Integrated Shift and Drift Control of a non-linear growth process

Soham Chakraborty

May, 2019

(Revised May, 2020)

Thesis submitted to the Indian Statistical Institute in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy in Quality, Reliability and Operations Research, 2018.

Indian Statistical Institute

203, B.T. Road, Kolkata – 108

West Bengal, India

Dedicated to my parents, Satyabrata and Sanga, my beloved wife Upasana and especially to my thesis advisor Dr. Pathik Mandal without whose help and support this would not have been possible.

Acknowledgements

Firstly, I would like to thank Indian Statistical Institute for this incredible opportunity. It has been a long, hard and invaluable journey here at ISI for the last seven years. I am indebted to ISI for providing me the research fellowship as well as a vibrant and wonderful atmosphere for my work. During my stay here at ISI, I was fortunate enough to know many faculty members, students, scholars and even office bearers who showed me the value of being a warm and compassionate human being, above everything else. So many have touched my life that it would be futile to thank every one of them individually. Instead, I would like to express my thanks and gratitude to every single one of them.

I am especially biased towards the SQC & OR Division at ISI where I spent the bulk of my time here. The faculty members and the staff have been nothing but generous with me. I owe them a lot. I would like to thank *Prof. Ashis Chakraborty*, *Prof. Arup Mukherjee*, *Prof. Prasun Das*, *Prof. Arup Das* and *Prof. Biswabrata Pradhan* for their support and guidance whenever I asked for. *Prof. Ashis Chakraborty* and *Prof. Arup Mukherjee* gave me their invaluable advice during the most difficult and trying phases of my research. I would also like to mention my co-researchers Ritwik, Moutushi, Sonal and Tanujit who were my default support system here at ISI. I wish all of them a happy and successful future.

I owe a lot to my thesis advisor *Dr. Pathik Mandal*. This surely would not have been possible without his help and determination. I hope I have done enough to earn his respect. My parents and my sister were always there for me. It has been a long, hard ride for them too. Lastly, I would like to thank my wife. She stuck by me during the most difficult days and always gave me hope when I needed it most.

Soham Chakraborty

Published papers from the thesis

- [with Mandal, P.] *SPC based on growth models for monitoring the process of hydrogenation of edible oil*. Journal of Food Engineering 146 (2015) 192-203. (**Relevant to Chapters 2, 4 and 5**)
- [with Mandal, P.] *Variance function estimation using signal-to-noise ratio*. IAPQR Transactions Vol.43 No.2 178-211. (**Relevant to Chapter 3**)
- [with Mandal, P.] *Control charts for detection of a dead batch during hydrogenation of edible oil.* International Journal of Quality & Reliability Management, 36(2019) 1804-1820 (**Relevant to Chapters 4,5 and 6**)
- [with Mandal, P.] *Stochastic optimization for controlling the end state of hydrogenation of edible oil in the presence of process failure*. International Journal of Operations Research, Vol. 16, No. 4, 91-104. (**Relevant to Chapter 6**)

List of Figures

List of Tables

Table of Contents

List of Notations

- Y_t is time indexed stochastic process
- D_t and S_t are deterministic and stochastic components of process drift
- T is the most preferred end state of the characteristic under regulation
- n_{max} is the maximum number of samples per batch
- f and g are the conditional mean and variance function with parameters β and θ respectively
- MP_f is the final melting point of a batch
- MP_H is the melting point at the end of the end of the hydrogenation cycle
- $\Delta MP_{cooling}$ is the growth in melting point during the cooling cycle
- IV is the iodine value of oil
- MP_{eq} is the equilibrium melting point
- MP_{in} is the initial melting point
- k is the growth rate during hydrogenation
- MP_{ij} is the melting point of the i^{th} sample drawn at time t_i from the j^{th} batch
- e,r,u & ε all denotes error terms
- δ is the autocorrelation coefficient
- $Z=MP(t+a)$ is the linear transformation of the monomolecular model
- a and b' are the parameters of the linear model
- d is the growth rate (per hour) of Z
- $G|(MP_t, \Delta t)$ is the growth of melting point for a given MP_t and sampling interval Δt
- d_0, d_1 are the parameters of the biased mean model
- I is the instrumental variable for correcting bias

- a_0 , a_1 are parameters of the 1st stage mean model
- b_0 ,b_1 are parameters of the 2nd stage mean model
- t_0 is the time of fresh catalyst addition
- Q_1 is the control statistic for monitoring an active batch
- Q_2 is the control statistic for detecting dead batch using the 2nd chart
- MP_C is the cooling limit
- NS is the number of samples till the end of cooling
- C_P is the path cost and C_E is the end cost
- t_d is the failure time of a batch
- FT is the failure time
- RT is the recovery time
- lag is the time required before the result of a sample becomes available
- λ is the quality loss coefficient
- k_1 , k_2 are shrinkage factors for policy P1 and P2 respectively
- m_0 , m_1 and m_2 are parameters of policy P2
- SN is the signal-to-noise ratio
- η is the sample signal-to-noise ratio
- μ and σ are the process mean and standard deviation respectively

Chapter 1 Introduction

In the field of Quality Control (QC), the term 'control' is used to mean many different things. For example, it may refer to an isolated control action (e.g. a process adjustment), a particular type of control architecture (e.g. feedback vs. feed forward control) or even a control discipline (e.g. Deming's PDCA cycle). So will be the case here. But the exact meaning will be clear from the context.

Since the development of control chart by Shewhart [1] and the method of acceptance sampling by Dodge and Romig [2] during 1920's - 1940's, the field of quality control has been greatly enriched by the development of many new principles and methods (in particular, the statistical methods for QC). Notable among these are the following:

(i) Development of the new philosophy of Japanese management during 1940 - 1950 [3];

(ii) Development of the concept of Total Quality Control (TQC) by Feigenbaum [4] in late 1940;

(iii) Widespread use of the simple tools of QC from 1950 onwards and the pioneering efforts of Deming [5], Juran [6] and Ishikawa [7] in achieving the same;

(iv) Development of the concept of robust design and the statistical methods for achieving the same by Taguchi (see ref. [8]) during 1970's and

(v) Recent popularization of the Six Sigma breakthrough strategy for process improvement by Harry (see ref. [9]).

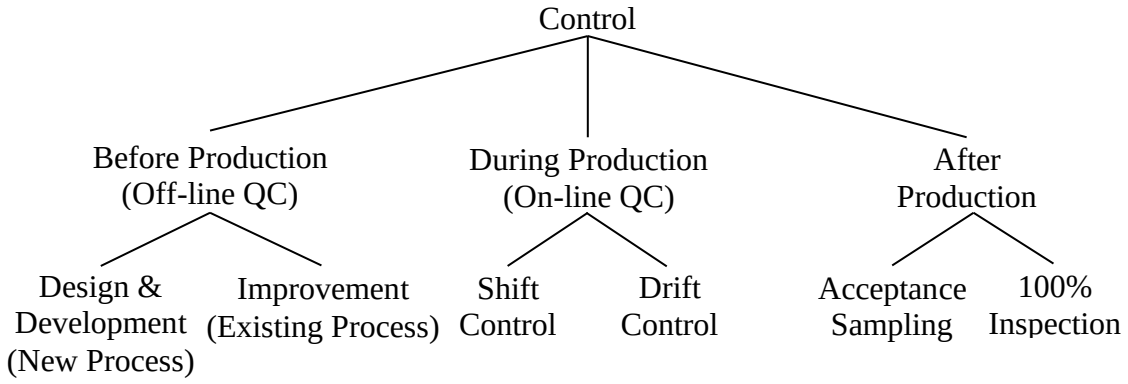
Of course, the above efforts have been greatly supported by many researchers and practitioners from academia and industry. In what follows, there will be many opportunities to have a look at some of these contributions, particularly in the area of Statistical Process Control (SPC).

1.1. Classification of control activities

The QC tasks performed at various stages of product/service realization before shipping can be broadly classified as shown in Figure 1.1. It should however be noted that all the six activities are not of equal importance. There is an inherent hierarchy among them. Anticipating customers' usage conditions of a product and making it robust so that it performs well under all such conditions are obviously more important activities than the activity of inspection for product acceptance/rejection. It is now widely recognized that the former has

got much higher potential to influence customer satisfaction than the latter (which, actually, is aimed at minimizing customer dissatisfaction). In fact, one can say that the control objective of each of these three stages of product realization is to reduce or even eliminate the need for control at the downstream.

Figure 1.1. Classification of control activities



This work is concerned essentially with the two control activities at the middle of Figure 1.1, i.e. shift control and drift control. These two terms are not unknown to the QC community at large. But here they will be used in a slightly different fashion as explained in the next section.

1.2. Terminology

1.2.1. Shift and drift

The core of SPC is control charting for detecting process abnormalities or 'shifts'. The presence of shifts refers to any type of process behaviour, which is different from the expected behaviour under an in-control state of the process. It is to be noted that ‘shift’ does not represent any particular type of variation. The presence of shift represents an abnormal state of the process, which may be manifested in the form of a simple change in process mean or variance or a complete change in the dynamics of the process. Also the effect of shift may be either sporadic or it may persist over time (chronic).

Drift is usually defined as a systematic and slow variation. However, there is lack of clarity over the meaning of the word 'slow' in the definition. Here we shall interpret it as smooth to distinguish drift from jump variation. However, ‘jump’ and ‘shift’ are not synonymous. As mentioned above, shift does not represent a particular type of variation, but

a jump refers to a sudden change in process mean. The time points at which jumps occur may be either fixed (e.g. seasonal effect) or random (e.g. jump diffusion model).

1.2.2. Stochastic and deterministic drift

Let a time indexed stochastic process $\{Y_t, 0 \le t < \infty\}$ be given by $Y_t = D_t + S_t$, where D_t is the deterministic component and the stochastic component S_t is mean stationary, i.e. $E(S_t)$ is constant. It is assumed that there is no shift in the process, i.e. Y_t has a predictable distribution. If the mean of the process varies systematically over time, then it will be said that the process has a deterministic component of drift. In time series literature, a non-periodic deterministic drift is called a secular trend. On the other hand, if the elements of S_t are correlated then it will be recognized as a stochastic component of drift. Of course, if D_t is constant and S_t is white noise then it is a drift-free pure random process.

1.2.3. Signal factor

The deviation of a process from its specified target due to drift is corrected by adjusting the levels of one or more judiciously selected process variables. In control engineering literature, these variables are called 'manipulated variables'. However, here a manipulated variable will be referred to as a 'signal factor', i.e. the factor that is used to realize the adjustment signal.

1.2.4. Integrated shift and drift control

Drift control is widely used in many areas of engineering, particularly for controlling continuous processes. However, of late, the QC people have also begun to take active interest in this area, although the first edition of the books by Box and Jenkins [10] and Åström [11] appeared way back in 1970. The primary motivation for the renewed interest has been the desire to integrate the two approaches to have better control over a process. Although all the integrated systems which have been developed so far have the same basic elements, the drift control part of the proposed integrated schemes has been given different names by different researchers, e.g. Automatic Process Control (APC) [12], Engineering Process Control (EPC) [13] and Statistical Process Adjustment (SPA) [14]. Thus, we have the names like ‘combined SPC and APC’ or ‘combined SPC and EPC’ for the integrated approach. In fact, the integrated version is also known as Algorithmic Statistical Process Control (ASPC) [15].

Each of the above names highlights a particular aspect of drift control. For example, the name EPC highlights the engineering origin of the approach. The name APC highlights the fact that most drift control procedures are automated. However, none of the above terminologies will be used here. Instead, we shall call it as ‘Integrated Shift and Drift Control’. However, the term SPC is retained (because of its long history) and it will be used interchangeably with 'shift control'.

1.2.5. Regulation

In this work, the control of a deterministic drift will be called regulation and it will be used to mean control of deterministic drift only. However, the term control will be used in the context of both deterministic and stochastic drift.

1.2.6. Active vs. passive regulation

If the control action taken at a time point influences the subsequent evolution of the process then it will be called an act of active regulation. In contrast, the growth path is not influenced by passive regulation. The growth process is regulated by drawing samples periodically and/or terminating the process at the right time solely for the purpose of controlling the end-state of the process.

1.3. Shift control vs. drift control

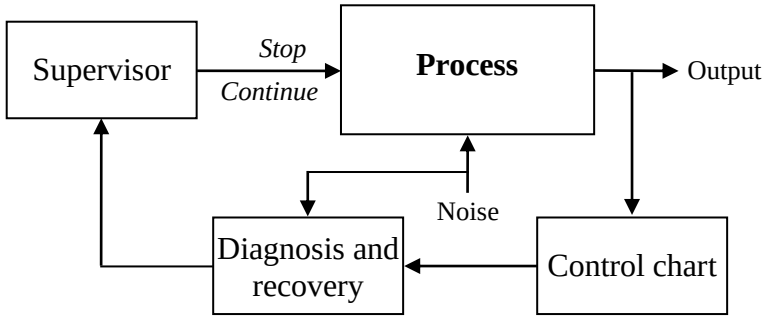
Both shift control and drift control have the same objective of reducing the variation in the process output. However, there is an important difference between the two aspects at the strategic level. While the former takes a medium term view of the process, the latter takes a short term view. Consequently, the tasks and the associated tools of the two approaches are also vastly different from each other. It is emphasized here that the primary objective of shift control is to achieve continuous improvement by identifying and removing the root causes of the process shifts. However, it is obvious that the root cause analysis cannot go very deep when the process is on-line. It is also permissible to use process correction as an interim measure till the cause is identified. There can also be a failure to identify the cause. However, if such failures become too frequent then a relook at the control chart becomes necessary.

In contrast, process correction is the fundamental mode of drift control. In fact, drift control may be viewed as a mechanism of transferring the variation in process output to the signal factor. It should however be noted that drift control in the presence of process shifts

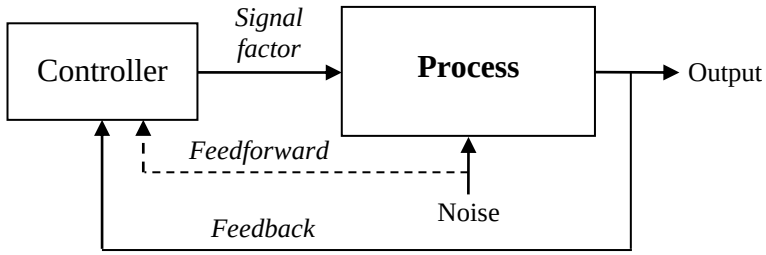
may not be very effective. For example, the process corrections made against sporadic shifts may actually increase rather than reduce output variation.

The operational difference between shift and drift control is shown in Figure 1.2. It may be noted that the shift control works only in the feedback mode, while drift control can be implemented in various modes (e.g. feedback, feedforward). Another difference between the two is in the length of the control cycle (time interval between two consecutive control decisions). In case of shift control, it usually varies from a few hours to a day, but the same for drift control may be as low as a few seconds.

Figure 1.2. Shift control vs. drift control



(2a) Shift control



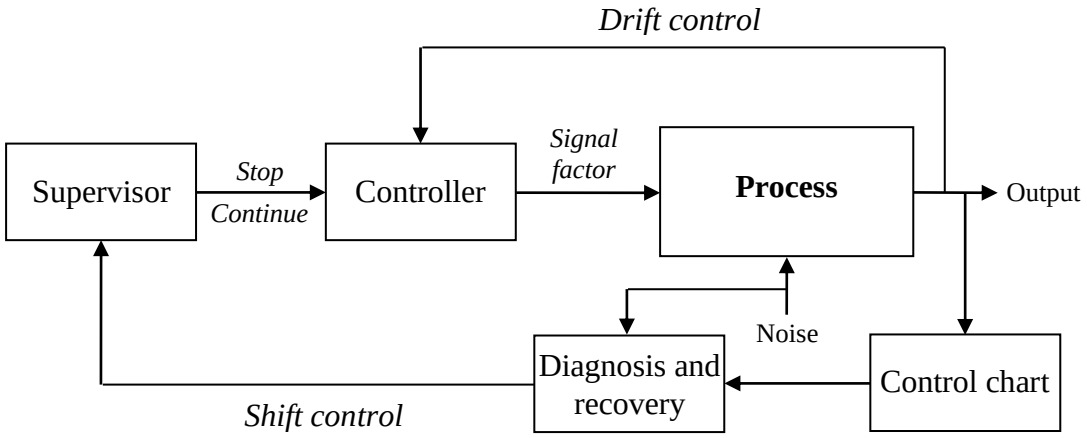
(2b) Drift control

1.4. Integrated shift and drift control: An illustrative example

The basic idea behind the integrated approach is to superimpose shift control over a process that is under drift control. For the purpose of shift control, one may monitor (using control charts) either the closed loop output of the process or the levels of the signal factor or both. The architecture of an integrated scheme is shown in Figure 1.3.

To illustrate the integrated approach, consider the following example:

Figure 1.3. Integrated shift and drift control with monitoring of the controlled output



Tyre pressure example: The tyre pressure of an automobile is measured once in every three days and cannot be measured in between. If the tyre deflates sufficiently, then the automobile remains idle till the next measurement. Assume that the initial pressure is on target. It is expected that the pressure will fall gradually (may be very slowly) over time. So, at every measurement, the pressure is adjusted to target by injecting appropriate amount of air.

Clearly, the above process is under pure drift control. Now, it is easy to recognize that occasionally the valve in the tube may not function properly or some pin hole may develop within the tube (either due to manufacturing defect or a nail piercing the tube). If the tyre always gets deflated quickly (say within a few minutes) due to the occurrence of such assignable causes, then the control system can be improved only by reducing the cost of recovery. But let us assume that on many occasions, it takes quite a few days of continuous leakage of air before it becomes non-functional. So what happens under such a scenario if the process is only under drift control? It is obvious that progressively more and more air needs to be injected to compensate for the leakage and ultimately, the tyre will fail. Since failure can occur at any point of time between two measurements, there will be a significant cost of failure, increased cost of refilling and perhaps also poor performance of the automobile for a few hours. Is there a better way to control the process?

Tyre pressure example (contd.): A dynamic model is developed for predicting the one step ahead tyre pressure. Based on this model, a suitable control chart is developed for monitoring the difference in pressure between the predicted and the observed value. If the same, i.e. (predicted pressure – observed pressure) is found to be above the upper control limit of the chart, then the wheel is examined thoroughly, repaired if needed and then the tube is refilled such that the pressure is on target. Otherwise, the pressure is adjusted as before and the cycle continues.

Note that the above is an integrated procedure where the drift control action is the adjustment of pressure and the shift control action is the repair of the tyre, if needed. Assuming that the control chart has been designed properly, the above integrated approach should be expected to perform better since the control chart acts like an early warning system. Also note that since the amount of air filled will be progressively higher after a leak develops, an appropriate control chart may be developed for monitoring the amount of air injected at every adjustment.

In the above, the diagnosis and recovery actions are nearly classical shift control actions. However, this may not always be the case. It can happen that the recovery action is not preceded by a diagnosis.

Tyre pressure example (contd.): Since the cost of diagnosis and repair is found to be very high, it is decided that whenever the chart gives an out-of-control signal (above UCL), the tyre will be examined for the presence of any foreign objects and the old tube will be replaced by a new defect free tube.

Can the above remedial action be considered as a valid shift control action? The answer is yes, since the process is expected to be restored back to its in-control state. In one of the two control charts proposed in this study, the remedial action is similar to the above (replacement of tube).

1.5. Role of the control chart and its integration in an integrated system

It is now widely recognized that making process adjustments for correcting small deviations from the target may not be advisable. In fact, a drift control scheme where no corrections are made within a specified band around the target has been called a ‘dead band’ policy by Box and Jenkins [16]. Taguchi's beta-correction factor [17] for process adjustment also has this feature. It is thus entirely conceivable to design an adaptive drift controller where the parameters of the controller are updated based on the signal received from a control chart. In fact, in the early days of SPC, Barnard [18] viewed the control chart primarily as a tool for estimating the process mean. Surely, there cannot be any objection to such uses of a control chart. But then it is also important to maintain the true spirit of shift control as noted in Section 1.3 for exploiting its full potential.

Depending on the role of the control chart, the integrated schemes that have been proposed so far can broadly be classified into two groups - one in which the control chart is used to select an appropriate mode of drift control [19, 20] and the other in which the chart is used

for monitoring the drift controlled process [15, 21]. Thus, in the first case, the control chart (shift controller) and the drift controller are functionally integrated at the level of system design. The two components can be integrated further at the level of parameter design by optimizing all the parameters of the control system (including the control limits) together. On the other hand, if the control chart is used for monitoring a drift controlled process then obviously there is no integration at the level of system design. However, the two components can be integrated at the level of parameter design [21].

It appears that integrated systems where the control chart is structurally integrated as in the first group and at the same time retains the true spirit of shift control as in the second group are yet to be developed.

It is obvious that integrating the two components is important for optimal performance of the system. But it is far more important to have both the controllers developed by the same team. The kind of process knowledge the team is expected to gain from modelling the process will be extremely useful in designing the control chart. Unfortunately, the QC practitioners are often in an unenviable situation, where the shift control system is to be designed for a process that is already under drift control (designed by somebody else). Thus, the drift control system becomes a part of the black box for the shift control system, which makes the task of shift control very difficult.

1.6. Growth process control

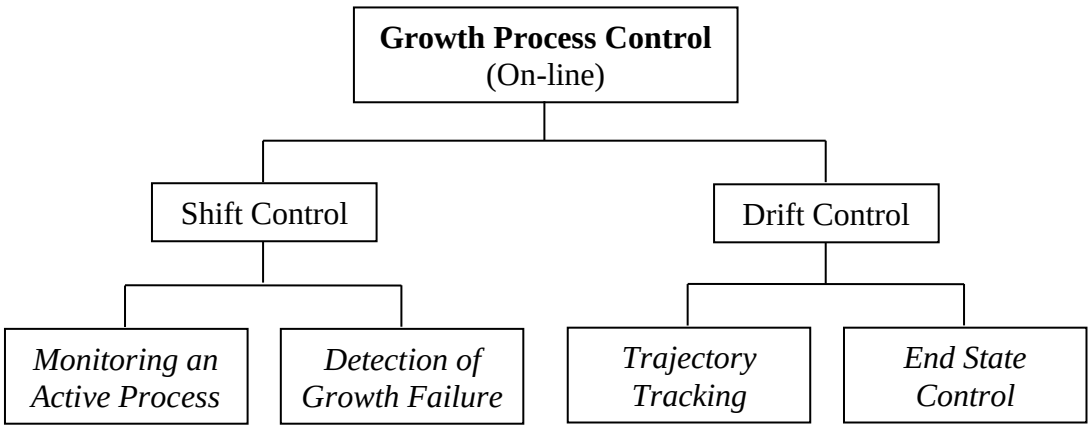
Let y_t be the observed value of an output characteristic of a process measured at time t. The characteristic evolves over time in such a manner that y_t may be represented as $y_t = f(t) + \varepsilon_t$, where ε is the random error term. We shall call the process $\{y_t\}$, a growth process. Thus, in accordance with the definition of deterministic drift given in section 1.2.2, it is seen that a growth process is characterised by the presence of a deterministic component of drift. In general, the growth process $\{y_t\}$ can be either discrete or continuous. However, in this work we shall consider only the continuous case.

It is assumed that the growth process operates in cycles, i.e. there is a beginning and an end of the process. The end point may be the one in which the process is either physically terminated or deemed to have been terminated since no control actions are required beyond this point. Assuming further that f is monotonic in the zone of practical interest, i.e. within the growth cycle, the four main tasks of growth process control can be classified as follows: (i) Monitoring an active process, (ii) Detection of growth failure, (iii) Trajectory tracking and

(iv) End state control. The classification scheme is shown in Figure 1.4. The task of trajectory tracking refers to regulation of the growth trajectory so that y_t remains as close as possible to a pre-specified reference trajectory. However, a growth process is not usually subjected to trajectory tracking. This is primarily due to the difficulty of translating the optimal process performance to a particular growth trajectory. Nevertheless, attempts have been made in the past to first specify a reference path and then to regulate the process to track the specified path, e.g. controlling the growth of broiler chicken [22] and the temperature profile (which is not the growth characteristic but a process variable) of a crystal growth process [23].

On the other hand, in the end state control, the focus is only on minimizing the deviation of the final value y_t, $t > 0$ from its specified target. However, if the process generates many cycles over time then in addition to the above four cycle specific tasks, it will also be necessary to monitor the overall health of the process using characteristics like the cycle time and the number of growth failures over a period.

Figure 1.4. Classification of growth process control tasks



Of course, it may not be necessary to have formal procedures for performing all the four tasks in each case. For example, if the occurrence of growth failure is obvious or if the failures occur only rarely with little cost implications, then one may not have a formal procedure for on-line detection of growth failure. Similarly, one may not resort to trajectory tracking, if the same is thought to be impracticable at the present state of process knowledge. In fact, some growth processes may not be subjected to drift control at all. To illustrate, let us consider the example of human fetal growth. In this case, the growth path can be customized, i.e. the expected growth path for an individual can be estimated with an acceptable band of error [24]. But no regulatory actions are taken to correct the deviations from the expected

path. Further, the end state is specified only in terms of an estimated due date and a variation of ± 2 weeks from the estimate is considered as normal. This is clearly a case of no regulation. However, shift control (monitoring using control charts) can play an important role here in deciding the need for medical interventions, if any.

This work is concerned with developing methods for performing three of the above four tasks (excluding trajectory tracking). The methods of trajectory tracking are not discussed since the same are found to be unsuitable for the growth process considered in this work (see Section 1.7). In what follows, we give an overview of the approaches for performing these three tasks.

1.6.1. Growth process monitoring

Let us assume that the growth process generates many cycles over time and the process is expected to grow in the same fashion within each cycle. Under such a situation, depending on the current status of the process, one may choose to monitor either the within or between cycle variation. If the cycle time of the process is very high, then the former approach becomes a necessity.

A growth curve can be considered as a special type of (process) 'profile'. Thus, the between cycle variation can be monitored following the well-known methods of ‘profile monitoring’, provided the data are available in a suitable format. Once a profile is found to be out-of-control then it may be examined to see if any particular region (or a particular parameter of the function describing the profile) is affected and planning for trouble shooting can be done accordingly.

Presently, profile monitoring is a very well-developed field, with numerous publications. However, according to Woodall [25], the type of problems considered in some of these publications, may not qualify as belonging to the class of ‘profile monitoring’. An overview of a few of the methods for profile monitoring is given in Chapter 5. The relevant references are also cited therein.

However, the number of published applications on monitoring the within cycle variation of a growth process is very limited. Recently Shore et al. [24] and Zhu et al. [26] have proposed control charts for monitoring the within cycle variation. Both the above studies make use of a set of initial observations to obtain an initial estimate of the parameters of the model. These estimates are then sequentially updated as new data arrive(s).

A significant part of this study is also concerned with monitoring within cycle variation. However, a major difficulty that has been faced in developing an appropriate procedure for

this purpose is the scanty nature of the data. For example, we have only 2-6 observations in each growth cycle, but the growth model selected for the process has three parameters. Clearly, the approach proposed by the previous authors cannot be adopted in the present case. In order to overcome this limitation, data from several cycles are pooled to estimate the parameters of the model. According to Woodall et al. [27], such an approach is permissible for defining the in-control state of a process, i.e. some amount of between profile variations may be included to define the common cause variation.

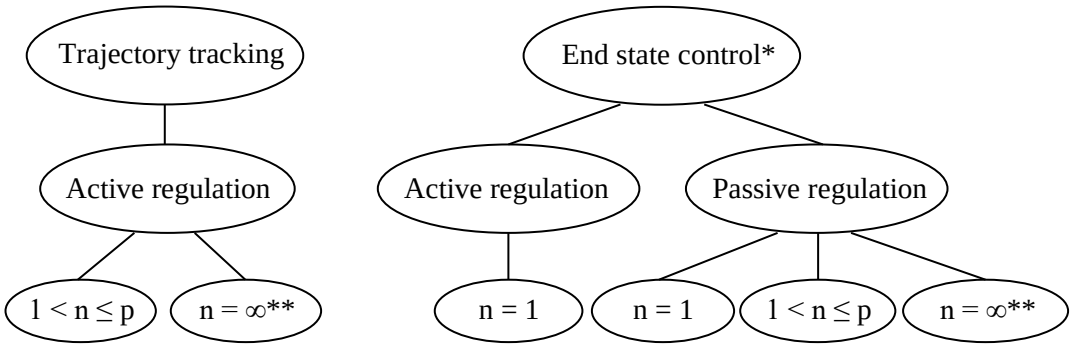
So far as detection of growth failure is concerned, it appears that this problem has received almost no attention in the past, except perhaps in one area, i.e. early detection of children having short stature. Grote et al [28] have proposed two sets of rules for two different age groups for detection of growth failure and subsequent referral for further diagnosis of the cause of suspected growth failure. However, the details of the methods used for arriving at the rules are not given by these authors. In this study, two control charts are proposed for detection of growth failure. One of these two charts is recommended for implementation based on a comparison of their power in detecting growth failure. It is interesting to note that the primary rule for referral for short stature [28] as mentioned above is similar to the one proposed here.

1.6.2. Growth process regulation

The drift of a growth process may be subjected to either active or passive regulation (as defined in section 1.2.6). This work however is not concerned with active regulation, whether the same is used for trajectory tracking or for end state control. Rather, the objective is to control the end state through passive regulation.

Depending on the kind of measurements available for the purpose of control, the methods of passive regulation can broadly be classified into three groups as shown in Figure 1.5. The control of coating thickness at electroplating is an example of passive regulation with $n = 1$, i.e. when only the end state of the process can be evaluated off-line at the end of a cycle. It is well known that the coating thickness increases with the duration of electroplating and the same has been found to be true for a variety of substrate-coat pairs [29-30]. Thus, one can develop a regression model of the type $y_t = f(t)$, and compute the coating time needed to achieve the target thickness. Once the model is developed, the same may be updated periodically using the latest measurements.

Figure 1.5. Approaches for growth process regulation



* Combined active and passive regulation is feasible (see section 6.5.7)
** n = Number of sampling points; Continuous measurement using on-line sensors

The third case represents situations where continuous measurement of the process is possible, e.g. in-situ real-time measurements for detection of the end point of a plasma etching process [31]. Usually the ability to monitor a process continuously in real-time is considered as a sign of technological progress.

In between the above two, we have processes where the status of the process can be evaluated based on samples drawn periodically from the process, but there may be a restriction on the maximum number of samples (say p) that can be drawn from each growth cycle. Such a control strategy is widely used in industry, agriculture and also in everyday life. For example, farmers decide on the date of harvesting based on periodic inspection. In manual cooking of rice, the end point is decided based on periodic sampling. In LD converter-based steel making, samples are drawn periodically to control the phosphorus content of steel. However, it is somewhat surprising that this particular control problem has received no attention in the past despite its great practical significance. No published work in this area has come to our notice while conducting a literature survey.

1.6.3. Integrating shift control and passive regulation

Integrating shift control of an active process with passive regulation based on periodic sampling is comparatively simpler than integrating the same with active regulation. Since the process path is not affected under passive regulation, the control chart for monitoring an active process and the passive regulator can be developed independently provided economic considerations are not involved. However, the control chart for detection of growth failure and the periodic sampling based passive regulator cannot be designed independently. This is

for two reasons. First, the performance of the control chart will depend on the sampling interval. The smaller the interval, the lower will be the probability of detection of failure. Secondly, it may be necessary to have more conservative rule for regulating the process post its recovery from growth failure, since the uncertainty associated with the dynamics of the process during this period is likely to be relatively higher.

1.7. Hydrogenation of edible oil: the motivating and working example

The melting point of a batch of edible oil rises gradually over time during its hydrogenation. Figure 1.6 shows two typical growth curves along with the important features of the growth process. The specifications for final melting point are 40.5 ± 0.5^{0}C. The upper specification limit of 41^{0}C is the statutory limit. The nominal or the target value and the lower specification limit are established internally by the company keeping the shelf-life of the end product and the capability of the hydrogenation process in mind.

The growth of melting point is monitored by drawing samples periodically from the batch under hydrogenation. As can be seen from Figure 1.6, the number of such samples drawn within a growth cycle is very limited. As soon as the observed melting point is found to be $\geq 40^0 C,$ the hydrogenation cycle is stopped, and the cooling cycle begins. The melting point of the batch also increases slightly during cooling. But no control can be exercised over the growth during cooling. Thus, the growth process considered here is a two-stage process where the first stage refers to the active hydrogenation phase and cooling is the second phase.

Figure 1.6. Two typical growth curves

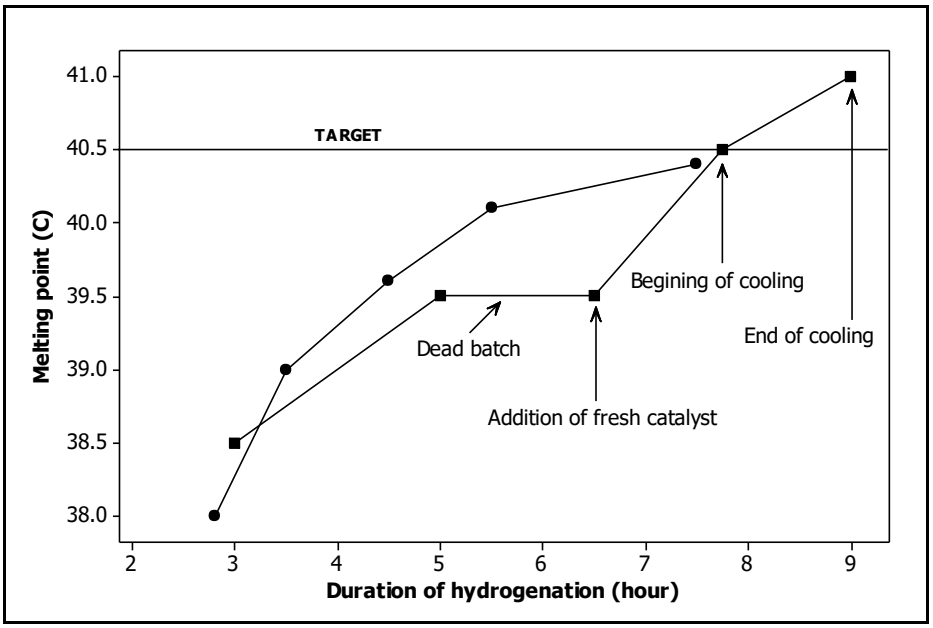


Figure 1.6 also shows that a batch, during hydrogenation, can become dead at a random point of time. In the existing system, the detection of a dead batch depends on the subjective judgment of the production supervisor. If a batch is judged to be dead, then it is recovered by addition of a certain amount of fresh catalyst.

The objective is to develop an integrated shift and drift control system for a two-stage growth process as shown in Figure 1.6. It may be noted that the integrated scheme should have modules for performing the following three tasks:

(i) Monitoring the growth of melting point in each batch till it remains active.

(ii) Efficient detection of a dead batch, and

(iii) Economic control of the final melting point around the specified target of $40.5^{o}C$.

So far as the integration of the three modules is concerned, the following approach is adopted. First, the control chart that is developed for detection of a dead batch is integrated with the drift control scheme to develop an economic scheme for controlling the final melting point in the presence of batch failure. It is to be noted that the hydrogenation process as above is a two-stage process. Thus, in addition to finding the optimal sequence of sampling intervals $\{\Delta t_i\}$, it is also necessary to determine the cooling limits, i.e. the melting point at which the cooling phase should begin. There are many constraints in the system like the measurement lag and the maximum number of samples that can be drawn in each cycle. This particular problem is formulated as a Stochastic Dynamic Program (SDP) and is solved following the method of Approximate Dynamic Programming (ADP). The optimized scheme thus obtained is then integrated with the control chart for monitoring an active batch to obtain the overall integrated scheme.

It is important to note here that the above example not only provided the motivation for this study, the same is also used as a working example for both developing and illustrating the proposed methodology. However, it will be seen that the proposed methodologies are fairly general and can easily be adopted do deal with other similar situations, particularly when a growth process is to be controlled based on periodic sampling. In fact, in other situations, the development of the control scheme is expected to be relatively easier if the process is a single-stage process (unlike the two-stage hydrogenation process as above).

1.8. Growth models

Let the growth characteristic of interest be denoted by Y, which evolves over time t. The basic growth model that is used for modeling the process is of the following form:

$$Y_t = f(t,\beta) + \varepsilon_t, \sigma_t^2 = g(t,\theta),$$

where f and g are the conditional mean and variance function with parameters β and θ respectively, i.e. $E(Y_t \mid t) = \mu_t$ and $V(Y_t \mid t) = \sigma_t^2$. It is to be noted that the error variance is not assumed as constant. This is because heteroscedasticity is common in applied regression. Moreover, it is noted below that the growth model to be considered here is non-linear in nature (with decreasing growth rate over time). Thus, one should expect the variance to decrease over time.

It is shown in Chapter 4 of this thesis that for the hydrogenation process, the growth of melting point (*MP*) during hydrogenation can be adequately described by the following well-known mono-molecular model [32, 33],

$$MP_t = MP_{eq} - (MP_{eq} - MP_{in})\exp(-kt), \tag{1.1}$$

where MP_{in} and MP_{eq} are respectively the initial and equilibrium melting point and k is the rate constant.

The static model (1.1) is used for developing the control chart for monitoring an active batch. However, it is not useful either for detection of growth failure or passive regulation. Since growth failure is a local phenomenon that is independent of the location of the process, it is necessary to have a dynamic model for this purpose. It is also obvious that a static model cannot be used for determining the sampling intervals. To illustrate, let T be the process target and the time to reach the target, as obtained from (1.1) be t_s. However, due to the noise present in the system, there will be situations when the observed value of *MP* at t_s is much less than T. Consequently, the time for the next sample after t_s becomes indeterminate. This also shows that a dynamic model is needed for the purpose of passive regulation.

It is also shown in Chapter 4 that the dynamic model corresponding to the static model (1.1) is given by

$$\begin{aligned} MP_{t+\Delta t} &= MP_t + G, \\ G &= (d_0)(\Delta t) - (d_1)(MP_t)(\Delta t), \end{aligned} \tag{1.2}$$

where G is the growth in the time interval Δt and d_0 and d_1 are model coefficients. The statistical model specified for estimating G is

$$G_i = (d_0)(\Delta t_i) - (d_1)(MP_i)(\Delta t_i) + \varepsilon_i, \tag{1.3}$$

where G_i is the growth in the i_{th} interval of length Δt_i and MP_i is the melting point at the beginning of the i_{th} interval. It is assumed that the error term ε_i is independently distributed as

$N(0, \sigma_i^2)$. The variance function specified for estimating σ_i^2 is of the form

$$\sigma_i^2 = g(MP_i, \Delta t_i, t_i - t_0), \tag{1.4}$$

where t_0 is the time of first catalyst addition.

1.9. Challenges in the growth process control

In course of the development of the control scheme for the hydrogenation process, the two main challenges faced are as follows: (i) Defining the in-control state of the process for monitoring an active batch and (ii) Developing an appropriate dynamic growth model for detection of growth failure and end-state control.

It is well known that the in-control state is specified by the location and chance-cause variation of the process. In case of a growth process, the location should be defined by the growth path of the optimized process since the same is unlikely to be customer specified. Although it may be possible to specify an optimal reference path in certain cases, such instances are rare since, in general, it is a difficult task. So, assuming that the reference path is not available, it becomes necessary to establish such a path as objectively as possible for the purpose of process monitoring. How can we achieve the same based on a sample of *k* (say, 50) growth curves? The other problem is related to establishing the chance cause variation. Should the same be estimated based on the variation of individual curves or a group of curves? In the above discussion, it is mentioned that a group of curves is used since only a limited number of observations are available in each growth cycle. However, this question is more fundamental and needs to be answered even when sufficient number of observations is available in each growth curve (discussed in more detail in Chapter 5). Assuming that it is decided to make use of a group of curves for estimating the chance cause variation but the observed variation in the sample curves is too large, then the problem becomes one of classification of a set of growth curves. However, depending on the nature of the data, this classification problem may at times be extremely difficult. No formal and appropriate method may be available for this purpose. In the present case, an informal graphical method has been used for classification of the sample growth curves (Chapter 4).

The challenges faced in developing the dynamic growth model are two-fold. First, the process variation is found to be non-constant but the available data for estimating the variance function is limited with inadequate number of near repeat points. A survey of the existing literature and the initial modeling effort in course of this work reveals that the existing methods for variance function estimation based on un-grouped data [34] are highly

problematic when the sample size is small. In particular, the variances are known to be highly underestimated. Driven by this limitation, a new approach is proposed for estimation of variance function based on signal-to-noise ratio (Chapter 3).

The second problem faced during modeling is in obtaining consistent estimates of growth within a given interval. It is easy to see from (1.2) that the growth over a longer interval is larger than the growth when the same interval is split into smaller intervals, i.e.

$$G_{\Delta t} \mid MP_t \geq (G_{\Delta t_1} \mid MP_t) + (G_{\Delta t_2} \mid MP_{t+\Delta t_1}),\ \Delta t_1 + \Delta t_2 = \Delta t, \tag{1.5}$$

where the equality holds only when $\Delta t_1 = \Delta t_2 = 0$. This makes the growth path dependent on the sampling scheme, which is clearly not desirable. It may be noted that most growth processes are non-linear and hence similar problem will arise in all such cases. This is the well-known problem associated with developing a continuous time dynamic model based on discrete time data. An innovative method of instrumental variable estimation is proposed here for reducing the bias as above (Chapter 4).

Other challenges faced in solving the passive regulation problem (Chapter 6) are primarily due to the fact that it is an unknown territory. So, it requires careful consideration of the requirements of the process while formulating the dynamic program and selecting the policy function for its solution.

1.10. Main contributions

The main contributions of this work are as follows:

I. A novel approach is proposed for variance function estimation based on SN ratio. It is shown that the variances can be estimated more accurately from the proposed bias corrected SN function than those obtained directly from an estimate of the corresponding variance function.

II. It is shown that the growth of melting point during hydrogenation can be described adequately by the monomolecular model.

III. An innovative method of instrumental variable estimation is proposed for developing a continuous time dynamic model based on discrete time data. It is shown that the proposed method provides reasonably consistent estimates of growth (from a given state over a given duration) within the working range.

IV. The proposed control chart for detection of growth failure is shown to be much more effective and efficient compared to the existing procedure that relies on subjective judgment of the process supervisor.

V. The problem of optimal end state control of a growth process in the presence of growth failure when the process is monitored through periodic sampling is formulated as a SDP (Stochastic Dynamic Programming). A simulation based methodology is proposed for solving the SDP. A comparison of the existing performance of the hydrogenation process with that expected under the proposed optimal scheme shows that the total cost will reduce by about 39% from the existing level. Besides, the average cycle time is also expected to reduce by more than half an hour.

1.11. Organization of the remainder of the thesis

It is almost customary to have a chapter on literature review in a PhD thesis. However, since the topics covered here is somewhat broad (Growth models - Variance function estimation - Shift control - Passive regulation), instead of having a separate chapter, the related work is reviewed at the beginning of each chapter dedicated to the above topics.

Next chapter (Chapter 2) gives a very brief account of the process of hydrogenation.

Chapter 3 gives a detailed account of the new method of variance function estimation.

Chapter 4 is devoted to the growth models which are used for developing the integrated scheme. As noted already, the results of the previous chapter are used extensively for developing the models.

Chapter 5 is about control charting. Different types of control charts for monitoring between cycle variation, within cycle variation and detection of growth failure are discussed in this chapter.

The economic passive regulation scheme, which is the main component of the proposed control scheme, is described in Chapter 6. The final integrated scheme is also presented in this chapter.

The main conclusions of this study are summarized in the concluding chapter (Chapter 7). The areas/directions for future research in this topic are also indicated here.

Chapter 2 Hydrogenation of edible oil

Cooking fat or what is known as 'vanaspati' in the Indian subcontinent is produced through partial hydrogenation of edible oil. The solid content (crystalline fat) of edible fats and oils increases due to hydrogenation. As a result, the resistivity of the oil to thermal and atmospheric oxidation is also improved. The melting point of oil is an indicator of its solid content and is a very important quality characteristic of the final product.

The remainder of this chapter gives a very brief account of the basics of hydrogenation along with a description of the industrial batch process of hydrogenation that is considered in this study. It is expected that the general discussion on hydrogenation will be helpful in having a better appreciation of the control system proposed for the process.

2.1. Chemistry

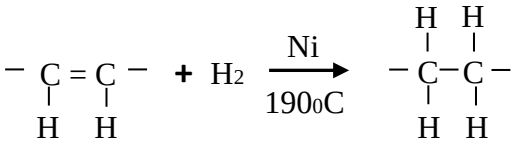
Dietary fats are usually classified as either saturated, monounsaturated or polyunsaturated, depending on the number of double bonds that is present in the fat's molecular structure. The saturated fat contains no double bond, while the mono and polyunsaturated fats contain one and more than one double bonds respectively.

Fatty acids are commonly designated as N: n, where N is the number of carbon atoms in the acid molecule and n is the number of C = C double bonds. In this study, we are primarily concerned with the following transformation through hydrogenation:

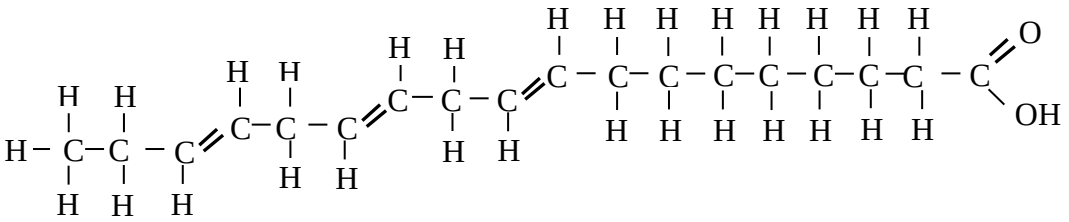
Linolenic (18 : 3) ⟶ Linoleic (18 : 2) ⟶ Oleic (18 : 1) ⟶ Stearic (18 : 0)

2.1.1. Chemical reaction

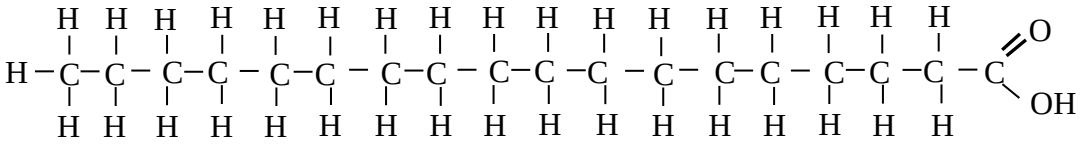
Hydrogenation reaction principally consists of addition of hydrogen atoms to the double bonds of various unsaturated fatty acids. The breakage of the double bonds is achieved in a heterogeneous catalyst system consisting of solid catalyst, liquid oil and gaseous hydrogen at an elevated temperature. The hydrogenation reaction may be represented as follows:



To take an example, the structure of a polyunsaturated fatty acid is transformed as shown below when the acid becomes saturated through hydrogenation:



Linolenic acid ($C_{18}H_{30}O_2$)



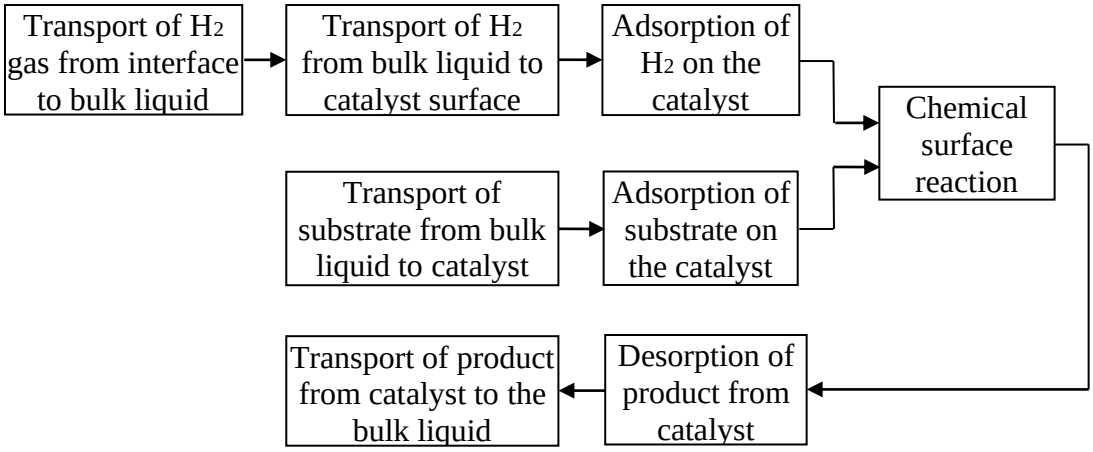
Stearic acid ($C_{18}H_{36}O_2$)

Note that the double bonds produce a bend in the fatty acid molecule, which prevents dense packing of the molecules. As a result, the oils and fats composed of unsaturated fatty acids melt at lower temperatures than those containing saturated fatty acids. The melting point of the saturated stearic acid is 69.6^0C, while that of linolenic acid is -11^0C. The melting points of the two intermediates 18:1 and 18:2 are 13.4^0C and -5^0C respectively. Further details on chemistry of fatty acid are available in [35].

2.1.2. Mechanism of catalytic hydrogenation

It is now commonly accepted that hydrogenation of unsaturated fatty acid in the presence of nickel catalyst takes place following the Horiuti-Polanyi mechanism [36]. According to this mechanism, both molecular hydrogen and the fatty acid with its double bond are first adsorbed onto the nickel surface. Next, the molecular hydrogen is dissociated into hydrogen atoms. One of these two atoms is first added to the fatty acid and forms a half hydrogenated intermediate. This step of the reaction is reversible and hence the intermediate can dissociate. However, if a second atom of hydrogen is added to the intermediate, the original double bond gets saturated and the transformation becomes virtually irreversible. The block diagram of the steps involved in catalytic hydrogenation as above is shown in Figure 2.1.

Figure 2.1. Mechanism of heterogeneous catalysis (Hydrogenation of carbon-carbon double bond)



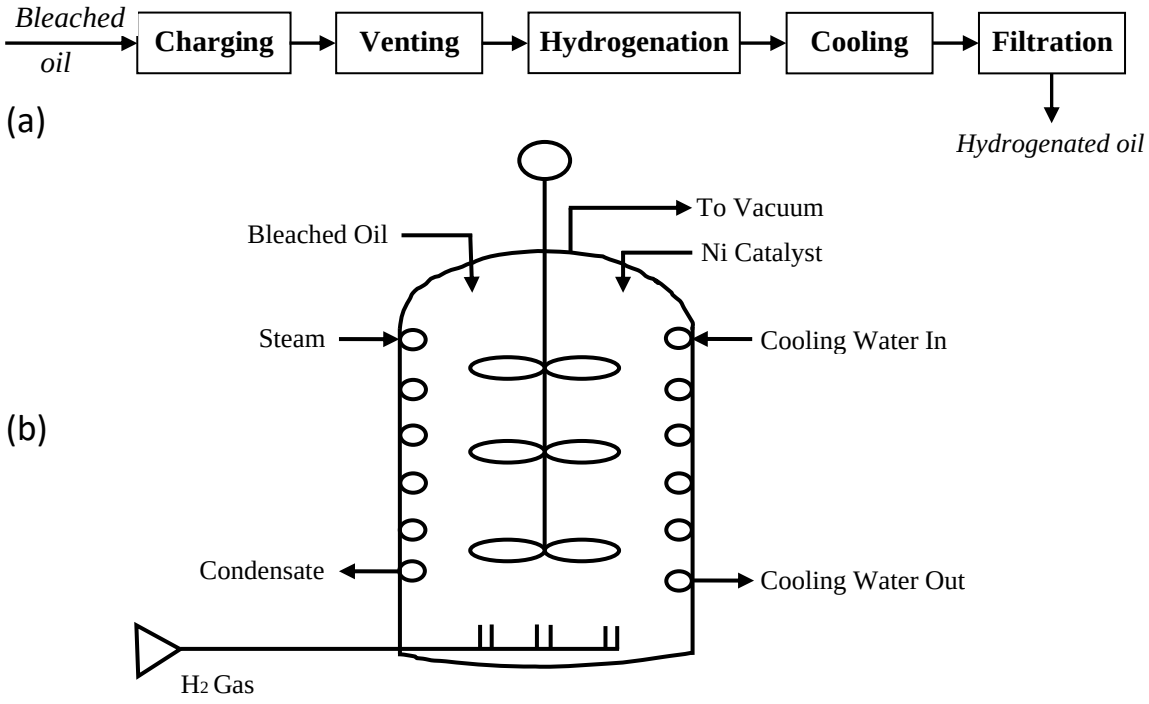
2.2. Industrial hydrogenation and its control

The basic process of hydrogenation of edible oil was invented by W. Normann [37] in 1902. Presently, many variants of the same are used for commercial production of cooking fat. Here, we are concerned with a batch process as shown in Figure 2.2a. First, a batch of bleached oil is charged into the reactor called autoclave (see Figure 2.2b). Then the vacuum pump is switched on and after attaining the desired pressure within the autoclave, the oil is slowly heated to raise its temperature to 165^0C. Next, the spent (recycled) catalyst is dozed into the autoclave and the hydrogen gas is injected through the bottom of the charge. Since the hydrogenation is an exothermic reaction, it becomes necessary to control the temperature of the charge within 170 - 210^0C by controlling the flow of cooling water. Usually, about one kilogram of fresh nickel catalyst is added to the charge after an hour of hydrogenation with the spent catalyst.

The process has two stages - active hydrogenation and cooling. The melting point of the oil increases as hydrogenation proceeds (as it should due to the conversion of polyunsaturated to monounsaturated/saturated acids). The specifications for final melting point, which is the most important quality characteristic that needs to be controlled, is 40.5 ± 0.5^0C. The upper specification limit of 41^0C is the statutory limit. The nominal or the target value and the lower specification limit are established internally by the company keeping the shelf-life of the end product and the capability of the hydrogenation process in mind. In order to control the final melting point, oil samples are drawn periodically from the process and its melting point is evaluated. If the observed melting point of the sample is found to be $\geq 40^0$C, then the

active hydrogenation cycle is immediately brought to an end and the cooling cycle begins. At the end of the cooling cycle, the cooled oil is filtered and sent for further processing. The melting point of the batch increases slightly also during the cooling cycle, but this increase cannot be controlled.

Figure 2.2. (a) Hydrogenation process flow chart and (b) sketch of the autoclave



Apart from controlling the final melting point, the sample measurements are also used to judge the status of a batch, since all the batches may not remain active throughout the hydrogenation cycle. Occasionally, a batch also becomes dead at a random time point. Such dead batches are recovered by addition of fresh catalyst (0.5 - 2 Kg).

It may be noted that if too many batches become dead or if a dead batch is not recovered quickly or if the growth rate of an active batch is not under control then the other two performance measures of the process, i.e. the amount of catalyst consumption and process cycle time will be adversely affected. The reactivity and the melting point of the input oil also play an important role in keeping the process under control. But, since oil is an agricultural input, it is difficult to exercise much control over it.

Another important point to be noted is that nickel is classified as a carcinogen. So the nickel catalyst is a hazardous substance and must be handled with great care. Usually, there will be a strict protocol for disposal of the spent catalyst.

2.3. Effect of important process variables

The key to analyse the effects of the process variables is to understand the concept called 'selectivity'. It is defined as the discrimination shown by a process condition in a competitive environment involving two or more reactions. Quantitatively, selectivity is expressed as ratios of two competing reactions. Now, during hydrogenation, all the four types of fatty acids, as shown above, are present in various proportions. Thus, depending on the selectivity of the process condition, one particular reaction may be favoured over the other. As a result, it is possible to have the same melting point with different proportions of the four compounds. Since this proportion is an integral part of the product specification, it is necessary to measure other quality characteristics in addition to melting point.

In the context of hydrogenation, two different types of selectivity are in use - 'linole(n)ic selectivity' and 'trans selectivity'. The linole(n)ic selectivity is the preference for transformation of a particular type of fatty acid over linole(n)ic. It is a common practice to use the term selectivity without using the qualifier 'linole(n)ic'. Trans selectivity is defined as the ratio of the percentage increase in trans isomer content and the decrease in iodine value (or the decrease in percentage double bonds, since Iodine Value (*IV*) of an oil is proportional to the number of double bonds present in it). Trans selectivity has the advantage that it remains nearly the same throughout the hydrogenation cycle commonly used in practice.

It is now generally accepted that the effects of the four important process variables, namely temperature, pressure, agitation speed and catalyst concentration on the process of hydrogenation can be summarized as shown in Table 2.1. The nature of the catalyst is also an important process variable. However, its effect is not as clearly established as the above four.

Table 2.1. Effect of important process variables

Factor	*Selectivity*	*Trans content*	*Reaction rate*
Temperature	↓	↑	↑
Pressure	↑	↓	↑
Agitation speed	↓	↓	↑
Catalyst conc.	↑	↑	↑

Chapter 3 Variance function estimation using signal-to-noise ratio*

The feature of non-constant error variance is common in applied regression. Since the standard estimation techniques like Ordinary Least Squares (OLS) are known to be less efficient in such cases, the first effort that is usually made is to select an appropriate transformation of the variables to reduce the extent of heteroscedasticity. However, as pointed out by Carroll and Ruppert [34], such an effort may not always be successful. An alternative and at times supplementary approach to transformation of variables is to model the variance explicitly as a function of one or more explanatory variables. It is customary to treat the estimated variance function thus obtained as known and then use the method of Weighted Least Squares (WLS), which usually provide a more precise estimate of the parameters. Moreover, the knowledge of the variance function allows us to construct the prediction intervals more accurately.

The approach of variance function modeling has been found useful in a variety of fields, including quality control. For example, Shoemaker et al. [38] have used the approach of variance function modeling for the purpose of process optimization. The variance function has also been used for constructing a control chart for monitoring a growth process [39]. A few other applications of the variance function in the field of quality control are highlighted in Section 3.4.

3.1. Problem statement

The problem of variance function estimation is usually stated as follows: Let Y be the response variable and the vectors X and Z represent the explanatory variables for the mean and variance function respectively. The variable set Z is usually a subset (but not necessarily a proper subset) of X. However, there can be situations where only a few or none of the elements of Z belong to X. The explanatory variables can be either fixed or random. The statistical model for the observation vector (Y_i, X_i, Z_i), $i = 1, 2, ..., n$ is then given by

$$Y_i = f(X_i, \beta) + \varepsilon_i, \quad \sigma_i^2 = g(Z_i, \beta, \theta), \tag{3.1}$$

* This chapter (excluding sections 3.7.5 - 3.7.7) is nearly the same as the following publication: S. Chakraborty, P. Mandal. “Variance function estimation using signal-to-noise ratio”. *IAPQR Transaction*, *43*(2019) 178-211.

where the errors $\varepsilon_i's$ are independent and have mean zero. The functions *f* and *g* having the parameters β and θ are the conditional mean and variance function, i.e. $E(Y_i|X_i) = \mu_i$ and $V(Y_i|Z_i) = \sigma_i^2$. The objective is to estimate *g* under the condition that *f* may be unknown. It is assumed that both *f* and *g* can be represented parametrically. Note that the popular power-of-the mean model for the variance function, using the above formulation, can be represented as $g(Z_i, \beta, \theta) = \theta_0(f(X_i, \beta))^{\theta_1}$.

The above variance function is somewhat special. It says that all the elements of *X* and no other variable affect the variance (i.e. *Z* = *X*) and this effect is a function of their joint effect on mean. For the purpose of this study, the following additional assumptions are made: (i) *X* and *Z* are uniformly distributed, (ii) $Y > 0$ and (iii) the errors are normally distributed. In addition, it is necessary to ensure that $\mu_i / \sigma_i > \tau$, where τ is called the threshold value. In practice, the threshold value is often violated while measuring very small values of *Y*. It is necessary to identify such cases and take appropriate remedial actions before embarking on the task of variance function estimation (see Example 1 in Section 3.4).

3.2. Methods of estimation - an overview

The methods of variance function estimation can be broadly classified into two groups - the methods based on grouped and ungrouped data. The details of the methods based on grouped-data along with numerical examples are available in [40] and [41]. However, it should be noted that these methods can be used only when sufficient number of repeat or near repeat points are available for forming the groups. This requirement however is unlikely to be met in many real-life situations.

The popular methods based on ungrouped data work as follows [34]: If the mean function is unknown (which is usually the case), then the same is estimated first. Then the absolute values of the residuals $r_i = y_i - f(x; \hat{\beta})$ or some functions of it like $|r_i|^2$ or $\log|r_i|$ is used as the response for the purpose of regression analysis. All such choices of the response variable are justified by the fact that $E(r_i^2) \approx \sigma_i^2$. The choice of the explanatory variables depends on the specification of the variance function. For example, if the power-of-the-mean model

$$\log(\sigma) = \log(\theta_0) + \theta_1 \log(\mu) \qquad (3.2)$$

is specified as the variance function, then $(|r_i|, \hat{\mu}_i)$ is typically used as the corresponding sample observations to develop the regression model

$$\log|r_i| = b_0 + b_1 \log(\hat{\mu}_i). \qquad (3.3)$$

The estimate b_1 is then taken as the estimate of θ_1. However $\hat{\theta}_0 = 10^{b_0}$ is known to be a highly biased estimate [42] and hence it is not used in practice. It is usually recommended that the estimates b_0 and b_1 be obtained following an iterative procedure wherein the residuals of the weighted mean model is used repeatedly to estimate the variance function till convergence (if convergence is achieved). However, here we shall estimate the variance function only once using the residuals of the un-weighted model for mean.

There are quite a few other methods for estimation of the variance function. A detailed discussion of these methods is available in the monograph by Carroll and Ruppert [34].

3.3. Inherent difficulty

It is important to realize that the estimation of the variance function is inherently more challenging than the estimation of the mean function [43]. To illustrate, let us assume that the variances are to be estimated from $|r_i| = \theta_0 + \theta_1 Z_i,\ \theta_0 \geq 0.$ Since the errors have been assumed to follow the Normal distribution, the distribution of $|r_i|$ is Half Normal with mean $\sigma_i\sqrt{2/\pi}$ and variance $\sigma_i^2(1-2/\pi).$ Thus, the Signal-to-Noise Ratio of $|r_i|$ is

$10\log[(\sigma_i\sqrt{2/\pi})^2/(\sigma_i^2(1-2/\pi))] = 2.4,$

which indicates that the observations $|r_i|$ are very noisy. This puts a serious restriction on our ability to develop a good predictive model for $|r_i|,$ particularly when the ratio $Z_{\max}/Z_{\min}$ is low. To analyze the importance of this ratio let us write the coefficient of determination of the variance function as $R^2 = 1.752[\theta_1/(\theta_0+\theta_1\bar{Z})]^2 V_Z$, where V_Z is the variance of Z. This shows that the maximum value of R^2 (corresponding to $\theta_0 = 0$) is $R^2_{\max} = 1.752(V_Z/\bar{Z}^2)$. Now, assuming that Z is uniformly distributed over (a, b), we have $V_Z/\bar{Z}^2 = (1/3)[(\gamma-1)/(\gamma+1)]^2$, where $\gamma = b/a.$ Thus, if $\gamma = 3,$ then $R^2_{\max} = 0.146.$ This shows that if $\gamma \leq 3,$ then even with a sample of moderate size (say 40), it will be extremely difficult to identify the variance function using $|r_i|$ as the response. The situation with respect to any other function of $|r_i|$ is not expected to be very different. Moreover, in practice, the task of estimation may get further complicated due to model misspecification error and the presence of outliers, if any.

3.4. Motivation for using signal-to-noise ratio

It is clear from the above discussion that the estimation of variance function, in general, is a hazardous task. Thus, it is important to use a robust procedure for the purpose of its estimation. However, here we are not interested in the standard techniques of robust estimation [34]. Rather it is proposed to estimate the *SN* function defined by

$$SN_i = 10\log(\mu_i^2 / \sigma_i^2) = g(Z_i, \beta, \theta'), \qquad (3.4)$$

where *SN* is the Signal-to-Noise Ratio. It may be noted that the above is the well-known *SN* proposed by Taguchi for nominal-the-best type of characteristic [44]. Here, we are interested in the estimation process whose primary outputs are $\hat{\mu}_i \mid X_i$. Clearly, this is a nominal-the best type of characteristic. Thus, the above *SN* measures the estimation process. It does not measure the performance of the process at different levels of *X*. For this latter purpose, it will be necessary to use the signal-to-noise ratio that is appropriate for the characteristic *Y* (see Example 4 in section 3.8.4).

The motivation for estimating the *SN* function as above comes from the works of Chen et al. [45] and Mandal [46], where it is shown that the *SN* is robust against model uncertainty and that it is a fundamental and physically meaningful characteristic of the process generating the data. The fundamental nature of the *SN* makes it a very valuable statistic for process characterization. Thus, modeling the above *SN* function itself may be the primary objective of a study, when *Y* is a nominal-the best type of characteristic. However, here the *SN* is used only as a proxy measure for estimating the variance (irrespective of the nature of *Y*). It will be shown that such an approach does provide more accurate estimates of the variances than those obtained directly from an estimate of the variance function.

3.5. Two major problems

Apart from the issue of robustness, the method of variance function estimation using ungrouped data also suffers from two major drawbacks. One is that the observations corresponding to the very low values of the residuals may appear as outliers, particularly when $\log|r_i|$ is used as the response. Usually, such outliers are trimmed out before performing the regression analysis. However, barring a few obvious cases, it is not easy to decide on optimal trimming on a case by case basis. It is thus important to study the effect of trimming on the accuracy of estimation.

The other problem, as already noted, is the large bias in the estimate of θ_0, which results in a gross underestimation of the variances. It is only in two special cases, when (3.2) becomes a single parameter model, i.e. when θ_1 is either 0 or 1, the standard error of the estimated mean function provides an unbiased estimate of θ_0. But, in general, it is necessary to correct the bias in the estimates of σ_i^2. Harvey [47] has proposed that if the variance function is given by $\sigma_i^2 = \exp(\theta Z)$, then $[\ln(r_i^2) - 1.27]$ should be used as the response for obtaining an unbiased estimate of θ. However, Ferrer and Romero [48] have shown that even after subtracting 1.27, the estimates obtained are biased when the sample size is small. Based on simulation studies, these authors have also suggested that the average of the maximum likelihood and the OLS estimate provides a nearly unbiased estimate of θ.

It is thus seen that the above two problems have received only a very limited attention in the past despite their great practical significance. An effort is made here to develop practical procedures to deal with these two problems; with a focus on small sample estimation since the two problems becomes acute when the sample size is small. The proposed methodology has two main steps - trimming of the observations followed by adjustment of the OLS estimates of the parameters of the variance (or *SN*) function. This makes the proposed estimator somewhat complicated and consequently theoretical examination of the properties of the estimator becomes extremely difficult. Thus, all the results reported are based on extensive simulation.

The rest of the chapter is organized as follows. The next section describes the *SN* and the variance functions that are considered in this study. This is followed by a detailed description of the proposed methodology (Section 3.7). In Section 3.8, four real life examples are discussed to illustrate the merits of the proposed methodology. Next, the robustness of *SN* is examined under three different types of model misspecifications (Section 3.9). The recommended zone for variance function estimation is discussed in Section 3.10. The chapter concludes with a brief summary.

3.6. *SN* and variance function

So far as performance assessment through simulation is concerned, the only SN function that is considered in this study is

$$SN = \theta_0' + \theta_1' \log(\mu), \quad 0 \le \theta_1' \le 20. \qquad (3.5)$$

Both the limits imposed on θ_1' as above are based on practical considerations. Out of the two, a violation of the upper limit is more problematic. Mandal [46] has argued that if the mean function is truly linear then we cannot have a situation where the process variance decreases with the increase in process mean. In fact, such a process, for all practical purposes, may be considered as unstable and hence the process should be improved first before undertaking any modeling exercise. However, if the mean function is non-linear then it is possible to have $\theta_1' > 20$. But such cases are not considered here. Thus, it is assumed that the mean function is linear, except in Example 4 (Section 3.8) where the mean function is linear in its parameters.

On the other hand, the lower limit is violated more frequently, particularly when the zone of study is very large (e.g. experimentation for product or process design). However, since the *SN* ratio decreases with the increase in mean, it becomes very difficult to obtain reliable estimates of the mean and variance beyond a certain limit. Thus, it will be preferable to partition the study zone and develop separate models for each zone without violating the lower limit (see Example 3, Section 3.8). It should however be noted that the violation of the limit in studies on process improvement does not pose much of a problem, since the objective is to find the best process condition (see Example 4, Section 3.8).

The *SN* function (3.5) may be estimated by developing the regression model

$$\eta_i = b_0' + b_1' \log(\hat{\mu}_i),\ \eta_i = 10\log(\hat{\mu}_i^2 / r_i^2). \qquad (3.6)$$

It will be shown (through simulation) that as in case of the variance function (3.3), b_1' is a reasonably good estimate of θ_1', but b_0' is a highly biased estimate of θ_0'.

Note that the *SN* function (3.5) is just a re-parameterized version of (3.2), where $\theta_0' = -20\log(\theta_0)$ and $\theta_1' = 20(1-\theta_1)$. So what do we gain by estimating the *SN* function? Actually, there is no gain if the residuals of all the observations are used to obtain either (3.3) or (3.6) and the same are considered as the final result. The gain is realized only when attempts are made to address the two problems mentioned in Section 3.5. The *SN* ratio provides a good basis for trimming (Section 3.7.1) and helps in correcting the bias more accurately (Section 3.7.2).

Many other variance functions similar to (3.2) have been proposed in the literature. For example, the variance function may be specified as (i) $\sigma^2 = a_0 + a_1\mu$, or (ii) $\sigma = b_0 + b_1\mu$, or (iii) $\log(\sigma) = c_0 + c_1\mu$. It will be shown in Section 3.9 (again through simulation) that the *SN* function (3.5) can be used even in such situations as an approximation provided the data does not show any lack of fit for the fitted *SN* function.

It is also possible to specify the *SN* function as a function of *Z*. Such *SN* functions can also be estimated following the method proposed in this study (see Example 4, Section 3.8).

3.7. Estimation - proposed methodology

The proposed methodology consists of four main steps - (i) Trimming for data preparation, (ii) Graphical analysis of the trimmed data for identification of the *SN* function, (iii) Estimation of the parameters following OLS and hypothesis testing, and finally (iv) Correction of bias. The relevant details of the four steps are discussed below under four subsections.

3.7.1. Trimming

According to Davidian and Carroll [43], it is essential to remove the smallest few absolute residuals for estimating the variance function (3.2). However, they have left it to the judgment of the practitioner to decide on the number of observations to be removed. It must also be noted that in some cases, trimming of just one observation can destroy the variance structure completely, particularly when the sample size is small (see Ryan [42] for an example). Thus, it is extremely difficult to decide the optimal trimming on a case by case basis unless one has substantial prior knowledge of the process. Moreover, for the purpose of examining the effect of trimming through simulation, it is necessary to have predefined trimming rules. Our objective here is to identify a good trimming rule.

3.7.1.1. Trimming rules

The various types of trimming schemes can broadly be classified into two groups: 'ar' trimming and 'SN' trimming. In 'ar' trimming, a few observations corresponding to the smallest absolute values of the residuals are removed. This is the most commonly used method of trimming. However, here we shall also consider 'SN' trimming, where a few observations corresponding to the highest values of η_i are removed. Each of these can be further subdivided into two groups: fixed and variable. In fixed trimming, a fixed percent of the observations is removed while in variable trimming, the observations to be removed are decided based on some predefined rules. For example, a rule for variable 'ar' trimming may be as follows: remove i_{th} observation if $|r_i| < 0.2s_0$, where s_0 is the standard error of the estimated mean function.

3.7.1.2. Existing status

As already noted, the existing practice is to perform 'ar' trimming using subjective judgment. In order to avoid subjectivity, let us consider fixed 'ar' trimming and optimize the level of trimming with respect to the variance function (3.2). The simulation experiment conducted for this purpose is as follows. The data are collected for twenty different combinations of n (= 20, 40, 60, 100) and θ_1 (= 0, 0.25, 0.5, 0.75, 1). The simulated data are first subjected to various levels of trimming (1 - 6% of n) and then the trimmed data is used to obtain (3.3). The accuracy of the estimates of θ_1' is evaluated in terms of p_1 and p_2 (defined in Table 3.1) as well as Mean Square Error (MSE) of the estimates.

Table 3.1. Accuracy measures

Name	*Notation*	*Description*
Probability 1	p_1	% of $\hat{\theta}_1'$ within ±5 of the true value θ_1'
Probability 2	p_2	% of $\hat{\theta}_1'$ within ±10 of the true value θ_1'
Error in $\hat{\sigma}_{ij}^2$	e_{ij}	$e_{ij} = (\sigma_{ij}^2 - \hat{\sigma}_{ij}^2)/\sigma_{ij}^2; i = 1,..,n; j = 1,2,..,N$
Bias of e_{ij}	B_E	Grand average of e_{ij}
Average within variation of e_{ij}	V_E^w	$V_E^w = \sum V_j / N$, V_j = Variance of the e_{ij} within a sample
Mean square error of e_{ij}	MSE_E	$MSE_j = \sum_i e_{ij}^2 / n$, $MSE_E = \sum_j MSE_j / N$

At each level of trimming, the accuracy of estimation is evaluated based on N = 10000 samples. It is observed that, in general, the accuracy deteriorates for higher values of θ_1' as the degree of trimming increases. The optimal trimming, based on a subjective judgment of the overall performance, is found to be 3%. Henceforth, this optimal ‘ar’ trimming of 3% will be called the method M1. The results obtained with M1 are summarized in Table 3.2. It is seen that even with optimal trimming the performance is not satisfactory, since both p_1 and p_2 varies significantly over different levels of θ_1'. Table 3.2 also shows that the overall bias (B_E, defined in Table 3.1) of the estimates of variance, as expected, is very large.

3.7.1.3. A Good trimming rule

After experimenting with various types of trimming schemes as above, the best performing trimming rule is found to be 4% 'SN' trimming (rounded to nearest integer),

which will henceforth be referred to as the method M2. The performance of the two methods (M1 and M2) with respect to p_1 and p_2 is summarized in Table 3.3. It is seen that the performance of M2, in comparison to M1, is much more consistent over all the values of θ_1'. So, henceforth the method M2 will be used for the purpose of estimating both θ_1 and θ_1'.

Table 3.2. Performance of M1

n	θ_1'	$\overline{\hat{\theta}_1'}$	$MSE(\hat{\theta}_1')$	p_1	p_2	B_E (%)	
						LAR	*SN*
20	20	16.9	31.1	0.74	0.90	60	61
	15	14.5	31.1	0.51	0.93	59	61
	10	11.7	39.5	0.50	0.82	59	60
	5	8.8	51.0	0.46	0.82	57	58
	0	6.7	79.1	0.45	0.71	54	55
60	20	18.1	11.6	0.86	0.98	64	64
	15	15.2	15.7	0.73	0.99	64	64
	10	11.2	20.9	0.71	0.96	63	64
	5	7.3	23.5	0.67	0.96	63	63
	0	4.1	31.8	0.62	0.91	62	62
100	20	18.5	7.1	0.92	0.99	64	65
	15	15.3	10.8	0.84	1.00	65	65
	10	11.1	14.1	0.82	0.99	64	64
	5	6.9	16.6	0.77	0.99	64	64
	0	3.4	21.5	0.70	0.97	64	64

However, it should also be noted from Table 3.3 that the estimate of θ_1' tends to get biased as one moves away from $\theta_1' = 10$ in either direction. Consequently, the bias becomes somewhat large at the two ends ($\theta_1' = 20$ and 0). This bias decreases with the increase in sample size, but for smaller samples the bias is substantial.

3.7.2. Graphical analysis

A scatter plot of r_i vs. $\hat{\mu}_i$ (or Z_i) is commonly used for identification of the nature of the variance function. Similarly, the scatter plot of η_i vs. $\log(\hat{\mu}_i)$ or Z_i can be used for identification of the *SN* function. A scatter plot of η_i vs. $\log(\hat{\mu}_i)$ of the trimmed data will be called a *SN-LFIT* plot. If the plot shows no correlation between *SN* and *LFIT* then it signifies

that σ_i is nearly proportional to μ_i. The other extreme is when the variance is constant. In this case the plot will exhibit positive correlation with moderate slope. However, since it is difficult to judge the slope of the *SN* function from the plot, the fitted line may be superimposed on the plot. Note that if the slope is found to be very high (much greater than 20) then it implies $\theta_1 < 0$. Similarly, if *SN* and *LFIT* are found to be negatively correlated, then it signifies that $\theta_1 > 1$. Of course, if the plot shows that the relationship between *SN* and *LFIT* is nonlinear then the specification of the *SN* function should be modified.

Table 3.3. Performance of M2

n	θ_1'	$\overline{\hat{\theta}_1'}$	$MSE(\hat{\theta}_1')$	p_1	p_2	B_E (%)	
						LAR	*SN*
20	20	15.8	46.1	0.66	0.84	60	61
	15	13.1	39.9	0.50	0.88	60	61
	10	10.1	38.6	0.51	0.83	59	60
	5	7.2	39.2	0.51	0.88	58	59
	0	5.4	59.2	0.55	0.79	55	56
60	20	16.9	21.1	0.74	0.95	61	61
	15	13.4	20.5	0.71	0.97	61	61
	10	9.5	19.2	0.74	0.97	60	61
	5	6.0	16.0	0.74	0.99	60	60
	0	3.7	19.0	0.76	0.96	59	60
100	20	17.5	14.3	0.82	0.98	63	63
	15	13.7	14.1	0.81	0.99	63	63
	10	9.3	13.4	0.83	0.99	63	63
	5	5.3	11.4	0.83	1.00	62	62
	0	2.2	11.9	0.84	0.99	62	62

Another use of the *SN-LFIT* plot is to judge whether any further trimming is necessary. Our experience suggests that such subjective trimming should be limited only to extreme cases like the presence of an isolated point either at the upper left or lower right corner of the plot. As far as possible, subjective trimming should be avoided when the sample size is small unless one has sufficient knowledge about the nature of the *SN* function.

The *SN-LFIT* plot can also be used for examining the residuals of the weighted mean function. Of course, the appropriate residuals (r_i^w) and fitted values $(\hat{\mu}_i^w)$, in this case, are $w_i^{1/2} r_i$ and $w_i^{1/2} \hat{\mu}_i$ respectively, where $w_i = 1/\hat{\sigma}_1^2$. Consequently, the signal-to-noise ratio is

given by $\eta_i^w = 10\log[(\hat{\mu}_i^w)^2/(r_i^w)^2]$, which is the same as that for the un-weighted model and *LFIT* is given by $\log(w_i^{1/2}\hat{\mu}_i)$.

3.7.3. Role of hypothesis testing

The hypothesis testing is used here only for the purpose of testing whether $\hat{\theta}_1'$ satisfies the restriction $0 \le \theta_1' \le 20$ or not. Thus, if the null hypothesis in any of the two cases

(i) $H_0 : \theta_1' = 20, H_1 : \theta_1' > 20,$ (ii) $H_0 : \theta_1' = 0, H_1 : \theta_1' < 0,$

is rejected at 2.5% level of significance, then the sample is rejected. Otherwise, the values of the estimates $\hat{\theta}_1' > 20$ and $\hat{\theta}_1' < 0$ are modified to twenty and zero respectively. Note that in practice the rejection of a sample will call for a detail investigation of the process including the method of data collection. All the results presented here, including those in Tables 3.2 and 3.3, are after modification of θ_1' as above.

3.7.4. Correction of bias

It may be noted from the last two columns of Table 3.2 that the estimates of the variances obtained from (3.3) and (3.6) are highly biased (underestimated by about 60%). In this section, simple methods are proposed for adjusting the estimates of θ_0' and θ_1', which reduce the bias considerably. However, the methods proposed are not aimed at obtaining unbiased estimates of the variances. Rather, the objective is to adjust the two parameters of the variance and the *SN* function so that the mean square error (*MSE*) of the estimates as measured by e_{ij} (defined in Table 1) is minimized. Since, bias is the major component of the *MSE*; its minimization automatically leads to the reduction of bias.

3.7.4.1. Adjustment of $\hat{\theta}_1$ and $\hat{\theta}_1'$

In order to reduce the *MSE* of e_{ij}, it is proposed to adjust the estimate of θ_1' as follows:

$$\begin{aligned} \hat{\theta}_1'^c &= 20, \text{ if } \hat{\theta}_1' > 17.5, \quad \text{and} \quad \hat{\theta}_1'^c = 0, \text{ if } \hat{\theta}_1' < 2.5, \\ &= \hat{\theta}_1', \text{ otherwise}, \qquad\qquad = \hat{\theta}_1', \text{ otherwise}, \end{aligned} \tag{3.7}$$

where the superscript *c* indicates that it is a corrected estimate. The estimates of θ_1 are adjusted in a similar fashion.

The effectiveness of the above method of adjustment is verified through simulation as follows. Ten thousand random samples of size 20 and 100 are generated by varying the parameters of the mean and the *SN* function randomly over a wide range. The fit ratio, for the sake of simplicity, is kept fixed at 4. Next, the mean square error of e_{ij} is computed for each of the samples under three conditions: (i) Using the estimate of θ_1' as obtained from the regression analysis of the trimmed data, (ii) Assuming $\theta_1' = 20$ and (iii) Assuming $\theta_1' = 0$. Let the mean square error MSE_j (defined in Table 3.1) for the j_{th} sample under the three conditions as above be denoted by $MSE0$, $MSE1$ and $MSE2$ respectively. The averages of $MSE0$, $MSE1$ and $MSE2$ are shown in Table 3.4. It is seen from this table that

$$\begin{aligned} Min(\overline{MSE0}, \overline{MSE1}, \overline{MSE2}) &= \overline{MSE1}, \ if \ \hat{\theta}_1' > 17.5, \\ &= \overline{MSE0}, \ if \ 2.5 \le \hat{\theta}_1' \le 17.5, \\ &= \overline{MSE2}, if \ \hat{\theta}_1' < 2.5. \end{aligned}$$

This shows that the bias correction as in (3.7) reduces the mean square error of the estimates significantly.

Table 3.4. Average *MSE*0, *MSE*1 and *MSE*2

$\hat{\theta}_1'$	*n* =20			*n* =100		
	$\overline{MSE0}$	$\overline{MSE1}$	$\overline{MSE2}$	$\overline{MSE0}$	$\overline{MSE1}$	$\overline{MSE2}$
> 17.5	0.52	0.24	0.60	0.060	0.052	0.496
17.5-12.5	0.22	0.47	0.48	0.054	0.109	0.372
12.5-7.5	0.19	0.67	0.42	0.057	0.314	0.223
7.5-2.5	0.23	0.99	0.40	0.057	0.869	0.118
< 2.5	0.78	1.79	0.37	0.101	1.758	0.079

3.7.4.2. Adjustment of $\hat{\theta}_0$ and $\hat{\theta}_0'$

The method proposed for adjusting the estimate of the intercept term is also very simple. Let T_1 and T_2 be the suitably chosen targets for the mean of the predicted values of *LAR* (Logarithm of Absolute Residuals) and *SN* respectively. Then

$$\begin{aligned} &\hat{\theta}_0^c = s_0, \ if \ \hat{\theta}_1^c = 0, \\ &\quad = s_1, \ if \ \hat{\theta}_1^c = 1, \\ &\quad = T_1 - \hat{\theta}_1^c \overline{\log(\hat{\mu})}, \ otherwise \ ; \\ &T_1 = \log(s_0), \end{aligned} \tag{3.8}$$

where s_0 and s_1 are the standard errors of the un-weighted and weighted mean function respectively.

Similarly, for the *SN* function we have

$$\hat{\theta}_0'^c = T_2 - \hat{\theta}_1'^c \overline{\log(\hat{\mu})}, T_2 = 10\log(\overline{\hat{\mu}}^2 / V). \quad (3.9a)$$

$$V = \overline{\hat{\sigma}_i^2(LAR)}, \text{ if } \hat{\theta}_1'^c = 0, \quad (3.9b)$$

$$= s_0^2, \text{ Otherwise,}$$

In the above, *LAR* and *SN* indicates that the variances are estimated from *LAR* and *SN* respectively.

3.7.4.3. Corrected estimates of the variances

Using the corrected estimates of the two parameters of the variance and the *SN* function, the final estimates of the variances are obtained as follows.

$$\hat{\sigma}_i^2(LAR) = (\hat{\theta}_0^c \hat{\mu}_i^{\hat{\theta}_1^c})^2, \quad (3.10)$$

$$\hat{\sigma}_i^2(SN) = \hat{\mu}_i^2 / 10^{\hat{\eta}_i^c/10}, \ \hat{\eta}_i^c = \hat{\theta}_0'^c + \hat{\theta}_1'^c \log(\hat{\mu}_i). \quad (3.11)$$

3.7.4.4. Effectiveness of adjustment

As before, the effectiveness of the above methods of adjustment is examined through simulation using N =10000 samples of different size (n). The results obtained are summarized in Table 3.5. It is evident from this table that the bias B_E has reduced considerably compared to the unadjusted estimates (from about 60% to a maximum of 35% and 21% for the variance and the *SN* function respectively). But the remaining bias is still somewhat large. In particular, since the biases at the two ends are large, the method of adjustment is modified by replacing T_1 by $f(\hat{\theta}_1)T_1$ with various choices for f. The target T_2 is also modified similarly. However, none of these efforts has been successful. In all such cases, the MSE_E is found to increase significantly. So, the methods of adjustment presented in (3.8) and (3.9) and the results shown in Table 3.5 are considered as final.

The other important conclusions which can be drawn from Table 3.5 are as follows: (i) The sign of bias changes from positive to negative in case of *SN*, but this is not so with the estimates obtained from *LAR*. (ii) $V_E^w(SN) < V_E^w(LAR)$ for all n and θ_1'. (iii) $MSE_E(SN)$ is much lower than $MSE_E(LAR)$ for $\theta_1' \geq 10$. The $MSE_E(LAR)$ tends to be slightly smaller than $MSE_E(SN)$ for $\theta_1' < 10$ and $n > 60$. This is to be expected. But it does not have much

practical significance (unless the sample size is very large) since the true value of θ_1' will never be known. (iv) Overall, $\hat{\sigma}_i^2(SN)$ are found to be more accurate than $\hat{\sigma}_i^2(LAR)$.

Table 3.5. Accuracy of the estimates of variances

n	θ_1'	B_E (%)		V_E^w (%)		MSE_E (%)	
		LAR	*SN*	*LAR*	*SN*	*LAR*	*SN*
20	20	-5.8	18.7	16.2	13.4	30.3	28.2
	15	-8.7	14.7	16.8	14.3	31.6	28.6
	10	-15.7	6.4	19.8	15.4	33.8	29.4
	5	-23.8	-4.7	29.1	17.8	57.9	33.2
	0	-34.7	-17.7	72.5	33.4	122.2	54.0
60	20	-2.8	20.2	5.4	3.4	9.3	10.1
	15	-6.3	17.0	7.1	4.6	11.9	10.8
	10	-15.1	8.0	7.9	6.1	15.7	11.6
	5	-22.0	-5.0	7.2	5.8	20.3	11.9
	0	-21.0	-16.6	10.6	6.0	28.4	14.7
100	20	-1.6	20.5	3.4	2.1	5.6	7.8
	15	-5.5	17.4	4.7	3.0	7.6	6.9
	10	-14.1	9.9	4.7	3.4	9.9	6.9
	5	-21.9	-4.8	4.7	4.2	14.7	8.3
	0	-16.6	-16.7	5.3	3.1	16.8	9.4

3.7.5. Comparison of variance of $\hat{\theta}_1$ with its asymptotic variance

The mean square error of adjusted $\hat{\theta}_1'$, as shown in Table 3.3, indicates that the values are sufficiently small to be useful in most practical situations even when the sample size is small. But what is the individual contribution of trimming and adjustment of the estimates towards reducing the variance of the estimates? Also, is the variance of $\hat{\theta}_1'^c$ comparable with the asymptotic variance of the usual estimate (i.e. $\hat{\theta}_1$ obtained from untrimmed data)? In order to answer these questions, we shall perform another simulation experiment. But let us first note the asymptotic distribution of $\hat{\theta}_1$.

The details of asymptotic analysis related to the problem of variance function estimation are available in [43]. Here we note only the relevant details related to the regression method of estimation based on the logarithm of the absolute residuals, as summarized in [34]. Using the same notation as in [34], let us write the variance function as

$Variance\ (y_i) = \sigma^2 g^2(\mu_i(\beta), z_i, \theta), i = 1, 2, ..., N,$

where N is the sample size. Let $\nu(i, \beta, \theta) = \log[g(\mu_i(\beta), z_i, \theta)], i = 1, 2, ..., N$ and ν_θ be its derivative. Also let $\xi(\beta, \theta)$ denote the sample covariance matrix of ν_θ. Then, if β is known, $\hat{\theta}$ is asymptotically normally distributed with mean θ and variance V_a, which is given by

$V_a = (Variance\ [\log(\varepsilon^2)]/(4N))\xi(\beta, \theta)^{-1},$

where ε is the standardized error component, i.e., $\varepsilon_i = [y_i - f(x_i, \beta)]/[\sigma g(\mu_i(\beta), z_i, \theta)]$.

It is also shown in [43] that even when β is estimated from a sample data, the asymptotic variance of $\hat{\theta}$ may be approximated as V_a, provided the variances do not depend on mean or if σ is small. But, in our case, the variance depends on the mean and the proposed method is not restricted to the small σ cases. The formulae for the asymptotic variance for the general case are also available in [43]. But these are very complicated for any practical use. So as an approximation, the asymptotic variance of $\hat{\theta}$ will be computed as V_a.

Now, for the power-of-the-mean model considered here, we have $g(\mu, \theta) = \mu^{\theta_1}$ and $\sigma = \theta_0$. Thus, $\nu_\theta = d[\theta_1 \log(\mu)]/d\theta_1 = \log(\mu)$ and hence $\hat{\xi}(\beta, \theta_1) = Variance\ (LFIT)$, where $(LFIT)_i = \log(\hat{\mu}_i)$. Further, since ε has mean zero and standard deviation unity, the variance of $\log(\varepsilon 2)$ is approximately 0.949. Thus, replacing N by n (in order to get back to our notation), we finally have

$\hat{V}_a(\theta_1) = 0.237/[n * Variance\,(LFIT)].$

It may be noted that the variance of $\hat{\theta}_1$ decreases with the increase in the variance of *LFIT*, which is consistent with our earlier observation made in Section 3.3 that the ability to identify the variance function critically depends on the fit ratio $(\hat{\mu}_{\max} / \hat{\mu}_{\min})$. The importance of this ratio is illustrated further in Section 3.10.

The main features of the simulation experiment that is conducted to answer the two questions posed at the beginning of this section are as follows. As before, the basic data (of size 20, 60 and 100) is generated by varying the two parameters of the mean function and the scale parameter θ_0 of the variance function randomly in a wide range such that the fit ratio remains approximately five in all cases and $y > 0$. From each such sample, the parameter θ_1 is estimated for the following four cases: (i) Untrimmed data-Unadjusted estimate, (ii) Untrimmed data-Adjusted estimate, (iii) Trimmed data-Unadjusted estimate and the (iv)

Proposed estimate, i.e. Trimmed data-Adjusted estimate. The proposed method M2 is used for trimming and $\hat{\theta}_1$ is adjusted following the method described in Section 3.7.4.1. The variance of $\hat{\theta}_1$ is estimated based on 10000 such samples. In addition, the asymptotic variance, as given above by $\hat{V}_a(\theta_1)$ is also obtained for each sample size. The results obtained are summarized in Table 3.6.

Table 3.6. Comparison of variances of $\hat{\theta}_1$

n	θ_1	*Variance* $(\hat{\theta}_1)$				
		Untrimmed & Unadjusted	*Untrimmed & Adjusted*	*Trimmed & Unadjusted*	*Trimmed & Adjusted*	*Asymptotic*
20	0	0.397	0.107	0.285	0.101	0.164
	0.25	0.380	0.134	0.277	0.128	
	0.5	0.390	0.151	0.271	0.135	
	0.75	0.395	0.142	0.268	0.121	
	1	0.389	0.114	0.261	0.090	
60	0	0.114	0.043	0.087	0.043	0.055
	0.25	0.113	0.078	0.084	0.070	
	0.5	0.113	0.094	0.082	0.077	
	0.75	0.116	0.078	0.081	0.060	
	1	0.117	0.044	0.083	0.033	
100	0	0.069	0.026	0.048	0.028	0.033
	0.25	0.067	0.053	0.046	0.046	
	0.5	0.067	0.066	0.045	0.047	
	0.75	0.068	0.055	0.047	0.041	
	1	0.067	0.025	0.045	0.018	

The following conclusions and recommendations may be made based on the results shown in Table 3.6.

(i) The unadjusted estimate of θ_1, as obtained from the untrimmed data should never be used. These estimates have very high standard error.

(ii) Both trimming and adjustment reduce the variance of $\hat{\theta}_1$. Adjustment reduces the variance by a large amount when θ_1 is either zero or one, as is to be expected since the adjustment is made only for the very low and high values of $\hat{\theta}_1$.

(iii) Adjustment has greater beneficial effect over trimming for sample size of 20. However, as sample size increases, its effect is mostly felt at the two extremes.

(iv) The asymptotic variance is significantly larger than the variance of the proposed estimator for the two extreme cases, i.e., $\theta_1 = 0 \; and \; 1$. But for $0.25 \le \theta_1 \le 0.75$ and sample size of 60 - 100, the former tends to become lower than the latter. Note that the variances of the unadjusted estimates are more or less the same for all θ_1.

(v) Finally, since the approximate asymptotic variance V_a is much lower than V_s (variance of the unadjusted estimates obtained from trimmed data), we shall not recommend the use of V_a for the purpose of inference on θ_1. Our experience suggests that the inference on θ_1 may be based on the standard error of $\hat{\theta}_1$ as obtained from the regression output, which usually comes close to V_s.

3.7.6. Large sample behavior of $\hat{\sigma}_i^2$

It has been noted in Section 3.2 that the residual based regression method of variation function estimation makes use of the fact that $E(r_i^2) \approx \sigma_i^2$. Also, $s_o^2 = \sum_{i=1}^{n} r_i^2 /(n-k)$, where s_0 and k are the standard error and the number of parameters of the mean function respectively. Thus, assuming k to be small compared to n, we have $E(s_0^2) = \overline{\sigma_i^2}$. Now, assuming further that $\hat{\theta}_1$ is unbiased, we should expect the estimates $\hat{\sigma}_i^2$ also to be unbiased if $\overline{\hat{\sigma}_i^2}$ is set to s_o^2 and this should be true irrespective of the form of the variance function and the method used for its estimation. The proposed method of bias correction is based on this simple idea. But the situation here is a little more complicated since we are equating the average of $(\hat{\log \sigma_i})$ with $(\log s_0)$. Consequently, one cannot expect the estimates to be unbiased. The results of simulation shown in Table 3.5 also show that the adjusted estimates are not unbiased. But the important point to note here is that the remaining bias is low and the same may be acceptable in most practical situations. However, two questions remain – (i) Does the observed bias of the unadjusted estimates decrease with increasing sample size and if not, then (ii) does the accuracy of the adjusted estimates improve with increasing sample size? It will be nice to answer these questions based on an appropriate asymptotic analysis. But it is evident from the discussion in the previous section that an exact asymptotic analysis is very difficult even when the estimates are obtained in a straightforward manner. The analysis will

be even more difficult in the present case since the proposed approach is more complicated. Consequently, we shall resort to simulation to answer the above two questions.

The simulated data are generated as before, but unlike in the previous case (Section 3.7.5), the fit ratio is also varied randomly in the rage of 4 to 10. The maximum sample size used in these simulation experiments is 10000. It is obvious from the simulation results (Table 3.7) that the bias in the uncorrected estimates remains very large. In fact, the observed biases are more than those reported in Table 3.3. This is because the bias figures reported in this table are for untrimmed data, while the values shown in Table 3.3 correspond to trimmed data. However, it is obvious that the source of bias is not related to sample size. It is structural in nature. In particular, it is shown in the next section that the bias is primarily due to the use of $\log|r_i|$ as the response for the purpose of regression.

It is also interesting to note from this table that when the sample size is 5000 or more, there is no beneficial effect of trimming. In fact, with $n \geq 5000$, the accuracy of the estimates decreases slightly (perhaps due to the reduction of sample size on account of trimming).

Table 3.7. Large sample behavior of the estimates of σ_i^2

n	θ_1'	*Untrimmed data – Uncorrected estimates (B_E %)*	*Untrimmed data – Corrected estimates (MSE_E %)*		*Trimmed data – Corrected estimates (MSE_E %)*	
			LAR	*SN*	*LAR*	*SN*
1000	20	71.9	0.29	0.39	0.87	1.59
	15	71.9	1.06	3.78	1.2	3.81
	10	71.9	2.49	1.74	2.55	1.77
	5	71.8	8.73	0.91	8.5	0.79
	0	71.8	1.16	0.44	0.56	0.29
5000	20	71.9	0.04	0.04	0.2	0.45
	15	71.9	0.25	3.32	0.74	3.55
	10	71.9	1.70	1.30	1.98	1.44
	5	71.9	7.81	0.17	7.86	0.2
	0	71.9	0.04	0.04	0.04	0.04
10000	20	71.9	0.02	0.02	0.07	0.15
	15	71.9	0.18	3.29	0.68	3.53
	10	71.9	1.61	1.23	1.91	1.38
	5	71.9	7.71	0.09	7.8	0.14
	0	71.9	0.02	0.02	0.02	0.02

Finally, it may be noted (from Table 3.7) that the MSE_E of the corrected estimates of the variances is reduced with the increase in sample size. This, of course, is to be expected. In fact, with sample size of 10000, it is possible to obtain nearly exact estimates when θ_1' is 20, 5

or 0. However, the rate of convergence of MSE_E for the other two cases is very slow. This shows that there is ample scope for improvement of the proposed method of adjustment of $\hat{\theta}_0$. Nevertheless, it is seen that the *SN* function continues to outperform the corresponding variance function even in case of large samples. But the problem of high variation of performance (over different values of θ_1) remains.

3.7.7. Origin of bias

It has been noted in Section 3.2 that the residual based methods for variance function estimation exploit the relationship $E(r_i^2) \approx \sigma_i^2$. This being an approximation, one may suspect this to be a source of bias. To examine the nature of the approximation, let us write the mean function in the matrix form as $E(Y) = X\beta,$ where β is the vector of parameters. Then, the exact expression for the variance of r_i is given by

$$E(r_i^2) = V(r_i) = (1-h_{ii})^2\sigma_i^2 + \sum_{k\neq i}\sigma_k^2 h_{ik}^2,$$

where h_{ik} are the elements of the hat matrix $H = X(X^TX)^{-1}X^T$. However, the contribution of the additional terms in the exact expression as above can be ignored for large *n*. So, the above approximation cannot be the main source of the observed bias, which is as large as 72%.

To look at the problem differently, it may be noted that if the parameters (β, θ) of (3.1) are estimated following the method of maximum likelihood, then the estimates are known to be asymptotically unbiased, although they may be biased in case of small samples [48]. This suggests that the bias, which persists even for large samples, must be related to the method of estimation. Since the method used in the present case is OLS, it becomes readily apparent that the problem of bias is a simple consequence of the Jensen's inequality, i.e. $E(\log|r_i|) \leq \log E(|r_i|).$ To explain, we note that since $\log|r_i|$ is used as the response for the purpose of regression, the fitted line is forced to pass through $(\overline{\log\hat{\mu}}, \overline{\log|r_i|})$. But, the proposed method of adjustment implies that there will be no need for such an adjustment if the line passes through $(\overline{\log\hat{\mu}}, \log s_0)$. In other words, since $\overline{\log|r_i|} \leq \log\overline{|r_i|} \leq \log(s_0),$ the estimates are bound to be highly underestimated.

The above discussion also suggests that if, for example, the parameters of the variance function $\sigma^2 = \theta_0 + \theta_1 X$ are estimated following OLS and r_i^2 is used as the response, then there will be no need for bias correction.

Thus, we may conclude that the main source of bias is the use of $\log|r_i|$ as the response in the regression model and the variation of bias over θ_1 is an artifact of the proposed method of adjustment, since the bias of the unadjusted estimate is same for all θ_1. But, can we estimate the variance function in a manner, which eliminates both the problems of bias and variation in bias over θ_1? The following algorithm does this job remarkably well. But it is applicable only for large samples. The main idea behind this alternative algorithm is to select a suitable set of observations from among the original data set such that the sampled values of signal-to-noise ratio (η) are nearly uniformly distributed and the average η of the sampled values is close to the average of the true values. Towards this end, we proceed as follows:

Step 1: Arrange the dataset $(\eta, \log \hat{\mu})$ in ascending order of $\hat{\eta}$.

Step 2: Trim the top 0.6% and bottom 28% of the observations. These trimming fractions are obtained through trial and error. The objective is to make the trimmed data nearly uniform with a mean close to the target. Let n_1 be the number of observations in the trimmed data set.

Step 3: Select $n_2 = n_1$ observations systematically from the trimmed data as follows. The performance of the algorithm does not vary much for $0.6n \leq n_2 \leq n_1$, where n is the number of observations in the original data set.

3.1: Select the first observation. Let η_1 be the signal-to-noise ratio of the first observation.

3.2: Select the $k = 2, ..., n_{2th}$ observation such that the value of η of the k_{th} observation is closest to but greater than $\eta_1 + (k - 1)(R/n_2)$, where R is the range of η of the trimmed data.

Step 4: Develop the SN function with the sampled data as obtained above. The estimates obtained of the parameters are final and require no further adjustment.

The performance of the above algorithm is judged through simulation (10000 samples of different sizes) by computing the MSE_E as before. Here we report only the values for n = 1000. The estimates of MSE_E (%) for θ_1' = 20, 15, 10, 5, 0 are found to be 1.18, 1.15, 1.30, 1.24 and 1.26. The average of these five values is 1.23. In comparison, the first five values in the last column of Table 3.7 show that the corresponding MSE_E of the proposed small sample estimator are 1.59, 3.81, 1.77, 0.79 and 0.29, with an average of 1.65. It is thus seen that the

above algorithm performs better than the previous one in terms of both average and variation of MSE_E.

However, the above algorithm performs poorly in case of small samples. For example, the MSE_E of the present procedure for n = 60 are found to be 17.9, 16.8, 16.3, 15.3 and 15.9 for θ_1' = 20, 15, 10, 5 and 0 respectively with an average of 16.44. But for the same samples, the corresponding values obtained using the previous procedure is 7.2, 10.0, 11.0, 10.1 and 12.5 respectively with an average of 10.16. Thus, the above algorithm may be used only for large samples. But, in case of large samples, one should expect to have near repeat points and consequently the variance function may also be estimated following the method of grouping. So, it is necessary to compare the performance of the above procedure with a suitable procedure based on grouping. Such a comparison is not made here, since this article is primarily focused on variance function estimation based on a small sample.

3.7.8. Summary of the proposed methodology for small samples

To summarize, the steps involved in obtaining the estimates $\hat{\sigma}_i^2(SN)$ are as follows:

Step 1: Fit the mean function and obtain $(\eta_i, \hat{\mu}_i)$.

Step 2: Arrange $(\eta_i, \hat{\mu}_i)$ in descending order of η_i and trim top 4% of the observations.

Step 3: Obtain the estimates of θ_0' and θ_1' following OLS regression

Step 4: Perform tests of hypothesis as in section 3.7.3. If H_0 is rejected, then STOP.

Step 5: If $\hat{\theta}_1' > 17.5$ then $\hat{\theta}_1'^c = 20$; if $\hat{\theta}_1' < 2.5$ then $\hat{\theta}_1'^c = 0$; otherwise $\hat{\theta}_1'^c = \hat{\theta}_1'$.

Step 6: If $\hat{\theta}_1'^c = 20$ then estimate the variances as $\hat{\sigma}_i^2 = s_0^2$, where s_0 is the standard error of the OLS.

Step 7: If $\hat{\theta}_1'^c = 0$ then perform weighted regression and obtain the standard error s_1 of the weighted mean function. Obtain the estimates of the variances as $\hat{\sigma}_i^2(LAR) = (s_1\hat{\mu}_i)^2$.

Step 7: If $\hat{\theta}_1'^c \neq 0$, then obtain the estimate $\hat{\theta}_0'^c$ using (3.9a) and (3.9b).

Step 8: Obtain the estimates of the variances as $\hat{\sigma}_i^2(SN)$ given by (3.11).

3.7.9. Numerical examples

Let the true mean and the variance function of a process be given by E(Y) = μ = 2 + 3X_1 + 4X_2 and σ = 0.6* $\mu_{0.5}$ respectively. It is assumed that X_1 and X_2 follows uniform distributions with parameters (2, 14) and (10, 70) respectively. A random sample of size 40 (shown in the

first six columns of Table 3.8) is generated from the above process, assuming that $Y \mid X$ follows normal distribution. The OLS estimate of the mean function obtained from this sample is $Y = 8.63 + 2.84X_1 + 3.84X_2$ with $R^2 = 97.89\%$ and $s = 9.56$.

The scatter plot (Figure 3.1) of the residuals vs. the fitted values of the above mean function shows that there are three extreme values at the right end of the plot (two at the top and one at the bottom). The standardized values of these residuals are 2.97, 2.58 and -2.32. Should the corresponding observations be considered outliers (assuming homoscedasticity) or inliers having larger variance? A good estimate of the *SN* function may be helpful in answering this question. But, in this particular case, it may be difficult to have a clear-cut answer, since the fit ratio $(\hat{Y}_{\max} / \hat{Y}_{\min})$ is very low (= 3.24).

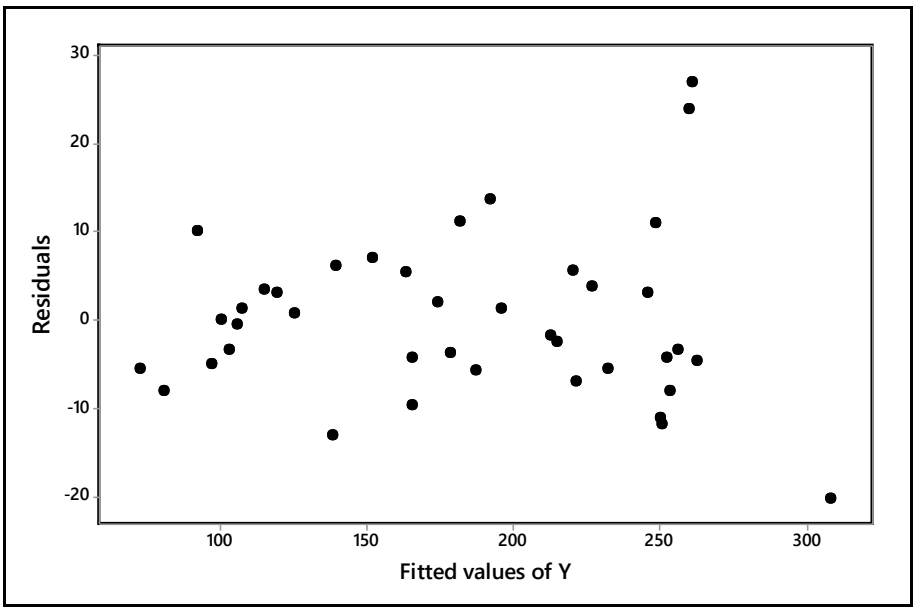
The SN function for the sample data is estimated following the proposed method as above, i.e. the relevant data are first trimmed following M2 and the parameters of (3.6) are estimated following OLS regression. This gives $SN = 10.17 + 8.89 \log(\hat{\mu})$, which indicates that the error variance may not be constant. However, the evidence is not conclusive, since the 95% confidence interval of the slope, i.e. (-4.7, 22.5) includes 20. Our experience suggests that if $n \leq 40$, $15 \leq \theta_1' \leq 5$ and fit ratio ≤ 5, then one should expect to have such results (i.e. failure to identify the variance function with high confidence) in many cases.

The OLS estimate of the variance function obtained from the same set of trimmed data is $\log | r | = -0.509 + 0.556 \log(\hat{\mu})$, where the *p*-value of the estimate of slope is 0.11. It is thus seen that the variance function also leads to the same conclusion with respect to the constancy of variance. In fact, both the SN and the variance function as above will provide the same but highly biased estimates of the variances. The SN function is expected to provide better estimates only after bias correction following the method proposed in Section 3.7.5. For example, in the present case, the $MSE_E\%$ (defined in Table 3.1) of the estimates obtained from the adjusted SN function is found to be much lower (8.7%) than that obtained from the adjusted variance function (23.2%).

Table 3.8: Simulated data for the numerical examples

First example						*Second example*							
X_1	X_2	Y	X_1	X_2	Y	X_1	X_2	X_3	Y	X_1	X_2	X_3	Y
13.6	51.8	249.3	9.4	34.0	156.2	6.7	6.9	18.2	64.9	22.1	2.9	12.3	107.9
5.6	21.3	105.5	12.6	53.8	239.2	6.6	9.0	14.7	71.5	7.5	8.6	23.0	79.7
13.6	36.6	181.8	12.1	24.9	125.6	22.3	3.2	18.9	113.7	5.9	5.6	4.7	59.2
7.0	13.7	73.1	3.5	55.7	226.9	16.5	2.9	7.5	89.9	14.6	6.7	6.7	84.5
13.2	47.1	230.8	4.9	37.3	161.4	22.7	8.6	14.3	148.6	17.4	5.8	9.9	102.4
5.4	30.2	145.8	13.4	38.0	206.1	18.2	6.0	9.7	111.0	19.9	5.5	22.2	119.4
10.7	56.0	245.8	5.3	62.4	258.4	22.7	6.1	22.1	128.7	10.1	2.3	9.1	58.5
12.1	34.2	176.3	4.7	24.2	118.5	7.5	2.3	8.1	49.5	9.1	9.9	21.1	91.6
3.5	34.8	159.0	10.2	47.6	225.9	14.0	3.1	14.1	74.7	17.0	4.1	10.8	93.6
8.9	22.4	123.0	6.0	58.2	259.9	10.9	6.6	5.0	89.3	21.9	7.0	15.9	128.4
8.7	59.1	284.0	13.5	13.0	92.1	13.3	8.9	12.4	108.1	10.4	4.8	12.7	62.6
9.1	19.0	108.8	5.9	20.3	99.7	15.3	9.0	17.3	107.2	6.4	8.3	19.2	73.6
12.9	56.2	288.1	8.0	10.9	67.4	20.3	9.3	6.5	150.1	13.1	2.4	21.2	68.2
4.6	41.7	193.1	8.3	47.7	212.6	5.3	8.6	23.6	67.4	15.0	3.3	11.2	84.0
7.2	35.1	169.3	9.1	23.7	126.3	17.5	7.6	16.3	112.5	22.5	4.7	4.5	115.8
3.4	53.0	214.7	6.4	58.9	248.4	4.9	7.8	3.7	62.5	15.5	4.2	16.5	85.5
5.9	17.6	102.8	5.1	59.2	239.2	21.3	6.4	15.1	116.6	10.3	8.6	3.0	87.5
3.3	50.8	211.0	13.5	68.2	288.1	8.0	9.1	18.3	90.8	7.0	2.7	8.9	46.1
6.9	39.2	175.1	2.5	62.6	252.9	23.5	7.0	10.9	150.5	23.6	3.8	4.9	118.0
2.5	22.2	100.8	11.0	40.7	197.3	19.2	8.9	23.8	126.6	17.1	4.6	9.4	98.3

Figure 3.1. Scatter plot (Numerical example)



To take another example, let the true mean and variance function of a process are given by $E(Y) = \mu = 5 + 4X_1 + 5X_2$ and $\sigma_Y = 2 + 0.2X_1 + 0.4X_2 - 0.2X_3$ respectively. Thus, in our notation of Section 3.1, $X = (X_1, X_2)$ and $Z = (X_1, X_2, X_3)$. It is assumed that all the three explanatory variables X_1, X_2, and X_3 follow uniform distribution with parameters (4, 24), (2, 10) and (3, 24) respectively. As before, a random sample of size 40 is generated using the above specifications and assuming normality of the error term. The random sample thus obtained (shown in the last eight columns of Table 3.8) gives an estimate of the mean function as $Y = -0.47 + 4.27X_1 + 5.47X_2$, with $R_2 = 96.3\%$ and $s = 5.406$. As before, the variance and the SN function are estimated using the trimmed data. The estimates obtained are $\log|r| = 0.263 + 0.01584X_1 + 0.0452X_2 - 0.02163X_3$ and $SN = 24.21 + 0.397X_3$. Unlike in the previous example, all the regression coefficients are found to be significant in this case. The only exception is the intercept term of the variance function. But it is retained since its omission leads to a large increase in the variance inflation factors (VIF) of the remaining terms. Note that the SN function is much simpler than the variance function, since the effects of X_1 and X_2 on standard deviation are approximately proportional to their effects on mean.

Next, the variances at forty different level combinations of (X_1, X_2, X_3) are estimated following the proposed method and it is found that the MSE_E% of the estimates obtained from the SN function is lower (43.3%) than that obtained from the variance function (51.9%). Finally, using the estimates of the variances obtained from the SN function, the weighted mean function is estimated as $Y = 2.57 + 4.17X_1 + 5.17X_2$. Note that the estimates of the parameters are now closer to their true values compared to the estimates obtained following OLS regression. Of course, one should not expect such an improvement in all cases. But, if the variance function is estimated sufficiently accurately, then one should expect a reduction in the standard error of the estimates of the parameters of the mean function. For example, the standard errors of the intercept and the coefficients of X_1 and X_2 for the un-weighted model are 3.44, 0.144 and 0.368, whereas the same for the weighted model are 2.35, 0.118 and 0.251 respectively.

3.8. Real life examples

In this section, the proposed methodology is illustrated with the help of four real life examples. Apart from the proposed methodology, the examples also illustrate the role of data transformation in modeling.

The three of the data sets used in these examples have also been analyzed previously by others and the models developed here are not the same as those reported by the previous authors. But the alternative models proposed here should not be considered as superior to the previous ones. As Carroll and Ruppert [34, p. 23] says "any real data analysis is highly context specific and claiming an improved analysis of someone else's data is usually as pointless as it is misleading". Our purpose here is only to illustrate the proposed methodology and to highlight the different perspective that the *SN* function brings to the problem of variance function estimation.

3.8.1. Example 1

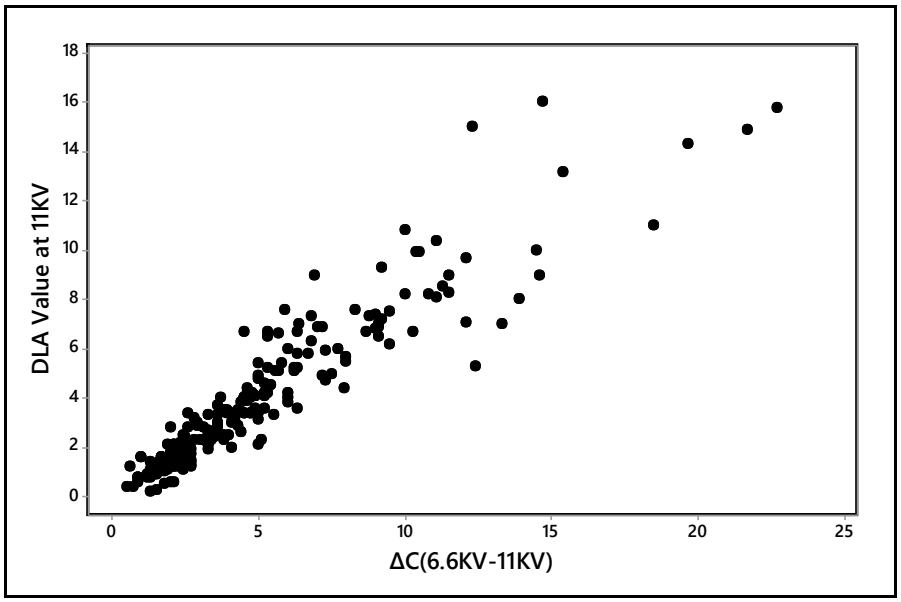
The measurements obtained from a Dielectric Loss Analyzer (DLA) are used to judge the overall quality of insulation of an electrical cable. For the sake of convenience, the measurements obtained from the DLA analyzer will also be referred to as *DLA*. There is an upper specification limit of *DLA*. The change in capacitance (ΔC) of the cable with the increase in applied voltage is another measure related to the quality of insulation. This later measurement is easy and fast while the *DLA* measurement is time consuming. The measurement of ΔC is made on 100% of the cables produced but the *DLA* values are measured on a sample basis (10% of the cables). However, the measurement of *DLA* even on a sample basis is found to be a troublesome exercise.

It is known that ΔC and *DLA* have strong positive correlation. So it is decided to develop a model for predicting *DLA* based on the corresponding measurement of ΔC. The objective is to perform *DLA* measurement only for those cables for which ΔC exceed a certain limit and thereby reduce the testing load. Here we discuss only the modeling part of the problem. Other relevant details of the study including the data set are available elsewhere [49].

The scatter plot of *DLA* vs. ΔC shows that the two are strongly and positively correlated but the variation of *DLA* increases as ΔC increases (Figure 3.2). The *SN* function for this data set is estimated as $\hat{\eta} = 13.3 + 6.52\hat{\mu}$, which corresponds to $\hat{\theta}_1 = 0.674$. So the weighted model for the mean function is developed using $(1/\hat{\mu}_i^{0.674})^2$ as the weights. Figure 3.3 shows the *SN-LFIT* plot of the weighted model along with the fitted *SN* function ($\hat{\eta}^w = 15.13 + 11.66\hat{\mu}^w$). The 95% upper confidence limit of $\hat{\theta}_1'$ is found to be approximately 21.5, which indicates that it is a borderline case. A closer look at the *SN-LFIT* plot shows that the plot is not very satisfactory. In fact, if only one observation corresponding to the lowest *SN* (= -2.64) is

removed, then the *SN* function becomes insignificant at 5% level of significance. This suggests that the weighted model does not satisfy the model assumptions well.

Figure 3.2. Scatter plot of *DLA* vs. ΔC (Example 1)



The scatter plot of log (*DLA*) vs. log (ΔC) is shown in Figure 3.4. It is apparent from the plot that the log (*DLA*) fluctuates much more at the lower level than the rest, where the variance is nearly constant. Further investigation reveals that the DLA instrument loses its sensitivity below one. Thus, to be on the safe side, all the observations corresponding to $\Delta C <$ 3 are removed and the mean function is estimated as Log(*DLA*) = -0.0714 + 0.962*Log(ΔC). For this log-log model with truncated data we have $\hat{\theta}_1' = 21.96$ and the corresponding *SN-LFIT* plot is found to be satisfactory.

Figure 3.3. *SN-LFIT* plot of the weighted model (Example 1)

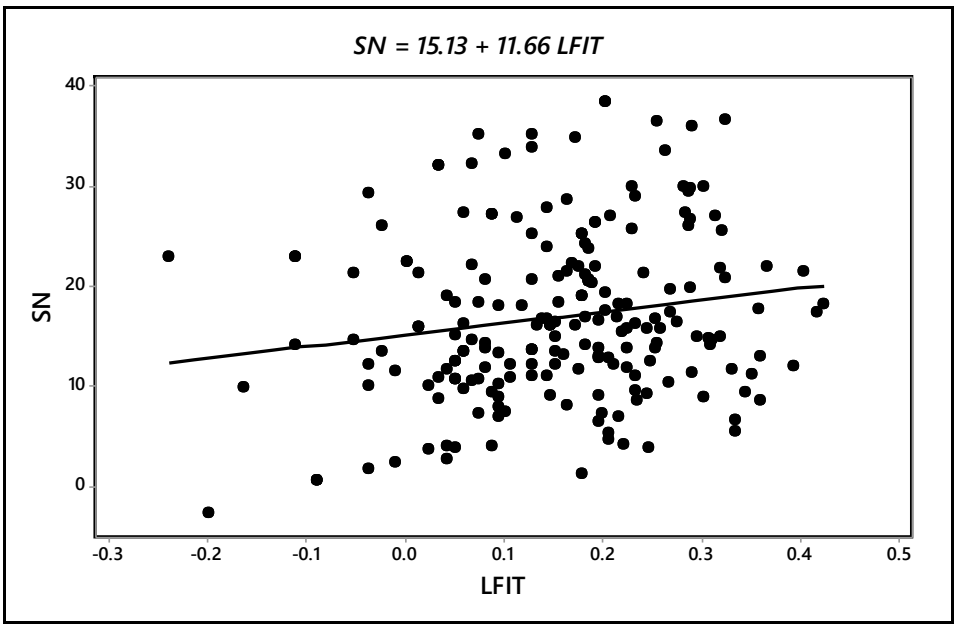
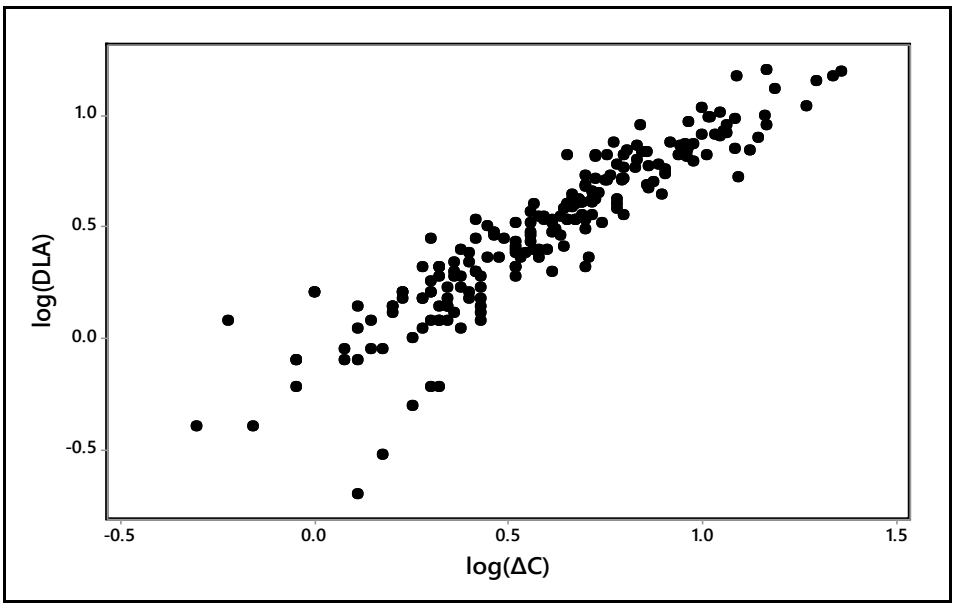


Figure 3.4. Scatter plot of log(*DLA*) vs. log(ΔC)



3.8.2. Example 2

Carroll and Ruppert [34] have analyzed the data (n = 108) of a calibration experiment involving radioimmunoassay (RIA). The calibration curve to be estimated is specified as $Y = \beta_0 + \beta_1 X$, where Y is the concentration of esterase (an enzyme) and X is the *RIA* count.

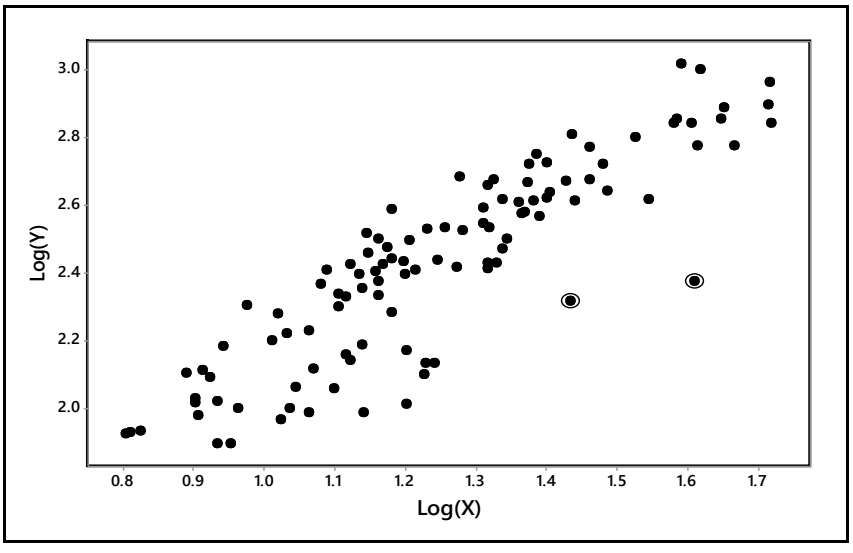
Once the calibration curve is established then the same will be used for inferring about the unknown concentration of a new sample based on the measurement of *RIA* count.

For the sake of simplicity, they consider $\hat{\theta}_1 = 1.0$ and the mean function is estimated following the method of weighted regression. However, no comment is made about the adequacy of the weighted model.

The proposed method M2 gives $\hat{\theta}_1 = 0.776,$ which is in line with the observation by the authors that $\hat{\theta}_1 = 1.0$ is a slight overestimate. An analysis of the residuals of the corresponding weighted mean model does not indicate any obvious sign of model inadequacy, but it does not provide enough confidence on the model either. So let us have a critical look at the data.

Figure 3.5 shows the scatter plot of Log(*Y*) vs. Log(*X*). It is obvious from the plot that at lower values of concentration, the measurements of counts are very erratic. As a remedial measure, the observations corresponding to Log(Y) < 2.2 (i.e. count < 159) and another two possible outliers (marked in Figure 3.5) are removed and the mean function is re-estimated using the truncated data. This gives $\log(Y) = 1.31 + 0.95\log(X), s = 0.08.$ The residuals of this model gives an excellent fit to the normal distribution (much better than the residuals of the weighted model).

Figure 3.5. Scatter plot of Log(*Y*) vs. Log(*X*) (Example 2)



Alternatively, if one wants to play safe (since there are outliers even after truncation), then a random coefficient model may be considered. For the truncated data, we have $Y / X = k$, where the observed values of k give an excellent fit to the log-normal distribution with location parameter 2.863 and scale parameter 0.1839.

Note that the situation here is similar to that in Example 1 above. However, there is an important practical difference. In Example 1, the truncation of data is not problematic since the interest lies in predicting *DLA* at the higher end. However, in this case the truncation as above may defeat the main purpose of the study, which is to measure the low concentrations. Thus, if it is decided to improve the process, then the objective should be to improve the signal-to-noise ratio of the measurement process, given by $SN = 10\log(\hat{\beta}^2 / s^2)$, where $\hat{\beta}$ is the slope and s is the standard error of the log-log model. It is found that the signal-to-noise ratio of the existing process is only 18.4 db, whereas the same for the truncated data is $10\log(.9509^2 / .0796^2) = 21.5$ db. Thus, the goal for process improvement should be to achieve $SN \geq 21.5$ db for the entire range of concentration that may be of interest. However, if the data are assumed to be sufficiently free from measurement error then it is necessary to change the model specifications.

3.8.3. Example 3

Draper and Smith [40] illustrate the method of weighted regression analysis using a set of 34 observations on (X, Y) having near repeat points. So the variance function is estimated following the method of grouping. They obtain the following estimates of the weighted mean function and the variance function,

$$\hat{\mu} = -.889 + 1.1648X, \quad \hat{\sigma}^2 = 1.5329 - 0.7334X + 0.0883X^2. \qquad (3.12)$$

The residuals (r^w) of the above mean function are plotted against $(w^{0.5}X)$ in Figure 3.6. The two left most points in the plot are beyond the range of X that is used for estimating the above variance function. So, the low spread of the residuals in this region is ignored and the authors conclude that the weighted model in (3.12) is adequate.

Let us now examine the model (3.12) a little differently. It may be noted that in our framework there is a restriction of $\theta_1 \leq 1$. However, the variance function in (3.12) violates the spirit of this restriction. In fact, the proposed method M2 (after removing the two problematic observations) gives an estimate of θ_1 as 1.63. To illustrate the difficulty of working with such a variance function, consider the mean and the *SN* function of the process

as shown in Figure 3.7. Here the mean and the signal-to-noise ratio are computed based on the same groupings as in (3.12). It is seen that there is a large drop in *SN* with $X > 3$. In our opinion, if the mean function is really linear in the region $3 \leq X \leq 5$ and the estimate of *SN* at $X = 3$ is reliable then the process offers significant scope for improvement (assuming that *Y* is a nominal-the-best type of characteristic). A difference of about 11db in *SN* within a narrow range of *X* should not be accommodated by accepting the estimates on their face value and modeling the mean and the variance function accordingly.

Figure 3.6. Scatter plot of the residuals vs. X for the weighted mean model in (3.12)

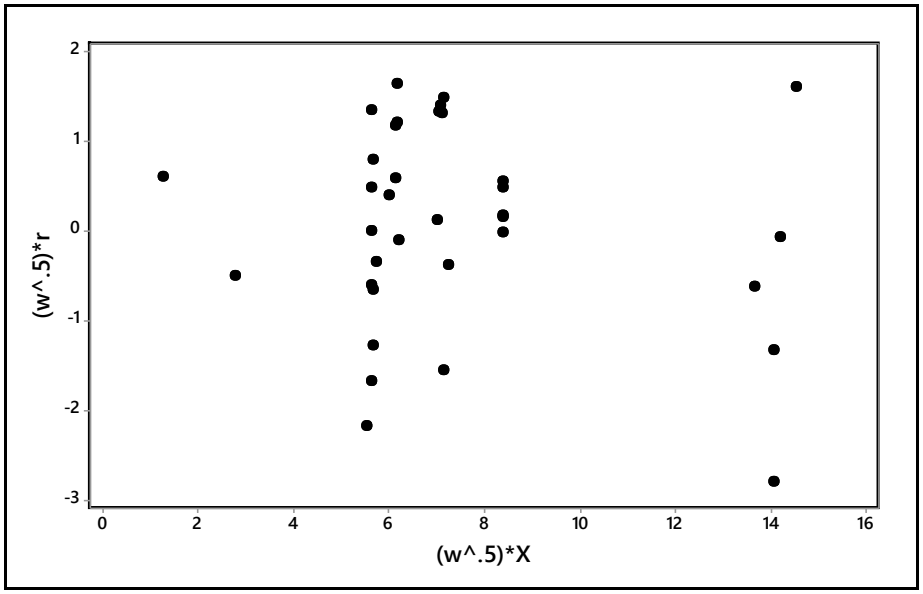
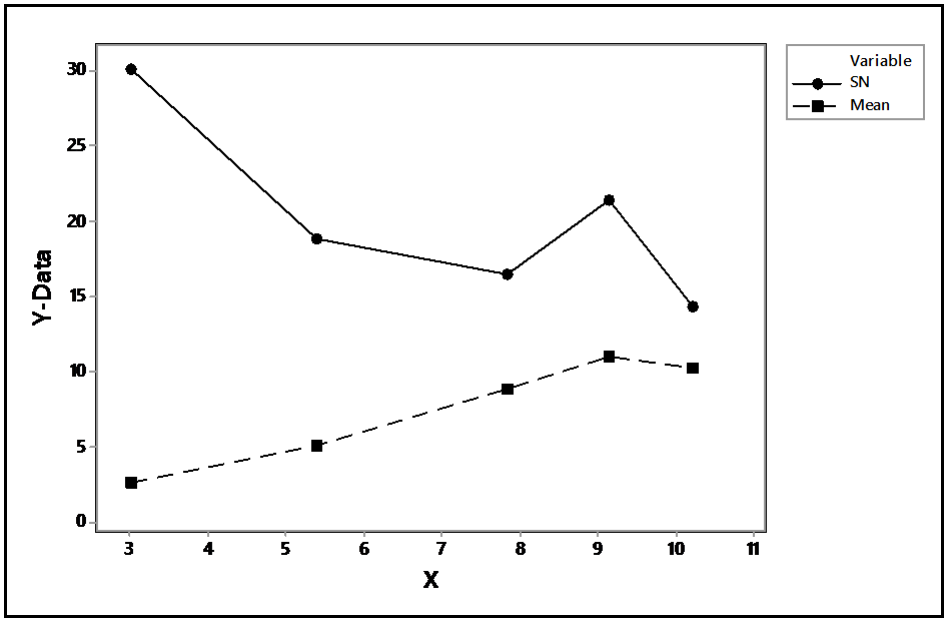


Figure 3.7. The mean and the *SN* function (Example 3)



However, given the data, it will be preferable to model the process separately for two regions: $X \leq 3$ and $X > 3$. For $X > 3$, the method M2 gives $\hat{\theta}_1 = 1.0$, which, in turn, leads to the following process model; $\hat{\mu} = 1.068X$, $\hat{\sigma} = 0.1653\hat{\mu}$. The average error variance $[= \sum r_i^2/(n-1)]$ of this model is 2.56. On the other hand, if the variance function in (3.12) is used, then for this restricted region we have $\hat{\mu} = -1.671 + 1.277X$ and an average error variance of 2.7.

It is also seen from Figure 3.7 that the *SN* at the highest level of *X* is slightly lower than the rest. Thus, as in the previous example, here also one may like to play safe and consider the random coefficient model $Y = kX$, where *k* follows normal distribution with mean 1.068 and standard deviation 0.177 as an alternative.

It should however be noted that the above analysis does not put a question mark on the linearity of the mean function, which may have a strong theoretical basis. Our limited conclusion is that the mean function may be actually linear but its parameters cannot be estimated accurately based on the measurements available. So, either the process needs to be improved or the mean function should be estimated for a restricted region. The situation here may be similar to the 'strong interaction' example given in Weisberg [50, pp. 88-102], where the specification of the mean function is not questioned, even though the experimental data shows significant lack of fit. But then the measurement process certainly becomes suspect (see also Weisberg et al. [51]).

3.8.4. Example 4

Taguchi and Wu [8, pp. 67-73] describe a welding experiment involving nine two-level factors (*A* - *I*). The design matrix is obtained by allocating the nine factors suitably to the different columns of the orthogonal array L_{16}, so that the main effects as well as four specific interaction effects can be estimated. This single replicate experiment has five responses, but here we shall consider only the data on tensile strength (*Y*). The original objective of the study is to find the best operating condition but here the data are used only for estimating the *SN* function, which will allow us to identify the factors having significant dispersion effects.

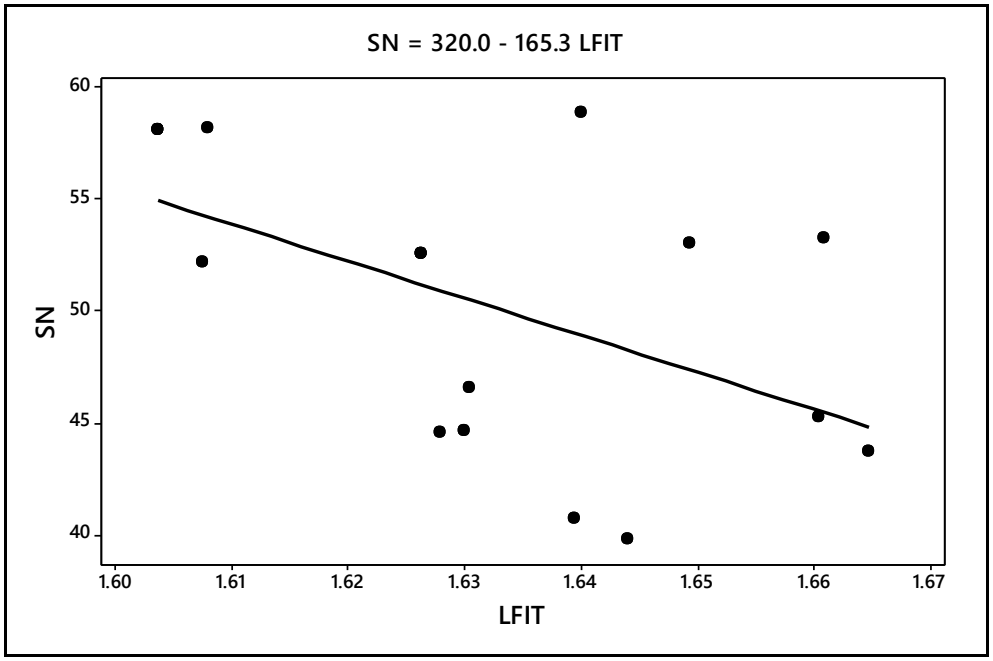
Box and Meyer [52], based on a graphical analysis of the data, conclude that the factors *B* and *C* have significant effect on mean and that the factor *C* has a strong dispersion effect. Later, Ferrer and Romero [48], based on an analysis of the same data, note that the factor *I* (in addition to *C*) also have significant dispersion effect.

Identification of the mean and the variance function in this case is expected to be a problematic task since the sample size is small and there are a large number of predictors to choose from. So it is necessary to exercise appropriate caution while developing the process model. Since the factor *C* has already been identified as having significant dispersion effect, let us partition the experiment into two with eight trials each corresponding to the two levels of C. Now, if the eight trials at the second level of *C* are considered as a 2_3 experiment involving the factors *AG*, *H* and *B*, then all the three factors are found to be highly significant. Similarly, at the first level of *C*, the effects of *A*, *B* and *F* are found to be significant at 10, 0.3 and 9% level of significance respectively. This shows that even on a conservative basis, a few more effects other than the main effects of *B* and *C*, are likely to be important. The standard ANOVA of the experimental data reveals that the effects of *A*, *AH*, *B*, *F*, *I*, *AC* and *C* are significant, which is the same set of effects reported as significant in Taguchi and Wu [8]. So, to begin with, we obtain the following mean function,

$$Y = 42.96 + 0.2A - 0.21AH + 1.08B - 0.2F - 0.19I + 0.19AC - 1.55C, \qquad (3.13)$$

where the first and the second level of the factors are coded as -1 and +1 respectively. The *SN-LFIT* plot of the above mean function is shown in Figure 3.8.

Figure 3.8. *SN-LFIT* plot of (3.13)



Usually one should not expect any significant pattern in the *SN-LFIT* plot with such a low fit ratio (= 1.15). However, in this case, a significant pattern is visible since the error variance

varies a great deal from trial to trial. In fact, the negative slope within such a narrow range of *LFIT* indicates that the noise conditions are totally different at the two levels of one or more factors and hence it should be easy to identify those factors responsible for such a large difference in variance. For example, if the design matrix along with the *LFIT* vector is sorted in ascending order of *LFIT*, then eight out of the top nine trials correspond to the first level of *C*. This clearly indicates that the factor *C* is the most important factor. It is also noted that the four smallest values of *LFIT* correspond to the second level of *D*, which indicates that the factor *D* may also be significant. However, since *D* has no significant effect on mean, the following SN function involving only the factor *C* is considered first,

$$SN = 49.31 + 3.54C. \tag{3.14}$$

Using the above *SN* function, the final weighted model for mean is obtained as

$$Y = 42.96 + 0.2A - 0.21AH + 1.03B - 0.2F - 0.14I + 0.19AC - 1.55C, \tag{3.15}$$

which is seen to be very similar to the un-weighted model (3.13). The *SN* function of the model (3.15) is given by $SN = 4.9 + 20.95LFIT$. This shows that the variance has been stabilized successfully.

Next, it is found from multiple regression analysis that both the factors *C* and *D* have significant effect on *SN*. But the *SN* function of the corresponding weighted mean model becomes slightly nonlinear. So, the models (3.15) and (3.14) are considered as the final mean and *SN* function respectively. Using (3.14) and (3.15), the estimates obtained of the variances at the two levels of the nine factors are shown in Table 3.9. As is to be expected, the factors *D*, *E*, *G* and *H*, which are not present in the mean function, have no dispersion effect. The remaining five factors are significant with respect to variance. In case of *A*, *B*, *F* and *I*, the standard deviation is proportional to mean but *C* has disproportionally large effect on variance. However, since the effects of *A*, *F* and *I* on mean are small, the differences in the estimates of the variances are also negligibly small. It should however be noted that since we are dealing with a case of $\theta_1' < 0$, the estimates obtained as above may not be as accurate as is to be expected otherwise.

Table 3.9. Estimates of variances (Example 4)

Level	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*
1	0.092	0.088	0.199	0.092	0.092	0.093	0.092	0.092	0.093
2	0.093	0.097	0.043	0.092	0.092	0.092	0.092	0.092	0.092

Finally, let us find the optimal level of *C*. Since tensile strength is a larger the better type of characteristic, the appropriate *SN* ratio [44] for comparing the two levels of *C* is given by $\eta = -10\log[\sum(1/y_i^2)/n]$, where y_i are the n (= 8) observations at a given level of *C*. This gives $\eta(C_1) = 32.96$ and $\eta(C_2) = 32.34$, which shows that even though the expected variance at C_1 is much larger than that at C_2, the optimum is obtained at C_1.

3.9. Robustness of the *SN* function

So far, we have assumed that the mean and the variance functions are specified correctly and have demonstrated the superiority of the SN function. However, since model misspecification error is commonplace in practice, it is important to evaluate the performance of the *SN* ratio under model misspecification. In the following, three types of misspecification errors are considered - misspecification of the (i) mean function, (ii) variance function and (iii) both the mean and variance functions. Thus, the true models are specified as

(i) $\mu = 2 + 10X + 0.05X^2$, $\sigma = 0.15\mu^{0.75}$,

(ii) $\mu = 2 + 10X$, $\log(\sigma) = 0.8 + 0.027X$, and

(iii) $\mu = 2 + 10X + 0.05X^2$, $\log(\sigma) = 0.8 + 0.027X$,

but the data are analyzed assuming that the mean function is linear, and the variance and the SN function are given by (3.2) and (3.5) respectively. The parameters of the above three models are selected in such a manner so that the nonlinearity of the mean and the SN function cannot be detected in most of the simulated samples when the maximum fit ratio is limited to 6. For example, in case (i), the fitted linear model for mean to the simulated data shows lack of fit in only 8% of the cases (for all *n*). The simulations are performed as before (keeping fit ratio in the range of 3-6) and the estimates $\hat{\sigma}_i^2(SN)$ and $\hat{\sigma}_i^2(LAR)$ are obtained following the methodology summarized in Section 3.7.5. The results of these experiments are summarized in Table 3.10.

It is seen from Table 3.10 that the results obtained are very similar to those shown in Table 3.5. This suggests that the SN function (3.5) can be used to approximate a wide variety of situations so long as there is no empirical evidence of non-linearity of the mean and the SN function. Note also that for a given sample size, the MSE_E values of the estimates of variances obtained from the SN function fluctuate much less compared to those obtained from the variance function. This shows that the SN function is relatively more robust than the corresponding variance function.

Table 3.10. Robustness of SN ratio

n	*Case*	B_E (%)		MSE_E (%)	
		LAR	*SN*	*LAR*	*SN*
40	(i)	-19.4	-4.3	22.0	16.8
	(ii)	-15.0	-3.3	21.5	16.1
	(iii)	-17.5	-4.5	24.8	16.9
60	(i)	-18.0	-3.8	18.6	12.8
	(ii)	-14.6	-2.7	15.9	11.7
	(iii)	-18.7	-3.5	19.8	13.0
100	(i)	-15.5	-3.1	16.6	9.0
	(ii)	-14.9	-2.6	12.2	8.5
	(iii)	-16.6	-3.3	15.9	8.7

3.10. Recommended zone of variance function estimation

The recommended zone is defined as a function of sample size and fit ratio. Such a zone is identified by noting the proportion of cases for which *MSE*0 > *MSE*1, where *MSE*0 and *MSE*1 are as defined in Section 3.7.4.1. The idea is that if the variances can be estimated more accurately following the proposed methodology than assuming constancy of variance in a reasonably large number of cases then it will be worthwhile to estimate the *SN* function. Thus, a simulation experiment is conducted for different combination of *n* and fit ratio and for each combination, the proportion (*p*) of cases (based on *N* = 10000) in which *MSE*0 is greater than *MSE*1 is noted. As before, the parameters of the mean and the variance function are varied in a wide range covering the entire range of θ_1' (from 20 to 0). However, the cases for which $\hat{\theta}_1' > 17.5$ are not considered for obtaining the value of *p* since the variance is assumed to be constant in such cases. The results of the experiment are summarized in Figure 3.9.

As expected, the value of *p* increases with the increase in *n* and fit ratio. Assuming a cut-off value of 82.5%, the minimum requirement of fit ratio for a given *n* is noted. These minimum values of fit ratio are plotted against *n* to obtain the recommended zone as shown in Figure 3.10. The plot also shows a preferred zone. The rate of increase in *p* below the preferred line is relatively higher than above it. However, the recommended zone as above should be used only as a guideline since it has been constructed based on subjective judgment.

Figure 3.9. Effect of *n* and fit ratio on *p*

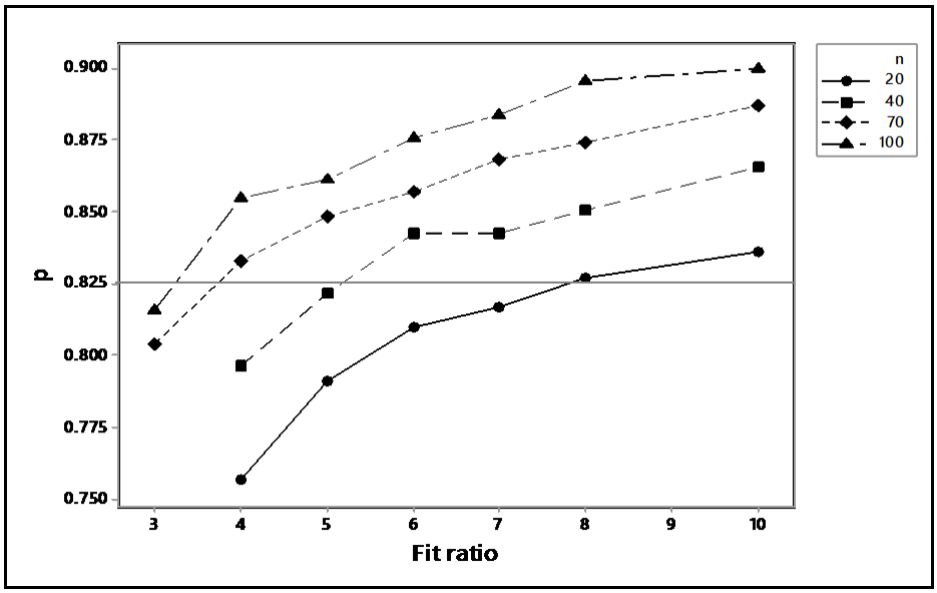
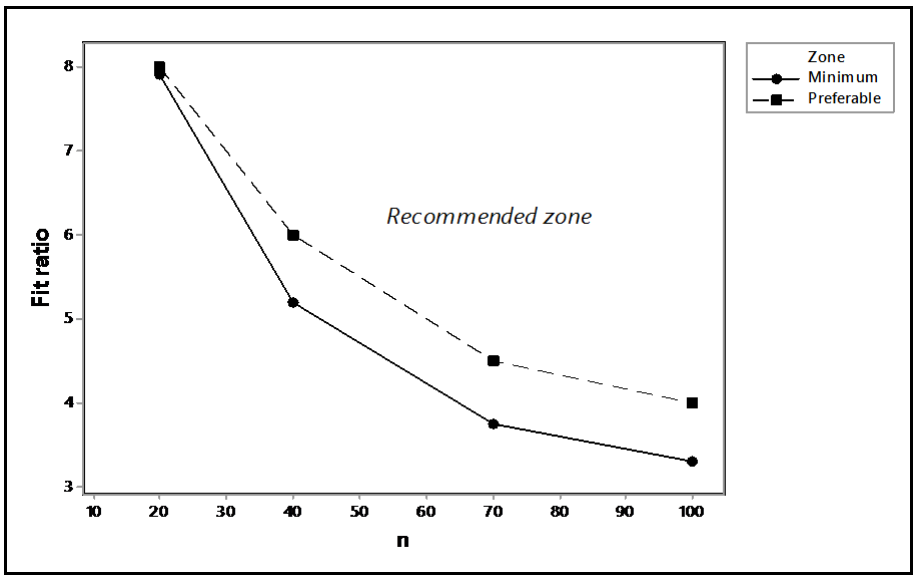


Figure 3.10. Recommended zone for variance function estimation



3.11. Conclusions

Modeling the *SN* function brings a new perspective to the problem of developing regression-based process models under heteroscedasticity. It is shown through extensive simulations that the adjusted *SN* function gives more accurate estimates of variances than the corresponding adjusted variance function. The results presented here are expected to be useful to practitioners in developing the variance function of a process based on a small sample.

Chapter 4 Growth models

All of us are familiar with the phenomenon of growth. We grow from our childhood and so many things around us also grow along with us. We put a great deal of effort to have growth. For example, the editor of a journal is trying hard to grow its readership. Most of us would like our investments and savings to grow and plan accordingly. It is thus obvious that we have an intuitive understanding of growth. This intuitive understanding is pictorially represented in Figure 4.1.

Figure 4.1. Intuitive depictions of growth*

* Source: Castellano and Ho [53]

However, the above intuitive understanding is not enough to deal with many other similar situations. It is true that we put a great deal of effort to have growth as above. But we also put serious effort to prevent growth. For example, nobody would like a tumour to grow, whether it is benign or malignant. But this distinction between desirable and undesirable growth is not process specific. India's population is growing, perhaps at an alarming rate. But another country may be worried about its population de-growth. Secondly, the pattern of growth, whether increasing or decreasing, need not be smooth as shown in Figure 4.1. Finally, the growth function, in many cases, is found to be non-monotonic. But, whatever be the nature of growth, it is clear that growth is not about status quo. It refers to change over time.

Unfortunately, such a broad definition of growth is likely to cause some confusion. The main source of this confusion is the two concepts related to growth, i.e. the concepts of *status* and *improvement*. Castellano and Ho [53] provide the following example from the field of education to differentiate the three terms. Table 4.1 gives the averages of marks obtained in mathematics by all the students of different grades (3 to 8) of a school in different years. A vertical slice of the data, i.e. a column refers to *status*; a horizontal slice, i.e. a row giving the change in average marks over time for a particular grade describes *improvement* over time, while the diagonal slice (as shown by the bold faced numbers in Table 4.1) represents *growth*.

Table 4.1. Example of school status scores across grade levels

	Year					
Grade	2007	2008	2009	2010	2011	2012
3	**320**	380	350	400	390	420
4	400	**450**	420	450	480	500
5	510	550	**600**	650	620	620
6	610	620	630	**620**	650	660
7	710	780	750	750	**800**	800
8	810	810	820	820	810	**840**

It may be noted that 'improvement' in the above example also refers to growth in a broader sense, since it can be thought of as a measure of performance growth of the school. However, in this case the group of students, i.e. the cohort changes over time. This constitutes a dilution of the concept of growth, which refers to the change of an individual or a group of fixed set of individuals over time. Of course, in practice, the cohort may change even while we are studying growth, but all such changes are to be viewed as aberrations.

The structure of Table 4.1 is very simple. But it is a generic and very useful way to summarize growth data. We shall also provide such a summary of our growth data in Section 4.3. However, for the purpose of this study, it is necessary to provide some more clarification regarding the nature of a growth process. Let Y_t represent the value of a growth characteristic at time t. If the change in Y_t is viewed in a purely deterministic fashion, then no further clarification is necessary. But, if Y_t is viewed as a random variable, then for $\{Y_t\}$ to be a growth process, it must have a deterministic component (D_t) that changes over time. For example, we may have $Y_t = D_t + S_t$, where $D_t = f(t)$ and the stochastic component S_t is mean stationary. In other words, if Y_t is a growth characteristic then $\{Y_t\}$ cannot be mean stationary. To illustrate, let us assume that the observed growth process for the education example is

given by {320, 300, 600, 250, 240, 320}. Clearly, there is no growth in the marks obtained by the students. Thus, if education is considered as a growth process, it is necessary to answer the following two questions before proceeding with further analysis of the data.

- Does the existing process serve the intended purpose?
- Is the marks scored in mathematics a proper growth characteristic?

It is important to answer the above two questions since 'education' as a growth process does not have as strong a theoretical foundation as those in the field of natural science.

4.1. Growth models: an introduction and a brief review

4.1.1. Definition

A growth model can be a simple statement like 'growth rate of Y is k unit/hour' and this simple statement may be sufficient for the desired purpose. In other cases, a growth model may take the form of a complicated statistical model. In this study, we shall define a growth model as follows:

Let Y be a proper growth characteristic. A growth model is a mathematical (or statistical) model that describes the change of Y over time and supports interpretations about one or more elements and/or the status of the growth process.

To explain, let us consider the hydrogenation process, the working example of this study. Here the characteristic of interest is melting point (Y) and various elements of the process as described in Chapter 2 are the input oil, the process variables like reaction temperature, and pressure, the control elements like the addition of fresh catalyst in the event of batch failure and of course the process output, i.e. the melting point at time t. The parameters of the growth model either directly or indirectly should tell us something about one or more of these process elements. The aspect of interpretation about the process status is discussed a little later in Section 4.1.3.

4.1.2. Classification of growth models

Depending on the nature of the data, the model structure and the method of development, growth models can broadly be classified as follows:

(i) *Continuous vs. Discrete*: If growth occurs at discrete points of time then the model is called discrete, otherwise it is continuous.

(ii) *Static vs. Dynamic*: let $\{Y_t, t > 0\}$ represents a growth process. If the prediction of Y_t at $t = t_1$ does not require the knowledge of past values, i.e. of Y_t, $t < t_1$, then the model is

called a static model. In dynamic models, such a prediction requires the knowledge of the initial value of Y_t, i.e. Y_t at $t = t_0 < t_1$. The choice between a static and dynamic model is made based on the intended purpose. As already noted in Chapter 1, a dynamic model is needed for the purpose of drift control of a growth process (in fact, any process). In this study, a model of the form $Y_t = f(t)$ is considered as static, while a dynamic model is represented as $Y_{t+\Delta t} = f(Y_t, \Delta t)$, where Δt is the time interval.

(iii) *Empirical vs. Mechanistic*: Empirical models are developed purely based on the data at hand (the so called best-fit models), but the development of a mechanistic model requires the knowledge of the underlying growth process. A simple mechanistic model may be developed based on the overall behaviour of the system (ignoring things happening at the subsystem level). Alternatively, one may also incorporate the effects of the processes at the subsystem level to develop the overall model. Such a model can be developed either following a top-down or bottom up approach [54]. The advantages of a detailed subsystem level modelling are that the model becomes process specific and hence requires less tuning. It can also answer questions related to the factors affecting growth and consequently one may feel more confident about the predictions made using such models. But, the development of such models requires specialised process knowledge, are more complex and in many cases the performance of the model can be studied only through simulation.

Let us illustrate the development of a former type of mechanistic model with the help of a well-known example given by Maino and Kearney [55]: "All living organisms must acquire energy and resources from their environment to fuel metabolism. The transformation of these resources into biomass and their ongoing accumulation is commonly referred to as growth. ....Several generic mechanistic models for animal growth are based on the simple differential equation

$$dm/dt = am^c - bm^d,$$

where a and b are coefficients, and c and d are the scaling exponents of the body mass m. The catabolism (maintenance) exponent d is typically taken as 1, whereas the anabolism (assimilation) exponent is assumed to take values of 2/3 and 3/4 on the basis of.....when the assimilation exponent is taken as 2/3, the von Bertalanffy function for mass m through time t emerges as

$$m = m_\infty[1-(1-(m_b^{1/3}/m_\infty^{1/3}))\exp(-c_v t)]^3,$$

where m_b is mass at birth, m_∞ is asymptotic mass, c_v is the von Bertalanffy growth rate."

The basic growth model for the process of hydrogenation, which is central to this study and is discussed in detail in Section 4.2.3, is developed following a similar approach as above.

Apart from the above classifications, the growth models within each of the above group can be classified further in the usual manner like univariate vs. multivariate, linear vs. non-linear, and whether the parameters of the model are treated as fixed or random.

4.1.3. Practitioner's checklist for describing a growth model

Castellano and Ho [53] propose a list of the following six questions for describing a growth model for its successful implementation in the field of education.

(i) What primary interpretation does the growth model best support?
(ii) What is the statistical foundation underlying the growth model?
(iii) What are the required data features for this growth model?
(iv) What kinds of group level interpretations can this growth model support?
(v) How does the growth model set standards for expected or adequate growth?
(vi) What are the common misinterpretations of this growth model and possible unintended consequences of its use in accountability system?

The answers to these questions for the models developed in this study are given near the end of this chapter. Here, we only note that the two important features of the growth models of this study lie in the answers to the 4th and the 5th questions. What kind of group level interpretation can a growth model support? For the example given in Table 1, one may define a suitable metric of growth for all the students belonging to a particular grade and estimate the mean and standard deviation of this metric. However, here a group may represent only the in-control state of the process. Thus, all the growth curves within the in-control group have the same inference. Finally, the answer to the fifth question is the control chart developed based on a group-level model. If the chosen growth metric does not fall within the control limits, then the process is said to be out-of-control. It is in this sense that the word 'status' has been used in the definition of growth model given in Section 4.1.1.

4.1.4. Applications

From the point of view quality control, a growth model can be used either for the purpose of process characterization, or improvement or control. In the past, growth models have been developed primarily for the purpose of process characterization [56-62] and in a relatively lesser number of cases for supporting process improvement in the long run (e.g. [55]). But the case studies on growth process improvement are very limited [63-64]. It may be noted here

that improving a growth process is expected to be more difficult than improving a process that is expected to be stationary. This is because the optimum level of some of the important control factors of a growth process, instead of being a single value, may itself be a trajectory like the growth trajectory.

So far as process characterization is concerned, growth models have been found useful in a diverse field of applications, as they should given the above definition of growth model. Consequently, growth process modelling is presently a very well developed field and the models proposed are so diverse and process specific that it is difficult to provide even a partial list of these models. So, we only indicate a few areas of applications to indicate the diversity of applications: Growth of forest [56], human population [57], crystal [58], tumour [59], nanowire [60], reliability [61] and a host of models for economic growth [62].

On the other hand, the number of studies on the use of growth models for process control is very limited. A few of these studies [22-24] have already been cited in Chapter 1 and a few more will be noted in the next two chapters.

4.2. Hydrogenation - growth data and preliminary analysis

This section describes the data used for developing the growth models for the process of hydrogenation and presents a summary of the main data set.

4.2.1. Growth and failure data

Two data sets are used for the purpose of developing the growth models. The main data set (that is used for developing the control system) and a supporting data set. The details of these two sets of data are as under.

Main data set: This set of data gives the history of growth of melting point over time for fifty consecutive batches (covering a period of about one month), which is obtained from the past QC records maintained by the concerned organization. In this data set, each growth record gives the following information - the batch number, date of processing, time of fresh catalyst addition for the first time and its amount, time of sampling, corresponding melting point and the time and amount of fresh catalyst addition for the second time, if any. The batch number is a unique identification number, which can be used to trace the supplier of the oil and also for referring to the production log book that records the process conditions used during its processing. The data thus obtained is shown in Appendix A. The last two

observations in this table refer to the time and melting point at the beginning and end of the cooling cycle respectively.

In the sample of fifty batches, there are seven batches that are identified as dead. The sampling intervals in which the batches failed and the end time of the hydrogenation cycle for the active batches gives the failure data (censored). This failure data is used in Chapter 6 for modelling the distribution of failure time, which is an essential input for developing the passive regulation scheme.

Let the growth data be represented as (MP_{ij}, t_{ij}), where MP_{ij} is the melting point of the sample drawn at time t_{ij}, i.e. the time at which the i_{th} sample is drawn from the j_{th} batch, $i = 1, 2, ..., n_j$, $j = 1, 2, ..., n (= 50)$. It may be noted that the time t_i varies from batch to batch and $2 \leq n_j \leq 7$. Further since the last observation does not belong to the hydrogenation cycle, effectively only 1- 6 observations per batch are available for modelling the growth curves. In other words, the data that is available is scanty and is in the form of a highly unbalanced panel data. It is thus acknowledged at the outset that due to this messy data environment, the analysis at times may not be a copybook exercise.

Supporting data set: This data set (given in Appendix B) is taken from [65] and is used for checking the adequacy of the basic growth models.

4.2.2. Summary of the growth data

Figure 4.2 shows the plot of the 50 sample growth curves based on the data given in Appendix A. The unbalanced nature of the data is evident from this plot. It is also seen (from the growth curves having more than three observations) that the growth process is non-linear. Table 4.2 gives a summary of the growth data in the form of a two-way table, which is similar to the one presented in Table 4.1 for the education process. Note that the row variable "Time of sampling" of Table 4.2 is analogous to the variable "Year" of Table 4.1. The column variable "Sample no." in Table 4.2 represents the order of the samples (i.e., first sample, second sample etc.) for a given growth curve. This variable is thus analogous to the variable "Grade" of Table 4.1. The cell entries are the averages of melting point. For example, the average value of 37.96 in the first cell of the first row is the average of all those observations, which represent the first sample for a batch and were drawn before two hours from the start of hydrogenation. The first obvious conclusion that can be drawn from a look at Table 4.2 is that a set of growth data may not always be as nicely structured as in Table 4.1. Nevertheless, it brings out the main features of the process quite nicely. For example, a

perusal of the figures in bold in the successive columns reveals that the initial growth rate is faster than that near the end of a growth cycle. However, the observed drop in melting point from 39.9 to 39.3 as one move from the third to the fourth column may be due to sampling fluctuation, since these cells contain only one observation each.

But what about the status and improvement - the two other concepts related to growth? It is to be noted that if the above table is filled up with experimental data, then under ideal situation, the melting point should increase from left to right and all the rows and the columns will be identical, since sampling is not expected to have any effect on the growth process. Keeping this in view, it is obvious from Table 4.2 that the status of the process is very poor across all the time intervals and naturally there is no sign of process improvement either.

Figure 4.2. Growth curves of fifty sample batches till the beginning of cooling

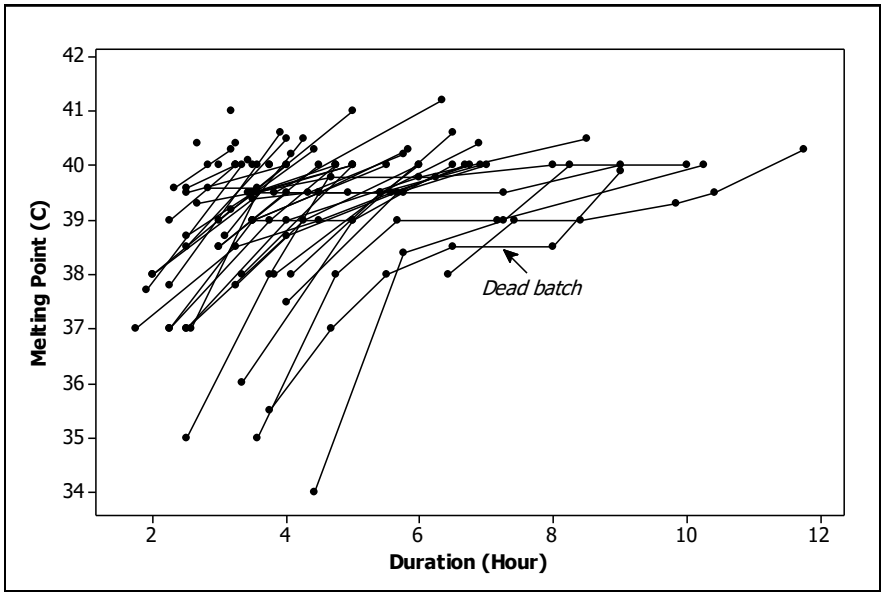


Table 4.2. Growth summary of the hydrogenation data

Sample no.	Time of sampling (hour)								
	<=2.5	2.5-3.5	3.5-4.5	4.5-5.5	5.5-6.5	6.5-7.5	7.5-8.5	8.5-9.5	> 9.5
1st	**37.96**	39.10	37.97						
2nd	39.00	**39.77**	39.52	39.00	39.88				
3rd	39.00	-	**39.90**	40.01	39.66	39.67			
4th				**39.30**	40.00	39.48	40.25		
5th				39.50	-	-	39.25	40.00	40.00
6th						**40.30**	-	39.90	40.00

4.3. Growth models for the process of hydrogenation

The final melting point (MP_f) of a batch of oil can be represented as

$$MP_f = MP_H + \Delta MP_{cooling}, \quad (4.1)$$

where MP_H is the melting point at the end of the hydrogenation cycle and $\Delta MP_{cooling}$ is the growth of melting point during the cooling cycle. The model for growth during cooling is developed in Chapter 6. This section is devoted to the growth models for the active hydrogenation phase.

4.3.1. Approach

The primary purpose of developing the growth models is to use them for process monitoring and regulation. Process monitoring calls for a specification of the in-control state of the process. Once the in-control state is specified then the same can be used for constructing the control charts. It is well known that the in-control state of a process is usually specified in terms of the process mean and its variance. In this case, since the object of control is a curve, it is necessary to specify the mean function and the variance function, if the variance is not constant. But, what should be considered as the process variance or the chance-cause variation? Should it be only the within batch variation or some amount of between batch variation can also be included in it? In the context of profile monitoring, Woodall et al. [27] make the following observations. "Common-cause variation is that variation considered to be characteristic of the process and that cannot be reduced substantially without fundamental process changes". Next, the nature of the variation is qualified as the "variation in the values of the parameters but not variation in the form of the model itself". Then these authors go on to say that "there is common cause variation between profiles if it is not reasonable to expect each set of observed profile data to be adequately represented using the same set of parameters for the assumed model". This last statement is somewhat confusing. Does it mean that any amount of variation in the parameters is permissible, so long as the form of the model remains the same? Clearly, it cannot be so. Perhaps, the expectation here is that the permissible amount of variation can be judged from an examination of the data in the light of process knowledge since the authors conclude by saying that "Process knowledge is always required in the decision whether or not to include profile-to-profile variation as common-cause variation". To summarize, in the opinion of these authors, the common-cause variation is that variation which, as judged by the process knowledge, (i) can be attributed to the variation in the model parameters and (ii) is the

minimum variation to be expected under the existing process design. But this is an extremely difficult task. In fact, the difficulty of the task becomes apparent, if one wants to define the common-cause variation for monitoring within batch variation.

In order to make the task of defining the common-cause variation more tractable, we shall come out of the fixation of minimum variation under a given process design. As an alternative, it is proposed to use the ability to identify the special causes as the basis for judging the permissible amount of common cause variation. This aspect is discussed in more detail in Chapter 5, Section 5.1. In fact, it is argued there that such an approach is expected to be more effective not only for monitoring profiles but also for monitoring a static characteristic.

In this approach, it is thus necessary to select a reasonably homogenous group of growth curves for estimating the process variance without any major violation of the model assumptions. Based on this initial selection, one can then note the out-of-control signals obtained from the chart and examine the extent of process knowledge one has for identifying the special or assignable causes. Depending on the results of such a test at the practical level, one may modify the in-control group, if necessary. Keeping the above in mind, the following approach is adopted for developing the process models.

Step1. Developing a mechanistic model for the growth of melting point during hydrogenation and validating the same using published data.

Step2. Clustering of the sample growth curves

Step3. Developing both static and dynamic models using the curves belonging to one or more of the four groups formed in step 2.

The details of the above three steps are discussed in the following sections.

4.3.2. The monomolecular model

It is noted in Chapter 2 that the Iodine Value (*IV*) of oil is directly proportional to the number of double bonds present in it. Since hydrogenation is a process of breaking the double bonds, the *IV* of the oil decreases progressively during hydrogenation.

In the past, the kinetics of hydrogenation has been modelled using the following first order model [66],

$$-d(IV)/dt = k(IV), \qquad (4.2)$$

where k is the rate constant. The above study is particularly relevant for this work since it makes use of spent catalyst, as is the case here.

Recently a slightly generalized version of the above has been used for the same purpose [67],

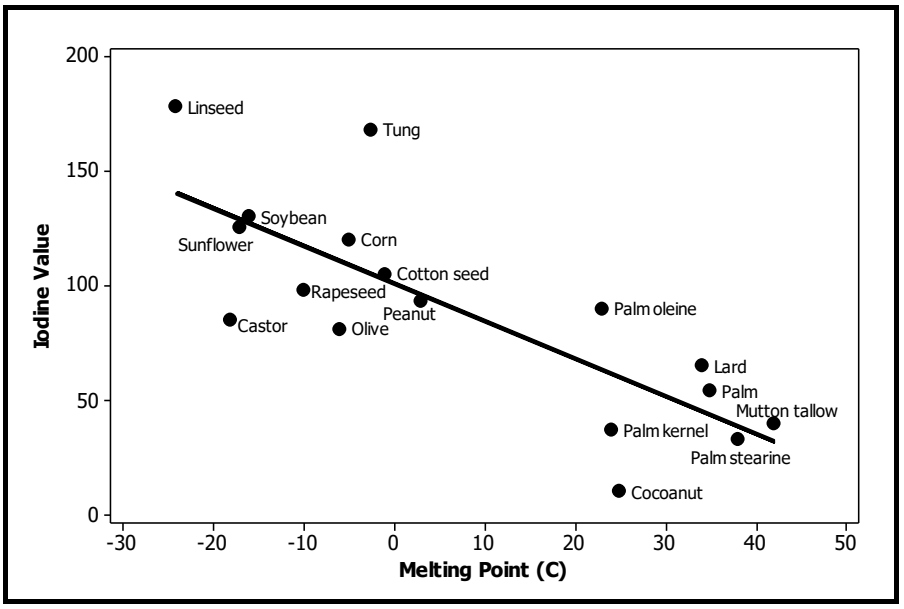
$$-d(IV)/dt = k'(IV)^n. \quad (4.3)$$

In the above, the growth kinetics is presented in terms of the Iodine Value, whereas this study is concerned with the growth of melting point (*MP*). So, the relationship between iodine value and melting point is examined next. The scatter plot of melting point versus iodine value of several different types of oil shows that the two are linearly related (Figure 4.3). In fact, the relationship can be modelled adequately as *IV = 99.9 – 1.60*MP*. Further, an analysis of the growth data reported by Yemiscioglu et al. [65] for partially hydrogenated oil shows that the relationship between the two is indeed linear (Figure 4.4). Thus, assuming that the relationship between *IV* and *MP* is linear, the equation (4.2) gives

$$d(MP)/dt = k(\nu - MP). \quad (4.4)$$

The parameter ν in (4.4) signifies the highest average melting point that can be achieved under a given process condition and will be referred to as the equilibrium melting point MP_{eq}.

Figure 4.3. Scatter plot showing the relationship between iodine value and melting point. *(Source of data: http://journeytoforever.org/biodiesel_yield.html)*



Integrating (4.4) between the initial melting point MP_{in} and the melting point MP_t at time *t*, we have

$$\ln(MP_{eq} - MP_t) = \ln(MP_{eq} - MP_{in}) - kt. \quad (4.5)$$

Simplification of (4.5) gives

$$MP_t = MP_{eq} - (MP_{eq} - MP_{in})\exp(-kt). \quad (4.6)$$

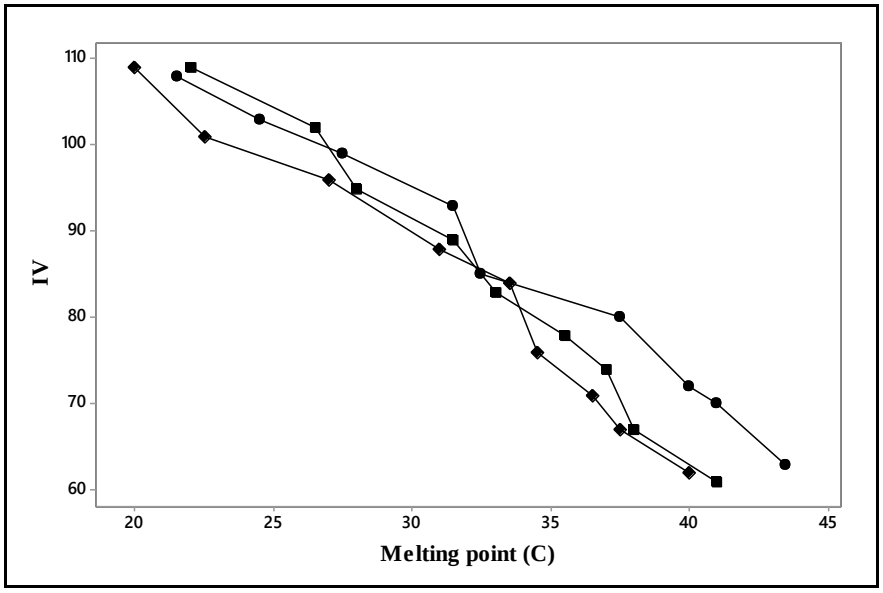
The equation (4.6) is the well-known monomolecular growth model (Spillman and Lang [32]), which is a special case of the Richards growth model [68]. It may also be noted here that instead of (4.2), if the growth rate of *IV* is generalised as

$$-d(IV)/dt = k_0(IV - IV_{eq}),\ IV_{eq} \geq 0, \quad (4.7)$$

then also the growth of melting point is given by (4.6). Integrating (4.7), the model obtained for *IV* is

$$IV_t = IV_{eq} + (IV_{in} - IV_{eq})\exp(-k_0 t). \quad (4.8)$$

Figure 4.4. Scatter plot showing the relationship between iodine value and melting point for partially hydrogenated oil



4.3.3. Validation of the monomolecular model

The parameters of the models (4.6) and (4.8) are estimated using the secondary data set, described above. It is assumed that the errors associated with the observations are uncorrelated and have mean zero and variance σ^2_{MP} and σ^2_{IV} respectively. The estimates obtained (using the commercial software MINITAB) are summarized in Table 4.3. The fits of all the models are similar to the one shown in Figure 4.5.

It is seen from the standard error of the estimates and the graphical fit that the models are adequate. The estimates of the initial melting point of oil are close to each other, as it should be since the same oil is used in all the three cases. The estimate of about 15^0C is also in the known range of melting point of cotton seed oil (though slightly on the higher side). It is also

seen that the value of the rate constants k and k_0 increases with the increase in reaction temperature, which is in line with its expectation (see Table 2.1). It is thus concluded that the monomolecular model can be safely used for modelling the growth of melting point during hydrogenation.

Figure 4.5. A typical fit of the estimated model to the supporting data

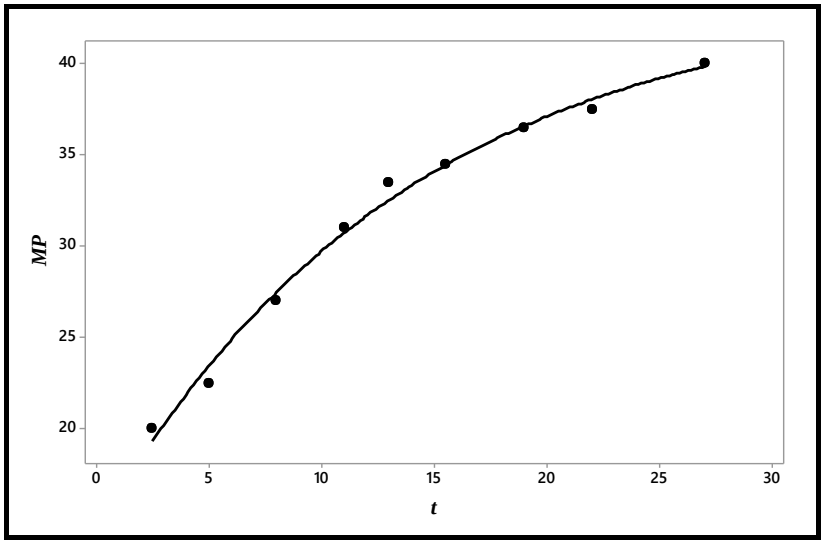


Table 4.3. Estimates* of the parameters of (4.6) and (4.8) for the supporting data

Temp. (^{0}C)	*Model (4.6)*				*Model (4.8)*			
	MP_{eq}	MP_{in}	k	$\hat{\sigma}_{MP}$	IV_{eq}	IV_{in}	k_0	$\hat{\sigma}_{IV}$
120	61.0 (9.7)	16.5 (1.4)	0.012 (0.0004)	0.9	0** (-)	117.8 (1.42)	0.008 (0.0003)	1.6
150	46.4 (3.6)	18.6 (1.3)	0.036 (0.0096)	0.9	22.3 (10.2)	116.5 (1.17)	0.021 (0.0035)	0.9
170	44.0 (1.8)	14.4 (1.1)	0.073 (0.011)	0.7	30.7 (12.2)	117.4 (2.0)	0.039 (0.0093)	1.5

*The figures within parentheses are standard error of the estimates
**Convergent estimate gives a negative value and hence locked at 0^0C.

4.3.4. Clustering of the growth curves

Clustering is a general classification procedure for assigning a set of objects to different classes, where the objects within each class are similar in some sense. Since no training data with known class membership are available as in case of classification, clustering is also called unsupervised classification. However, in some clustering problems, the number of classes may be known. In most other cases, the number of classes is decided based on the

data. Clustering problems arise in a wide variety of fields like pattern matching, image processing, data mining and of course also in process control.

The two most commonly used approaches of clustering are hierarchical and partitional clustering. In partitional clustering, the objects are simply partitioned into several non-overlapping subsets such that each object is in exactly one subset. But in hierarchical approach, the objects are classified into a set of nested clusters that are organized in the form of a tree.

Within the methodological universe of partitional clustering, the *k*-means algorithm (Mac Queen, [68a]) appears to be the most popular. The basic idea behind the *k*-means method is to begin with the assumption of the existence of *k* clusters and to choose the initial centroids of these *k* clusters. Next, each data point is allocated to one of these clusters based on their distances from the chosen centroids. A point is considered to be in a particular cluster, if it is closer to that cluster's centroid than any other cluster. In subsequent iterations a new set of *k* centroids are chosen to improve upon the allocation as judged by certain criteria till convergence. A very popular algorithm in which the centroids at the j_{th} iteration are updated as the means of the clusters formed at the $(j\text{-}1)_{th}$ iteration, is known as Loyd's algorithm [68b]. Presently, a variety of other k-means algorithms are available in the literature. For example one can employ Mahalanobis distance for discovering hyper ellipsoidal clusters [68c], or use the algorithm of Huang [68d] for dealing with categorical data.

The *k*-means technique is a nonparametric procedure. But the parametric approach based on mixture distribution model is also being increasingly used nowadays for the purpose of classification. It is interesting to note that the Loyd's algorithm mentioned above can also be viewed as an EM (Expectation Maximization) algorithm applied to the Naive Bayes model.

Hierarchical clustering methods can be either agglomerative or divisive [68e, 68f]. The former is a bottom-up approach while the latter is based on a top-down approach. In the bottom-up approach, the individual objects are at the bottom and clusters are formed as we go up resulting in a single cluster at the top.

The problem of clustering growth curves comes under the broad class of clustering functional data. For such objects like curves, the similarity measure of objects within a class may be defined either based on the location or the shape of the curves. If similarity is judged by location then one can employ any of the methods described above, but only after making some suitable adjustments to take care of the functional nature of the data. The shape-based clustering is the most natural approach for clustering of curves [69, 69a, 69b]. However, this branch is not as well developed as the location based clustering methods.

From the mathematical and problem solving perspective, the methods of clustering functional data can broadly be classified into four groups [70] - (i) Raw data clustering, (ii) Two-step method, (iii) Nonparametric clustering and (iv) Model based clustering. Such a partitional clustering of the methods is somewhat arbitrary and not very informative. Perhaps a better way to form the clusters (groups) will be to adopt a divisive hierarchical approach. For example, at the first level, the methods may be classified into two groups - namely the clustering of raw and dimensionally reduced data. Raw data clustering can then be subdivided into nonparametric and model based methods, while the methods involving dimension reduction may be classified into two-step and simultaneous methods. In the two-step method, the dimension reduction is the first step and actual clustering is done in the second step.

It may be noted that for the hydrogenation data, the clusters need to be formed based on the raw data, since it is a very low dimensional data set. In particular, we have only a maximum of six observations per curve and there are a large number of cases with either one or two observations (Figure 4.2). Consequently, it is not even possible to fit the monomolecular model (which has three parameters) to many of the curves.

The problem of missing data is common in practice, particularly in a multivariate set up. There are a variety of methods proposed in the literature for handling missing data in the context of functional data analysis. However, in most of these cases, it is assumed that the observations are Missing Completely At Random (MCAR), as defined by Little and Rubin [71]. However this assumption of MCAR cannot be justified for the hydrogenation data. The missing observations at the higher end (say $t > 3$ hours) is indicative of the fact that the average growth rate for the batch is high, which is an important feature of the growth curves.

Liebl and Rameseder [71a] propose a method for classifying functional data in the presence of systematically missing observations, but their method is applicable only when the observed part of the curve is fully observed, i.e. the observations are not noise contaminated and there are no missing observations within the observed part. Our hydrogenation data, as noted above, do not satisfy this restriction. On the other hand, the method of clustering sparsely sampled data (i.e. containing missing observations within the observed part) has been discussed by James and Sugar [71b], but their method is applicable only when the MCAR assumption is satisfied. It thus appears that a suitable method for clustering the growth curves is currently not available in the literature. So, an informal graphical approach has been used here for the purpose of clustering. The steps involved in this three-step method are as follows.

Step 1: Separation of the dead batches.

Step 2: Normal Probability Plot (NPP) of $t40$, the time taken by an active batch to reach the melting point of 40^0C. This step gives us two groups of active batches.

Step 3: Scatter plot of growth rate vs. melting point for one of the two groups obtained in Step 2.

4.3.4.1. Separation of the dead batches

The dead batches are identified through inspection. All the batches for which the two consecutive observations are same or the difference is only 0.1^0C are classified as dead. Of course, identification of a dead batch is not such a trivial exercise (as will be seen in the next chapter). But at this stage, it will suffice.

4.3.4.2. NPP of t40

This is the main step of the clustering scheme, which makes use of average growth rate as a criterion of classification. This metric is chosen based on the following two practical considerations. First, it is an important performance measure of the active hydrogenation phase and secondly, an inspection of the growth data reveals that the value of $t40$ can be obtained directly from the observed data for many batches. For the batches, where the same is not available, it is estimated by adjusting the value of t at which the observed MP is nearest to 40^0C. The adjustments are made using an estimated average growth rate of 0.65^0C per hour around 40^0C. The Normal Probability Plot (NPP) of the estimates of $t40$ thus obtained is shown in Figure 4.6.

It is seen from Figure 4.6 that there is a step in the plot at around 4 hours. Thus the observations are classified into two groups: $t40 \leq 4$ hours and $t40 > 4$ hours. The right panel of Figure 4.6 shows that the NPP of $t40 \leq 4$ hours can be considered as more or less homogeneous. Henceforth, this group will be referred to as Group A.

4.3.4.3. Scatter plot of growth rate vs. melting point

Let the group of active batches for which $t40 > 4$ be denoted by A1. Figure 4.7 shows the scatter plot of average growth rate $[= (MP_{i+1} - MP_i)/(t_{i+1} - t_i)]$ versus average melting point $[= (MP_{i+1} - MP_i)/2]$ for the group A1. It is seen that there are four batches (# 6, 16, 21, and 42), which can be considered as outliers. These four batches are classified into a separate group (Group C). The rest of the batches in this group are classified as belonging to Group B.

Figure 4.6. Normal probability plots of *t*40

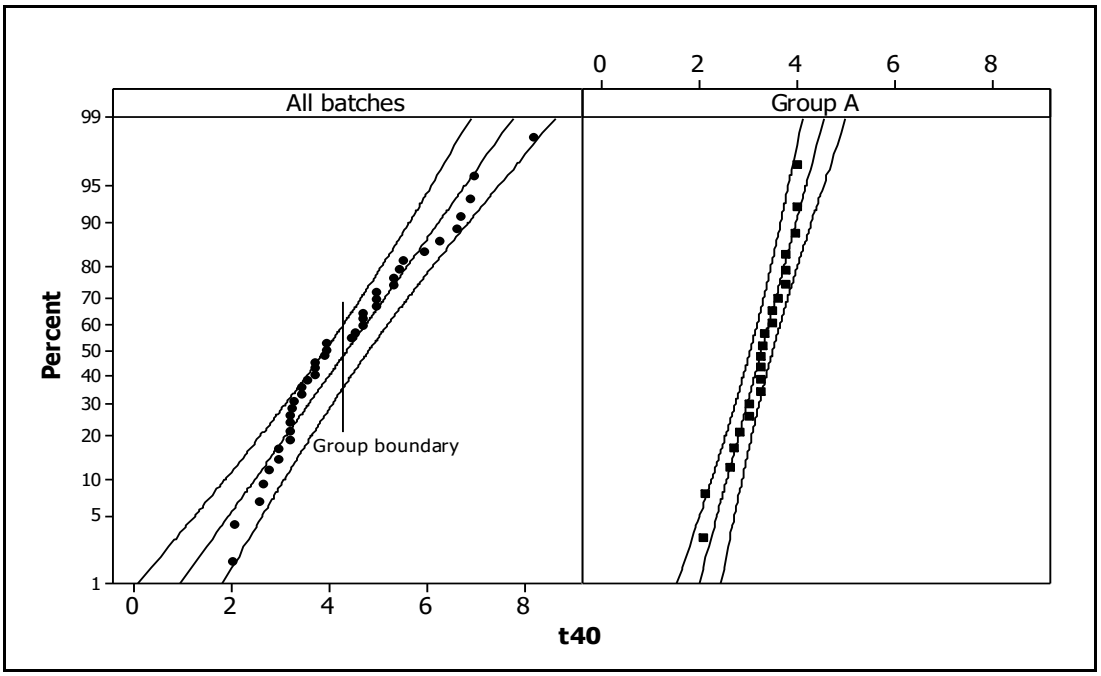
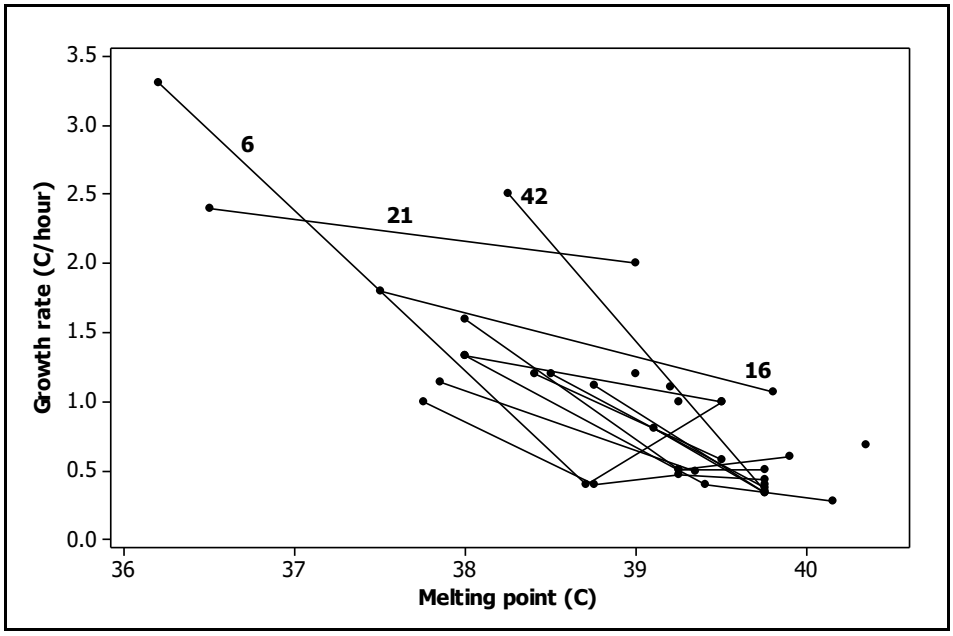


Figure 4.7. Scatter plot showing the linear relationship between growth rate and melting point and the outliers (# 6, 16, 21, and 42)



4.3.4.4. Final result

The clusters obtained following the above three-step-method are shown Table 4.4 and the essence of the approach is illustrated in Figure 4.8. It may be noted that the approach adopted is essentially that of hierarchical clustering, but at each level a different measure of similarity has been used. At the top level, it is a shape based measure of similarity (dead or active); at the second level the measure *t*40 separates the clusters based on their levels, while at the last level the clusters are identified again based on the similarity of shape.

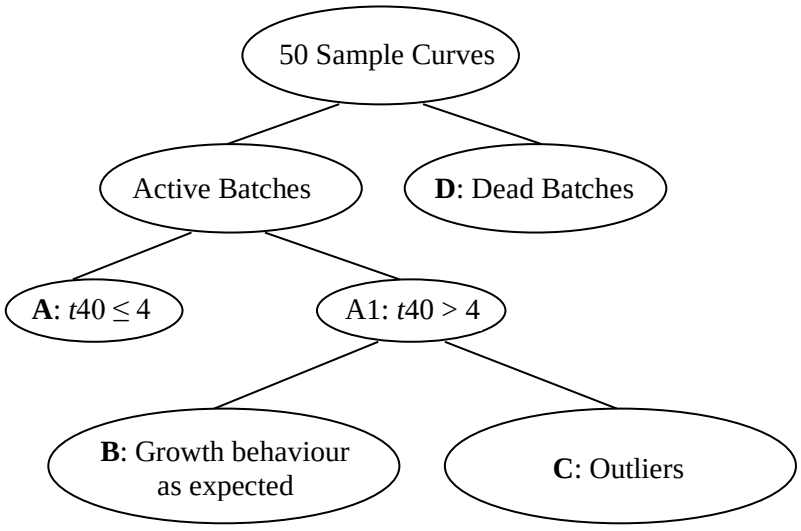
It is interesting to note here that an examination of the oil procurement and production records reveals that the oil used for the Group C batches is likely to have been procured from a different source. However, the causes responsible for the difference between the Groups A and B remain unknown.

Table 4.4. Clustering of the fifty sample batches

Group	*Source of input oil*	*Batch number*	
		Active batches	*Dead batches*
A	Regular	2, 9, 11, 27, 39, 40, 43, 44, 45, 48, [12, 18, 19, 24, 26, 28, 37, 38, 41, 46, 47, 50]	1, 4, 10 20, 22, 30, 31, 32, 34
B	Regular	3, 5, 7, 8, 13, 14, 15, 17, 23, 25, 29, 33, 35, 36, 49	
C	Special	6, 16, 21, 42	

[Batches having only one observation per batch]

Figure 4.8. The clustering scheme



4.3.5. Selection of the in-control group

Since the Group A represents batches with $t40 \leq 4$ hours, it may be selected as a potential candidate for the in-control process without further analysis. Still, for the sake of completeness, the performance of the three groups is compared with respect to both cycle time and catalyst consumption (Table 4.5). The variation of the final melting point is not considered since it will depend significantly on the performance of the drift control system (discussed in Chapter 6). As expected, the performance of the Group A is found to be better than the other two in terms of both cycle time and catalyst consumption. Thus, Group A is selected as a potential candidate for defining the in-control state of the process. It should however be noted that Group A is a suitable candidate for the in-control group only for monitoring an active batch. The in-control group for detecting a dead batch needs different consideration (see Section 4.3.7.1).

Table 4.5. Comparison of performance of the three groups

Group	*Number of batches*	*Cycle time (hrs.)*		*Catalyst consumption (kg.)*	
		Mean	*Standard deviation*	*Mean*	*Standard deviation*
A	22	5.31	0.66	1.75	0.65
B	15	7.73	1.18	1.97	0.95
C	4	7.86	1.75	1.73	0.96

4.3.6. Static models for Group A

Let the statistical model for MP_{ij} of a group of batches be given by

$$MP_{ij} = f(t_{ij}; \beta_j) + e_{ij};\ i = 1, 2, ..., n_j;\ j = 1, 2, ..., n, \qquad (4.9)$$

where f is given by (4.6), β_j is the unknown parameter vector (MP_{in}, MP_{eq}, k) of the j^{th} batch and e_{ij} is the random error term. It is assumed that $E(e_{ij}) = 0$. The parameter vector β_j can be viewed as either fixed or random. If β_j is viewed as fixed, then the parameters of the model may be estimated as in [72], while the procedure popularly known as Non-linear Mixed Effect Model (NONMEM) may be used if β_j is assumed to have a distribution [73, 74].

However, given the problematic structure of the data as discussed above in the context of clustering, it is extremely difficult to develop a NONMEM model for the growth curves. Even otherwise the NONMEM framework does not appear to be particularly attractive for the stated purpose of the growth models. In this framework, it is not only troublesome to estimate

the prediction error variance for a future observation, there is also no straightforward way to judge the adequacy of a group of curves to represent the in-control state of the process. Keeping this in view, it is proposed to adopt the framework of a marginal model, as described below. It should however be noted that the specified model and the method used for estimation of its parameters are far from ideal. It involves a lot of approximations and subjective judgement.

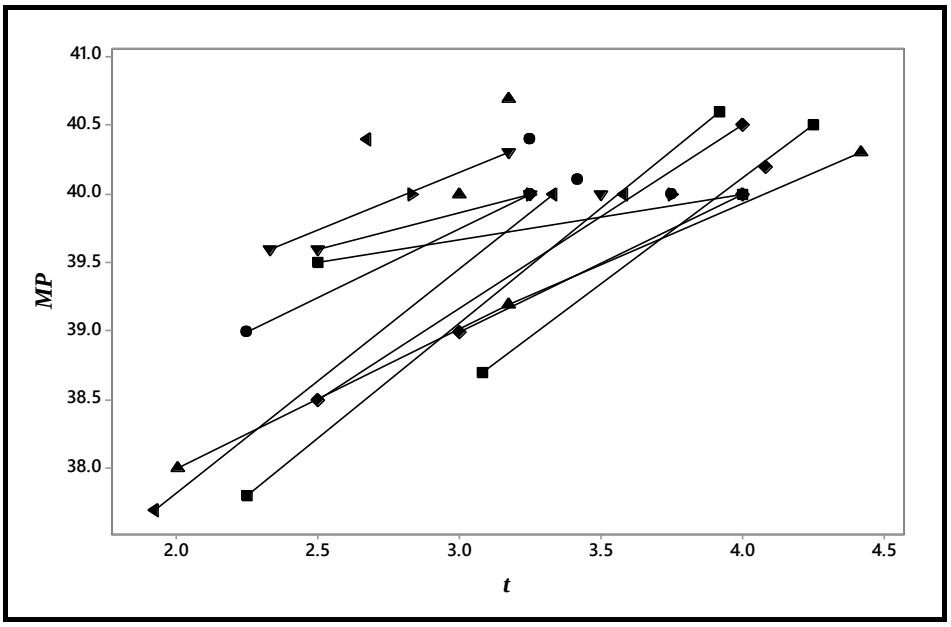
4.3.6.1. Marginal model with monomolecular growth

Let the population level model for the growth curves be given by

$$MP_{ij} = f(t_{ij}, \beta) + \varepsilon_{ij},\ j \in A. \qquad (4.10)$$

where the variance of the error term ε_{ij} now includes both the within and between batch variation. However, it now becomes extremely important to specify an appropriate structure for the error term. Figure 4.9 shows the scatter plot of the growth curves belonging to Group A. Unfortunately, due to the unbalanced nature of the design matrix, the underlying error structure of (4.10) is not brought out well by this figure.

Figure 4.9. Growth curves of group A



In order to have a better idea of the nature of variation exhibited by a group of curves, a small simulation experiment is conducted. Thirty growth curves are simulated assuming that the three parameters MP_{in}, MP_{eq} and k of (4.9) are independent and uniformly distributed

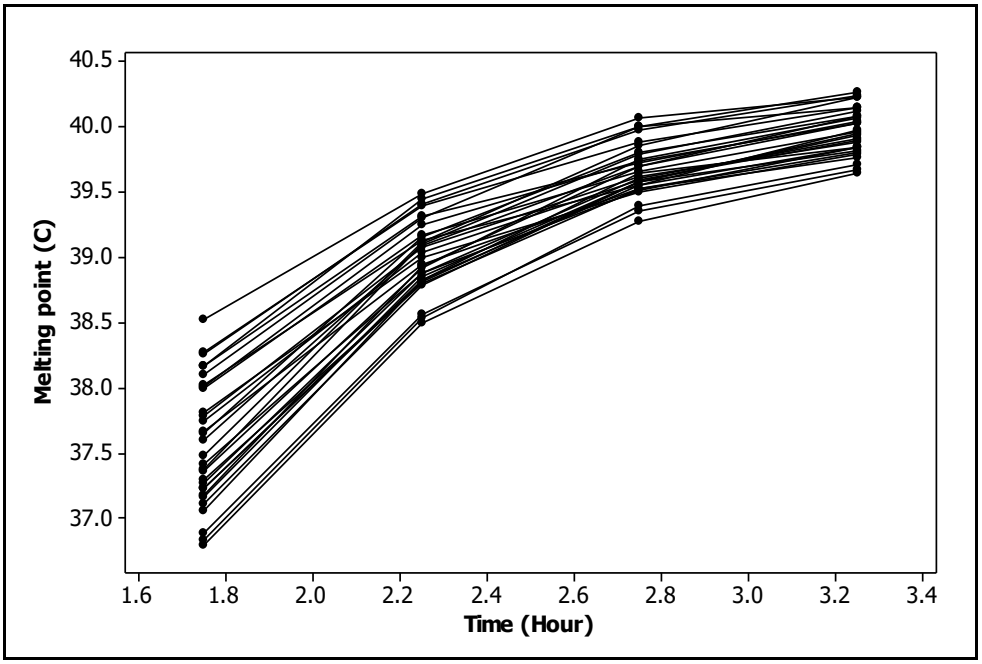
over (-10, 10), (40.0, 40.5) and (1.5, 1.6) respectively. The error term e_{ij} is assumed to be independently distributed as N(0, 0.03). Figure 4.10 shows the plot of the simulated growth curves. It is evident from this figure that (i) the spread of the curves decreases over time and (ii) the curves are spatially separated. Since the assumed error variance in the experiment is very small, there are no curves which run from one corner of the plot to the opposite corner at the other end (as in Figure 4.8 for the observed batches). However, even if such curves are present, one should expect the presence of spatial autocorrelation among the errors of (4.10).

Considering the above, the following statistical model is specified for the growth curves of group A:

$$\begin{aligned} MP_{ij} &= MP_{eq} - (MP_{eq} - MP_{in})\exp(-kt) + \varepsilon_{ij}, \\ \varepsilon_{ij} &= \delta\varepsilon_{(i-1)j} + r_{ij}, \\ \sigma_i &= a - bt_i, \end{aligned} \qquad (4.11)$$

where r_{ij} are assumed to be independently and normally distributed as $N(0, \sigma_i^2)$. The error structure as specified above for the spatial correlation is admittedly very simple, particularly because the time interval between two measurements is not constant. Strictly speaking, it is necessary to make use of a continuous auto regressive process in this case [75]. However, considering the limited and ill structured nature of the data, it is decided to keep it simple.

Figure 4.10. Thirty simulated monomolecular growth curves



The attractiveness of the marginal model for defining the in-control state is apparent from the above formulation. It provides the mean and variance of the process in a straightforward manner. More importantly, the parameter δ, which should ideally be zero, can be used to judge the adequacy of the group to represent an in-control process. Of course, a monitoring scheme based on $\delta \neq 0$ (or any other specification for the correlation structure) can also be developed, provided we are ready to take a sufficient number of samples from each batch and it can be supported by an effective troubleshooting process (see the next chapter for more on this later aspect of process monitoring).

In principle, the six parameters of (4.11) can be estimated following the method of weighted regression. However, in order to simplify the estimation process further, the variance function is estimated first, following the method of grouping (as noted in Chapter 3). This gives (after some trial and error based on subjective judgment)

$$s_i = 1.16 - 0.234 t_i, \qquad (4.12)$$

where s_i is the estimate of σ_i. With such a simplification, it now becomes easy to estimate the remaining four parameters following the method of weighted least squares. The details of this method are described in Appendix C and the estimates obtained are given in Table 4.6.

It is seen from Table 4.6 that the confidence interval of the initial estimate of MP_{in} is very large. This is because there are no observations near $t = 0;$ the minimum value of observed t is 1.92 hours. Accordingly, this set of initial estimates is disregarded and the parameters are re-estimated assuming that $MP_{in} = 0.$ This assumption is made based on technical knowledge of the process. The new set of estimates thus obtained is shown in the last row of Table 4.6, which is judged to be reasonable.

Table 4.6. Estimates of the parameters of (4.11) for group A

$\hat{MP}_{in}$	$\hat{MP}_{eq}$	$\hat{k}$	$\hat{\delta}$	$\hat{\sigma}^2$ *	*Remark*
-32.1 (-186.1, 121.9)	40.30 (40.07, 40.53)	1.72 (0.72, 2.72)	0	0.988	Initial estimate
0	40.352 (40.338, 40.367)	1.45 (1.37, 1.53)	0	0.965	Final estimate

(Approximate 95% confidence interval) * Variance of the weighted model

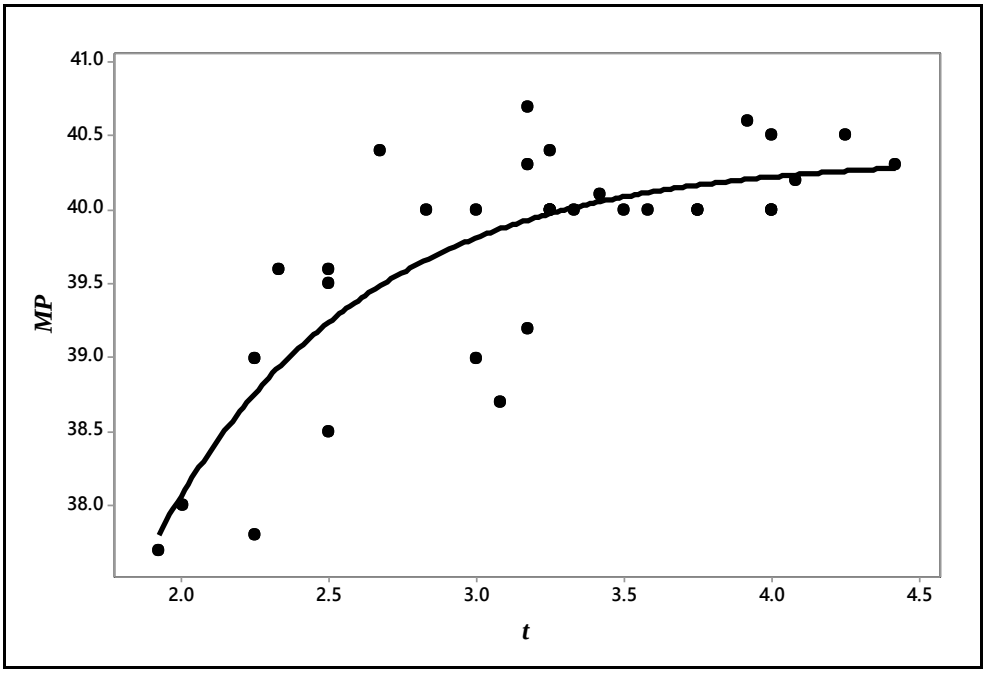
However, the estimate of MP_{eq} appears to be slightly on the lower side, which may be due to the non-availability of observations at the higher end of t. A limited analysis of the effect

of truncation (of the range of *t*) and the fluctuation of the parameters within a batch on estimation of the parameters of a monomolecular model is presented elsewhere [39].

The fit of the final estimate of the model to the data is shown in Figure 4.11. The fitted line plot does not show any obvious sign of lack of fit. A routine diagnostic check of the residuals also does not indicate any sign of model inadequacy.

However, the subjective judgments made in arriving at the variance function do not give sufficient confidence in using it for the purpose of constructing the control chart, which requires a good estimate of the process variance. In fact, the development of the new method for variance function estimation (Chapter 3) is partly motivated by this experience. However, as noted in Chapter 3, the proposed method is neither optimized nor its performance evaluated under such situations as above, i.e. when the mean function is intrinsically non-linear and the variance decreases with the increase in mean. Consequently, as an alternative, it is decided to find a suitable linear approximation of the non-linear model, which is discussed in the next section.

Figure 4.11. Scatter plot showing the fit of the monomolecular model (Group A)



4.2.6.2. A linear approximation of the monomolecular model

It is easy to show (by retaining only the first order terms in Taylor series expansions of te^{-kt} and e^{-kt}) that $t\exp(-kt) = [1-\exp(-kt)]/k$. Next, multiplying both sides of the monomolecular model (4.6) by t, and after some simplification using the above relationship, we have

$$MP(t+1/k) = MP_{eq}t + MP_{in}/k. \tag{4.13}$$

Now, assuming $MP_{in} = 0,$ the equation (4.13) can be written as

$$MP(t+a) = bt, \tag{4.14}$$

where $a = 1/k$ and $b = MP_{eq}$. It should be noted that the estimates of a and b may be significantly different from the estimates of $1/k$ and MP_{eq} obtained from (4.6), since the higher order terms have not been considered in obtaining (4.14). A standard nonlinear regression exercise (using MINITAB) gives $\hat{a} = 0.2$. Henceforth, this estimated value of a will be assumed as known. This gives the following linear approximation of (4.6) for the group A,

$$Z = b't, \tag{4.15}$$

where $Z = MP(t+0.2).$

So let the statistical model for the group A be given by

$$Z_{ij} = b't_{ij} + u_{ij},\ j \in A, \tag{4.16}$$

where the errors u_{ij} are assumed to be independently distributed as $N(0,\sigma_u).$ Since the spatial autocorrelation is found to be zero for the model in (4.11), this aspect is not investigated further. It should however be mentioned that even though $\hat{\delta} = 0,$ it would have been better to consider only one or two observations (which are far apart from each other) from each batch for estimating (4.16). However, due to the scarcity of data, it is not possible to try that here.

The parameters of (4.16) are estimated following OLS to obtain

$$Z_t = 42.35t \ \ [\hat{\sigma}_u = 1.44]. \tag{4.17}$$

The above model is obtained after eliminating batch number 43, since one of the two observations of this batch is found to be an outlier.

Let us now examine the statistical adequacy of the above model. Table 4.7 gives the one-way ANOVA of the batch-wise residuals. It is seen that there is no significant spatial separation of the growth curves.

Table 4.7. ANOVA of batch-wise residuals of (4.17)

Source	*df*	*SS*	*MS*	*F*	*p*
Batch	20	46.05	2.302	1.42	0.289
Error	10	16.20	1.620	-	-
Total	30	62.25	-	-	-

Next, the *SN-LFIT* plot of (4.17) as shown in Figure 4.12 indicates no abnormal pattern. In fact, the 95% confidence interval of the slope of the fitted line in this plot is found to include 20, indicating constancy of variance. Further, it is seen that the residuals fit the Normal distribution well (Figure 4.13). To summarize, the model passes the standard adequacy tests well and hence it can be used for constructing the control chart (Chapter 5).

Figure 4.12. *SN-LFIT* plot of (4.17)

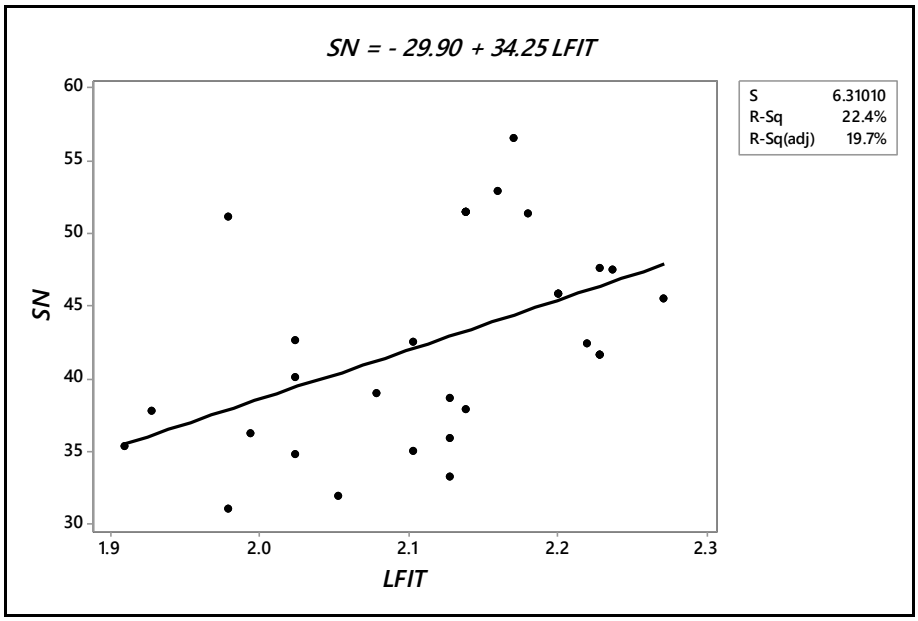
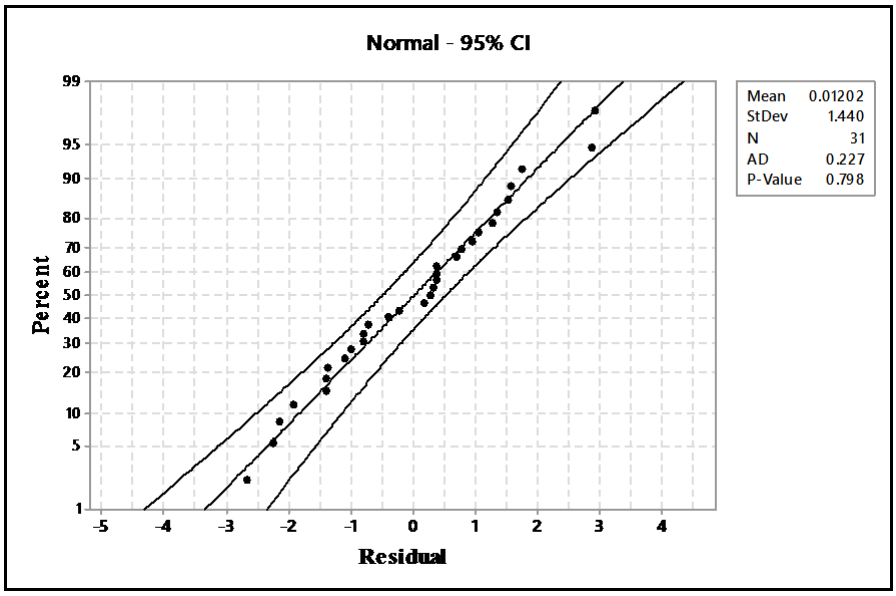


Figure 4.13. Normal probability plot of the residuals of (4.17)



4.3.7. Dynamic models

It has already been noted in Chapter 1 that in order to perform the tasks of detecting a dead batch and controlling the end state, we need to make use of dynamic models. The need for a dynamic model for controlling the end state may be appreciated better by noting that the standard error of the static model (4.17) is 1.44. This means it will be necessary to hydrogenate a batch for 3.4 hours to achieve the average melting point of 400C and an approximate estimate of standard deviation of melting point after 3.4 hours is 0.4. This, in turn, implies that the variation in melting point of the batches at the end of the active hydrogenation phase will be about ± 1.20C, whereas the specified tolerance is only ± 0.50C. This shows that the process cannot be controlled based only on one measurement. The key to exercise better control is to draw samples periodically from the process and use a dynamic model for the purpose of control. Two such dynamic models are discussed in this section, which are just the dynamic versions of the static models (4.6) and (4.15).

4.3.7.1. Data requirement for developing the dynamic models

Let us first consider the type of data that need to be used for developing a dynamic model for the purpose detection of growth failure. Since the process can fail at any time within a

growth cycle irrespective of whether it is a slow or fast growing process, the in-control state is defined by the growth within all the active intervals. All these intervals should represent the entire spectrum of variation of the growth curves expected from the process. To illustrate, let us consider that the five consecutive measurements obtained for a particular batch is given by {37, 38, 38, 39.5, 40}. Thus, the observed growths for the four consecutive intervals are given by 1, 0, 1.5 and 0.5. Now, if the above pattern is obtained without any intervention in between, then all the four growth values should be used for developing the model. On the other hand, if the process is recovered at the end of the second interval (whether rightly or wrongly), then the growth figure of zero should not be used for model development. However, no such intervals are present in the growth data shown in Appendix A.

What about the data requirement for developing a dynamic model for controlling the end state? This will depend entirely on the nature of the process and individual preference. A very general approach is to update the growth rate at the end of each interval. The other extreme is to use a single predetermined model till there is enough evidence to believe that the dynamics of the process has changed. In between these two extremes, one can identify the clusters (similar to the one used for developing the static model) and develop separate models for each of the clusters and to have a rule for switching from one cluster to the another (e.g. switching from an initial high growth regime to a low growth regime in the event of a point falling below the lower control limit of a suitable control chart). In the present case, only a single model is proposed to be used for controlling the end state. This may seem unreasonable since we have already identified clusters based on $t40$. But $t40$ is a measure of average growth rate between the initial melting point ($\approx 0^0$C) and 40^0C, whereas the feasible control region is from about 37^0C to 40^0C. It is shown that within this operating range there is no significant difference in the growth rate between the two groups.

4.3.7.2. Dynamic model corresponding to the static linear model

The linear model (4.15) gives

$$Z_{t+\Delta t} = Z_t + d(\Delta t),$$

where d is the growth rate of Z. It is to be expected that the growth rate of the two groups A and B will be different, with $d_A > d_B$. Thus, the statistical model specified for estimating d_A is

$$Z_{t+\Delta t} = Z_t + d_A(\Delta t) + \varepsilon_{\Delta t}, \qquad (4.18)$$

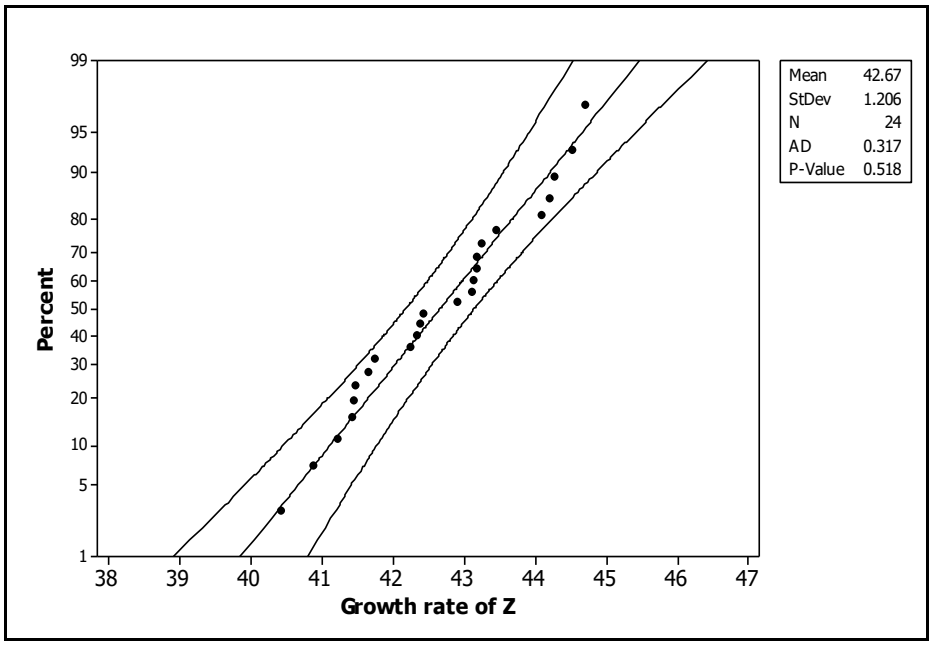
where the error term $\varepsilon_{\Delta t}$ signifies that its variance is dependent on Δt. Similar model is specified for the other group. An analysis of the residuals of the model for group B shows

that the *SN* function is constant, as is to be expected. However, the variance function for the group A cannot be studied since the sample size is very small. Assuming that the standard deviation of growth, $\Delta Z = Z_{t+\Delta t} - Z_t$, is proportional to Δt, the weighted regression of ΔZ vs. Δt gives $\hat{d}_A = 42.9$ and $\hat{d}_B = 42.5$ with standard errors as 0.35 and 0.23 respectively. In other words, the difference in growth rates for the two groups is marginal. This suggests that the growth curves are nearly parallel in the data range ($t \geq 1.92$ hour). So it is decided to develop a common dynamic model for all the batches.

However, the development of a common model as above runs into a problem - the regression of ΔZ vs. Δt shows that the intercept is significant. So, it is decided to examine the suitability of treating *d* as random (instead of fixed as above). Unfortunately, it appears that the variation of *d* is somewhat lower for the larger intervals compared to the smaller ones. Although the evidence is not conclusive, it is decided to limit the value of Δt to a maximum of 1.5 hour. However, this is not a serious restriction, since most of the sampling intervals are expected to lie within this limit. It is seen from Figure 4.14 that the observed values of *d* for $\Delta t \leq 1.5$ give a good fit to the Normal distribution. The distribution of *d* without any restriction on Δt also gives an equally good fit to the Normal distribution. But to be on the safe side, the following dynamic model is specified for the process,

$$Z_{t+\Delta t} = Z_t + d(\Delta t),\ \Delta t \leq 1.5,$$
$$d \bigcap N(42.67, 1.21). \qquad (4.19)$$

Figure 4.14. Normal probability plot of *d*



4.3.7.3. Dynamic model corresponding to the monomolecular model

The differential equation (4.4) that describes the dynamics of the monomolecular model can be approximated by the following forward difference equation

$$MP_{t+\Delta t} - MP_t = d_0(\Delta t) - d_1(MP_t)(\Delta t). \tag{4.20}$$

This gives us the following random walk model with variable drift,

$$MP_{t+\Delta t} = MP_t + G \mid (MP_t, \Delta t), \tag{4.21}$$

where G is the growth during a time interval Δt,

$$G = d_0(\Delta T) - d_1(MP_t)(\Delta t). \tag{4.22}$$

Thus, the main task is to estimate the parameters of (4.22). To begin with, the observed growth data are used to obtain a set of 51 observations on $(G, MP, MPS, \Delta t, t - t_0)$, where t_0 is the time of fresh catalyst addition for the first time and $MPS = (MP_t - 38.488)/1.082$ is the standardized value of *MP*. This set of 51 observations represents only genuine active intervals. Using this set of data, the parameters of (4.22) are estimated as follows. Let the statistical model for G is given by

$$G_{ij} = d_0(\Delta t_{ij}) - d_1(MP_{ij})(\Delta t_{ij}) + \varepsilon_{ij}, \tag{4.23}$$

where $\Delta t_{ij} = t_{(i+1)j} - t_{(i)j}$ and MP_{ij} is the melting point at the beginning of the i^{th} sampling interval of the j^{th} batch. It is assumed that ε_{ij} are uncorrelated, have mean zero and variance σ_{ij}^2. The assumption of ε_{ij} being uncorrelated is reasonable since it has already been noted that there is no significant difference in growth rate (d) between the two groups. For the sake of convenience, it is assumed that the data are in stacked form and the above model can then be written as

$$\begin{aligned} G_i &= d_0(\Delta t_i) - d_1(MP_i)(\Delta t_i),\ i = 1, 2, ..., \textstyle\sum_j (n_j - 1), \\ \sigma_i^2 &= g(MP_i, \Delta t_i, t_i - t_{0i}), \end{aligned} \tag{4.24}$$

where t_{0i} is the time of catalyst addition, which is same for all i representing the same batch.

The corresponding *SN* function (as defined in Chapter 3) is given by

$$SN_i = h(MP_i, t_i - t_{0i}). \tag{4.25}$$

Note that the *SN* function is simpler than the variance function, since the former has one parameter less than the latter.

The *SN* function (4.25) is estimated following the method described in Chapter 3. Using the estimate of the *SN* function, the mean function is then estimated following the method of WLS. This gives

$$E(G) = 0.852\,(\Delta t) - 0.4563\,(MPS)\,(\Delta t), \tag{4.26}$$

$$SN(G) = 152.351 - 3.915(MP_t) + 3.38(t - t_0). \tag{4.27}$$

However, it is recognized that the estimates of G given by (4.26) are biased in the sense that

$$G\,|\,(MP_t, \Delta t) > G\,|\,(MP_t, \Delta t_1) + G\,|\,(MP_{t+\Delta t_1}, \Delta t_2),\ \Delta t = \Delta t_1 + \Delta t_2. \tag{4.28}$$

Consequently, the growth path becomes dependent on the choice of $\{\Delta t_i\}$. Of course; there are processes such as manual cooking of rice, where the growth after sampling may slow down due to the opening of the lid for the purpose of sampling. But that is not the case here. It is extremely important for our purpose to develop an appropriate method to take care of this bias. So an innovative method based on the approach of instrumental variable estimation is used for estimation of (4.23). It will be shown that the proposed method provides nearly unbiased estimates of G. The instrument is constructed as follows: Let an interval Δt_i be partitioned into n_1 number of smaller intervals of length 0.5 hour and another interval of length $l_i < 0.5$ hour such that $\Delta t_i = 0.5 n_1 + l_i$ and g_k be the estimate of growth obtained from (4.26) for the k^{th} subinterval. Then the instrumental variable is defined as $I = \sum g_k$.

Having constructed the instrument I, the estimate of G is obtained following the method of two stage regression. The estimates of the components of the two-stage model are obtained as

First stage model:

$$E(G_1) = -0.0482 + 1.2819 I, \tag{4.29}$$

$$SN(G_1) = 20.99, \tag{4.30}$$

where $G_1 = \hat{G}$ is given by (4.26).

Second stage model:

$$E(G^I) = 0.0805 + 0.9089 \hat{G}_1, \tag{4.31}$$

$$SN(G^I) = 113.091 - 2.85(MP_t) + 2.57(t - t_0), \tag{4.32}$$

where the superscript *I* is used only to distinguish it from the estimates obtained from (4.26) and (4.27). As before, the mean function (4.29) and (4.31) are estimated following the method of WLS.

The average error variance of $\hat{G}^I$ is found to be 0.162, which is slightly higher than the error variance of the estimates obtained from (4.26). But there is a big gain with respect to the bias of the estimates. Table 4.8 shows a comparison of the estimates of growth obtained from the two equations under two conditions - using a single interval of length Δt and two intervals of length $\Delta t/2$. It is apparent from the table that the instrumental variable approach has reduced the bias significantly. Although, the bias is not completely eliminated, it gives a workable model for $\Delta t \leq 2$ hour.

Table 4.8. Comparison of the estimates of growth

MP_t	*M1: Equation (4.28)*		*M2:Equation (4.32)*	
	$\Delta t = 2$	$\Delta t = 1+1$	$\Delta t = 2$	$\Delta t = 1+1$
37	2.963	2.337	2.541	2.465
38	2.118	1.671	1.827	1.778
MP_t	$\Delta t = 1.5$	$\Delta t = 0.75+0.75$	$\Delta t = 1.5$	$\Delta t = 0.75+0.75$
37	2.222	1.870	2.117	2.055
38	1.589	1.337	1.524	1.486

Finally, in order to examine the implication of the *SN* function (4.32), let us assume that the melting point increases from 35^0C at (t - t_0) = 0 to 40^0C at (t - t_0) = 4. The *SN* function shows that the signal-to-noise ratio of the process deteriorates significantly from 13.3 to 9.4 due to the increase in melting point. This means that the controllability of the process is poor near the target melting point of 40^0C. Of course, the controllability can be improved by increasing t, i.e. by slowing down the process.

4.3.8. Description of the proposed growth models from a practitioner's point of view

A practitioner's checklist, as proposed by Castellano and Ho [53], for describing a growth model is given at the beginning of this chapter, in Section 4.1.3. This checklist consists of six questions. Here we answer the six questions in this checklist for describing two of the four models developed in the previous sections, since only these two are finally used for developing the integrated monitoring and regulation scheme. The two models are (a) the linear static model described in Section 4.2.6.2 and (b) the non-linear dynamic model described in Section 4.3.7.3.

Q1: What primary interpretation does the growth model best support?

(a) The mean and variance of melting point at time t of an in-control process; (b) expected growth of melting point and its variance from a given melting point over a given time interval, assuming that the batch remains active throughout the interval.

Q2. What is the statistical foundation underlying the growth model?

(a) The static model is the marginal model for a group of growth curves; (b) the dynamic model is a random walk model with variable drift.

Q3. What are the required data features for this growth model?

Observational data on (MP_{ij}, t_{ij}) for a group of growth curves representing the present status of the process, where both MP and t are continuous variables. Although the data used for developing the models is highly unbalanced, it is preferable to have at least three observations for each curve.

Q4. What kinds of group level interpretations can this growth model support?

Both the models support only group level interpretations. The individual level interpretations are obtained by comparing observations with the group level estimates.

Q5. How does the growth model set standards for expected or adequate growth?

The standards are set by using the models for constructing appropriate control charts.

Q6. What are the common misinterpretations of this growth model and possible unintended consequences of its use in accountability system?

The common misuses are likely to come from the usage of the models for predicting the melting point or the growth in melting point for (i) an individual batch, (ii) an arbitrary group of batches and (iii) a situation where the ranges of MP, t and Δt are far beyond the data range (Appendix A).

4.4. Proposed instrumental variable method and its generalization

The use of instrumental variables and its construction based on observed data is a well-established procedure for estimating the parameters of a dynamic panel model (Anderson and Hsiao [77]). However, here the source of inconsistency is fundamental in nature, since our objective is to obtain a continuous time model based on discrete time data. Note that as $\Delta t \to 0,$ the least square estimates of the parameters of (4.24) will tend towards their true values, i.e. the estimates are asymptotically consistent (Bigi et al [78]). However, the estimates of the parameters of the model, as shown in (4.26) are obtained with $\Delta t >> 0.$ Hence the need for instrumental variable estimation. The basic idea behind this method of estimation is as follows. Let us first note that the observed values of melting points are

generated by a continuous time process. So the initial estimate of the mean function (4.26) is used to generate pseudo observations that mimic a continuous time process. These pseudo observations are represented by the instrumental variable *I*. Next, the estimates obtained from (4.26) are corrected by the features of continuity at the first stage. The estimates obtained from the first stage model can then be considered as observations that incorporate the features of continuity. So these estimates are then calibrated in the second stage by the original observations generated through a continuous time process to obtain the final estimates.

It should also be noted that the above procedure does not give a general purpose continuous time dynamic model. It only allows us to obtain a reasonable model within the working range of $0.5\text{h.} \leq \Delta t \leq 2.0\text{h.}$

It is however acknowledged that the sub-interval length of 0.5 hour that has been used in this study is obtained through a little trial and error. Also the form of the growth model, in the present case, can be considered as more or less known. But, in general, this may not be the case. Further studies are needed to examine the above aspects to generalize the proposed approach.

4.5. Summary

To begin with, this chapter gives a brief introduction to growth models and growth process modelling. This is followed by a detailed description of the growth models developed for the hydrogenation process. The basic growth model, i.e. the monomolecular model is derived from theoretical consideration. The appropriateness of the model is also validated with published growth data before using it for modelling the growth data of this study. Since the growth models are developed primarily for the purpose of process control, all the models developed are for a group of batches. The first group level model that is developed is for a group of monomolecular growth curves, where the group is selected based on an informal graphical approach of clustering of the fifty sample batches. However, due to the lack of confidence in the estimate of the variance function associated with this model, it is not used for further analysis. So, an alternative linear model based on an approximation of the monomolecular model is developed. This model is judged to be adequate for defining the in-control state of active batches.

Next the development of two dynamic models corresponding to the two static models as above is discussed in detail. An innovative method of instrumental variable estimation is used for estimating the parameters of the latter model. The proposed method of variance function

estimation (Chapter 3) is also used here for estimating the parameters of this model following the method of weighted least squares. It is shown that the proposed instrumental variable approach gives nearly unbiased of growth from a given melting point over a given time interval.

Chapter 5 Shift control*

5.1. Basic premise of SPC

Shewhart [1] characterized the basic premise of SPC with his three famous postulates. Based on these three postulates, the classical definition of an in-control process is usually stated as follows: A process is said to be in-control (or under statistical control), if no assignable causes are present in the process (or conversely, if the process is under the influence of chance causes only). A control chart is a statistical tool for detecting the presence of assignable causes in a process.

The third postulate of Shewhart [1] states that "assignable causes of variation may be found and eliminated". The implication of this postulate is that if an out-of-control signal from a control chart cannot be attributed to a removable cause then it is not an assignable cause and hence one cannot say anything about the state of control of a process even though the chart gives an out-of-control signal. In other words, the effective rate of false alarm may be much higher than that obtained from theoretical considerations. Of course, there can be occasional failure to identify and remove an assignable cause. But, in general, there is a need to be adequately conservative while defining the in-control state of a process. Unfortunately, the 'removable' character of an assignable cause is often ignored in practice. This lack of appreciation for the need to be conservative may be one of the reasons for the gap between theory and practice of SPC, as noted in [79]. Accordingly, an effort has been made here to take a more practical approach, which may not be always fully in-line with some of the prevalent practices. In particular, no effort is made here to have one chart at the beginning of the implementation of SPC based on the so called inherent variation of the process. Rather, a sequential approach is adopted. For example, if hour-to-hour variation becomes unmanageable then one can study the day-to-day variation and if even this becomes problematic then an examination of the month-to-month variation may be more appropriate. Once stability is achieved over month-to-month variation, then the control chart should be redesigned for the next level and so on.

The following two articles are based on this and the previous chapter: (i) S. Chakraborty & P. Mandal. “SPC based on growth models for monitoring the process of hydrogenation of edible oil”. *Journal of Food Engineering, 146*(2015) 192-203, (ii) S. Chakraborty & P. Mandal. “Control charts for detection of dead batches during hydrogenation of edible oil”. *International Journal of Quality & Reliability Management, 36*(2019) 1804-1820. However, a few additional and related issues are also discussed in these two chapters.

In order to implement the above liberal view of an in-control process, it may be useful to think in terms of the sampled process rather than the process itself. So let $(y_{1j}, y_{2j}, ..., y_{pj})$, $j = 1, 2, ..., n,$ is a random sample representing the process with respect to the characteristic of interest *Y*. The n samples are drawn over time. The *p* measurements in each sample may be either on p different variables (multivariate *Y*), locations or time points. The basic premise of SPC may then be stated as follows:

(i) The in-control state of a process can be defined in terms of predictable distributions of y_{ij} $\forall i, j.$ In practice, it will suffice to be able to predict the mean and variance (finite) of y_{ij} and have an approximate idea about the statistical form of its distribution.

(ii) The out-of-control states can be attributed to assignable causes (in almost all cases). In other words, it should be possible to identify and eliminate the root cause(s) of the out-of-control states in an economic manner based on the available process knowledge and other resources.

It is obvious that the control chart cannot be made to give signals only when the cause is removable. The point is, one cannot respond to too many out-of-control signals at a time. Moreover, it is also necessary to verify the effectiveness of the corrective actions. By limiting the number of out-of-control signals, the chart only facilitates these tasks of removal and verification. Otherwise, the exercise can be demoralizing.

Thus, the classical definition of an in-control process given at the beginning of this section is considered as the ultimate goal of SPC and the above is a practical framework for achieving the same. It is expected that if the above two requirements of predictability and removability are kept in mind while designing a control chart then it will have a greater chance of success. In other words, a control chart is best viewed as a tool for testing the amount of process knowledge one has in obtaining predictable process output. Of course, the lower the output variation for a given process design, the greater is the process knowledge.

5.2. Literature review: from Shewhart charts to growth process monitoring

There are three basic elements of a control chart, which distinguishes one control chart from another. These three elements are (i) the sampling scheme, (ii) the control statistic and (iii) the out-of-control rules. The designer of a control chart selects the above elements judiciously to deal with various situations.

The control statistic (Q) plays a central role in designing a control chart. It is important to choose a control statistic that is simple and easy to interpret, apart from having favorable sampling properties. The value of Q is computed based on the sample observations and other information, which may be either known (e.g. a model describing a physical law) or assumed to be known (e.g. an estimate of a parameter is treated as known).

Various types of control charts have been developed in the past to deal with a wide variety of situations encountered in practice. A part of this rich body of work is reviewed here in brief under three sections: (i) Basic control charts, (ii) Advanced SPC and (iii) Growth process monitoring. The objective is to cover only the main thread of development of SPC. Thus, we begin with the control charts based on the Shewhart model and then the subsequent developments to deal with situations when the assumptions of the above model are violated. Other parallel developments like nonparametric, Bayesian and economic design of control charts or extensions like the charts for non-normal processes or those based on big data (e.g. data acquired through multiple on-line sensors) are not discussed. Also, a discussion on the evaluation and comparison of the performance of various types of charts is kept outside the scope of this review.

5.2.1. Basic control charts

Let Y be the characteristic selected for monitoring a process using a suitable control chart. Also let the in-control state of the process be given by

$$Y_t = \mu + u_t, \tag{5.1}$$

where, μ is the process mean and u_t are independently distributed as $N(0, \sigma^2)$. The basic control charts for variables are designed to monitor processes for which (5.1) holds at least approximately. The popular among these charts are the following set of univariate charts for continuous data: X, $\bar{X} - R$, $\bar{X} - s$, *CUSUM* (Cumulative Sum) and *EWMA* (Exponentially Weighted Moving Average) charts. A similar set of basic charts for monitoring defectives and defects are (*p*, *np*) and (*c*, *u*) charts respectively. The construction, use and properties of these charts are discussed in great detail in standard text books on SQC (e.g. [80]).

5.2.1.1. Basic multivariate control charts

Often, more than one characteristic is measured on each unit and one may like to monitor all of them. A straightforward solution is to use a separate univariate chart for each of the characteristic. However, such an approach is not only inconvenient but also inefficient. A

better alternative is to make use of a multivariate chart. The popular multivariate charts are T^2 [81], multivariate *CUSUM* (*MCUSUM*) [82] and multivariate *EWMA* (*MEWMA*) [83] charts. Several other variants of *MCUSUM* have also been proposed subsequently. It may be noted that the above three charts are multivariate counterparts of $\overline{X}$, *CUSUM* and *EWMA* charts respectively.

The main difficulty in using the above and for that matter any other multivariate control chart is that the results are very hard to interpret. Although methods have been proposed to facilitate interpretation [84], the difficulty still remains. Moreover, the proposed methods are very complicated – perhaps more complicated than the construction of the chart. However, on the positive side, it appears that if a series of out-of-control signals are interpreted correctly then the gain in process knowledge is expected to be much greater than what will be possible otherwise.

5.2.2. Advanced SPC

In the real world, it is hard to find processes, for which the in-control state of the process can be represented as in (5.1). The two most commonly observed violations of the assumptions of (5.1) are the following: (i) Y_t are autocorrelated and (ii) The mean and/or variance of Y_t varies systematically over time or space. The problem arises when such systematic components of variation are to be considered as inherent feature of an in-control process for the purpose of shift control. This section gives a summary of the methods available in literature to deal with various such situations.

5.2.2.1. Autocorrelated process

In our language, SPC for an autocorrelated process can be thought of as *shift control in the presence of stochastic drift*. It is extremely important to recognize this drift for the purpose of designing a control chart for such processes. It has been shown by Bethouex et al. [85] and later by many others that the a control chart designed under the assumption of independence gives far too many false alarms in the presence of even moderate level of autocorrelation. There are three basic approaches two deal with autocorrelation. An obvious way is to increase the time interval between the samples so that the sampled process behaves like (5.1). Alternatively, one can adjust the control limits [86]. However, the above two approaches can be used only when the underlying process $\{Y_t\}$ is stationary. The general approach to deal with autocorrelation is to develop an appropriate time series model like an Autoregressive-

Integrated-Moving-Average (ARIMA) model to capture the autocorrelations and monitor the uncorrelated residuals of the model using an appropriate control chat [87]. However, such residual based charts may not be efficient in detecting small shifts [88].

The well-known EWMA charts also perform well in the presence of autocorrelations if the process can be represented by a first order integrated moving average model [89].

5.2.2.2. Systematic variation of process mean

Let the model of the process be given by $Y_t = f(X_t) + e_t$, where the mean of the process depends on *X*, which varies over time. Then the general procedure for monitoring *Y* is to develop the functional relationship f (if unknown) and then monitor the residuals e_t using an appropriate control chart. Depending on the nature of the residuals, an appropriate control chart may be selected from among those mentioned above.

If the variables *X* are controllable and are associated with the sub-processes preceding the one under investigation, then the pair of control charts *X*-*e*, for monitoring *X* and *e* respectively, is called 'cause selecting control charts' [90, 91].

The above regression-based control charting procedure can be generalized to situations where, for an in-control process, a particular relationship $Y_i = f(X_i)$ within a given domain of *X* is expected to be repeated over time. Thus, the data in this case are represented as $y_{it} = f(x_{it}; \beta_t) + \varepsilon_{it}$, where y_{it} is the observation made at time *t*, keeping *X* at x_i; β is the parameter vector of the model and ε_{it} satisfy the assumptions of classical SPC. In the SPC literature, such data for a given t is said to describe a 'profile'. The profiles thus obtained, which may be either linear or nonlinear may be monitored either by monitoring the residuals $\hat{\varepsilon}_{it}$ [92] or by monitoring the parameter vector $\hat{\beta}_t$ [93-98]. Usually, a multivariate chart is used for the purpose of monitoring the profiles. However, since a multivariate chart is, in general, difficult to interpret, univariate charts may also be used for monitoring each of the parameters (including process variance) separately. Other approaches for profile monitoring that have been proposed include (i) monitoring a linear profile based on change point model [99, 100], (ii) monitoring complex non-linear profile using wavelet transform [101, 102] and (iii) use of both wavelet and B-splines for monitoring a nonlinear profile [103].

All the above methods involve modeling the profile in some form. However, a T^2 chart based on the original observations y_{it} has also been used for monitoring a linear profile

[104]. The feasibility of such an approach for monitoring a monomolecular growth curve is also examined here (see section 5.3).

A generalized version of profile monitoring is surface monitoring (Zang and Qiu [105]). The principle used here is the same as above. However, the methodology is usually computation intensive.

5.2.3. Shift control of a growth process

In the last two decades or so, profile and surface monitoring have been the most popular area of research in the field of SPC. This is exemplified by the fact that as many as 22 articles on this topic have been published in *Journal of Quality Technology* during 2003-2018, which is about 13.6% of all the articles on SPC and related subjects (excluding measurement system analysis). The corresponding figure for the journal *International Journal of Quality and Reliability Management* is 10.6% for the period 2011-2018. The contributions to other important process situations like healthcare (Thor et al. [106]) or multistage processes (Shu et al. [107], Shang et al. [108]) and attributed network (Mostafa and Paynabar [109]) have been far less in comparison. The situation with respect to growth process monitoring is even worse despite its great practical importance. So a legitimate question is - will growth process monitoring emerge as the new frontier of SPC in near future? Only time will tell.

There are two aspects to monitoring a growth process: (i) Monitoring an active batch and (ii) Detection of growth failure. The past work on these two topics, which is very limited, is reviewed in this section.

5.2.3.1 Monitoring an active growth process

Let the growth process of interest has a beginning and an end. The evolution of the process within these boundaries will be called a growth cycle. The growth cycles are repeated over time. It is also assumed that the characteristic grows monotonically within each cycle. Thus, depending on the length of the cycle and the existing status of the process, it may be decided to monitor either the within or between cycle variation. However, note that even when it is decided to monitor the within cycle variation, it is possible to have some idea about the between cycle variation when the results for several cycles are plotted on the same chart.

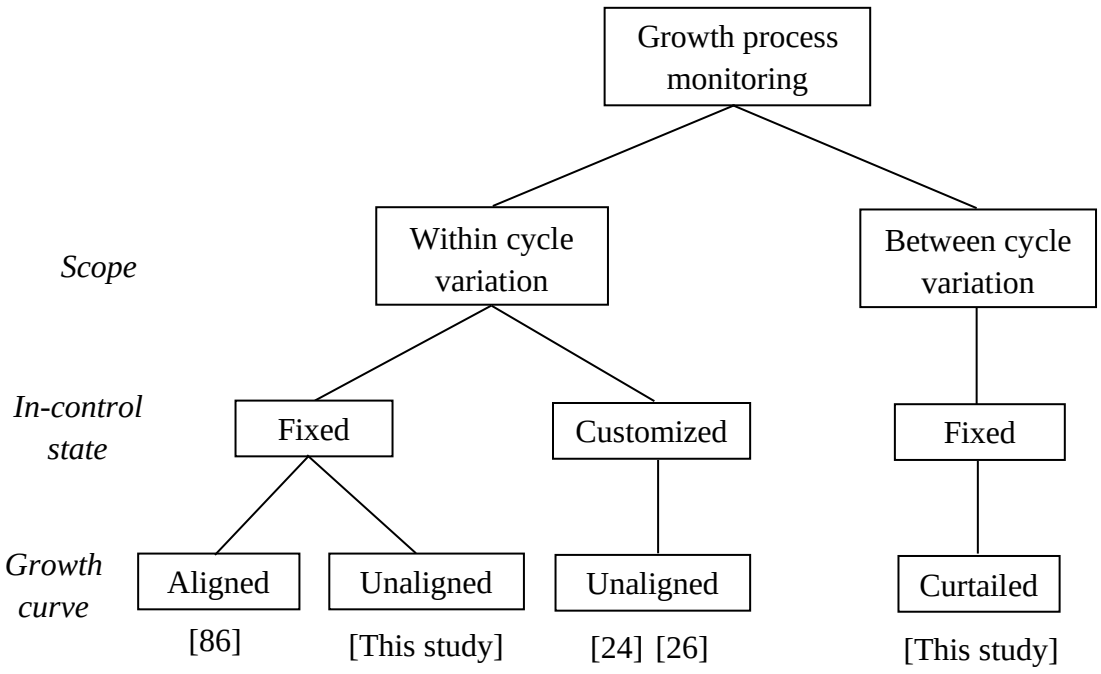
A growth curve can be considered as a special type of (process) profile, i.e. $y_{ij} = f(t_{ij})$, where y_{ij} is the observation made on the i_{th} cycle at the j_{th} instant. So, the between cycle variation of the process can be monitored following the methods of profile monitoring discussed in the previous section, provided the data are available in a suitable format.

However, in general, neither the end point of a growth cycle nor the sampling points within a cycle may remain fixed. For example, as noted previously, both the end point and the sampling instances vary from profile to profile in our growth data. It is obvious that the approach of profile monitoring is not directly applicable in such cases. So far no published study has come to our notice that deals with growth process monitoring based on between cycle variations.

It appears that all the studies on growth process monitoring that have been published so far aims at monitoring a growth process as it evolves. Shore et al. [24] proposed a methodology for monitoring human foetal growth and Zhu et al. [26] considered the case of monitoring a single crystal ingot growth process. Dai, Wang and Jin [110] also investigated the ingot growth process following a different methodology. In the first of the above three studies, the data are generated through periodic measurement, while the other two studies make use of continuous data obtained through on-line sensors.

Figure 5.1 shows the classification of the approaches proposed so far (including this study) for growth process monitoring. The second level of classification as shown in this figure is an important issue and this is discussed in detail in Section 5.4. Here we only note that in the customized approach, the in-control state is defined separately for each cycle based on an initial set of observations. The last level of classification is based on whether the growth curves are aligned or not before proceeding with further analysis. In [110], the variation in cycle length is adjusted following the method of dynamic time warping.

Figure 5.1. Classification of approaches for growth process monitoring



It may also be noted here that in all the studies mentioned above, the cycle time varies from cycle to cycle. However, it will be seen in the next chapter that there are processes where the growth rate is extremely high and consequently the process is designed to operate within a fixed time interval as a basic means of controlling the end state of the process. Such processes can be monitored only by monitoring the end state of the process.

5.2.3.2. Control charts for detection of growth failure

The detection of growth failure, when the characteristic(s) of interest can be monitored continuously using on-line sensors, is likely to be a simple problem. However, in many cases the process is monitored through periodic sampling/measurement, where the sampling or measurement intervals may vary over time. Furthermore, in many of these cases only a very few samples can be drawn within each growth cycle. In such cases, detection of growth failure at the earliest may pose a great challenge. To illustrate, consider the case of foetal growth monitoring, where the doctor monitors the conditions of the mother and the foetus periodically following a prescribed protocol. Under such a situation, the detection of foetal death is known to be an extremely difficult problem [24]. This is particularly so since the costs associated with the two types of errors in diagnosis are very high.

Unfortunately, like the other aspects of growth process control, the problem of detection of growth failure has also received almost no attention in the past. The only exception is perhaps the problem of early detection of children having short stature. It has already been noted in Chapter 1 that Grote et al [28] have proposed two sets of rules for two different age groups for detection of growth failure and subsequent referral for further diagnosis of the cause of suspected growth failure. Although these authors propose no control chart, the rules proposed for detection of growth failure are similar to the out-of-control rules of a Shewhart chart.

5.3. Control charts for monitoring between cycle variations of monomolecular growth cycles

In an industrial scenario, it is often the case that the number of measurements (p) made within a growth cycle is very limited. Assuming $p = 4$, it is obvious that the parameters of the monomolecular model cannot be estimated with any reasonable degree of accuracy. The desire to monitor the process by monitoring the parameters of the growth model is very reasonable and may also be very helpful in identifying the assignable causes (see Section 5.6). But such an approach cannot be implemented when the sample size is very small. So,

the following two alternative approaches are considered: (i) T^2 chart based on the original observations and (ii) *X*-chart of the Area Under Growth Curve (AUGC).

It is assumed that the four observations are made at fixed time points, *t* = 1.5, 2.0, 2.5 and 3.0 hour and that all the batches are active throughout this period.

5.3.1. T^2 Chart

The T_2 chart for the phase I study is constructed as follows. Let $Y_j' = (MP_{1j}, MP_{2j}, MP_{3j}, MP_{4j})$ be the vector of four observations for the j^{th} batch. Then, the control statistic T_2 for the i_{th} batch is given by

$$T_j^2 = (Y_j - \bar{Y})' S^{-1} (Y_j - \bar{Y}), \qquad (5.2)$$

where $\bar{Y}$ is the mean vector and *S* is the sample covariance matrix. The lower control limit (*LCL*) of the T_2 values is zero and the upper control limit (*UCL*) for the Phase I and Phase II study are respectively given by [111]

$$UCL_I = [(n-1)^2 / n]\, Beta_\alpha(p/2, (n-p-1)/2) \qquad (5.3)$$

$$UCL_{II} = [p(n+1)(n-1)/(n^2 - np)]\, F_\alpha(p, n-p) \qquad (5.4)$$

where *n* is the number of batches and *p* is the number of observations per batch. The $Beta_\alpha(a,b)$ and $F_\alpha(\nu_1, \nu_2)$ are (1- α)th quantile of the $Beta(a,b)$ and $F(\nu_1, \nu_2)$ distributions respectively.

For the purpose of illustration, the same thirty simulated growth curves considered in the previous chapter (Figure 4.8) are used to define the in-control state of the process. Although a much larger sample (in excess of 100) is recommended for constructing a T_2 chart [112], a sample size of (30×4) will suffice for the purpose of illustration. The mean vector $\bar{Y}$ and the sample covariance matrix S are obtained from this data set and then the values of T_j^2 for the thirty batches are computed using (5.2). The resulting T_2 chart with UCL given by (5.3) for $\alpha = 0.05$ is shown in Figure 5.2, where the last four points correspond to the out-of-control batches having abnormal growth parameters (see Table 5.1).

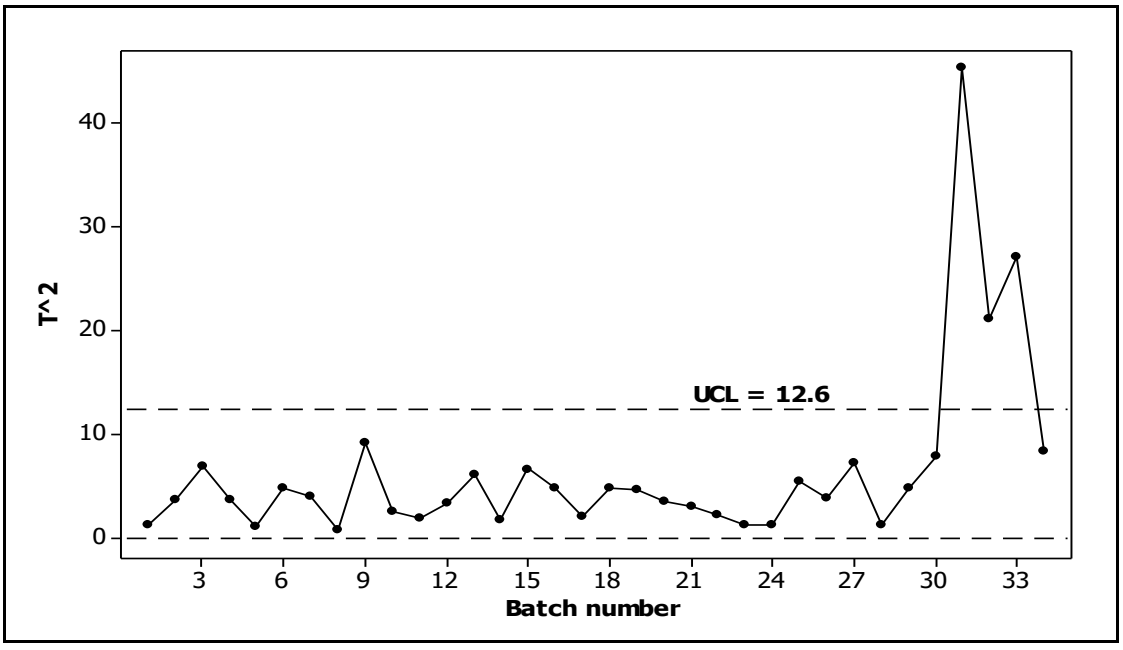
It is seen from Figure 5.2 that the chart provides no false alarm and is able to detect three out of the four out-of-control situations. The last case which the chart fails to detect is actually a complicated case. All the three parameters of this batch are significantly different from the mean of the in-control process (see Table 5.1). But, the chart has failed to detect it as an out-of-control batch since the growth behaviour of this batch, in the range of $1.75 \le t \le 3.25$, is similar to that of the in-control batches. Detection of such out-of-control situations

will require more samples covering a wider range of *t*. However, the failure to detect such complicated out-of-control situations may not be a serious problem. If the root causes for the first three simple cases are eliminated effectively, then the last case is likely to get solved automatically.

Table 5.1. Parameters of the in-control and out-of-control batches for the T^2 chart

Sample No.	*Status of the process*	$MP_{in}(C)$	$MP_{eq}(C)$	*k*
1-30	In-control	0±10	40.25±0.25	1.55±0.05
31	High MP_{eq}	0	41.25	1.55
32	High MP_{in}	30	40.25	1.55
33	Low *k*	0	40.25	1.30
34	Low MP_{in}, low MP_{eq}, high *k*	-20	39.5	1.75

Figure 5.2. T^2 Chart



Strictly speaking, the T^2 chart as constructed above is valid only when *Y* follows multivariate normal distribution. However, this assumption is not satisfied in the present case since the observations are generated using the growth model $MP_t = MP_{eq} - (MP_{eq} - MP_{in})\exp(-kt) + \varepsilon_t$, where the three parameters (MP_{eq}, MP_{in}, *k*) are assumed to be uniformly distributed and the error term ε is normally distributed. In particular, the marginal distributions of the first two of the four observations from each growth curve are found to be somewhat skewed. However, it

is known that the T^2 - tests are fairly robust against the departure from normality [112a]. So, usually it would be appropriate to use a T^2 chart for monitoring a monomolecular growth process. For example, in the above case, the chart is found to perform reasonably well.

5.3.2. Control chart of AUGC

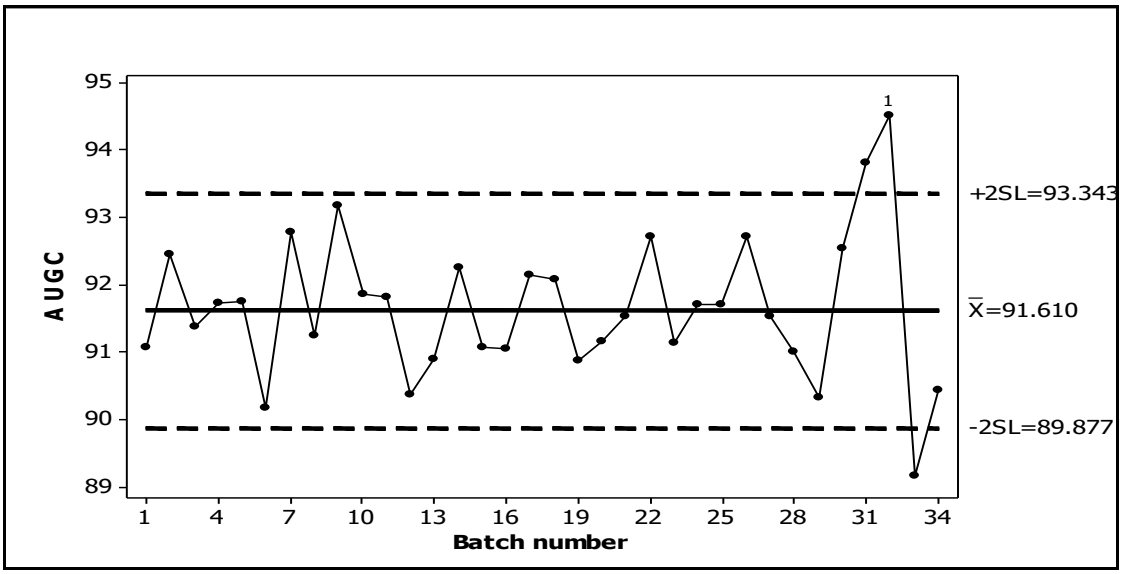
The area under a growth curve (*AUGC*) is frequently used as a summary measure of non-linear growth. The metric *AUGC* has proved useful for classification and comparison of growth curves in various fields, e.g. kinetics of tumour growth [113], intensity of a disease in the rice field [114] and the growth of algal species [115]. Here, our objective is to examine the feasibility of using *AUGC* for monitoring the hydrogenation process.

Let (MP_i, t_i) be the observations of a growth curve, where $t_{i+1} > t_i$ and $MP_{i+1} \geq MP_i$. Then the *AUGC* up to time t_n may be obtained as

$$AUGC(t_n) = \sum_{i=0}^{n} [MP_i(t_{i+1} - t_i) + (1/2)(t_{i+1} - t_i)(MP_{i+1} - MP_i)]. \qquad (5.5)$$

Figure 5.3 shows the *X*-chart of the *AUGC* values computed as above for the same thirty-four simulated batches, which are used previously for constructing the T^2 chart. It is seen that the in-control and the out-of-control signals are exactly the same as those obtained from the T^2 chart. However, the *AUGC* has two practical advantages over the T^2 – it is much simpler to compute and all the measurements need not be made at the same time points. But, as with the T^2, the time interval for computing the *AUGC* should remain the same for all the batches.

Figure 5.3. *X*-chart of area under the growth curve (AUGC) for thirty-four simulated batches. The control limits are ± 2-sigma limits.



It is thus seen that both the charts perform fairly well. Of course, it is necessary to examine the distributions of both the control statistics, namely T_2 and *AUGC*, and evaluate the performance of the above two charts in more detail before making any recommendation. Nevertheless, the results show that the charts pass the preliminary feasibility test.

5.4. Defining the in-control state for monitoring an active batch

In Chapter 1, Section 1.9, it has been noted that the two main challenges in the growth process control are in modeling and in defining the in-control state. The issues related to modeling are addressed in the previous chapter. The problem related to the specification of the in-control state is also discussed in the previous chapter, but the same is discussed here in more detail. A practical approach for specifying the in-control state is also proposed at the end of this section.

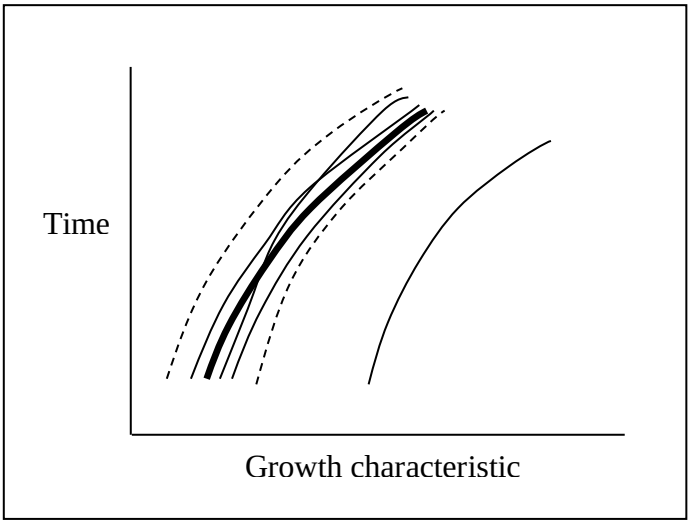
The first problem in defining the in-control state is the absence of a target or the reference growth path. This is primarily due to the difficulty of translating the optimal process performance to a particular growth path. An obvious way to deal with this problem is to consider that in each growth cycle the process will have its own mean trajectory, as defined by a particular set of values of the parameters of the underlying growth model. Thus, we may begin with an initial set of estimates (either based on past experience or an initial set of observations) and update the same as new data arrive. But a growth trajectory is also an indicator of process performance. So, it is obvious that we cannot adopt an unbounded approach as above. To illustrate, consider the four growth curves (solid normal lines) as in Figure 5.4. There is no difficulty in using the above approach for the three curves within the two dotted lines. But what about the isolated curve at the far right of this figure? Normally, one would like to consider it as an out-of-control situation. In order to have such a decision, it is then necessary to impose some bounds on the acceptable growth trajectories at the population level. The pair of dotted lines in Figure 5.4 is an example of such bounds. Shore et al. [24] used the above approach for monitoring fetal growth.

It is thus seen that the need for a set of population level bounds cannot be avoided even when it is assumed that each individual curve has its own mean trajectory. But, once the bounds become available, the possibility of adopting other approaches for monitoring the process also emerges. One such approach that has been used in this study, is discussed below.

It is assumed that if the population level bounds are not available from previous studies then the same can be established by evaluating the performance of the process for a given set

of sample growth curves. It is further assumed that a preferable zone can be identified within the acceptable zone of growth trajectories. So, in order to define the in-control state, we may begin with a set of curves which lie not only in the acceptable zone but the mean trajectory of the set lie in the preferable zone. This initial set of curves can then be modified either by contracting or expanding it in an iterative manner to arrive at the optimal choice. Assume that the mean trajectory for the optimal in-control set of curves be given by the bold line as shown in Figure 5.4. So, an appropriate monitoring scheme may be developed for monitoring the growth curves based on the deviations of the observations from this population level mean trajectory.

Figure 5.4. Illustration of growth process monitoring



It is obvious from the above discussion that in the population level approach, some amount of between individual variation is also included in the common cause variation. But how do we find the optimal set of in-control curves? The methodology for finding such a set is discussed a little later in Section 5.4.2. Here we only note the three criteria that may be used for arriving at the optimal in-control set: (i) The mean trajectory of the optimal set must lie in the acceptable zone, (ii) The amount of between curve variation that is included in the common cause variation does not exceed a specified limit and most importantly (iii) It is verified at the practical level that the assignable causes have been found for almost all the out-of-control signals obtained from the control chart.

5.4.1. Individual vs. population level approach: practical considerations

The motivation for adopting the individual level approach may come from the following cause and effect model, where it is assumed that there are three types of factors in the system

– The set of fixed factors *FX*, whose level remain fixed throughout a growth cycle; the set of control factors *X*, which fluctuate around their local mean (or mean trajectory) and the set of noise factors *N*. The mean trajectory of a growth cycle is dependent on the levels of *FX* and *X*, but their levels are determined by a mechanism which is either unknown or beyond our control. Thus, the common cause variation of the process is due to the fluctuation of *X* around its local mean and the fluctuations of *N*. An additional assumption here is that the mean level of an individual can be estimated either based on past experience or the data obtained from the beginning of the present growth cycle and updating the same as the growth cycle evolves. Thus, the monitoring scheme may be devised based on the common cause variation around the individual-level mean trajectory. For example, if a student who usually scores around 50 but is found to have scored 100 on a certain occasion, then the score becomes suspicious. However, if a student who usually scores around 90, there may not be enough reason to suspect a score of 100 obtained by this student. Also, if a score is found to be below the population level bound, say 30, then serious intervention, including termination of the growth process in the form of failing the student, may be called for.

The individual-level approach seems to be very reasonable in cases such as the above. But there are practical difficulties in its implementation. In order to appreciate the difficulties involved, let us first recognize two different types of situations. One in which the individuals themselves behave like a system, as in the medical field (excluding public health), and the other in which the individuals are treated as just one of the elements of the system, as in case of the education process. Of course, the human body is also composed of many subsystems, but the search for an assignable cause is usually limited to the subsystem of interest. However, there may be difficulty in finding the assignable cause even within a subsystem. Note that the common cause variation, in this case, includes only the effect of the fluctuation of the *X* around its local mean. It is thus necessary to define a medical condition also in terms of the deviation of *X* from the local mean. But, if in the existing practice, the medical condition is defined by a fixed zone X_0 around the population level mean of *X*, then the individual level approach will lead to a large number of false alarms. On the other hand, since the common cause variation in case of a population level approach includes the variation of *X* from its global mean, the use of X_0 for diagnosing a medical condition will be appropriate.

Let us now consider the education process. Suppose the assignable cause of a student scoring unusually high marks has been identified as the leakage of the question paper. To take another example, suppose the assignable cause for a student scoring unusually low marks is found to be his or her illness on the day of examination. Note that the assignable

causes in both the cases do not belong to the core process of learning. They belong to peripheral systems related to the main system over which the process owner may have very little or no control. On the other hand, mass illness may be identified as the assignable cause in a population-level approach and the process owner can do many things to prevent its recurrence. Most importantly, the population level approach provides opportunities to investigate the deficiencies related to the core elements of the learning process like the class timings, coverage of the syllabus, difficulty level of the question paper and many others.

It is also to be noted that whatever be the approach, there may be important process variables over which it may not be possible to exercise any control within a growth cycle, but control can be exercised at the population level. The initial melting point of the input oil and the catalyst condition are examples of such process variables for the hydrogenation process. The implication of this restriction is that the on-line search for assignable causes should exclude these factors. Thus, the search for assignable causes in the individual level approach may be expected to be comparatively easier. But, this will be so only when the interaction effect between a fixed factor with other factors that vary within a growth cycle is not very high. In the presence of such strong interaction effects, the control chart is likely to give many false alarms.

Finally, let us emphasize the need to be conservative in defining the in-control state with the help of an example. Shore et al [24], while outlining their proposed monitoring scheme, suggest that if an out-of-control signal cannot be attributed to a medical condition then "consider deleting this observation from further estimation due to estimation error". This, in our opinion, is an aggressive approach. The point should be deleted only when the source of the measurement error, if any, has been fixed. If not, then a conservative approach will be to redefine the in-control state to account for this lack of process knowledge to control measurement error.

5.4.2. Proposed methodology

It is assumed that the available growth data is in the form (y_{ij}, t_{ij}), $i = 1, 2, ..., n_i$, $j = 1, 2, ..., k$. In general, there is no restriction either on t or n, but it is preferable to have data on $k > 50$ curves and $n_i \geq 3$, where n_i is the number of observations in each curve. Since our objective is to use a population level approach for monitoring the process, the common cause variation is given by the deviations of each of the observations from the respective population mean. Further, it is desired to limit the contribution of the between cycle variation towards the common cause variation. In order to evaluate the contribution of the between cycle variation,

it is proposed that the prediction error variances may be compared under two situations - maintaining the cycle identity of the observations and without maintaining the same. These two error variances may be estimated by developing the following two models.

Model 1:

$$
\begin{aligned}
&y_{ij} = f(t_{ij};\beta) + r_{ij},\\
&r_{ij} = f_1(r_{i-1(j)}, r_{i-2(j)},..., \Delta t_{ij};\delta) + \varepsilon_{ij},\ \ \Delta t_{ij} = t_{ij} - t_{i-1(j)},\\
&\varepsilon_{ij} \sim N(0, \sigma_i),\\
&\sigma_i = g(t_i, \beta;\theta),
\end{aligned}
$$

where β, δ and θ are the parameter vectors of the mean, correlation and variance function respectively.

Model 2:

$$
\begin{aligned}
&y_{ij} = f(t_{ij};\beta) + \varepsilon_{ij},\\
&\varepsilon_{ij} \sim N(0, \sigma_i),\\
&\sigma_i = g(t_i, \beta;\theta),
\end{aligned}
$$

The errors ε_{ij} are assumed to be independent in both the models. The second model is the same as Model 1 except that the spatial correlation term in the second line is omitted and r_{ij} is replaced by ε_{ij}. Ideally, the parameters of Model 2 are to be estimated by selecting only one observation randomly from each curve. In case sufficient data are not available, a few observations from each curve, which are far apart from each other, may also be selected for developing the model. As such, there is no need to develop Model 2, since both the error variances can be estimated from Model 1. However, it is our belief that the estimate of error variance obtained from Model 2 will be more robust, provided the sampling scheme is chosen judiciously. In case the sampling scheme does not provide a satisfactory model then the estimates of f and g obtained from Model 1 may be considered as the desired estimates for Model 2 as well.

Structurally, Model 1 is a marginal model for longitudinal data [116]. Note that we have already used the above models for modeling the hydrogenation data; the models (4.11) and (4.16) in Chapter 4 are special cases of Model 1 and Model 2 respectively. Of course, one can also make use of other models as noted in Chapter 4. But the above models are specified primarily from the considerations of flexibility and simplicity.

Let the average error variance, i.e. the average of $(y_{ij} - \hat{y}_{ij})^2$ for the two models be denoted by v_1 and v_2 respectively. In order to limit the contribution of between curve variation in the estimate of common cause variation, we shall impose the restriction $F = v_2/v_1 \leq F_0$ (the second criteria for arriving at the optimal set of in-control curves). Now, it is reasonable to assume that the error variances ν_2 and ν_1 will be estimated based on at least twenty and thirty

degrees of freedom respectively. Thus for F to be significant at 95% level of significant it should be greater than 2.2.Consequently, as an approximate thumb rule, we may take $F_0 = 2$. If required, its value can be relaxed further, if the other two criteria of optimality are satisfied. The proposed 20-step methodology can now be stated as follows.

Step1. Select a sample of growth curves judiciously to capture the variation in the growth behavior (normal/abnormal) of the existing process.

Step2. Conduct a preliminary analysis of the growth data to understand the basic features of the process. The simple tools that may be useful for this purpose are the scatter plot of the growth curves, histograms of t given y, plot of the growth curves over time (if the growth curves are not too long) and a tabular summary as in Table 4.2 (Chapter 4).

Step3. Identify the preferable and permissible zones of acceptable growth trajectories based on an (subjective) assessment of the production cost and the quality of the process output for the sample curves. If no acceptable zone is found or the zone is found to be too narrow, then abandon the idea of monitoring the present process and initiate special study on process stabilization and improvement as noted in ***Step19***.

Step4. Select a subset of curves from the k sample curves or modify the previous selection (to satisfy the requirement in ***Step7***). This set of curves should be close to each other and similar in shape. The mean trajectory of these curves should lie in the preferable zone. If necessary, perform cluster analysis for this purpose. In case of a well-controlled process, this subset may consist of all the sample curves.

Step5. Develop Model 1 using the growth curves selected in ***Step4***.

Step6. Develop Model 2.

Step7. If $F \leq 2$ then continue, otherwise go to ***Step4***.

Step8. Construct a suitable control chart using the residuals of the model obtained in ***Step6***.

Step9. Examine the status of all the k curves using the control chart. Let the set of in-control group obtained at this stage be denoted by S.

*******End of Initial Iteration*******

Step10. If all the curves are found to be in control then go to ***Step16***, otherwise continue.

Step11. Search assignable causes for the out-of-control signals. If assignable causes are found for all then go to ***Step16***, otherwise continue.

*******Further Iterations in the Expansion Mode*******

Step12. Select one or more curves from just outside the present in-control zone for which no assignable causes have been found. Consider these as in-control curves and add them to *S*. If no such curves are available, then go to ***Step19***.

Step13. Redevelop Model 1 and Model 2 using the curves selected in ***Step12*** (enlarged *S*).

Step14. If $F > 2$ or the mean trajectory is outside the acceptable zone, then go to ***Step19***.

Step15. Construct the control chart. Place reasonable control limits so that all the curves used for developing the process model are in control. If no such control limits are found, then go to ***Step19***. Otherwise update *S* and go to ***Step12***.

*******Further Iterations in the Contraction Mode*******

Step16. Select one or more curves from *S*, which are at the two extremes of the present in-control zone. Consider these as out-of-control curves and examine whether any assignable causes can be established for them.

Step17. If assignable causes are not found for any of these curves, then go to ***Step19***. Otherwise remove the curves for which assignable causes are found and redevelop Model 2.

Step18. Construct the control chart. Place reasonable control limits to obtain out-of-control signals for the batches removed in ***Step17***. If no such control limits are found, then go to ***Step19***. Otherwise update *S* and go to ***Step16***.

*******Exit to Phase II Monitoring or Process Stabilization and Improvement*******

Step19. If all the sample curves are in control or the assignable causes have been found for almost all the out-of-control signals, then initiate Phase II monitoring.

Step20. If assignable causes are not found for many of the out-of-control curves but these curves belong to a particular time period or if there is a systematic pattern in the time of their occurrence then also initiate phase II monitoring but support it with a special data collection format for identifying the special causes. Otherwise initiate special studies on process stabilization and improvement.

The above twenty step methodology is somewhat long but not complicated. It may be seen from Figure 5.5 that the underlying structure is in fact very simple.

5.5. Control charts for monitoring the process of hydrogenation

The following three control charts are developed for monitoring the process of hydrogenation:

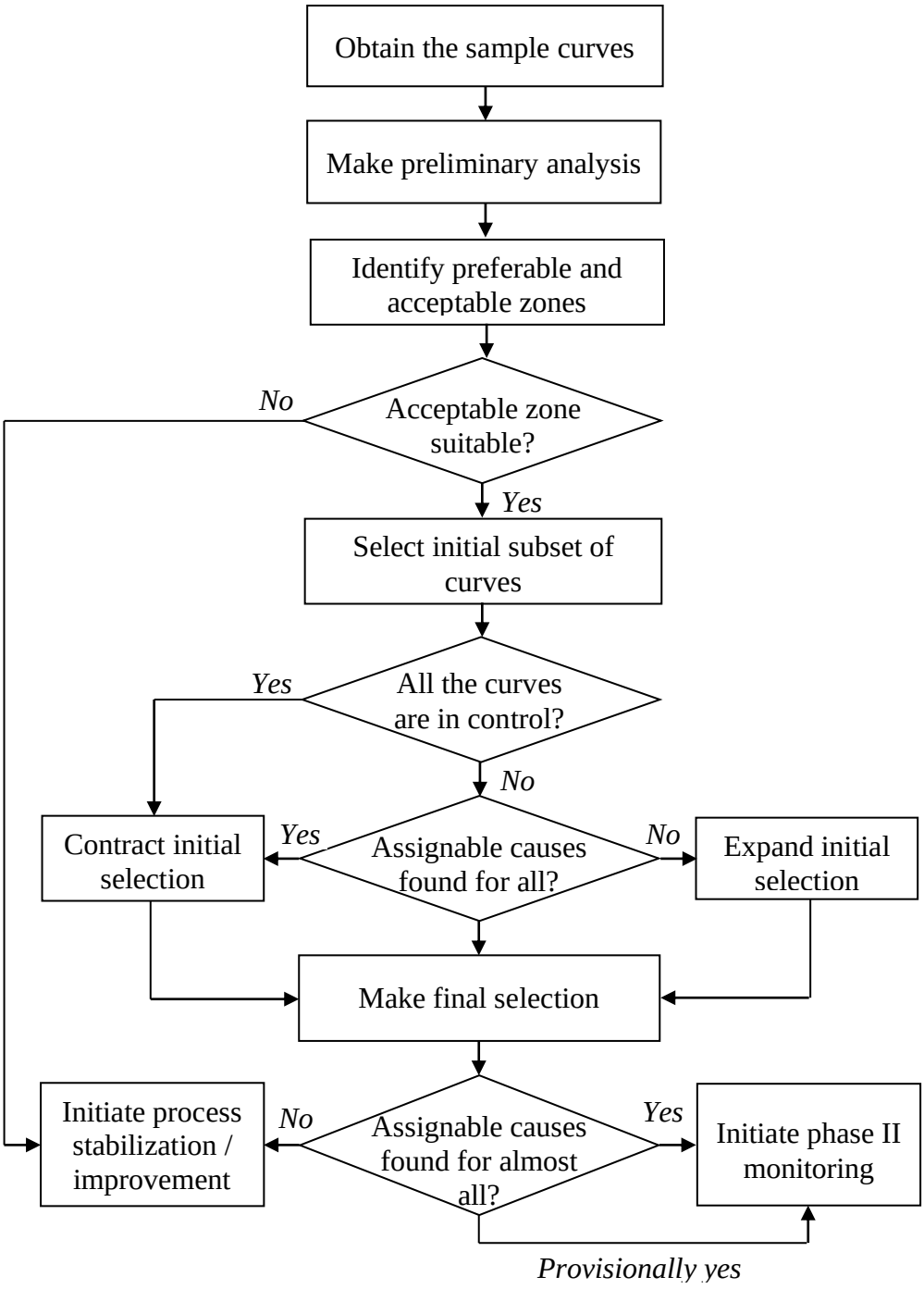
(i) Individual chart for monitoring an active batch. This chart is based on a linear approximation of the monomolecular model.

(ii) Individual chart for detection of a dead batch. Like the previous one, this chart is also based on the same linear approximation of the monomolecular model.

(iii) The second individual chart for detection of a dead batch, but unlike the previous one, this chart is based on the original monomolecular model.

One of the two charts for detection of dead batches as mentioned above is recommended for implementation based on a comparison of power of the two.

Figure 5.5. Flowchart of the twenty step methodology



5.5.1. Control chart for monitoring an active batch

The control chart for detecting shifts in the growth parameters is constructed using the static growth model (4.17), which is given by

$$\hat{Z}_t = 42.35t,\ Z_t = MP_t(t+0.2),\ [\hat{\sigma} = 1.44], \qquad (5.7)$$

where $\hat{\sigma}$ is the standard error of $\hat{Z}_t$. The control statistic chosen for monitoring the process is

$$Q1 = (Z_t - \hat{Z}_t)/\hat{\sigma}. \qquad (5.8)$$

Thus, the control chart is a simple *X*-chart of the standardized residuals of (5.7). The above control statistic is intended to act as an omnibus test for the shifts in initial melting point (MP_{in}), growth potential (MP_{eq}) and rate constant (k).

The control limits are placed at ± 3 (usual Shewhart limits), and the resulting control chart is shown in Figure 5.6, where the residuals corresponding to the same batch are connected by straight lines. Also, for the sake of clarity, the status of the four batches belonging to group C and the nine apparently dead batches are not shown in this chart. All these batches, except batch #10 give very strong out-of-control signals. The status of batch number 10, i.e. whether it can be considered as dead or not is examined later.

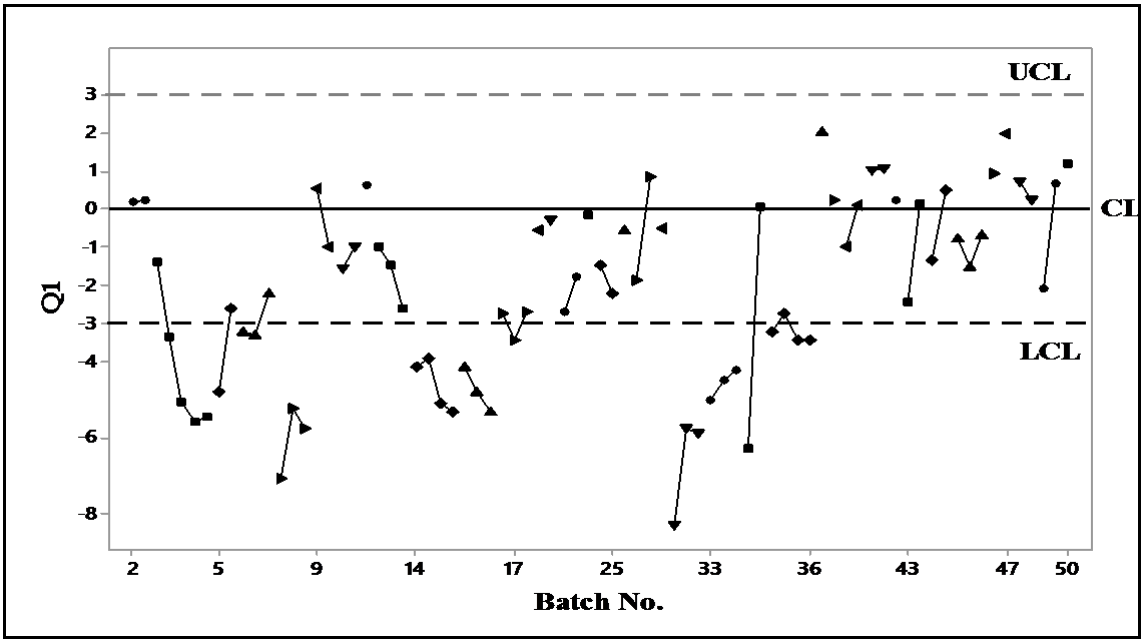
Let us now examine whether the process model (5.7) based on which the control chart is constructed can be considered as representing the in-control state of the process. We shall make use of the three criteria of optimality as noted in Section 5.4 for this purpose. The model obviously satisfies the first two criteria since it represents group A, which is the best performing group (see Table 4.5, Chapter 4) and the spatial correlation among the curves within this group is insignificant (see Table 4.6, Chapter 4). However, it is evident from the chart that the process is very unstable. This is to be expected since the in-control process is defined by group A and there are too many batches in group B compared to group A. Even though a few of the batches belonging to group B are now seen to be in control, there are still too many out-of-control signals. So, the question arises - does it satisfy the third requirement? It may be noted from the chart (Figure 5.6) that the process shows a distinct sign of improvement from batch number 37 onwards. Thus, essentially there is only one major out-of-control signal. It is possible that a majority of the out-of-control signals prior to batch #37 are due to the same assignable cause. In fact, subsequent investigation following such a characterization of the problem makes it possible to identify the assignable cause for the process being out of control till batch #37. However, the details of this study and the results obtained after correcting the process are not discussed here.

However, it may be noted here that once the process stability is achieved, it may be useful to construct out-of-control rules based on the consecutive values of $Q1$ for the same batch.

5.5.2. Control charts for detection of a dead batch

A naive approach for detection of a dead batch will be to consider a batch as dead if two consecutive measurements are found to be the same. However, two consecutive measurements can be the same just on account of measurement error (since the least count of measurement here is 0.1°*C*). Thus, the naive approach can be improved further by requiring that the sampling interval should be a minimum of Δt minutes for declaring a batch as dead. But then a batch may be dead even when the observed growth is greater than zero, since the growth failure can occur anywhere within an interval. It is thus necessary to device a procedure to detect such situations. An obvious way to achieve this is to predict the amount of growth that should be expected within a given time interval along with the band of prediction error (including the measurement error) and then compare the same with the observed growth. A large negative deviation of the observed growth from that predicted may be considered as an indication of growth failure within the interval concerned. The same approach has been adopted here.

Figure 5.6. *X*-chart of *Q* for monitoring the active batches



Further, since growth-failure is a sharp local feature that is independent of the location of the growth curve, it is necessary to make use of a dynamic model for prediction of growth.

However, since the growth rate of an active batch may reduce significantly for various reasons even when it remains active, it is necessary to make use of appropriate data for developing such models. This aspect of the modelling has already been discussed in detail in the previous chapter (Section 4.3.7.1) and the models have been developed accordingly. These dynamic models are now used in this section to develop the control charts.

5.5.2.1. The first chart

The first control chart is constructed based on the model (4.19), which is given by

$$Z_{t+\Delta t} = Z_t + d(\Delta t),\ \Delta t \le 1.5,\quad d \bigcap N(42.67, 1.21). \tag{5.9}$$

As before, a simple *X*-chart of d is proposed for detection of a dead batch. Thus, the control statistic d_{ij} is given by

$$d_{ij} = (z_{ij} - z_{i-1,j})/(t_{ij} - t_{i-1,j}),\ i = 2,3,...,n_j, \tag{5.10}$$

where z_{ij} are computed from the sample observations (MP_{ij}, t_{ij}). The control statistic d has an interesting property. In general, $d \ge MP_t$, the melting point at the beginning of an interval. The equality holds when the growth within an interval is zero. This has two practical implications. First is that the lower control limit (LCL) of *d* should be 40, since if LCL < 40, then the chart will not be able to detect growth failure in the range of *MP* > LCL and there is not much to be gained by detecting a dead batch beyond the cooling limit of 40^0C. The other implication is that the chart will always signal a growth failure for any value of Δt if the observed growth within an interval is zero. It is thus important to have a lower bound on Δt to avoid false alarm due to measurement error. This lower bound is obtained as follows. It is assumed that a sampling interval Δt that satisfies $Prob(MP_{t+\Delta t} - MP_t \ge 0.1) = 0.99$ should be satisfactory. Based on this criterion, it is found from (5.9) that the sampling interval should be about one hour or more in the range of $MP \ge 38.5^o$C.

It is found from the distribution of *d* given in (5.9) that LCL = 40 is a 2.2 sigma limit. Using the same factor for the upper control limit (UCL), we have UCL = 45.34. Figure 5.7 shows the resulting control chart along with a plot of the observed values of *d* for all the intervals (both active and dead). For the sake of simplicity, the restriction of $\Delta t \le 1.5$ hours (see Equation 5.9) is ignored for constructing this chart. However, it is checked that all the out-of-control signals satisfy the restriction.

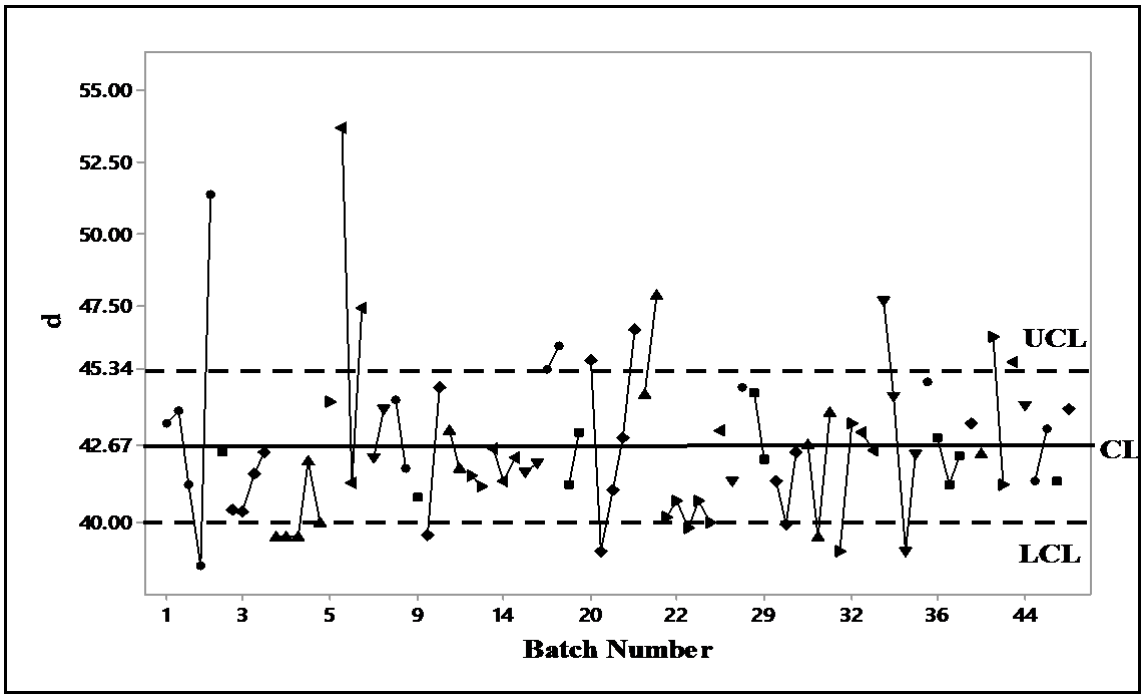
Since a point falling below the LCL is an indication of a dead batch, it is seen that the chart identifies nine batches as dead. A scrutiny of the growth data corresponding to these nine dead batches reveals the following:

(i) For all the nine intervals indicating growth failure, $\Delta MP \leq 0.1^0$C. However, it will be seen later that the chart can detect growth failure even when $\Delta MP > 0.1^0$C, but such cases are not available in the sample.

(ii) Out of the nine dead batches, two batches (# 10 and 30) do not satisfy the restriction of $\Delta t \geq 1$ hour. Thus, the benefit of doubt should be given to these two batches.

(iii) There are two extremely slow growing batches (# 1 and 22) which are not dead but can be considered as 'practically dead'. It is advisable to add fresh catalyst to such batches to increase the growth rate.

Figure 5.7. *X*-chart of *d*



However, the main problem with the chart is that there are too many out-of-control signals above the UCL. Out of the eight such signals, only four (batch # 6, 16, 21 and 42) can be attributed to a different type of oil that has been used for these batches. The cause of the remaining four remains unknown and hence for all practical purposes these signals should be considered as false alarms.

Considering the above, it is decided to develop an alternative control chart for the same purpose. The details of this chart are given in the next section.

5.5.2.2. The second chart

The second chart is based on the dynamic model of the process given by (4.20) and (4.21), for which the growth within an interval is estimated from the instrumental variable model given by (4.29) - (4.32). This model is restated below in a compact form,

$$E(G^I) = 0.0805 + 0.9089\,\hat{G}_1,$$

$$\hat{G}_1 = -0.0482 + 1.2819I$$

$$I = \sum_{i=1}^{n2} g_i,\ g_i = 0.852\,(\Delta t_i) - 0.4563\,(MPS)(\Delta t_i),\ MPS = (MP - 38.488)/1.082, \qquad (5.11)$$

$$\Delta t = \sum_{i=1}^{n2} \Delta t_i,\ \Delta t_{n2} = (\Delta t - 0.5(n2-1)) < 0.5,$$

$$SN(G^I) = 113.091 - 2.85(MP) + 2.57(t - t_0) = 10\log[E(G^I)^2 / V(G^I)].$$

The control statistic chosen for the chart is given by

$$Q2 = (g - E(G^I))/\sigma(G^I), \qquad (5.12)$$

where g is the observed value of growth, $E(G_I)$ is given by the mean function in the first line of (5.11) and $\sigma(G_I)$ is obtained by solving the SN function in the last line of (5.11). Considering the estimates of the mean and variance of growth as known, the distribution of $Q2$ for an in-control process (i.e. active process) is given by $N(0, 1)$. The NPP of $Q2$ for the 51 active intervals, which were used for developing the model (5.11), shows that the above assumption is very reasonable (Figure 5.8).

Figure 5.8. NPP of standardized growth ($Q2$)

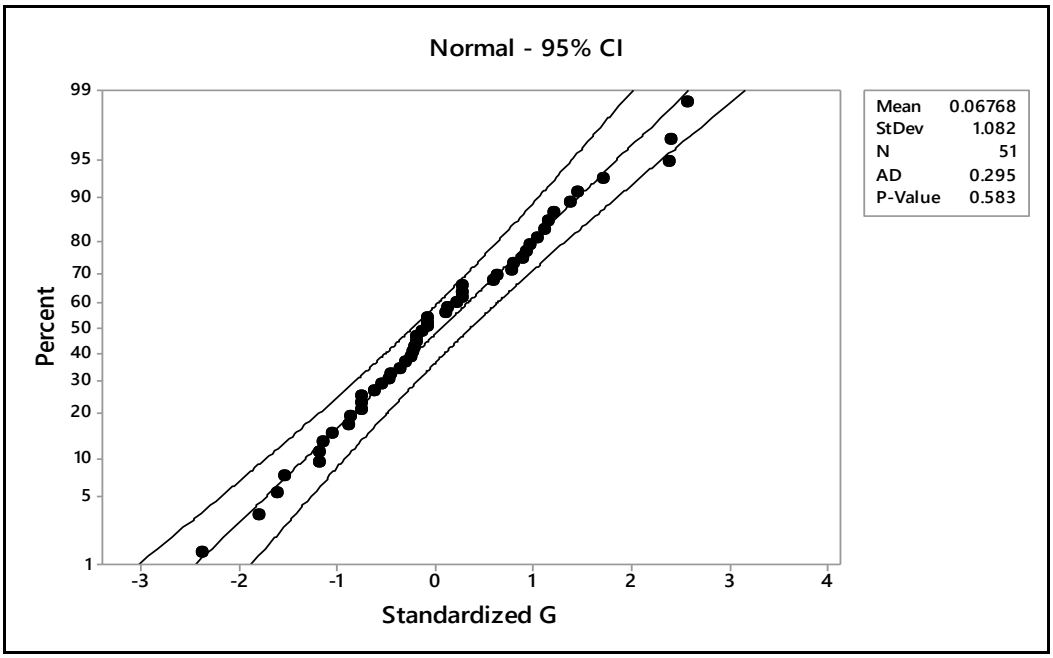
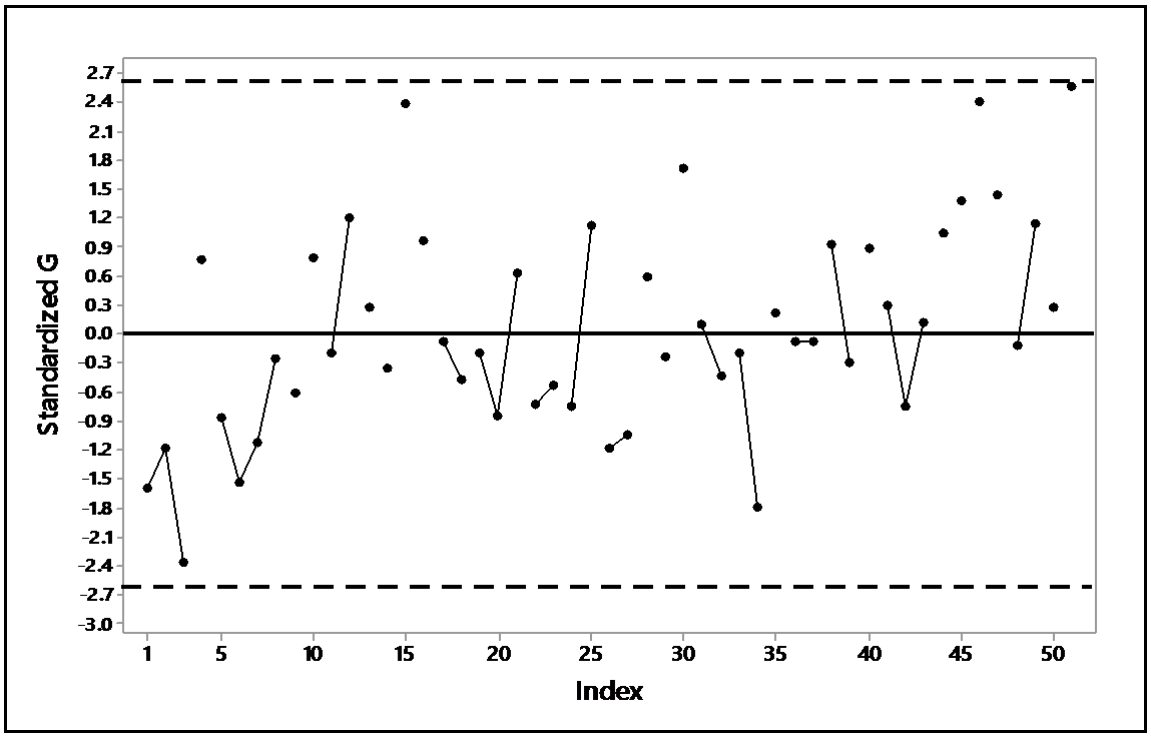


Figure 5.9 shows the individual control chart for $G|(MP, \Delta t)$ with control limits at $\pm 2.6\sigma$. It is seen that the process can be broadly considered as in-control, except that the batch # 1 is a doubtful case, since all the three points representing this batch are far below the central line. However, it is explained below that the chart with control limits at ± 2.6 will be inefficient. Accordingly, the control limits are modified slightly to obtain the final control chart.

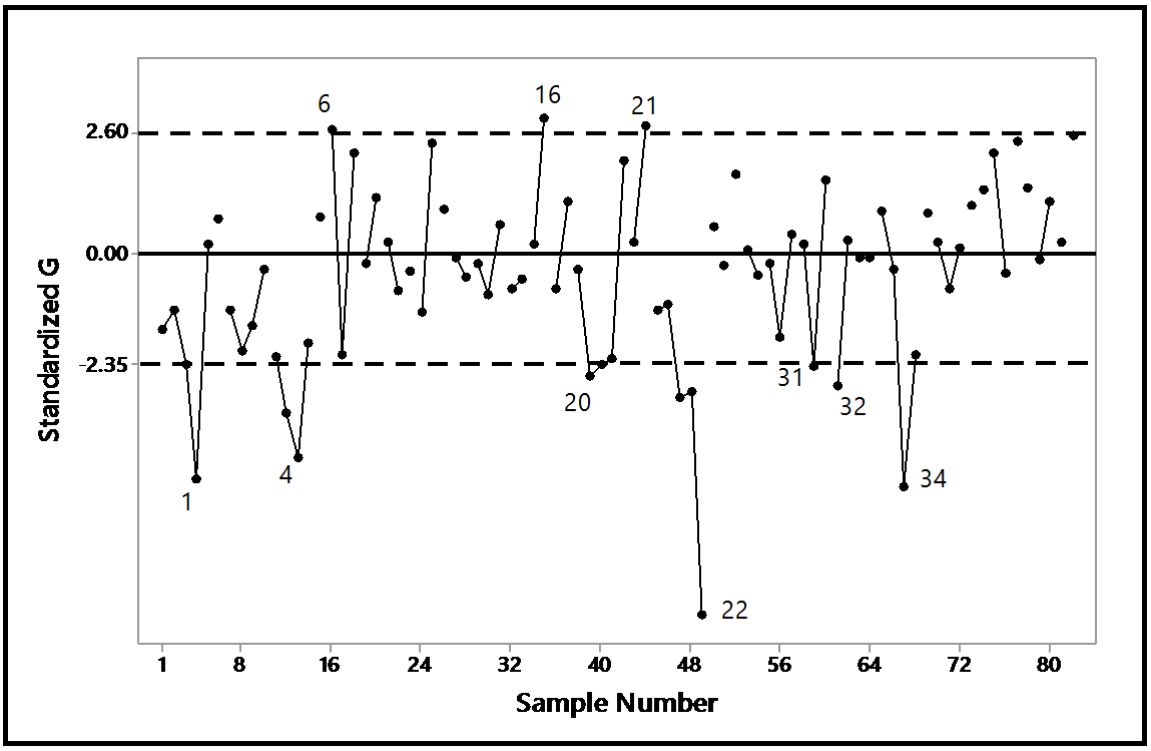
Figure 5.10 shows the control chart for all the intervals (both active and presumed to be dead) with lower limit at -2.35. It is seen that there are seven instances of growth failure. However, with LCL at -2.6, the batch # 31 will be identified as active throughout. A close scrutiny of the growth curve for this batch suggests that it might not have reached the target without the addition of fresh catalyst. Consequently, it is decided to have the lower control limit at -2.36, which corresponds to an expected false alarm rate of less than 0.1% on a sample basis. However, the UCL is kept unchanged at 2.6. It is noted that the two batches which are given benefit of doubt in the first chart are now found to be active throughout. Most importantly, unlike the previous chart, there are now only three out-of-control signals above the UCL and all the three signals correspond to the batches belonging to the group C.

Figure 5.9. *X* - chart of growth within an interval (active intervals only)



A comparison of the results obtained from the control chart in Figure 5.10 and the existing system shows that fresh catalyst has been added on as many as twelve occasions. Such additions have been made either presuming that the batch is dead or to increase its growth rate. On the other hand, there are two extremely slow growing batches to which no fresh catalyst has been added. Thus, as per the proposed control chart, fresh catalyst has been added unnecessarily on seven occasions. The values of Q2 for these seven suspected intervals are (-0.77, -1.24, -0.47, -0.21, -0.74, -1.79, -0.08). Since all the values are well above the LCL, it is safe to conclude that the addition of fresh catalyst should have been avoided in these seven cases. It is also found that in one of the five cases of genuine dead batches, the addition of fresh catalyst is delayed by as much as 2.33 hours. Thus, the implementation of the proposed chart may be expected to improve the performance of the process significantly.

Figure 5.10. *X* - chart of growth within an interval (all intervals)



5.5.2.3. Comparison of power

The following two performance measures are used for comparing the power of the two charts proposed as above for detection of a dead batch:

(i) Average delay $(\overline{D})$ in detection based on the first sample with $0.75 \le \Delta t \le 2.25$ hour, and

(ii) Maximum melting point $(MP_{\max})$ for first sample detection with Δt as above.

Table 5.2 summarizes the performance of the two charts evaluated through simulation. It is seen that the first chart performs better than the second only when the observed growth is zero, i.e. when an active batch fails immediately after sampling. In all other situations, the second chart performs better than the first one with respect to first sample detection. Moreover, it is seen that the first sample detection is possible in the entire working range with the second chart. But the first chart performs very poorly in this respect.

However, a closer examination of the results of simulation reveals that there is a region where the first chart performs better than the second. This region is $MP_d \geq 39.5^0 C$ and $t_d \leq 2.5$ hours, i.e. if a fast growing batch becomes dead early. This fact is not captured in Table 5.2. However, such situations are expected to occur only rarely. Thus, considering the overall performance of the two charts, the second chart is considered to be superior to the first one. Hence, the second chart is used for developing the integrated control scheme discussed in the next chapter.

Table 5.2. Comparison of power of the two charts for detection of a dead batch

Growth of MP (^{0}C)	*First chart*		*Second chart*	
	$\overline{D}$	$MP_{\max}$	$\overline{D}$	$MP_{\max}$
0.0	0.75	40	0.80	40
0.1	0.87	39.7	0.83*	40
0.2	0.82	39.5	0.80*	40
0.3	0.86	39.0	0.80*	40

*Average delay corresponds to the MP_{max} of the first chart. Average delay for MP_{max} of 40 are 0.86, 0.92 and 0.97 for growth level of 0.1, 0.2 and 0.3 resp.

5.6. Identification of assignable causes using the monomolecular model

It is well known and is also evident from the previous discussion that obtaining out-of-control signals from a control chart is only the first step towards developing a rational approach for shift control. It is important to have effective process diagnostics for identifying the assignable causes for the out-of-control signals. In the absence of such diagnostics, effective implementation of SPC on a continuing basis is difficult to achieve. So, an attempt is made here to indicate the possibility of developing such a diagnostic based on the molecular model for controlling a hydrogenation process. The main idea is that the parameters of the monomolecular model are expected to be strongly correlated with the important process variables, as already noted at the beginning of Chapter 4. Of course, it will be very useful to have a formal model showing the relationship among the model parameters

and the process variables. But we do not have adequate data to develop such a model. So, the discussion here is essentially qualitative in nature. But it does show that even the estimates of the parameters of the molecular model can be helpful in identifying the assignable cause for a given out-of-control situation.

So far as the rate constant k is concerned, it is known that the greater the reaction temperature, pressure, agitation speed and catalyst activity, the greater will be the value of k. The estimate of the rate constant k, given in Table 4.1, also shows that its value increases with the increase in reaction temperature. However, the effect of the process variables on MP_{eq} is not known. So, a qualitative analysis is made here to analyze this aspect of the problem purely from an engineering point of view.

It is well known that the relationship between IV and MP is highly influenced by hydrogenation process conditions [117]. For example, Rangxuan [118] reports that for a given iodine value, the melting point of partially hydrogenated rapeseed oil decreases as reaction temperature increases. In other words, the slope of the relationship $IV = \alpha - \beta(MP)$ increases with the increase in hydrogenation temperature. Now, it is easy to see from (4.4) that $MP_{eq} = \alpha/\beta$. Using $\alpha = IV_{in} + \beta(MP_{in})$, we have $MP_{eq} = \alpha/\beta = IV_{in}/\beta + MP_{in}$. This shows that for given oil, i.e. with IV_{in} and MP_{in} remaining fixed, the value of MP_{eq} should decrease with increasing hydrogenation temperature. However, if the growth dynamics of IV is given by (4.7) then we have $MP_{eq} = (\alpha - IV_{eq})/\beta = (IV_{in} - IV_{eq})/\beta + MP_{in}$. Thus, if IV_{eq} increases with increasing temperature, as is to be expected due to enhanced catalyst selectivity, then MP_{eq} should decrease with increasing hydrogenation temperature. This is what has also been noted in this study (Table 4.1).

Thus, if the estimates of k and MP_{eq} for an out-of-control batch are found to be significantly lower and higher respectively than that of the in-control group then a dip in reaction temperature may be responsible for the out-of-control situation (provided the input oil is the same and there is no within batch instability). On the other hand, if the estimates of both k and MP_{eq} are found to be significantly lower, then a fall in either pressure or agitation speed or both may be suspected. This is because the amount of saturate formed is expected to decrease with decreasing pressure and agitation speed. However, more studies are necessary to establish the effect of pressure and agitation speed on MP_{eq} before we can draw such a conclusion.

Finally, it may be noted that if sufficient data are available then the three parameters of the model can be estimated at the end of a cycle and hence the search for assignable causes can

be initiated at the end of a growth cycle. Otherwise, it is necessary to pool data from several cycles and the methods used in Chapter 4 may be used for estimating the parameters at the group level.

5.7. Conclusion

The review of literature reveals that growth process monitoring is yet to receive adequate attention of the researchers. However, it is shown here that an active growth process may be monitored either based on an individual-level approach or a population-level approach. The pros and cons of these two approaches are discussed in detail and a methodology is proposed for growth process monitoring based on the population-level approach.

It is also argued that both predictability and removability are important criteria for defining the in-control state of a process. The proposed methodology makes an attempt to achieve a balance between these two opposing requirements.

It is shown that the process of hydrogenation can be monitored effectively either at the batch level or during the evolution of the growth process within a batch. However, the proposed control chart for monitoring the hydrogenation process is meant for monitoring the process as it evolves. Two simple X-charts are also constructed for detection of a dead batch. It is shown that the control chart that is constructed based on the original non-linear model of the process is more efficient than the other chart based on a linear approximation of the non-linear model.

Presently, detection of a dead batch during hydrogenation is heavily dependent on the subjective judgement of the process engineer. The use of the proposed control chart is expected to greatly reduce the subjectivity involved in decision making in this regard.

Chapter 6 Integrated shift and drift control*

It is said that 'control is ubiquitous' [119]. For example, our body is a huge collection of control systems. Without control our heart would malfunction. The home appliances like refrigerators, televisions and audio players are all controlled according to some control laws. Automobiles and airplanes are also full of control systems. This suggest that control is indeed ubiquitous. Still, the abundance of control is not as widespread as it should have been. Note that the examples cited above and taken from [119] are all examples of drift control! Why are not every refrigerator fitted with a shift controller that gives an early warning signal of impending failure and identifies the assignable cause for the same? Why do we have to wait till the malfunctioning of the refrigerator becomes apparent and then spent quite some time in identifying the assignable cause? The answer perhaps lies in the economics of the business and lack of process knowledge.

In the previous reference [119], it is also noted that the application of first feedback control is the ancient water clock of Ktesibios in Alexandria Egypt, which dates back to around the third century B.C. On the other hand, sharpening of the cutting/hunting tools or resorting to more practice in response to poor cutting/hunting performance can be considered as a crude form of shift control and such a control system is likely to be in place since prehistoric days. However, the following timeline shows that the development of shift control has lagged far behind compared to drift control.

Prehistory	: Applications of shift control
Third century B.C	: Application of feedback control
1868	: Foundation of mathematical control theory by J. C. Maxwell (better known for the discovery of Maxwell electromagnetic field)
1923-1941	: Foundation of stochastic control (Wiener-Kolmogorov theory)
1924- 1931	: Foundation of modern shift control by Shewhart [1]
1952-1954	: Foundation of optimal control theory (Development of Dynamic Programming (DP) by Bellman [120])
1956	: Development of methods for economic design of control charts by Duncan [121]

*The following paper is based on this chapter: S. Chakraborty and P. Mandal. "Stochastic optimization for controlling the end state of hydrogenation of edible oil in the presence of process failure". *International Journal of Operations Research, 16(*2019*)* 91-104.

It is evident from the above timeline that modern shift control has been developed much later since fundamentally it is a stochastic control system. So, unlike drift control it has no history of deterministic control. So far as stochastic control is concerned, both the shift and drift control approaches have been developed around the same time. In fact, prior to the publication of the book in 1931, Shewhart used control charts during mid 1920s at AT&T Bell Labs, USA. The subsequent developments of both the fields are also seen to have taken place in a similar pace. It is interesting to note that the above general trend of development, i.e. the transition from a deterministic system to a stochastic system is not maintained at the micro level. For example, DP has been originally developed for stochastic processes. That the methodology is also applicable for a deterministic system has been realized much later [122].

Both shift and drift control have the same objective, which is to reduce the variation of process output. However, since the methodology and the tools used in the two approaches are different, they have developed nearly independently. The developments in drift control have been driven primarily by the control engineering group, while shift control has been primarily driven by the statisticians and the quality control group. Conceptually speaking, there is also a difference in the nature of the two approaches - drift control is aggressive, while shift control is conservative. The aggressive nature of drift control is manifested by the fact that it is entirely model driven. Once the suitable control architecture and the control scheme are developed, there is no looking back at the system. In contrast, the shift control also makes use of a process model but the purpose here is only to specify the process dynamics for a given process condition. Subsequently, efforts are spent in testing whether the specified model is correct or not and finding the assignable causes, if the model is found incorrect. Thus, in drift control, there is a predefined mapping of the state space to the control space. But in shift control, the control space is much larger and hence a predefined map as in case of drift control is usually absent. However, the presence or absence of the mapping from the state space to the control space is as the defining feature of the two types of control. For example, in the control chart for detection of growth failure (discussed in the previous chapter), the mapping from the state space (active vs. dead) to the control space (add a certain amount of fresh catalyst in case of a dead signal) is predefined. The distinguishing feature of this chart as a shift controller is that the out-of-control signals are to be investigated off-line (perhaps using another control chart) to find the causes of failures and prevent their recurrence. If this feature is absent, then of course, the control chart acts merely as a drift controller.

The importance of controlling both shift and drift is now well recognized and this has been illustrated with the help of the tyre-pressure example in Chapter 1. To highlight the

importance of shift control further, let us consider a slightly modified version of this example. Instead of the earlier scheme of pressure adjustment once in every three days, let us now assume that the pressure can be adjusted as frequently as is practicable. What is going to happen now in the presence of a tiny pin hole? Can we ignore the existence of the pin hole and continue forever? It will obviously not be prudent to do so. For an initial period, the adjustment will do its job and the shift controller (monitoring pressure) will fail to provide any out-of-control signal. But over time, the pin hole may be expected to become larger and eventually the tube will fail. If one is unlucky, the tyre may even burst causing greater damage. There are two important points to note from this example. First, it is difficult to monitor a drift controlled process. Perhaps monitoring of both the input and the output signal may be more effective than monitoring only one of the two [123]. Secondly, in the presence of an assignable cause, the drift controller may even act like an additional source of noise.

Realizing the potential benefits as above of an integrated approach, a large variety of integrated controllers has been developed in the past. A few of the early studies [12-15] in this area have already been indicated in Chapter 1. Some additional references are [123-132], which cover a wide variety of situations and approaches - feedback and feed forward architecture, univariate and multivariate set up, parametric and nonparametric process models, manufacturing and process industry, batch and continuous process, economic controller and even integration of on-line with off-line QC. However, it must be mentioned here that not all of the above studies uphold the true spirit of shift control. In many cases, the control chart is used only to modify the parameters of the drift controller. As already noted in Chapter 1, such control schemes are better viewed as supervised drift control rather than integrated shift and drift control.

Furthermore, all the studies mentioned above deals only with static characteristics and the process drift is considered as noise that can be corrected. However, our interest lies here in growth process regulation. The related past work in this area is briefly reviewed in the next section.

6.1. Related work

There are two types of tasks in growth process regulation - (i) Trajectory tracking and (ii) End state control. The second problem is a sub-problem of the first one. But, there are many processes (as in this study) where trajectory tracking may not be feasible but it is desired to control only the end state of the process. It should be noted here that by trajectory tracking we

mean the tracking of the trajectory of the growth characteristic (process output) and not the trajectory of an input process variable. In fact, there are many processes where the levels of an input variable may be specified in the form of a path or a trajectory even though the output characteristic of interest is a static characteristic. The examples of industrial processes, where such control approaches are widely used include crystallization [132a], polymerization [132b] and fermentation [132c, 132d]. It may be noted further that many growth processes exhibit more than one distinct intermediate phases within a growth cycle. In such cases, the detection of the onset of a new phase is often a critical issue in process control. However, such problems are not discussed separately since, in principle, intermediate state control is similar to end state control.

The main difficulty in implementing trajectory tracking for a growth process, as already noted, is the absence of a reference path. However, if an appropriate reference path can be provided then there are a large number of algorithms for trajectory tracking to choose from. Aerts et al. [22] have adopted such a strategy for regulating the growth path of broiler chicken. In fact, this is the only study on trajectory tracking of a growth process that has come to our notice. In this study, the authors first specify the reference path based on the generally accepted view that a good growth trajectory should consist of an initial regime of slow growth followed by faster compensatory growth. Then the observed growth is regulated by varying the feed stock following the method of Model Predictive Control. Based on the results obtained, the authors opine that it is very hard to specify a feasible optimal growth trajectory, i.e. a growth trajectory that can be tracked and is optimal in some sense. In particular, the birds can be grown following a predefined trajectory provided the trajectory is biologically feasible. It is also reported that the average weight of the birds at the end of 42 days of growth is significantly lower for the group that is subjected to feed control than the control group (not subjected to feed control). This is despite the fact that there is no significant difference in the feed conversion ratio (Kg feed/Kg weight gain) for the two groups. The above results clearly indicate that active regulation of the growth trajectory, in this case, may not be appropriate at the present state of our process knowledge. This is a very general observation in the sense that one needs to have sufficient process knowledge for implementing control in any form, whether it is for controlling the drift or process shifts.

From the point of view of controlling the end state of a growth cycle, where many similar cycles of growth are observed over time, two different scenarios appear in practice – one in which the start and the end time in each cycle remain the same and the other in which the end time may vary from cycle to cycle. In case of the former, if there is an additional constraint

that the state of the process can be measured only at the end, then the end state can be controlled following the method of Terminal Iterative Learning Control (TILC) proposed by Xu et al. [133] and enhanced recently by Chi et al. [134]. TILC is a special case of Iterative Learning Control [135]. It is to be noted that the TILC belongs to the category of active regulation, since the control action(s) are meant to influence the subsequent evolution of the growth process.

On the other hand, if the end time is flexible then the end state may be controlled by terminating the process at the right time, i.e. through passive regulation. Depending on the number of measurements (n) that can be made in each cycle of the process (excluding the initial state), the passive regulation problems can broadly be classified into three groups: (i) $n = 1$, (ii) $n = \infty$ (representing continuous measurement) and (iii) $1 < n \leq nmax$.

If $n = 1$, then one may use the observations on (Y_f, X, Δt), where Y_f is the end state corresponding to the cycle time Δt and X is the vector of relevant process variables to develop a suitable model for predicting $\Delta t \mid (X, T)$, where T is the target for Y_f. The current and the past observations can then be used to develop a suitable scheme for updating the model. This is a low cost approach and is used, for example, in electroplating for controlling the coating thickness.

The second case represents situations where continuous monitoring of the process is possible using on-line sensors, e.g. in-situ real time measurements for detection of the end point of a plasma etching process. It is obvious that if the growth characteristic of interest can be measured continuously with sufficient accuracy as it evolves then end point detection becomes a trivial problem. However, in most practical situations this may not be feasible. So, the end point is detected by monitoring one or more process variables that show significant variation at the end point. For example, in case of plasma etching, the process variables that exhibit a significant change in their behaviour are process pressure, overall impedance, concentration of the plasma reactants and a few others [136]. Thus the end point detection becomes the problem of change point detection in one or more of these variables, which is certainly not a trivial problem. The degree of difficulties involved depends upon the choice of the variables, the amount of noise present in the data and the criticality of application. Thus, in some applications the end point detection may be a simple graphical exercise [137] , while in another situation it may involve the use of sophisticated data analytic tools like Principal Component Analysis and Independent Component Analysis [136].

The third case represents passive regulation based on periodic sampling and measurement. Such control problems are encountered widely in industry, agriculture and also in everyday

life. For example, farmers decide on the date of harvesting based on periodic inspection. In manual cooking of rice, periodic sampling is used to decide the end point of cooking. In LD converter based steel making, samples are drawn periodically to control the phosphorus content of steel. However, to the best of our knowledge, no previous work on this particular problem has been reported so far. In this work an attempt is made to formulate and solve this particular problem. As before the proposed methodology is illustrated with the help of the hydrogenation example.

6.2. Passive regulation based on periodic sampling

As before, let the random variable $Y_t > 0$ represents the state of the growth characteristic at time t and $E(Y_t)$ is a continuous and monotonically increasing function of $t > 0$. Let T be the target of Y_t. It is assumed that the process can be observed at arbitrary time points (t_1, t_2, ..., t_k), $k \leq nmax$ to obtain a sequence of observations $\{(y_1, t_1), (y_2, t_2), ..., (y_k, t_k)\}$. In its simplest form, i.e. assuming a single stage process without the possibility of growth failure, the control problem is to determine the sequence (t_1, t_2, ..., t_k) so that if the process is terminated at t_k, then y_k will be closest to T. But, how do we make the decision of terminating the process? If the state of the process can be evaluated instantaneously and have a policy to take most of the samples near the end of the cycle (without violating the limit of *nmax*) then we may take $y_k \geq T$ as the process termination rule. Note that such a strategy is based on a practical approximation of the process situation by the second case mentioned at the beginning of Section 6.1, i.e. as if an on line sensor is available for measuring the growth characteristic directly near the end of the cycle. Clearly, such a strategy will not work in most practical situations. For example, there may be a significant lag in measurement, which will impose a natural restriction like $\Delta t_i = t_{i+1} - t_i \geq lag$, where Δt_i is the i_{th} sampling interval and *lag* is the measurement lag. So, in general, it is necessary to have a stopping rule other than $y_k \geq T$. Let us consider a simple stopping rule like $y_k \geq T_1$, where $T_1 < T$ is the pseudo process target. The control problem can now be stated as follows: Determine (T_1, t_1, t_2, ..., t_k), $k \leq nmax$ that minimizes the average of total cost TC = Sampling cost + End cost [or a similar quantity like $E(TC^2)$].

Note that it is not possible to satisfy the constraint on k if we have a stopping rule $y_k \geq T_1$. There will be occasions when $y_{nmax} < T_1$. It is thus necessary to treat one of the two constraints as a soft constraint and impose a penalty cost for violating the constraint. Let us assume that the restriction on k is the soft one. Thus, the control problem now becomes -

Determine (T_1, t_1, t_2, ..., t_k), that minimizes the average of total cost TC = Sampling cost + End cost + Penalty cost.

The above problem can be generalized to a m-stage process in a straightforward manner. However, here we shall consider only a simple extension of the above, i.e. a two-stage process where we have no control over the second stage. The control problem for such a two-stage process remains the same as above but now the end cost needs to be computed based on the deviation of ($y_k + \Delta y$) from T instead of the deviation of y_k from T, where Δy is the growth during the second stage of the process.

Let us now generalize the above two-stage process to consider the occurrence of growth failure. It is assumed that the growth failures are detected with the help of a control chart like those discussed in the previous chapter. Note that if a control chart is used for detection of growth failure, then the longer the sampling interval, the easier will it be to detect failure based on the first sample taken after failure. Also, the longer the delay in detection of growth failure, the larger will be the failure cost. Under such a situation, it is obvious that no naive approach should be expected to provide even a near optimal solution. However, the basic control problem remains the same. But, now we have additional constraints in the form a distribution of failure time and a procedure for detection of growth failure. Accordingly, the objective function also gets modified as the average of TC = Sampling cost + End cost + Failure cost + Recovery cost.

Finally, it is necessary to specify the initial state of the process to complete the specification of the control problem. If the initial state at $t = 0$ is unknown (as here), then the same needs to be specified for some other value of $t > 0$. So the determination of t_1 becomes a part of the problem specification. It is assumed that the distribution of Y_t at $t = t_1$ can be specified.

Clearly, the above problem can be formulated as a Stochastic Dynamic Program (SDP), although it may not be possible to find an exact solution for the same. So, by way of approximation, let us assume that the sampling points (t_2, t_3, ..., t_k) and the pseudo process target T_1 for a given (y_1, t_1) are generated by the policy functions P1 and P2 respectively. Thus, the final control problem can now be stated as follows: Given the process model, the distribution of Y_t at $t = t_1$, the distribution of failure time, the control chart for detection of growth failure, the distribution of Δy during the second stage and the policy functions P1 and P2, estimate the parameters of P1 and P2 that minimizes E(TC).

In principle, the above problem can be solved by estimating the parameters of P1 and P2 simultaneously. But it will be much simpler to estimate the two policies separately. In this

study, the later approach is adopted for the sake of simplicity. The policy P2 is determined first by solving a single-stage stochastic optimization problem. The optimal P2 thus obtained is then used to obtain P1 following a simulation based method of optimization. The policy P2 is obtained in a tabular form, while P1 is obtained in a closed form. The details of the problem formulation and the proposed methodology of optimization are illustrated with the help of the hydrogenation example in Section 6.4.

6.3. Developing an integrated shift and drift control system

The three components of an integrated system are : (i) The control chart for monitoring an active batch (C1), (ii) The control chart for detection of a dead batch (C2) and (iii) The passive regulator (R) for determining P1 and P2. The above three components can be integrated in a varieties of ways. For example, the regulator R may consist of three sets of policies (P1, P2) and the chart C1 may be used for choosing one of the three sets of policies depending on whether $y_i \geq$ UCL or LCL $< y_i <$ UCL or $y_i \leq$ LCL, where UCL and LCL are upper and lower control limits respectively of C1. However, as already noted in Chapter 1, such a system is better viewed as a supervised drift control system rather than an integrated system. In this study, the chart C1 is proposed to be used only for finding the assignable causes. Of course, one may have more than one set of policies (P1, P2), if the dynamic behaviour of the process is found to be significantly different in different zones. But such zones need to be identified separately.

The control chart C2 can however be used to define two different sets of (P1, P2) depending on the nature of the signal obtained from the chart (active or dead). This is because the method of recovery remains the same (addition of fresh catalyst in case of the hydrogenation process) whenever the 'dead' signal is obtained. With such an approach, C2 becomes an integral part of R. Again, as noted before, the control limits of C2 may be obtained simultaneously with P1 to minimize E(TC). However, for the sake of simplicity and also for using C2 as an independent tool, the control limits of C2 are determined separately, i.e. without considering its impact on P1.

6.4. Integrated control system for the process of hydrogenation

The main features of the hydrogenation process have already been discussed in Section 1.7, Chapter 1. However, for the sake of convenience, the problem is restated here with some additional details.

During hydrogenation, the melting point of a batch of oil rises gradually over time. The specifications for final melting point, which is the most important quality characteristic that needs to be controlled, is 40.0 ± 0.5^{0}C. This is achieved by terminating the growth process at an appropriate time. In other words, the approach of passive regulation is used for controlling the process.

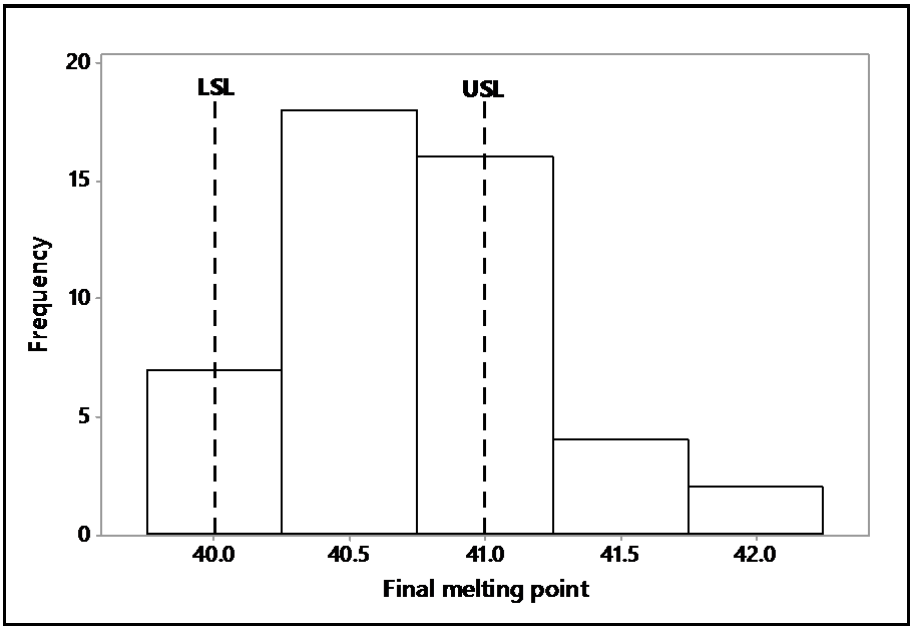
However, the hydrogenation process also suffers from occasional growth failure at a random time point leading to the occurrence of what is known as a 'dead batch'. It is obvious that such failures need to be detected as quickly as possible so that the cost of production remains under control.

In order to perform the twin tasks of detection of a dead batch and controlling the final melting point, samples are drawn periodically from a batch and the melting point of the samples are determined off-line to judge the status of the batch.

Thus, so far as process regulation is concerned, the three important decision variables for the control system are (i) The sampling points $\{t_i\}$, (ii) A state variable representing the status of the batch (whether dead or active) and (iii) The cooling limit (MP_C), i.e. the melting point at which the active hydrogenation phase should be terminated. In the notation of Section 6.2, MP_C is the same as T_1.

An important constraint of the system is the maximum number of samples that can be drawn from each batch ($nmax$ = 6). This constraint represents the capacity constraint of the testing laboratory and also the hazard associated with frequent sampling.

Figure 6.1. Histogram of final melting point (^{0}C) along with its lower and upper specification limits



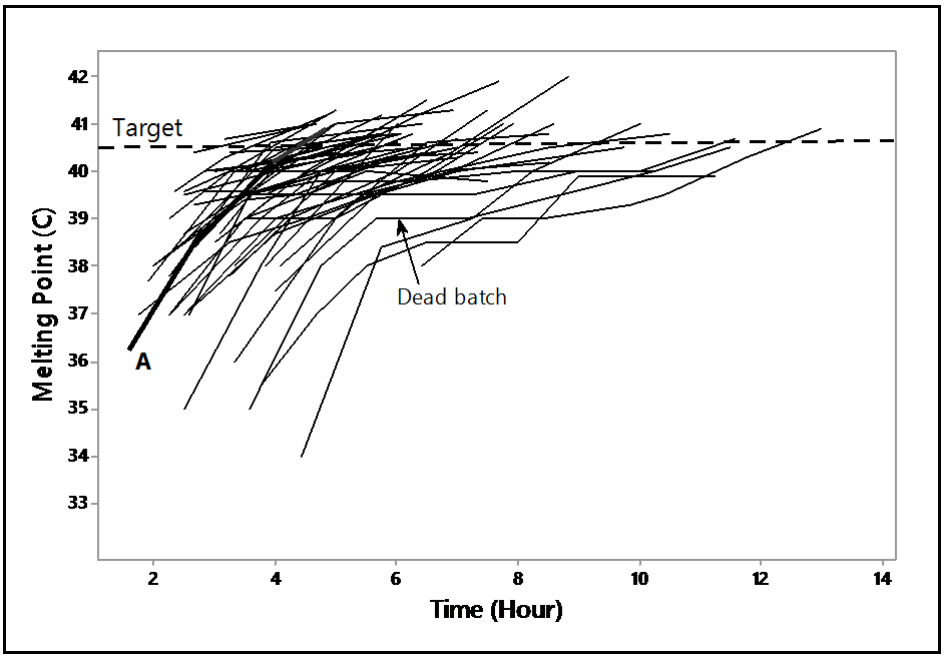
In the existing system the process engineer decides on the most appropriate course of actions for controlling the process based on his or her subjective judgment. Also, there is no formal system for controlling the process shifts. As a result, the performance of the process with respect to both final melting point and cycle time is found to be very poor. It is seen from Figure 6.1 that the distribution of final melting point is positively skewed and the melting point exceeds the upper specification limit on many occasions. The cycle time (till cooling) is also found to vary widely from 2.5 – 10 hours.

Our objective is to develop an integrated shift and passive regulation system for the process of hydrogenation.

6.4.1. Why passive regulation?

In general, it is risky to regulate a growth characteristic to track a specified growth trajectory. As noted in Section 6.1, the past experience of Aerts et al. [22] in this regard is also not encouraging. Implementation of trajectory tracking for the hydrogenation process also does not appear to be an attractive option. To illustrate, let the ideal or the reference trajectory for the process be given by curve A as shown in Figure 6.2. The other curves shown in this figure are the growth curves for the fifty sample batches. The question is - should the growth trajectory of all the batches be regulated to follow curve A?

Figure 6.2. Growth curves of fifty sample batches and an arbitrary reference path A



It is obvious from Figure 6.2 that very large corrections need to be made on many occasions to achieve the same and this may not be feasible in actual practice. Further, even if making such large corrections is technically feasible, it may not be advisable to do so for the following reasons: (i) Designing the controller will be a very difficult task due to the unstable nature of the process, (ii) It may even be counterproductive, since other process outputs may get adversely affected due to the large corrections that will need to be made, and (ii) Opportunities for reducing the process variation by removing the assignable cause(s) will be lost. In other words, if the choice is between the control of drift (whether deterministic or stochastic) of an unstable process and the same after a period of shift control, then, in general, the latter should be preferred. Note that the approach of passive regulation in the present case will provide full opportunity for shift control. Of course, the scope for active regulation may arise in future, once the process is sufficiently stabilized.

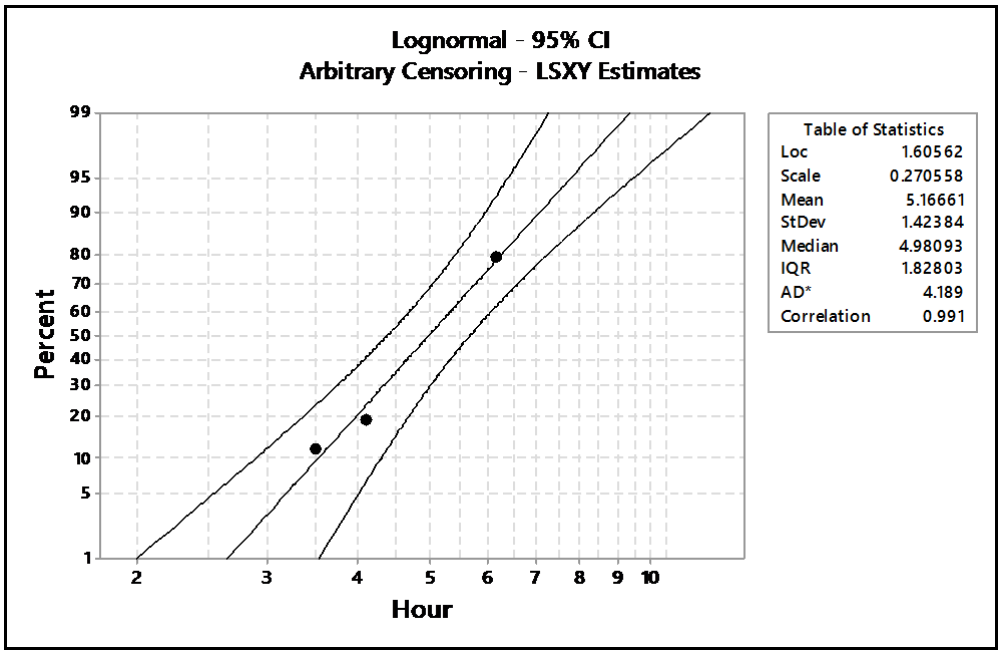
6.4.2. Main elements of the system

As noted in Sections 5.2 and 5.3, the five main elements of the system are (i) The growth model for the hydrogenation phase, (ii) The growth model for the cooling phase, (iii) The failure model for the occurrence of dead batches, (iv) The control chart for detection of dead batches and (v) The initial condition of the process. The dynamic model (5.11) will be used here as the growth model for the hydrogenation phase and the control chart given in Section 5.5.2.2 will be used for detection of dead batches. The remaining three elements are described in this section.

6.4.2.1. Distribution of failure time

The control chart in Figure 5.8 shows that seven out of the fifty sample batches can be considered as dead. The three batches with points falling above the *UCL* are not considered. Thus, the failure data consists of seven observations that are interval censored (only the interval of failure is known, not the exact time of failure) and forty observations that are right censored (only the times till the batches are active are known). It is further observed that the failure times (t_d) for a few batches are exceptionally high due to the inordinate delay in the addition of fresh catalyst. This suggests that the process has remained idle for some time, perhaps due to some process problems like malfunctioning of the vacuum pump. Consequently, it is decided to model the distribution of $(t_d - t_0)$ instead of (t_d), where t_0 is time at which fresh catalyst is added for the first time.

Figure 6.3. Lognormal probability plot of failure time (t_d - t_0)



The methods of estimation of the parameters of a life distribution with censored data are available in standard text books on reliability (for example, see [138]). Here the commercial software MINITAB is used for developing the model. It is seen from Figure 6.3 that the Log-normal distribution gives a good fit to the observations. The least square estimates of the location and scale parameters are found to be 1.606 and 0.271 with 95% confidence intervals as (1.476, 1.735) and (0.185, 0.395) respectively.

6.4.2.2. Distribution of growth during cooling

Let the growth during cooling be denoted by y'. Also let the cooling decision be based on the melting point $MP_{t'}$ of the last but one sample (taken at t'). Thus, since there is a measurement lag of twenty minutes,

$$MP_f = MP_{t'} + \Delta MP_{lag} + y' = MP_{t'+lag} + y', \qquad (6.1)$$

where MP_f is the final melting point of the batch. Now, assuming that a batch is active throughout the growth process, the estimates of $MP_{t'+lag} | (MP_{t'}, \Delta t = lag)$ are obtained using the process model (5.11). These estimates are then substituted in (6.1) to have the estimates of y'. As before, the three out-of-control batches (that violate UCL) are not considered for modeling the distribution of y'. However, it is observed that in six out of the remaining forty-seven batches, the growth during cooling is zero, i.e. $MP_f = MP_{t'}$. It is seen from Figure 6.4

that if these six batches are excluded, then the estimates of $y' > 0$ gives a very good fit to the two-parameter Weibull distribution. Thus, the following mixture of Bernoulli and Weibull is specified as the distribution of y',

$$\begin{aligned} f(y') &= p', && y' = 0, \\ &= (1-p')\, f_1(y'), && y' > 0, \end{aligned} \tag{6.2}$$

where p' is the probability of no growth during cooling (=6/47) and f and f_1 are the probability density functions. The density $f_1(y')$ of the Weibull distribution is given by

$$f_1(y') = (\beta/\alpha)(y'/\alpha)^{\beta-1} \exp[-(y'/\alpha)^{\beta}],\ \alpha > 0,\ \beta > 0,\ y' > 0. \tag{6.3}$$

It is seen from Figure 6.4 that the maximum likelihood estimates of α and β are 1.504 and 0.588 respectively. The mean (μ') and variance (σ'^2) of $y' > 0$ are given by

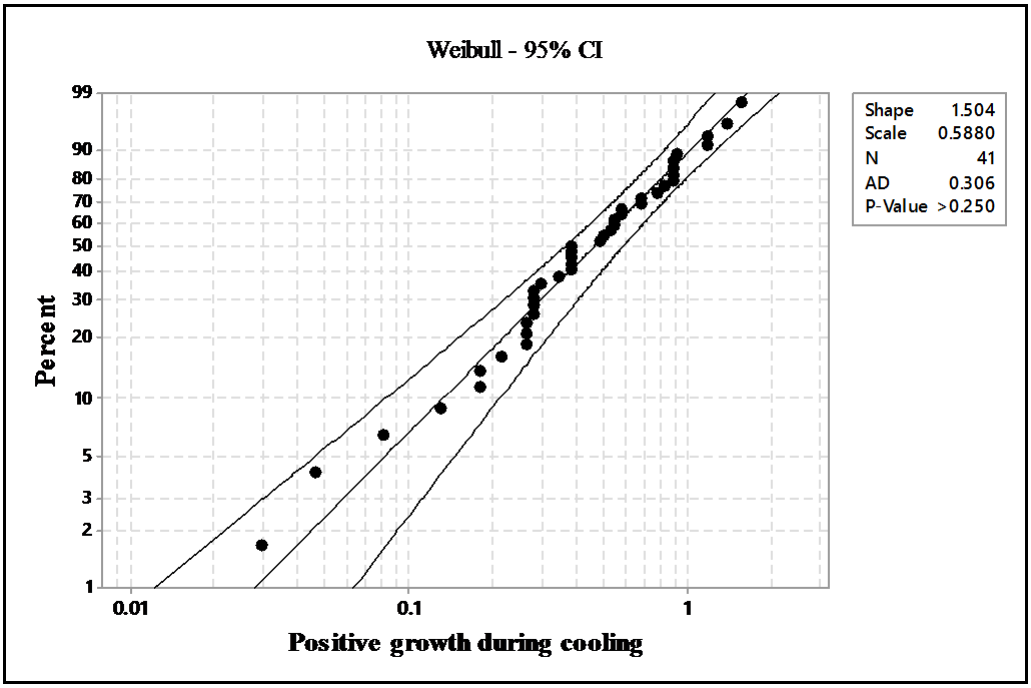
$$\mu' = \beta\,\Gamma(1+1/\alpha),\ \ \sigma'^2 = \beta^2[\Gamma(1+2/\alpha) - \Gamma^2(1+1/\alpha)]. \tag{6.4}$$

This gives, $\hat{\mu}' = 0.5306$ and $\hat{\sigma}'^2 = 0.1292$. Further, it can be easily shown that

$$E(y') = (1-p')\mu' \text{ and } Var(y') = (1-p')(p'\mu'^2 + \sigma'^2). \tag{6.5}$$

Using (6.5), the estimates of mean and variance of y' are obtained as 0.4628 and 0.144 respectively.

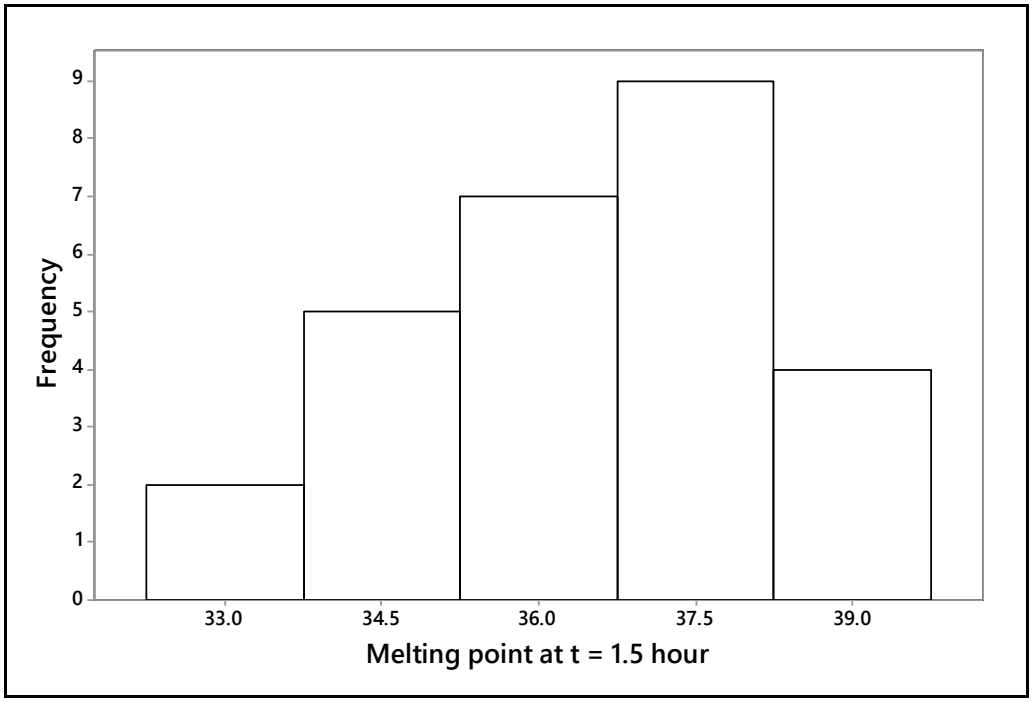
Figure 6.4. Weibull probability plot of $y' > 0$



6.4.2.3. Initial condition

The initial melting point of a batch (at $t = 0$) is neither measured nor known. The earliest a sample has been drawn in the fifty sample batches is at $t = 1.75$ hour. However, considering the importance of detecting the status of the process as early as possible, it is decided that the first sample will be taken at $t = 1.5$ hour. The distribution of failure time as estimated above shows that the probability of failure of a batch before $t = 1.5$ hour is nearly zero. The estimates of the melting point at $t = 1.5$ for a few of the sample batches are obtained as follows. Let MP_1 be the melting point of the first sample drawn at $t_1 > 1.5$ for a given batch. Then the desired estimate MP_0 is such that the expected value of MP given by the process model (5.11) for $\Delta t = t_1 - 1.5$ is MP_1. Also, considering the risk of extrapolation and the fact that the model (5.11) is valid only for $\Delta t \leq 2.0$ hours, the backward extrapolation is restricted to only those batches for which $t_1 \leq 3.5$. The distribution of the estimates thus obtained is shown in Figure 6.5.

Figure 6.5. Histogram of melting point at $t_1 = 1.5$ hour



It is seen that the distribution is negatively skewed. Although the sample size here is very small, the pattern of variation of the growth curves in Figure 6.2 suggests that the long left tail is likely to be due to the occasional occurrence of out-of-control batches. Consequently, a stable distribution of melting point at $t = 1.5$ may not exist. But it is necessary to specify a reasonable distribution to develop the control scheme that will be valid under a wide range of

operating conditions. Considering the above and the advantage of having a bounded distribution for the purpose of simulation, it is decided to use the following truncated normal distribution as the distribution of $MP_{t=1.5}$,

$$MP_{t=1.5} \cap TN(36.8, 2, 32, 39.6), \quad (6.6)$$

where the parameters signify that a normal distribution with mean 36.8 and standard deviation 2 is truncated at 32 and 39.6. The mean and standard deviation of $MP_{t=1.5}$ with the parameters as above are very close to the mean and standard deviation of the observed distribution (Figure 6.5).

6.4.3. Problem formulation

This section consists of four subsections describing the state variables, the development of the cost function, the penalty function and the formal mathematical formulation of the problem as a Stochastic Dynamic Program (SDP).

6.4.3.1. State variables

Let S_0 represents the true status of a batch being hydrogenated, i.e. S_0= {Active, Dead}. The observed status of the batch is defined similarly and is denoted by S_1. However, since the observation on S_1 is made by a control chart, it will be unrealistic to neglect the error involved in this measurement. So let the four possible scenarios related to this measurement be denoted by $S = \{s_1, s_2, s_3, s_4\}$, where s_1 = Dead and detected, s_2 = Dead but undetected, s_3 = Active and not identified as dead, and s_4 = Active but identified as dead. The quality cost associated with a particular growth trajectory will be evaluated taking all these four possible scenarios into account.

The measurement error associated with the other state variable *MP* is a component of the error variance of the process model given by (4.32).

6.4.3.2. Cost function

Let the total quality cost (C) associated with a growth path (ω) be given by

$$C(\omega) = C_P(\omega) + C_E(\omega), \quad (6.7)$$

where C_P is the path cost and C_E is the end cost. The path cost C_P has two components,

$$C_p = C_1 + C_2, \quad (6.8)$$

where C_1 is the cost of sampling and testing and

$$C_2 = C_d + C_r, \quad (6.9)$$

is the failure cost. The failure cost has two components - detection cost (C_d), i.e. the cost of the delay in detecting a dead batch and the cost of recovering (C_r) a batch that has been identified as dead, whether rightly or wrongly.

Let a batch that becomes dead at t_d is detected at the j_{th} sample. Since there is a measurement lag of twenty minutes (i.e. lag = 1/3 hour), the delay in detection or the failure time (FT) is given by $FT = t_j + lag - t_d$. It is possible that a batch fails between (t_j, t_j + lag) and it is decided to cool the batch based on the sample drawn at t_j. There is a cost (in the form of loss of information) associated with such failures but such costs (being very small) are not taken into account.

The cost of recovery is assumed to be proportional to the time lost in making arrangement for fresh catalyst addition, which is estimated approximately as 10 minutes, i.e., the recovery time (RT) per failure is 1/6 hour. The cost of catalyst is not considered since the catalyst remains in the system (recycled). Of course, the catalyst does suffer from some deterioration in its performance over time, but for the sake of simplicity, cost of deterioration associated with recycling is ignored. Further, it is assumed that a batch can become dead at most once during the entire cycle.

The end cost is the quality loss component, which is assumed to be given by the quadratic loss function

$$C_E = \lambda(MP_f - 40.5)^2, \tag{6.10}$$

where λ is the loss coefficient and 40.5 is the target of final melting point MP_f. It is to be noted that if a dead batch is not recovered before cooling then MP_f is the MP at t_d, otherwise it is given by (6.1).

The total cost of quality for a batch is then given by

$$C = c_1(NS + 1) + c_2(FT + RT) + \lambda(MP_f - 40.5)_2, \tag{6.11}$$

where c_1 is the cost of sampling and testing per sample, NS is the number of samples drawn till cooling and c_2 is the hourly cost of downtime. The economics of the process suggests that c_2 is approximately 100 times of c_1 and λ is approximately 3.5 times of c_2. Thus, the cost function becomes

$$C(\omega) = (NS + 1) + \sum_{j=1}^{NS(\omega)} [100(a_j l_j + b_j(RT)] + 350(MP_f - 40.5)^2, \tag{6.12}$$

where t_j, j = 1, 2, ..., NS are the time points at which the samples are drawn and a_j = 0, if the batch is active at t_j and 1, otherwise. The failure time l_j for the j_{th} interval (t_{j-1}, t_j) is given by

$$l_j = t_j + lag - t_d, \ if \ t_{j-1} < t_d \le t_j \ [i.e.\ S_j = s_2 \ or \ s_3],$$
$$= t_j - t_{j-1}, \qquad if \ t_d < t_{j-1} \ [i.e.\ S_{j-1} = s_2]. \tag{6.13}$$

Also, $b_j = 1$, if $S_j = s_1$ or s_4 and $MP_j < MP_C$, where MP_C is the cooling limit.

6.4.3.3. Penalty function

It has already been noted in Section 6.2 that the constraints on both MP_C and *nmax* cannot be satisfied simultaneously. There will be instances of MP_j being less than MP_C at $j = nmax$. So it is decided to treat the limit on *nmax* (= 6) as a soft constraint, where *nmax* includes the last sample taken at the end of the cooling phase. Thus, for the purpose of optimization, the following penalty function is added to the original cost function [Equation (6.12)],

$$C_p(\omega) = 100(NS(\omega) - 5)^3, \ if \ NS > 5,$$
$$= 0, \ otherwise. \tag{6.14}$$

6.4.3.4. Mathematical formulation of the problem

Let, for a given batch, the sampling intervals are defined by $\Delta t_j = t_j - t_{j-1}$, $j = 2, 3, ..., NS$. Also let A, D and R be the set of all active, dead and recovered intervals respectively, i.e. A contains the intervals in which no failure occurs, D contains those intervals in which the batch fails at time $t_{j-1} \le t_d \le tj$ and R is the set of intervals in which the control chart gives an out-of-control signal. For a given batch, the set B is either empty or has at most one element (by assumption). Further, the variable t_0 is generalized to t_{0j} to consider the time of fresh catalyst addition for each interval. The SDP can now be formally stated as follows:

Minimize $E[C(\omega) + C_P(\omega)]$

subject to

(i) $t_1 = 1.5$.

(ii) The distribution of MP at t_1 is given by (6.6).

(iii) S_0= Active, at t_1.

(iv) $MP_j = MP_{j-1} + G_j, \ j = 2, 3, ..., NS,$

$G_j \bigcap N(\mu_j, \sigma_j)$, where μ_j and σ_j are given by (5.11), with

$$\Delta t_j = t_j - t_{j-1}, \ j \in A,$$
$$= t_d - t_{j-1}, \ \ j \in D,$$
$$t_{0j} = 0, \ j - 1 \in R,$$
$$= 1, \ otherwise.$$

(v) The distribution of t_d is given in Figure 6.3.

(vi) The mapping $S_{0j} \rightarrow S_{1j}, j = 2,3,....,NS$ is obtained by the control chart.

(vii) $MP_{NS} \geq MP_C$, where the cooling limits MP_C are to be specified (see Section 6.4.4.1)

(viii) $MP_f = MP_{t'+lag} + y'$; y' is defined in (6.1).

(ix) lag = 1/3.

(x) $0.5 \leq \Delta t_j \leq 1.5; \Delta t_j = 0.5,\ j-1 \in R$.

It may be noted from the last constraint that the length of the sampling intervals post recovery has been restricted to thirty minutes. This is to have some safeguard against the possibility of greater variation in growth behaviour of the process when fresh catalyst is added for the second time, particularly when such additions are made unnecessarily.

6.4.4. Solution methodology

It is noted in Section 6.2 that for the sake of simplicity, the two policies P1 (giving sampling intervals) and P2 (giving the cooling limits) are determined separately. Such an approach has an added advantage. If the cooling limits are found to be too low (say < 39.5^{0}C), then it will be necessary to put a constraint on MP_C and then solve the SDP. This is because if a batch becomes dead immediately after cooling then the end cost cannot be evaluated properly by the quadratic loss function in such cases. The losses are expected to be much higher than that given by the loss function. An independent exercise to find the cooling limits will then be helpful in specifying a suitable constraint on MP_C, if needed. However, in the present case, the standalone cooling limits obtained from an independent exercise are judged to be satisfactory.

6.4.4.1. Determination of optimal cooling limits

Let the present state of the process be denoted by MP_t. Then the determination of cooling limit involves a comparison of cost of cooling at MP_t with that of going ahead by one more step. It is decided to use the minimum permissible step size, i.e. Δt = 1/2, so that the limit becomes closer to the target. Thus, the determination of cooling limit becomes a single stage stochastic optimization problem with only two alternatives - start cooling at MP_t or continue hydrogenation for another thirty minutes. Of course, the costs will vary depending on whether the batch, at time t, is identified as active or dead. If it is identified as dead, then the options are either to start cooling at MP_t without recovering or to recover the batch and then continue hydrogenation for thirty minutes.

The scenario tree for the case when the batch at (*MP*, *t*) is identified as active by the control chart is shown in Figure 6.6. The two simplifying assumptions made in constructing the scenario tree are as follows: (i) The control chart cannot detect a dead batch if the sample is drawn less than half an hour after the occurrence of death and (ii) Second sample detection of growth failure is certain. The study on the power of the control chart (described in Section 5.5.2.3) reveals that the above two assumptions are very reasonable. Thus, the expected total cost of cooling at MP_t is given by (see also Figure 6.6)

$$E(TC_1) = (1 - p_1)C_A + p_1 C_D,$$

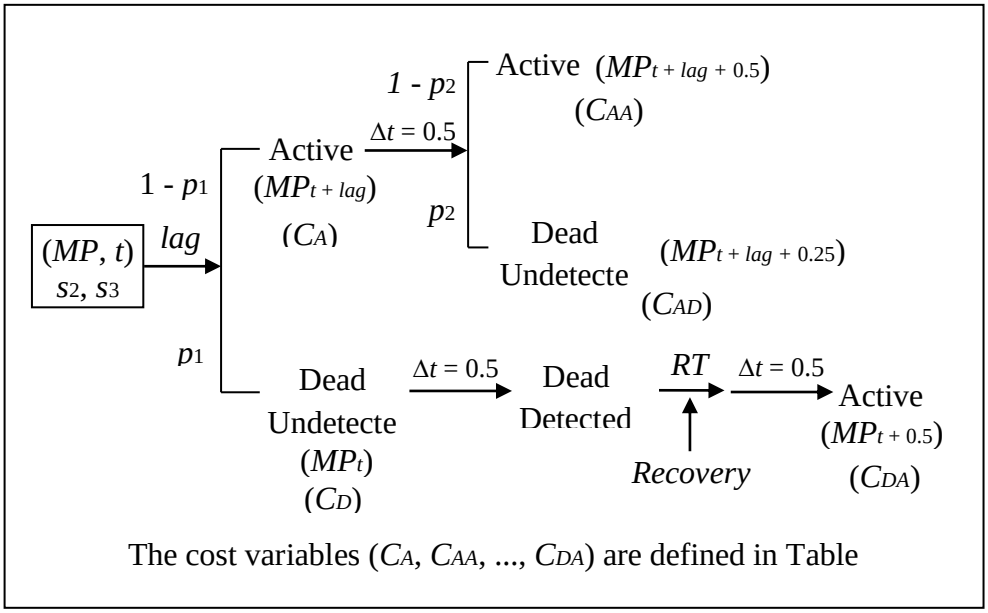
where p_1, by virtue of assumption (i) above, is approximately the probability of a batch becoming dead in the interval (t – 0.5, t + *lag*), given that it is active at t – 0.5.

On the other hand, instead of cooling at $t + lag$, if it is decided to continue hydrogenation for another thirty minutes then the expected cost is given by (see Figure 6.6)

$$E(TC_2) = (1 - p_1)[p_2 C_{AD} + (1 - p_2)C_{AA}] + p_1 C_{DA},$$

where p_2 is the probability of failure in the interval (t + *lag*, t + *lag* + 0.5), given that the batch is active at t + *lag*. The cost variables (C_A, C_D, ..., C_{DA}) involved in the above two equations are described in Table 6.1.

Figure 6.6. Scenario tree showing the possible states at the time of cooling when a batch is identified as active based on the sample drawn at time *t*



The cost component C_A is given by

$$C_A = C_1 + C_d + C_r + C_E = 350[\mathrm{E}(MP_f - 40.5)_2]$$

$$= 350[\{\mathrm{E}(MP_{t+lag}) + \mathrm{E}(y') - 40.5\}_2 + \mathrm{V}(MP_{t+lag}) + \mathrm{V}(y')],$$

where E(MP_{t+lag}) and V(MP_{t+lag}) are obtained from (5.11) and E(y') and V(y') are obtained from (6.5). The other costs are evaluated similarly.

The desired optimal cooling limit for a given t is given by the minimum melting point at which $E(TC_2)$ becomes greater than $E(TC_1)$.

Table 6.1. Cost variables for determining the cooling limits

Cost variable	*Description*	*Cost component**			
		C_1	C_d	C_r	C_E
C_A	Expected cost of cooling an active batch at ($t + lag$)	0	0	0	α_2E(MP_f - 40.5)^2 $MP_f = MP_{t+lag} + y'$
C_D	Expected cost of cooling a dead batch at ($t + lag$)	0	0	0	α_2(MP_f - 40.5)^2 $MP_f = MP_t$
C_{AA}	Expected cost of cooling an active batch after hydrogenation for 0.5 hour	1	0	0	α_2E(MP_f - 40.5)^2 $MP_f = MP_{t+lag+0.5} + y'$
C_{AD}	Expected cost of cooling a dead batch after hydrogenation for 0.5 hour given that it was active at the beginning	1	$0.25\alpha_1$	0	α_2E(MP_f - 40.5)^2 $MP_f = MP_{t+lag+0.25}$
C_{DA}	Expected cost of cooling a recovered batch after hydrogenation for 0.5 hour given that it was dead at 0.5 + RT hour before	2	α_1(0.5 + lag)	$k_1(RT)$	α_2E(MP_f - 40.5)^2 $MP_f = MP_{t+0.5} + y'$ $t_0 = 0$

*C_1 = Sampling cost, C_d = Detection cost, C_r = Recovery cost, C_E = End cost, α_1 = 100, α_2 = 350

The scenario tree for the case of a dead batch, i.e. when $S = \{s_1, s_4\}$ is constructed similarly. Here the first two branches at $t + lag$ correspond to the batch being genuinely dead or it is a false alarm. It may be noted that the probability of false alarm at the lower limit of the control chart (= -2.35) is very low (< 1%). Hence for all practical purposes this branch may be ignored. However, no such approximation is made here for computing the expected costs. The method of computation remains the same as before.

6.4.4.2. Determination of optimal sampling intervals

It is obvious from the formulation of the SDP that even if the cooling limits are specified beforehand, some sort of approximation needs to be made to determine the optimal sampling intervals. Powell [139] has classified the approximations comprising the body of ADP into

four categories: (i) Myopic cost function approximation, (ii) Look-ahead policy (iii) Policy function approximation and (ii) Value function approximation.

The specification of t_1 = 1.5 hour can be thought of as an outcome of the myopic cost function approximation. The focus here is to find the time where the batch can be considered as active in almost all cases, irrespective of its implications on future decisions. The optimal sampling intervals are obtained following a policy function approximation described below.

Policy functions

The policies chosen for determining the sampling intervals (*DT*) are as follows:

$$P_1 : DT = k_1(PT),\ 0 < k_1 \le 1, \tag{6.15}$$

$$P_2 : DT = k_2(PT),\ \ k_2 = m_0 + m_1(MP_t) + m_2(t - t_0),\ 0 < k_2 \le 1, \tag{6.16}$$

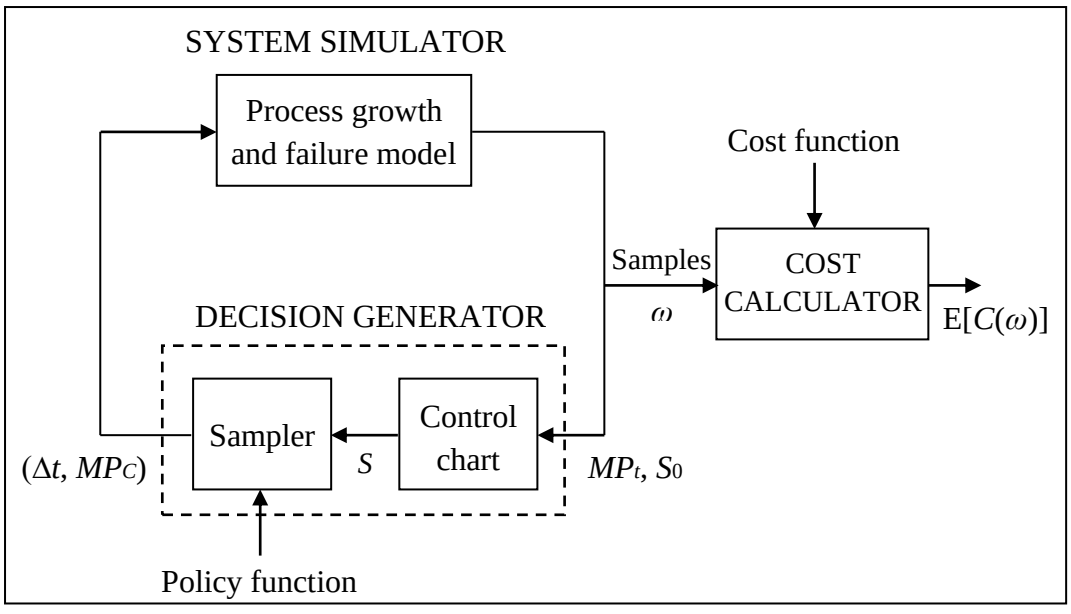
where *PT* is the predicted time to reach the melting point of 40^0C from the present state MP_t and k_1, k_2 are shrinkage factors. The predicted time is obtained from the process model (5.11). However, in order to satisfy the constraint $0.5 \le \Delta t \le 1.5$, the final sampling intervals are obtained as $\Delta t = Min\ (DT, 1.5)$, if $DT > 1.5$ and $\Delta t = Max\ (DT, 0.5)$, if $DT < 0.5$.

Thus the task now reduces to estimating the parameters of P_1 and P_2. The details of the simulation based optimization method used for this purpose are described below.

Simulation algorithm

The expected total cost of the control system for a given policy is evaluated through simulation. The high-level diagram of the simulation process is shown in Figure 6.7.

Figure 6.7. High level diagram of the simulation process



The flow chart for simulation of the growth curves and computation of the total quality cost is given in Figure 6.8. An R-code is developed for implementing the process as shown in Figure 6.8.

Figure 6.8. Flow chart for growth curve simulation and cost computation

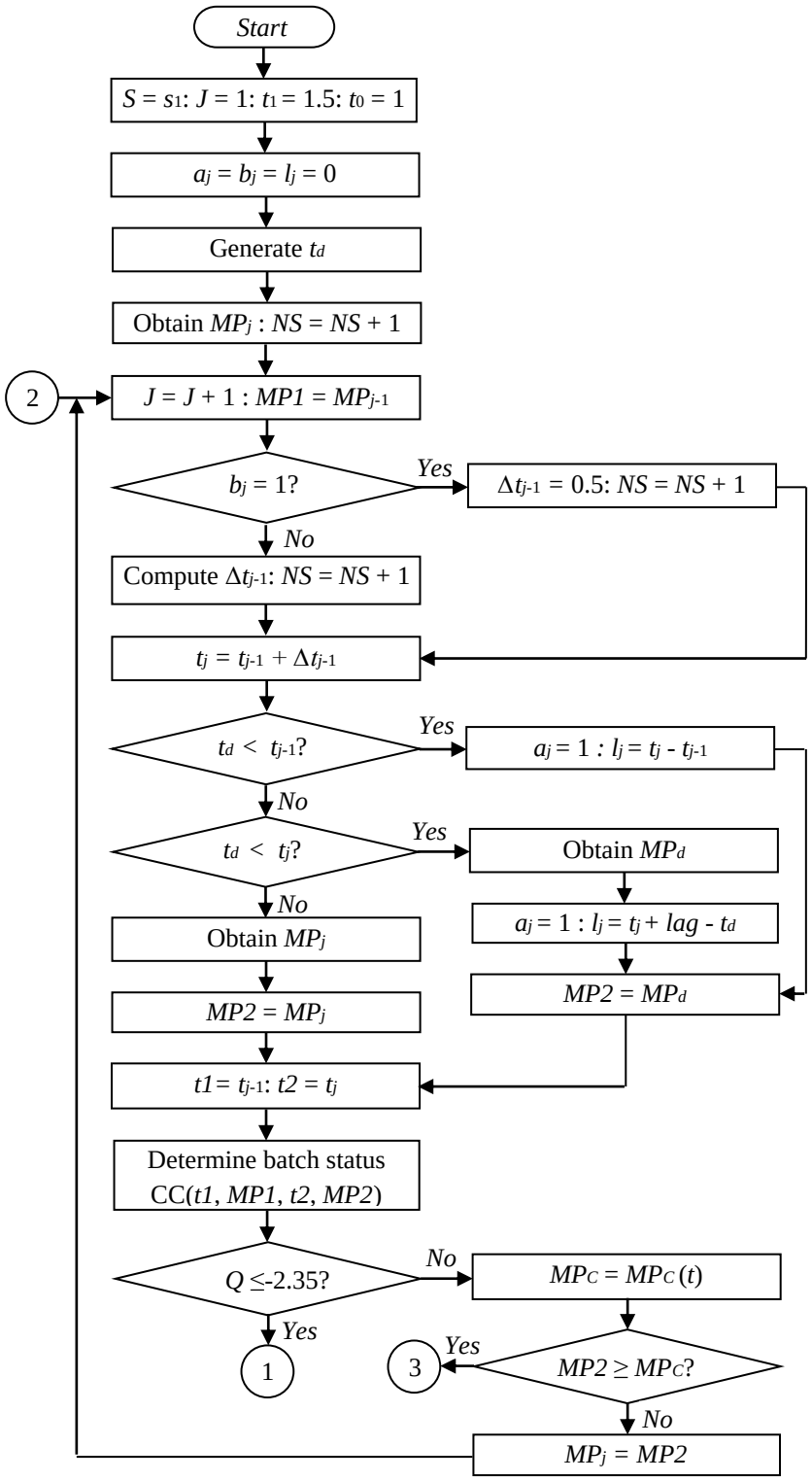
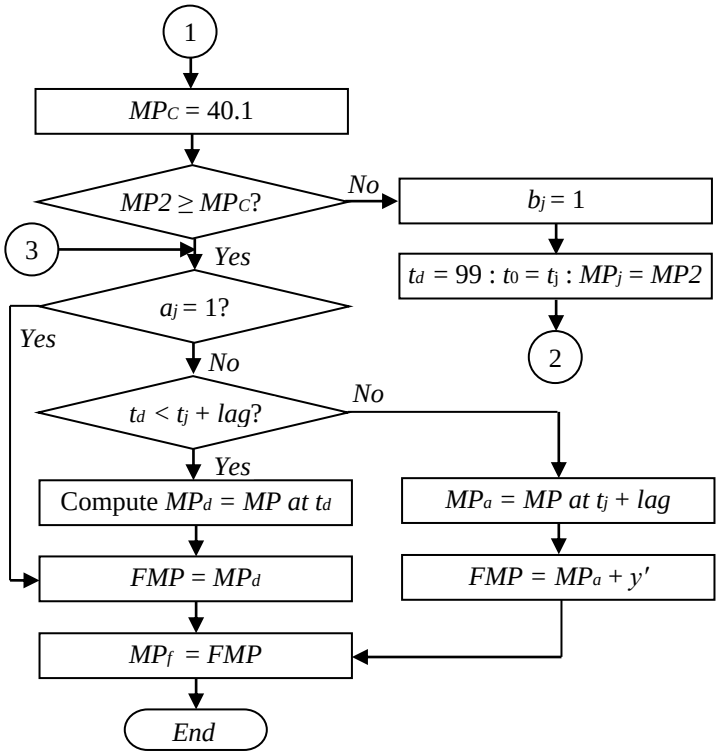


Figure 6.8. (..contd.)



Optimization of the parameters of P_1 and P_2

The parameter k_1 of P_1 is estimated in a straightforward manner. The average of $C(\omega)$ is computed based on 25000 samples for a series of values of k_1 with a step size of 0.05. The value of k_1, at which the average cost becomes the lowest, is the desired estimate.

However, the estimation of the parameters of P_2 is not such a simple task. The method of Experimental Regression Analysis proposed by Taguchi [140] is used for this purpose. In this method, the parameter space within which the optimal is expected to lie is systematically searched using a statistical design to locate the optimum. At the first iteration, the levels of the parameters chosen judiciously to cover the entire space that needs to be searched. In subsequent iterations the levels of the parameters are selected around the best levels obtained from the previous iteration. The process ends when there is no significant improvement in response. Here, the 3_3 design is used for the purpose of experimentation. In the first iteration, the total cost is evaluated based on 5000 samples, which is increased gradually to 25000 samples at the last (4_{th}) iteration.

6.4.5. Results and discussion

The main results of the study are the optimal cooling limits, optimal sampling intervals and the performance of the process under the optimal scheme. A summary of these results are

given below. In addition, two important aspects of the proposed integrated regulator are also discussed in the last two subsections.

6.4.5.1. Cooling limits

The cooling limits are obtained following the methodology described in Section 6.4.4.1. The comparison of costs for the two options when a batch is identified as active is shown in Table 6.2 for various values of *MP* and *t*. The corresponding cooling limits are also indicated in this table. A similar table is constructed for determining the cooling limits when a batch is identified as dead. The cooling limits thus obtained are summarized in Table 6.3. It is seen from this table that the cooling limit increases with the increase in *t*. This is to be expected since the variance of melting point decreases with the increase in *t* (see Equation (4.32)). The cooling limit for a dead batch is the highest since a dead batch will have no growth during the cooling phase.

Table 6.2. Comparison of costs for determining the cooling limits when the batch at MP_t is identified as active.

t	*Cost of cooling at MP_t*					*Cost of continuing hydrogenation*				
	MP					*MP*				
	39.6	*39.7*	*39.8*	*39.9*	*40.0*	*39.6*	*39.7*	*39.8*	*39.9*	*40.0*
2	**78.9**	65.9	57.8	54.6	56.2	79.2	74.5	72.4	73.2	76.7
3	76.6	**63.8**	55.9	52.9	54.8	68.9	65.3	64.4	66.3	70.9
4	87.9	72.4	**61.9**	56.4	56.0	74.3	68.9	66.4	66.7	69.9
5	121.7	98.5	80.7	**68.3**	61.2	92.0	84.2	79.3	77.3	78.2
6	153.1	122.8	98.2	**79.4**	66.2	106.5	97.5	91.3	88.1	87.9
7	172.8	138.1	109.3	**86.4**	69.4	114.9	105.4	98.9	95.2	94.6
8	184.2	147.0	115.7	**90.5**	71.2	119.6	110.0	103.2	99.4	98.6

*The melting points corresponding to the bold-faced numbers are the cooling limits

Table 6.3. Cooling limits

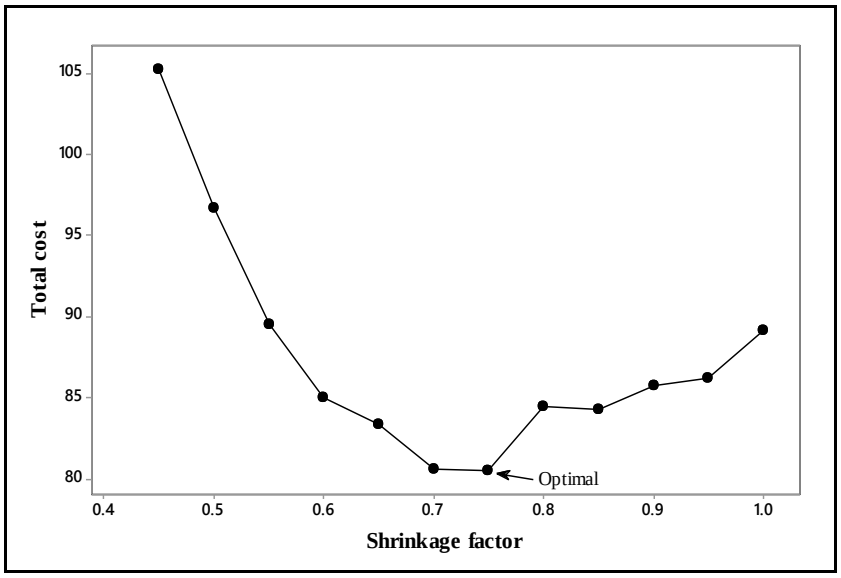
Active batch				*Dead batch*
$1.5 \le t \le 2.5$	$2.5 < t \le 3.5$	$3.5 < t \le 4.5$	$t > 4.5$	All *t*
39.6	39.7	39.8	39.9	40.1

6.4.5.2. Optimal policies

Figure 6.9 shows the plot of $E(C(\omega))$ versus k_1. It is obvious from the figure that the optimal value of k_1 is 0.75. The expected cost for $k_1 < 0.5$, which are much higher than that for $k_1 = 0.5$, are not shown in this figure.

The optimal policy P_2 turns out to be simpler than specified, since the optimal value of m_1 is found to be zero. The optimal shrinkage factor k_2 is given by $k_2 = 0.65 + 0.02(t - t_0)$.

Figure 6.9. Optimum value of the shrinkage factor k_1



6.4.5.3. Comparison of the performance of P_1, P_2 and the existing procedure

Table 6.4 summarizes the performance of the two policies (optimized) and the existing procedure. It is seen that there is no significant difference in performance between P_1 and P_2. However, considering all the three important performance measures namely total cost, cycle time and % detection of dead batches, the policy P_2 is recommended for implementation.

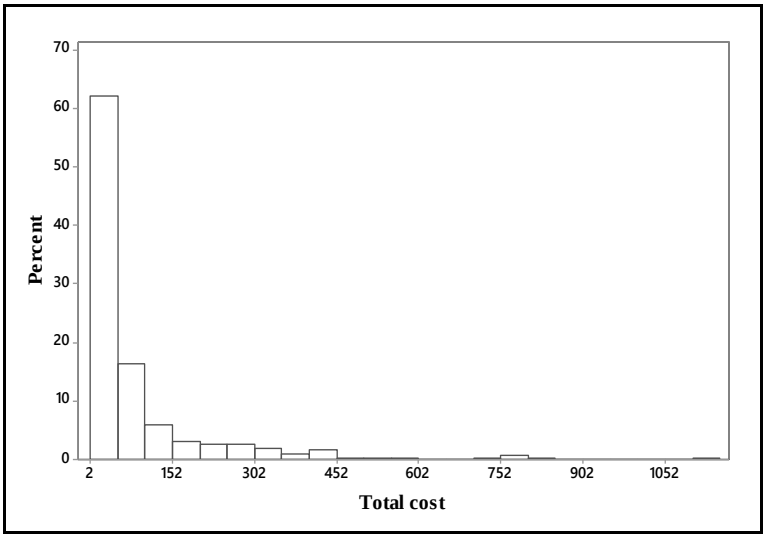
Table 6.4. Comparison of the performance of P_1, P_2 and the existing system

Item	*Average*		
	P_1	P_2	*Existing*
Total cost	80.54	79.03	128.95
Penalty cost	8.56	8.9	6.38
Sampling cost	4.34	4.34	3.6
Failure cost	15.47	14.57	20.29
Recovery cost	1.46	1.46	4.26
End cost	51.71	50.76	94.42
Mean (MP_f)	(40.49)	(40.48)	(40.71)
MSE (MP_f)	(0.148)	(0.145)	(0.27)
Cycle time*	(3.47)	(3.39)	(4.0)
% Detection	(71.76)	(75.51)	-

*From first t_0 to the beginning of cooling

Next, the performance of P_2 is compared with that of the existing system (E). As is to be expected, the proposed policy is seen to be far superior to the existing one. A comparison of the individual cost and other elements reveals the following. (i) The similarity of the penalty cost of P_2 and E suggests that the concern related to the workload of the testing laboratory has been addressed adequately. (ii) The sampling cost of P_2 is marginally higher than that of E. This is to be expected since an extra sample is taken at t = 1.5 hour for every batch. However, the additional information provided by this extra sample about the growth behavior of the process over a larger range of t, is likely to be very valuable in developing a better process model in future. (iii) Both the failure and recovery cost of the existing system is significantly higher than that of P_2. This also indicates the effectiveness of the control chart (compared to the existing naive approach) in detecting a dead batch. (iv) The main contribution of the proposed system is that the quality loss is expected to reduce by about 46%. However, even under P_2, the quality loss remains the main contributor (64.2%) to the total cost. Therefore, the future improvement effort should be directed towards reduction of this component. (v) The policy P_2 is also expected to reduce the cycle time by more than half an hour, which is very significant from the point of view of production cost. This reduction in cycle time is a combined effect of the lower failure time, less overshooting of the final melting point and the recovery of the extremely slow growing batches.

Figure 6.10. Histogram of total cost



The average values of the total cost and its components as reported in Table 6.4 however do not tell the full story. The distribution of total cost (Figure 6.10) is seen to be highly skewed (estimated skewness = 4.1). Moreover, the distribution is more or less bounded at the

lower end, which is nearly the same for both P_2 and the existing system. The lower bound for the existing system is 2, which corresponds to a minimum of two samples, no failure and the final melting point at target. The same for P_2 is 3 since an additional sample is taken at t = 1.5. So, the reduction of mean is achieved not by a shift of the distribution of total cost towards the left but by a favorable redistribution of the costs within the lower and a soft upper bound. This is shown in Table 6.5.

It is obvious from Figure 6.10 that even under P_2 the total cost varies a great deal from batch to batch. Broadly, the two main sources of this variation, which can be controlled, are (i) The variation in the growth path of an active batch and (ii) The process failures. It is expected that successful implementation of the proposed control charts will facilitate gradual reduction of the variation of total cost over time.

Table 6.5. Comparison of the distributions of total cost under P_2 and the existing system

Interval	*Relative frequency (%)*	
	P2	*Existing system*
2 - 82	72.55	46.81
82 - 162	11.59	25.53
162 - 242	6.22	12.77
242 - 322	4.84	6.38
322 - 402	2.53	2.13
> 402	2.26	6.38

6.4.5.4 Look-up table

In order to facilitate implementation of the proposed control scheme, a two-way look-up table giving the values of $\Delta t \mid MP_t$ is also prepared. An abridged version of the same is shown in Table 6.6.

Table 6.6. Look-up table for Δt in minutes (when the batch is identified as active*)

MP	*t*				
	1.5	2.5	3.5	4.5	6
32-38.5	90 for all *t*				
38.6	87	90	90	90	90
38.8	79	82	84	86	90
39.0	70	72	74	76	79
39.2	59	61	63	65	67
39.4	54	56	58	59	61
39.6	41	42	43	44	46
39.8-39.9	30 for all *t*				

*Use Δt = 30 for a recovered batch

6.4.5.5. Undetected dead batches

Table 6.4 shows that about 25% of the dead batches remain undetected at the time of cooling. Thus, one should expect the distribution of y' to be a mixture distribution as given by (6.2). However, it is also possible to have $y' = 0$ on account of the poor least count of measurement (= 0.1_0C). Since these two sources of variation cannot be separated, the use of the mixture distribution for generating y' (as has been done here) is a safer strategy. Note that if the value of y' for an active batch is generated from the Weibull distribution, then the optimal total cost will be lower than that shown in Table 6.4.

6.4.5.6. Integration of SPC and passive regulation

It has been noted in Section 1.5 that the integrated shift and drift control schemes that have been developed so far can broadly be classified into two groups - one in which the control chart is used for selecting the appropriate mode of drift control and the other in which the control chart is used for monitoring the drift controlled process. The system proposed here has features of both the above groups. Like the former, the control chart is used for selecting one of the two modes of control (the safe mode having smaller intervals and the normal mode), and at the same time it retains the classical flavour of SPC, where an out-of-control signal from the chart is followed by a diagnosis and recovery of the process to its in-control state. In the present case, however, the process diagnosis step is absent.

It is to be noted that the passive regulator and the shift controller (control chart) are integrated since the output of the former depends on the output of the latter and vice versa (length of the sampling interval determines the chance of detecting a dead batch). Of course, the integrated scheme has been optimized only locally, i.e., for a given set of control limits of the control chart. As already mentioned, this has been done to have the flexibility of using the control chart as a standalone tool.

6.4.5.7. Combined active and passive regulation

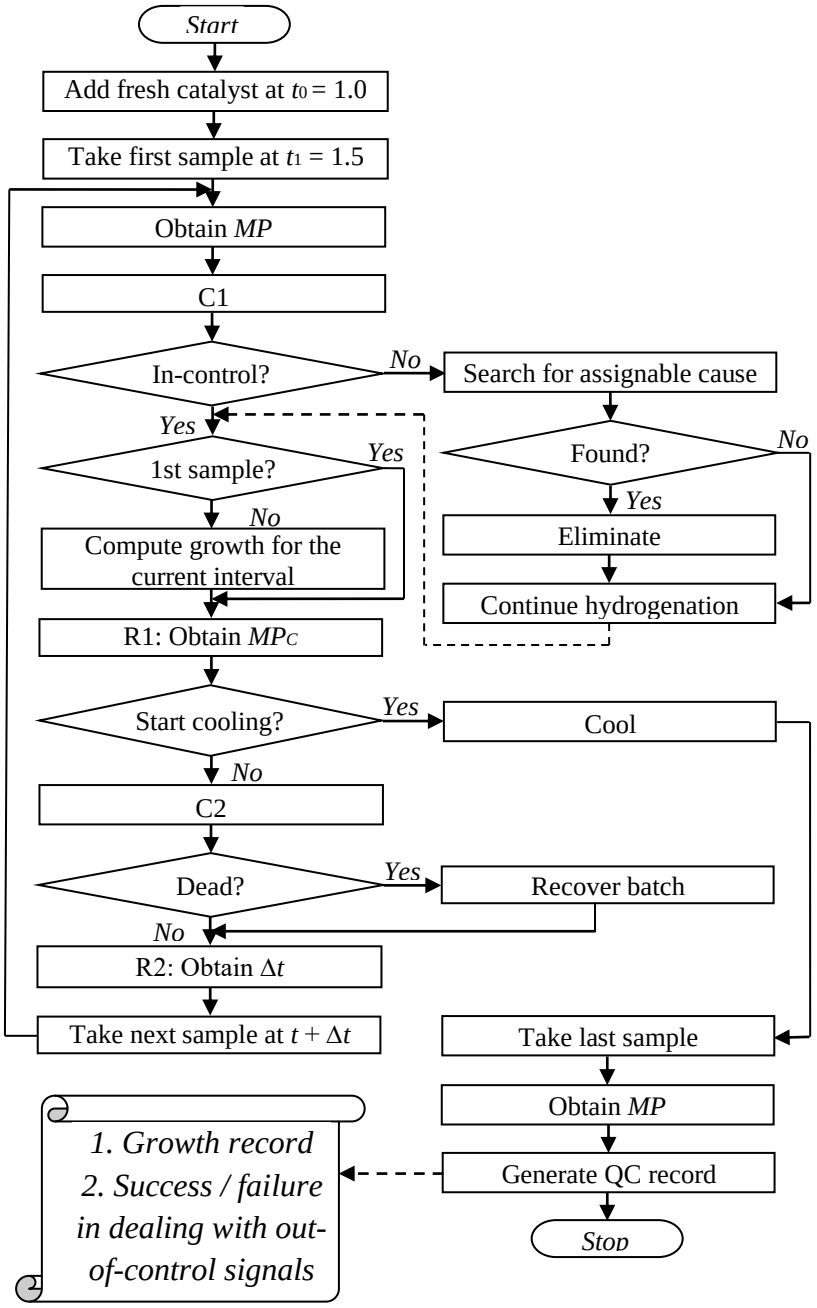
As noted in Section 6.1, a growth process is usually not subjected to active regulation. However, active regulation may be employed in a limited fashion to supplement the basic scheme of passive regulation. For example, the growth rate of the process can be deliberately slowed down towards the end to achieve better control over the end state of the process. Such an approach is commonly used in practice (e.g. manual cooking of rice). The proposed approach can easily be adopted to develop such a hybrid control scheme. However, this will

require data from a controlled experiment for developing an appropriate model for the plant and much more information to modify the cost function suitably.

6.4.6. Proposed control scheme

The integrated shift and drift control scheme that is proposed for controlling the process of hydrogenation is shown in Figure 6.11. In this figure, the control chart for monitoring an active batch is denoted by C1 and the other chart for detection of a dead batch is denoted by C2. The two components of the drift regulator, i.e. the one giving the cooling limits and the other determining the sampling intervals are denoted by R1 and R2 respectively.

Figure 6.11. Proposed integrated control scheme



6.5. Conclusion

The control of the end state of a monotonic growth process without regulating the growth path is defined as a problem of passive regulation. The most interesting problems of passive regulation are those that involve periodic sampling and when there is a significant lag in measurement. It is shown with the help of the hydrogenation example that the problem of passive regulation can be formulated as a SDP, which can be solved effectively following the method of approximate dynamic programming. The proposed approach is fairly general and can be used in many other similar situations. It can also be generalized to incorporate active regulation in a limited way.

It is surprising that the problem of passive regulation has not received any noticeable attention in the past despite its great practical significance. Of course, if an economic method for continuous on-line measurement becomes available then the process may be controlled much more effectively compared to the method of passive regulation as described here. But, in the real world, there will always be a gap between the need and the availability of suitable measurements due to techno commercial reasons. This methodology proposed in this chapter is a contribution towards objective management of this gap.

Chapter 7 Conclusion

In general, a growth process may be represented as $Y_t = f(t) + S_t$, where the random variable $Y_t > 0$ represents the value of the growth characteristic at time t and S_t is a mean stationary stochastic process. In this work, it is assumed that the deterministic growth function $f(t)$, $t \geq 0$ is continuous and monotonic. It is also assumed that the growth process has a beginning and an end so that the start and the end points of the process define a growth cycle. A collection of several such growth cycles defines the growth process.

The start and the end time of the growth cycles may either remain fixed over all the cycles or they may vary from cycle to cycle. At the production level, the most stringent requirement that may be specified for controlling a growth process is a reference growth trajectory for Y_t. However, such a specification for a growth process is rare.

Commonly the requirements for a growth process are specified in terms of an end state target T for Y_t, the cycle time and other quality requirements of the process output. In order to meet these specifications, it is necessary to perform the following three control tasks:

(i) Monitoring an active process

(ii) Detection of growth failure and

(iii) Controlling the end state of the process

An active process may be monitored in two ways depending on whether the control actions are to be taken within a cycle as the process evolves or at the end of a cycle. Since a growth curve can be considered as a special type of process profile, the approach of profile monitoring may be used for monitoring the between cycle variation, i.e. when the control actions are taken at the end of a growth cycle, provided the growth data are available in suitable format.

Depending on the nature of the process, the end state of the process may be controlled in two ways: (i) by fixing the end time (say t_e) and then taking suitable control actions at the beginning of the cycle so that the process remains close to T at time t_e or (ii) by terminating the process at the right time. Here, this latter approach of regulating a growth process has been called ‘passive regulation’. Depending on the nature of the data that may be obtained from the process, a growth process can be passively regulated in various ways. In this study the interest lies in passive regulation based on periodic sampling.

A review of the existing literature on growth process control suggests that if the possibility of using profile monitoring is excluded, then the past work in this area is indeed very limited. In all, we have been able to identify only three publications on monitoring an active process, one on detection of growth failure (with a few related works) and none on passive regulation based on periodic sampling. It is thus obvious that the growth process control has received very little attention in the past, despite its great practical significance. An attempt has been made here to give a comprehensive account of growth process control covering all the three control tasks as mentioned above.

The data on Y_t and other process variables for modelling a growth process may be obtained either in the form of a continuous stream through on-line sensors or by measuring the process at discrete time points. The discrete time points may either remain fixed over all the cycles or they may vary from cycle to cycle. The growth data used in this study is in the latter form, which may be represented as (y_{ij}, t_{ij}), where y_{ij} is the observed value of Y for the j_{th} growth cycle observed at the i_{th} instant (at time t_{ij}), $i = 1, 2, ..., n_j$, $j = 1, 2, ..., N$, where n_j is the number of observations corresponding to the j_{th} cycle and N is the total number of cycles. For the growth data used in this study $1 \leq n_j \leq 6$ and $N = 50$. Thus the growth data is in the form of a highly unbalanced panel data. Clearly, the approach of profile monitoring is not suitable for such a process. The available data also suggests that the approach of passive regulation based on periodic sampling needs to be adopted for controlling the end state of the process.

Throughout this work, it is assumed (and validated later) that the distribution of Y_t is approximately given by $N(\mu_t, \sigma_t)$, where μ_t and σ_t are the mean and standard deviation respectively of the process at time t. Since the process variance is assumed to vary over time, it is necessary to develop an appropriate variance function for the process. However, a review of the existing literature on variance function estimation reveals that the recommended methods based on individual observations (when repeat points are not available) do not perform well when the sample size is small. In particular, the least square estimates of the variances are found to be highly biased. It is shown that the said bias is not related to sample size. The bias persists even when the sample size is as large as 10000. So a new approach is proposed for estimation of the variance function based on SN ratio and adjustment of the OLS estimates of the parameters of the SN function. It is shown that the adjusted SN function gives more accurate estimates of the variances than those obtained directly from a similar estimate of the variance function. However, the accuracy of the estimates varies over different values of the slope coefficient θ_1 (Chapter 3).

Presently growth process modelling is a very well developed field with diverse applications. The areas of applications include the growth of human beings, foetus, tumour, forest, algae, crystals, silicon ingot, nanowire, population, wealth and learning. However, most of these studies conducted in the past are primarily concerned with process characterization. In contrast, the growth models are developed here for the purpose of process control.

Developing growth models for process control requires a somewhat different perspective from the usual approach of modelling for process characterization. This is for the following reasons. First, in case of modelling for process characterization, the focus is primarily on estimating the mean function (variance being treated as a nuisance parameter). However, for the purpose of process control, it is essential to have a good estimate of the process variance as well. Thus, estimation of both the mean and the variance function are equally important in process control. Secondly, a good estimate of process variance in the context of process control means that it should be consistent with the existing process knowledge. In other words, the process variance should be used as a measure of ignorance in maintaining Y_t exactly on its target T_t. In case of a growth process, this may include the ignorance about T_t. In our opinion, it does not make much practical sense to be in a situation where the estimate of process variance gets inflated every now and then and we cannot eliminate the causes or their inflationary effects on the output variance. In this work, this has been an overriding consideration in developing the models and the associated methods of process control. In fact, we have argued at several places in this thesis in favour of such a characterization of process variance.

Thirdly, in studies on process characterization one usually does not employ dynamic models unless forced to do so due to the limitations of data. However, a dynamic model may be a model of choice in case of process control. For example, if it is desired to detect location free sharp changes in a growth process then dynamic models may be preferred. The two control charts developed here for detection of growth failures are based on dynamic models.

Throughout this work, the growth of melting point during hydrogenation is used as the working example for illustrating the proposed methodology of integrated shift and drift control. It is shown that the growth of melting point during hydrogenation can be described by the well known monomolecular model. The model is derived based on theoretical consideration and validated using published data (Chapter 4). The growth data of this study as mentioned above, where each growth curve represents the process performance with

respect to a given batch of oil, also suggests that the monomolecular model is appropriate for describing the growth of melting point during hydrogenation. A linear approximation of the monomolecular model is also proposed for simplifying the task of modelling.

Among others, the following two population level growth models are developed for the process of hydrogenation.

Model A: An approximate linear model for a suitably chosen group of batches. The groups are formed heuristically using a graphical approach of clustering.

Model B: A random walk model with variable drift, which is the dynamic version of the monomolecular model.

In general, the development of a continuous time dynamic model (such as Model B) based on discrete time data when the underlying growth function is nonlinear is a difficult task. In the present case, an innovative instrumental variable method of estimation is used for this purpose (Chapter 4). Although the estimates obtained (of the amount of growth from a given state over a given duration) are not highly consistent, the proposed method does give a working model.

It is already noted that the approach of growth process monitoring based on between cycle variations is not appropriate for the process of hydrogenation. However, for the sake of completeness, the suitability of this approach for monitoring monomolecular growth is examined based on simulated data. It is found that a T^2 - chart based on only four observations at fixed time points and a simple *X*-chart of the area under the growth curve (*AUGC*) within a fixed time interval can be used effectively for this purpose.

Next, two simple *X*- charts are developed based on Model A and Model B for monitoring an active process and detection of growth failure respectively (Chapter 5). In both the cases the control statistic used is $Q = (q_t - \hat{q}_t)/s_t$, where q is a generic random variable and s_t is the estimate of standard deviation of q_t. We have used $q_t = MP_t\,(t + 0.2)$ and $q_t = y_{t+\Delta t} - y_t$ for monitoring an active process and detection of growth failure respectively. It is found that the proposed *X*-chart for detection of growth failure is much more effective and efficient compared to the existing method that relies on the personal judgement of the process supervisor.

Having developed the control chart for detection of growth failure, an economic passive regulator is also developed for controlling the end state of the process (Chapter 6). The control chart for detection of growth failure is made an integral part of this regulator. The problem of optimal control of the end state in the presence of growth failure is broken down

to the following two relatively simpler problems: (i) A single stage stochastic optimization problem to determine the process state at which further growth should be terminated and (ii) A stochastic dynamic program for finding the optimal sampling scheme for monitoring the process. The second problem is solved following a simulation based method of optimization that minimizes the total cost given by the sum of sampling, failure, recovery and end costs. A comparison of the performance of the existing process with that expected under the proposed economic regulator reveals that the total cost is expected to reduce by about 39% from the existing level. Besides, the average cycle time is also expected to reduce by more than half an hour.

Finally, the economic regulator obtained as above is integrated with the control chart for monitoring the active process to develop an integrated shift and drift control system for the hydrogenation process.

To conclude, it is noted that growth process control is still in its infancy. Thus, although an attempt is made here to give a comprehensive account of growth process control, it makes only a few scratches on the surface of a largely barren land.

7.1. Main contributions

The main contributions of this work are as follows:

I. A new approach is proposed for variance function estimation based on SN ratio. It is shown that the variances can be estimated more accurately from the proposed bias corrected SN function than those obtained directly from an estimate of the corresponding variance function.

II. It is shown that the growth of melting point during hydrogenation can be described adequately by the monomolecular model.

III. An innovative method of instrumental variable estimation is proposed for developing a continuous time dynamic model based on discrete time data. It is shown that the proposed method provides reasonably consistent estimates of growth (from a given state over a given duration) within the working range.

IV. The proposed control chart for detection of growth failure is shown to be much more effective and efficient compared to the existing procedure that relies on subjective judgment of the process supervisor.

V. The problem of optimal control of the end state of a growth process in the presence of growth failure when the process is monitored through periodic sampling is formulated as a SDP. A simulation-based methodology is proposed for solving the SDP.

7.2. Area/Direction of future research

Since growth process control is still in its infancy, the opportunities for future research in this area are huge. A few of the important issues are noted below.

(i) The fundamental issue of growth process monitoring is yet to be resolved. The issue is - how to define the in-control state of the process? An approach has been proposed here for this purpose and the same has been used for the hydrogenation process. However, many other practical situations need to be examined to judge its usefulness and improved upon, if necessary.

(ii) Another important issue is related to integrating a shift controller with a drift controller. There will be many situations in practice, where it will be beneficial to allow the shift controller to have some supervisory role over the drift controller. But how can the same be achieved optimally keeping the true spirit of shift control?

On the methodological front, the following issues need further investigation:

(i) The variable $Z_t = MP_t(t + a)$, where a is a constant has been used for obtaining a linear approximation of the monomolecular growth function. This variable has an interesting physical interpretation. Note that if the growth function is linear and the initial melting point is zero, then MP_t*t is proportional to the Area Under the Growth Curve ($AUGC$) till time t. The constant 'a' is an adjustment coefficient that takes care of the non-linearity of a growth function. This shows that $AUGC_t$ may also be used for monitoring the within cycle variation of a monomolecular growth process. However this needs further investigation and studies on comparison of the performance of Z_t and $AUGC_t$ will be useful, particularly when $AUGC_t$ is used for detection of growth failure.

(ii) Generalization of the proposed method of instrumental variable estimation is another area of further study.

(iii) In this study, dynamic growth models have been used only for detecting extreme shifts like growth failure. The possibilities of using dynamic models for detecting other types of shifts may also be explored in future.

(iv) The proposed approach of variance function estimation provides accurate estimates of the variances, but the level of accuracy varies over the slope coefficient θ_1. In future, variance function estimation may focus on this aspect of the problem.

(v) In many practical cases of growth process control, the growth data may be sparsely sampled and the missing observations cannot be assumed to be missing at random. Appropriate methods are needed for clustering such a set of growth curves for the purpose of defining the in-control state of a growth process.

Appendix A. Growth data for fifty consecutive batches

Batch #	Duration (Hrs)	Melting point ©	Fresh catalyst (Kg)
1	2.50	-	0.5
	3.75	35.5	-
	4.67	37.0	-
	5.50	38.0	-
	6.50	38.5	-
	7.25	38.5	-
	9.00	39.9	-
	11.25	39.9	-
2	1.00	-	1.0
	2.25	39.0	-
	3.25	40.0	-
	5.33	40.0	-
3	1.75	37.0	Nil
	3.25	38.5	-
	4.50	39.0	-
	5.58	39.5	-
	6.75	40.0	-
	8.50	40.2	-
4	1.00	-	1.0
	3.42	39.5	-
	4.92	39.5	-
	5.75	39.5	-
	7.25	39.5	1.0
	9.00	40.0	-
	10.25	40.0	-
5	1.00	-	1.0
	3.33	38.0	-
	5.00	40.0	-
	7.00	40.3	-
6	2.50	-	1.0
	4.42	34.0	-
	5.75	38.4	-
	7.25	39.0	-
	8.25	40.0	-
	10.00	41.0	-
7	1.00	-	1.0
	2.25	37.0	-
	3.75	39.0	-
	4.75	40.0	-
	6.25	40.5	-
8	1.25	-	1.0
	4.08	38.0	-
	5.42	39.5	-
	5.75	-	0.5
	6.92	40.0	-
	8.58	41.0	-
9	1.00	-	1.0
	2.50	39.5	-
	4.00	40.0	-
	5.25	40.0	-
10	2.17	-	1.0
	2.83	39.6	-
	3.58	39.6	-
	4.16	-	0.5
	5.00	41.0	-
	6.92	41.3	-
11	1.00	-	1.0
	3.00	39.0	-
	4.00	40.0	-
	6.00	40.6	-
12	1.25	-	1.0
	3.00	40.0	-
	5.00	40.4	-
13	1.00	-	1.0
	2.50	38.7	-
	3.50	39.5	-
	4.00	-	1.0
	5.00	40.0	-
	7.00	40.7	-
14	1.00	-	2.0
	2.50	37.0	-
	4.00	39.0	-
	4.50	-	1.0
	6.00	39.8	-
	8.50	40.5	-
	10.50	40.8	-
15	1.00	-	2.0
	2.50	37.0	-
	4.00	38.7	-
	5.50	-	1.0
	6.67	40.0	-
	7.92	41.0	-
16	1.00	-	2.0
	3.33	36.0	-
	5.00	39.0	-
	6.50	40.6	-
	8.50	40.8	-
17	1.00	-	1.0
	3.50	39.0	-
	4.50	39.5	-
	5.83	40.3	-
	7.33	40.4	-

Batch #	Duration (Hrs)	Melting point ©	Fresh catalyst (Kg)
18	2.50	-	2.0
	3.75	40.0	-
	5.75	40.8	-
19	0.75	-	1.0
	3.58	40.0	-
	6.08	40.8	-
20	4.92	-	2.0
	6.42	38.0	-
	7.42	39.0	-
	8.42	39.0	1.0
	9.84	39.3	-
	10.42	39.5	-
	11.75	40.3	-
	13.00	40.9	-
21	1.25	-	3.0
	2.50	35.0	-
	3.75	38.0	-
	4.75	40.0	-
	6.25	40.8	-
22	0.75	-	2.0
	2.67	39.3	-
	3.50	39.5	-
	4.67	39.8	-
	6.25	39.8	-
	8.00	40.0	-
	10.00	40.0	-
	11.58	40.7	-
23	1.00	-	2.0
	3.00	38.5	-
	4.50	40.0	-
	6.50	41.5	-
24	0.75	-	2.0
	3.50	40.0	-
	5.00	41.0	-
25	1.00	-	3.0
	3.50	39.5	-
	4.75	40.0	-
	6.50	40.4	-
26	1.75	-	3.0
	3.75	40.0	-
	5.50	40.4	-
27	0.75	-	3.0
	2.25	37.8	-
	3.92	40.6	-
	6.42	41.0	-
28	2.58	-	2.0
	4.08	40.2	-
	5.75	41.2	-
29	2.00	-	2.0
	4.00	37.5	-
	5.67	39.5	-
	7.00	40.0	-
	9.75	40.5	-

Batch #	Duration (Hrs)	Melting point ©	Fresh catalyst (Kg)
30	1.00	-	2.0
	2.00	38.0	-
	3.50	39.4	-
	4.33	39.5	-
	4.83	-	1.0
	5.75	40.2	-
	7.00	40.5	-
31	1.25	-	2.0
	3.00	38.5	-
	4.00	39.5	-
	5.42	39.5	-
	6.17	-	1.0
	6.89	40.4	-
	8.84	42.0	-
32	0.75	-	2.0
	3.50	39.0	-
	5.00	39.0	-
	5.50	-	1.0
	6.50	40.0	-
	7.75	41.0	-
33	1.00	-	2.5
	3.25	37.8	-
	4.25	39.0	-
	5.00	-	1.0
	6.00	40.0	-
	7.50	41.3	-
34	1.00	-	2.5
	3.58	35.0	-
	4.75	38.0	-
	5.67	39.0	-
	7.17	39.0	-
	8.50	-	1.0
	10.25	40.0	-
	11.50	40.5	-
35	1.50	-	2.5
	3.83	38.0	-
	6.00	40.4	-
	7.50	41.0	-
36	1.00	-	2.0
	2.25	37.0	-
	3.50	39.0	-
	4.50	39.5	-
	5.50	40.0	-
	7.50	39.8*	-
37	1.17	-	2.0
	3.17	40.7	-
	4.67	41.0	-
38	1.50	-	1.0
	3.25	40.0	-
	4.75	40.0	-
39	1.00	-	1.0
	1.92	37.7	-
	3.33	40.0	-
	4.83	40.5	-

Batch #	Duration (Hrs)	Melting point ©	Fresh catalyst (Kg)
40	0.75	-	1.0
	2.33	39.6	-
	3.17	40.3	-
	4.83	41.2	-
41	0.83	-	2.0
	3.42	40.1	-
	5.42	40.7	-
42	1.08	-	1.0
	2.58	37.0	-
	3.58	39.5	-
	5.00	40.0	-
	6.67	40.4	-
43	1.00	-	2.0
	3.08	38.7	-
	4.25	40.5	-
	5.75	40.8	-
44	1.00	-	2.0
	2.50	38.5	-
	4.00	40.5	-
	6.00	40.8	-
45	1.17	-	2.0
	2.00	38.0	-
	3.17	39.2	-

Batch #	Duration (Hrs)	Melting point ©	Fresh catalyst (Kg)
45	3.50	-	0.5
	4.42	40.3	-
	6.42	40.5	-
46	0.67	-	2.0
	2.83	40.0	-
	4.67	40.3	-
47	1.00	-	2.0
	2.67	40.4	-
	4.67	41.0	-
48	0.50	-	2.0
	2.50	39.6	-
	3.25	40.0	-
	5.00	41.3	-
49	0.83	-	2.0
	3.83	39.5	-
	6.33	41.2	-
	7.67	41.9	-
50	0.75	-	2.0
	3.25	40.4	-
	4.75	40.5	-
Note: The last two rows corresponding to each batch give the status of the batch at the start and end of cooling			

* Melting point decreases during cooling. Growth during cooling is assumed to be zero.

Appendix B. Growth data on hydrogenation of cotton seed oil

Temperature: 120C				*Temperature: 150C*				*Temperature: 170C*			
Time (Mts.)	*IV*	*MP (C)*	*Saturate 18:0 (%)*	*Time (Mts.)*	*IV*	*MP (C)*	*Saturate 18:0 (%)*	*Time (Mts.)*	*IV*	*MP (C)*	*Saturate 18:0 (%)*
9	108	21.5	3.42	4	109	22.0	3.11	2.5	109	20.0	2.81
16	103	24.5	3.73	8	102	26.5	3.50	5	101	22.5	2.75
24	99	27.5	4.91	13	95	28.0	3.60	8	96	27.0	3.08
32	93	31.5	5.79	17	89	31.5	4.23	11	88	31.0	3.15
40	85	32.5	7.28	21	83	33.0	4.34	13	84	33.5	3.30
48	80	37.5	8.66	25	78	35.5	5.39	15.5	76	34.5	3.68
57	72	40.0	10.50	29	74	37.0	6.36	19	71	36.5	4.60
67	70	41.0	12.42	38	67	38.0	7.35	22	67	37.5	5.85
76	63	43.5	15.00	42	61	41.0	11.49	27	62	40.0	9.08

Appendix C. Estimation of parameters of the monomolecular growth model for a group of batches

The general nonlinear regression model is given by $y = f(x;\theta) + \varepsilon$, where the errors ε are not necessarily independent (see Seber & Wild, [76]). In our case, we have

$$MP_{ij} = f(t_{ij};\theta) + e_{ij}; e_{ij} = \delta e_{(i-1)j} + r_{ij} \quad i = 1,2,...n_j, \quad j = 1,2,...n \quad , \qquad \text{(C.1)}$$

We may define $f' = MP_{eq} - (MP_{eq} - MP_{in})\exp(kt_{ij}) + \delta\varepsilon_{(i-1)j}$ as the nonlinear function of t_{ij} and $e_{(i-1)j}$, $\theta^* = (MP_{eq}, MP_{in}, k, \delta)$ as the parameter vector and r_{ij} as the uncorrelated errors distributed as $N(0, \sigma_{ij}^2), \sigma_{ij}^2 = \sigma^2 / w_{ij}$. So $Var(w_{ij}^{1/2} r_{ij}) = \sigma^2$. Therefore θ^* is estimated by minimizing $S = \sum_{i=1}^{n_j} \sum_{j=1}^{n} w_{ij} r_{ij}^2$. This is known as weighted least square regression, where the weights are $w_{ij} = 1/\sigma_{ij}^2$. Since σ_{ij}^2 are unknown, we shall replace these by their estimates $s_{ij}^2 = s_i^2$ (assumed). The residuals r_{ij} are obtained from (C.1) by setting $\varepsilon_{0j} = 0, \forall j$. The sum *S* is minimized following the Newton-Raphson iterative equation

$$\theta^{*(i+1)} = \theta^{*i} - H^{-1}(S)G(S), \qquad \text{(C.2)}$$

where *H* is the hessian matrix and *G* is the gradient vector. The initial value of θ^{*0} needs to be provided in order to begin the above iteration. Our experience with the monomolecular model suggests that if data are available covering the entire period of growth then the procedure converges for any reasonable choice of θ^{*0}. However, if data near *t* = 0 are not available then the choice of θ^{*0} becomes critical and the same needs to be found by trial and error. It is difficult to compute the exact confidence intervals for the parameters of a non-linear model such as above. Accordingly, as is the standard practice, the approximate confidence intervals are obtained as follows. Let $C = F'_{p\times m} F_{m\times p}$, $F(\theta^*) = \left[\left(\partial f'(t,\varepsilon;\theta^*) / \partial \theta_r^*\right)\right]$ where $r = 1, 2 \ldots p$ (the number of parameters) and $m = \sum_{j=1}^{n} n_j$ (*n* batches with the j_{th} batch having n_j observations) is the sample size. Then, if *m* is large and under certain regularity conditions, we have (Seber & Wild, [76])

$$a'\hat{\theta}^* \sim N(a'\theta^*, \sigma^2 a' C^{-1} a), \qquad \text{(C.3)}$$

where *a* is an arbitrary column vector of size *p* and $\hat{\theta}^*$ is the convergent estimate of θ^*. Thus, $T = [a'\hat{\theta}^* - a'\theta^*] / [s(a'C^{-1}a)^{1/2}]$ will follow the *t* distribution with degrees of freedom (*m* – *p*),

where $s^2 = S/(m - p)$ is the unbiased estimate of σ^2. Consequently, an approximate $100(1 - \alpha)\%$ confidence interval of $a'\theta^*$ is given by $a'\theta^* \pm t_{\alpha/2,m-p} s(a'C^{-1}a)^{1/2}$.

The confidence intervals of the individual parameters θ_r^* are then obtained by setting $a' = (0,0,...,1,0,...)$, where the r_{th} element is one and the rest are zero. However, in order to use the above, we need an estimate of *C*. This is obtained as $\hat{C} = F'(\hat{\theta}^*)F(\hat{\theta}^*)$.

The computations as above are performed using an R-program, which has been specially developed for this work.

References

[1] W. A. Shewhart. *Economic Control of Quality of Manufactured Product*. D. Van Nostrand, New York, 1931.

[2] H. F. Dodge and H. G. Romig. "A method of sampling inspection". *The Bell System Technical Journal*, *8*(1929) 613-631.

[3] T. Olejniczak. "Japanese Management: 50 years of Evolution of the Concept". *Acta Asiatica Varsoviensia*, *26*(2013) 7-25.

[4] A. V. Feigenbaum. *Total Quality Control*. McGraw-Hill, New York, 1991. (First edition in 1961)

[5] W. E. Deming. *Out of The Crisis*. MIT Press, Cambridge, 2000.

[6] J. M. Juran (Ed.). *Juran's Quality Control Handbook*. McGraw-Hill, New York, 1988. (First edition in 1951)

[7] K. Ishikawa. *Guide to Quality Control*. Asian Productivity Organization, Tokyo, 1986.

[8] G. Taguchi, Y. Wu. *Off-line Quality Control*. Japanese Standards Association. Japan, 1977.

[9] M. Harry, R. Schroeder. *Six Sigma*. Doubleday, New York, 2000.

[10] G. E. P. Box, G. M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, California, 1970.

[11] K. J. Åström. *Introduction to Stochastic Control Theory*. Academic Press, California, 1970.

[12] G. E. P. Box, T. Kramer. "Statistical process monitoring and feedback adjustment: a discussion". *Technometrics*, *34*(1992) 251-267.

[13] D. C. Montgomery, J. B. Keats, G. C. Runger, W. S. Messina. “Integrating statistical process control and engineering process control”. *Journal of Quality Technology*, *26*(1994) 79-87.

[14] E. D. Castillo. *Statistical Process Adjustment for Quality Control*. John Wiley & Sons, New York, 2002.

[15] S. A. Vander Weil, W. T. Tucker, F. W. Faltin, N. Doganaksoy. “Algorithmic statistical process control: concepts and an application”. *Technometrics*, *34*(1992) 286-297.

[16] G. E. P. Box, G. M. Jenkins. “Further contributions to adaptive quality control: Simultaneous estimation of dynamics: no-zero costs”. *ISI Bulletin*, 34th session, Ottawa, Canada, (1963)943-974.

[17] G. Taguchi. *On-line Quality Control During Production*. Japanese Standards Association, Japan, 1981

[18] G. A. Barnard. “Control charts and stochastic processes”. *Journal of Royal Statistical Society, Ser. B, 21*(1959) 239-271.

[19] P. K. Mozumder, S. Saxena, D. J. Collins. “A monitor wafer based controllers for semiconductor processes”. *IEEE Transactions on Semiconductor Manufacturing*, *7*(1994) 400-411.

[20] E. Sachs, A. Hu, A. Ingolfsson. “Run-by-run process control: combining SPC and feedback control”. *IEEE Transactions on Semiconductor Manufacturing*, *8*(1995) 26-43.

[21] M. Park, J. Kim, M. K. Jeong, A. M. S. Hamouda, K. N. Al-Khalifa, E. A. Elsayed. “Economic cost models of integrated APC controlled SPC charts”. *International Journal of Production Research, 50*(2012) 3936-3955.

[22] J. Aerts, V. Buggenhout, E. Vranken, M. Lippens, J. Buyse, E. Decuypere, D. Berckmans. “Active control of the growth trajectory of broiler chickens based on online animal responses”. *Poultry Science, 82*(2003) 1853-1862.

[23] L. Niu, D. Yang. “Dynamic simulation and optimization for batch reactor control profiles”. *International Proceedings of Computer Science and Information Technology (IPCSIT)*, *22*(2012) 58-68. Available at http://www.ipcsit.com

[24] H. Shore, D. Benson-Karhi, M. Malamud, A. Bashiri. “Customized fetal growth modelling and monitoring - a statistical process control approach”. *Quality Engineering, 26*(2014) 290-310.

[25] W. H. Woodall. “Current research on profile monitoring”. *Production*, *17*(2007) 420-425. [on-line] http://dx.doi.org/10.1590/S0 103-65132007000300002.

[26] L. Zhu, C. Dai, H. Sun, W. Li, R. Jin, K. Wang. “Curve monitoring for a single crystal ingot growth process”. [on-line] doi:*10.2991/iemi-14.2014.36*, 2014.

[27] W. H. Woodall, D. J. Spitzner, D. C. Montgomery, S. Gupta. "Using control charts to monitor process and product quality profiles". *Journal of Quality Technology, 36*(2004) 309-

[28] F. K. Grote, P van Dommelen, W. Oostdijk, S. M. P. F. de Muinck Keizer-Schrama, P. H. Verkerk, J. M. Wit, S van Buuren. "Developing evidence-based guidelines for referral for short stature". *Arch. Dis. Child., 93*(2008) 212-217.

[29] H. A. Wahab, M. Y. Noordin, S. Izman, D. Kurniawan. "Quantitative analysis of electroplated nickel coating on hard metal". *The Scientific World Journal*, (2013). Available at http://dx.doi.org/10.1155/2013/631936.

[30] A. Mahapatro, S. K. Suggu. "Modelling and simulation of electro deposition: Effect of electrolyte content density and conductivity on electroplating thickness". *Advanced Materials Science, 3*(2018) 1-9.

[31] H-T. Noh, D-I. Kim, S-S. Han, "Real time end point detection in plasma etching using real-time decision making algorithm". China Semiconductor Technology International Conference, 15-16 march (2015), Shanghai. *DOI: 10.1109/CSTIC.1015.7153380*.

[32] W. J. Spillman, E. Lang. *The Law of Diminishing Returns*. World Book Company, New York, 1924.

[33] J. France, J. Dijkstra, MS. Dhanoa. "Growth functions and their application in animal science". *Ann. Zootech., 45 Suppl.*(1996) 165-174.

[34] R. J. Carroll, D. Ruppert. *Transformation and Weighting in Regression*. Chapman & Hall / CRC, New York, 1988.

[35] C. Scrimgeour. "Chemistry of fatty acid". In: F. Shahidi (Ed.), *Bailey's Industrial Oil and Fat Products*, Sixth edition, John Wiley, New York, 2005.

[36] L. Horiuti, M. Polanyi. "Exchange reactions of hydrogen on metallic catalysts". *Transactions of Faraday Society, 30*(1934) 1164-1172.

[37] W. Normann. "Process for converting unsaturated fatty acids or their glycerides into saturated compounds". British patent 1515, 1903.

[38] A. C. Shoemaker, K-L. Tsui, C. F. J. Wu. "Economical experimentation methods for robust design". *Technometrics, 33*(1991)415-427.

[39] S. Chakraborty, P. Mandal. "SPC based on growth models for monitoring the process of hydrogenation of edible oil". *Journal of Food Engineering, 146*(2015) 192-203

[40] N. R. Draper, H. Smith. *Applied Regression Analysis,* 3rd edition. John Wiley & Sons, NJ, 1998.

[41] D. C. Montgomery, E. A. Peck, G. G. Vining. *Introduction to Linear Regression Analysis*. 5th edition. John Wiley & Sons, NJ, 2012.

[42] T. P. Ryan. *Modern Regression Methods*. John Wiley & Sons, NJ, 1997.

[43] M. Davidian, R. J. Carroll. "Variance function estimation". *Journal of American Statistical Association, 82*(1987) 1079-1091.

[44] G. Taguchi, S. Chowdhury, Y. Wu. *Taguchi's Quality Engineering Handbook*, John Wiley & Sons, NJ, 2005.

[45] G. Chen, R. A. Lockhart, M. A. Stephens. “Box-Cox transformation in linear models: large sample theory and tests of normality”. *The Canadian Journal of Statistics*, *30*(2002) 177-234.

[46] P. Mandal. “Signal-to-noise ratio: a fundamental and broad process performance measure”. *Journal of Engineering Design*, *23*(2012) 927-944.

[47] A. C. Harvey.“Estimating regression models with multiplicative heteroscedasticity”. *Econometrica*, *44*(1976) 461-465.

[48] A. J. Ferrer, R. Romero. “Small samples estimation of dispersion effects from unreplicated data”. *Communications in Statistics– Simulation*, *22*(1993) 975-995.

[49] N. N. Shah. *Fitting a Calibration Curve for DLA - A Modelling Experience*. Project report, SQC & OR Unit (Baroda), Indian Statistical Institute, Kolkata, India, 1986.

[50] S. Weisberg. *Applied Linear Regression*, 3rd edition. J. Wiley & Sons, NJ, 2005.

[51] H. Weisberg, E. Beier, H. Brody, R. Patton, K. Roychaudhuri, H. Takeda, R. Thern, R. Van Berg. “s-dependence of proton fragmentation by hardons. II. Incident laboratory momenta 30-250 GeV/c". *Physical Review, D, 17*(1978) 2875-2887.

[52] G. E. P. Box, R. D. Meyer. “Dispersion effects from factorial designs”. *Technometrics*, *28*(1986), 19-27.

[53] K. E. Castellano, A. D. Ho. *A Practitioner's Guide to Growth Models*, Council of Chief State School Affairs, 2013.

Available at. http://scholar.harvard.edu/files/.../a_practitioners_guide_to_growth_models.pdf.

[54] H. Poorter, N. P. R. Anten, L. F. M. Marcelis. "Physiological mechanisms in plant growth models: do we need a supra-cellular systems biology approach". *Plant, Cell & Environment*, *36*(2013) 1673-1690.

[55] J. L. Maino, M. R. Kearney. "Testing mechanistic models of growth in insects". Proceedings of Royal Society, B, *282*(2015).

[56] J. K. Vanclay. *Modelling Forest Growth and Yield: Application to Mixed Tropical Forests*. Cab International, Wellington, UK, 1994.

[57] D. Neal. *Introduction to Population Biology*. Cambridge University Press, UK, 2004.

[58] P. M. Martins, F. Rocha. “New developments on size dependent growth applied to the crystallization of sucrose”. *Surface Science*, *601*(2007) 5466-5472.

[59] P. L. Bonate, D. L. Howard. “Modelling tumor growth in oncology”. In: P. L. Bonate, D. L. Howard (Eds.), *Pharmakokinetics in Drug Development*, Springer, Berlin, 2011.

[60] H. Qiang, L. Wang, T. Dasgupta, L. Zhou, P. K. Sekhar, S. Bhansali, Y. An. “Estimation for silica nanowire growth catalyzed by Pd thin films”. *IEEE Transactions on Automation Science and Engineering, 8*(2011) 303-310.

[61] A. Wood. *Software Reliability Growth Models*, Technical Report 96.1, TANDEM, 1996. Available at: http://www.hpl.hp.com/techreports/tandem/TR-96.1.pdf.

[62] Ł. Piętak. "Review of theories and models of economic growth". *Comparative Economic Research, 17*(2014) 45-60.

Available at: http://dx.doi.org/10.1098/rspb.2015.1973.

[63] Fssai. Using models to control pathogen growth, 2017.

Available at: http://www.fssai.gov.in/jcr../Case_Studies_Risk_Assessment_09_01_2017.pdf.

[64] W. Chen, B. Zhou, J. Ding, Y. Yu, H. Dong, G. Zhong, X. Huang. "Growth process improvement for casting high-performance multi-crystalline silicon ingots for solar cells". *Journal of Materials Science and Engineering, B 6*(2016) 201-210.

[65] F. Yemiscioglu, M. Bayaz, A. S. Gumuskesen. "Effect of process conditions on physical and chemical properties of hydrogenated fats: monitoring of cottonseed oil hydrogenation process". *GIDA/The Journal of Food Technology,35*(2010) 13-19.

[66] L. F. Albright, M. M. Win, J. M. Woods. "Hydrogenation of cottonseed oil with reused catalyst". *The Journal of American Oil Chemical Society, 42*(1965) 556-560.

[67] M. A. Tike, V. V. Mahajani. "Kinetics of hydrogenation of palm stearic fatty acid over Ru/Al_2O_3 catalyst in presence of small quantity of water". *Indian Journal of Chemical Technology, 14*(2007) 52-63.

[68] F. J. Richards. "A flexible growth function for empirical use". *Journal of Experimental Botany, 10*(1959) 290-301.

[68a] J. MacQueen. "Some methods for classification and analysis of multivariate observations". *Proc. Fifth Barkley Symposium on Math. Satist. and Prob., 1*(1967) 281-297.

[68b] S. P. Loyd. "Least squares quantization in PCM". *IEEE Trans. on Inform. Theory, 28*(1982) 129-137.

[68c] J. Mao, A. K. Jain. "A self-organizing network for hyperellipsoidal clustering (HEC)". *IEEE Trans. on Neural Networks, 7*(1996) 16-29.

[68d] Z. Huang. "Extension to the k-means algorithm for clustering large data sets with categorical values". *Data Mining and Knowledge Discovery, 2*(1998) 283-304.

[68e] R. J. Gil-Garcia, J. M. Bodia-Contelles, A. Pons-Porrata. "A general framework for agglomerative hierarchical clustering". *In*: The Eighteenth Int. Conf. on Pattern Recognition, IEEE, 2006, p. 1-4.

[68f] M. Roux. "A comparative study of divisive and agglomerative hierarchical clustering algorithms". *Journal of Classification, 35*(2018) 345-366.

[69] B. C. Heggeseth. *Longitudinal cluster analysis with applications to growth trajectories*. Ph. D Thesis, University of California, Berkeley, 2013.
Available at: https//:escholarship.org/uc/item/10p932sm.

[69a] I. Epifanio, N. Ventura-Campos. "Functional data analysis in shape analysis". *Computational Statistics and Data Analysis, 55*(9) 2758-2773.

[69b] I. Epifanio. "Shape descriptors for classification of functional data". *Technometrics, 50*(2008) 284-294.

[70] J. Jacques, C. Preda. "Functional data clustering: a survey". In: *Advances in Data Analysis and Classification*, Springer Verlag, 8 (3), 2014, pp.24.

[71] R. J. A. Little, D. B. Rubin. *Statistical Analysis with Missing Data*, 2nd edition. J. Wiley & Sons., NJ, 2002.

[71a] D. Liebl, S. Rameseder. "Partially observed functional data: the case of systematically missing parts". *Computational Statistics and Data Analysis, 131*(2019) 104-115.

[71b] G. M. James, C. A. Sugar. "Clustering for sparsely sampled functional data". *Journal of the American Statistical Association, 98*(2003) 397-408.

[72] G. C. White, J. T. Ratti. "Estimation and testing of parameters in Richards growth model for western grebes". *Growth, 41*(1977) 315-323.

[73] S. L. Beal. "Population pharmacokinetic data and parameter estimation based on their first two statistical moments". *Drug Metabolism Review, 15*(1984) 173-193.

[74] L. B. Sheiner, T. H. Grasela. "Experience with NONMEM: analysis of routine phenytoin clinical pharmacokinetic data". *Drug Metabolism Review, 15*(1984) 293-303.

[75] R. H. Tones, F. Boadi-Boateng. "Unequally spaced longitudinal data with AR(1) serial correlation". *Biometrics, 47*(1991) 161-175.

[76] G. A. F. Seber, C. J. Wild. *Nonlinear Regression*. John Wiley, New York, 1989.

[77] T. W. Anderson, C. Hsiao. "Formulation and estimation of dynamic models using panel data". *Journal of Econometrics, 18*(1982) 47-82.

[78] S. Bigi, T. Söderström, B. Carlsson. “An IV-scheme for estimating continuous-time stochastic models from discrete-time data”, *IFAC Proceedings Volumes, 27*(1994) 1561-1566.

[79] Z. G. Stoumbos, M. R. Reynolds Jr., T. P. Ryan, W. H. Woodall. “The state of statistical process control as we proceed into the 21st century”. *Journal of American Statistical Asso*ciation, *95*(2000) 992-998.

[80] D. C. Montgomery. *Introduction to Statistical Quality Control*. John Wiley & Sons, New York, 1996.

[81] H. Hotelling. “Multivariate quality control”. In: C. Eisenhart, M. W. Hastay, W. A. Wallis (Eds.), *Techniques of Statistical Analysis*, McGraw-Hill, New York, 1947.

[82] W. H. Woodall, M. M. Ncube. “Multivariate CUSUM quality-control procedures”. *Technometrics*, *27*(1985) 285-292.

[83] C. A. Lowry, W. H. Woodall, C. W. Champ, S. E. Rigdon. “A multivariate exponentially weighted moving average control chart”. *Technometrics*, *34*(1992) 46-53.

[84] R. L. Mason, D. N. Tracy, C. J. Young. “A practical approach for interpreting multivariate T_2 control chart signals”. *Journal of Quality Technology*, *29*(1997) 396-406.

[85] P. M. Berthouex, W. G. Hunter, L. Pallesen. “Monitoring sewage treatment plants: some quality control aspects”. *Journal of Quality Technology*, *10*(1978) 139-149.

[86] A. V. Vasilopoulos, A. P. Stamboulis. “Modification of control chart limits in the presence of data correlation”. *Journal of Quality Technology, 10*(1978) 20-30.

[87] T. J. Harris, W. H. Ross. “Statistical process control procedures for correlated observations”. *The Canadian Journal of Chemical Engineering, 69*(1991) 48-57.

[88] D. G. Wardell, H. Moskowitz, R. D. Plante. “Run length distribution of special cause control charts for correlated processes”. *Technometrics*, *36*(1994) 3-17.

[89] D. C. Montgomery, C. M. Mastrangelo. “Some statistical process control methods for autocorrelated data”. *Journal of Quality Technology, 23*(1991) 179-204.

[90] G. Zhang. “A new type of quality control chart - the cause selecting control charts and a theory of diagnosis with control charts”. In: Proceedings of World Quality Congress, 1984, p. 175-185.

[91] M. R. Wade, W. H. Woodall. “A review and analysis of cause selecting control charts”. *Journal of Quality Technology*, *25*(1993) 161-169.

[92] L. Kang, S. L. Albin. “Online monitoring when the process yields a linear profile”. *Journal of Quality Technology*, *32*(2000) 418-426.

[93] F. S. Stover, R. V. Brill. “Statistical quality control applied to ion chromatography calibrations”. *Journal of Chromatography* A, *804*(1998) 37-43.

[94] K. Kim, M. A. Mahmoud, W. H. Woodall. “On the monitoring of linear profiles”. *Journal of Quality Technology*, *35*(2003) 317-328.

[95] M. A. Mahmoud, W. H. Woodall. “Phase I analysis of linear profiles with calibration applications”. *Technometrics*, *46*(2004) 380-391.

[96] J. D. Williams, W. H. Woodall, J. B. Birch. “Statistical monitoring of nonlinear product and process quality profiles”. *Quality and Reliability Engineering International*, *23*(2007) 925-941.

[97] R. Noorossana, A. Amiri, P. Soleimani. “On the monitoring of autocorrelated linear profiles”. *Communications in Statistics – Theory and Methods*. *37*(2008) 425-442.

[98] A.Vaghefi, S. D. Tajbakhsh, R. Norrosona. “Phase II monitoring of nonlinear profiles”. *Communications in Statistics – Theory and Methods*, *38*(2009) 1834-1851.

[99] C. Zou, Y. Zhang, Z. Wang. “A control chart based on a change point model for monitoring linear profiles”. *IIE Transactions*, *38*(2006) 1093-1103.

[100] M. A. Mahmoud, P. A. Parker, W. H. Woodall, D. M. Hawkins. “A change point method for linear profile data”. *Quality and Reliability Engineering International*, *23*(2007) 247-268.

[101] J. Jin, J. Shi. “Feature-preserving data compression of stamping tonnage information using wavelets”. *Technometrics*, *41*(1999) 327-339.

[102] J. Lee, Y. Hur, S. H. Kim, J. R. Wilson. “Monitoring nonlinear profiles using a wavelet-based distribution-free CUSUM chart”. *International Journal of Production Research*, *50*(2012) 6574-6594.

[103] S. I. Chang, S, Yadama. “Statistical process control for monitoring non-linear profiles using wavelet filtering and B-spline approximation”. *International Journal of Production Research*, *48*(2010) 1049-1068.

[104] O. Mestek, J. Pavlik, M. Suchánek. “Multivariate control charts: Control charts for calibration curves”. *Fresenius’ Journal of Analytical Chemistry*, *350*(1994) 340-351.

[105] Y. Zang, P. Qiu. “Phase II monitoring of free-form surfaces: An application to 3D printing”, *Journal of Quality Technology*, *50* (2018) 379-390.

[106] J. Thor, J. Lundberg, J. Ask, J. Olsson, C. Carli, P. Härenstam, and M. Brommels. “Application of statistical process control in healthcare improvement: systematic review”, *Quality and Safety in Healthcare*, *16*(2007) 387-399.

[107] L. J. Shu, F. Tsung, K. C. Kapur. “Design of multiple cause-selecting charts for multistage process with model uncertainty”, *Quality Engineering*, *16*(2004) 437-450.

[108] Y. Shang, F. Tsung, C. Zou. “Statistical process control for multistage processes with binary outputs”, *IIE Transactions*, 45(2013) 1008-1023.

[109] R. G. Mostafa, K. Paynabar. “Change detection in a dynamic stream of attributed networks”, *Journal of Quality Technology*, 50(2018) 418-430.

[110] C. Dai, K. Wang, R. Jin. “Monitoring profile trajectories with dynamic time warping alignment”. *Quality and Reliability Engineering International*, *30*(2014) 815-827.

[111] G. Nedumaran, J. J. Pignatiello Jr. “On constructing T^2control chart for on-line process monitoring”. *IIE Transactions*, *31*(1999) 529-536.

[112] J. Pan, S. Chen. “Determining optimal number of samples for constructing multivariate control charts”. *Communications in Statistics – Simulation and Computation*, *40*(2011) 216-228.

[112a] W. Y. Lin. "An overview of the performance of four alternatives to Hotelling *T* square". Educational Research Journal, 7(1992) 110-114.

[113] F. Duan, S. Simeone, R. Wu, J. Grady, I. Mandoiu, P. K. Srivastava. “Area under the curve as a tool to measure kinetics of tumour growth in experimental animals”. *Journal of Immunological Methods*, *382*(2012) 224-228.

[114] B. Nuryanto. “Relationship between area under disease progress curve and disease severity of sheath blight with rice field”. In: F. Kasim, A. Widjono, Sumano and Supriyano (Eds.). *Rice Industry, Culture and Environment*, Book 2, Indonesian Centre for Rice Research, Sukamandi, Indonesia, (2007) 545-548.

[115] OECD. *OECD Guidelines for Testing Chemicals - Alga Growth Inhibition Test.* Available at http://www.oecd.org/chemicalsafety/risk-assessment/1948257.pdf. (Accessed on 30 Dec. 2013), 1984.

[116] J. Serroyen, G. Molenberghs, G. Verbeke, M. Davidian. "Non-linear models for longitudinal data". *The American Statistician*, *63*(2009) 378-388.

[117] R. D. O’Brien. *Fats and Oils: Formulating and Processing for Applications*. Taylor and Francis Group, New York, 2009

[118] C. Rangxuan. “Hydrogenation characteristics of high erucic rapeseed oil”. *Lipid Science*, *97*(1995) 142-145.

[119] Website of IEEE Control System Society. Available at: www.ieeecss.org

[120] R. Bellman. *The Theory of Dynamic Programming*. Report P-550, The Rand Corporation, California, 1954.

[121] A. J. Duncan. "The economic design of $\overline{X}$ charts used to maintain current control of a process". *Journal of the American Statistical Association*, *51*(1956) 228-242.

[122] R. E. Bellman, S. E. Lee. "History and development of dynamic programming". *IEEE Control Systems Magazine*, *4*(1984) 24-28.

[123] F. Tsung, J. Shi. "Integrated design of run-to-run PID controller and SPC monitoring for process disturbance rejection". *IIE Transactions*, *31*(1999) 517-527.

[124] W. Jiang, K.L Tsui. "An economic model for integrated APC and SPC charts". *IIE Transactions*, *32*(2000) 505-513.

[125] D. Montgomery, J. B. Keats, M. Yatskievitch, W. Messina. "Integrating statistical process monitoring with feed forward control". *Quality and Reliability Engineering International*, *16*(2000) 515-525.

[126] W. A. J. Schippers. "An integrated approach to process control". *International Journal of Production Economics*, *69*(2001) 93-105.

[127] S. O. Duffuaa, S. N. Khursheed, S. M. Noman. "Integrating statistical process control, engineering process control and Taguchi's quality engineering". *International Journal of Production Research*, *42*(2004) 4109-4118.

[128] W. Jiang, J. V. Farr. "Integrating SPC and EPC methods for quality improvement". *Quality Technology and Quantitative Management*, *4*(2007) 345-363.

[129] L. Yang, S-H. Sheu. "Economic design of the integrated multivariate EPC and multivariate SPC charts". *Quality and Reliability Engineering International*, *23*(2007) 203-218.

[130] A. S. Matos, J. G. Requeijo, Z. L. Pereira. "Integration of engineering process control and statistical process control in pulp and paper industry". In: B. Braunschweig, X. Joulia, (Eds.). 18th European Symposium on Computer Aided Process Engineering, Elsevier, 2008.

[131] W. Hachicha, I. Moussa, R. Kolsi. "Integration of statistical and engineering process control in a batch process monitoring: case of alkyd polymerization reactor". *International Journal of Control and Automation*, *5*(2012) 45-62.

[132] X. Zhang, Z. He. "An integrated SPC-EPC study based on nonparametric transfer function model". *Applied Mathematics and Information Science*, *6-3S*(2012) 759-768.

[132a] Z. Q. Yu, J. W. Chew, P. S. Chow, R. B. H. Tan. "Recent advances in crystallization control - an industrial perspective". *Chemical Engineering Research and Design*, *85*(2007) 893-905.

[132b] W. J. Yoon, Y. S. Kim, I. S. Kim, K. Y. Choi. "Recent advances in polymer reaction engineering: Modeling and control of polymer properties". *Korean J. Chem. Eng.*, *21*(2004) 147-167.

[132c] G. A. Montague, A. J. Morris, A. C. Ward. "Fermentation monitoring and control: A perspective". *Biotechnology and Genetic Engineering Reviews*, *7*(1989) 147-188.

[132d] A. Johnson. "The control of fermentation processes". *In*: Modelling and Control of Biotechnological Processes, IFAC, 1985, p. 1-12.

[133] J-X. Xu, Y. Q. Chen, T. H. Lee, S. Yamamoto. "Terminal iterative learning control with an application to RTPCVD thickness control". *Automatica*, *35*(1999) 1535-1542.

[134] R. Chi, Z. Hou, S. Jin, D. Wang, C. Chien. "Enhanced data-driven optimal terminal ILC using current iteration control knowledge". *IEEE Transactions on Neural Networks and Learning Systems*, *26*(2015) 2939-2948.

[135] A. Hyo-Sung, Y. Q. Chen, K. L. Moore. "Iterative learning control: Brief survey and categorization". *IEEE Transactions on Systems, Man, and Cybernatics. - Part C: Applications and Reviews*, *37*(2007) 1099-1121.

[136] C. J. Pug. *End Point Detection in Reactive Ion Etch*ing (PhD thesis). 2013. Available at http://dicovery.ucl.ac.uk/1398304/1/Chris%20Pug%20Thesis%20amended.pdf

[137] W. Sutomo, X. Wang, D. Bullen, S. K. Braden, C. Liu. "Development of an end- point detector for parylene deposition process". *Journal of Microelectromechanical Systems*, *12*(2003) 64-69.

[138] W. Q. Meeker, L. A. Escobar. *Statistical Methods for Reliability Data*. J. Wiley & Sons., New York, 1998.

[139] W. B. Powell. *Approximate Dynamic Programming*. J. Wiley & Sons., New York, 2010.

[140] G. Taguchi. *System of Experimental Design*, Vol. 1. UNIPUB/Kraus International Publications, New York, 1987.