The case of tin-plating of surface mounted glass diodes

Jaap E. Wieringa*

March 10, 1998

Abstract

One of the steps in the production of diodes involves applying a protective tin/lead layer. In order to improve the quality of this process step at the plant of Philips Semiconductors Stadskanaal, a so-called *Process Action Team* (PAT) was instituted. This team, consisting of operators, a quality engineer, a service mechanic, a manager, and a statistician, worked together with the common objective to improve the quality of the tin-lead layers of the diodes. In this report I discuss some of the problems we had to beat when I had the opportunity to provide statistical assistance to this PAT. The tin/lead layer is applied in a chemical bath that needs adjustments from time to time. A *Linear Programming* model was developed to compute optimal adjustments. The remainder of the report deals with the subject when and how to adjust the bath, based on production data. Since these data appeared to be serially correlated, some form of regulation was needed in order to meet specifications. However, it turned out that the process was overregulated. It is discussed how we moved from overregulation (tampering with the process) to a process that is controlled by a sensible combination of regulation and SPC techniques. Due to the presence of serial correlation, it was not possible to apply the standard control chart techniques. We discuss what modifications were needed in order to be able to monitor correlated process data.

Keywords: Statistical Quality Control, Correlated observations, Time Series, Tampering, Linear Programming.

*Department of Econometrics, Economic Faculty, University of Groningen, PO Box 800, 9700 AV Groningen, The Netherlands. E-mail `j.e.wieringa@eco.rug.nl`

1 Introduction

The plant of Philips Semiconductors Stadskanaal is a leading supplier of diodes. Customers of Philips Semiconductors Stadskanaal are amongst others the automotive industry, the communications sector and manufacturers of consumer electronics. These customers are producers themselves, whose product quality is partly determined by the quality of the diodes. It is therefore not surprising that the customers of Philips are demanding with respect to the quality of the diodes. They require reliable, well functioning diodes that are easy to process.

In most cases the customers solder the diodes on a printed circuit board, so that ease of processing is to a large extend determined by solderability of the diodes. In order to ensure solderability of the diodes, Philips Stadskanaal is applying a protective tin/lead layer.

Insufficient layer thickness or wrong composition of the layer has been the cause of several customer complaints in the past. Philips Stadskanaal is therefore looking for ways to improve the process of applying the tin/lead layer, the objective being a better solderability.

In the last five years, Philips Stadskanaal acquired valuable experience with process improvement through successful application of *Statistical Process Control* (SPC) techniques (see Does, van Oord and Trip (1994)). The key to this success may be found in the approach that was chosen towards implementation of SPC.

At Philips Stadskanaal SPC techniques are implemented by so-called *Process Action Teams* (PAT's). A PAT is constituted as follows. Operators are important members because they are heavily involved in the process. The team is chaired by a responsible technical engineer. The team may be complemented with a quality engineer, a service mechanic and/or a developer. A neutral outsider with profound knowledge of and experience with *Statistical Process Control* completes the team. A PAT receives a clear mission what to improve, and the means to realize their plans (see Does et al. (1996)).

A PAT was started to improve the process step of tin-plating diodes. In this report I want to share some of the experiences I have gained when I had the opportunity to assist this PAT. In the sections 1.1–4 we will introduce respectively the product, the process, and the data that are gathered during the production process. Thereafter we will describe more or less chronologically the developments around this process step.

1.1 The product

A diode is an important electrical component that has the special property that it conducts current in only one direction, while it has a high resistance in its reverse direction. Diodes are used in all kinds of electrical circuits such as TV sets, computers, automotive ignition systems, telecommunication apparatus, power supplies for X-ray generators, and a great variety of consumer electronics.

The plant of Philips Semiconductors Stadskanaal makes four different types of glass-encapsulated diodes. Each type is made using a different manufacturing process. In this report we will restrict ourselves to one type of diodes, the so-called *Surface Mounted Implosion Diodes* (SMID's). An exploded view of a SMID is depicted in figure 1.1.

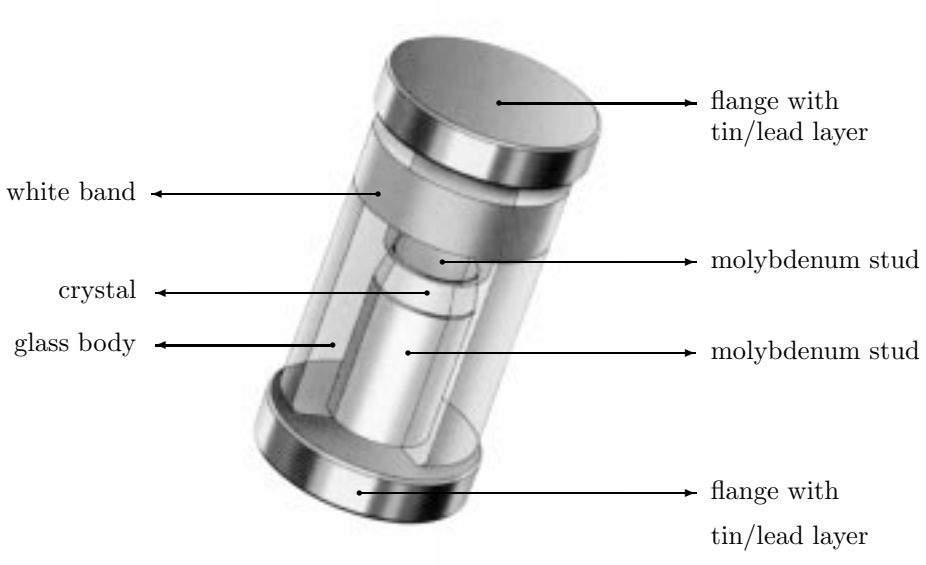

Figure 1.1: A surface mounted implosion diode.

One of the things that may strike an outsider in figure 1.1 is that the connection points of the diode are flanges rather than leads. This makes this type of diodes fit for surface mounting. These diodes are therefore called *surface-mounted* diodes. Surface mounted diodes are smaller than leaded diodes and easier to process in automated industry. Philips Semiconductors Stadskanaal produces both types of diodes, and data is available for both types of diodes. However, the tin/lead layer is more critical for solderability of surface-mounted diodes than for solderability of leaded diodes, so that we will restrict ourselves to (data of) surface-mounted products.

1.1.1 The crystal

The crystal is the heart of a diode. It is made of a small slice of the semiconducting material silicon. By impurifying both sides of a silicon wafer, one side with phosphorus, the other with boron, the silicon conducts power in one direction and blocks power in the other direction. The impurities are brought into the silicon wafer by a diffusion process. Subsequently, several crystals are produced from one wafer.

1.1.2 Assembly of the diodes

The crystal of the diode of figure 1.1 is placed between two studs/flange pairs. An implosion process follows that tightly fits a glass body around the crystal and the studs. The implosion process takes place in vacuum so that the edge of the crystal

only contacts the nonconductor glass. This prevents charge carriers from traveling through the diode in an other way than through the crystal.

The studs are made of molybdenum, the flanges are made of copper. Molybdenum is a metal that has a good thermal conductivity and has the additional property that (when oxidized) it adheres well to the glass body. The adherence of the glass body with the molybdenum studs provides mechanical strength and (more importantly) a hermetic sealing of the diode. The latter is very important since exposing the crystal to open air distorts the electrical properties of the diode.

1.1.3 The tin/lead layer

While oxidation of the molybdenum studs is helpful, this is not the case for oxidation of the copper flanges. One of the properties of oxidized copper is that the adherence with tin is bad. The customers of the type of diodes of figure 1.1 solder the diodes with tin, so that solderability (and hence ease of processing of the diodes) is adversely influenced by the oxidation of the copper flanges. To overcome this problem, the oxidized copper is removed, and a tin/lead layer that prevents the flanges from contacting oxygen is applied.

If the diodes were soldered right away, any thickness of the tin/lead layer would provide sufficient protection from oxidation. However, the diodes are shipped to factories and warehouses all over the world, where they may be kept in stock for some time. If the thickness of the tin/lead layer is not sufficient, diffusion of copper atoms in the tin/lead layer will result in copper atoms reaching the surface of the layer, and these atoms will oxidize. This phenomenon is one of the causes of bad solderability. For this reason, a lower specification limit is set for the thickness of the layer. A minimal thickness of 1μm (1×10^{-6}m) proves to be sufficient to warrant good solderability after a two years stay in any warehouse (provided some conditions considering relative humidity and temperature are met).

1.1.4 Inspection

After tin-plating, each diode is inspected. Several electrical characteristics of *each* diode are measured in this process step. According to these measurements a decision is made whether to accept or to reject the diode. For some types of diodes, a classification in sub-types is made.

The tin/lead layer is not only important for solderability but also facilitates the inspection. If it were absent, a potential difference would arise between copper-oxide on the surface of the flange and the pure copper inside the flange. This would introduce a large measurement error.

The inspection step makes some additional demands upon the tin/lead layer of the diodes. Firstly, the thickness of the tin/lead layer may not be too large, otherwise the diodes will not pass the sieve that is used to prevent crooked diodes from entering the measuring apparatus. In practice, there appears not to be a rigid upper limit on the thickness of the tin/lead layer, but layers thicker than 50μm are thought of as blobs of tin. A second, more rigid requirement concerns the portion of lead in the

tin/lead layer. If this portion becomes too large, the layer will be too soft. Putting a diode with too soft a layer through the measuring apparatus will grind off some of the tin/lead layer. These grindings smudge the glass body of following diodes, which may lead to leakage of current over the body. Such diodes are rejected, while the electrical characteristics might have been perfect before entering the measuring apparatus. To guard against waste caused by this phenomenon, the lower specification limit of the tin portion in the tin/lead layer is set to 77%.

1.1.5 Coding and packing

In the following process step, a white band is painted on the glass body of the accepted diodes (see figure 1.1). The placement of the band indicates the cathode side of the diode. The electrical characteristics can be deduced from a code that is also printed on the glass body.

As a final process step, the diodes are packed in such a way that the customers can process the diodes in an automated way.

1.2 Relevant quality characteristics of the tin/lead layer

During our investigation, we confined ourselves to only one of the process steps that were described in the previous section, namely applying a tin/lead layer. As we have seen, this tin/lead layer is important for the customers, since it greatly improves solderability of the diodes. The layer also facilitates inspection, which is the process step directly following the tin-plating of the diodes. As a result of these two fitnesses for use, the quality of the tin/lead layer is determined by two characteristics: solderability and composition.

The first quality characteristic is not directly observable, since it is the net result of a complex of determinants such as thickness of the layer, composition of the layer, pollution of the layer with organic materials, and so on. The unobservable demand "*a good solderability*" is therefore translated into a requirement on one of its observable determinants: the thickness of the layer. A lower specification limit on the thickness of the layer is set to 1μm (see subsection 1.1.3). Although excessive thickness also causes problems, a strict upper specification limit is not used in practice, and will therefore not be considered in the remainder of this report. The target thickness is 10μm.

The second quality characteristic, the composition of the layer, concerns the fraction of tin in the tin/lead layer. This characteristic is directly observable. Again, only a lower specification limit is formulated for this characteristic: the fraction tin should exceed 77% (see subsection 1.1.3). The target value is 80%. In the next section we will describe how the tin/lead layer is applied.

1.3 The tin-plating process

The layer is applied by means of an galvano electro chemical process. A schematic view of the tin-plating process is depicted in figure 1.2.

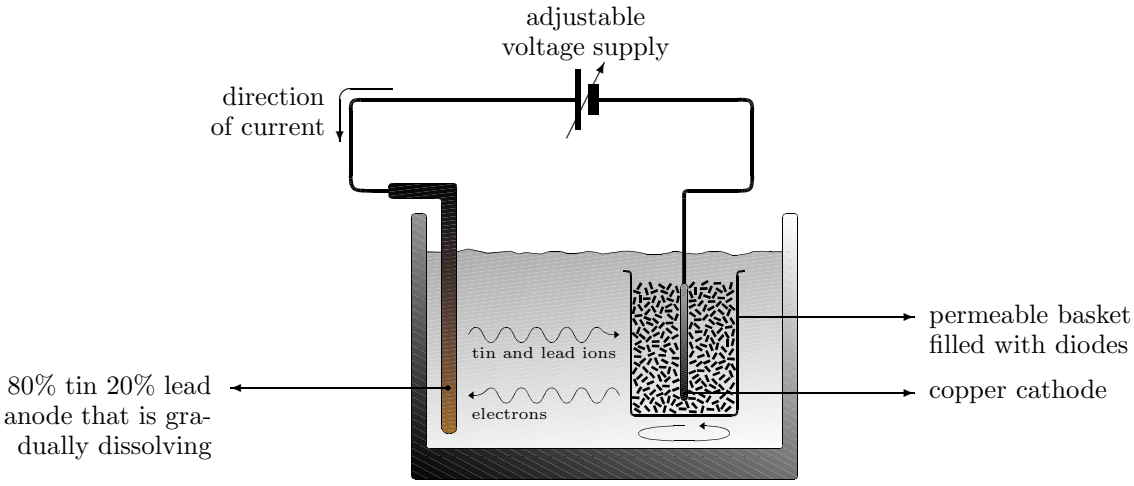

Figure 1.2: A schematic view of the tin-plating process.

The tin-plating takes place in a conducting chemical bath. As we can see in figure 1.2, a basket that is permeable with respect to the tin-plating liquid is filled with diodes and placed into the bath. Subsequently, current is run through the bath in such a way that the amperage is kept at a constant value. This is done by automatic adjustment of the voltage. On the cathode side of the voltage supply, metal ions that are dissolved in the liquid meet electrons and precipitate on flanges that contact the cathode. That is, on the cathode side we have the following chemical reactions:

$$\mathrm{Sn}^{2+} + 2\mathrm{e}^{-} \longrightarrow \mathrm{Sn}$$

and

$$\mathrm{Pb}^{2+} + 2\mathrm{e}^{-} \longrightarrow \mathrm{Pb},$$

where Sn is the chemical symbol for tin, Pb is the chemical symbol for lead, and electrons are denoted by e^{-}.

As a result of this reaction on the cathode side of the voltage supply, the concentration of metal ions in the liquid will decrease. Fortunately, by using an anode of a 80% tin, 20% lead alloy (see figure 1.2) it is possible to replenish the metal ions. With current running through the bath, electrons of tin and lead molecules in the anode are sent away in the direction of the voltage supply and the remaining metal ions dissolve in the liquid. Hence, the reverse of the chemical reaction that takes place at the cathode will take place at the anode. The net effect is that tin and lead molecules are transferred from the anode of the voltage supply to diode flanges.

Since only flanges contacting the cathode of the voltage supply receive a layer, the basket is rotated continuously to give each flange a chance to contact the cathode.

Important parameters of the process that influence the quality of the tin/lead layer are amongst others the temperature of the bath, the current density, the length of

the time span the diodes stay in the bath, the voltage and the chemical composition of the bath. To maintain a high quality of the tin/lead layer, it is desired to control each of these parameters. This is done easier for some parameters than for others. Despite the fact that a lot of effort was put in maintaining a stable bath, the chemical composition of the bath appeared to be the most difficult to control. In practice it turned out to be very difficult to derive a proper control strategy that could deal with exhaustion of the bath due to production. This is the problem we will focus on in the remainder of this report. For the combined effect of other parameters we will assume that they do not have a systematic effect on the tin/lead layer.

The chemical bath of figure 1.2 is a mixture of five components:

1. acid;
2. an Sn^{2+} solution;
3. a Pb^{2+} solution;
4. brightener;
5. formalin.

The acid is thought of as an important component since it takes care of the conductivity of the bath. The two metal solutions are also considered to be important since the Sn^{2+} and Pb^{2+} ions therein precipitate on the diodes to form the tin/lead layer. The remaining components, brightener and formalin, are considered to be necessary for the tin-plating process, but, as we will see in section 4.3, their importance was underrated. The brightener takes care of a smooth layer, while the formalin helps to control the proportions tin and lead in the resulting layer.

During the production process, the composition and the volume of tinning liquid changes. Changes in the composition are due to chemical reactions that take place in the production process. Furthermore, some components (e.g. formalin) evaporate quicker than other ones so that their relative proportions change. The overall volume of the bath decreases due to evaporation and tinning liquid being dragged out together with the diodes. As a result, the tin-plating bath needs to be replenished regularly.

Before the PAT was started the bath used to be replenished with fixed additions of the separate components, independent of the actual composition of the bath. This replenishment strategy was not satisfactorily, since it resulted in a high level of chemicals being used, while the bath did not appear to be stable. The instability of the bath affected the quality of the tin/lead layers on the diodes in such a way that customer complaints were received. For this reason, a new replenishment strategy was adopted that used the actual contents of the bath to determine the additions. To this end, the bath was analyzed every day. Based on the results of the analysis, an online computer computed what additions were necessary to ensure that the bath was fit for production again. The new strategy resulted in a large reduction in the amount of chemicals being used. The quality of the tin/lead layer however remained unsatisfactory.

At this point we got involved in the problem. The first improvement we could suggest concerned the way in which the additions were computed. We will see in chapter 2 that it is possible to formulate this computational problem as a linear programming problem, for which an optimal solution (in the sense of minimal costs of additions) can be found.

However, even before this computational improvement was implemented in the software, we obtained further insight in the process by studying measurements that were taken from the production process. This led to another replenishment strategy, which we will describe in chapter 5. But before doing so, we will first discuss the data that were available during our investigation.

1.4 The data

We have two kinds of observations available from the tin-plating process. We have measurements on the products at the end of the process, and measurements on the process itself.

Initially, the product measurements consisted of samples of size 10 that were taken from every batch of diodes (a batch may contain up to 31000 diodes and about 60 batches are processed each day in three shifts). From each of the sampled diodes, the thickness and the composition of the tin/lead layer is measured. The frequency of sampling and the sample size was set some years ago, and nobody could remember the exact reasons for this sampling strategy. Taking these measurements took considerable time of the operators. During the PAT meetings, we first decided to take a sample every other batch, and later on to reduce the sample size to five as well.

The size of the samples is sufficiently large to obtain an impression of the performance of the process. This makes the data fit for *process monitoring*. However, the data were used for *acceptance sampling*. On the basis of a sample of size 10 it was decided to reject or accept a batch of 31000 diodes! This was done in the following way: if one or more of the sampled diodes had a tin/lead layer smaller than 3μm, or if the sample mean was smaller than 4.85μm, it was decided to tin-plate the whole batch again. Otherwise, the batch was sent on to the inspection department. This sampling plan is not sufficient to ensure a high quality level of the outgoing batches. For example, assuming independence and normality of the observations and a (realistic) standard deviation equal to 4, a batch with 5% of the tin/lead layers smaller than 1μm passes this test with a probability of about 25%. Indeed, if one of the conditions is not met, then there are good reasons to suspect that something is wrong. However, to ensure an outgoing quality level of only a few defective parts per million, the sample size must be increased drastically.

The process measurements consist of a chemical analysis of the tin-plating bath. By means of a titration the concentrations of acid, Sn^{2+} and Pb^{2+} are measured. The concentration of brightener can only be roughly determined, while there is no measurement device available for establishing the concentration of formalin. The analysis is performed once a day by the operators. In section 1.3 we discussed that this

analysis is the input for a computer program that computes the necessary additions in order to obtain a bath that is fit for applying a good tin/lead layer. In the next chapter we will describe how we improved this procedure.

2 Computation of additions: an application of linear programming

2.1 Introduction

The chemical composition of the tin-plating bath changes due to production and evaporation of the tin-plating liquid. In order to maintain a good quality of the tin/lead layer, the concentrations of the components of the bath must fall between certain limits. In section 1.3 we saw that the bath consists of a mixture of the following five components:

1. acid;
2. an Sn^{2+} solution;
3. a Pb^{2+} solution;
4. brightener;
5. formalin.

For each of these components, a lower limit is set on the concentration. If the concentration of one of the components falls below its limit, the quality of the tin/lead layer deteriorates. On the concentrations of acid, Sn^{2+} solution, and Pb^{2+} solution also an upper limit is set.

If the composition of the bath is such that one or more of the concentrations fall outside the limits, action is required. Fortunately, there are highly concentrated solutions for each of the components which can be added in case its concentration has become too low. If the concentration of one of the components is too high, demineralized water can be added to lower the concentration.

Until so far the problem seems quite straightforward: concentrations that are too low can be raised by adding some of the appropriate solutions. Concentrations that are too high can be lowered by thinning the bath. However, computing how much is needed of every solution is not as straightforward as it may seem at first sight. The problem is that all five concentrations change if one of the components is added. This may cause trouble if we compute the additions for each of the components one by one. This can be illustrated by means of an example.

Suppose that we have computed how much of the first four components must be added so that their requirements are met. It then may happen that the addition of the fifth component lowers the concentrations of the first four in such a way that one or more fall below their lower limits. This means that an extra addition of such a component is required. But these additions in their turn lower the concentrations of all other components, which again may lead to extra additions, etc. In this way, this computation may result in an infinite loop, where large amounts are added to the bath.

This is also what happened in practice. The operators often had the feeling that the program prescribed far too large additions, relative to the results of the bath

analysis. It regularly happened that addition of the prescribed amounts would have resulted in an overflowing bath. This is of course an unwanted situation, since the tin-plating liquid is expensive. Moreover, the liquid contains the heavy metals tin and lead, and any surpluses have to be sent to a purifying installation before it is allowed to drain them off. The costs associated with purifying one liter of tin-plating liquid are also relatively high, comparable to the costs of buying a liter. So, throwing away tin-plating liquid is very expensive, and it was felt that the way the additions were computed was partly responsible for the large amounts of chemicals used in this department.

In the next sections, we will take a cost-minimizing approach to our replenishment problem. It turns out that it can be formulated as a linear programming model. The input of the model is a chemical analysis of the bath, and the output is a prescription of how much of each component must be added so that the bath meets all requirements. The output will be an *optimal* solution to our problem in the sense that it is not possible to find a cheaper set of additions that will bring all the concentrations between their limits.

At this point we want to make two remarks. Firstly, it is important to realize that we are discussing a deterministic problem in this chapter. We suppose that we precisely know the concentrations of the components of the bath, and all we want to do is to find a way to compute the cheapest additions that will make our bath fit for production again. Secondly, our model will not result in a *control strategy*: the behavior of the process in the course of time is not incorporated. We merely developed a tool that enables us to get from an 'unwanted' situation to a 'more wanted' situation in the cheapest way. In this sense, it can become a tool that facilitates a control strategy. Depending on the behavior of the process in time, a control strategy can define 'more wanted situations', and with our LP model we are able to compute the cheapest way to get there.

In section 2.2 we will introduce some notation, which will be followed by a mathematical formulation of the problem in section 2.3. In section 2.4 we will compare some of the results of our model to the results of the current computer program. A short conclusion in section 2.5 closes this chapter.

2.2 Notation

As the input of our model we have the concentrations of the bath components. We let $m_1, m_2, \cdots, m_5$ denote these five values:

m_1 :	concentration of acid	in (gr/l);
m_2 :	concentration of Sn^{2+} solution	in (gr/l);
m_3 :	concentration of Pb^{2+} solution	in (gr/l);
m_4 :	concentration of brightener	in (ml/l);
m_5 :	concentration of formalin	in (gr/l).

For all of the components, a lower limit is set on the concentration. On the concentration of acid, Sn^{2+} solution, and Pb^{2+} solution also an upper limit is set. We denote the lower limit for component i with l_i and the upper limit for component i with u_i. Hence for our components we have:

l_1, u_1 : lower and upper limit for the concentration of acid;
l_2, u_2 : lower and upper limit for the concentration of Sn^{2+} solution;
l_3, u_3 : lower and upper limit for the concentration of Pb^{2+} solution;
l_4 : lower limit for the concentration of brightener;
l_5 : lower limit for the concentration of formalin.

For each of the five components, we have highly concentrated solutions that can be added in case its concentration in the tin-plating liquid has become too low. This results in five *decisions*: for each component we have to determine the volume we must add in order to bring all concentrations within their ranges. With decision i we associate a *decision variable* x_i that indicates the number of liters we add from component i:

x_1 : addition of acid (in liters);
x_2 : addition of Sn^{2+} solution (in liters);
x_3 : addition of Pb^{2+} solution (in liters);
x_4 : addition of brightener (in liters);
x_5 : addition of formalin (in liters).

If the concentration of one of the components is too high, demineralized water can be added. This induces the need for a sixth decision variable x_6 indicating the number of liters of demineralized water we decide to add. Furthermore, in cases where the bath volume is high, some tin-plating liquid may have to be drained off before making the additions associated with $x_1, x_2, \cdots, x_6$ in order to prevent the bath from overflowing. To this end, we define x_7 as the number of liters we decide to drain off before making the additions. So we enlarge the list of decision variables above with

x_6 : addition of demineralized water (in liters);
x_7 : the number of liters of tin-plating liquid we choose to drain off *before* the additions are made.

We emphasized 'before' in the declaration of x_7 since the number of liters to be drained off was not taken in consideration in the old situation. The new bath volume was reported, whether the bath was overflowing or not. So, *after* the additions were computed it was known how many liters had to be drained off. This does in general not equal the number of liters to be drained off *before* making additions, since draining off affects the additions to be made.

In order to be able to compute how much of the highly concentrated solutions should be added, we need to know what their concentrations are. We denote these concentrations by $d_1, \cdots, d_5$:

d_1 : concentration of acid solution (in grammes/liter);
d_2 : concentration of Sn^{2+} solution (in grammes/liter);
d_3 : concentration of Pb^{2+} solution (in grammes/liter);
d_4 : concentration of brightener (in milliliter/liter);
d_5 : concentration of formalin (in grammes/liter).

With each of the decision variables, costs are associated. For the first six variables this is simply the price per liter, while for the cost associated with x_7 the cost price of purifying one liter of tin-plating liquid is used. We denote these costs by $c_1, \cdots, c_7$:

c_1 : costs of adding one liter of acid;
c_2 : costs of adding one liter of Sn^{2+} solution;
c_3 : costs of adding one liter of Pb^{2+} solution;
c_4 : costs of adding one liter of brightener;
c_5 : costs of adding one liter of formalin;
c_6 : costs of adding one liter of demineralized water;
c_7 : costs of purifying one liter of drained off tin-plating liquid.

2.3 Mathematical formulation of the problem

Our objective is to minimize the costs of replenishing the bath, such that the concentrations fall between predetermined limits. With the symbols introduced in the preceding section we can express our objective as follows

$$\min\ c_1x_1 + c_2x_2 + c_3x_3 + c_4x_4 + c_5x_5 + c_6x_6 + c_7x_7.$$

It is also possible to formulate our restrictions in mathematical terms. If we let V_{old} denote the volume of the bath before replenishing, our new bath volume will be $V_{\text{old}} + x_1 + x_2 + x_3 + x_4 + x_5 + x_6 - x_7$. Since the new bath volume may not exceed a certain maximum bath volume $b_{\max}$, we have as a first constraint

$$V_{\text{old}} + x_1 + x_2 + x_3 + x_4 + x_5 + x_6 - x_7 \leq b_{\max}.$$

The new bath volume must exceed a certain minimal value $b_{\min}$, leading to the following constraint:

$$V_{\text{old}} + x_1 + x_2 + x_3 + x_4 + x_5 + x_6 - x_7 \geq b_{\min}.$$

The tin-plating liquid contains $m_2(V_{\text{old}} - x_7)$ grammes of tin directly after an amount of x_7 liters is drained off and just before any additions are made. We make the decision to add d_2x_2 grammes of tin, so that after replenishment the bath contains $m_2(V_{\text{old}} - x_7) + d_2x_2$ grammes of tin. If we divide this by the new bath volume, we have the concentration of tin after replenishment. Since this concentration must exceed l_2, we arrive at the following constraint

$$\frac{m_2(V_{\text{old}} - x_7) + d_2 x_2}{V_{\text{old}} + x_1 + x_2 + x_3 + x_4 + x_5 + x_6 - x_7} \geq l_2$$

which can be rewritten as

$$l_2 x_1 + (l_2 - d_2)x_2 + l_2 x_3 + l_2 x_4 + l_2 x_5 + l_2 x_6 + (m_2 - l_2)x_7 \leq (m_2 - l_2)V_{\text{old}}.$$

Analogously, we find the following constraint associated with the upper limit of the tin concentration in the bath

$$-u_2 x_1 + (d_2 - u_2)x_2 - u_2 x_3 - u_2 x_4 - u_2 x_5 - u_2 x_6 + (u_2 - m_2)x_7 \leq (u_2 - m_2)V_{\text{old}}.$$

Following exactly the same line of reasoning, we find constraints associated with the lower and upper limit of the concentration of lead, and constraints associated with the lower limit of the concentration of brightener.

Until so far, we skipped the constraints associated with limits of the concentration of acid and formalin. The reason for this is that these constraints are different from the constraints above.

For the constraints on the concentration of acid, the difference is due to the fact that there are two ways to add acid to the bath. First, like all other components, we have acid available as a separate component for making additions to the tin-plating liquid. But secondly, acid is also the solvent of the Sn^{2+} solution, so that with each addition of Sn^{2+} solution, acid is added to the bath, too. Therefore, besides the addition of acid, which we denoted by x_1, we have to take the amount of Sn^{2+} solution added (x_2) into account in controlling the concentration of acid. First, let us define

$$d_* : \quad \text{concentration of acid in } Sn^{2+} \text{ solution (in grammes/liter).}$$

The number of grammes of acid we decide to add to the tin-plating liquid can then be written as $d_1 x_1 + d_* x_2$. Just before any additions are made, $m_1(V_{\text{old}} - x_7)$ grammes of acid are present in the bath, so that after addition we can write the concentration acid as

$$\frac{m_1(V_{\text{old}} - x_7) + d_1 x_1 + d_* x_2}{V_{\text{old}} + x_1 + x_2 + x_3 + x_4 + x_5 + x_6 - x_7}.$$

The lower limit on the concentration of acid then leads to the following constraint

$$(l_1 - d_1)x_1 + (l_1 - d_*)x_2 + l_1 x_3 + l_1 x_4 + l_1 x_5 + l_1 x_6 + (m_1 - l_1)x_7 \leq (m_1 - l_1)V_{\text{old}},$$

while the upper limit can be expressed as

$$(d_1 - u_1)x_1 + (d_* - u_1)x_2 - u_1 x_3 - u_1 x_4 - u_1 x_5 - u_1 x_6 + (u_1 - m_1)x_7 \leq (u_1 - m_1)V_{\text{old}}.$$

Regarding the constraint on the lower limit of formalin, one may have noticed that we described in section 1.4 that there is no measurement device available to determine

the concentration of formalin. In practice, the concentration of formalin is estimated at its lower limit and enough formalin is added to ensure that its concentration does not fall below the lower limit if additions of the other components are made.

If the concentration of formalin were observable we would have had the constraint

$$l_5x_1 + l_5x_2 + l_5x_3 + l_5x_4 + (l_5 - d_5)x_5 + l_5x_6 + (m_5 - l_5)x_7 \leq (m_5 - l_5)V_{\text{old}},$$

but by setting $m_5 = l_5$ this reduces to

$$l_5x_1 + l_5x_2 + l_5x_3 + l_5x_4 + (l_5 - d_5)x_5 + l_5x_6 \leq 0.$$

Note that this constraint is independent of $V_{\text{old}} - x_7$, the volume after draining off and before adding. It only depends on the volume of the additions. If for the unknown concentration of formalin holds that $m_5 > l_5$, this constraint is more restrictive than the constraint we would have had if m_5 was available, and less restrictive if $m_5 < l_5$. For this reason, combined with the quick evaporation of formalin it is felt that at least 0.5 liters of formalin should be added, so that we include the following constraint in our model

$$x_5 > 0.5.$$

To complete the model, we add a few trivial constraints. The first is that it is not possible to drain off more tin-plating liquid than the old volume of the bath:

$$x_7 \leq V_{\text{old}}.$$

And finally, we cannot add or drain off negative amounts. That is, we add the following nonnegativity constraints

$$x_i \geq 0 \qquad \text{for } i = 1, 2, \cdots, 7.$$

We have formulated the replenishment problem as a problem of minimization of a linear objective function, subject to linear constraints. In summary, the complete model is

$$
\begin{array}{lrrrrrrrrl}
\min & c_1x_1 & + c_2x_2 & + c_3x_3 & + c_4x_4 & + c_5x_5 + & c_6x_6 & + c_7x_7 & & \\
\\
\text{s.t.} & - x_1 & - x_2 & - x_3 & - x_4 & - x_5 & - x_6 & + x_7 & \leq & V_{\text{old}} - b_{\min} \\
& x_1 & + x_2 & + x_3 & + x_4 & + x_5 & + x_6 & - x_7 & \leq & b_{\max} - V_{\text{old}} \\
\\
& (l_1 - d_1)x_1 + & (l_1 - d_*)x_2 & + l_1x_3 & + l_1x_4 & + l_1x_5 + & l_1x_6 + & (m_1 - l_1)x_7 & \leq & (m_1 - l_1)V_{\text{old}} \\
& (d_1 - u_1)x_1 + & (d_* - u_1)x_2 & - u_1x_3 & - u_1x_4 & - u_1x_5 - & u_1x_6 + & (u_1 - m_1)x_7 & \leq & (u_1 - m_1)V_{\text{old}} \\
\\
& l_2x_1 + & (l_2 - d_2)x_2 & + l_2x_3 & + l_2x_4 & + l_2x_5 + & l_2x_6 + & (m_2 - l_2)x_7 & \leq & (m_2 - l_2)V_{\text{old}} \\
& - u_2x_1 + & (d_2 - u_2)x_2 & - u_2x_3 & - u_2x_4 & - u_2x_5 - & u_2x_6 + & (u_2 - m_2)x_7 & \leq & (u_2 - m_2)V_{\text{old}} \\
\\
& l_3x_1 & + l_3x_2 + & (l_3 - d_3)x_3 & + l_3x_4 & + l_3x_5 + & l_3x_6 + & (m_3 - l_3)x_7 & \leq & (m_3 - l_3)V_{\text{old}} \\
& - u_3x_1 & - u_3x_2 + & (d_3 - u_3)x_3 & - u_3x_4 & - u_3x_5 - & u_3x_6 + & (u_3 - m_3)x_7 & \leq & (u_3 - m_3)V_{\text{old}} \\
\\
& l_4x_1 & + l_4x_2 & + l_4x_3 + & (l_4 - d_4)x_4 & + l_4x_5 + & l_4x_6 + & (m_4 - l_4)x_7 & \leq & (m_4 - l_4)V_{\text{old}} \\
\\
& l_5x_1 & + l_5x_2 & + l_5x_3 & + l_5x_4 + & (l_5 - d_5)x_5 + & l_5x_6 & & \leq & 0 \\
\\
& & & & & - x_5 & & & \leq & -0.5 \\
\\
& & & & & & & x_7 & \leq & V_{\text{old}}
\end{array}
$$

$$x_1,\ x_2,\ x_3,\ x_4,\ x_5,\ x_6,\ x_7 \geq 0.$$

In this form, the replenishment problem is a standard *Linear Programming* problem, that can be solved by using for example the *Simplex Algorithm*. An optimal solution of the problem is a set of values for the decision variables $x_1, x_2, \cdots, x_7$, such that the costs of replenishment are minimal while all restrictions are fulfilled.

Each time the model is run, it may differ from previous models because of different results of the analysis $(m_1, \cdots, m_5)$ and a different volume of the bath (V_{old}). Hence, assuming that the lower and upper limits and the concentrations of the additions remain unchanged, the model has six parameters.

If the model is solved by means of the Simplex Algorithm, a starting point (a so-called *basic feasible* solution) is needed. Generating a basic feasible solution can be done for example by means of the Big M method, or by the first phase of the Two Phase Simplex Method (see e.g. Sierksma (1996)).

However, some computing efficiency could be gained if we had a starting point for the Simplex Algorithm that is independent of the measurements. Then it would not be necessary to start a procedure each time a new analysis becomes available to find a basic feasible solution. We could just plug in our 'standard' basic feasible solution, and start the Simplex Algorithm right away, whatever the measurements are.

Fortunately, for our problem a measurement independent basic feasible solution exists if it is possible to create a new bath out of the separate components. This is intuitively clear if we realize that this corresponds to draining off the whole bath, i.e. setting x_7 equal to V_{old}, and create a new one out of the separate components. If such a solution exists, it is independent of the measurements since we started with an empty bath. In mathematical terms, if we set $x_7 = V_{\text{old}}$, our model reduces to

$$
\begin{array}{lrrrrrrcl}
\min & c_1x_1 & + c_2x_2 & + c_3x_3 & + c_4x_4 & + c_5x_5 + & c_6x_6 & & \\
\\
\text{s.t.} & - x_1 & - x_2 & - x_3 & - x_4 & - x_5 & - x_6 & \leq & -b_{\min} \\
& x_1 & + x_2 & + x_3 & + x_4 & + x_5 & + x_6 & \leq & b_{\max} \\
\\
& (l_1 - d_1)x_1 + & (l_1 - d_*)x_2 & + l_1x_3 & + l_1x_4 & + l_1x_5 + & l_1x_6 & \leq & 0 \\
& (d_1 - u_1)x_1 + & (d_* - u_1)x_2 & - u_1x_3 & - u_1x_4 & - u_1x_5 - & u_1x_6 & \leq & 0 \\
\\
& l_2x_1 + & (l_2 - d_2)x_2 & + l_2x_3 & + l_2x_4 & + l_2x_5 + & l_2x_6 & \leq & 0 \\
& - u_2x_1 + & (d_2 - u_2)x_2 & - u_2x_3 & - u_2x_4 & - u_2x_5 - & u_2x_6 & \leq & 0 \\
\\
& l_3x_1 & + l_3x_2 + & (l_3 - d_3)x_3 & + l_3x_4 & + l_3x_5 + & l_3x_6 & \leq & 0 \\
& - u_3x_1 & - u_3x_2 + & (d_3 - u_3)x_3 & - u_3x_4 & - u_3x_5 - & u_3x_6 & \leq & 0 \\
\\
& l_4x_1 & + l_4x_2 & + l_4x_3 + & (l_4 - d_4)x_4 & + l_4x_5 + & l_4x_6 & \leq & 0 \\
\\
& l_5x_1 & + l_5x_2 & + l_5x_3 & + l_5x_4 + & (l_5 - d_5)x_5 + & l_5x_6 & \leq & 0 \\
\\
& & & & & - x_5 & & \leq & -0.5 \\
\\
& x_1,\ x_2,\ x_3,\ x_4,\ x_5,\ x_6 \geq 0, & & & & & & &
\end{array}
$$

where we suppressed the term c_7V_{old} in the objective function, since this is a constant and therefore does not influence the optimal solution of the problem. Solving this

model provides a 'standard' basic feasible solution to the replenishment problem that is independent of the measurements. This solution can be interpreted as the cheapest way to compose an admissible new tin-plating bath out of its separate components.

Note that the inclusion of the trivial constraint $x_7 \leq V_{\text{old}}$ in the original model (which is not relevant for the optimal solution) makes such a solution a *basic* feasible solution.

2.4 Comparison of results

In this section we will present a small comparison of the old method and the solution of the LP model. To this end, we use four *real* analyses, and compare the output of the method currently used with the output of the LP model. The method currently used basically computes the additions one by one. In section 2.1, we discussed why this approach may result in too large additions.

The following comparison is based on analyses that are taken from the tin-plating bath on the 10th of July, the 29th of August, the 27th of September, and the 19th of October of 1995. These were handed to us by the responsible technical engineer, whom we asked for a few typical measurements. He came up with these four analyses, which, according to him, represented combinations of concentrations that were encountered frequently in practice. As we will see, for all four measurements, the LP method results in considerable savings.

2.4.1 The analysis of July 10, 1995

On the tenth of July, the analysis indicated that the concentration acid fell between its upper and lower limits, but that the concentrations of Sn^{2+} solution, Pb^{2+} solution, and brightener were too low. The bath volume was low (470 liters), so that a considerable addition was possible before the bath would overflow.

component	**results of old method**	**results of old method***	**results of LP-problem**
acid	114.6 liters	71.4 liters	59.16 liters
Sn^{2+} solution	135.0 liters	84.1 liters	74.99 liters
Pb^{2+} solution	26.9 liters	16.8 liters	15.36 liters
brightener	27.3 liters	17.0 liters	16.00 liters
formalin	3.5 liters	2.2 liters	1.67 liters
demineralized water	0.0 liters	0.0 liters	0.00 liters
drain off	0.0 liters	177.3 liters	37.18 liters
new bath volume	777.30 liters	484.20 liters	600.00 liters
total costs:	*f* 8317.48	*f* 6853.01	*f* 3088.33

*draining off before making additions allowed

Table 2.1: Comparison of replenishing strategies for July 10, 1995.

In table 2.1 we present the results of the old method and the output of the LP model. Note that there are two columns for the solution of the old method. In the first column, indicated by 'results of old method' the rough results of the old method can be found. From this we see that, since the maximum bath volume $b_{\max} = 600$ liters, the bath would overflow with 177.3 liters. This suggestion is therefore not admissible. Furthermore, since this method does not take draining off into account, the additions are computed on the basis of a high volume bath. Apart from any computing inefficiencies, this is another reason why the additions are so high.

The second column presents the results of running the program with the same initial concentration measurements, but now it is allowed to drain off some tin-plating liquid first. Usually the operators drained off an amount equal to the surplus of the first suggestion (in this case 177.3 liters).

When draining off before making additions is allowed, both the amount of chemicals used and total costs are reduced. Total costs decrease with approximately 1500 guilders. However, even a greater cost reduction is possible if we use the results of the LP-problem, as we can see from the last column of table 2.1. The main part of the cost reduction stems from finding a combination of the decision variables so that the number of liters to drain off is kept at a minimum. As a result, the new bath volume is maximal, while the new bath volume of the old method is close to its minimal value of $b_{\min} = 450$ liters.

2.4.2 The analysis of August 29, 1995

The concentration of Sn^{2+} solution in the tin-plating bath on the 29th of August was a little too high. All other concentrations were acceptable. The bath volume was 575 liters, so that only small additions would result in a non-flooding bath. Fortunately, the bath only had to be thinned a little bit. In table 2.2 we present the results of the two strategies. Note that in this table only one column is presented for the output of the old method since no draining off appears to be necessary.

The differences between the methods are less spectacular in absolute sense as in the previous case, but the cost reduction is enormous if expressed in percentages. The total costs of the LP solution are only **0.6%** of the total costs of the old method.

The LP solution adds the obligatory 0.5 liters of formalin, and some demineralized water. The old method suggests to add more water, together with Pb^{2+} solution and brightener. The last two additions explain the main part of the cost difference between the methods. Note that the old method has no lower limit on the addition of formalin.

2.4.3 The analysis of September 27, 1995

On the 27th of September the tin-plating bath met all but one of the requirements: the concentration Sn^{2+} solution was too low. The concentration of Pb^{2+} solution was found exactly at its lower limit, so that with any addition of another component, addition of Pb^{2+} solution would become necessary. The bath volume was 475 liters, so that quite large additions were possible. In table 2.3 the results of the two methods can be found. Since the output of the old method initially results in an overflowing

component	results of old method	results of LP-problem
acid	0.0 liters	0.00 liters
Sn^{2+} solution	0.0 liters	0.00 liters
Pb^{2+} solution	2.2 liters	0.00 liters
brightener	4.4 liters	0.00 liters
formalin	0.4 liters	0.50 liters
demineralized water	6.0 liters	2.90 liters
drain off	0.0 liters	0.00 liters
new bath volume	588.00 liters	578.40 liters
total costs:	*f* 111.15	*f* 0.69

Table 2.2: Comparison of replenishing strategies for August 29, 1995.

bath, again a second column is presented where it is allowed to drain off the surplus of the first solution before computing additions. The last column contains the solution of the LP model.

component	results of old method	results of old method*	results of LP-problem
acid	67.2 liters	61.0 liters	32.16 liters
Sn^{2+} solution	80.0 liters	72.5 liters	43.82 liters
Pb^{2+} solution	15.0 liters	13.6 liters	6.76 liters
brightener	4.9 liters	4.5 liters	0.00 liters
formalin	2.0 liters	1.8 liters	0.84 liters
demineralized water	0.0 liters	0.0 liters	0.00 liters
drain off	0.0 liters	44.2 liters	0.00 liters
new bath volume	644.10 liters	584.20 liters	558.60 liters
total costs:	*f* 3149.17	*f* 2961.59	*f* 1032.04

*draining off before making additions allowed

Table 2.3: Comparison of replenishing strategies for September 27, 1995.

The LP solution is approximately 2000 guilders cheaper than the solution of the old method. The cost reduction is caused by a reduction in the use of chemicals. The effect of this is twofold: not only do we have a reduction in the costs of the additions, it also becomes unnecessary to drain off tin-plating liquid, which is very expensive.

2.4.4 The analysis of October 19, 1995

The analysis of the tin-plating bath of the final example indicated that the concentration of Sn^{2+} solution was too low, while the concentration Pb^{2+} solution was too high. The other concentrations fell within their limits. The bath volume was 575 liters, so that there was not much room for making additions. According to table 2.4 both methods appear not to be able to find additions such that draining off can be avoided.

component	**results of old method**	**results of old method***	**results of LP-problem**
acid	24.6 liters	23.1 liters	12.99 liters
Sn^{2+} solution	33.9 liters	31.9 liters	20.44 liters
Pb^{2+} solution	0.0 liters	0.0 liters	0.00 liters
brightener	0.0 liters	0.0 liters	0.00 liters
formalin	0.0 liters	0.0 liters	0.50 liters
demineralized water	0.9 liters	0.9 liters	2.69 liters
drain off	0.0 liters	36.4 liters	11.62 liters
new bath volume	636.44 liters	596.50 liters	600.00 liters
total costs:	*f* 1675.07	*f* 1628.83	*f* 743.03

*draining off before making additions allowed

Table 2.4: Comparison of replenishing strategies for October 19, 1995.

Also in this case, the LP solution results in a considerable cost reduction of replenishment.

2.5 Conclusions

In this chapter, we discussed a problem we encountered at Philips Stadskanaal. The problem concerns a chemical bath, which is a mixture of several components. For each of these components, limits are set on their concentrations in the bath. If all these restrictions are met, it is possible to apply a good tin/lead layer.

However, due to production and evaporation, the volume and the composition of the bath changes. Then the problem arises how to compute additions such that the bath meets its requirements. The method currently used computed far too large additions. We have formulated this problem as a Linear Programming problem, which can be solved by the Simplex Algorithm.

The result is a method that allows Philips Stadskanaal to replenish its bath in such a way that it is fit for production again, *and* that the costs of replenishment are minimal. In fact, it can be shown that no cheaper set of additions can be computed that will result in a bath meeting all requirements.

To illustrate what cost reductions were possible in practice, we compared the method currently used to the LP solution, using data of four days. The total costs of the

method currently used were 13,252.87 guilders if no draining off before adding was allowed, and 11,558.58 guilders if draining off was allowed. In contrast, the total costs of the LP solution were 4,864.09 guilders for these four days. This means that a cost reduction of 63% and 58%, respectively is possible! Philips Stadskanaal has several of such baths, which are replenished on a regular basis.

Cost reduction is not the only advantage of the LP method. It has also environmental implications. In the past, it regularly happened that large additions caused a surplus of tin-plating liquid. This liquid contains the heavy metals tin and lead. Although it is possible to purify the liquid to a certain extend, the limitations on draining off are becoming more and more strict. The solution of the LP model will always try to find a combination of the additions such that a minimum of liquid has to be drained off.

This methodology can be used in all chemical baths where a mixture of components has to be adjusted, so that the concentrations of the components satisfy certain requirements. We have shown how such a problem can be formulated as an LP model. For a solution of such a model it can be proven that it is not possible to find a cheaper solution that makes the bath meet all requirements.

When all the components are separately available, the corresponding constraints follow straightforward. However, if some of the components can only be added by adding a mix of components, as was the case with the Sn^{2+} solution, the constraints can easily be adapted to deal with this situation. If some of the components are separately available, but also in some pre-mixed form (which may be cheaper), the model can consider all of these as possible additions, and choose the cheapest combination.

Depending on the number of components, the model is typically small. With the aid of modern computer equipment such a model is solvable within a second. However, it is possible to gain a little computation time if a standard starting solution for the Simplex Algorithm is used.

Once more, we would like to emphasize that this LP model is only a tool for computing the cheapest way to create a bath that satisfies certain predefined limits. We did not yet discuss how these limits should be set. Often, such limits are set by engineers. Sometimes these limits have a rational foundation, in other situations no one can remember where the limits came from. In the next chapter we will discuss making additions based on statistical arguments.

3 Adjusting the tin-plating bath

As discussed previously, the adjustments to the bath were computed on the basis of a comparison of an analysis of the bath with predefined limits that are set for the concentrations of each of the components. If all of the concentrations fell within their limits, no adjustments were made. If one or more of the concentrations did not meet the requirements, additions were made to bring all of the concentrations within their limits. No target was specified, so that a concentration close to one of the limits was considered to serve equally well as a concentration in the middle of the interval of admissible concentrations. The limits that were used to control the concentrations were *process limits*: during production, the measurements were not supposed to exceed these limits. However, due to controlling the concentrations with these limits, the process was functioning outside its process limits half of the time according to the measurements. Fortunately (but surprisingly), the tin/lead layers did not deteriorate in such situations. This raised the question of how the process limits were set. It turned out to be the result of a combination of prescription by the supplier of the tin-plating chemicals, experience of the operators, and superstition. There did not seem to be a relation between the setting of the limits and the quality of the tin/lead layer.

Furthermore, relative to the variation in the measurements, the limits do not seem to be equally tight for all of the components. This is illustrated by figures 3.1 and 3.2, where we depicted the concentrations of Sn^{2+} solution and Pb^{2+} solution respectively, resulting from consecutive analyses over a three weeks period.

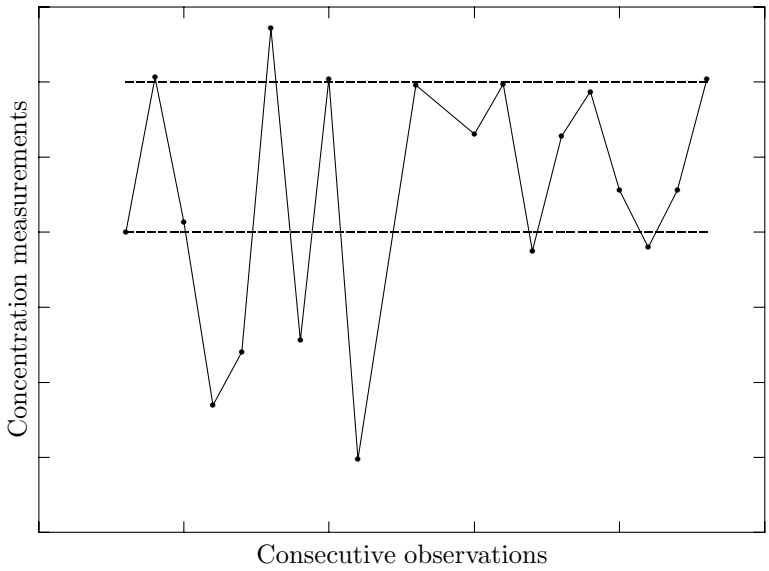

Figure 3.1: The behavior of Sn^{2+} solution in time.

From these figures we see that the process limits for the Sn^{2+} solution are more tight relative to the variation in the measurements than the limits for Pb^{2+} solution. It is to be expected that the concentration of Sn^{2+} solution needs to be adjusted much

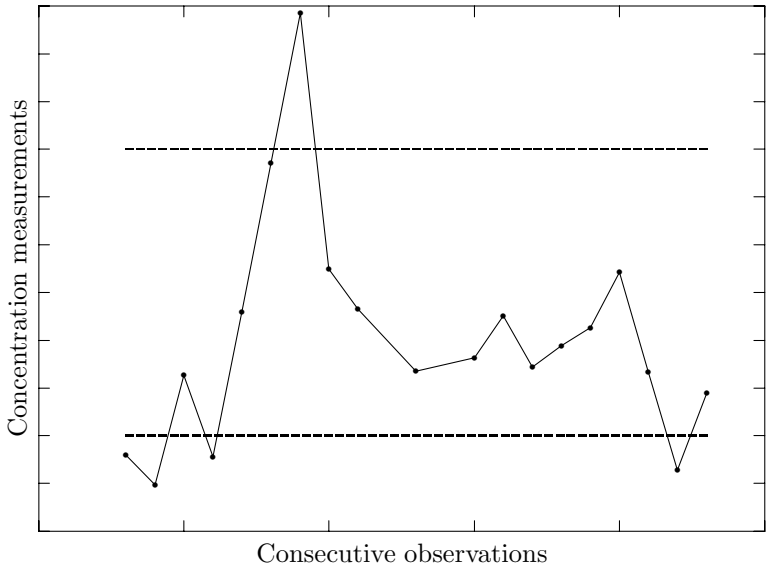

Figure 3.2: The behavior of Pb^{2+} solution in time.

more often than the concentration of Pb^{2+} solution.

Another peculiarity of figures 3.1 and 3.2 is that the the phenomenon of bath exhaustion is not noticeable from these figures. Instead of concentrations jumping up and down more or less randomly, we rather would have expected a trend or some other form of deterministic behavior (possibly depending on replenishments). We return to this subject later (see subsection 5.3.2). For the moment we assume that the need for control actions is indeed present. How these control actions can be performed is discussed in the next two subsections. First we will describe Proportional Integral Derivative (PID) control in section 3.1, thereafter we will discuss minimal mean squared error (MMSE) control in section 3.2.

3.1 PID control

Controllers of the PID type are widely used in process-control applications in industry. We will briefly discuss this type of controllers, and how they can be used together with the LP application that was developed in the previous chapter.

Figure 3.3: Block diagram of a process.

Consider figure 3.3, where we depicted a block diagram of a process. The system is affected by inputs, which are transferred through the process into outputs.

If the outputs are to show some desired behavior, a *controller* may be used to generate inputs that are designed to produce desired process outputs. Some of the

inputs may not be accessible (e.g. disturbances). For a controller it is only possible to regulate the accessible inputs. Figure 3.4 illustrates the function of a controller. Note that a controller may be viewed as a sub-process itself, having desired process output as its input, and controlled inputs as its output.

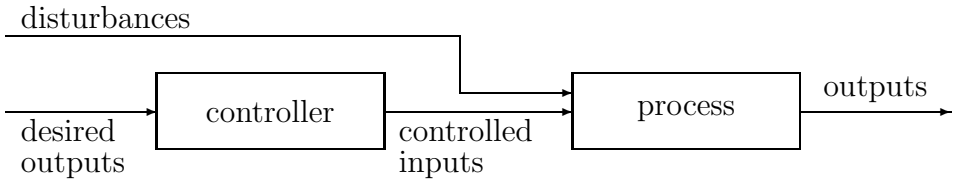

Figure 3.4: An open loop controlled process.

If the inputs of the controller are not influenced by the process outputs, the system is said to be regulated by open *loop control*. If there is feedback around the process, the system is said to have *closed loop* or *feedback* control, see figure 3.5.

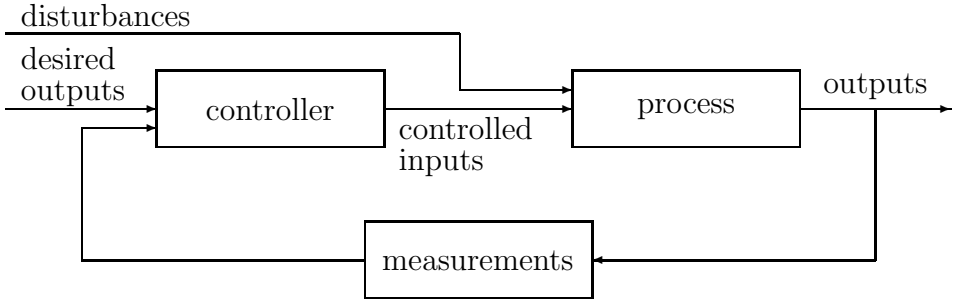

Figure 3.5: A feedback controlled process.

In some cases it is possible to compensate some of the disturbance inputs directly by making appropriate compensatory changes in other input variables. This type of control is referred to as *feedforward* control, and can be used in conjunction with feedback control. In this report however, we will restrict ourselves to feedback control. It is our objective to keep the output (the thickness of the tin/lead layer) as close as possible to a specified target value, and adjust our inputs (the concentrations of the components of the tin-plating bath) to compensate for disturbances in the process. In the literature, this type of problem is known as a *regulator* problem. For illustration purposes we will restrict ourselves to the single input, single output case.

In the following, the output of the process as a continuous function of time is denoted by $y(t)$, whereas the input is written as $x(t)$, and the deviation of $y(t)$ from the target value as $e(t)$, see figure 3.6.

If the controller is of the PID-type, the controlled input of the process is related to $e(t)$ in the following way

$$x(t) = k_P e(t) + k_I \int_{-\infty}^{t} e(s)ds + k_D \frac{d}{dt} e(t),$$

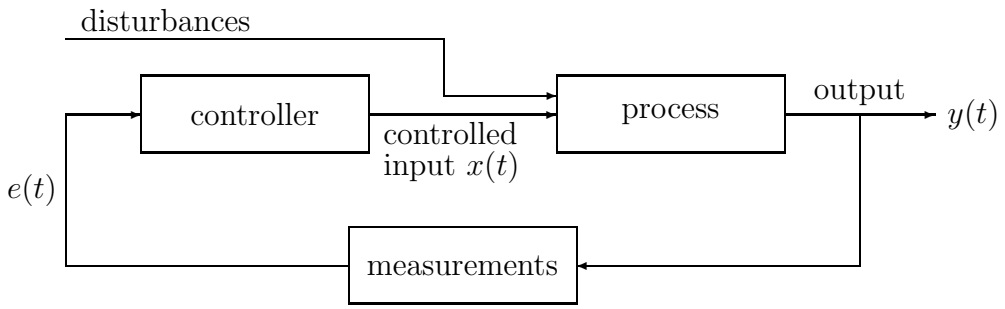

Figure 3.6: A single input, single output feedback controlled process.

for some constants k_P, k_I, and k_D. That is, the control action $x(t)$ is *proportional* to the deviation from the target value, to its *integral*, and to its *derivative*. The constants k_P, k_I, and k_D can be chosen such that the system shows some sort of desired behavior (often in terms of *stability* and *transient behavior*, see Stefani et al. (1982)).

For the controller described above, the assumption is made that measurements and control actions can be taken continuously. In our case, only discrete observations are available, which are approximately equally spaced in time. In Box and Jenkins (1976) the discrete analogue of continuous PID control is discussed. The control action at time t, which we will denote by X_t, is related to present and past deviations from target in the following way

$$X_t = k_P e_t + k_I \sum_{i=1}^{t} e_i + k_D \nabla e_t,$$

where e_t is the deviation of the output from target at time t, and ∇ is the backward difference operator: $\nabla e_t = e_t - e_{t-1}$. The constants k_P and k_I determine the amount of *proportional*, *integral*, *differential* control, respectively.

The PID feedback control schemes were originally developed empirically, and can be shown to have desirable properties in the sense of stability (the response to an impulse input decays asymptotically to zero with time) and convergence to a desired steady state value when different types of input are applied to the system.

It is possible to take a different objective in controlling discrete processes, one that is based on statistical arguments: finding the control equation that minimizes the mean squared deviations of the output from a target value. This type of discrete control is called *minimum mean square error* (MMSE) feedback control and will be discussed in the next section. A more complete discussion of this subject can be found in Box and Jenkins (1976).

3.2 MMSE feedback control schemes

Following Box and Jenkins (1976), we will use a notation that is slightly different from the notation used before. Here Y_t denotes the effect of the past and present

control actions on the output at time t. Furthermore, the joint effect of unobserved disturbances on the output at time t is denoted by N_t. Hence, the net output at time t can be written as $Y_t + N_t$, and will be denoted by e_t. As before, the control action at time t is denoted by X_t.

Suppose that Y_t, X_t, and N_t are defined as deviations from reference values, such that if the conditions $X = 0, N = 0$ were continuously maintained, then the process would remain in an equilibrium state such that the output was exactly on the target value: $Y = 0$. Note that in this situation the sequence $\{N_t\}$ can be interpreted as the behavior of the process if no control actions were applied.

Furthermore, we introduce B, the *backward shift operator* which is defined as $BY_t = Y_{t-1}$, so that $B^i Y_t = Y_{t-i}$. A fairly general class of process dynamics models relating present output to past and present control actions can easily be written down using polynomials in B. Suppose that the process dynamics can be modelled as

$$L_1(B)Y_t = L_2(B)B^{f+1}X_t,$$

where $L_1(B)$ and $L_2(B)$ are polynomials in B. The effect of previous control actions at the output of time t can then be written as

$$Y_t = L_1^{-1}(B)L_2(B)B^{f+1}X_t. \tag{3.1}$$

In words, equation (3.1) states that at time t, the effect on the output of all previous control actions can be written as a weighted sum of control actions before time $t - f$. In this formulation, every control action affects the output with a *lead time* of $f + 1$ units of time.

Suppose that the effect of the disturbances on the output at time t can be modelled as an ARMA model:

$$N_t = \varphi^{-1}(B)\theta(B)a_t, \tag{3.2}$$

where $\varphi(B)$ and $\theta(B)$ are polynomials in B and the sequence $\{a_t\}$ is a *white noise process* (i.e. $a_t \overset{\text{i.i.d}}{\sim} N(\mu, \sigma^2)$ $t = 0, 1, 2, \cdots$). An equivalent formulation of the disturbance process is given by

$$N_t = \psi(B)a_t = \left\{1 + \sum_{i=1}^{\infty} \psi_i B^i\right\} a_t, \tag{3.3}$$

where $\psi(B)$ is a power series in B with coefficients $\psi_1, \psi_2, \cdots$, and $\{a_t\}$ the same white noise process as in formula (3.2).

Note that not only the notation has changed, but that the process model also differs slightly from figure 3.6. The process model and the corresponding notation are illustrated in figure 3.7.

At time t, the effect of the disturbances on the output would be canceled if $Y_t = N_t$. For X_t, the control action at time t, this would mean

$$X_t = -L_1(B)L_2^{-1}(B)N_{t+f+1}. \tag{3.4}$$

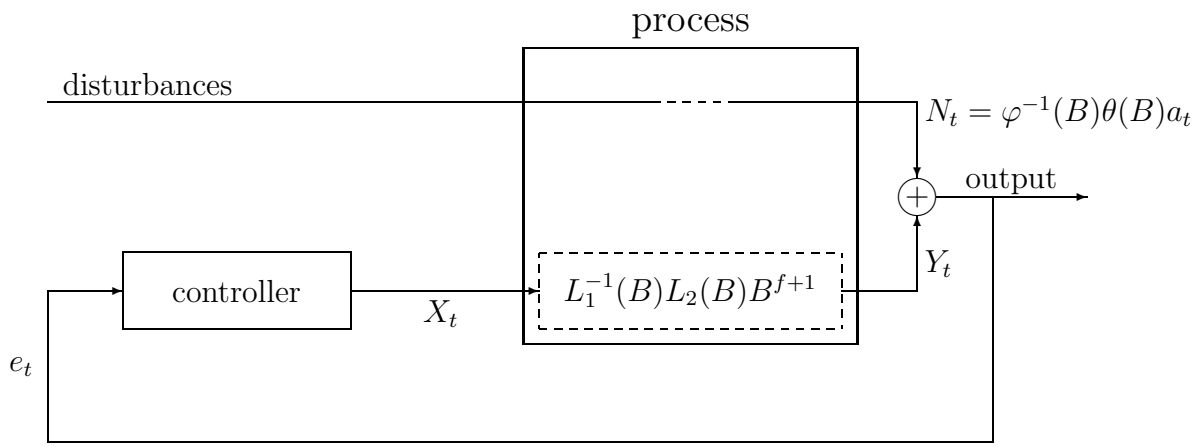

Figure 3.7: A single input, single output feedback controlled process.

However, at time t, only the effects of the disturbances up until time t are known, so that N_{t+f+1} is not known. Minimum square control error can be obtained by replacing N_{t+f+1} by its optimal predictor, the expectation of N_{t+f+1}, conditional on the information available at time t: $\hat{N}_{t+f+1|t} = \mathrm{E}(N_{t+f+1}|a_t, a_{t-1}, a_{t-2}, \cdots)$.

From equation (3.3) we have

$$\begin{aligned}
\hat{N}_{t+f+1|t} &= \mathrm{E}\left(\left\{1 + \sum_{i=1}^{\infty}\psi_i B^i\right\} a_{t+f+1}\,\middle|\, a_t, a_{t-1}, a_{t-2}, \cdots\right) \\
&= \mathrm{E}\left(\{a_{t+f+1} + \psi_1 a_{t+f} + \cdots + \psi_f a_{t+1}\}\,|\, a_t, a_{t-1}, a_{t-2}, \cdots\right) \\
&\qquad + \mathrm{E}\left(\{\psi_{f+1}a_t + \psi_{f+2}a_{t-1} + \cdots\}\,|\, a_t, a_{t-1}, a_{t-2}, \cdots\right) \\
&= \psi_{f+1}a_t + \psi_{f+2}a_{t-1} + \cdots \\
&= L_3(B)a_t,
\end{aligned}$$

where $L_3(B)$ is a power series in B whose coefficients are known if the disturbance model is known.

For the error in the forecast of N_t we can write

$$\begin{aligned}
N_t - \hat{N}_{t|t-f-1} &= a_t + \psi_1 a_{t-1} + \cdots + \psi_f a_{t-f} \\
&= L_4(B)a_t,
\end{aligned}$$

where $L_4(B)$ is a polynomial in B, with known coefficients if the model for the disturbances is known. This forecast error is also the output of the process at time t, hence,

$$e_t = L_4(B)a_t.$$

This implies that

$$\hat{N}_{t+f+1|t} = L_3(B)L_4^{-1}(B)e_t.$$

And we conclude that the control action

$$X_t = -L_1(B)L_2^{-1}(B)L_3(B)L_4^{-1}(B)e_t \tag{3.5}$$

results in minimum mean square error control.

As an example we discuss a slight modification of one of the models considered by Box and Kramer (1992). In this article, a control scheme is sought for controlling Y_t, the level of viscosity, by changing X_t, the level of catalyst. A first order model is used to describe the relation between X_t and Y_t:

$$Y_t = \delta Y_{t-1} + g(1-\delta)X_{t-1}, \qquad \text{with } 0 \leq \delta \leq 1.$$

In terms of the notation used above, in this example we have that $L_1(B) = 1 - \delta B$, $L_2(B) = g(1-\delta)$, and $f = 0$.

The disturbances are modelled with an IMA(1,1) model:

$$N_t = N_{t-1} + a_t - \theta a_{t-1},$$

so that $\varphi(B) = 1 - B$, $\theta(B) = 1 - \theta B$, and

$$\psi(B) = 1 + (1-\theta)B + (1-\theta)B^2 + \cdots.$$

Note that $f = 0$, $L_4(B) = 1$, and

$$\begin{aligned} L_3(B) &= \psi_1 + \psi_2 B + \psi_3 B^2 + \cdots \\ &= (1-\theta)(1-B)^{-1}. \end{aligned}$$

Using these expressions in equation (3.5), we have that the MMSE control action for this example can be written as

$$\begin{aligned} X_t &= -(1-\delta B)\frac{1}{g(1-\delta)}(1-\theta)(1-B)^{-1}e_t \\ &= -\frac{1-\theta}{g(1-\delta)}(1-\delta B)(1+B+B^2+\cdots)e_t \\ &= -\frac{1-\theta}{g(1-\delta)}\left(1+(1-\delta)\sum_{i=1}^{\infty}B^i\right)e_t \\ &= -\frac{1-\theta}{g(1-\delta)}e_t + \frac{1-\theta}{g}\sum_{i=1}^{\infty}e_{t-i}. \end{aligned}$$

It follows that for these specific model assumptions, MMSE control reduces to *proportional-integral* (PI) control with constants

$$k_P = -\frac{1-\theta}{g(1-\delta)} \quad \text{and} \quad k_I = \frac{1-\theta}{g}.$$

Box and Jenkins (1976) find that many simple situations lead to control equations containing discrete analogues of proportional, integral, or derivative control terms. However, not all control actions called for by (3.5) can be produced by some form of PID control.

In figure 3.8 the effect of the preceding control scheme on a simulation of IMA(1,1) observations is illustrated. The uncompensated simulations are driven by iid simulations from N(0,1). The subsequent IMA(1,1) observations are then computed as follows:

$$N_t = a_t + (1-\theta)\sum_{i=1}^{t} a_t. \tag{3.6}$$

We used $\theta = 0.8$ as value for the MA parameter.

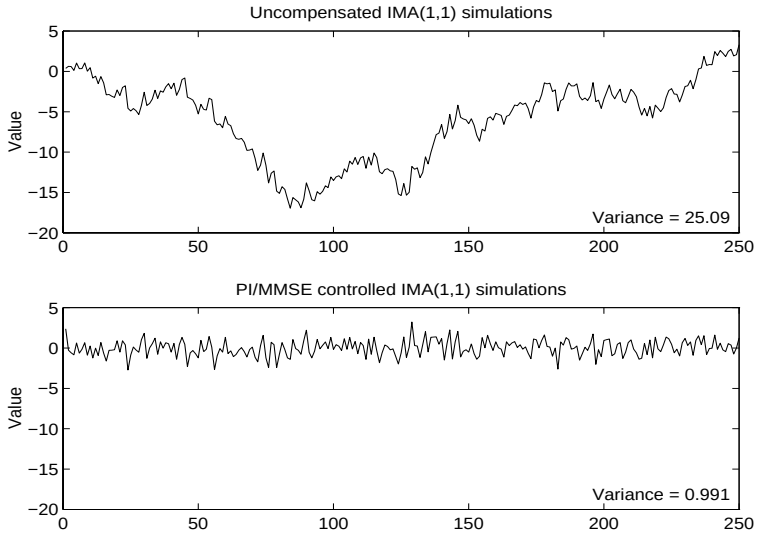

Figure 3.8: The effect of PI/MSSE control on IMA(1,1) observations.

Note that in figure 3.8 also variance computations are presented. They are the result of computing the sample variance

$$S_N^2 = \frac{1}{n-1}\sum_{i=1}^{n}(N_i - \bar{N})^2$$

for the uncompensated series, where $\bar{N} = 1/n\sum_{i=1}^{n} N_i$, and n is the number of simulations. The variance of the compensated simulations is computed analogously.

However, an IMA(1,1) process is not stationary and has infinite variance. The variance computation of the uncompensated simulation is therefore not to be interpreted as an estimator of the variance, since it can be shown to have expectation

$$\mathrm{E}(S_N^2) = \left(\frac{(1-\theta)^2}{2} \left(\frac{n^2 - 19}{3(n-1)} \right) + 1 \right) \sigma_a^2,$$

where the sequence $\{N_1, N_2, \cdots N_n\}$ is generated by (3.6), and σ_a^2 is the variance of the white noise process. For our choice of parameters we have $\mathrm{E}(S_N^2) = 27.766$. The realization of S_N^2 is 25.089. Again, this number is not to be interpreted as an estimate of the variance of N_t, but merely facilitates the illustration of the reduction in variability brought about by the controller.

3.3 Conclusion

In the preceding sections, we have discussed PID control and MMSE control, a multivariate extension of which can be used to complement the LP procedure, developed in chapter 2.

The concentrations of the bath components will serve as controllable inputs, and the thickness and the composition of the tin/lead layer are the characteristics we want to control. At each time point t, the control equation (a multivariate extension of (3.5)) computes desired values for the concentrations, such that the variation in the thickness and the composition of the tin/lead layer is minimal. Thereafter, the LP procedure can be employed to calculate the cheapest combination of additions, so that these new values are attained.

In order to be able to implement such a control strategy, a relation between the inputs and the outputs must be identified and estimated. Also, the behavior of the disturbances needs to be modelled. This will be the subject of the next chapter.

4 Implementing the control strategy

In the previous chapter we discussed the theoretical framework of the control strategy that can be used to replenish the tin-plating bath. In this chapter we will discuss the practical implementation of this strategy. To this end, will determine the dynamic relation between the measurements of the contents of the bath and the product measurements.

4.1 The relation between process inputs and outputs

Based on discussions with chemical engineers, we expect to find that some of the inputs are correlated with some of the outputs of the process. Such relations are the basis of our control strategy. In this section, we will verify whether these supposed relations are reflected in input and output measurements. For that purpose, a data set of 106 observations of bath analyses and corresponding product measurements is available. These are daily consecutive observations taken in the last five months of 1995.

The bath analyses are approximately equally spaced, and have values for the concentration acid, Sn^{2+} solution, Pb^{2+} solution and brightener. Recall from section 1.4 that the concentration of brightener can only be roughly determined, while the concentration of formalin is not measured at all. Furthermore, the volume of the bath is available. These measurements are the inputs, the thickness and composition of the tin/lead layers are the outputs of the process. The product measurements are taken from the last produced batch before analyzing the bath.

On investigating several (dynamic) relations between inputs and outputs of the tin-plating process, it became clear that the presumed mechanics of the process were not reflected in the data. To be more precise, we were not able to find any significant relation between inputs and outputs that had a practical importance. Note that the lack of correlation between process inputs and outputs could be suspected from chapter 3, where we observed that the bath was frequently operating outside its process limits, while the tin/lead layers were satisfactorily. The way the quality of the tin/lead layers was controlled up until now hinged on the relation between bath and product measurements. Since such a relation was not traceable in the data it is not surprisingly that the quality of the output was sometimes disappointing.

As an illustration of the lack of correlation, we depicted in figure 4.1 a scatter plot of the ratio of the Sn^{2+} /Pb^{2+} concentrations in the bath, against the Sn^{2+} /Pb^{2+} ratios of the product measurements. From this picture, it becomes clear why controlling the ratio of Sn^{2+} and Pb^{2+} concentrations on the product by controlling Sn^{2+} /Pb^{2+} in the bath is not a good idea.

After investigating several (lagged) relationships between inputs and outputs, suggested by the chemical reactions taking place in the bath, we reached the conclusion that attempting to control the quality characteristics of tin/lead layers on the basis of the bath measurements is asking for trouble. This conclusion meant a breakthrough in our quality improvement attempts, since it opened the way for other means of

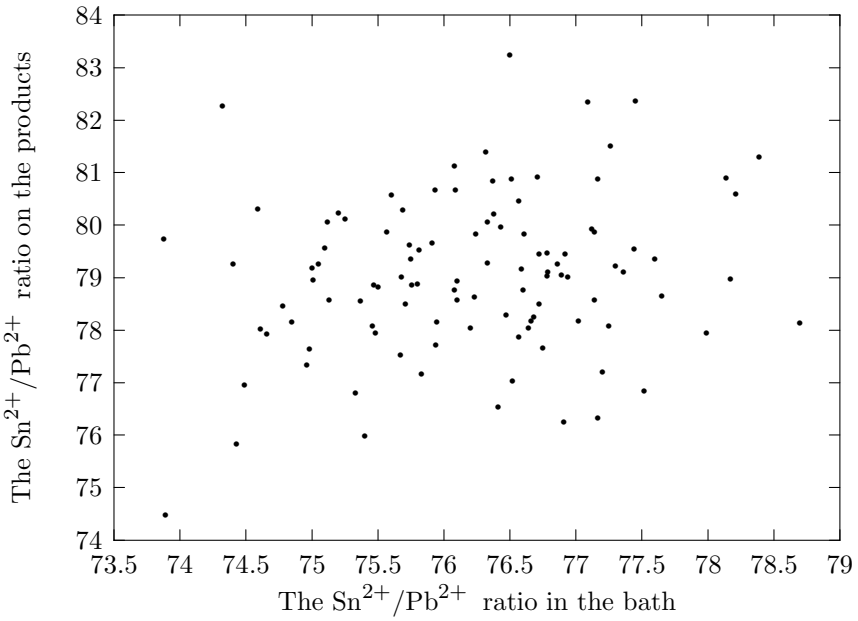

Figure 4.1: The relation between Sn^{2+}/Pb^{2+} in the bath and on the product.

controlling the output of the process.

Before developing a new control mechanism, it had to be understood what causes the lack of correlation between inputs and outputs. To this end we did two things: we informed ourselves of the accuracy of the measurements, and we returned to the chemical experts with graphs similar to figure 4.1, and brainstormed about a possible explanation.

4.2 Measurement errors

Based on the methodology described in Does, Roes and Trip (1996), or Banens et al. (1994), an R&R (Repeatability & Reproducibility) study for the product measurements had already been performed. It turned out that the 'gage R&R' was 1.120 for the thickness of the tin/lead layers, while the 'gage R&R' was 2.346 for the concentration measurements. The 'gage R&R' is computed as 5.15×(total measurement variance), and under the assumption of normally distributed observations it is the width of a 99% confidence interval. To judge the accuracy of the measurement device, the 'gage R&R' is compared to the width of the specification limits. In our case, the product measurements were sufficiently accurate.

Considering the bath measurements, we reached another conclusion. To determine the concentrations of the bath components, operators were to take exactly 1 ml of bath liquid using a pipette, whereafter an automized titration device determined how much (in mol) of each component was present in the liquid. The results of these measurements were imported into an online computer, which divided these numbers by 1 ml and presented the outcomes as concentration measurements. Clearly, the accuracy of this procedure is largely determined by the ability of the operators to

pipette exactly 1 ml of bath liquid. To investigate this, operators were asked to pipette at least five times 1 ml of demineralized water. To determine the pipetted quantity, it was weighted on a very accurate pair of scales. The results for 21 operators are depicted in figure 4.2.

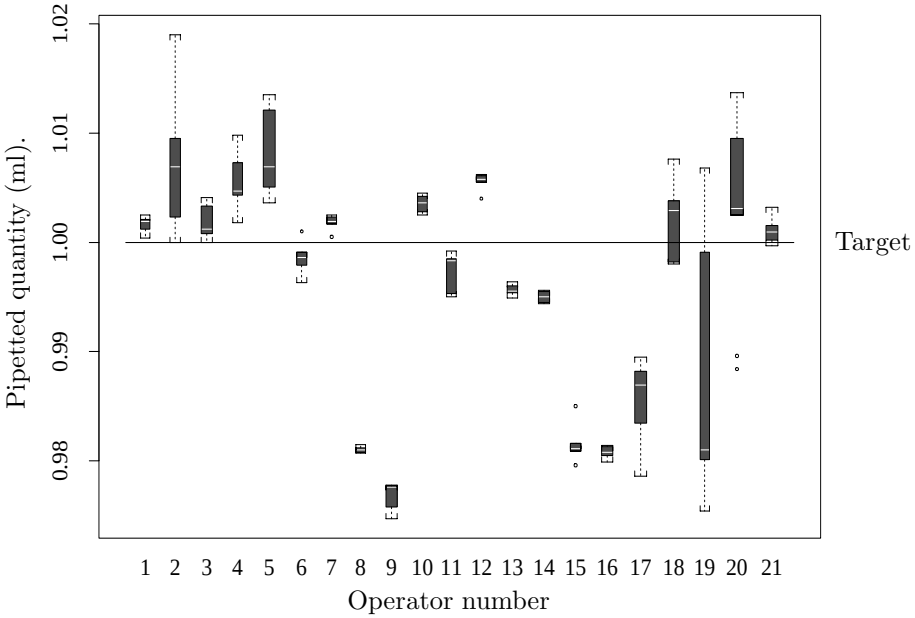

Figure 4.2: Box plots of pipetting results.

In the Box plots of figure 4.2 the median of the observations per operator is indicated by a white line in the box. The length of the boxes indicates the interquartile range, and the connected hooks (if present) show the minimum and the maximum values. Observations falling outside the range of 1.5×(interquartile range) are considered as outliers, and are plotted as disconnected dots. The width of the boxes is proportional to the square root of the number of observations per operator. The target value of 1 ml is indicated by a solid line.

Figure 4.2 shows that there is considerable variation between operators. Often, the target value of 1 ml is not even in the range of the observations. Some operators are able to take the amount of water with high precision, but in most of these cases the level is off target. In other cases the range is quite large, which is not always the result of a larger number of observations.

These observations were taken in a laboratory at the end of an operator course. It is therefore to be expected that the results on the work floor will be less accurate. Furthermore, figure 4.2 only shows the variation due to pipetting. There is also additional variation, for example variation due to measurement error of the titration device. Hence, figure 4.2 only gives a flattering idea of the total variation present in the analysis measurements. However, if we use the results of this experiment, and assume that not water but bath liquid was pipetted, and that the true concentration values were all exactly on target (i.e. in the middle of specification limits), then we are able to compute what the effect of the pipetting inaccuracy would have been on

the analysis results. For the concentration of Sn^{2+}, this is depicted in figure 4.3.

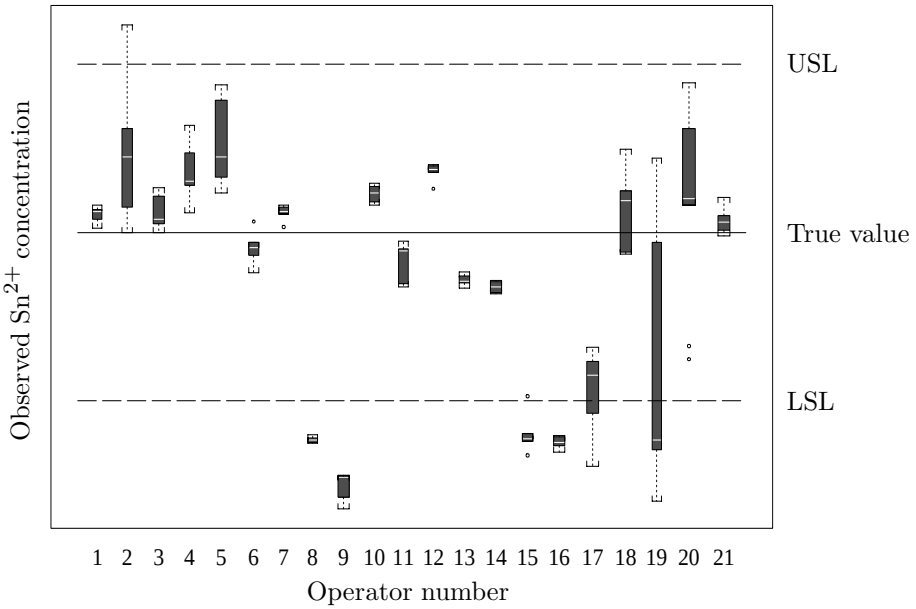

Figure 4.3: Box plots of Sn^{2+} measurement inaccuracies.

From figure 4.3 we may conclude that the measurement error due to pipetting is unacceptably high. The analysis results of a bath whose true concentration values are exactly in the middle of the specification limits (an ideal bath) can lead to the absurd conclusion that the bath is not fit for production. We must conclude that the analyses tell us more about the large measurement error than about the contents of the bath. And hence, if we are controlling on the basis of analysis results, we are mainly controlling on the basis of measurement error.

4.3 Process mechanics

In our search for an explanation for the lack of correlation between observed process inputs and process outputs, we also discovered that the assumptions concerning how the inputs affect the outputs were not entirely correct. Since such assumptions form the basis for any control strategy, this gives another explanation for the malfunctioning of the current replenishment strategy.

Remember from section 1.3 that the tin-plating bath consists of five components: acid, an Sn^{2+} solution, a Pb^{2+} solution, brightener, and formalin. The current strategy calls for action if one or more of the concentrations of Sn^{2+} solution, Pb^{2+} solution, or acid is out of range. These three parameters are thought to have a great impact on the composition and the thickness of the tin/lead layers. The concentration of brightener is only very roughly determined, while it is not possible to determine the concentration of formalin. Hence, controlling basically takes place on the first three parameters. However, in brainstorming sessions together with chemical engineers, it turned out that the thickness and the composition of the layer was mainly determined

by the two other parameters. Brightener not only takes care of a shiny tin/lead layer, it also influences the thickness of the layer. The concentration of formalin is an important determinant of the composition of the tin/lead layers.

The other components are not unimportant: the acid takes care of a good conductivity of the bath, while the concentrations of the Sn^{2+} and Pb^{2+} solutions must be large enough to ensure that the chemical reactions as described in section 1.3 can take place. However, none of these does directly influence the output quality characteristics.

Furthermore, it was assumed that due to the primary chemical reactions in production, the concentrations of Sn^{2+} and Pb^{2+} solution would change. If we take another look at the chemical reactions, this does not make sense. For every Sn^{2+} or Pb^{2+} ion that precipitates on a diode, another one dissolves from the anode (see figure 1.2). So, apart from secondary chemical reactions, there is no reason to assume that the concentration of Sn^{2+} solution or the concentration of Pb^{2+} solution will change with the volume of production.

4.4 Conclusion

From the discussion in this chapter, it is easy do deduce why the existing control strategy did not lead to the desired results. Firstly, control actions took place on the basis of relative unimportant process parameters. And secondly, the measurements of these parameters are highly inaccurate, so that the control actions are not based on process information, but rather on noise. The result of this is that the control actions which were meant to stabilize the bath, cause the bath to destabilize. This type of overacting is called *tampering*, and is a well known phenomenon in the field of statistical quality control. It is one of the main arguments for using SPC techniques. In the next chapter we will see that it is possible to obtain a more stable process by simply not reacting on input measurements.

As a consequence of the lack of an input-output relationship, a control strategy of the type discussed in chapter 3, was not applicable with these input and output measurements. In the next chapter, we will propose a control strategy, based on leaving the process alone except in situations when there is statistical evidence of the presence of an additional, abnormal source of variation.

5 Replenishing the tin-plating bath

In the last chapter we discussed the control strategy that was used up until now. We reached the conclusion that this strategy, due to *tampering*, is most likely to have a destabilizing effect on the bath. Furthermore, we argued that from the chemical reactions discussed in section 1.3, it is not clear why replenishments are necessary at all.

5.1 A new control strategy

Therefore, we experimented with producing without replenishments (except for formalin for reasons of quick evaporation, and brightener). Not to our surprise, the variation in the measurements of the tin/lead layers reduced. During the period of not replenishing, we closely monitored the product measurements to see whether this strategy had the expected effect. Details can be found in sections 5.2 and 5.3 . After four weeks of applying the new strategy, we saw that one of the quality parameters, the composition of the tin/lead layer, was drifting away. From a reliable analysis of the bath, performed by the laboratory, the level of acid and Sn^{2+} solution turned out to be too low for a stable process. Very likely, this was due to secondary chemical reactions in the bath.

Based on the experiment described above, we dropped the daily analysis by the operators and devised a new replenishment strategy, based on *not* replenishing except in situations where there is statistical evidence of something abnormal affecting the process (and except for daily formalin and brightener additions). However, the experiment showed that four weeks without replenishing was too long.

Once a week the bath is filtered to clean the tin-plating liquid from precipitations, caused by secondary reactions. This was a natural moment for a weekly reliable bath analysis, performed by a laboratory employee. On the basis of this analysis, appropriate replenishments can be determined with the aid of the LP model, discussed in chapter 2, to make sure that the concentrations are within their limits.

During the week, the product measurements are monitored with the aid of a control chart, to check whether there is evidence of an additional, abnormal source of variation. How this is done will be discussed in the following sections.

5.2 Monitoring the process

In figures 5.1 and 5.2 the product measurements of the period March 4, 1996 through March 31, 1996 are depicted. This was the period wherein it was decided to experiment with producing without replenishing.

Figure 5.1 shows the successive means of thickness measurements, and figure 5.2 shows the successive means of Sn^{2+} /Pb^{2+} ratios. Remember that a sample of size 10 is taken from every batch of approximately 31,000 diodes.

Despite the agreement not to make additions, it can be suspected from both figures that around observation 300 something unusual has happened to the process. Both

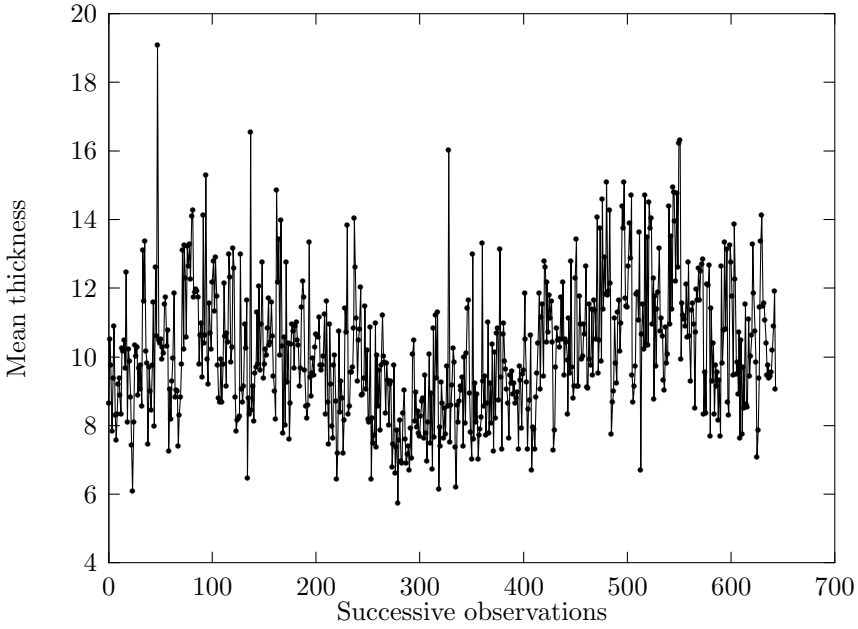

Figure 5.1: Successive mean thickness observations.

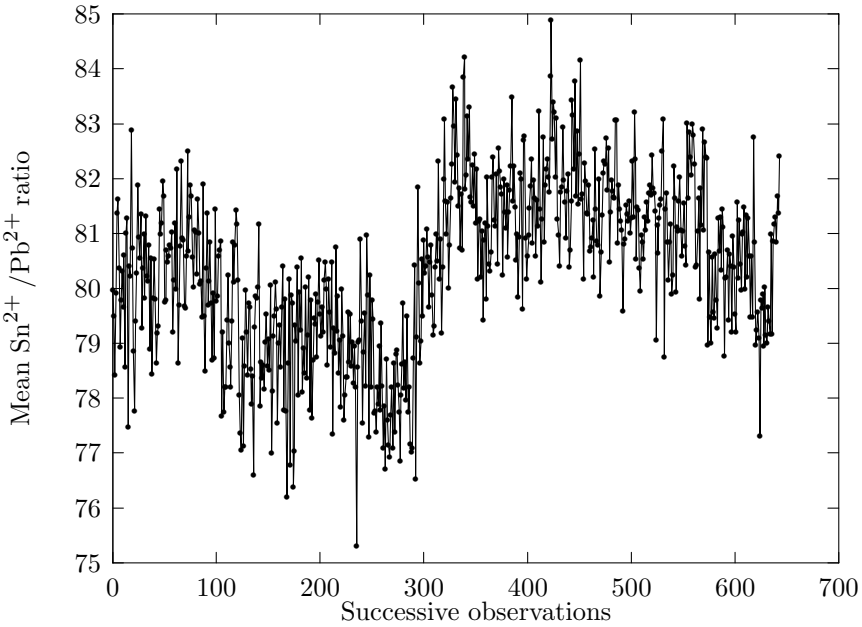

Figure 5.2: Successive mean Sn^{2+} /Pb^{2+} ratio observations.

the layer thickness and composition values seem to be affected by some special cause of variation.

Figure 5.2 also shows some other interesting behavior. The mean composition values are on average slowly decreasing. This trend is only interrupted by the same sudden jump upwards around observation 300. We will discuss these phenomena in more detail in subsection 5.3.2.

Figure 5.3 shows a scatter plot of layer thickness and composition. No strong relationship seems present. The correlation coefficient between thickness and composition measurements can be computed as 0.21. We therefore decided to monitor the process with two univariate control charts.

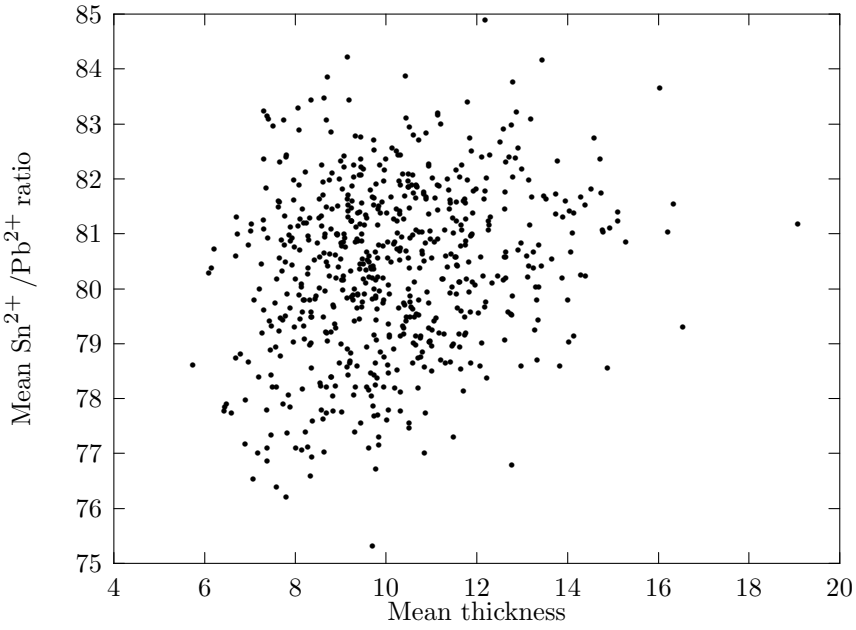

Figure 5.3: Scatter plot of mean Sn^{2+} /Pb^{2+} ratios and mean layer thickness.

The application of standard control charts hinges on two important assumptions: normality and independence of successive observations. Since serial correlation distorts various tools for checking normality of the data (such as normal probability plots), the correlation structure of the data is explored first, and the exploration of normality is deferred to subsection 5.3.1. A useful tool to explore the correlation structure of the data is the sample autocorrelation function. Figures 5.4 and 5.5 depict the sample autocorrelation functions for mean layer thickness, and mean layer composition, respectively.

It can be shown (see e.g. Anderson (1971)) that a fixed number of sample autocorrelation coefficients of white noise are asymptotically normally and independently distributed with zero means and standard deviations equal to $1/\sqrt{N}$, where N is the number of observations, in our case equal to 644. This can be used to judge the sample autocorrelation coefficients.

The dashed lines in figures 5.4 and 5.5 are drawn at $1.96 \times 1/\sqrt{N}$ as an approximate 95% confidence interval for individual sample autocorrelations with expectation

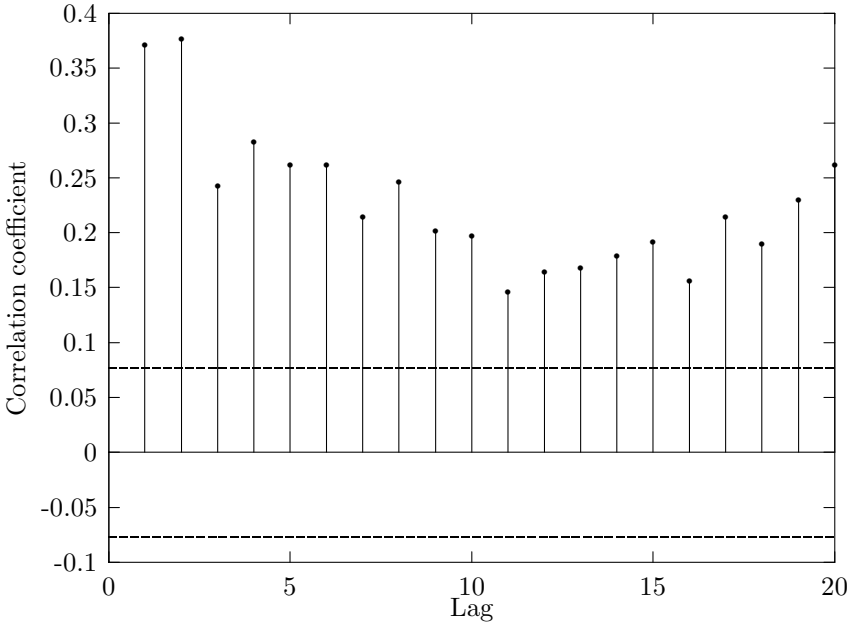

Figure 5.4: Autocorrelation function of mean thickness values.

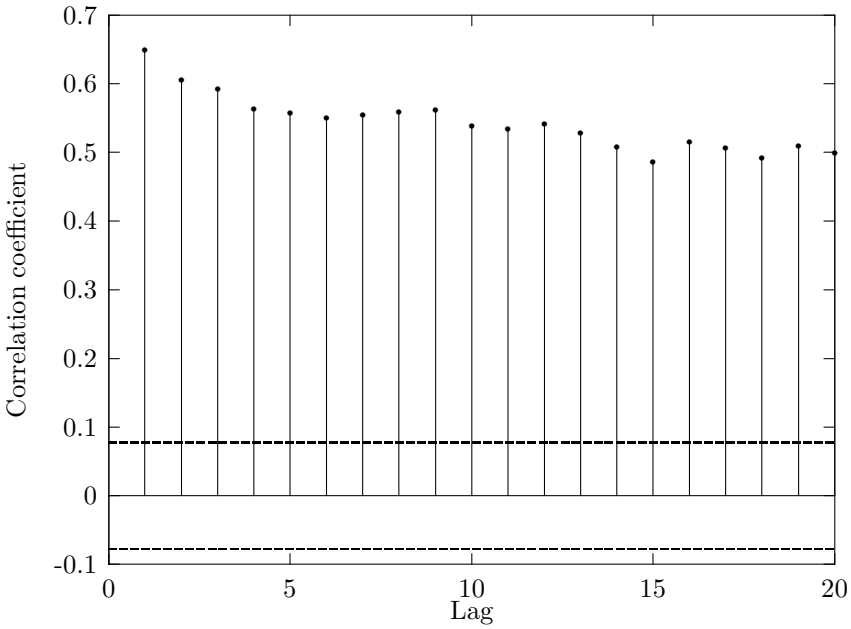

Figure 5.5: Autocorrelation function of mean composition values.

zero. Note that since we are plotting a number of autocorrelation coefficients at once, we expect to find on average one out of twenty outside these limits if the observations are uncorrelated! In addition, if the covariance between successive observations is nonzero, the sample autocorrelations are also correlated. These phenomena may seriously distort the interpretation of a sample autocorrelation function, and conclusions based upon this graph must be drawn with care.

Notwithstanding the above remarks, the conclusion from figures 5.4 and 5.5 is that both layer thickness and layer composition are highly autocorrelated. Moreover, both autocorrelation functions suggest nonstationarity. When constructing control charts, we must take these findings into account. In the next section we will discuss why this is necessary.

5.3 Control charts and correlated data

First we will ignore the serial correlation and set up a control chart for the mean of mean layer composition in the usual way. That is, assuming independence and normality of the observations.

Let us denote the observed mean composition value on time t by $\bar{y}_t$. Remember that means are taken over samples of size $n = 10$. Trial control limits with tail probabilities of $\frac{1}{2}\alpha$ are determined as

$$\text{LCL} = \bar{\bar{y}} + \Phi^{-1}(\tfrac{1}{2}\alpha)\frac{\bar{s}}{c_4(n)\sqrt{n}},$$

and

$$\text{UCL} = \bar{\bar{y}} + \Phi^{-1}(1 - \tfrac{1}{2}\alpha)\frac{\bar{s}}{c_4(n)\sqrt{n}},$$

where LCL and UCL stand for *Lower Control Limit* and *Upper Control Limit*, respectively, and $\bar{\bar{y}}$ is the overall mean of $N = 644$ observations of $\bar{y}_t$ ($t = 1, \cdots, 644$). Furthermore, $\Phi^{-1}(\frac{1}{2}\alpha)$ is the $\frac{1}{2}\alpha$-th percentile of the normal distribution, $\bar{s}$ is the mean of the sample standard deviations $s_1, \cdots s_N$, and $c_4(n)$ is a constant such that $\bar{s}/c_4(n)$ is an unbiased estimator of the standard deviation of the individual composition observations. From table E in Ryan (1989), we find that $c_4(10) = 0.9727$. The computed control limits for $\alpha = 0.002$ are depicted in figure 5.6.

Figure 5.6 shows a lot of out of control signals. This can be explained as follows. A Shewhart control chart for the mean assumes that observations of the quality characteristic under consideration can be viewed as realizations of independently distributed random variables. Furthermore, the corresponding distribution functions are assumed to be identical, except for a possible shift in location. To check whether such a shift has occurred, an estimator for location (usually the mean of a sample) is monitored in time. Due to sample variation, a certain amount of variation in the value of the estimator must be allowed for. However, if this value exceeds a certain upper or lower

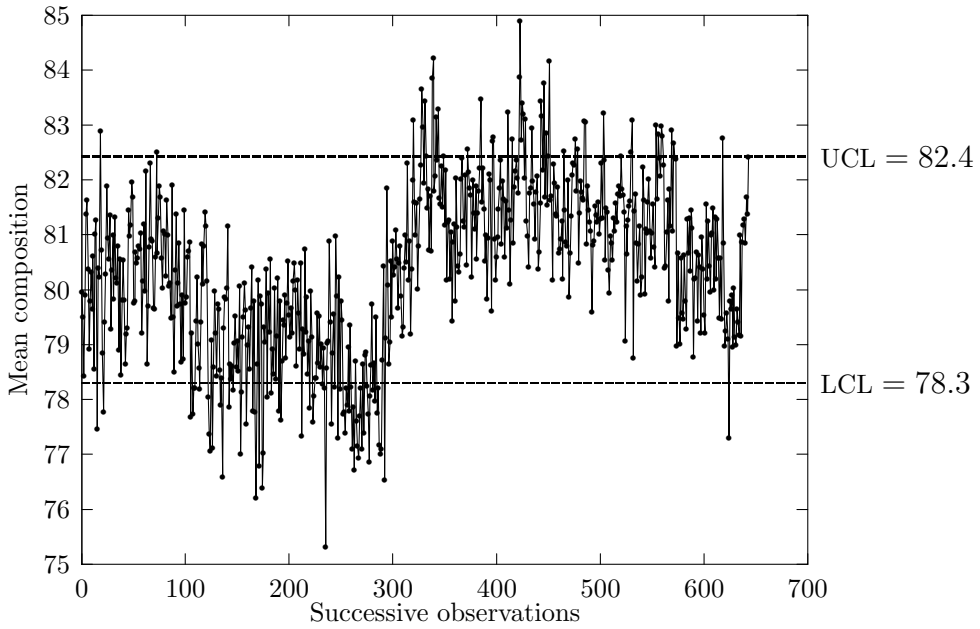

Figure 5.6: Trial control limits for monitoring the mean of composition.

bound (control limits), it becomes very unlikely that all observations are realizations of random variables with identical distributions.

In general, a shift in the mean will increase the probability of observing an out of control signal. As we will see in the next subsection, an IMA(1,1) model is the most appropriate model for our data in the class of ARIMA(p, d, q) models. A feature of an IMA(1,1) process is that its level is changing constantly. It is therefore to be expected that, in testing for stability of the mean, a lot of out of control signals will be generated by such a data set.

A vast number of articles has appeared on the subject of monitoring a process with correlated observations. Roughly speaking, two ways of monitoring correlated observations have become prevalent in the treatment of correlated measurements in quality control schemes.

The first approach is to model the correlation structure in the data with an appropriate time series model. Fitting an adequate model to the data should remove serial correlation in the sense that the resulting residuals should not exhibit significant serial correlation. The standard control charts can then be used to monitor the process by its residuals.

In the second approach an *exponential weighted moving average* (EWMA) is used as a one-step-ahead forecast, and the process is monitored by control charts applied to forecast errors.

In the following subsection we find that an IMA(1,1) model is useful for our data. In subsection 5.3.2 it is shown that the optimal one step ahead forecast of IMA(1,1) observations is an EWMA of previous observations, so that the two approaches coincide in our case.

Fitting of a time series model can be viewed as recursively computing one step ahead forecasts. For this reason, for the remainder of this report, no sharp distinction

will be made between the terms *fitted value* and *one step ahead forecasts* . The same holds for *residual* and *forecast error* . Furthermore, the terms *prediction* and *forecasting* will be used interchangeably.

5.3.1 Modelling the data

The autocorrelation function of figure 5.5 shows considerable autocorrelation which seems to decrease slowly and linearly with the length of the lags, an indication for the presence of a unit root. Box and Jenkins (1976) explain why this is so.

Let a sequence of random variables $\{z_t\}$ follow a stationary ARMA(p,q) model. Assume that $\mathrm{E}(z_t) = 0$. Then we have that z_t can be written as

$$z_t = \phi_1 z_{t-1} + \cdots + \phi_p z_{t-p} + a_t - \theta_1 a_{t-1} - \theta_q a_{t-q}, \tag{5.1}$$

where $\{a_t\}$ is a *white noise* process. When B, the backward shift operator is used, (5.1) can be written as

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) z_t = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q) a_t,$$

or

$$\phi(B) z_t = \theta(B) a_t,$$

where $\phi(B)$ and $\theta(B)$ are polynomials in B of order p and q, respectively.

Multiplying both sides of equation (5.1) by z_{t-k} and taking expectations, we find that

$$\gamma_k = \phi_1 \gamma_{k-1} + \phi_2 \gamma_{k-2} + \cdots + \phi_p \gamma_{k-p} \qquad \text{for} k \geq q+1, \tag{5.2}$$

where γ_k is the covariance between z_t and z_{t-k}. For $k \geq q+1$ the covariance between z_t and a_{t-k} is zero. Dividing both sides of equation (5.2) by γ_0, the variance of z_t, this results in

$$\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} + \cdots + \phi_p \rho_{k-p} \qquad \text{for} k \geq q+1, \tag{5.3}$$

where ρ_k is the autocorrelation coefficient of z_t and z_{t-k}. And hence, the autocorrelation function satisfies the difference equation

$$\phi(B) \rho_k = 0 \qquad \text{for} k \geq q+1. \tag{5.4}$$

Furthermore, if it is assumed that $\phi(B) = \prod_{i=1}^{p} (1 - G_i B)$, with $G_1, \cdots, G_p$ distinct, so that $1/G_1, \cdots, 1/G_p$ are distinct roots of $\phi(B)$, then ρ_k is of the form

$$\rho_k = A_1 G_1^k + A_2 G_2^k + \cdots + A_p G_p^k \qquad \text{for} k \geq q+1.$$

If one or more of the roots of $\phi(B)$ approaches one, say $G_1 = 1 - \delta$, then

$$\rho_k \approx A_1(1 - k\delta),$$

and the autocorrelation function will not die out quickly. Instead, it will be linearly decreasing in k.

On the other hand, for a model to be stationary, it is required that the roots of $\phi(B)$ must lie outside the unit circle, so that $G_1, \cdots, G_p$ must lie inside the unit circle. Hence, if none of the roots of $\phi(B)$ is close to one, ρ_k will damp out quickly.

Hence, figure 5.5 shows the typical behavior of a nonstationary process, and we decide to take first differences of the mean composition values. The autocorrelation function of the new series will give insight in the correlation structure of the first differences. If taking first differences does not seem to remove nonstationarity, we take first differences once more, and so on until the series displays stationary behavior. The sample autocorrelation function of first differences of the mean composition values is depicted in figure 5.7.

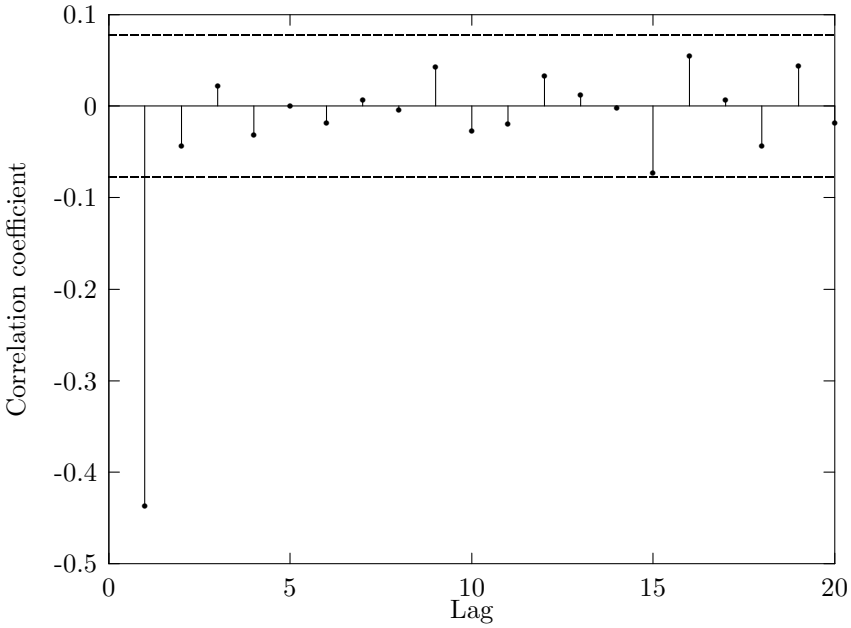

Figure 5.7: Autocorrelation function of differenced mean composition values.

The autocorrelation function of figure 5.7 shows a single spike at lag 1, indicating a *moving average* (MA) component of order 1 in the first differences. This can be seen from the theoretical autocorrelation function of a first order MA process. For a sequence of random variables $\{z_t\}$, generated by a MA(1) model, i.e.

$$z_t = \varepsilon_t - \theta\varepsilon_{t-1},$$

with $\{\varepsilon_t\}$ *white noise*, we have for the first order correlation coefficient

$$\rho(1) = \frac{\mathrm{E}(z_t z_{t-1})}{\mathrm{V}(z_t)} = \frac{-\theta}{1+\theta^2},$$

whereas $\rho(2) = \rho(3) = \cdots = 0$. Furthermore, a first order MA process can also be viewed as an infinite order *autoregressive* (AR) process with coefficients whose values decrease exponentially in absolute value:

$$\begin{aligned} & z_t &=& \varepsilon_t - \theta\varepsilon_{t-1} &=& (1-\theta B)\varepsilon_t \\ \iff & (1-\theta B)^{-1} z_t &=& \varepsilon_t & & \\ \iff & z_t &=& \varepsilon_t - \theta z_{t-1} - \theta^2 z_{t-2} - \theta^3 z_{t-3} + \cdots, & & \end{aligned}$$

so that the theoretical partial correlation function of a first order MA process is exponentially declining in absolute value. So, if a MA(1) model is useful for modelling the first differences, we expect to see such behavior in the corresponding partial autocorrelation function. The observed partial autocorrelation function of first differences of the mean composition values of figure 5.8 does indeed show such behavior.

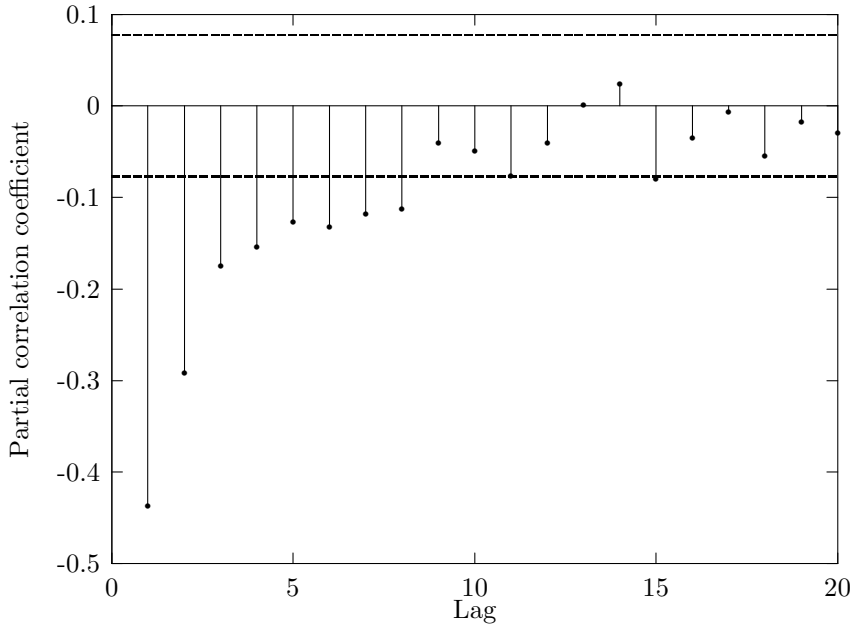

Figure 5.8: Partial autocorrelation function of differenced mean composition.

Hence, based on the behavior of the (partial) autocorrelation functions of the data, we decide to select an IMA(1,1) model in the class of ARIMA(p, d, q) models. The maximum likelihood estimate of the moving average coefficient was computed with the aid of SPLUS as $\hat{\theta} = 0.811$, with a standard error of 0.023, so that we have as a model for the series of mean composition values which, as previously, is denoted by $\{\bar{y}_t\}$:

$$\bar{y}_t = \bar{y}_{t-1} + \varepsilon_t - \underset{(0.023)}{0.811}\ \varepsilon_{t-1}. \tag{5.5}$$

When model (5.5) is fitted to the data, the sample autocorrelation function of the residuals, as depicted in figure 5.9 may be used to check whether there is still some remaining serial correlation present in the residuals.

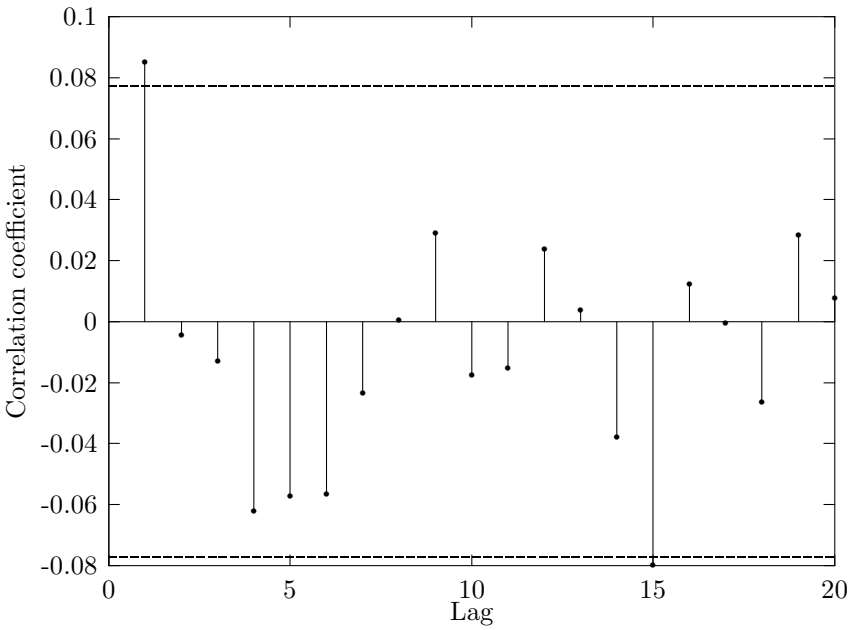

Figure 5.9: Autocorrelation function of residuals.

A warning regarding the interpretation of the residual sample autocorrelation function must be made at this point. Granger and Newbold (1986) cite references from which it follows that the asymptotic standard deviations of sample autocorrelation coefficients of *residuals* may be smaller than $1/\sqrt{N}$. Hence, the sample autocorrelation coefficients of figure 5.9 need careful interpretation. Nevertheless, the lines drawn in this figure at $1.96 \times 1/\sqrt{N}$ can provide a crude check for model adequacy. These limits are exceeded only just at lag 1 and 15.

However, the autocorrelation coefficients are small, especially when compared to the autocorrelation function of figure 5.5, so that the correlation is considerably reduced.

Also, evaluation of the (modified) Portmanteau test statistic at various lags does not lead us to reject the hypotheses of zero correlations. Hence, for this moment, we proceed as if the residuals are uncorrelated.

In section 5.2, we deferred testing for normality because of the presence of serial correlation. Since the residuals of model (5.5) behave more like an uncorrelated sequence than the original observations, its normal probability plot is more reliable for judging normality than a normal probability plot of the original observations. In figure 5.10 a normal probability plot of the residuals of model (5.5) is shown.

The normal probability plot does not indicate deviations of normality. Also more formal tests do not reject the hypotheses that the residuals are normally distributed. For example, a test based on observed kurtosis and skewness of the empirical distribution function as described by Harvey (1993), p. 45 results in a p-value of 0.330.

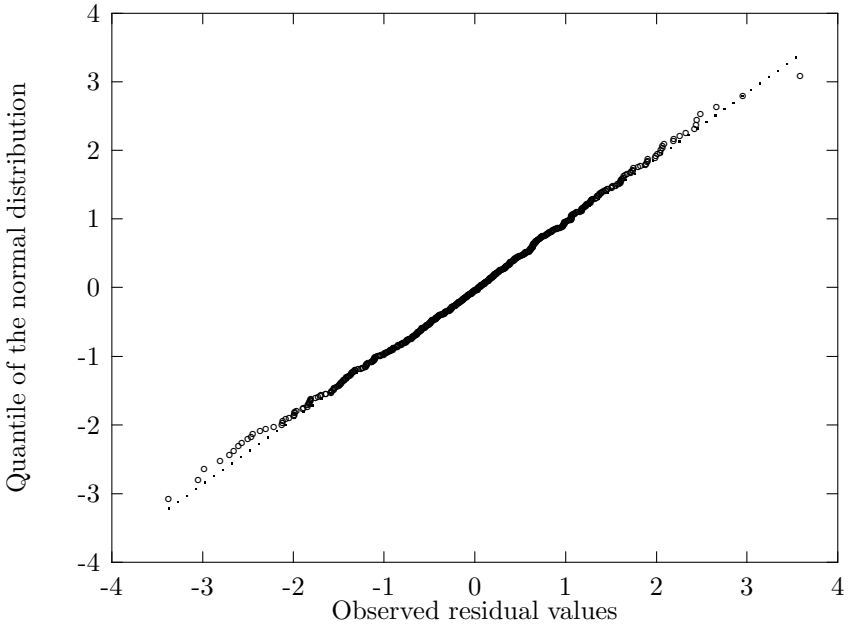

Figure 5.10: Normal probability of residuals.

For the remainder of this report we assume that model (5.5) is the correct model for our data. The residuals of the fitted model behave more or less as a sequence of uncorrelated normally distributed variates. Hence, both conditions discussed on page 37 are fulfilled, and the standard control charts can be used to monitor the residuals. As we will see, not only the residuals contain information about changes in the process. It is wise to monitor the fitted values as well. In the next subsection we will discuss how the results of the data analysis can provide us with information about changes in the process.

5.3.2 Monitoring fitted values and residuals

In the previous subsection, we modelled the mean composition observations using an IMA(1,1) model:

$$\bar{y}_t = \bar{y}_{t-1} + \varepsilon_t - \theta\varepsilon_{t-1}, \tag{5.6}$$

where the maximum likelihood estimate of the MA parameter turned out to be $\hat{\theta} = 0.811$. Using the backward shift operator B notation in (5.6), we can write the IMA(1,1) model as

$$(1 - B)\bar{y}_t = (1 - \theta B)\varepsilon_t,$$

or

$$(1 - \theta B)^{-1}(1 - B)\bar{y}_t = \varepsilon_t,$$

so that we can write for $\bar{y}_t$

$$\bar{y}_t = (1-\theta)\sum_{j=1}^{\infty}\theta^{j-1}\bar{y}_{t-j} + \varepsilon_t. \tag{5.7}$$

For $\hat{\bar{y}}_t$, the optimal one-step ahead forecast of $\bar{y}_t$, we write

$$\begin{aligned}\hat{\bar{y}}_t &= \mathrm{E}(\bar{y}_t|\bar{y}_{t-1}, \bar{y}_{t-2}, \bar{y}_{t-3}, \cdots) \\ &= (1-\theta)\sum_{j=1}^{\infty}\theta^{j-1}\bar{y}_{t-j}, \end{aligned} \tag{5.8}$$

so that the optimal forecast of $\bar{y}_t$ is an *exponentially weighted moving average* (EWMA) of previous observations. The forecast can be easily updated as new observations become available, since from (5.8) we find that $\hat{\bar{y}}_t$ can be written as

$$\hat{\bar{y}}_t = \theta\hat{\bar{y}}_{t-1} + (1-\theta)\bar{y}_{t-1}.$$

A plot of the EWMA of the mean composition values is presented in figure 5.11.

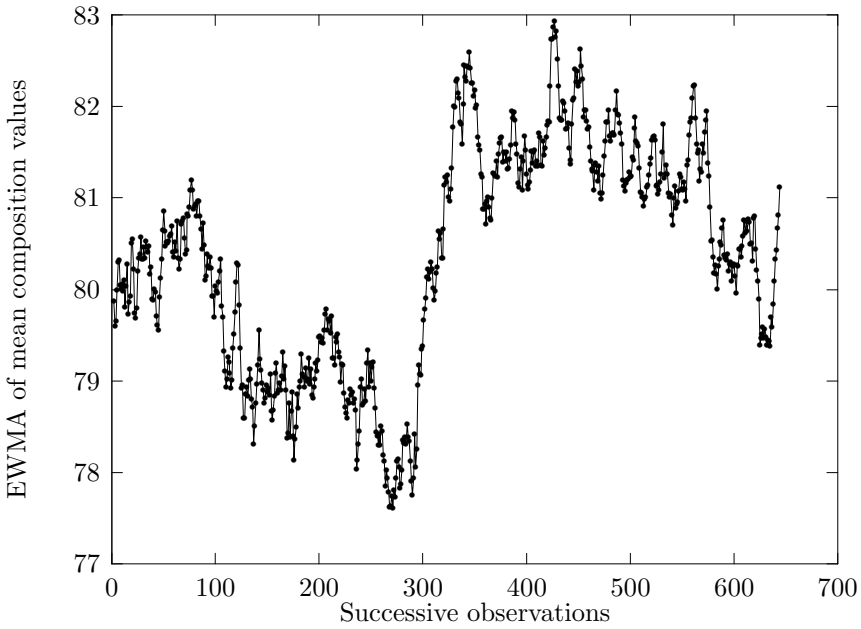

Figure 5.11: EWMA of mean composition values.

Note that figure 5.11 is a smoothened version of figure 5.2. With the short-term variation removed, figure 5.11 shows even more clearly the two phenomena already described in section 5.2.

As was remarked there, the first peculiarity that catches the eye is the sudden jump upwards around observation 300. The second one is the slow downward movement on the left and on the right of this jump.

The explanation for the sudden jump upwards was found in the log book that is kept by the operators. In spite of the agreement not to react on bath measurements,

low levels of Sn^{2+} solution that were reported by bath analyses around observation 300 worried the operators. Combined with the low level of composition, it was decided to add 100 liters of new tin-plating liquid, 6.5 liters of formalin, and 6 liters of brightener. Since no immediate improvement was observed, four more liters of formalin, twenty more liters of Sn^{2+} solution and 1 more liter of brightener were added the same day. Eventually, there was some reaction which is clearly visible in figure 5.11, but also can be observed in figures 5.1 and 5.2.

The foregoing explanation may also give a hint for explaining the slow downward trend in the composition values. Due to secondary chemical reactions, it is possible that dissolved Sn^{2+} ions form Sn^{4+} ions and precipitate. This may cause a slowly decreasing Sn^{2+} concentration in the bath, with a likewise effect on the composition observations on the products. However, since with the current process, it is not possible to obtain accurate measurements of these bath parameters on a regular basis, this hypothesis cannot be verified. On the other hand, adding highly concentrated Sn^{2+} solution to the bath breaks the downward trend, and may therefore be considered as an indication for a too low Sn^{2+} concentration. For a complete understanding of the process mechanism, further research into this is needed.

In the database of process measurements, it was the first time that we were able to observe a long term downward movement of the mean composition values. Previously, such movements were distorted by short term movements, induced by making additions to the bath. Knowledge of the cause of this trend may prove to be very useful. It may for example lead to constant addition of highly concentrated Sn^{2+} solution to the bath, to compensate for the downward trend. Thereby stabilizing the production conditions, and reducing the variation in the product measurements.

More generally, the graph of figure 5.11 shows the long term movement of the level of the process, free of short term disturbances. If it is known how to influence the level of the process, such a graph may form the basis of a control strategy. In addition, Box and Kramer (1992) have shown that control actions triggered by an EWMA crossing certain boundaries are a cost efficient way to control IMA(1,1) processes.

Alwan and Roberts (1988) call such a graph a *common cause* graph, since it shows the variation due to causes of variation that are inherent to the process. In situations with uncorrelated observations, a quality characteristic x_t is assumed to be generated by the following model

$$x_t = \mu + \varepsilon_t,$$

with $\{\varepsilon_t\}$ a white noise process. A graph of the one step ahead forecasts would be a constant, an estimate of μ. Therefore, in such cases, the common cause graph is not plotted.

In the present case, the behavior of the data is such that the assumption of a stationary mean cannot be maintained. The process generates observations with a wandering level, also in cases where only common causes of variation are affecting the process. Therefore, a model was fitted that allows for this type of behavior. A graph of the level of the process now contains nontrivial information, namely how the

process is affected by common causes. If the process is not allowed to wander too far from a certain target value, some kind of control must be applied to assure that product measurements do not fall outside specification limits.

However, like in situations with uncorrelated observations, special causes of variation can influence the output of the process.

Usually, in the uncorrelated case, this is narrowed down to a persistent shock in the mean level, or a persistent shock in the dispersion of the process. Suppose that the mean of the process shifts from μ to $\mu+\delta$ on time $T+1$. The mean of deviations from the (estimated) model then shifts from 0 to δ, which will lead to a higher probability of crossing one of the control limits of a control chart for the mean.

In the case of IMA(1,1) observations, an impulse shock to the level of the process by an amount of δ results in a persistent change in the level of subsequent realizations by the same amount. To see this, compare two IMA(1,1) processes that are exactly equal for $t = 1, 2, \cdots, T$. That is, we have a sequence $\{y_t\}$ generated by

$$y_t = y_{t-1} + \varepsilon_t - \theta\varepsilon_{t-1} \tag{5.9}$$

for $t = 1, 2, \cdots, T$, and a sequence $\{z_t\}$ that is generated exactly the same way for $t = 1, 2, \cdots, T$

$$z_t = z_{t-1} + \varepsilon_t - \theta\varepsilon_{t-1}. \tag{5.10}$$

Assume that θ and y_0 are known, and $\varepsilon_0 = 0$. Then it is possible for $t = 1, 2, \cdots$ to compute disturbance ε_t as soon as observation y_t becomes available.

Suppose that at time $T + 1$, an impulse shock of size δ is applied to y_{T+1}, so that

$$y_{T+1} = y_T + \varepsilon_{T+1} - \theta\varepsilon_T + \delta$$

while for z_{T+1} we have

$$z_{T+1} = z_T + \varepsilon_{T+1} - \theta\varepsilon_T$$

so that $y_{T+1} = z_{T+1} + \delta$. For y_{T+2} we have

$$\begin{aligned} y_{T+2} &= y_{T+1} + \varepsilon_{T+2} - \theta\varepsilon_{T+1} \\ &= y_T + \varepsilon_{T+1} - \theta\varepsilon_T + \delta + \varepsilon_{T+2} - \theta\varepsilon_{T+1}, \end{aligned}$$

while for z_{T+2} we have

$$\begin{aligned} z_{T+2} &= z_{T+1} + \varepsilon_{T+2} - \theta\varepsilon_{T+1} \\ &= z_T + \varepsilon_{T+1} - \theta\varepsilon_T + \varepsilon_{T+2} - \theta\varepsilon_{T+1}, \end{aligned}$$

so that $y_{T+2} = z_{T+2} + \delta$ since $y_t = z_t$ for $t = 1, 2, \cdots, T$. With induction it is not difficult to show that $y_{T+k} = z_{T+k} + \delta$ for every $k = 1, 2, \cdots$. If the forecasts errors e_t are computed as

$$e_t = y_t - y_{t-1} + \theta\varepsilon_{t-1},$$

we find that

$$\begin{aligned} e_T &= y_T - y_{T-1} + \theta\varepsilon_{T-1} \\ &= z_T - z_{T-1} + \theta\varepsilon_{T-1} = \varepsilon_T. \end{aligned}$$

For e_{T+1}, the forecast error at the time the shift occurs, we have

$$\begin{aligned} e_{T+1} &= y_{T+1} - y_T + \theta\varepsilon_T \\ &= z_{T+1} + \delta - z_T + \theta\varepsilon_T = \varepsilon_{T+1} + \delta. \end{aligned}$$

One period after the shock to the process we find as error

$$\begin{aligned} e_{T+2} &= y_{T+2} - y_{T+1} + \theta\varepsilon_{T+1} \\ &= z_{T+2} + \delta - z_{T+1} - \delta + \theta\varepsilon_{T+1} = \varepsilon_{T+2}. \end{aligned}$$

Hence, a persistent change in the level of an IMA(1,1) process is only once encountered in the residuals. Detecting a shock in the level of an IMA(1,1) process by monitoring its residuals is therefore doomed from the start, unless the shock is so large that the control limits are reached in one observation.

This is not much to our surprise since we chose the IMA(1,1) model for its ability to capture the nonstationarity of the mean of the observations. As a consequence, the model allows the mean of the process to wander. Detecting a change in the mean by looking at the residuals is then asking for trouble.

However, a persistent trend in the observations of an IMA(1,1) process *will* be reflected in the residuals. Consider again two IMA(1,1) processes $\{y_t\}$ and $\{z_t\}$ that are exactly equal for $t = 1, 2, \cdots, T$ as in (5.9) and (5.10). Assume now that y_t is generated as

$$y_t = y_{t-1} + \varepsilon_t - \theta\varepsilon_{t-1} + \delta$$

for $t = T+1, T+2, \cdots$, while z_t continues to be generated by (5.10) for $t = T+1, T+2, \cdots$. We then find that $y_{T+1+k} = z_{T+1+k} + k\delta$, so that y_t behaves as an IMA(1,1) process with a deterministic trend added.

For the one step ahead forecast errors we find at time T

$$\begin{aligned} e_T &= y_T - y_{T-1} + \theta\varepsilon_{T-1} \\ &= z_T - z_{T-1} + \theta\varepsilon_{T-1} = \varepsilon_T. \end{aligned}$$

For e_{T+1}, the error at the time the trend starts, we have

$$\begin{aligned} e_{T+1} &= y_{T+1} - y_T + \theta\varepsilon_T \\ &= z_{T+1} + \delta - z_T + \theta\varepsilon_T = \varepsilon_{T+1} + \delta. \end{aligned}$$

On time $T+2$ we find that

$$\begin{aligned} e_{T+2} &= y_{T+2} - y_{T+1} + \theta\varepsilon_{T+1} \\ &= z_{T+2} + 2\delta - z_{T+1} - \delta + \theta\varepsilon_{T+1} = \varepsilon_{T+2} + \delta. \end{aligned}$$

Repeating this argument shows that a trend applied to IMA(1,1) observations is encountered as a persistent change in the level of the residuals.

The errors, made in the EWMA forecasts of mean composition values, are depicted in figure 5.12. The control limits are computed with equal tail probabilities of 0.001.

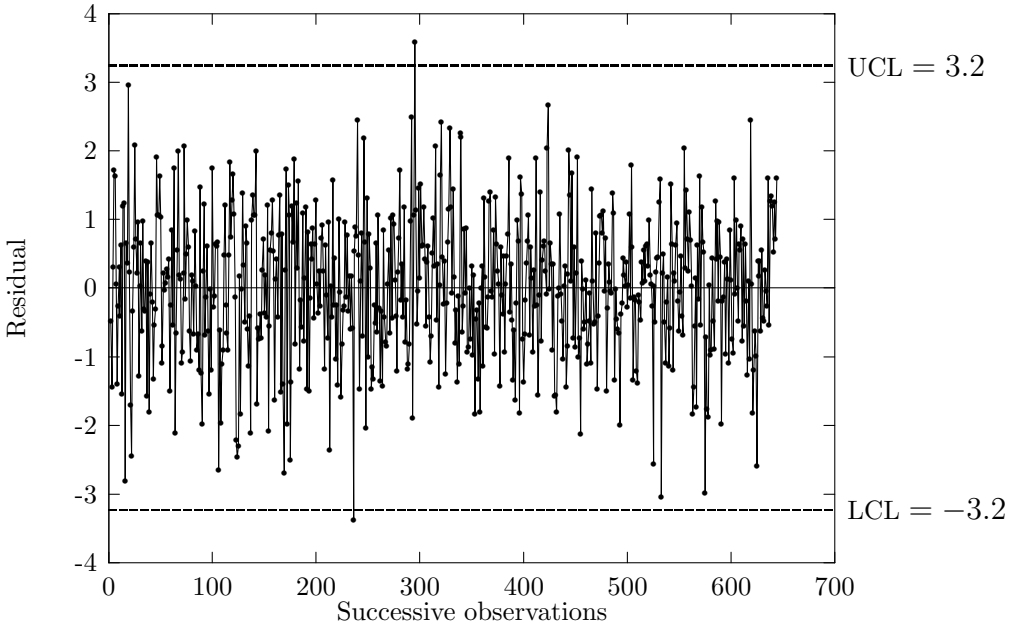

Figure 5.12: Residuals of one step ahead forecasts.

Since we observed two persistent downward trends and one upward trend in figure 5.2, we would, in light of the previous arguments, expect to see some changes in the mean level of the residuals of figure 5.12.

And indeed, there is an out of control signal that could be linked to a positive shift in the mean level of the residuals due to the trend around observation 300. Also, an out of control signal indicating a possible negative shift in the mean level of the residuals is observed. However, figure 5.12 does not show convincingly that there is something out of the ordinary happening to the process.

The latter can be explained as follows. If we make rough estimates of the trends from the raw data, we find that for the first downward trend the level drops approximately three units in 300 observations, resulting in a decline of about 0.01 units per observation. The upward trend raises the level about six units in 60 observations, resulting in an increase of about 0.1 units per observation. The last downwards trend can also be roughly estimated as a decline of 0.01 units per observation. Hence, in figure 5.12, we expect to see a negative shift in the level of the residuals of about 0.01 for the two

downwards trends, and a positive shift in the mean of 0.1 due to the upward trend. Since the sample standard deviation of the residuals equals $\hat{\sigma}_\varepsilon = 1.0479$, the change in the level of the residuals induced by the downwards trends is approximately $-0.01\sigma_\varepsilon$, and the change in the level induced by the upward trend equals approximately $0.1\sigma_\varepsilon$. Since these shifts are small relative to the standard error of the residuals, a regular Shewhart control chart is not the proper tool to detect such a shift.

Suppose that control limits are established for i.i.d. drawings from $N(\mu, \sigma^2)$ with tail probabilities $\frac{1}{2}\alpha$ as

$$\text{LCL} = \mu + \Phi^{-1}(\tfrac{1}{2}\alpha)\,\sigma$$

and

$$\text{UCL} = \mu + \Phi^{-1}(1 - \tfrac{1}{2}\alpha)\,\sigma.$$

Suppose that a shift in μ of size $\delta\sigma$ has occurred. Then we can write for $\text{P}(\delta)$, the probability that X, a drawing from the shifted distribution, will fall between the control limits

$$\begin{aligned}
\text{P}(\delta) &= \text{P}(X < \text{UCL}) - \text{P}(X < \text{LCL}) \\
&= \text{P}\left[\frac{X - \mu - \delta\sigma}{\sigma} < \Phi^{-1}(1 - \tfrac{1}{2}\alpha) - \delta\right] \\
&\qquad - \text{P}\left[\frac{X - \mu - \delta\sigma}{\sigma} < \Phi^{-1}(\tfrac{1}{2}\alpha) - \delta\right] \\
&= \Phi\Big(-\Phi^{-1}(\tfrac{1}{2}\alpha) - \delta\Big) - \Phi\Big(\Phi^{-1}(\tfrac{1}{2}\alpha) - \delta\Big)\,.
\end{aligned}$$

An out of control signal is observed at the first observation with probability $1 - \text{P}(\delta)$. The probability that an out of control signal is firstly observed at the second observation is $\text{P}(\delta)(1 - \text{P}(\delta))$, and the probability that the run length is k equals $\text{P}(\delta)^{k-1}\,[1 - \text{P}(\delta)]$. Hence, we have for the *average run length* (ARL)

$$\begin{aligned}
\text{ARL}(\delta) &= \sum_{i=1}^{\infty} i\text{P}(\delta)^{i-1}\,[1 - \text{P}(\delta)] \\
&= [1 - \text{P}(\delta)] \sum_{i=1}^{\infty} i\text{P}(\delta)^{i-1} \\
&= [1 - \text{P}(\delta)]\,\frac{1}{[1 - \text{P}(\delta)]^2} \\
&= \frac{1}{1 - \text{P}(\delta)}.
\end{aligned}$$

The ARL equals 499.74 if $\delta = -0.01$ and $\alpha = 0.002$, so that on average 500 observations are needed before a shift of the size of the slow downward shifts is detected for the first time with a Shewhart chart such as figure 5.12. The value of ARL(0.1)

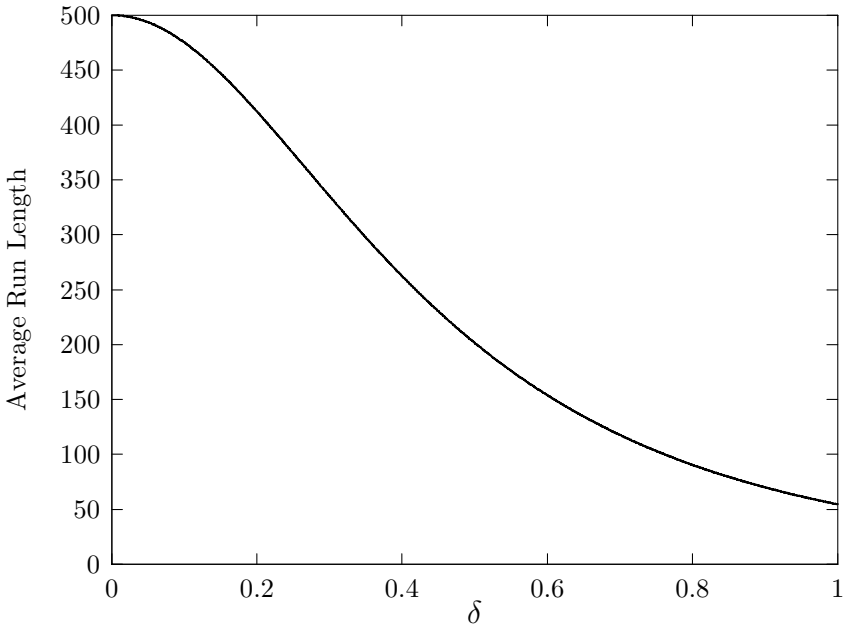

Figure 5.13: ARL curve of a Shewhart control chart ($\alpha = 0.002$).

equals 475.15 if $\alpha = 0.002$ so that detecting a shift of the size of the upward shift takes on average about 475 observations. The ARL curve for $\alpha = 0.002$ is depicted in figure 5.13.

The computations above illustrate the well known disadvantage of the Shewhart control chart that it is not very efficient in detecting small shifts in the mean. With additional run rules or warning limits its performance for detecting small shifts can be improved (see Does and Schriever (1992)). However, there are alternatives such as the EWMA and the CUSUM control chart. These control charts are more efficient in detecting small shifts in the mean.

The most common way to judge a CUSUM chart is using a *decision interval* (see for example Montgomery (1996)). Both the CUSUM for detecting positive shifts in the mean

$$S_{H_i} = \max\left[0, S_{H_{i-1}} + z_i - k\right]$$

(where z_i is the standardized observation at time i, and k a design parameter) and the CUSUM for detecting negative shifts

$$S_{L_i} = \max\left[0, S_{L_{i-1}} - z_i - k\right]$$

are compared to some threshold value h. One of the sums exceeding h is an indication for an out of control situation. The values of k and h can be chosen such that the two sided CUSUM is as efficient as possible (in terms of the ARL) in detecting a predescribed shift in the mean, while maintaining a certain value of the in-control ARL. In figure 5.14 it is illustrated how k can be chosen to design the CUSUM for efficiently detecting a shift of size $0.1\sigma_\varepsilon$, while maintaining an in-control ARL of 500.

For each $k \in [0, 1]$ we computed what value of h is needed to maintain an in-control ARL of 500. Subsequently, for this choice of h and k the ARL for $\delta = 0.1\sigma_\varepsilon$ was computed. In figure 5.14 these ARL(0.1) values are depicted against k.

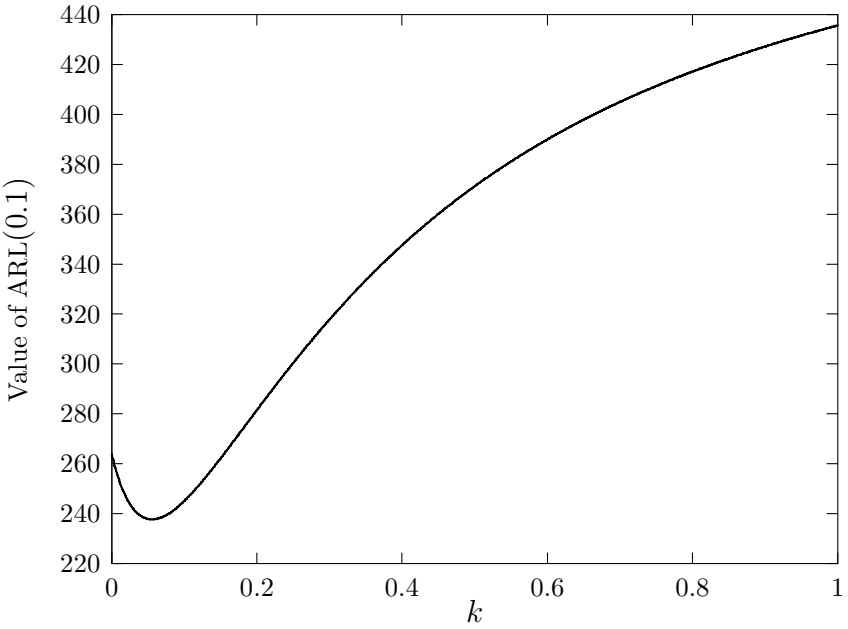

Figure 5.14: ARL(0.1) as a function of k, with h such that ARL(0)=500.

The out-of control ARL is minimized for $k = 0.055$. The corresponding value of h equals 19.025, and ARL(0.1) = 237.7. Hence, this CUSUM chart will on average detect a shift of $\delta = 0.1\sigma_\varepsilon$ twice as fast as a Shewhart chart. Note that this value of k agrees reasonably well with the general recommendation to set k equal to half the size of the shift we want to detect. In figure 5.15, the ARL curve of a CUSUM chart with $k = 0.055$ and $h = 19.025$ is depicted.

The ARL values of the CUSUM are evaluated using the recursive integrals approach, see for example Lucas and Crosier (1982).

Comparing figure 5.15 with figure 5.13 shows that the ARL of the CUSUM is also smaller than the ARL of the Shewhart chart for other small δ. However, numerical calculations show that the ARL curves of the two control charts intersect at $\delta = 1.7\sigma_\varepsilon$. The corresponding ARL value is 12.26. For shifts in the mean larger than $1.7\sigma_\varepsilon$ the ARL of the Shewhart chart is smaller than the ARL of the CUSUM chart. The value of $\text{ARL}_{\text{CUSUM}}$-$\text{ARL}_{\text{Shewhart}}$ is maximized for $\delta = 2.8\sigma_\varepsilon$. The ARL of the Shewhart chart then equals 2.6, while the ARL of the CUSUM equals 7.5. For larger values of δ this difference reduces to zero since both ARL curves converge to one.

Hence, designing the CUSUM chart so that it is as sensitive as possible for detecting small shifts of size $0.1\sigma_\varepsilon$ in the mean results in smaller sensitivity for larger shifts as compared to the Shewhart chart.

The CUSUM chart for the mean composition values is depicted in figure 5.16. An out of control signal on the high side is observed at observation 331, indicating the upward trend mentioned earlier. This control chart shows more convincingly than

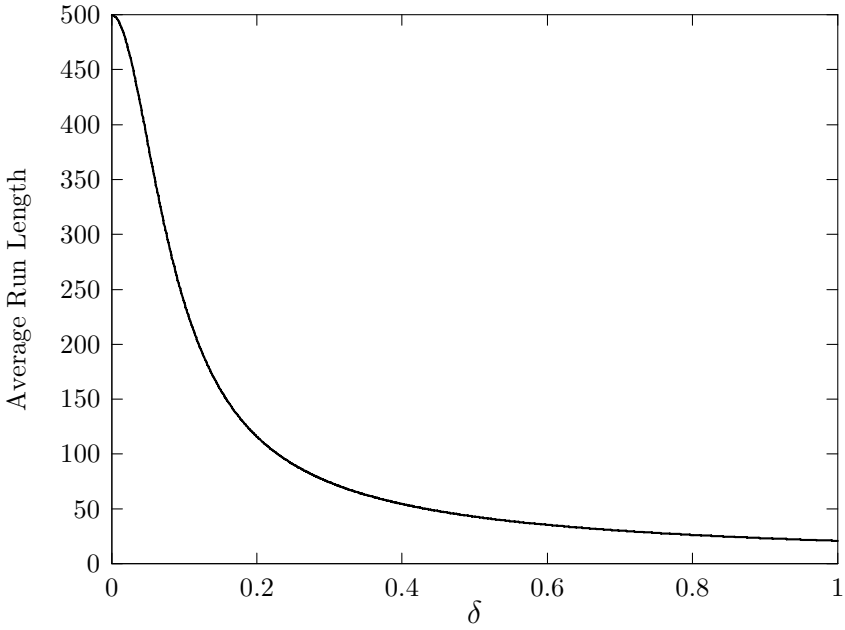

Figure 5.15: ARL curve of a CUSUM control chart ($k = 0.055$ and $h = 19.025$).

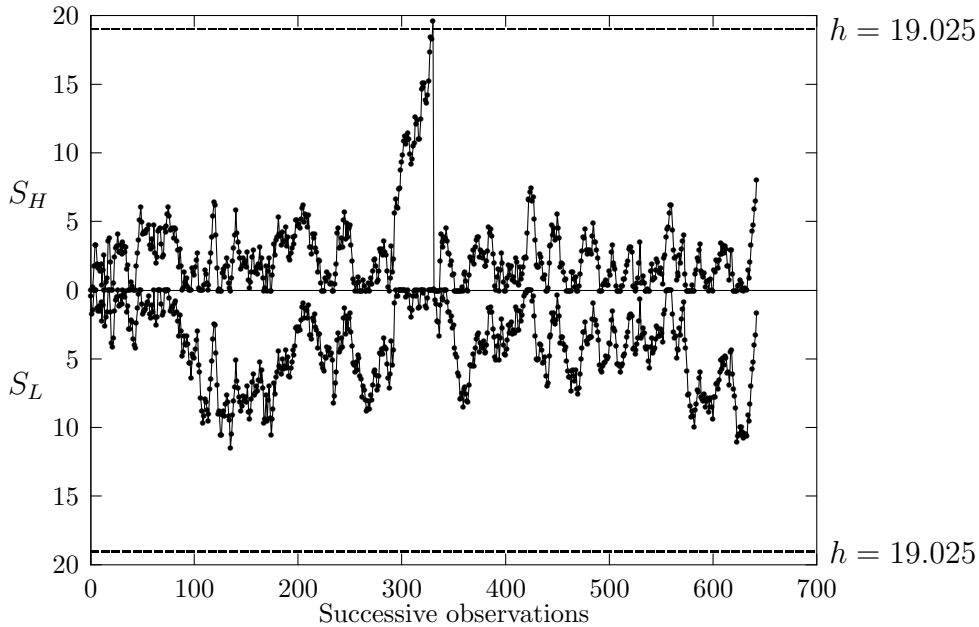

Figure 5.16: CUSUM chart of residuals of one step ahead forecasts.

the Shewhart chart that the trend in the data distorts the behavior of the one step ahead forecasts.

Obviously, the trend could be observed from figure 5.11, but from this figure it is not clear whether the upward movement is a result of the stochastic trend of model (5.5) or whether the data contain a deterministic trend. An out of control signal on a residuals control chart of model (5.5) provides statistical evidence of a trend in the observations in a way that is familiar to quality engineers.

While Alwan and Roberts (1988) advocate the use of two separate charts, the so-called *common cause chart* (figure 5.11) and a so-called *special cause chart* (figure 5.12 or 5.16), other authors (e.g. Montgomery and Mastrangelo (1991)) recommend combining these two charts into a control chart with a moving centerline, following variation induced by the the common causes of variation. However, at this point we feel it provides more insight to present the two charts seperately.

The previous analysis raises the question of how the model identification process was influenced by the presence of a deterministic trend. To investigate this, the data were detrended and the sample autocorrelation function of the resulting series was studied. The nonstationary behavior still appeared to be present. The estimate of the MA parameter changed from 0.811 to 0.878.

In order to estimate θ and the two trend parameters simultaniously, two dummy variables were added to model (5.5), one for the slow (linear) downward trends, and one for the steeper upward (linear) trend. The dummies both start from zero, and are increased by one for each observation where the trend is supposed to be active.

This resulted in the following fitted model

$$\bar{y}_t = \bar{y}_{t-1} + \varepsilon_t - \underset{(0.021)}{0.854}\ \varepsilon_{t-1} - \underset{(0.007)}{0.010}\ u_t + \underset{(0.024)}{0.079}\ v_t, \tag{5.11}$$

where u_t and v_t are dummies associated with the slow downward trends, and the upward trend, respectively. The value of u_t is increased by one for $t = 1, \cdots, 262$ and for $t = 425, \cdots, 624$. The value of v_t is increased by one for $t = 293, \cdots, 339$.

5.4 Implementation and results

During the four weeks of the experiment, the mean thickness and composition observations were closely monitored. In the previous subsection we discussed in detail how this was done for the composition data. The thickness data were monitored roughly in the same way, after making a log transform in order to be able to assume normality. The variability in the measurements was monitored using regular s-charts, since there appeared to be no correlation in subsequent sample standard deviations.

Studying the data as described in this chapter provided valuable information that helped to achieve a better understanding of the process. In practice, however, the procedure is too laborious to be implemented, and requires knowledge of statistical tools that did not come up in the education of the operators. A Shewhart chart of (means of) observations is much easier to work with than a CUSUM chart of

residuals of one-step ahead predictions. Therefore, the following procedure is now used at Philips Stadskanaal.

From the experiment it became clear that the output of the process will slowly deteriorate if the bath is not replenished regularly. Not replenishing for a period of two weeks will bring output measurements outside specification limits. Since the bath is filtered once a week, this is a natural moment to take a sample and have it analyzed accurately by the laboratory department. Based on these measurements, the bath is replenished such that all concentrations fall amply within their limits. For the rest of the week, apart from daily additions of formalin and brightener, additions are only allowed if this follows from the *out of control action plan* (OCAP), which is integrated in the automized SPC software. In the OCAP the process knowledge of the operators and chemical engineers as well as process supervisors is combined into a set of questions which initiate a systematic search for the cause of the out of control observations, and advise a remedy that in most cases can be applied by the operator to remove the cause that is responsible for the out of control signal.

The OCAP is triggered by out of control signals on regular Shewhart charts with widened control limits to allow for a slight wandering of the mean. This type of process monitoring is applicable since only relative short series (one week's data) are considered so that in 'normal' situations only a small deviation *due to common causes* in the mean will occur. We acknowledge that this approach could be improved upon, given the serial correlation present. However, in practice a balance must be struck between what is possible and what is optimal.

The result of the new replenishing strategy is illustrated by figures 5.17 and 5.18. Figure 5.17 shows layer thickness measurements of one weeks production before changing the replenishment strategy. The production data of the same week one year later, after changing the replenishment strategy, is depicted in figure 5.18.

We see that the variation in the individual observations reduced considerably, and that there is a substantial decrease in the number of outliers. The outcomes of the process are much more predictable. In addition, before the experiment, the mean level of layer thickness was raised to ensure that no individual measurement was found below 1μm. In the new situation, it was possible to reduce the mean level of layer thickness, without observing layer thicknesses below 1μm. This can be illustrated by comparing the average of the observations of figure 5.17(15.2350μm) to the average of figure 5.18, which equals 9.3990μm.

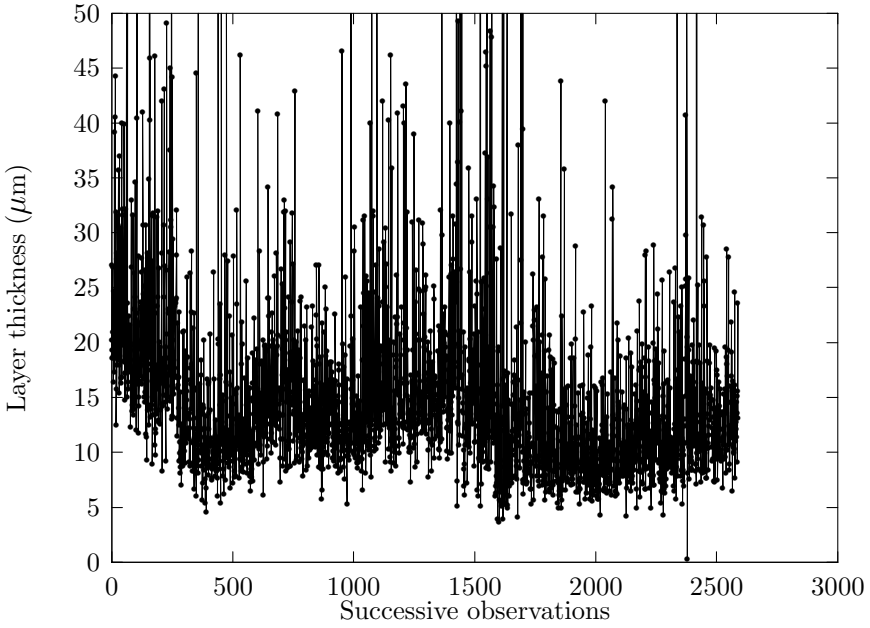

Figure 5.17: Layer thickness measurements before experiment.

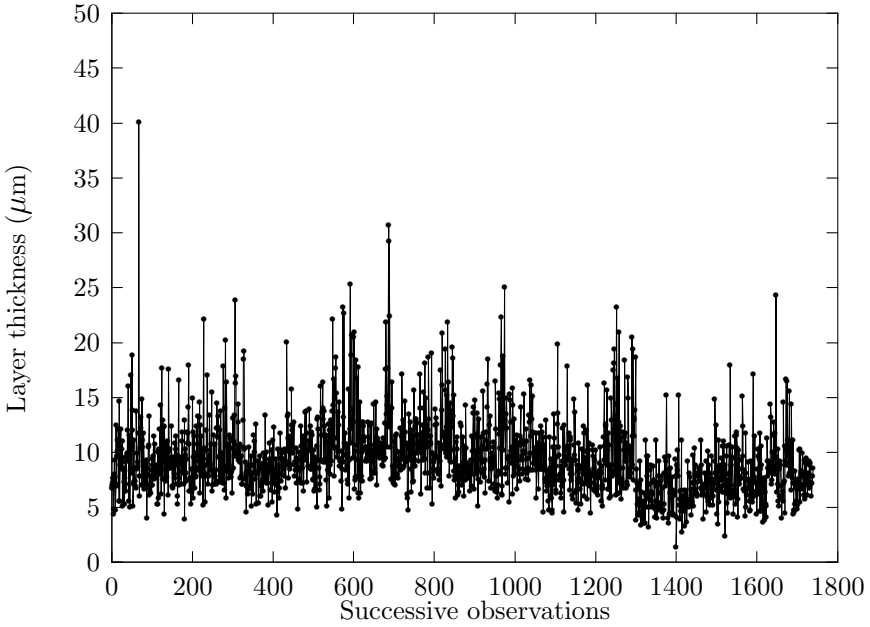

Figure 5.18: Layer thickness measurements after experiment.

6 Conclusions

In this report we discussed the process of tin-plating surface mounted diodes. The objective was to improve solderability. This was achieved by controlling the thickness and composition of the tin/lead layers.

Previously, the quality of the tin/lead layers was controlled by trying to keep the tin-plating bath stable. For several reasons described in chapter 4, this strategy did not work. It merely resulted in a less stable bath. In addition, the strategy was not understood. Replenishments were made ad hoc in extreme situations. Also, the way the replenishments were computed proved to be very costly.

In chapter 2 the latter problem is formulated as a LP problem. Contrary to the computation strategy that was used before, technical requirements were used as restrictions, while the main objective was to minimize a cost function. The old strategy only tried to fulfil the technical requirements. The replenishments that are computed with the LP model are optimal in the sense that there exists no cheaper set of additions such that all concentrations in the bath meet their requirements. In chapter 2 we illustrated by four real life examples that considerable savings were possible.

The other problem, the malfunctioning of the control strategy, was picked up in chapters 3–5. In chapter 4 it is discussed why the control strategy did not work. From an analysis of production data, it followed that the supposed relation between bath measurements and product measurements was not reflected in the observations. This relation was the basis for the control strategy. We argued that this relation could not be found in the data because of two reasons. Firstly, the measurement error on bath analyses proved to be too large. The variation due to measurement error was so large that an analysis of a perfect bath could well lead to measurements exceeding specification limits. Secondly, the concentrations of the two components that could not be measured, proved to be the most important for controlling the product measurements.

When we realized that the process was mainly controlled by measurement errors of more or less insignificant parameters, we decided to carry out an experiment where it was not allowed to make replenishments without statistical evidence of the need to do so. This was judged by means of control charts.

Monitoring the process was complicated by serial correlation in the measurements. In chapter 5 it is discussed what techniques we used to allow for serial correlation. Using these advanced techniques is more elaborate, but also provides more insight in the process.

After four weeks of experimenting with the new replenishment strategy, the product measurements were compared to product measurements before the experiment. The variation in the measurements reduced considerably, and the process has become more predictable. Therefore, it was decided to use the new replenishment strategy in the future.

Predictability means that Philips can keep promises towards customers concerning specification limits. Also, in the new situation, operators less often have to face situations where the outcomes of the process are unsatisfactorily, seemingly without

a specific reason. If an out of control situation occurs, the OCAP initiates a systematic search for an assignable cause. The result is that operators understand what causes the trouble and why certain actions are necessary to remove or to prevent unwanted behavior of the process. Before the experiment there used to be prescribed actions for several calamities, but there were very few people that were able to explain why these worked and who invented them.

The most important quality improvement stems from terminating overregulation (*tampering*). As a byproduct the use of chemicals in the tin-plating bath reduced by a considerable amount. Concerning the objective "improving solderability": we are not aware of any customer complaints about solderability after the introduction of the new replenishment strategy.

However, there are still some points that deserve further investigation. Firstly, taking the education in statistics of the operators into account, a comprehensive way of allowing for serial correlation in process monitoring schemes has yet to be developed.

Secondly, we studied a trended sequence of observations, and we concluded that both a a stochastic and a deterministic trend were present in the data. Since the presence of a deterministic trend often has entirely other practical implications than a stochastic trend, it is very important to be able to make a proper distinction.

Acknowledgements

I would like to thank the people who cooperated in this investigation. First of all, I would like to thank the people at Philips Semiconductors Stadskanaal, especially my colleagues from the Industrial Support Group. They showed great interest in the problems I was working on, and provided help whenever needed. In particular I would like to thank Albert Trip and Klaas Huisman for sharing their wit and wisdom in the numerous discussions we have had. Also, I would like to thank the PAT for their hospitality and allowing me to work on such interesting problems.

I acknowledge Gert Tijssen from the University of Groningen for his work on the LP model of chapter 2.

Finally, I would like to thank my supervisor, Ton Steerneman, from the University of Groningen. He was also actively involved in this investigation. He followed the progress with enthusing constructive criticism. I am very grateful for his contributions.

References

Alwan, L. C. and H. V. Roberts (1988), "Time-series modeling for statistical process control", *Journal of Business & Economic Statistics*, **6**(1), 87–95.

Anderson, T. W. (1971), *The Statistical Analysis of Time Series*, John Wiley & Sons, New York.

Box, G. and T. Kramer (1992), "Statistical process monitoring and feedback adjustment—a discussion", *Technometrics*, **34**(2), 251–285.

Box, G. E. P. and G. M. Jenkins (1976), *Time Series Analysis, Forecasting and Control*, revised edition, Holden-Day, San Francisco.

Does, R. J. M. M., K. C. B. Roes, and A. Trip (1996), *Statistische Procesbeheersing in Bedrijf: Implementeren en verankeren van SPC*, Kluwer Bedrijfswetenschappen, Deventer, The Netherlands [In Dutch].

Does, R. J. M. M. and B. F. Schriever (1992), "Variables control chars limits and tests for special causes", *Statistica Neerlandica*, **46**(1), 229–245.

Does, R. J. M. M., M. L. van Oord, and A. Trip (1994), "Een succesvolle implementatie van statistische procesbeheersing (SPC)", *Sigma*, **40**(6), 10–13 [In Dutch].

Granger, C. W. J. and P. Newbold (1986), *Forecasting Economic Time Series*, second edition, Academic Press, San Diego.

Harvey, A. C. (1993), *Time Series Models*, second edition, Harvester Wheatsheaf, New York.

Lucas, J. M. and R. B. Crosier (1982), "Fast initial response for CUSUM schemes: Give your CUSUM a head start", *Technometrics*, **24**(3), 199–205.

Montgomery, D. C. (1996), *Introduction to Statistical Quality Control*, third edition, John Wiley & Sons, New York.

Ryan, T. P. (1989), *Statistical Methods for Quality Improvement*, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, New York.

Sierksma, G. (1996), *Linear and Integer Programming: Theory and Practice*, Marcel Dekker Inc., New York.

Stefani, R. T., C. J. Savant, Jr., B. Shahian, and G. H. Hostetter (1982), *Design of Feedback Control Systems*, third edition, Saunders College Publishing, Boston.