4.16}$$

Using the α and β values the MSE_k is computed from 4.10. This is repeated for each time shift $-K \leq k \leq K$ and the lowest MSE_k is taken as the MMSE. For practical purposes, $K = 3$ is found to be a reasonable time shift range. Observed MMSE values range from 0 to approximately 1.

4.3 Signal Preprocessing

To help overcome the data issues outlined in section 3.5 some techniques were adopted to process modem data signals in circumstances that assume reliable data. At the same time, information collected during these preprocessing steps may be regarded as features of the signals themselves.

4.3.1 Flat Levels

Flat levels are considered as a modem power signal feature *flatness* and are also filtered from the signals before analysis when appropriate. Each sample time $t \in T_m$ for modem m falls within a flat level ($F_m(t) = 1$), or it does not ($F_m(t) = 0$). Regions of the signal where the power level is constant for S or more consecutive samples are considered flat levels. In the analysis of the modem data a value of $S = 24$ (one day) is used. For the ordered sequence of sample times $t_i \in T_m$,

$$F_m(t_i) = \begin{cases} 1, & if\ P_m(t_i) = P_m(t_l); \text{forall } l \in [j,k], k-j \geq S\ and\ j \leq i \leq k \\ 0, & otherwise. \end{cases} \tag{4.17}$$

flatness_m is the proportion of modem power samples that lie in flat levels. This is equal to the average value $F_m(t)$ for all values of $t \in T_m$.

$$flatness_m = \frac{\sum_{t \in T_m} F_m(t)}{\|T_m\|} \tag{4.18}$$

The flat level detector therefor serves to classify each power sample and it also servers as the basis for another modem feature.

4.3.2 Zero Levels

Zero levels may signify network faults but are usually filtered out before signal analysis. Unlike flat levels, a zero levels do not require multiple consecutive samples since they are easily detected in the data. For a modem power signal $p_m(t)$ the function $Z_m(t) \in \{0, 1\}$ characterizes each time sample $t \in T_m$ as being zero according to

$$Z_m(t) = \begin{cases} 1, & if\ p_m(t) = 0, \\ 0, & otherwise. \end{cases} \tag{4.19}$$

This rather obvious classification is formalized to be consistent with the other data filters for reasons that will be made clearer in section 4.3.4. Another modem feature is now defined as

$$zeros_m = \frac{\sum_{t \in T_m} Z_m(t)}{\|T_m\|} \tag{4.20}$$

4.3.3 Clipped Data

Modem power signals are sometimes clipped at an apparent maximum measured value. As mentioned in section 3.1.1.2, this likely does not reflect the real modem power variation and certainly not the temperature variation of the plant. Oddly, different modems have slightly different clipping levels, although all are near the 5600 dBmV mark.

For a modem power signal $p_m(t)$ the function $C_m(t) \in \{0, 1\}$ specifies which samples are clipped. For each sample time $t \in T_m$

$$C_m(t) = \begin{cases} 1, & if\ p_m(t) = \max(p_m(t))\ and\ \max(p_m(t)) > 5500, \\ 0, & otherwise. \end{cases} \tag{4.21}$$

The real clipping level is unknown, yet this function assumes that the maximum observed power level is the clipping level, provided it is high. It is possible that some samples above 5500 are considered clipped even when they are not, but it is unlikely that many will be since they must all be exactly the same maximum power level.

A feature defined similarly to $flatness_m$ and $zeros_m$ is

$$clipped_m = \frac{\sum_{t \in C_m} Z_m(t)}{\|T_m\|} \tag{4.22}$$

4.3.4 Filtering Invalid Modem Data

Sections 4.3.1, 4.3.2 and 4.3.3 describe functions which characterize each power sample of a modem. Combining these functions, a function describing which samples are *unusable* in that for at least one reason they do not reflect the real modem power. Collectively they suggest which power samples should be removed from a signal in an *invalid data* function

$$I_m(t) = \max(F_m(t), Z_m(t), C_m(t)) \tag{4.23}$$

Clipped power samples or samples at zero are also commonly within a flat level. This function provides a convenient way to measure the proportion of a signal that is unusable, namely $\overline{I_m}$, without overcounting. This function is used to specify a filtered set of modem power samples with times T'_m

$$T'_m = \{t | t \in T_m\ and\ I_m(t) = 0\} \tag{4.24}$$

A new modem power signal is created using only those samples which are valid according to the $I_m(t)$ function. Use of this filtering process is implied whenever the modem power signal is said to be filtered of unusable data.

4.3.5 Missing Data

Section 3.5.1 discussed missing data in the data sources, including modem signals. Since missing samples do not have timestamps, they are not detected in the normal sense. Missing data is not a feature since it is effectively the same as the $num\_samples_m$ feature with respect to other modems from the same plant. Missing data is a third class of data, independent of valid data and invalid data.

4.3.6 Valid Data Measure

Modem power signals are also assigned a validity measure, which is a collective measure rather than a per-sample measure. In essence, it represents the portion of the total signal that is considered usable. Since there is often a period for which all modems in a plant are missing data (due to a data collection interruption for example), data availability is best measured relative to the largest observed modem signal rather than the theoretical maximum. This allows for sound comparisons against absolute thresholds (which is done in section 4.7), making the measure insensitive to any amount of common missing data.

The valid data measure is normalized against the largest number of valid samples of any modem in the analysis time period, $\|T'_{max}\|$. For modem m, with valid samples T'_m, the validity feature is

$$valid\_data_m = \frac{\|T'_m\|}{\|T'_{max}\|} \tag{4.25}$$

This feature is also used when computing aggregate statistics of modem signals at a higher level, where is important to use only modems with sufficient reliable data, such as those in section B.1.

4.3.7 Common Time Base and Signal Resampling

A sample-to-sample comparison of two signals assumes that they have equal sampling times. Modem signals each have a one hour sampling period, yet two arbitrary modems

will have a phase offset, making them awkward to compare. This is overcome by resampling the signals to a common time base. Resampling a signal introduces some noise, but resampling a fairly continuous signal to the same sampling frequency should not introduce much noise. Although pre-filtering and higher order polynomial interpolation can reduce noise introduced by resampling, a linear resampling method was adopted for the purpose of resampling the modem signals.

The common sample time base chosen for this purpose is one sample every hour on the hour starting at the first hour of the analysis time period (which is typically chosen to start on the day boundary).

Individual modems might not have samples at each of the common sample times due to missing data. Time gaps will persist in the resampled signals and thus any given hour may have samples present from only a subset of modems. It is important to filter bad data from the modem power signals prior to resampling them, otherwise many samples would no longer be considered invalid according to the classification functions 4.17, 4.19, and 4.21. For example, isolated zeros would be resampled to non-zero values, and the boundary samples of flat or clipped regions would change.

The linear resampling technique is applied to the modem power signal P_m as follows, producing new power time series $P'_m(t)$ with sample times only from the common time base. Due to missing samples, resampling occurs between each adjacent pair of valid sample times, $t_a, t_b \in T'_m$, that are within 90 minutes of one another, i.e. $t_a \leq t_h \leq t_b$ and $t_b - t_a \leq 90m$. The power level for the resampled time stamp is then

$$P'_m(t_h) = P_m(t_a) + (P_m(t_b) - P_m(t_a))\frac{t_h - t_a}{t_b - t_a} \tag{4.26}$$

The time series generated this way for each qualified hourly time stamp $t_h \in T'_m$ is the resampled version of the original signal P_m.

4.4 Temperature Estimation using Modem Power

This section demonstrates that plant temperature can be reliably estimated from modem power signals in networks where temperature data is not readily available. Modem power signals collectively reflect the plant temperature although no individual signal necessarily reflects this trend accurately. Taking a large sample of modem power signals from a plant, they are preprocessed and selectively combined into a single temperature estimate.

This method is validated against a temperature average generated from temperature readings measured directly from the SMTs on trunk amplifiers distributed throughout a plant. The consistency between the temperature signals produced by the two techniques strongly implies that the modem power fluctuations are primarily due to the changing plant temperature, and that the modem power signals alone provide an estimate of the plant temperature variation.

Furthermore, the pervasive temperature influence is enough to suppose a model for normal modem power variation, one that follows the outside temperature variation closely. This model is employed to assign each modem a measure of normalcy that is the basis for identifying regions of the cable network that exhibit questionable behaviour.

For this process to be applied, the modem signals are first preprocessed to remove samples that are considered invalid and are then resampled to a common time base to facilitate a per-sample analysis. A similarity measure is selected that suits the signal comparisons involved in this process.

These steps are elaborated in the sections that follow. The resulting signals are used to generate a modem power MMSE feature which is central to the analysis in later chapters.

The following method is used to form an estimate of plant temperature using only power signals of modems in a plant. Variations of the method are compared and the chosen technique is validated by comparing this temperature estimate to the SMT derived estimate from appendix A.

For these tests one month of modem data is used from the target plant. Although a

different duration could be used, one month strikes a good balance between reliability and computational cost.

A hindrance to developing this technique was the restricted time frame when both modem and SMT data were available (see figure 3.17). Even then only two plants had sufficient data to permit the analysis, with a total of four different one month periods over which the temperature signals could be derived. Nevertheless, these few opportunities bring credibility to the method, and suggest that a reliable temperature estimate can be derived from the modem data alone. The four time periods supply the signals used to refine the estimation procedure, and are shown in table 4.1.

Table 4.1. *One Month Periods with Modem and SMT Data*

Plant	Start Date	End Date
Plant R	January 17 2001	February 17 2001
Plant R	June 1 2001	July 1 2001
Plant R	July 1 2001	August 1 2001
Plant D	May 20 2001	June 20 2001

The method used to combine each collection of modem samples[1] was progressively refined until further estimation improvements seemed unlikely. Many variations were investigated, yet only those which improved performance are described here. Descriptions of the refinements are presented in the following sections.

To compare SMT temperature estimates and modem power estimates using the MMSE technique, the SMT estimates are resampled to hourly samples to coincide with the modem signals. Information is lost in the downsampling process but this is required for signal comparison on a per-sample basis.

[1] Attempts will be made to keep the intended meaning of the word *sample* clear from context.

4.4.1 Selection of Modems

A random sample of 1000 modems from the target plant is selected. This number was chosen after initially observing an unacceptable level of signal variation in successive estimates using early versions of this technique. Larger collections of modems incur a higher computational cost without significantly improving consistency. Table 4.2 shows the proportion of each plant 1000 modems represents. It will be shown that this sample size is sufficient in large plants. Even in smaller plants, there is no harm in using a majority of modems.

Table 4.2. *Proportion of Total Plant Modems Used for Estimate*

Plant	Plant Modems	% of Plant Modems
Plant R Jan 2001	7337	13.6
Plant R June 2001	7215	13.9
Plant R July 2001	7351	13.6
Plant D May 2001	2290	43.7

4.4.2 Modem Signal Preprocessing

The selection of modem signals is first preprocessed. Several features are filtered from the modem power signals as per section 4.3 before incorporating the signals into a temperature estimate because they do not represent real power variation, including embedded status data, clipped power levels, and flat regions (see section 4.3). Status signals are resampled to the same hourly time base as described in section 4.3.7. The resulting data is viewed as a collection of samples for each hour in the time period, with up to but not necessarily 1000 samples each because of filtered and missing data.

It is important to monitor the quantity of samples dropped from the procedure because it is possible that the amount of data left is too small to support a reliable estimate. Each hourly sample is required to have at least half of the initial number of samples (500), otherwise the sample is dropped from the estimate altogether.

4.4.3 Trimming the Distribution Tails

Table 4.3. *MMSE of Modem Power Distributions Trimmed at 1.5 Standard Deviations*

Plant	Median			Mean		
	MMSE	Avg Samples/Hr	Estimate Samples	MMSE	Avg Samples/Hr	Estimate Samples
Plant R Jan 2001	0.2219	814.0	709	0.3075	805.9	709
Plant R June 2001	0.0815	748.3	647	0.0460	756.6	647
Plant R July 2001	0.1882	798.4	734	0.1983	809.5	734
Plant D May 2001	0.0464	818.2	696	0.0432	805.2	696
Average	0.1345	794.7	696.5	0.1488	794.3	696.5

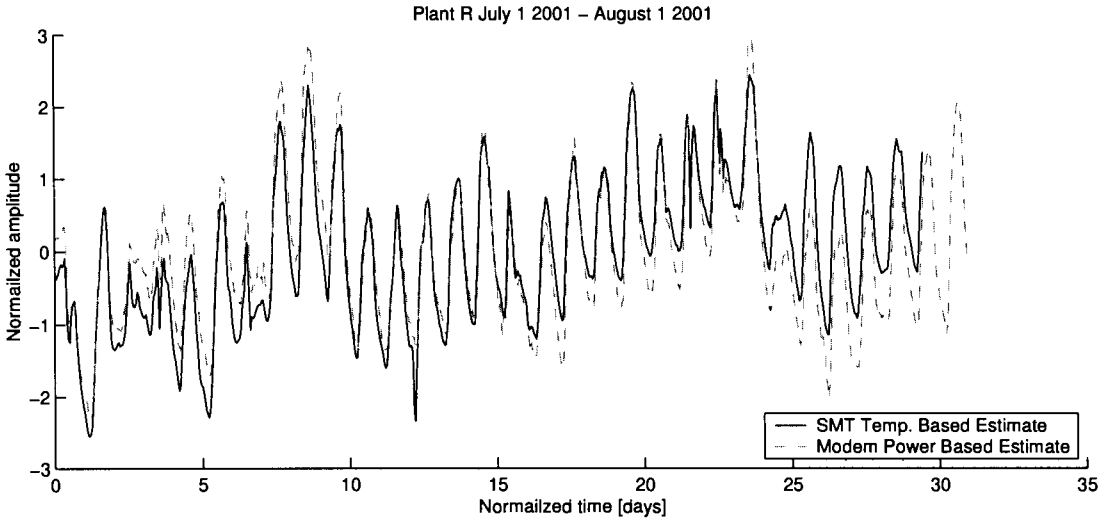

Figure 4.1. *Temperature Estimate Using Trimmed Modem Power Distribution*

To select a representative temperature estimate for each hour, the distribution of modem power levels at that time is considered, and the samples that lie at the extremes of the distribution are discarded. The motivation for this is that the majority of modem signals follow the temperature trend, and those which deviate from this behaviour are likely influenced by factors which interfere with the estimate, and moreover, they may be indicative of abnormal behaviour. Inspection of sample histograms and numerical tests showed that 1.5 standard deviations from the mean of the distribution for each hour was a good threshold.

The estimated temperature level for an hour is taken from the set of samples for that hour that lie within 1.5 standard deviations of the mean. Two candidate signals are generated from the sequence of sample distributions, one using the mean of each trimmed distribution, and one using the median. The modem power based estimate using this method is shown against the SMT derived estimate (appendix A) in figure 4.1.

The MMSE errors for each of the eight estimated signals are given in table 4.3, and an average is given over the four plants for each of the two combining methods. The average number of samples per hour used in the estimate is given to show that of the potential 1000 samples, close to 80% were available (not missing, filtered, or trimmed). These numbers are not the same for the mean and median derived signals because each estimate signal is derived from a different set of random modems. The estimate samples are the length of the resulting estimate signals, and it is the number of hourly distributions with at least 500 samples. These are usually equal for all signals generated for a specific plant because certain time gaps affect all modems at once and during these times no power based estimates can be made.

While two of the estimates have MMSEs less than 0.1, the two over 0.18 are not adequate. Further refinement of the method is necessary.

4.4.4 High Pass Filtering

Table 4.4. *MMSE of High Pass Filtered Estimates*

Plant	Median			Mean		
	MMSE	Avg Samples/Hr	Estimate Samples	MMSE	Avg Samples/Hr	Estimate Samples
Plant R Jan 2001	0.0670	814.0	709	0.0586	805.9	709
Plant R June 2001	0.0359	748.3	647	0.0249	756.6	647
Plant R July 2001	0.0549	798.4	734	0.0410	809.5	734
Plant D May 2001	0.0261	818.2	696	0.0216	805.2	696
Average	0.0460	794.7	696.5	0.0365	794.3	696.5

The initial estimates show promise, as the variation of the modem power derived esti-

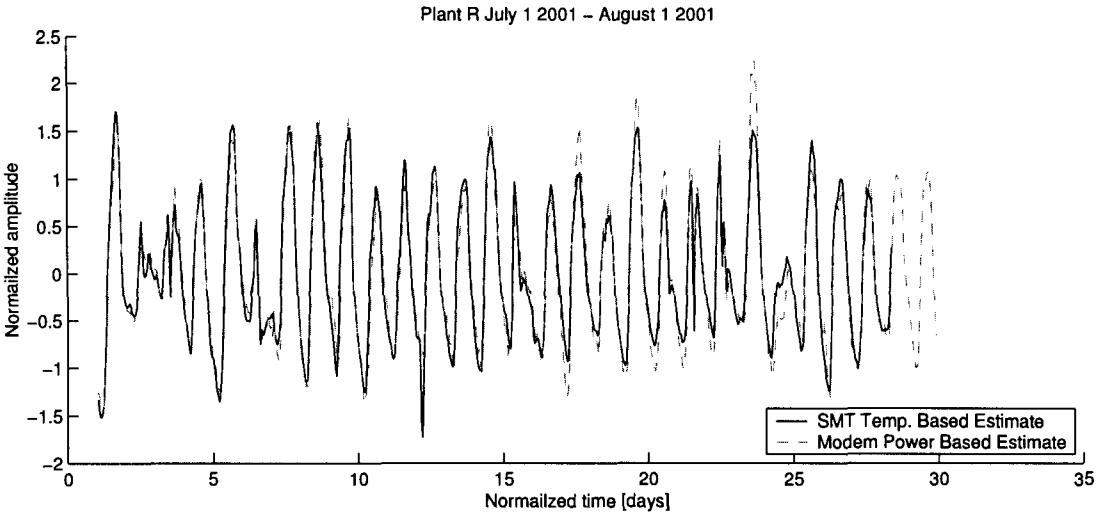

Figure 4.2. *High Pass Filtered Temperature Estimates*

mate matches closely with the SMT derived estimate. The major source of error, particularly with the two highest MMSE errors, appears to be caused by an offset in the estimate rather than the shape of the signal variation. It is believed these offsets are caused by level shifts in the modem power signals (see section 3.1.1.5), despite the trimming efforts mentioned above. These level shifts are almost certainly not sudden temperature changes and are accounted for separately in a level shift feature attributed to each modem. This prompted the use of a high pass filter on the two estimates before comparison. Eliminating any drift between the two compared signals is bound to reduce the MMSE error, and does not reflect an improvement in the signal estimation method. It merely increases the sensitivity of the estimation error to the daily variation by eliminating the component of error caused by the less interesting long term discrepancy. Consistent application of a filter should not impede the goal of separating abnormal from normal modem power signals. Note that the high pass filter is applied only for calculating the error, and does not change the estimated signal itself.

After signal generation but prior to comparison, the modem derived estimate and the SMT derived estimate are subject to a high pass filter. The filter is designed to maintain

variation in the signals at the 24 hour scale but to eliminate variation that persists over longer periods, hence removing lower frequency variations. The filter is achieved by first generating a lowpass filtered signal containing just the slow drift component, and then subtracting it from the original signal, leaving a signal with no drift. The lowpass filtered signal is formed by averaging the original signal over $I + 1$ consecutive samples ($I/2$ samples on either side plus the sample itself). Valid for even I,

$$y_{hpf}(n) = \frac{1}{I+1} \sum_{i=-I/2}^{I/2} y(n+i) \tag{4.27}$$

Visually and empirically, a value of 50 was deemed appropriate for I. Performance of the filtered estimates are summarized in table 4.4. The estimation error is reduced, as expected, and the filtering has made the errors more consistent between test signals, owing to the elimination of the low frequency drift. The mean number of samples retained is the same as in table 4.3 because the two methods produce the same initial signal.

4.4.5 DC Filtering

Table 4.5. *MMSE of DC Filtered Estimates*

Plant	Median		Mean	
	MMSE	Estimate Samples	MMSE	Estimate Samples
Plant R Jan 2001	0.0561	709	0.0557	710
Plant R June 2001	0.0223	647	0.0196	647
Plant R July 2001	0.0314	734	0.0277	735
Plant D May 2001	0.0235	696	0.0225	696
Average	0.0333	696.5	0.0314	697.0

In an attempt to narrow the distribution of power samples for each hour, the power signals have their DC component removed by subtracting out the mean of the signal prior to collecting the hourly distributions of samples. The motivation for this is that the distribution of the average of the selected modem power signals is interfering with the distribution

of modem power levels in each hourly distribution, while the goal is to measure the difference in signal variation and not the difference in average amplitude. The improvement is significant, as shown in table 4.5[2].

4.4.6 Exclusion Filter

Table 4.6. *MMSE Using Different Exclusion Window Sizes*

	-100			-200			+/-100		
Plant	MMSE	Samples/Hr	Samples	MMSE	Samples/Hr	Samples	MMSE	Samples/Hr	Samples
R Jan	0.0507	663.7	705	0.0886	655.8	705	0.0487	644.3	703
R Jun	0.0210	641.3	645	0.0202	606.6	643	0.0213	595.3	642
R Jul	0.0268	635.3	734	0.0280	639.7	734	0.0285	617.8	734
D May	0.0222	630.0	696	0.0430	577.3	630	0.0373	569.9	527
Avg	0.0302	642.6	695	0.0340	619.9	678	0.0339	606.8	652

A final attempt to improve estimation accuracy stemmed from the observation that modem signals, even the well behaving ones, have occasional bouts of erratic behaviour. For example, a power spike is a short but substantial signal change. A more stringent hourly distribution based filter could help remove the influence of occasional artifacts from the well behaved signals. This filter considers the distribution of hourly samples for the current hour as well as the distributions in a time window surrounding it. Any modem which falls outside the 1.5 standard deviation mark for any of the distributions in the time window are removed from consideration, leaving only those modem samples which are consistently close to the average.

Several windowing strategies were evaluated: one including (up to) the previous 100 hours, one including the previous 200 hours, and one including both the previous and next 100 hours. Results are summarized in table 4.6. In each case only the mean of the remain-

[2]The average number of samples used was not recorded for this experiment, however, the final method which also uses DC filtering shows that sufficient samples are retained.

Table 4.7. *MMSE of Modem Power Temperature Estimates*

Plant	MMSE
Plant R Jan 2001	0.0507
Plant R June 2001	0.0210
Plant R July 2001	0.0268
Plant D May 2001	0.0222
Average	0.0302

ing samples is given[3]. Larger window sizes introduced the danger of dropping many or all of the modems, as can be seen by the smaller average number of samples used. Windows larger than those mentioned here, including one that spanned the whole signal, tended to drop the entire signal. The window of 100 hours into the past produced the best result.

4.4.7 Summary

Ultimately, the estimated temperature level for each hour is taken as the mean of the DC filtered modem power signals who do not fall outside 1.5 standard deviations of the overall sample mean each of the previous 100 hours. The resulting ensemble of samples is a reliable estimate of plant temperature.

Figure 4.3 shows temperature estimates against a measured temperature signal. In any case, the SMT derived signal is an estimate derived from a different source and has an inherent error itself, suggesting a level beyond which the estimates cannot agree. Table 4.7 shows that size and consistency of the modem power temperature estimation error is quite acceptable, making these signals a valid substitute for real temperature signals in further plant analysis.

[3] The mean performed better than the median for each of the window sizes so it is removed for clarity.

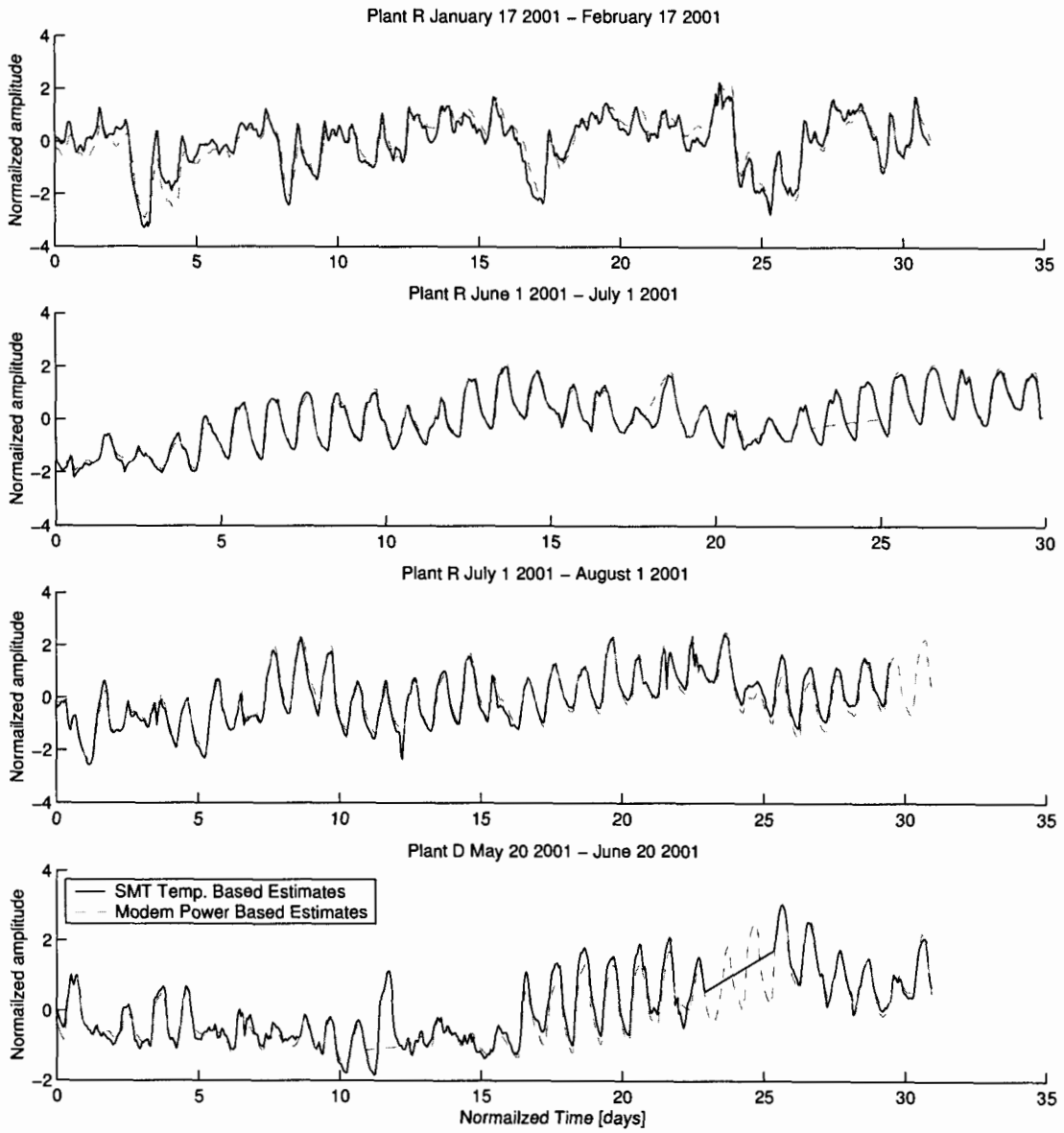

Figure 4.3. *Temperature Estimates: The two signals are shown on top one another. The straight lines are regions of missing data.*

4.5 Modem Power MMSE Feature

The modem power derived temperature estimate from the previous section effectively represents the average "normally" behaving modem. The MMSE between an arbitrary modem power signal and the temperature estimate is taken to represent a metric indicating how normal each modem behaves. This provides a valuable feature for differentiating modems in fault detection analysis.

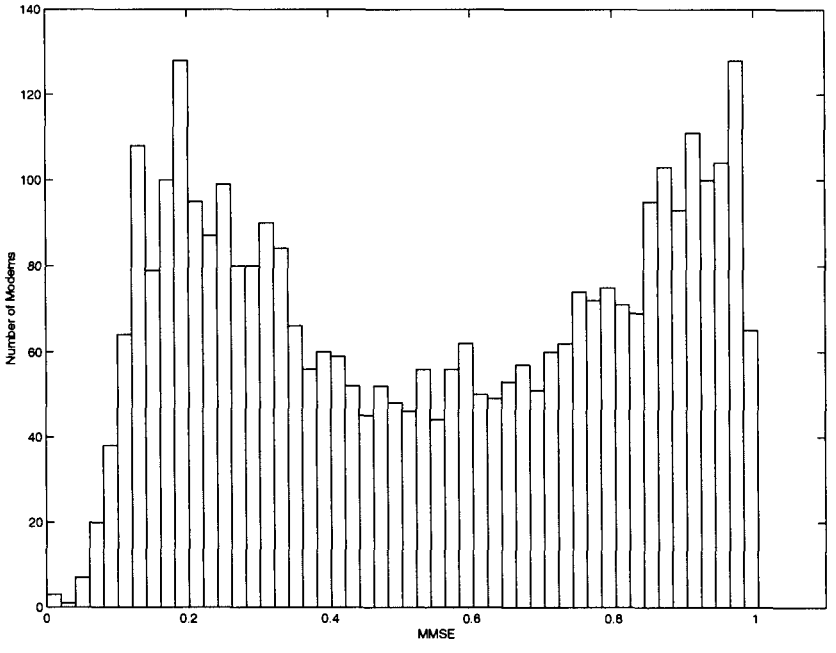

Figure 4.4. *Histogram of Modem Power MMSEs within a Cable Plant*

A typical histogram of MMSE scores for all modems in a plant is shown in figure 4.4. The distribution is bimodal, showing "good" modems which resemble the normal temperature on the left and "bad" modems which have aberrant behaviour on the right[4]. The shape of the distribution for a particular network gives an idea of the overall plant health according to the normality of modem power variation.

[4]The terms *good* and *bad modem* refer only to the behaviour of the modem's power signal and do not imply a problem with the modem itself.

The following chapter will explore how bad modems cluster in different plants and at different levels within the plants and will compare how the identified bad regions compare in the stability data.

4.6 Feature Summary

The previous sections described how to generate a set of features for each modem based on the power and CRC signals. Together these features form a feature vector that characterizes the modem, to some extent, providing a concise representation that may be used for easier analysis. The feature vector F_m for a modem is described by feature elements with indices n given in table 4.8.

Table 4.8. *Modem Feature Vector Structure*

n	Feature
1	m
2	id_m
3	$num\_samples_m$
4	$mean\_power_m$
5	std_power_m
6	$mean\_crc_m$
7	std_crc_m
8	$num\_crc\_spikes_m$
9	$power\_temp\_corr_m$
10	$power\_temp\_corr\_std_m$
11	$flatness_m$
12	$zeros_m$
13	$clipped_m$
14	$valid\_data_m$
15	$MMSE_m$

The first two entries, the modem index and identifier, are not normal features. However, they are useful to have in the feature vector for for tracking purposes. There is also the odd chance that the index and names actually show relationships to other features, which might lead to additional insights of how the data and network is organized.

These features along with the stability fields from section 3.2 (which are segment features), are analyzed in appendix B. As that analysis did not reveal satisfactory trends lower than at the plant level, it is summarized and placed in the appendix so that it does not sidetrack the present analysis.

4.7 Valid Data Modems

Prior to generating collective features from the multitude of modem power signals it is important to consider the data issues described in section 3.5. The idea is to generate higher quality statistics by excluding modems with less valid data. Initial investigations showed a wide range of MMSE values for modems within a plant, but upon closer inspection for some modems this score was misleading and further refinement appeared necessary. The problem was that many modems at the extremes of the MMSE distribution had little data that was telling of the true power temperature relationship due to significant clipping, flat regions, and zero levels. As described in section 4.3, methods of detecting bad unusable samples provide a data validity measure. Modems with little valid data detract from the reliability of the present analysis that makes inferences from the number of poorly behaving modems. The valid data measure from section 4.3.6 was designed to detect modems of this nature, and is now used to separate out the modems with unreliable data.

The valid data feature is now used to classify modems into two classes, those with "valid data" and those with "invalid data".

$$VD_m = \begin{cases} 1, & valid\_data_m > 0.5 \\ 0, & otherwise. \end{cases} \tag{4.28}$$

The threshold of 0.5 seemed reasonable since it eliminated the modems at the low end of the MMSE scale whose data did not appear reliable while at the same time it did not capture too many modems in total. Although many modems have more than 90% valid data, the few between 50% and 90% may be significant, such as a group that are in a faulty network region that also has some missing data. The majority of modems have enough

valid data so the effect of using only valid data modems is that it cleans up calculations using a lot of modem signals.

Figure 4.5 shows the distribution of the valid data feature for all modems in all plants present in the April 2002 data. Of the 42569 total modems, 36932 (86.8%) have valid data.

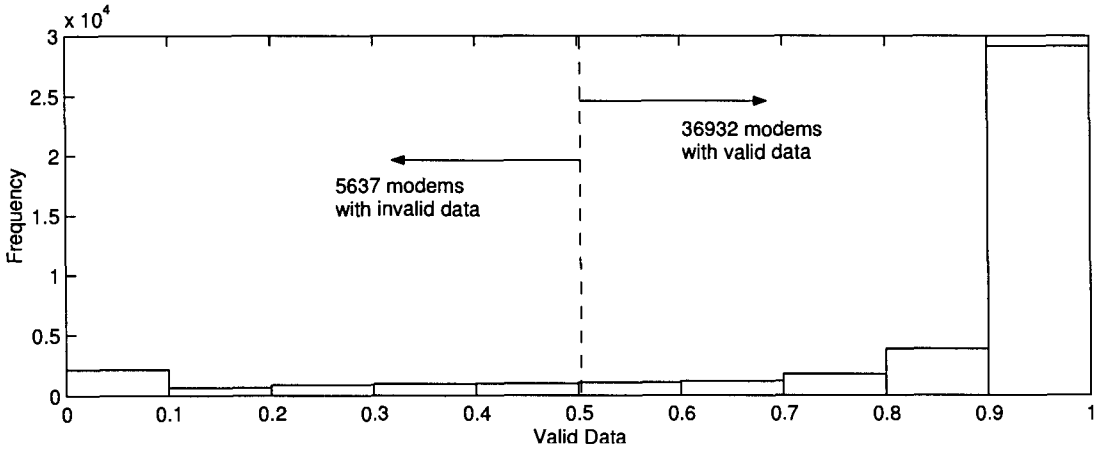

Figure 4.5. *Modem Valid Data Histogram*

The histogram shows the mass majority of modems with a very high percentage of valid data. This is encouraging since a large amount of reliable data remains for analysis after the unreliable modem signals are removed.

Table 4.9. *Modems with Valid and Invalid Data*

Total Modems	Valid Data Modems	Bad Data Modems
42569	36932	5637

4.8 Modem Behaviour Classification

Looking back at the MMSE histogram in figure 4.4, two different classes of modems are apparent. It will simplify matters to classify individual modems as "good" and "bad". The classification is used to help simplify analysis at the large scale. The presence of bad

modems in a segment is not interpreted the same as a larger number of mediocre modems as they are in a simple averaging process. Instead, an interest metric is used, which uses the bad modem concentration. This metric is introduced in chapter 5. In the meantime, a method is presented for classifying a modem as being "good" or "bad". Essentially, modems whose behaviour is far removed from the "normal" temperature behaviour are considered "bad".

4.8.1 Threshold Determination

To classify good and bad modems, an MMSE threshold value is needed. Rather than setting the threshold arbitrarily, or by eye, an information theoretic strategy using all modems from all plants is used. The intention is to use the data available to automatically determine which threshold separates the classes the most effectively.

The number of bad modems which belong to each segment, those which have MMSE values above a threshold T, is used to calculate the information H_T generated from the variability of bad modem concentration over all segments. For a segment s, b_s is the number of modems with MMSE values lower than T and m_s is the number of modems in the segment.

$$H_T = -\sum_{s=1}^{S} \frac{b_s}{m_s} log_2 \frac{b_s}{m_s} \tag{4.29}$$

The threshold T is chosen so that the information metric H_T is maximized. Each of the twenty T values $0.05, 0.10, \ldots, 1.0$ are considered. Using data from February 10th to April 3rd 2002 from 18 plants, the threshold that maximizes H_T is 0.6. Figure 4.6 shows the relationship between the information H_T and the threshold.

4.8.2 Bad Modem Classification

With an established MMSE threshold for splitting the modems into two classes, the modems whose MMSEs lie above the 0.6 threshold value are considered "bad" for the remainder of

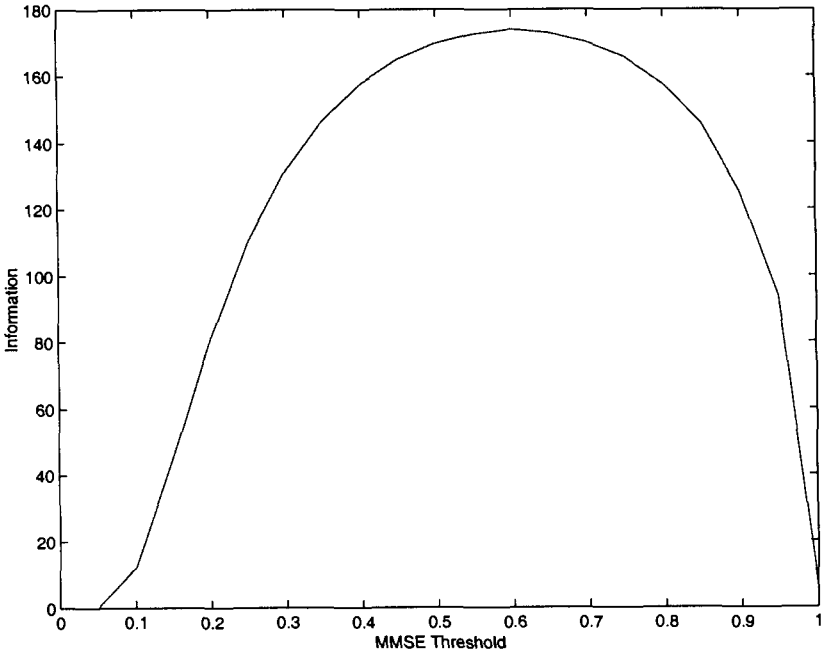

Figure 4.6. *Modem MMSE Threshold vs Information Content*

this analysis. For modem m,

$$B_m = \begin{cases} 1, & MMSE_m > T, \\ 0, & otherwise. \end{cases} \tag{4.30}$$

Figure 4.7 shows where the threshold splits the MMSE distribution. 62.8% of the modem power signals are "good" and the other 37.2% are "bad".

Table 4.10. *Modems with Good and Bad Behaviour*

Valid Data Modems	Good Behaviour	Bad Behaviour
36932	23189	13743

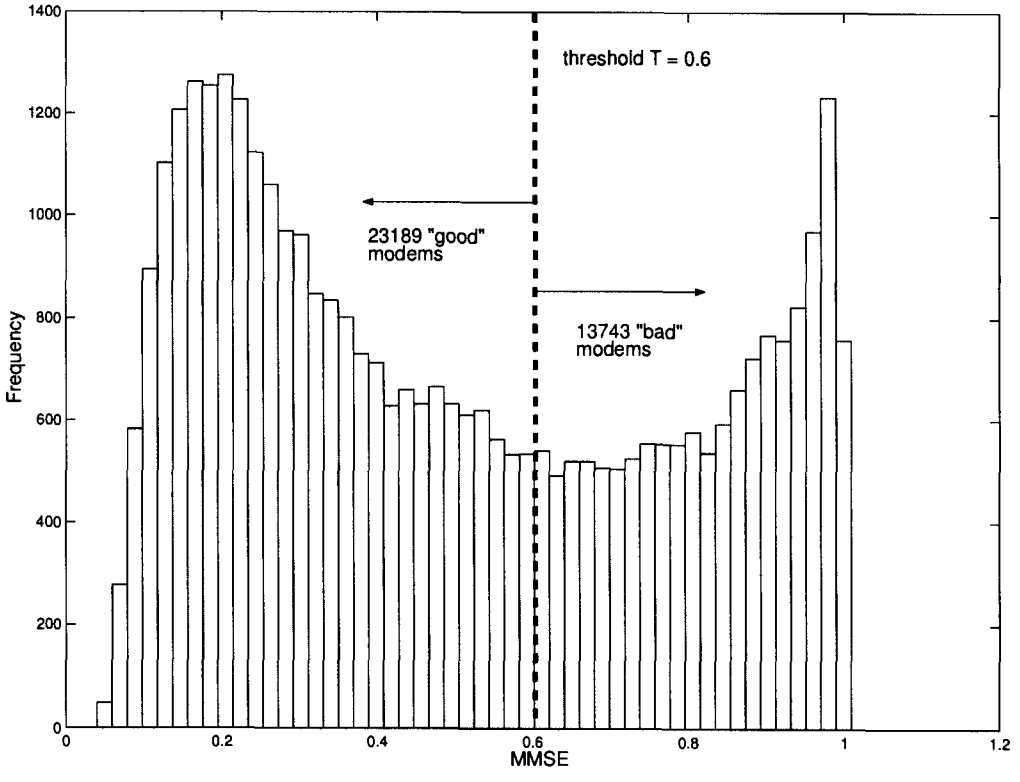

Figure 4.7. *Modem MMSE Distribution and Threshold*

Chapter 5

Fault Determination

The data sources and feature extraction methods described in earlier chapters are applied in an attempt to define a pattern which allows regions of poor plant stability to be predicted from modem data alone. To discover trends that allow for such inferences, modem data from February 10th to April 3rd 2002 and stability data from March 7th to April 7th is used for the analysis. From this, the various modem and stability features are computed for each plant.

A correlational analysis of the cable modem and stability data (appendix B) indicated the existence of a link between poor modem power behaviour and plant stability. However, the correlation apparent at the plant level was not observed at the segment and lower levels. It is desirable to find a more specific link between these factors, and the focus of this chapter is an analysis to uncover a pattern at the segment level between badly behaving modems and segment stability.

5.1 Segment Bad Modem Interest Measure

This section introduces a measure specifically designed to compare feature values from groups of modems of different sizes on equal terms. When comparing different sized segments in terms of the number of bad modems or number of WSRs, the total count or average per customer are not good measures. Instead, a measure of *interest* which takes into account both group size and feature count, is used. This uses the average proportion of bad

modems over all plants to produce a degree of interest due to unusually high concentrations of bad modems in a segment. The goal is to assign high interest scores to segments that are both large and have high counts of modems, without relying on hard thresholding techniques.

Figure 5.1 shows the number of bad modems in each segment of all plants combined, plotted against segment size. It is clear that large and small segments may host a large proportion of bad modems. In general, the larger the segment, the larger the number of bad modems it contains. The interesting ones are those at the top right of the plot since they are large and also have more bad modems than one would expect based on the regular proportion. The interest measure will identify these segments without having to draw a strict threshold based on either behaviour or size.

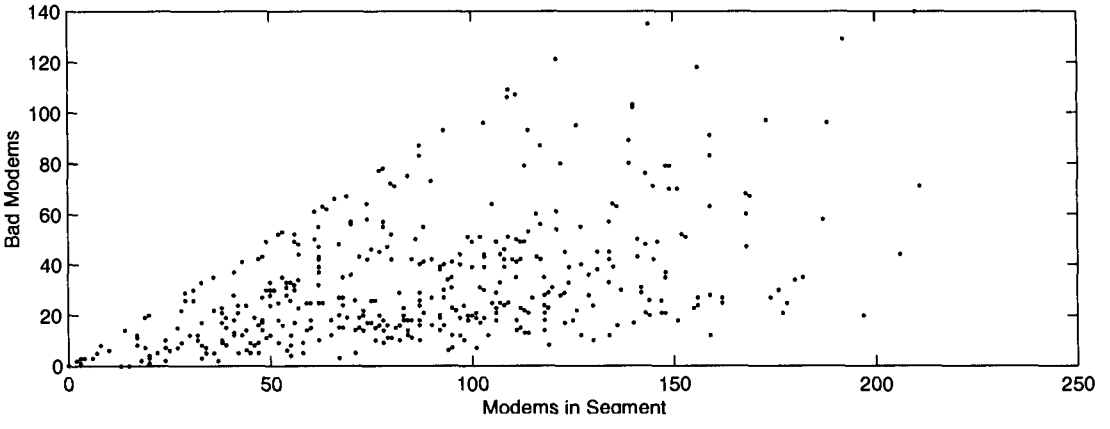

Figure 5.1. *Bad Modem Count vs Segment Size*

A segment-level bad modem interest score is computed using the modem behaviour classification. This figure takes into account the number of bad modems (with reliable data) in a segment, the size of the segment, and the overall proportion of bad modems from all plants.

5.1.1 Bad Modem Proportion

The interest measure requires a parameter that represents how common bad modems are in general, i.e. the proportion of all modems that are bad. This is computed simply by counting all the modems in the bad class of all plants, and the total count of modems in all plants. Again, only modems with valid data are considered. From table 4.10 the parameter is 0.372. Symbolically it is B/M, where B is the total number of bad modems in all plants while M is the overall total number of modems.

5.1.2 Interest Measure Calculation

With the global bad modem proportion, an interest measure is computed for each segment s based on its own bad modem proportion.

$$BMI_s = M_s log_2 \frac{B_s/M_s}{B/M} \tag{5.1}$$

Figure 5.2 shows the distribution of segment bad modem interest scores. Segments with infinitely negative interest (no bad modems) are shown at -400 to give an idea of proportions. There is a peak at zero interest, which corresponds to the majority of segments where the bad modem count is close to the global proportion.

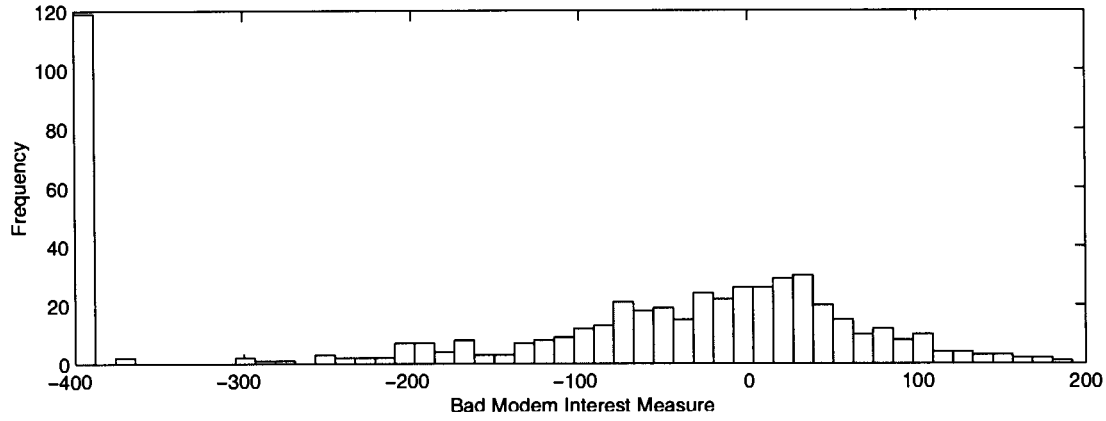

Figure 5.2. *Segment Bad Modem Interest Histogram*

For a segment with the average number of bad modems this metric gives zero interest, while segments with greater than average bad modem concentrations yield a positive interest. The larger and worse a segment is, the more interesting it is. Small bad segments are not very interesting because they are more likely to occur due to chance and would confuse a simple averaging metric.

5.2 Segment WSR Interest Measure

Using the segment stability figures representing the number of work service requests for each segment the WSR interest is computed in the same manner as the bad modem interest but using WSR counts in place of bad modem counts. WSR interest indicative of network performance as experienced by the customers and the more interesting segments are those which are large and also have a high WSR count for their size. In the subsequent sections, 5.2.1 and 5.2.2, the WSR interest measure is developed.

5.2.1 Global WSR Rate

For the interest measure, a value relative to the entire set of cable networks is required. In this case the factor is the one that represents the probability that a randomly chosen customer was the origin of a WSR. Each customer is considered equally, regardless of the size of the plant. Therefore the calculation is the total number of WSRs recorded for the April 2002 time frame over the total number of customers. This is also considered the rate that WSRs are generated per month across all plants. The exact numbers are not given due to their classified nature.

The plot in figure 5.3 shows the WSRs for each segment against the segment size, normalized against the maximum number of segment WSRs. Clearly some segments have more WSRs than is expected from the overall rate, and the WSR interest will be high for these segments.

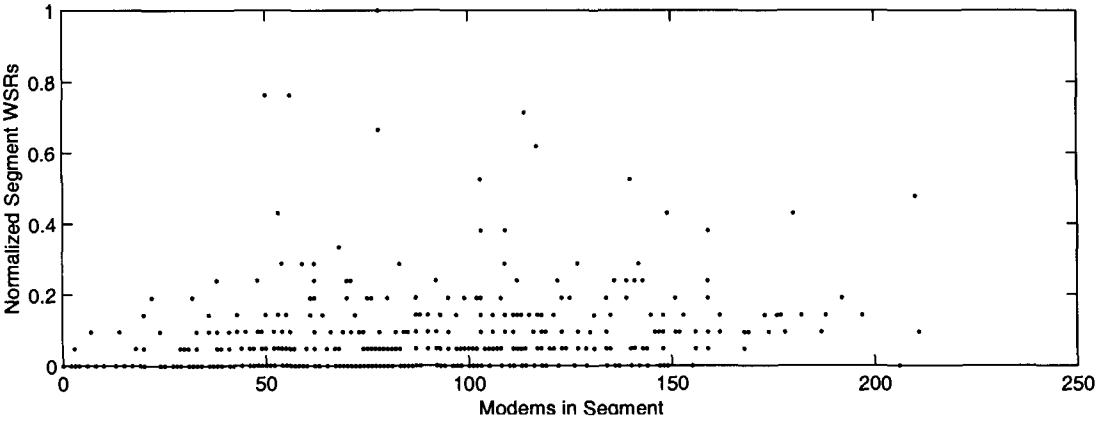

Figure 5.3. *WSR Count vs Segment Size*

5.2.2 Interest Measure Calculation

With the overall WSR and modem counts W and M, each segment WSR interest score WI_s is computed from the segment WSR and modem counts W_s and M_s as

$$WI_s = M_s log_2 \frac{W_s/M_s}{W/M} \tag{5.2}$$

The resulting distribution is shown in figure 5.4. Once again, segments with zero WSRs are shown at -400 instead of negative infinity.

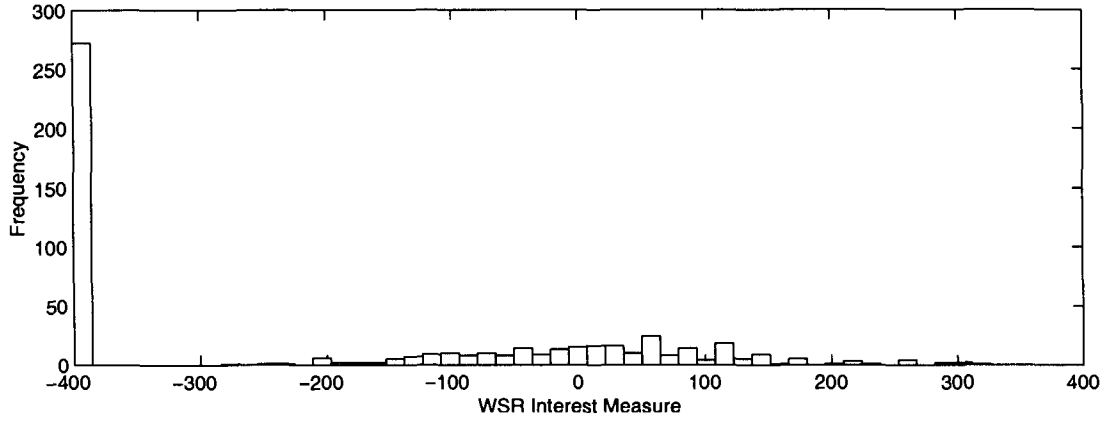

Figure 5.4. *Segment WSR Interest Histogram*

Some segments fall far to the right of the zero interest level. These are the ones that

produce many WSRs for their size and are also large.

5.3 Comparison Between Segment Bad Modem and WSR Interest Measures

The bad modem and WSR interest distributions provide measures of segment behaviour from two entirely different data sources. The histograms show that some segments stand out as unusually bad. These are the ones with both high concentrations of bad modems or WSRs and are also of significant size. This leads to the question of whether or not there is a relationship between a segment's two interest measures.

As shown in figure 5.5, the segments which are the most interesting in terms of WSRs (Region I) are also very interesting in terms of bad modems. Additionally, low power interest implies low WSR interest[1]. However the converse is not true, meaning that a large power interest does not imply a large WSR interest (Region III). This might be explained in that WSRs are generated only when a threshold of user inconvenience is exceeded. Also, the time scales of the stability data and modem data do not totally overlap, and there is an unknown delay between a malfunction and the rolling of a truck. If this is the reason, a high power interest is indeed a predictive metric that can be used to prevent further network deterioration.

This differentiation led to the observation that bad modems tend to cluster in the plants with poor stability, but not necessarily in the good plants. This might be explained by considering that network issues do not provoke customer complaints unless there is a significant loss of service or that network performance deteriorates rapidly.

[1] Segments with no bad modems are shown at -400 instead of $-\infty$ so that they may be seen. There is a very large concentration at the lower left.

5.4 Other Bad Modem Thresholds

In section 4.8.1, the bad modem MMSE threshold was set to 0.6 using an information theoretic strategy. This determined the bad modem interest scores for each segment. The correlation between segment WSR and bad modem interest scores (WI_s and BMI_s) is 0.2795, excluding the points at negative infinity. This is computed using a formula equivalent to 4.7. There is a question of whether the bad modem threshold could be set differently to enhance this correlation or better separate the bad segments. To investigate this, the segment interest comparison from the previous section was repeated for different bad modem threshold values.

Figure 5.6 shows the segment WSR to bad modem interest correlations for 20 bad modem thresholds between 0 and 1. The plot peaks at a threshold of 0.65, and the information theoretic derived threshold is very near that top value. The threshold selection strategy was successful in producing a threshold for modem classification without using segment WSR data. This is encouraging since these fault detection methods aimed to identify plant faults using the modem data sources alone.

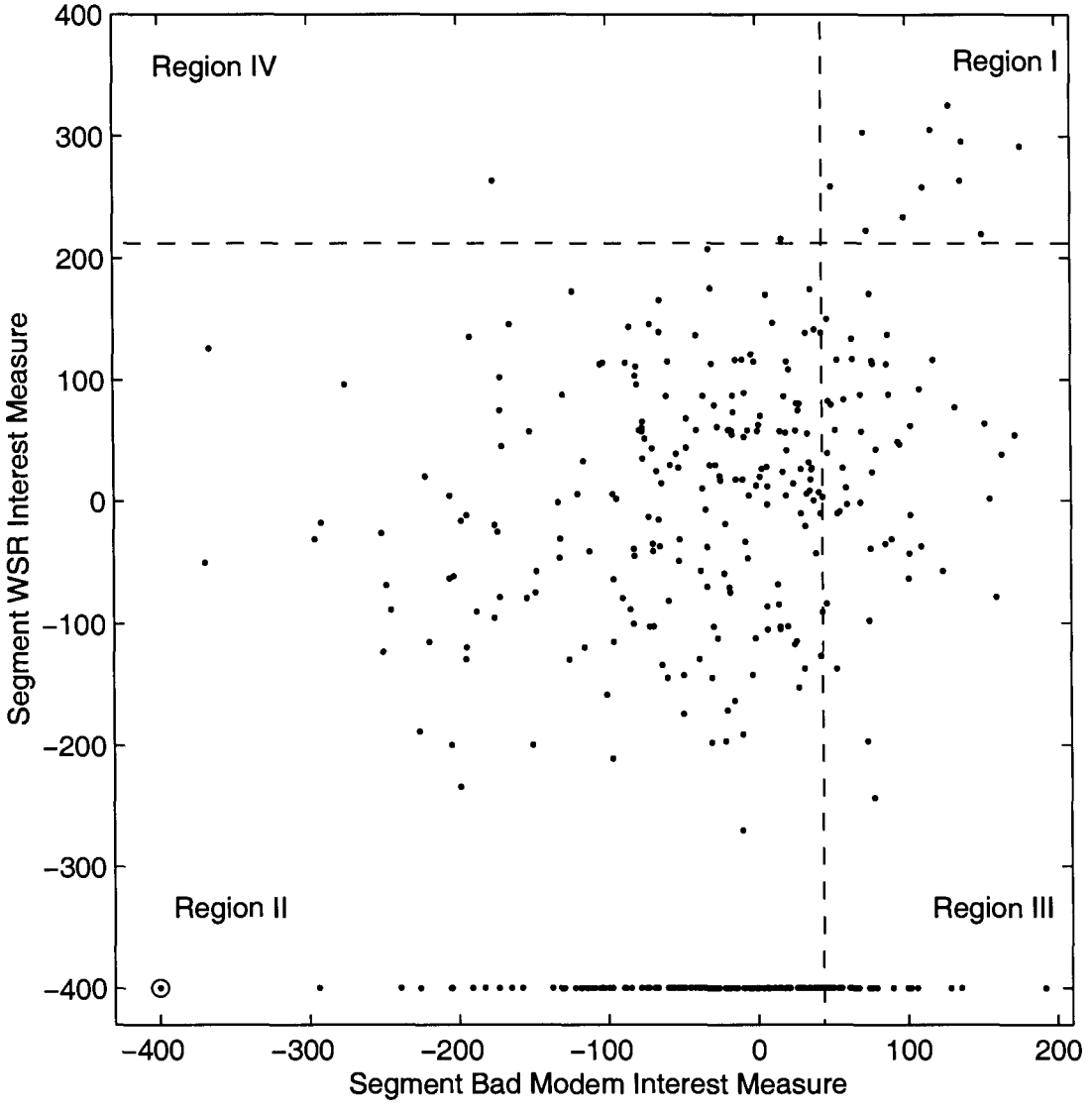

Figure 5.5. *Segment WSR Interest vs Bad Modem Interest*

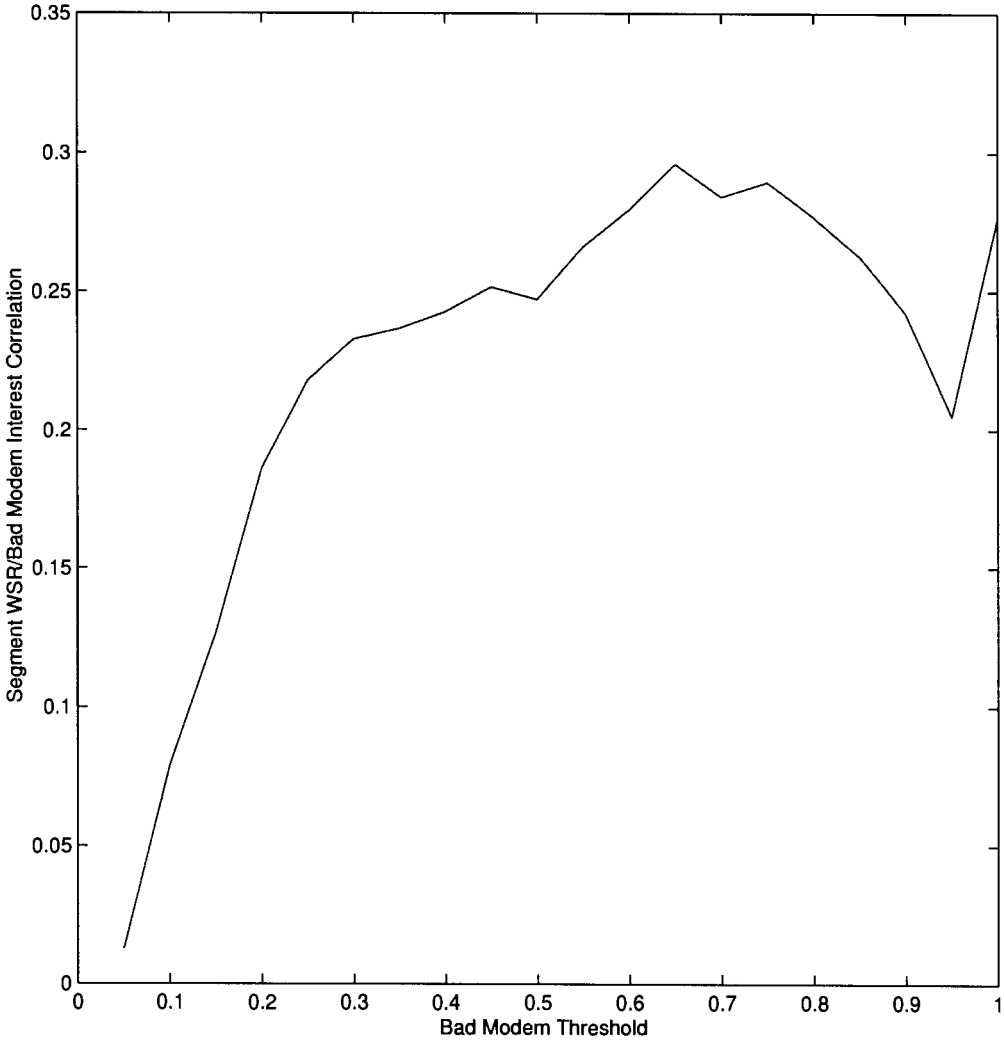

Figure 5.6. *Interest Correlation vs Bad Modem Threshold*

Chapter 6

Conclusions and Future Work

The cable modem networks investigated showed a variety of modem power status signal behaviours. The majority vary in accordance to plant temperature and this behaviour is accepted as the norm. Signal statistics and other features were extracted from the signals, providing informative yet convenient quantitative information for detailed analysis.

Processing of the modem data status signals posed several challenges. The frequent gaps in the data sources and the limited availability of some sources restricted the analysis time frame. The modem data signals contained many regions that did not reflect the true signal being measured. These problems were handled by selecting and performing analyses over the largest data sets available, and various methods were employed to detect invalid samples and clean the data prior to evaluation.

It was determined that a directly measured plant temperature signal may be replaced by an estimate derived from a large number of modem power signals with a high degree of accuracy. Individual modems are classified as having normal or abnormal behaviour using a calculated distance between the plant temperature estimate and each modem power status signal.

The clustering of pathological modem signals within the network could be suggestive of localized faults degrading the performance of the digital and analog transmission capabilities of the network, and might also be predictive of the component failure seen in related research. There is a correspondence between large abnormal modem concentrations and customer complaints at the plant level, and a relationship was also found at the segment

level.

The results show that plant health as measured by work service request (WSR) frequency is related to the modem power signal behaviour at the segment level (i.e. beneath a single head end cable modem). The discovered pattern showed that a high degree of WSRs in a segment implied a high number of poorly behaving modems. On the other hand, a high number of deviant modems did not necessarily imply a large number of WSRs. Since the worst WSR segments had high bad modem interests, it appears that the modem data does, through the power behaviour, reflect WSR inducing problems. In any case, this result helps narrow the search for faults within the network by identifying segments with high bad modem interests, which are more likely to cause customer disturbances.

There are several possible explanations for the discrepancy with segments with high modem power interests but low WSR interests. Some may have relatively new manifestations that have not persisted long enough to prompt customers to speak out. There are certainly different kinds of network faults, and although different kinds influence the modem signals, some may not significantly affect the subscriber. Although WSRs are an informative network measure, they are not a definitive measure.

A more detailed discrimination of plant health using WSR interest measures was not possible since the cable plant stability data provides only this level of detail. Similar techniques could be applied at a lower level, such as that of SMTs, given more information about the actual plant health. It is not certain if clusters of poorly behaving modems directly correspond to regions of the network with fault manifestations, although this investigation could be performed on the lower network levels using techniques described in this thesis and with more detailed stability data.

Further work may include an analysis to determine if regions of the network with high WSR levels had many poorly behaving modems in the months prior to the complaints. Segments with well behaving modems in earlier months might correspond to those without WSRs in the analyzed time period. This would help clarify the effectiveness of the technique as a predictive tool.

There is also a question of how well the interest measure scales between large and small segments. It may be worth dividing the segments into two groups, large and small, and developing two two different metrics based on different thresholds. Since interest scores from the bad modem counts in smaller segments are more prone to noise, perhaps focusing them separately would produce more telling results. Also, the threshold based bad modem classification presented in thesis ignores each modem's distance from the MMSE threshold. This information could be included in a modified interest measure calculation, which might prove to be more accurate.

Other techniques could be applied to search for patterns relating the segments bad modem and WSR characteristics. Neural networks have proven effective in classifying cable network elements in related research. Their ability to learn mapping from input and output data might be successful in separating the segments with many WSRs from those without based on the modem data features alone. This could provide a more accurate tool for identifying bad segments.

In conclusion, the intended outcomes of this thesis were reached. Methods of using operational status data collected from cable modem networks were developed that help identify regions of those networks with abnormal behaviour and customer complaints. Further investigations in to network fault detection using cable modem status signals are supported by the techniques developed in this thesis.

Bibliography

[1] T. F. Baldwin and D. S. McVoy, *Cable Communicatoin*, 1st ed. IEEE Press, 1986.

[2] D. J. Buerger, *Latex for Scientists and Engineers*, 1st ed. McGraw-Hill, 1990.

[3] K. T. Deschler, *Cable Television Technology*, 1st ed. McGraw-Hill, 1987.

[4] J. L. Devore, *Probability and Statistics for Engineering and the Sciences*, 3rd ed. Wadsworth, 1991.

[5] G. Donaldson and D. Jones, "Cable television broadband network architectures," *IEEE Communications*, pp. 122–126, June 2001.

[6] J. T. Dorocicz, "Asymptotitically stable recurrent neural networks: Theory and application," MASc thesis, University of Victoria, 1997.

[7] I. E. Frank and R. Todeshini, *The Data Analysis Handbook*, 1st ed. Elsevier Science B.V., 1994.

[8] R. G. Gallager, *Information Theory and Reliable Communication*, 1st ed. John Wiley & Sons, 1968.

[9] C. S. Hood and C. Ji, "Proactive network fault detection," 1997, pp. 1147–1155.

[10] I. Katzela and M. Schwartz, "Schemes for fault identification in communication networks," 1997.

[11] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Addison-Wesley, 1999.

[12] N. P. Kourounakis, "Improved fault detection in cable television networks," MASc thesis, University of Victoria, 1998.

[13] N. P. Kourounakis, S. W. Neville, and N. J. Dimopoulos, "A model based approach to fault detection for the reverse path of cable television networks," *IEEE Transactions on Broadcasting*, vol. 44, no. 4, pp. 478–487, December 1998.

[14] J. C. Madden, "Taming the *rogue*," *Cablecaster*, pp. 48–53,56, March 2000.

[15] S. W. Neville, "A prototype expert system based diagnostic tool for cable trunk amplifier networks," MASc thesis, University of Victoria, 1992.

[16] S. W. Neville, "Approaches for early fault detection in large scale engineering plants," Ph.D. dissertation, University of Victoria, 1998.

[17] R. Ng and R. Yee, "Teracomm data over cable access system characterization test results," Rogers IP Services Technology, Tech. Rep., May 2000.

[18] P. Z. J. Peebles, *Probability, Random Variables, and Random Signal Principles*, 3rd ed. McGraw-Hill Inc, 1993.

[19] L. Persons, N. J. Dimopoulos, K. F. Li, E. Manning, C. Somers, A. Schoorl, N. Kourounakis, S. Neville, A. Watkins, R. Glendenning, B. Anderson, D. Vyfhuis, R. Kovalik, and J. Madden, "How rogers cable systems improved network reliability with the university of victoria expert network analyzer," in *1999 Summer Cable Conference*, Vail, CO, July 1999.

[20] D. Picker, "Design considerations for a hyrid fiber coax high-speed data access network," 1996, pp. 45–50.

[21] A. P. Schoorl, N. P. Kourounakis, C. D. A. Somers, and N. J. Dimopoulos, "Using statistics and neural networks in fault determination," in *1999 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE'99)*, Edmonton, Alberta, May 1999.

[22] C. Somers, N. Dimopoulos, and S. Neville, "Cable network fault detection using cable modem status signals," in *2003 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, Victoria, B.C., August 2003.

[23] S. B. Weinstein, *Getting the Picture*, 1st ed. Prentice Hall, 1983.

[24] C. Westphal and T. Blaxton, *Data Mining Solutions: Methods and Tools for Solving Real-World Problems*, 1st ed. John Wiley & Sons, Inc., 1998.

Appendix A

Plant Temperature Estimation from SMT Temperature Data

This section describes a method of estimating a cable plant's temperature from the temperature readings from status monitoring transponders (SMTs) in the case of each cable trunk amplifier. The method was developed by Dr. Stephen Neville and has not been significantly modified.

Earlier cable network fault detection systems relied heavily on the SMT temperature data. When the SMT data was no longer supplied, a replacement temperature estimate was needed. This section details how the original temperature estimate from the SMT data was computed. Plant temperature estimates generated using this method are the reference temperature signals used in section 4.4.

The plant temperature estimation from SMT temperature signals consists of two steps. First, the SMTs are classified into above and below ground units. Second, a set of above ground amplifier temperature signals is combined into a single temperature estimate.

A.1 SMT Classification

SMT temperature signals vary with the outside temperature of the plant, in addition to factors which influence their individual temperatures such as heat dissipation from electrical currents. Above ground SMT temperatures vary more significantly than those below

ground as the earth acts as an insulator from temperature change. Although it is not indicated in the SMT data whether an amplifier is above or below ground, this property is inferred from the SMT temperature signal. Figure A.1 shows an above ground SMT temperature signal.

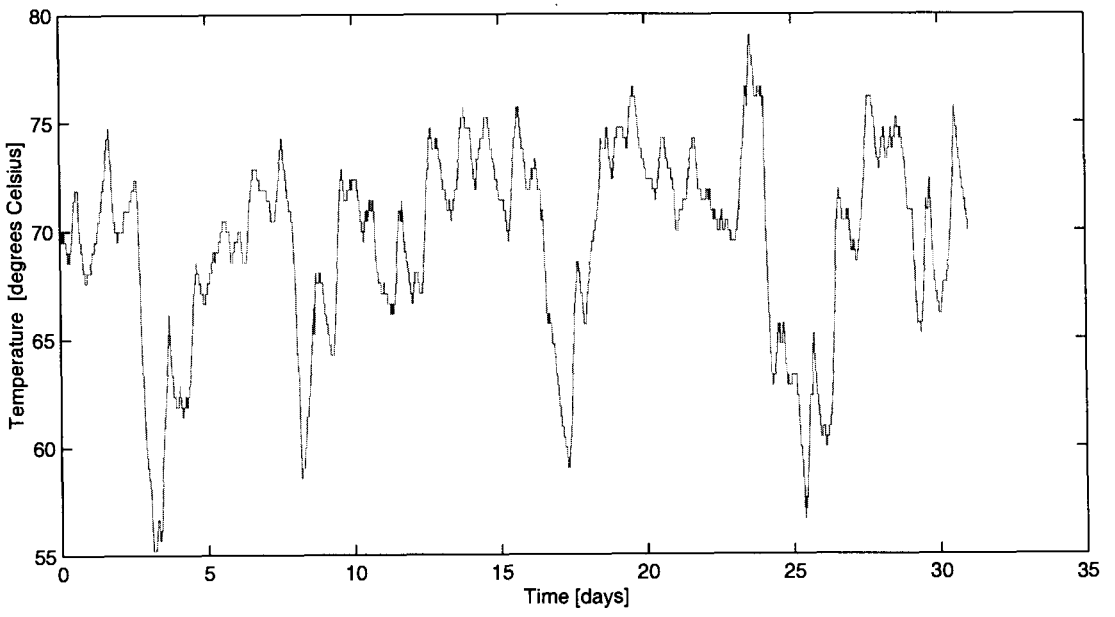

Figure A.1. *An Above Ground SMT Temperature Signal*

The SMT temperature standard deviations shows a clear separation between two classes of signals. Those with high standard deviations are above ground amplifiers, and the others are below ground. This is understandable since above ground signals are more indicative of the ambient temperature swings because they are not sheltered by the ground. Figure A.2 shows a histogram of SMT temperature signal standard deviations for one plant. These are the 485 SMTs from Plant R for which data was available in the January 17 to February 17 2001 time frame.

An experimentally determined threshold on the standard deviation of the temperature signal is used to classify each SMT. For reliability, the chosen threshold must be effective for different locations plants and for different times of the year, since the average tempera-

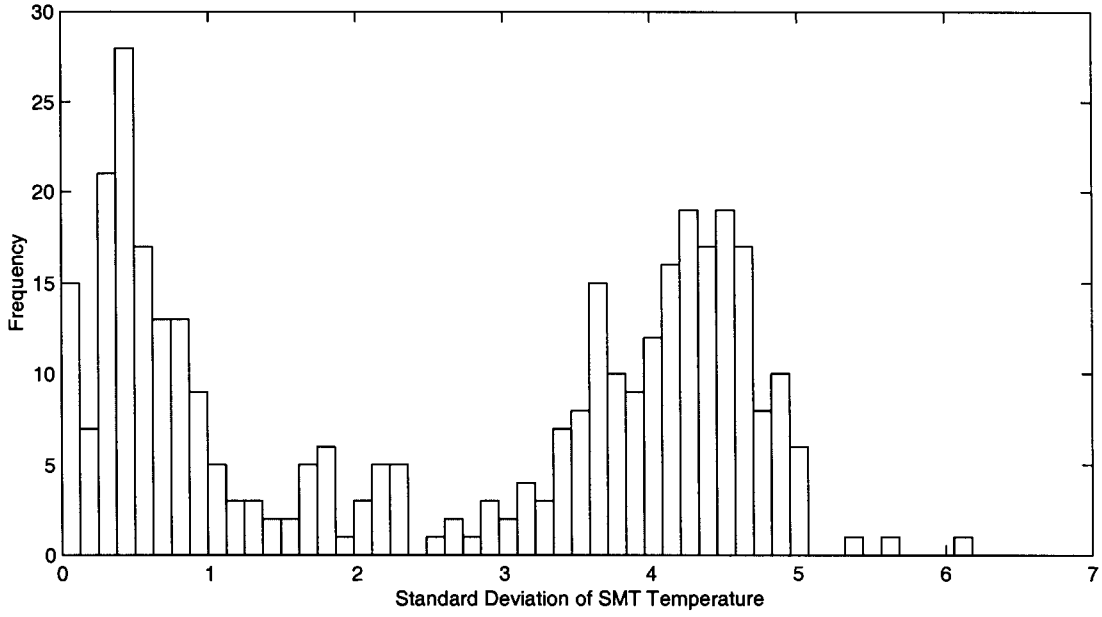

Figure A.2. *Histogram of SMT Temperature Signal Standard Deviations*

ture and temperature variation will vary. A good value for this threshold was found to be a standard deviation of three degrees Celsius. If the standard deviation of an SMT temperature signal is greater than the threshold, it is considered to be above ground, otherwise it is considered to lie below ground.

A.2 Signal Selection, Preprocessing, and Estimation

Only above ground SMT temperature signals with 80% or more of the maximum number of samples present in any temperature signal are considered. From this set, 20 signals are chosen at random. This number was verified to be sufficient for consistency in a separate analysis which is not covered in this thesis.

The SMT temperature signals do not require much preprocessing, unlike the modem power signals, since they are not host to many signal artifacts. Prior to combination, the temperature signals are simply filtered of zeros, which do not represent the actual temper-

ature. The 20 signals are resampled to a common time base so that their sample values may be averaged. Sample times that are not present for all 20 modems are dropped from the final estimate. This is a side effect of the method implementation (the modem power derived estimate in section 4.4 avoids this problem).

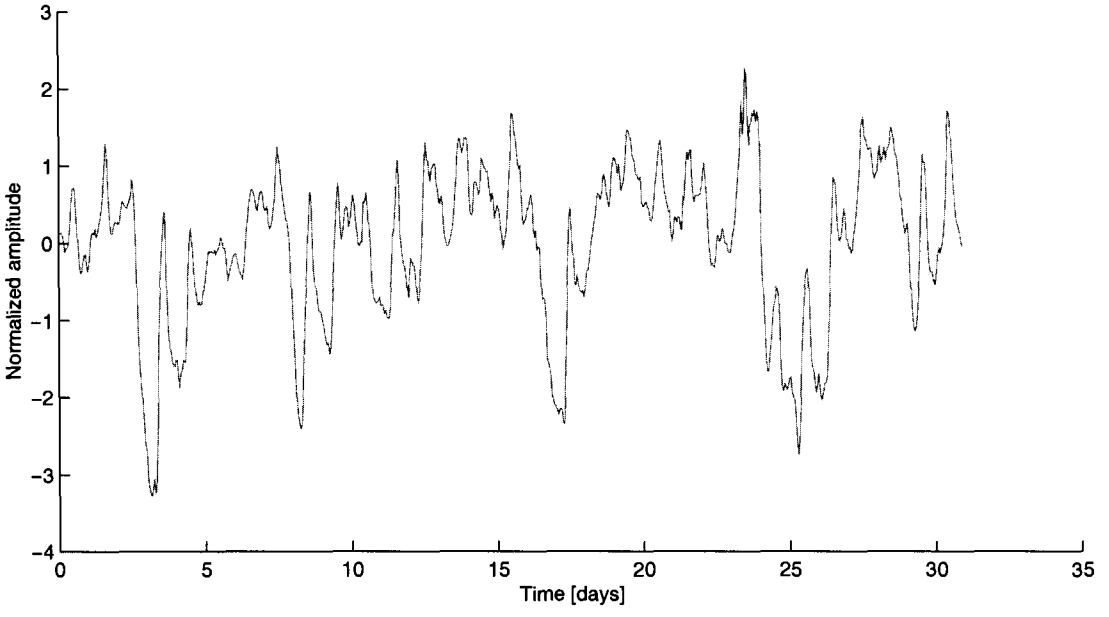

Figure A.3. *Plant Temperature Estimate*

The plant temperature estimate is taken as the sample by sample average of these resampled and filtered above ground SMT temperature signals. The resulting signal for Plant R, January 17 to February 17 2001, is shown in figure A.3. Although this signal strongly resembles the single SMT temperature signal in figure A.1, it is more representative of the true plant temperature since it is generated from 20 independent sources.

Appendix B

Feature Analysis

The features described in chapter 4 provided different views into the data which are more amenable to analysis than the modem data signals themselves. In an attempt to find high level patterns that relate certain modem behaviours to others as well as to features of plant health the features for the various data sources are generated and compared.

The primary feature comparison method used is the cross correlation matrix, which attempts to make obvious any correlation between any two features. In addition to the modem level features, the stability data outlined in chapter 3 provides a set of stability features at the segment level. The following describes how these features are compared with the modem data to suit a cross correlation analysis such as that described in [24].

B.1 Higher Level Features

Modem features defined earlier in this thesis are used to characterize properties of cable modems which lie at the lowest level of the network topology. Features for higher level network components are needed if fault detection analysis is to be applied at larger scales. Without high certainty of what different modem behaviours imply, the perspective from higher network levels may be definitive. The following sections describe how these features are obtained.

B.1.1 SMT Level Features

SMT features characterize the network at the level of individual cable trunk amplifiers. The modem data sources contain topological information but no signal information at this level. There is also no network stability information available at this level. SMT features are therefore generated indirectly using the modems which they feed and their modem level features. Still, a view of the network at this level may expose abnormalities that are not visible from any other level. Topology inferred from the modem data is used to determine which modems belong to which SMTs.

Each feature at the modem level is averaged to make a corresponding feature at the SMT level. Additional SMT features are obtained by taking standard deviations of modem features, taking a total modem count, and correlating modem power and CRC signals. Some maximum and minimum feature values from SMT modems are used.

B.1.2 SHUB Level Features

Again, data is not directly available at the SHUB level. Features describing network SHUBs are generated using averages of those features for all the modems within the SHUB. Averages of features for SMTs in a SHUB are not used because different SMTs have different modem counts and thus modems would not have equal representation in the SHUB average. Topology information from the modem data is used. Modem counts and signal correlations are also calculated.

B.1.3 Segment Level Features

Both modem and stability data can be used at the segment level. Modem data derived features are calculated for segments using aggregate values for the segment modems. The segment modem populations are identified through the stability data SHUB lists, combined with the modem topology SHUB to modem mapping. Stability features are provided at the segment level in the data sources, so these lead naturally to segment level feature values.

One consideration however is the fixed time frames over which these features apply. For comparison purposes, ranges of modem data are chosen to match segment stability time frames as closely as possible. For example, when using the 5 week WSR count from the stability data which ends April 30th, the modem data from March 24th to April 30th should be used to extract the corresponding modem features.

B.1.4 Plant Level Features

Plant level modem data features are generated in the same manner as lower level aggregate features, and features such as total number of modems, SMTs, etc are counted. Stability WSRs are averaged over the number of modems, and segment counts are totalled.

B.1.5 Multi-Plant Level Features

Some analyses call for features derived from data combined from multiple plants over an analysis time frame. Methods for feature calculation are as they are for lower level aggregate features. Multi-plant level features are useful for assessing plants in relation to the norm established over all plants. For example, a particularly troublesome plant is identified as one with a high WSR per customer ratio. Features at this level are generated from such a large source of data they may to some extent be considered "normal" feature values, and are useful for comparing to plant level features. When possible, all 18 plants are used.

B.2 Correlation Analysis

The method of cross correlation analysis takes two feature matrices and correlates each vector in the first matrix with each vector in the second matrix. The result is a matrix of correlation coefficients called the cross correlation matrix. Features which correlate highly imply that there is a relationship in the data between those features, while features that have

zero correlation are independent.

The two argument matrices are comprised of feature vectors for all the elements at some level of the plant topology. For example, a feature matrix consisting of the 15 modem features from section 4.6 for a plant with 6000 modems is a 6000 by 15 matrix, where each column is a modem feature vector and each row is a vector of a single feature across all modems.

A correlation analysis was performed using the modem and stability data. The feature matrices generated from the modem data were modem level features, SMT level features, SHUB level features, segment level features, and plant level features. Features generated from the stability data were segment level features, plant level features, and singleton SHUB level features. Segments normally feed several SHUBs, but when there is only one SHUB in a segment it is referred to as a singleton SHUB. These are important SHUBs because the stability data for that segment is known to have come from that SHUB. Normally it is not possible to see stability data for individual SHUBs.

B.3 Summary of Results

This high level analysis revealed that there was indeed a relationship between plant stability and modem MMSE levels. This is a most promising relationship in that it confirms what was expected from the outset, that there is some predictive power from comparing abnormally behaving modems to plant stability. This promising pattern does not hold up at the segment levels through the cross correlation analysis. Further analysis at the segment level is needed to reveal a more focused predictive mechanism relating these behaviours. That is the subject of chapter 5.