CHAPTER 2

Merits of the MT System and Its Computation Methods

A series of introductory articles on the MT System, created by Dr. Genichi Taguchi, appeared in the Journal of Quality Engineering Society (Japan) starting in the 1990s.[2] The MT System is short for the Mahalanobis-Taguchi System. And, as the name suggests, Mahalanobis Distance was used as the computational mathematical foundation of the system. In the years since, a number of new computation methods have been proposed by Dr. Taguchi.

Currently, the following four computation methods are operational in association with the MT System:

MT Method
MTA Method
TS Method
T Method

The T Method breaks down into three types of computational methods: T Method-1, T Method-2, and T Method-3. At present, the MT Method, T Method-1, and T Method-3 are on record as the unchallenged, most widely used methods. For this reason, those three methods will be discussed in some detail from the standpoint of their merits and some of the computational methods they employ.

2.1 CHARACTERISTICS SHARED BY ALL MT SYSTEM COMPONENTS

2.1.1 THE "UNIT SPACE" CONCEPT

The fact that a "Unit Space" is defined is the most important feature of the MT System. A Unit Space is a population that is homogeneous with respect to the target object, which is a normal population in many cases.

The conception of the MT System by Dr. Taguchi is said to trace its origin to the line, quoted below, from the opening passage of Leo Tolstoy's novel Anna Karenina:

> *Happy families are all alike, but every unhappy family is unhappy in its own way…*

The concept of "alike" makes up the basis of the MT System; its significance cannot be overemphasized. "Alike" points to the essential characteristic of a homogeneous population, a condition interchangeably, and broadly, describable as a constant and normal state, a state associated with high frequency, or an average state.

One might with some credibility claim that unhappy families are more homogeneous, but we must say at the end of the day that that is a misguided outlook. If we make a mistake in the way we look at things, we cannot achieve accurate results. The definition of homogeneity is the most important point in the MT System.

2.1.1.1 Reliability of Pattern Recognition is Easily Ensured

In measuring the degree of healthfulness, the traditional methodology would closely examine the state of the illness (abnormality) and devote the greatest part of medical attention to explaining the nature and characteristics of the illness on hand. In other words, those about whom data would be collected were mainly sick people. However, since the types and degrees of illness are of uncountable variety and complexity, making it extremely difficult to cover every detail in a thorough manner, it would be a challenge to define a stable frame of coordinates for pattern identification. Compared with this, *healthy people are homogeneous in nature so that they can be defined as a robust and steady entity*.

Such is the case of some of the persistent problems besetting the industrial sector, as well. Unacceptable products are not of one standard variety and the manufacturing situation, with regard to quality, is seldom stable; whereas when the situation is stable, a homogeneous, reliably stable measuring yardstick will prevail.

2.1.1.2 Sensitivity to Unknown Abnormalities Can Be Ensured

If we define a normal state as the Unit Space, and the object data is far removed from that space, we can determine the situation to be abnormal. There are a number of types and degrees of abnormality, and it is even possible that unknown types of abnormality exist. Under the MT System, all that one needs to be able to do in order to determine whether a case is normal or abnormal is simply define the Unit Space.

2.1.2 EXAMPLES OF UNIT SPACE

With the MT System, the Unit Data is defined based on homogeneity. A homogeneous state, in many cases, is a normal state. But a homogeneous group cannot necessarily be equated to a normal state. Unit Space examples appear with comments below under the following three broad headings:

2.1.2.1 When a Normal State Constitutes a Homogeneous Population

PRODUCT INSPECTION

Few would disagree that homogeneity is expected to prevail throughout the lot of parts that an inspector has determined to be normal. If one inspector's findings are considered insufficient, one solution would be to assign a team of inspectors to the case; and, if the majority of them determine something to be normal, to designate that product as the Unit Space. When the MT System was introduced to the United States in 1998, Unit Space was referred to as "Normal Space."

HEALTH EXAMINATIONS

It is natural for us to believe that healthy people would form a homogeneous grouping. If necessary, people who have been determined by multiple physicians to be in good health can be defined as truly healthy people. If an even more cautious definition is required, the second-year data of a person who has been diagnosed as healthy for three consecutive years would be appropriate as the Unit Data.

It is often argued that "A strapping student with an athletic background and a regular office worker do not belong together as members of a homogeneous population." Deciding which of the two types should be considered Unit Data material for the purposes of measuring healthfulness seems like a puzzling matter. However, as long as the Unit Space is bias-free, in many cases it really does not matter which is chosen.

MONITORING BUILDINGS AND OTHER STRUCTURES

It happens once in a while that, all of a sudden, a bridge collapses. Buildings and other structures, which would cause enormous damage if they collapsed, should be placed under the watchful eye of sensors. If the structures are judged by specialists to be in fine condition one year after they are built, the state in which they have stood can be considered to be the homogeneous grouping. This concept can be applied to giant structures such as dams and large buildings.

2.1.2.2 When a Normal State or an Ideal State Is Not Necessarily a Homogeneous Population

Let's consider a situation in which a certain manufacturer's state of production is stable but is beset by a particular off-specification output rate. The mold casting process at a factory has been experiencing a 10% rate of deformed products coming off the line for a number of years. Malformed products occur because an ideal casting setup is not in place. In circumstances such as these, the current casting process, even if not ideal, constitutes a homogeneous population. And even if there happens to be a day that the casts are completely free of deformities, this would have occurred *by accident*, and no homogeneity can be claimed by such a happenstance.

This state may be conceptually illustrated as in Figure 2.1. What this figure shows is that, although the current production process is basically stable in terms of quality, the quality, as such, is not ideal and leaves room for improvement.

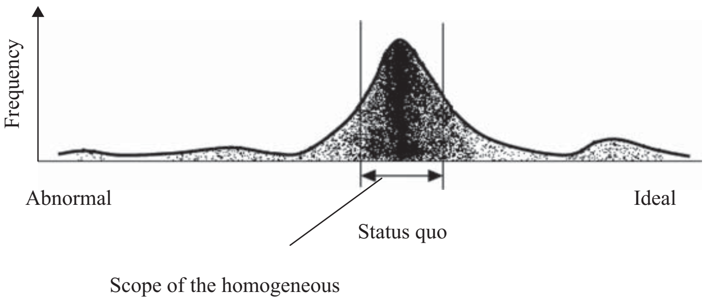


Figure 2.1. Conceptual image of a homogeneous population differing from the ideal.

2.1.2.3 Other Cases

PREDICTION OF NATURAL PHENOMENA

If we take weather forecasts for example, we can think of several states of homogeneity. In the case of Hawaii, or a desert, an overwhelming majority of days are clear, which should be regarded as forming a homogeneous population. Generally speaking, the most commonly occurring state should be defined as the homogeneous state. In some cases, it may become necessary to prepare separate, season-by-season Unit Spaces. As for earthquake prediction, a quake-free state (a state characterized by as little earth tremor as possible) is the homogeneous group.

CHARACTER RECOGNITION

In the case of character recognition, no such thing as a normal, ideal character pattern exists. We have no choice but to settle for this statement: A collection of a variety of written characters, which are all readable as "5," therefore will be considered the homogeneous population.

In the case of the MT System, one Unit Space constitutes one homogeneous population at a time so that, if numeral recognition is to be performed for 0 to 9, ten Unit Spaces will have to be prepared.

ESTIMATION OF ECONOMIC VALUE

In the prediction of a sales trend or corporate performance, a stable state, which withstands minimal change, forms a homogeneous population. This would be a state that flows with the times within a set range of fluctuation based on economic indicators (of, for example, ±5%). In the case of real estate prices, the use of *average values* is fairly common.

Figure 2.2 shows a stable state, followed by an unstable state undergoing the phenomenon of changes in a temporal sequence. As represented in the graph, the stretch immediately preceding the transfer to a state of instability would be included within the bounds of "what can be regarded as a state of stability."

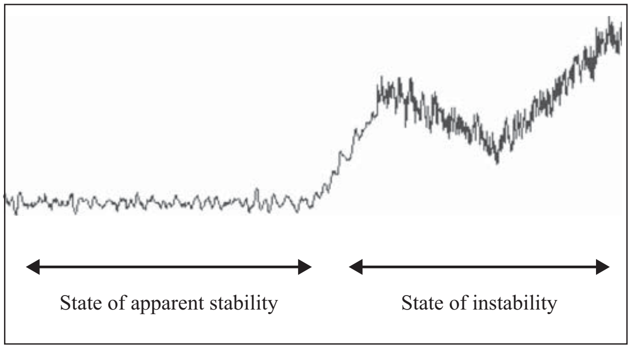


Figure 2.2. Definition of a state of stability in temporal sequence data.

2.1.3 ITEM SELECTION AND CAUSE DIAGNOSTICS

Under the MT System, item selection and cause diagnostics, which is an application related to item selection, are performed. These processes are not part of traditional statistical processing or pattern recognition, but are of a highly practical value.

2.1.3.1 Concept and Method of Item Selection

With the MT System, as with all types of pattern recognition, it is necessary to use *variables (items) valid for the purposes of recognition*. If *invalid variables* are included, recognition detection and reliability will suffer. For this reason, it is important to set a sufficient quantity of variables that are considered valid. But in many cases, the validity of variables is not known at the initial stage. As shown in Figure 2.3, the task of choosing valid variables out of a multitude of prospective variables is like unearthing *gems* from a mixed bag of gems and pebbles.

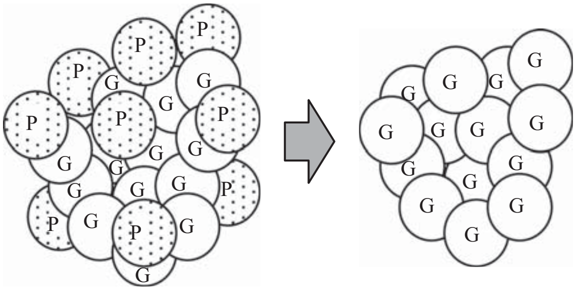


Figure 2.3. Concept of item selection (P: pebble; G: gem).

With the MT System, valid items (variables) can be selected using a procedure called "item selection." At the time of item selection, the validity of items is evaluated from one or the other of the following standards:

1. Using data that is known not to be included in the Unit Space, evaluating to what extent each item in question is contributing to the increase in MD;
2. Using data, the degree of abnormality of which is known in advance, evaluating to what extent each variable in question is contributing to abnormality detection.

The idea is as outlined above, but when it comes to actual application, various hurdles lie ahead that are not easily surmountable. If we, for example, have 100 prospective variables on hand and set out to perform an evaluation as set forth in 1) or 2) above, we will have to deal with a large number of combination evaluations. It is not even known in advance if the number of valid items is 10 or 50.

To solve this problem, we will make use of an orthogonal array. An orthogonal array is a chart utilized to compress a large number of combinations rationally. For item selection purposes, a two-level orthogonal array is used. An example of this, orthogonal array L_8, is shown in Table 2.1.

Using this chart, the seven variables A through G are allotted to the appropriate columns. The number of all possible combinations of the seven variables works out to $2^7 = 128$, but the use of the orthogonal array makes it possible to compress the number of evaluations to just eight. The orthogonal array is of an infinite scale and can flexibly assume any size according to the number of variables.

Table 2.1. Orthogonal Array L_8

Alloted variables in this tier

Number of tests

No.	A	B	C	D	E	F	G
1	1	1	1	1	1	1	1
2	1	1	1	2	2	2	2
3	1	2	2	1	1	2	2
4	1	2	2	2	2	1	1
5	2	1	2	1	2	1	2
6	2	1	2	2	1	2	1
7	2	2	1	1	2	2	1
8	2	2	1	2	1	1	2

In proceeding to item selection with the help of the orthogonal array, we will perform the evaluations according to the following:

When a variable is to be used → Level 1

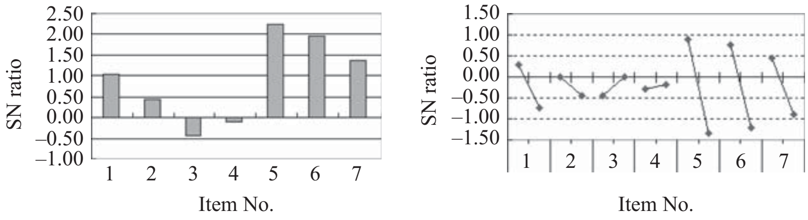
When a variable is ***not*** to be used → Level 2

Therefore, when the orthogonal array in Table 2.1 is utilized, for the tests with the items in the first row, all variables will be used to create the Unit Space; for the tests with the items in the second row, the three variables A, B, and C will be used to create the Unit Space. And, for example, to perform an evaluation according to item 1) mentioned above, the evaluation will be made of the Unit Space using data known to be abnormal.

If among the variables being used there are any variables or sets of variables with sensitivity to abnormalities, the distance from the Unit Space will increase. Conversely, if such variables are not included, the distance will be short. After eight tests have been performed, all combination tests will have been completed in simulated form, allowing the question regarding the validity of each variable to be answered.

2.1.3.2 Results of Item Selection

Figure 2.4 shows the results of an item selection run. The horizontal axis lists item numbers (variable numbers). In the said figure, (a) is a bar graph representation of the results of analysis performed with the help of the orthogonal array, while (b) is a representation of the same using a broken-line graph. Note, when reading the graphs, the greater the positive value of the variable in display (a), and the steeper the drop to the right of the broken line in display (b), the greater the degree to which the variable is valid for abnormality determination purposes. Use of valid variables selected on the basis of the results shown in graphs such as these will enable the creation of a Unit Space that has excellent recognition ability.



(a) Bar graph display

(b) Broken-line graph display

Figure 2.4. Sample results of item selection.

Note, however, that there exist variables in Figure 2.4 (a) that carry a negative value. The meaning of these negative-valued variables is that "use of such variables will degrade the sensitivity of abnormality detection." It is therefore better to avoid using them. But it remains to be theoretically explained if conspicuously negative variables really are devoid of any validity at all.

2.1.3.3 Points to Heed concerning Item Selection; What Countermeasures to Take

A Unit Space created on the basis of the results of item selection will logically have a heightened sensitivity to abnormalities. A few points need to be heeded, however. They are discussed below together with what, if any, measures to take, should attention be required.

ENUMERATION OF DATA OUTSIDE THE UNIT SPACE

As for data not included in the Unit Space prepared for item selection, all states should be enumerated. Item selection should be performed only after all abnormal data has been prepared. If these are not enumerated, there is the danger that blind spots (undetectable abnormalities) could remain intact.

NON-ABSOLUTE NATURE OF ITEM SELECTION

The result of item selection is not absolute. That is so for the following two reasons:

1. Due to the effects of reciprocal action and other interactions within orthogonal arrays, errors occur that affect the results. Accordingly, instead of selecting variables simply in the descending order of more effective ones, it is better to proceed with the selection allowing a range for each variable.
2. Because the orthogonal array-based effects represent average values, variables with great effects may in some cases fail to appear prominently.

MEASURES TO TAKE IN CONNECTION WITH THE POINTS TO HEED

The following measures may be taken to solve the apparently contradictory issues mentioned above. One is to prepare *both* a Unit Space created on the basis of the results of item selection *and* a Unit Space composed of all the variables that existed before the selection (Figure 2.5). This will provide a single solution to two contradictory issues, highly sharpened recognition accuracy and preparedness for unknown abnormalities. Unless a computation speed-related issues surfaces, this is a highly effective solution.

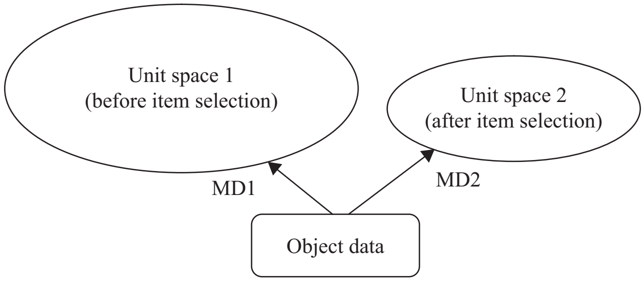


Figure 2.5. Method of using multiple unit spaces for one object.

2.1.3.4 Cause Diagnostics

When the Mahalanobis Distance (MD) turns out to be large, which item, or set of items, is responsible for having made these values so large can be diagnosed. The result of the diagnostic can be put to use toward **identification of the type of abnormality or prevention of the occurrence of future abnormalities** and is of inestimable practical value.

The basic computation procedure for cause diagnostics is exactly the same as for item selection. With item selection, multiple abnormal data is utilized, but with cause diagnostics only one abnormal item of data is focused on. That is the only difference.

2.1.4 MT SYSTEM ARCHITECTURE

The MT System architecture is shown in Figure 2.6. This chart shows which method is indicated for use depending on the conditions of the data to be treated. The most widely used today of the MT System components are: the MT Method, the T Method-1, and the RT Method = T Method-3. Table 2.2 lists the items required of the pattern recognition tool and further describes the characteristics of the MT System computation method most indicated for each of them.

The MT* Method, alluded to in Figure 2.6 and Table 2.2, is basically an MT Method with a singular solution processing unit incorporated capable of performing accurate computations even when multicollinearity presents.

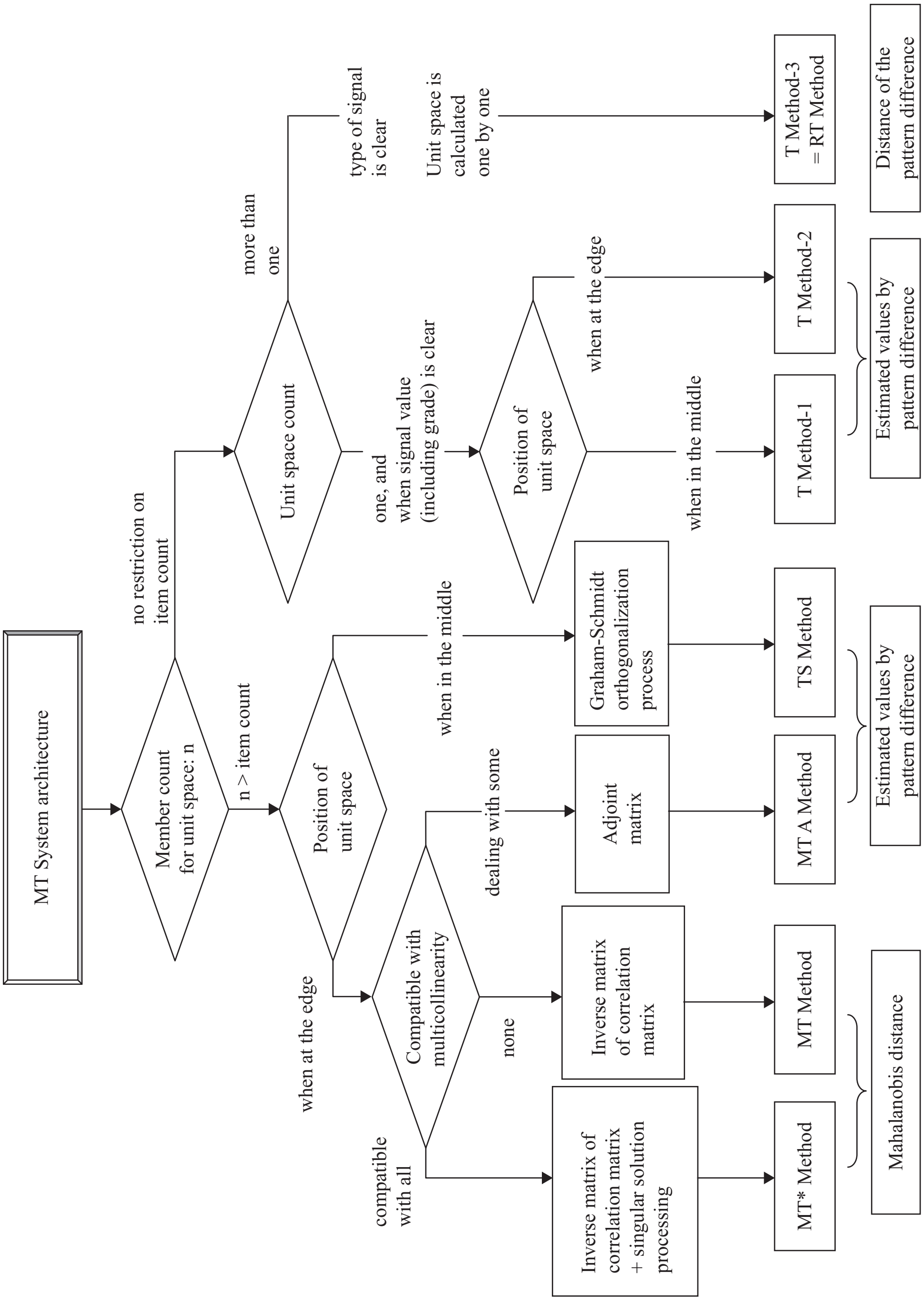


Figure 2.6. MT system architecture.

Table 2.2. Features of the MT system at a glance

	Method		
Characteristic	**MT/MT***	**T Method-1**	**T Method-3 = RT**
Breakdown	Inverse matrix of correlation matrix	Use of SN ratio and Sensitivity	Compress two variables
Use of correlations	Used	Not used (Only main effect)	Used a little
With or without code	Without	With	Without
Necessity of signal	Unnecessary	Necessary	Unnecessary
Compatible with multicollinearity	Compatible	Compatible	Compatible
Amount of data	Medium	Large	Large

Note 2.1 "The larger, the better" SN ratio

It is better that the MD of data that is understood not to belong to a Unit Space be as large as possible. "The larger, the better" characteristic is a concept considered unique to Quality Engineering. Wire rope and structural strength are further concrete examples of "the larger, the better" concept. SN ratio is none other than a quantification of this type of desirability.

The computational formula for "the larger, the better" SN ratio is given below. "The greater the SN ratio, the higher performing it is" is the meaning. In it, D^2 is Mahalanobis Distance (in squared value), and m, the number of data that do not belong to the Unit Space.

"The larger, the better" SN ratio $\eta = -10 \log\{(1/D_1^2 + 1/D_2^2 + \ldots + 1/D_m^2)/m\}$.

Note 2.2 SN ratio of dynamic characteristics

If a value that was input appears in the output intact, a linear correlation between input and output must have occurred. Some further concrete examples of such a correlation are faucet valve turn angle and the amount of water discharge, amount of gas pedal acceleration and the number of car engine rotations, and so on. But in commonly seen real-life situations, the input-output relation, impacted by disturbances of exterior origin, energy loss, and other events deviates from linearity. One measure of evaluating the "insignificance" of such deviations—to what extent a linear relation is secured—is the quantified SN ratio of a dynamic characteristic (Figure 2.7).

Note 2.2 SN ratio of dynamic characteristics (Cont.)

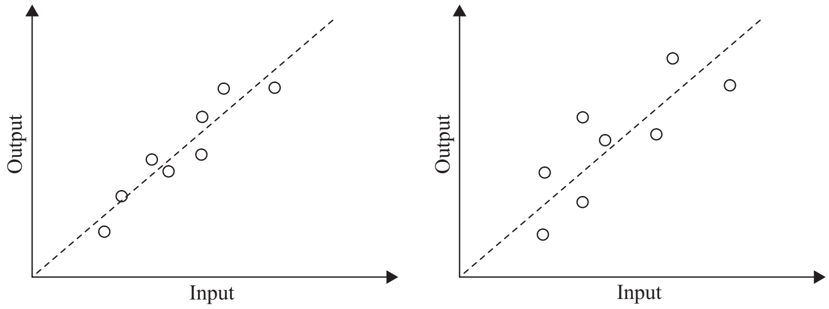


Figure 2.7. Concept of dynamic characteristics.

In the case of pattern recognition as well, if the value of a signal and the result of recognition are in a linear correlation, it can be said that a good recognition system is in place. Thus, an evaluation of the quality of the recognition function can be made using the SN ratio of the result of the dynamic characteristics.

2.2 FEATURES OF THE MT METHOD

2.2.1 OVERVIEW

The MT Method is a method for gaining cognizance of whether the target data (otherwise referred to as unknown data) belongs to the same standard, homogeneous, group. The MT Method defines a homogeneous population as the Unit Space. And it finds the distance between the center of the Unit Space and the target data in the form of Mahalanobis Distance (MD).

As shown in Figure 2.8, it is recognized that, if the MD turns out to be short, the pattern is close to the Unit Space, and that, if the MD turns out to be long, the pattern is distant. When the Unit Space is a normal population, if the MD is short, there is a strong possibility that the target data is normal.

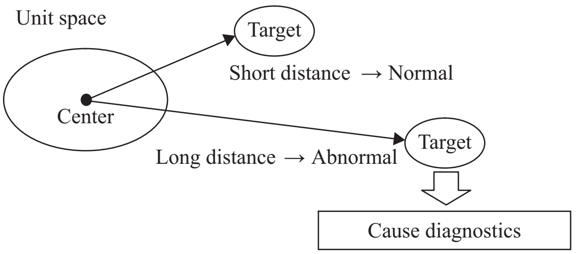


Figure 2.8. Conceptual drawing of the MT method.

The MT Method stands on a simple concept and is very easy to use; among the components of the MT System, the MT Method has the greatest number of practical applications. Mahalanobis Distance itself was proposed by Indian mathematician P. C. Mahalanobis. In India, bones and fossils of a wide variety of animals had been discovered, and Mahalanobis introduced a new statistical concept as a rational approach to the classification of these finds.

2.2.2 EXAMPLE OF AMOUNT OF RAINFALL AND NUMBER OF UMBRELLAS SOLD

The basic meaning of Mahalanobis Distance is explained citing a specific example. Table 2.3 shows the correlation between the monthly average rainfalls in Tokyo from 1971 to 2000 and a local umbrella shop's tally of umbrellas sold. The table shows that the rainfall is heaviest in September, and umbrella sales for that month are the briskest, as well. On the other hand, the rainfalls for December and January are the lightest, as umbrella sales for those months are slowest.

Table 2.3. Relationship between amount of rainfall and umbrella sales

Month	Amount of rainfall (Normalized)	Numbers of umbrella sold (Normalized)
1	48.6	3,400
2	60.2	5,642
3	114.5	7,769
4	130.3	9,854
5	128.0	6,950
6	164.9	10,876
7	161.5	13,007
8	155.1	10,857
9	208.5	14,595
10	163.1	9,410
11	92.5	8,475
12	39.6	3,772
Average	122.2	8,717
Standard deviation	50.6	3,294

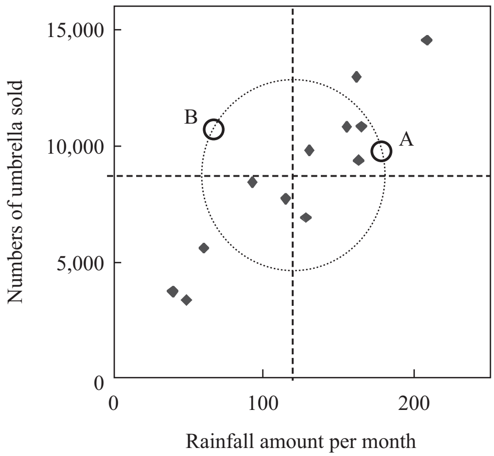


Figure 2.9. Unknown data A and B and degree of membership.

A scatter diagram that graphically illustrates this relationship is shown in Figure 2.9. From this graph, we can see that there is a strong *correlation* between the rainfall and the amount of umbrellas sold. The correlation coefficient between the two is 0.936. The cause-and-effect relationship in the proposition that rainier months see more umbrellas sold is intuitively easy to understand as a matter of ordinary common sense. From the viewpoint of the MT Method, a normal (standard) state such as this is defined as a Unit Space.

Now, let us assume that, for certain months we are looking at, that the following amounts of rainfall were recorded and the umbrellas sold during the same months (hereafter called "umbrellas sold") amounted to the following numbers:

A [Rainfall: 175 mm; umbrellas sold: 10,500]

B [Rainfall: 91 mm; umbrellas sold: 12,000]

And let us consider whether A and B are close or far from the normal (standard) state.

First, we consider what positional relationship B has to the Unit Space. For B, the rainfall is 91 mm. If we assume that the standard deviation of the group of 12 dots is σ, B is about 0.7σ away. The number of umbrellas sold for B is 12,000, which indicates a position approximately 1 σ removed from the average. In other words, the rainfall and umbrellas sold for B, as such, do not count as particularly abnormal values.

But, when looking at the figure, if we ask ourselves whether B can then be regarded as a member of the group of 12 dots, we will get the feeling that "something seems strange." B seems left out of the collection of 12 dots, unlike A, which seems to belong with the group of 12 dots.

2.2.3 COMPUTATION OF THE MAHALANOBIS DISTANCE

The Mahalanobis Distance (MD) provides an appropriate solution for problems such as the one just described. The MDs of points A and B, when calculated, work out as follows:

MD of point A: $D_A^2 = 1.30$

MD of point B: $D_B^2 = 10.14$

What these results tell us is that the shorter the MD, the nearer the target to the Unit Space. Therefore, given these MD data, we can acquire the degrees of kinship of points A and B to the normal state. In other words, we can say that point B is distant from the normal state, while point A is by far closer to the normal state than B. If we look into why it is that point B is distant, we realize that it is so because umbrella sales are high in spite of a relatively modest amount of rainfall. It may be that the month in question happened to be packed with outdoor events with sudden rain surprising visitors to the area on many days.

Next, we will take a look at what sort of computation leads us to the MD values such as the 1.30 or 10.14 that we have arrived at here. The computation processes involved are examined below.

2.2.3.1 Normalize the Data

The first step is to *normalize* the data for the rainfall and number of umbrellas sold. What is meant by normalization is the mathematical processing of making heterogeneous measurement values to the same dimensional values. For example, we note in the above situation that the rainfall is represented in values of the order of tens to hundreds of millimeters, while the number of umbrellas sold is on the order of thousands to upwards of ten thousand items. The numerical values are predicated in different units of measurement and scale of digits. Normalization is processed as follows:

$$\text{Normalized value} = (\text{Raw data} - \text{average value})/(\text{standard deviation}) \qquad (2.1)$$

For the standard deviation, we use a formula (degrees of freedom = number of data points) that treats data as the whole of the statistical population.

Introducing actual numerical values, the values for rainfall and umbrellas sold, say, for January are normalized as follows:

$$\text{Rainfall (normalized)} = (48.6 - 122.2)\ 50.6/50.6 = -1.46 \qquad (2.2)$$

$$\text{Umbrellas sold (normalized)} = (3{,}400 - 8{,}717)/3{,}294 = -1.61 \qquad (2.3)$$

Normalizing the data for February and beyond, we obtain the values shown in Table 2.4.

As Table 2.4 shows, the variables, as normalized, both come out as 0.00 for the average and 1.00 for the standard deviation. This means that variables of different kinds have now been

Table 2.4. Normalized amount of rainfall and quantity of umbrellas sold

Month	Amount of rainfall (Normalized)	Numbers of umbrella sold (Normalized)
1	−1.46	−1.61
2	−1.23	−0.93
3	−0.15	−0.29
4	0.16	0.35
5	0.11	−0.54
6	0.84	0.66
7	0.78	1.30
8	0.65	0.65
9	1.71	1.78
10	0.81	0.21
11	−0.59	−0.07
12	−1.63	−1.50
Average	0.00	0.00
Standard deviation	1.00	1.00

converted so that they can stand on equal footing. Plotting the normalized data on the graph, we obtain Figure 2.10. Notice the scale marks on the y-axis and x-axis of this figure.

The meaning will become clear as to why the amount of rainfall and number of umbrellas sold have been converted so they can "stand on the same page." The elongated oval shape in the figure outlines the limits of group membership. By rule measurement, the distance from the center point to A and that to B are identical, and both A and B orbit in one and the same circle. The oval and the circle should aid the understanding of Mahalanobis Distance.

2.2.3.2 Calculate the Correlation Matrix

Using Table 2.4, we will calculate the correlation matrix between amount of rainfall and umbrellas sold. Since this case involves two variables, the size of the correlation matrix will be 2×2. There is only one correlation coefficient to calculate, and the value is 0.936. Therefore, correlation matrix R will be found as follows:

$$R = \begin{pmatrix} 1.000 & 0.936 \\ 0.936 & 1.000 \end{pmatrix} \tag{2.4}$$

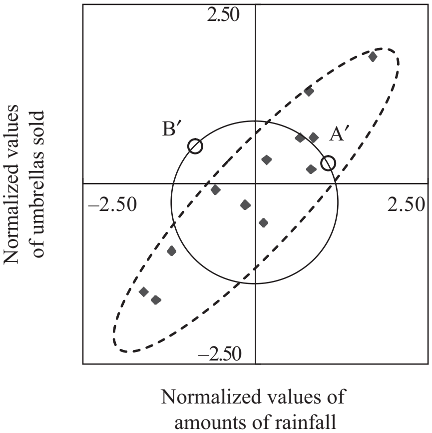


Figure 2.10. Group of 12 dots and points A and B, normalized.

2.2.3.3 Calculate the Inverse Matrix

The inverse matrix $\boldsymbol{R}^{-1}$ of the correlation matrix equation (2.4) is calculated.

$$R^{-1}=\begin{pmatrix} 8.029 & -7.153 \\ -7.153 & 8.029 \end{pmatrix} \tag{2.5}$$

2.2.3.4 Normalize the Target Data

Normalize the data pertaining to point B [75 mm of rainfall, 12,000 umbrellas sold] according to the formula given (2.1).

$$Y_{1B}=(91.0-122.2)/50.6=-0.617 \tag{2.6}$$

$$Y_{2B}=(12{,}000-8{,}717)/3{,}294=0.966 \tag{2.7}$$

2.2.3.5 Calculate the Mahalanobis Distance

Calculate the Mahalanobis Distance according to the formula given below. Note that k is the number of variables, which in this case is $k = 2$. Note also that $\boldsymbol{Y}^T$ is a notation for "transposition."

$$\begin{aligned} D_B^2 &= \mathbf{YR}^{-1}\boldsymbol{Y}^T/k \\ &= \begin{pmatrix} -0.934 & 0.996 \end{pmatrix} \begin{pmatrix} 8.029 & -7.513 \\ -7.513 & 8.029 \end{pmatrix} \begin{pmatrix} -0.617 \\ 0.996 \end{pmatrix} \Big/ 2 \\ &= 10.14 \end{aligned} \tag{2.8}$$

For point A, as well, $D_A^2=1.30$ is obtained in a similar manner.

2.2.4 SUPPLEMENTARY NOTES

2.2.4.1 Number of Correlation Coefficients

Regarding the meaning of the Mahalanobis Distance, explanatory notes have been given in connection with an example with two variables. In reality, there can be tens of variables involved, or more. As for the number of correlation coefficients, one set will suffice for two variables, but if ten variables are involved, 45 sets ($_{10}C_2$) will be calculated; if 100 variables are in question, a good 4,950 sets ($_{100}C_2$) will have to be used. However, even if the number of variables grows large, what we have is an aggregation of correlations between two variables.

2.2.4.2 The Gram-Schmidt Orthogonalization Process

In connection with MT Method computational methodology, Dr. Taguchi demonstrated a computation method based on the Gram-Schmidt orthogonalization process. In making use of this computation method, Dr. Taguchi suggested: "One can find practically beneficial solutions by utilizing "high-level variables considered important by engineers." But it is often difficult to decide on importance, and to date, few if any instances of actual use have been reported.

If computations not only of high-level variables, but also of an entirety of items are executed all the way to their logical conclusions using Gram-Schmidt's orthogonalization process, the solutions obtained prove equivalent to those provided by the Mahalanobis Distance. If all items are processed using a computation method based on principal component analysis, here, likewise, an MD-equivalent solution is obtained.

Note 2.3 Advances in the practical applications of Mahalanobis distance made possible by the sophistication of information technology hardware

MT System computation would be next to impossible if it were to be done manually. Mahalanobis Distance was proposed by P.C. Mahalanobis in 1936, and its theoretical rationality was recognized at an early stage in the field of statistics. But not much use was made of it except as applied to the internal operations of the multivariate control chart and discriminate function.

In the 1980s, as ever-higher-speed personal computers became available, as did ever-larger memory capacities, it became easier to perform multivariate problem solving tasks. The same can be said of problems involving artificial intelligence (AI) and image processing. With PCs becoming more and more readily available, it became possible for many researchers and engineers to tackle AI with relative ease. This development benefited image processing as well, and heavy-duty computation processing was made possible thanks to increasingly sophisticated digital cameras and image processing software.

Information technology hardware will continue to leapfrog ahead, ever widening the range of issues that can be solved using the MT System.

2.3 FEATURES OF THE T METHOD[3]

2.3.1 FEATURES OF THE T METHOD-1

The T Method-1 represents a theory concerning the estimation of an output (objective variable) from a multivariate, which renders computation feasible even with a limited quantity of already-known data. We will explain this point below using a concrete example.

When collecting new data is possible, the application of traditional Quality Engineering is appropriate, but if an accumulated body of data pertaining to day-to-day manufacturing activities is available, it is desirable to pull all available information from the data using the MT System.

A case in 1998 where the MT Method was put into practice was presented under the title of "Wafer Yield Prediction Based on the Mahalanobis Distance." "Wafer yield" refers to the ratio of good-quality chips to the entire lot of chips produced from one wafer. In the case presented, a "group of high-yield wafers" was selected as the Unit Space.

This case was one of the trailblazing case studies that demonstrated an actual application of the MT Method. Dr. Taguchi made some key observations, which are summarized in Figure 2.11. In a nutshell, he said that how a Unit Space is set is the most important part of the task of designing an information system. The Unit Space should be filled with data that meets the following criteria:

1. The results should be as homogeneous as possible
2. A dense population should be selected

In connection with the example cited, Dr. Taguchi further stated that "if the Unit Space is placed in the medium position, the accuracy of recognition will certainly be further enhanced." This case dates back to right after the inception of the MT Method, and the concept of placing the Unit Space in the medium position was not yet the current practice.

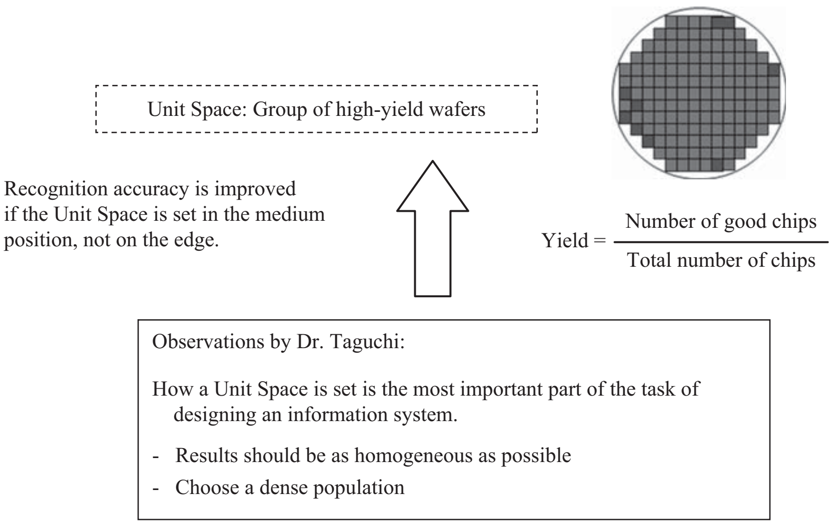


Figure 2.11. How to select a unit space source: Cf. 1) under *Sources*.

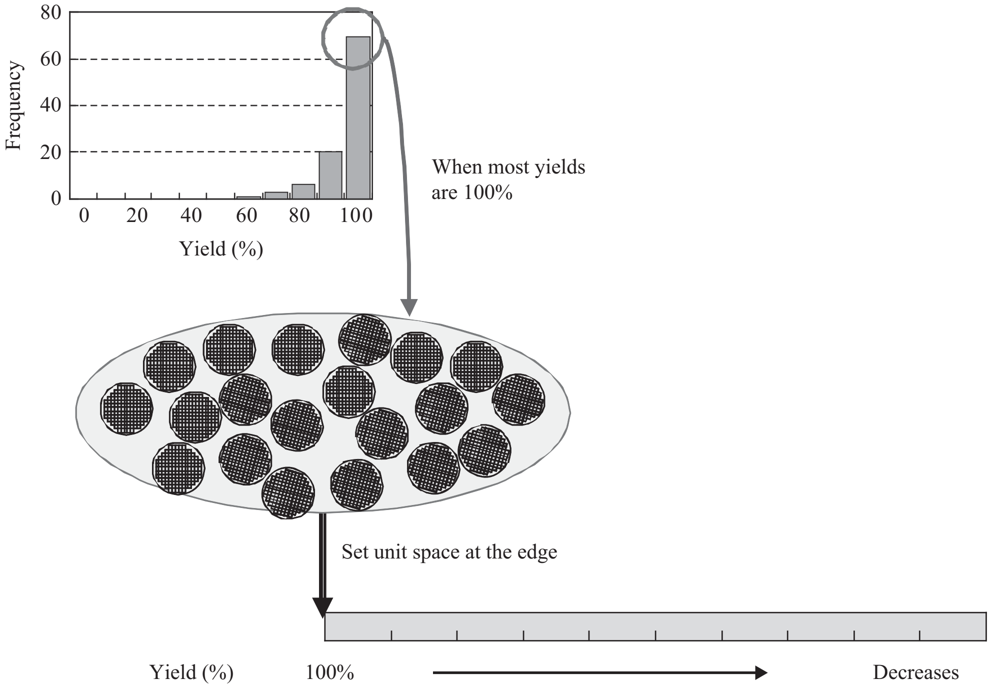


Figure 2.12. MT method-based approach.

Let's explain the concepts referred to in the above discussion in a bit more detail. In the graph in Figure 2.12, the *x*-axis represents the yield, and the *y*-axis, the frequency. If, as in the case at hand, the majority of the yields are 100%, the proper procedure is to place the Unit Space at the end because a concentration of homogeneous data exists close to the edge. The deviation from the pattern of the Unit Space placed at the end is then measured with a ruler that uses Mahalanobis Distance.

In the graph for Figure 2.13, as well, the *x*-axis represents the yield, and the *y*-axis, the frequency. In this graph, the number of yields is greatest where the yield is not 100%, but 70%. That means that the optimal placement of the Unit Space is in the vicinity of 70%. Three wafers or so are selected from the vicinity of the average value and set as the Unit Space, and average figures for the items and output (yield) are found. In some cases, one average sample should suffice as the Unit Space.

The T Method-1 is a method that performs the analysis with the Unit Space in the medium position, including the directionality of the +/– sign.

2.3.2 FEATURES OF THE RT METHOD (= T METHOD-3)

The T Method-3 is most appropriate when many Unit Spaces exist making it useful as a multipurpose recognition technique capable of handling character recognition, person recognition, image recognition, etc.

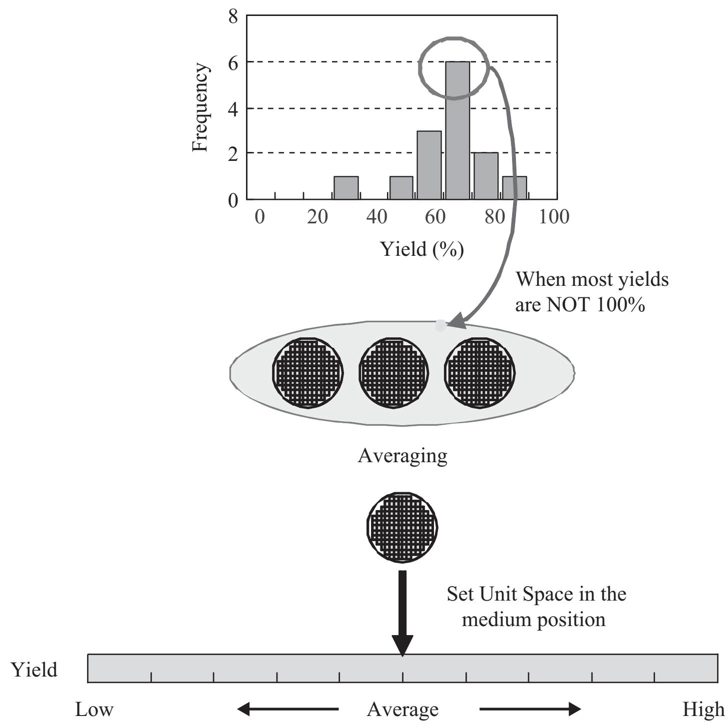


Figure 2.13. T Method-1-based approach.

This method converts the variables, regardless of how many there are, into a mere pair of variables. This feature, which makes do with a small amount of data for the Unit Space, makes it possible to define large numbers of Unit Spaces. The advantage of this reduced data size proves itself, for example, when the method is used for character recognition. For character recognition, large numbers of Unit Spaces, running to tens of thousands in the case of some languages, may become necessary in addition to the regular alphanumeric characters.

Use of alternative recognition techniques would be futile because an inordinate amount of data would nonetheless have to be dealt with, not to mention the high cost of processing time, collaterally incurred. Thus, the obvious merit of the RT Method is that, even if a large number of Unit Spaces must be dealt with, the amount of data processed per case is minimal.

The RT Method is free of all problems arising from the occurrence of multicollinearity and features the ability to perform computations even when the samples in the Unit Space are outnumbered by the variables (items).

The essence of the RT Method can be explained with the use of a simple example—fruit image pattern recognition. There is a variety of fruit, including apples, pears, persimmons, oranges, and many more, which makes it necessary to create an object in the Unit Space for each of the wide variety of possible fruits. Figure 2.14 shows what happens if the apple is to serve as the Unit Space.

For each sample in the Unit Space, the image data is summarized into two variables, and, with the use of the MT Method, the distance, D, from the center of each Unit Space is found for each

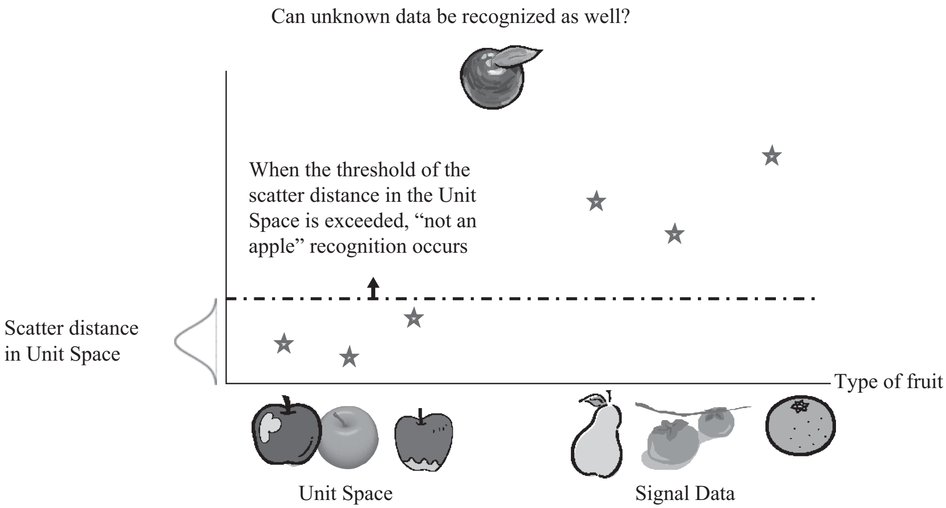


Figure 2.14. T Method-3-based approach.

individual sample of the Unit Space. Likewise, the image data for pears, persimmons, oranges, etc. can be summarized into two variables so that their distance, *D*, from the center of the Unit Space, can be found.

If distance *D* falls within the threshold or the bounds of scatter distance in the Unit Space (for a variety of apples), it will be recognized as an apple, and if it falls outside the threshold of the scatter distance, it will be recognized as a non-apple. Furthermore, if the image data of an unknown fruit (an apple, in this case) is brought in, whether it will be recognized as not a pear, persimmon, orange, etc., but correctly as an apple will be evaluated.

In the preceding paragraphs, we have explained the T Method-3 or RT Method with pattern recognition of fruit as an example. When a wide variety of abnormal products shows up at the actual

Note 2.4 Straightforward and problematic output definitions

For the T Method-1, "Signal Data" is required when determining an estimate equation. Signal Data is accompanied by output values, which correspond to the objective variables in multivariate analysis. In working with problems involving value estimation, estimated object values, or output values, are defined in addition to variables (items). For instance, if we are to estimate a temperature in a weather forecast, the temperature will be the output value; if we are to estimate an adequate price in a real estate transaction, the price will serve as the outcome value.

But defining output values proves difficult for the purpose of character recognition, degree of healthfulness determination, etc. For example, it may be possible to determine the ***order*** of "character readability" based on a comparison of characters, or the "seriousness of a head cold," etc., but it is difficult to assign values to these items (Figures 2.15 and 2.16). The same can be said of equipment diagnostics. In many cases, an order may be assignable to degrees of abnormality, but values cannot be assigned.

Note 2.4 Straightforward and problematic output definitions (Cont.)

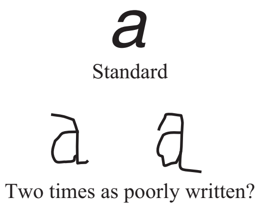


Figure 2.15. Handwriting ability.

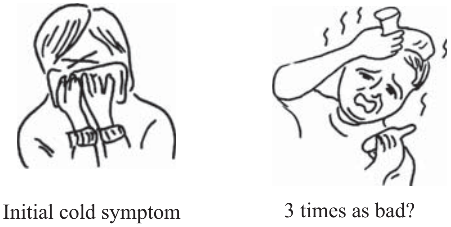


Figure 2.16. Degrees of illness.

The MT Method and RT Method may be utilized in cases where there is difficulty in assigning output values, and the T Method-1 in cases in which values *can* be assigned.

production site, if a Unit Space for each category of abnormality can be created, it is possible to apply this method to the analytical task of identifying a variety of types of abnormalities.

2.4 THE MT SYSTEM COMPUTATION FORMULAS

In this section, we will conduct a detailed discussion of the three most frequently utilized MT System-associated computation methods, namely the MT Method, the T Method-1, and the T Method-3.

2.4.1 COMPUTATION FORMULA FOR THE MT METHOD

2.4.1.1 Unit Data

Let us suppose that, as Unit Data, k variables and n samples have been acquired, as given in Table 2.5. The Table also shows the average value and standard deviation of each variable.

Table 2.5. Unit data

	Variable				
Sample No.	**1**	**2**	**3**	**…**	**k**
1	x_{11}	x_{12}	x_{13}	…	x_{1k}
2	x_{21}	x_{22}	x_{23}	…	x_{2k}
3	x_{31}	x_{32}	x_{23}	…	x_{3k}
…	…	…	…	…	…
N	x_{n1}	x_{n2}	x_{n3}	…	x_{nk}
Average	$\overline{x}_1$	$\overline{x}_2$	$\overline{x}_3$		$\overline{x}_k$
Standard deviation	σ_1	σ_2	σ_3	…	σ_k

2.4.1.2 Data Normalization

We will compute the normalized value X for each variable as follows:

$$X_{ij} = \frac{x_{ij} - \overline{x}_j}{\sigma_j} \quad \left(i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, k\right) \tag{2.9}$$

The normalized data is repeated in Table 2.6. The average value of each variable after normalization is 0, and its standard deviation is 1. Note that the *standard deviation* σ in the MT Method is found by the following computation formula in which all of the data form the population. Through this step, as explained under Item 1.3.2, the average value of the MD of the Unit Data will be 1 (one).

2.4.1.3 Computations of Correlation Matrices and Inverse Matrices

As concerns the k varieties of variables in Table 2.6, we will calculate the correlation matrix $\boldsymbol{R}$. A correlation matrix is a $k \times k$ square matrix.

$$\sigma = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n}} \tag{2.10}$$

Table 2.6. Normalized data

	Variable				
Sample No.	**1**	**2**	**3**	**…**	**k**
1	X_{11}	X_{12}	X_{13}	…	X_{1k}
2	X_{21}	X_{22}	X_{23}	…	X_{2k}
3	X_{31}	X_{32}	X_{23}	…	X_{3k}
…	…	…	…	…	…
n	X_{n1}	X_{n2}	X_{n3}	…	X_{nk}
Average	0	0	0	…	0
Standard deviation	1	1	1	…	1

$$\boldsymbol{R}=\begin{pmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{21} & 1 & \cdots & r_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ r_{k1} & r_{k2} & \cdots & 1 \end{pmatrix} \tag{2.11}$$

where

$$r_{ij}=\frac{\sum(X_{pi}\times X_{pj})}{n} \quad (p=1\ldots n) \tag{2.12}$$

Next, we calculate the inverse matrix $\boldsymbol{R}^{-1}$ of correlation matrix $\boldsymbol{R}$, and call it matrix $\boldsymbol{A}$.

$$\boldsymbol{A}=\boldsymbol{R}^{-1}=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{pmatrix} \tag{2.13}$$

2.4.1.4 Mahalanobis Distance Computation

Target data (unknown data) is represented as follows:

$$\boldsymbol{y}=(y_1,\ y_2,\ldots,y_k) \tag{2.14}$$

This data is normalized into Y in accordance with the equation given above (2.9):

$$\boldsymbol{Y} = \left(\mathrm{Y}_1, \mathrm{Y}_2, \ldots, \mathrm{Y}_k \right) \tag{2.15}$$

whereupon the Mahalanobis Distance is found through:

$$\begin{aligned} \mathrm{D}^2 &= \mathbf{YAY}^{\mathrm{T}}/k \\ &= \sum_{i=1}^{k} \sum_{j=1}^{k} a_{ij} Y_i Y_j / k \end{aligned} \tag{2.16}$$

With the MT Method, note that the Mahalanobis Distance (MD) obtained is given in a squared value, D^2. Unless otherwise noted, the expression in the form of a squared value will be used throughout the rest of this book.

2.4.2 COMPUTATION FORMULA FOR THE T METHOD-1[3]

The T Method-1 defines the Unit Space where the output value is in the medium position and homogeneous (densely populated).

The computation procedure of the T Method-1 is explained below.

2.4.2.1 Definition of the Unit Space and Computation of the Average of Relevant Items and Outputs

Let's suppose that, as shown in Table 2.7, *n* number of data have been obtained for the Unit Space. All the items of the data must be in same dimension as image density or must be no dimension data.

Table 2.7. Data for the Unit Space and averages values of the items and outputs

	Item/variable				
Data No.	**1**	**2**	**…**	***k***	**Output value**
1	x_{11}	x_{12}	…	x_{1k}	y_1
2	x_{21}	x_{22}	…	x_{2k}	y_2
…	…	…	…	…	…
n	x_{n1}	x_{n2}	…	x_{nk}	y_n
Average	$\overline{x}_1$	$\overline{x}_2$	…	$\overline{x}_k$	$\overline{y} = M_0$

From the n number of samples in the Unit Space, we find average values $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_k$ and average output value $\overline{y} = M_0$ for all items. Accordingly, the average values work out as follows:

$$\overline{x}_j = \frac{1}{n}\left(x_{1j} + x_{2j} + \cdots + x_{nj}\right) \quad (j = 1, 2, \ldots, k) \tag{2.17}$$

$$\overline{y} = M_0 = \frac{1}{n}\left(y_1 + y_2 + \cdots + y_n\right) \tag{2.18}$$

2.4.2.2 Definition of Signal Data

All data items marked l, left unselected for the Unit Space are treated as Signal Data. Signal Data is shown in Table 2.8. "Signal Data" refers to all data used for finding the proportional coefficient β and SN ratio η, which will be discussed in Section 2.4.2.4.

Table 2.8. Signal data

	Item/variable				
Data No.	**1**	**2**	**…**	***k***	**Output value**
1	x'_{11}	x'_{12}	…	x'_{1k}	y'_1
2	x'_{21}	x'_{22}	…	x'_{2k}	y'_2
…	…	…	…	…	…
l	x'_{l1}	x'_{l2}	…	x'_{lk}	y'_l

2.4.2.3 Normalization of Signal Data

Signal Data is normalized using the average values of items and the output values of samples in the Unit Space. Normalization is performed by subtracting the average value $\overline{x}_j$ of item j in the Unit Space from value x'_{ij} of item j of the i-th Signal Data.

$$X_{ij} = x'_{ij} - \overline{x}_j \quad (i = 1, 2, \ldots, l) \quad (j = 1, 2, \ldots, k) \tag{2.19}$$

Likewise, normalization is performed by subtracting average value M_0 of the output from the Unit Space from output value y'_i of the i-th Signal Data.

$$M_i = y_i' - M_0 \quad (i = 1, 2, \ldots, l) \tag{2.20}$$

Normalized Signal Data is shown in Table 2.9.

Table 2.9. Normalized signal data

	Item/variable				
Data No.	**1**	**2**	**…**	***k***	**Output value**
1	X_{11}	X_{12}	…	X_{1k}	M_1
2	X_{21}	X_{22}	…	X_{2k}	M_2
…	…	…	…	…	…
l	X_{l1}	X_{l2}	…	X_{lk}	M_l

2.4.2.4 Computation of Proportional Coefficient β and SN Ratio η (in Duplicate Ratio) for All Items

We will next compute proportional coefficient β and SN ratio η for all items. How the computation is performed is explained with item 1 as an example:

$$\text{Proportional coefficient} \quad \beta_1 = \frac{M_1 X_{11} + M_2 X_{21} + \cdots + M_l X_{l1}}{r} \tag{2.21}$$

$$\text{SN ratio } \eta_1 = \begin{cases} \dfrac{\frac{1}{r}\left(S_{\beta 1} - V_{e1}\right)}{V_{e1}} & \left(\text{when } S_{\beta 1} > V_{e1}\right) \\ 0 & \left(\text{when } S_{\beta 1} \leq V_{e1}\right) \end{cases} \tag{2.22}$$

where:

$$\text{Effective divider} \quad r = M_1^2 + M_2^2 + \cdots + M_l^2 \tag{2.23}$$

$$\text{Total variation} \quad S_{T1} = X_{11}^2 + X_{21}^2 + \cdots + X_{l1}^2 \quad (\mathrm{f} = l) \tag{2.24}$$

$$\text{Variation of Proportional term} \quad S_{\beta 1} = \frac{\left(M_1 X_{11} + M_2 X_{21} + \cdots M_l X_{l1}\right)^2}{r} \quad (\mathrm{f} = 1) \tag{2.25}$$

$$\text{Error variation} \quad S_{e1} = S_{T1} - S_{\beta 1} \tag{2.26}$$

$$\text{Error variance} \quad V_{e1} = \frac{S_{e1}}{l-1} \tag{2.27}$$

From item 2 up to item k, we will likewise find proportional coefficient β and SN ratio η. This operation yields the results that are shown in Table 2.10.

Table 2.10. Proportional coefficient β and SN ratio η, item by item

	Item/variable			
B, η	**1**	**2**	**…**	***k***
Proportional β	β_1	β_2	...	β_k
SN ratio η	η_1	η_2	...	η_k

2.4.2.5 Computation, Signal by Signal, of Integrated Estimate Value $\hat{M}$ of Output

An item-by-item estimated value is found for each piece of Signal Data using the proportional coefficient β and SN ratio η, item by item. The estimated value of the output of item 1 for the i-th Signal Data is:

$$\hat{M}_{i1} = \frac{X_{i1}}{\beta_1} \tag{2.28}$$

An estimation is likewise made of item 2 through item l for the i-th Signal Data. And finally an integration of the result is performed by weighting it with $\eta_1, \ldots, \eta_l$, which is the estimated measure of precision of each item.

Thus, the integrated estimate value $\hat{M}_i$ of the output of the i-th Signal Data becomes:

$$\hat{M}_i = \frac{\eta_1 \times \dfrac{X_{i1}}{\beta_1} + \eta_2 \times \dfrac{X_{i2}}{\beta_2} + \cdots + \eta_k \times \dfrac{X_{ik}}{\beta_k}}{\eta_1 + \eta_2 + \cdots + \eta_k} \quad \left(i = 1, 2, \ldots, l\right) \tag{2.29}$$

Table 2.11 shows the real values (measured values) of the Signal Data $M_1, M_2, \ldots, M_l$ and the integrated estimate values $M_1, M_2, \ldots, M_l$.

Table 2.11. Measured values and integrated estimate values of signal data

Data No.	Measured value	Integrated estimate value
1	M_1	$\hat{M}_1$
2	M_2	$\hat{M}_2$
:	:	:
:	:	:
l	M_l	$\hat{M}_l$

2.4.2.6 Computation of Integrated Estimate SN Ratio

The Integrated Estimate SN Ratio is computed using the following equation based on Table 2.11. The result of the computation will be used in Section 2.4.2.7 "Evaluation of the Relative Importance of an Item," comes up as the next step.

$$\text{Integrated Estimate SN Ratio} \quad \eta = 10\log\left(\frac{\frac{1}{r}\left(S_\beta - V_e\right)}{V_e}\right) \quad (db) \tag{2.30}$$

where:

$$\text{Linear equation} \quad L = M_1\hat{M}_1 + M_2\hat{M}_2 + \cdots + M_l\hat{M}_l \tag{2.31}$$

$$\text{Effective divider} \quad r = M_1^2 + M_2^2 + \cdots + M_l^2 \tag{2.32}$$

$$\text{Total variation} \quad S_T = \hat{M}_1^2 + \hat{M}_2^2 + \cdots + \hat{M}_l^2 \quad (\mathrm{f} = l) \tag{2.33}$$

$$\text{Variation of proportional term} \quad S_\beta = \frac{L^2}{r} \quad (\mathrm{f}=1) \tag{2.34}$$

$$\text{Error variation} \quad S_e = S_T - S_\beta \quad (\mathrm{f}=l-1) \tag{2.35}$$

$$\text{Error variance} \quad V_e = \frac{S_e}{l-1} \tag{2.36}$$

2.4.2.7 Evaluation of the Relative Importance of an Item

The relative importance of an item is evaluated in terms of the extent to which the Integrated Estimate SN Ratio ratio deteriorates when the item is not used. For the evaluation, a two-level orthogonal array (a 4 × prime version of the two-level series is advisable) is used. Use of an orthogonal array allows a comparison to be made of the SN ratio η of the integrated estimate under various conditions.

Let's suppose we have 11 items before us, $X_1, X_2, \ldots, X_{11}$. We assign the 11 items to Columns 1 to 11 in Table 2.12.

Table 2.12. Orthogonal array L_{12} and assignment of items

	Item No.											
	X_1	**X_2**	**X_3**	**X_4**	**X_5**	**X_6**	**X_7**	**X_8**	**X_9**	**X_{10}**	**X_{11}**	**Integrated Estimate**
No.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**SN Ratio η(db)**
1	1	1	1	1	1	1	1	1	1	1	1	η_1
2	1	1	1	1	1	2	2	2	2	2	2	η_2
3	1	1	2	2	2	1	1	1	2	2	2	η_3
4	1	2	1	2	2	1	2	2	1	1	2	η_4
5	1	2	2	1	2	2	1	2	1	2	1	η_5
6	1	2	2	2	1	2	2	1	2	1	1	η_6
7	2	1	2	2	1	1	2	2	1	2	1	η_7
8	2	1	2	1	2	2	2	1	1	1	2	η_8
9	2	1	1	2	2	2	1	2	2	1	1	η_9
10	2	2	2	1	1	1	1	2	2	1	2	η_{10}
11	2	2	1	2	1	2	1	1	1	2	2	η_{11}
12	2	2	1	1	2	1	2	1	2	2	1	η_{12}

Where the two levels of the orthogonal ray mean the following:

Level 1: Item will be used.

Level 2: Item will *not* be used.

In Test 1, with all Columns, Column 1 through Column 11, being on Level 1, it is shown that items X_1, X_2, ..., X_{11} will all be used, and that the SN ratio (db) of the integrated estimate works out to η_I. Note in passing that the SN ratio η(db) of the integrated estimate computed in Section 2.4.2.6 is used for η_1 in Table 2.12.

In Test 2, with Columns 1 through 5 being on Level 1, it is shown that five items, X_1, X_2, …, X_5 will be used and that the SN ratio (db) of the integrated estimate come out to be η_2. In the same manner as in Test 12, it is shown that, with Columns 3, 4, 6, 8, and 11 being on Level 1, five items, X_2, X_4, X_6, X_8, and X_{11}, are used, and the SN ratio (db) of the integrated estimate work out to be η_{12}.

With regard to the SN ratio of the integrated estimate, we find the difference between the averages of the SN ratio for Level 1 (with the items to be used) and that for Level 2 (with the items *not* to be used), item by item, and on that basis determine the relative importance of the items.

2.4.2.8 Computation of the Integrated Estimate Value of Unknown Data

Item-by-item normalization of data with unknown output value in the Unit Space is performed through the use of the average value, x_j(f = 1, 2, …, k). Computation is performed on the integrated estimate value, $\hat{M}$, of the normalized unknown data as in the case of Signal Data.

2.4.2.9 Computation of the Pre-Normalization Integrated Estimate Value $\hat{y}$ based on the Integrated Estimate Value $\hat{M}$

Now that the output value of the Signal Data has been normalized, the integrated estimate value $\hat{y}$ of the pre-normalization (actual) output is computed on the basis of the normalized integrated estimate value $\hat{M}$, as follows:

$$\hat{y}_i = \hat{M}_i + M_0 \quad (i = 1, 2, \ldots, l) \tag{2.37}$$

The same method may be used for unknown data in computing the integrated estimate value $\hat{y}$ of the pre-normalization (actual) output.

2.4.3 COMPUTATION FORMULA FOR THE RT METHOD (T METHOD-3)

The T Method-3 is the T Method for recognition that has the ability to classify objects into two categories, one inside and the other outside the Unit Space. Unlike the T Method-1, the T Method-3 proves of use when the real value (measured value of the output) of a signal is unknown, but the category to which it belongs is clear and when multiple Unit Spaces (more than just one) exist. The procedure for the method's use is described below:

2.4.3.1 Definition of the Unit Space, and Computation of the Average Value of Each Item

Let's suppose that, as shown in Table 2.13, *n* number of samples have been acquired for the Unit Space. The k-items have to be in same dimension or in no-dimension. For example, image data is in same dimension.

Table 2.13. Average values of samples and items in the unit space

Data No.	Item/variable 1	2	…	*k*	Linear formula L
1	x_{11}	x_{12}	…	x_{1k}	L_1
2	x_{21}	x_{22}	…	x_{2k}	L_2
…	…	…	…	…	…
n	x_{n1}	x_{n2}	…	x_{nk}	L_n
Average	$\overline{x}_1$	$\overline{x}_2$	…	$\overline{x}_k$	

We then find the average values $\overline{x}_1, \overline{x}_2, \ldots, \vec{x}_k$ for each item, from *n* number of samples in the Unit Space. These average values are shown in Table 2.14.

$$\overline{x}_j = \frac{1}{n}\left(x_{1j} + x_{2j} + \cdots + x_{nj}\right) \quad (j = 1, 2, \ldots, k) \tag{2.38}$$

Linear formula *L* can be found through formula (2.40), procedure (2.4.3.2) in the next paragraph.

2.4.3.2 Computation of Sensitivity β and Standard SN Ratio η, Sample by Sample, from the Unit Space

We explain below how to find the sensitivity β and standard SN ratio η of the first sample in the Unit Space.

$$\text{Sensitivity} \quad \beta_1 = \frac{L_1}{r} \tag{2.39}$$

where:

$$\text{Linear equation} \quad L_1 = \bar{x}_1 x_{11} + \bar{x}_2 x_{12} + \cdots + \bar{x}_k x_{1k} \tag{2.40}$$

$$\text{Effective divider} \quad r = \bar{x}_1^2 + \bar{x}_2^2 + \cdots + \bar{x}_k^2 \tag{2.41}$$

Standard SN ratio η_1 is computed after, as shown below, all variations, S_T, variation of the proportional item S_β, etc. have first been found.

$$\text{Total variations} \quad S_{T1} = x_{11}{}^2 + x_{12}{}^2 + \cdots + x_{1k}{}^2 \quad (\mathrm{f} = \mathrm{k}) \tag{2.42}$$

$$\text{Variation of proportional term} \quad S_{\beta 1} = \frac{L_1^2}{r} \quad (\mathrm{f} = 1) \tag{2.43}$$

$$\text{Error variation} \quad S_{e1} = S_{T1} - S_{\beta 1} \quad (\mathrm{f} = \mathrm{k} - 1) \tag{2.44}$$

$$\text{Error variance} \quad V_{e1} = \frac{S_{e1}}{k-1} \tag{2.45}$$

Accordingly, standard SN ratio $\boldsymbol{\eta}_1$ is given as the following formula:

$$\text{Standard SN ratio} \quad \eta_1 = \frac{1}{V_{e1}} \tag{2.46}$$

Use of the expression “standard SN ratio” reflects the fact that we have decided for these purposes to treat the average values of the items of the Unit Space as the standard signals; the dividend in formula (2.46) might as well be represented by r, but the numeral 1 (one) has been chosen because it is common to all members.

Sample by sample in the Unit Space, sensitivity β and standard SN ratio η are then found in a similar manner, with the result shown in Table 2.14.

2.4.3.3 Computation of Two Variables Y_1 and Y_2, for Each Sample in the Unit Space

Using the two items, sensitivity β and standard SN ratio η in Table 2.14, the two variables Y_1 and Y_2 are computed. For the new variable Y_1, β, is used, as is. Y_2 will first be converted as follows

Table 2.14. Sensitivity β and standard SN ratio η (duplicate ratio) for all samples in the unit space

Data No.	Sensitivity β	SN ratio η
1	β_1	η_1
2	β_2	η_2
:	:	:
:	:	:
n	β_n	η_n

to allow an evaluation of any scatter from the standard conditions (Unit Space item by item average value in Table 2.13):

$$Y_{i1} = \beta_i \quad (i = 1, 2, \ldots, n) \tag{2.47}$$

$$Y_{i2} = \frac{1}{\sqrt{\eta_i}} = \sqrt{V_{e_i}} \quad (i = 1, 2, \ldots, n) \tag{2.48}$$

Y_1 and Y_2 are computed for all the samples of the Unit Space to find the average value for either of them. The results are displayed in Table 2.15.

$$\overline{Y}_1 = \frac{1}{n}\left(Y_{11} + Y_{21} + \cdots + Y_{n1}\right) \tag{2.49}$$

$$\overline{Y}_2 = \frac{1}{n}\left(Y_{12} + Y_{22} + \cdots + Y_{n2}\right) \tag{2.50}$$

2.4.3.4 Computation of Distances of Samples in the Unit Space

Based upon the above steps, the k number of items are understood to have been condensed into two variables, Y_1 and Y_2, with respect to the individual samples of the Unit Space. Next, using the MT Method, we will measure the Mahalanobis Distance, D, of the individual samples of the Unit Space from the center of the Unit Space $(\overline{Y}_1, \overline{Y}_2)$. How to find Mahalanobis Distances is described in Sections 2.2.2 and 2.4.1. Distance measurements of samples in the Unit Space taken individually using the MT Method are shown in Table 2.16.

Table 2.15. Y_1 and Y_2 for all samples in the unit space

Data No.	Y_1	Y_2
1	Y_{11}	Y_{12}
2	Y_{21}	Y_{22}
:	:	:
:	:	:
n	Y_{n1}	Y_{n2}
Average	$\bar{Y}_1$	$\bar{Y}_2$

Table 2.16. Distances of samples in unit space

Data No.	Distance D^2	Distance D
1	D_1^2	D_1
2	D_2^2	D_2
…	…	…
n	D_n^2	D_n

2.4.3.5 Definition of Signal Data

Let's suppose that we have acquired l amount of Signal Data, as shown in Table 2.17, toward evaluating discriminating ability. Linear formula L can be found with formula (2.52) given in Section 2.4.3.6.

Table 2.17. Signal data items and linear formula

	Item/variable				Linear
Data No.	1	2	…	k	equation L'
1	x'_{11}	x'_{12}	…	x'_{1k}	L'_1
2	x'_{21}	x'_{22}	…	x'_{2k}	L'_2
…	…	…	…	…	…
l	x'_{l1}	x'_{l2}	…	x'_{lk}	L'_l

2.4.3.6 Computation of Sensitivity β and Standard SN Ratio η (Duplicate Ratio), Individually, for Each Piece of Signal Data

We will explain next how to find sensitivity β and standard SN ratio η for the first Signal Data item.

$$\text{Sensitivity} \quad \beta_1 = \frac{L_1'}{r} \tag{2.51}$$

where:

$$\text{Linear equation} \quad L'_1 = \bar{x}_1 x'_{11} + \bar{x}_2 x'_{12} + \cdots + \bar{x}_k x'_{1k} \tag{2.52}$$

$$\text{Total variations} \quad S_{T1} = x'^2_{11} + x'^2_{12} + \cdots + x'^2_{1k} \quad (\mathrm{f=k}) \tag{2.53}$$

$$\text{Variation of proportional term} \quad S_{\beta 1} = \frac{L_1'^2}{r} \quad (\mathrm{f=1}) \tag{2.54}$$

$$\text{Error variation} \quad S_{e1} = S_{T1} - S_{\beta 1} \quad (\mathrm{f = k-1}) \tag{2.55}$$

$$\text{Error variance} \quad V_{e1} = \frac{S_{e1}}{k-1} \tag{2.56}$$

Accordingly, the standard SN ratio η_1 is given as the following equation:

$$\text{Standard SN ratio (duplicate ratio)} \quad \eta_1 = \frac{1}{V_{e1}} \tag{2.57}$$

Sensitivity β and standard SN ratio η are found for each set of Signal Data individually, as shown in Table 2.18.

Table 2.18. Sensitivity β and standard SN ratio η (duplicate ratio) for each signal data

Data No.	**Sensitivity *β***	**SN ratio *η***
1	β_1	η_1
2	β_2	η_2
:	:	:
:	:	:
l	β_l	η_l

2.4.3.7 Computation of Y_1 and Y_2 for Each Separate Signal Data Set

Using the two categories of sensitivity β and standard SN ratio η in Table 2.18, Y_1 and Y_2 are computed. Here, again, the new variable Y_1 will retain sensitivity β in its intact form, while Y_2 will not retain standard SN ratio η in its intact form but will first undergo conversion as follows to allow an evaluation of any irregular deviation from the standard condition (Unit Space category-by-category average value, as in Table 2.13):

$$Y_{il} = \beta_i \ \left(i = 1, 2, \ldots, l\right) \tag{2.58}$$

$$Y_{i2} = \frac{1}{\sqrt{\eta_i}} = \sqrt{V_{ei}} \ \left(i = 1, 2, \ldots, l\right) \tag{2.59}$$

Y_1 and Y_2 are computed for all Signal Data. The result of the computation is as shown in Table 2.19.

Table 2.19. Y_1 and Y_2 for each separate signal

Data No.	$\mathbf{Y_1}$	$\mathbf{Y_2}$
1	Y_{11}	Y_{12}
2	Y_{21}	Y_{22}
:	:	:
:	:	:
l	Y_{l1}	Y_{l2}

2.4.3.8 Computation of Mahalanobis Distance of Each Separate Signal Data Set

Following the preceding operations, the k variables of individual Signal Data sets have been reduced to Y_1 and Y_2. Next, using the MT Method, Mahalanobis Distance D from the center of the Unit Space $\left(\overline{Y}_1, \overline{Y}_2\right)$ is found for each individual Signal Data set (Table 2.20).

2.4.3.9 Computation of Distance D from the Center of the Unit Space to Unknown Data

With unknown data just as with Signal Data, sensitivity β and standard SN ratio η are computed with the use of the average value of each item in the Unit Space. The two variables Y_1 and Y_2 are

Table 2.20. Distance of each signal data item

Data No.	Distance D^2	Distance D
1	D_1^2	D_1
2	D_2^2	D_2
…	…	…
l	D_l^2	D_l

computed from sensitivity β and standard SN ratio η of the unknown data; then, distance D from the center of the Unit Space $(\bar{Y}_1, \bar{Y}_2)$ is computed.

2.4.3.10 Evaluation of Discrimination Ability Based on Standard Deviation σ of the Unit Space

The method described below provides one means of evaluating ability to discriminate:

Standard deviation σ is found from the Mahalanobis Distance of each sample in the Unit Space, and an evaluation is performed to see what values are yielded by the distances of signals or unknown data in relation to σ.

For instance, if the distance of the unknown data falls within 2σ, then it will be determined to belong to the Unit Space; if it is in excess of 2σ, then it will not. It will be necessary, on a case-by-case basis, to study and determine whether 2σ is adequate as the threshold.

The following formulas are used to find the standard deviation, σ, of a Unit Space.

$$\sigma^2 = \bar{D}^2 = \frac{1}{n}\left(D_1^{\,2} + D_2^{\,2} + \cdots + D_n^{\,2}\right) \tag{2.60}$$

$$\sigma = \bar{D} = \sqrt{\frac{1}{n}\left(D_1^{\,2} + D_2^{\,2} + \cdots + D_n^{\,2}\right)} \tag{2.61}$$

$$\sigma = \sqrt{\frac{1}{n}\left\{\left(D_1 - 0\right)^2 + \left(D_2 - 0\right)^2 + \cdots + \left(D_n - 0\right)^2\right\}} = \bar{D} \tag{2.62}$$

The reason is that, standard deviation σ can be found using the following equation since the distance of the center of the Unit Space is zero.

CHAPTER 3

Data Handled by the MT System and Feature Extraction

For pattern recognition purposes, measured data are utilized in an unmodified form in some cases, but in many cases feature values that are extracted from measured data are utilized. Especially in cases where time-series data or image data are being handled, feature extraction outcome plays a major role in determining the success or failure of pattern recognition. In this chapter, we explain the types of variables handled by the MT System and feature extraction.

3.1 USE OF MEASURED VALUES IN AN UNMODIFIED FORM

Table 3.1 shows an example of Unit Space data pertaining to healthy people collected from a medical examination. Across rows, information such as age, gender, and biochemical inspection items has been entered, and, going down, in the columns, there are inspection object sample numbers. Table 3.1 contains variables that assume only one of two states such as gender, namely, *category data*.

When data of this type is used, it is common practice to employ measured values in their raw form. In the field of industrial production as well, measured values pertaining to individual mass-produced products are used, unmodified in many cases.

3.2 PERFORMING FEATURE EXTRACTION

When we look at something, we are *characterizing* the object. When we see the character in Figure 3.1, which stands for the number "five," we are processing in our mind something along the lines of: "a horizontal bar at the top; a circle at the bottom with an opening on its left side, etc." If the opening of the circle were closed, the character might be readable as "six." What we

Table 3.1. Example of physical examination data

Case No.	1 Age	2 Sex	3 TP	4 Alb	5 ChE	… …	12 TCh	13 TG	14 PL	15 Cr	16 BUN	17 UA
1	35	1	8.0	5.5	526	…	171	55	186	1.0	12	4.3
2	48	1	7.9	5.2	520	…	183	63	196	1.0	17	3.2
3	45	1	8.1	5.5	594	…	192	66	202	0.8	13	2.9
4	40	1	8.3	5.6	486	…	179	77	196	1.0	13	2.9
…	…	…	…	…	…	…	…	…	…	…	…	…
197	56	10	8.6	5.5	630	…	188	134	215	1.5	19	3.9
198	46	10	8.3	5.7	589	…	190	74	200	1.7	18	4.5
199	57	10	8.4	5.4	612	…	138	85	178	1.4	18	5.2
200	34	10	8.2	5.6	294	…	157	71	174	1.4	13	6.3

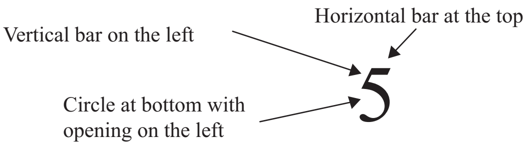


Figure 3.1. Feature expressions of the numeral.

are doing here, in other words, is extracting features (characteristics) from the object character while referring it to, and comparing it against, a pattern that we learned in our early years.

In the case of human beings, feature extraction is either a genetically acquired, inborn ability (*cf*. Note 3.1), or a learned ability acquired through study. Whichever the case, we perform feature extraction processing in our brains unconsciously.

In having the computer execute pattern recognition, it is up to us humans to provide the characterization rules. It has been said in the matter of pattern recognition by computer that “there is no royal road to characterization.” Historically, for each of the problems at hand, a feature specific to it has had to be defined. It would be ideal if a characterization method could be worked out that were universally applicable to as many different kinds of tasks as possible. Under the umbrella of the MT System, a highly multipurpose method that also competently addresses the characterization issue has been proposed.

When we draw a picture, we normally start by drawing a line. When portraying a face, cloud, or tree, as it turns out, we unconsciously perform feature extraction. A person’s cognitive ability is not restricted to pattern recognition in the sense of reference processing, but it extends further afield to include a wider gamut of information processing we refer to as “feature extraction.”

It is human beings that “teach” the computer how to perform feature extraction or pattern recognition. There is ongoing two-way contact between scientific research into the brain’s

Note 3.1 The feature extraction ability of higher level animals

The feature extraction ability of higher-level animals, including man, is an ability acquired over the long course of the evolution of life. Tests using monkeys have shown that in their brain (nerve cells) there is a specific portion that reacts in a specific manner to the sight of a certain shape. For instance, when they were shown a set of lips and a small wavy bell pepper, an identified portion of their nerve cells was observed to show a common reaction to both objects.[4]

Furthermore, it has been elucidated that this particular reaction of the monkeys occurs to a horizontally placed, elliptical shape with a dark line dividing the object into upper and lower, lighter-shaded portions. A glance at the images in Figure 3.2 shows how the two objects resemble each other.

Lips Green pepper

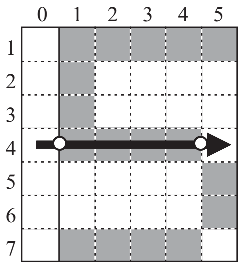
Figure 3.2. Lips and green pepper.

information-processing ability and engineering research dedicated to the development of information-processing technology that can emulate the brain, and this rapport between the two will last forever. It is said that "science pursues one truth while technology finds one solution from endless possibilities." If science is to end as a science, it will end in intellectual satisfaction only. As for the role of the engineering expert, it may be said that continued, relentless pursuit, based on scientific accomplishments, of the best of all the limitlessly available technical solutions is the best fit.

3.3 FEATURE EXTRACTION TECHNIQUE FROM CHARACTER PATTERN

In the MT System, variation value (differential characteristics) and abundance value (integral characteristics) have been proposed as feature extraction techniques.[5] These feature values are highly versatile and lend themselves to the solution of a great variety of tasks. In this section, taking character pattern as an example, we will study how to define them and what significance, etc. they may have.

Figure 3.3 shows the numeral "5" laid out on a 5-column-by-7-row grid. If we direct our attention to the black arrow traversing the fourth row, we notice that there are two places marked with a small circle where a white-to-gray or gray-to-white switch occurs. The number of such switches, we will define as so many instances of "variation."



Variation	Abundance
1	5
2	1
2	1
2	4
1	1
1	1
2	4

Figure 3.3. Variation and abundance in character patterns.

Next, we count the gray squares that the black arrow strikes through. There are four of them. This number, we label as an index of "abundance." Note that, when counting the instances of variation in the case at hand, we assume a "zero" column to the left of the arrow. Furthermore, we stipulate that the tip of the arrow shall stop at the fifth column. These are stipulations necessary to our study and operations; it is known that these stipulations contribute to the amplification of information to support pattern recognition.

After conducting similar count processing on all the seven rows, we obtain results, which are tabulated in the right-hand chart in Figure 3.3. That is to say, the character pattern for "5" has been characterized in terms of a total of 14 numerical values. The "straight, horizontal bar at the top" is described by a combination of indices of abundance and variation on the first row. The characteristics of a circle open on its left side are described with indices of variation on the fourth through seventh rows.

And as we experiment with various other patterns, we are made aware that variation and abundance actually make up a well thought-out method of feature extraction from character patterns and image data. The characterization processing of which we have just seen one example makes it possible to *transfer processing-ready character patterns to the computer.*

"Variation and abundance" is a truly multipurpose feature extraction technique applicable also to waveform detection, which we discuss in the next section. The features dealt with here have been referred to as "differential" and "integral" characteristics, but on these pages we will refer to them in terms of "variation" and "abundance," respectively. We will take a look at some specific examples of MT Method applications that make use of feature extraction from characters in Chapter 4.

3.4 FEATURE EXTRACTION TECHNIQUE FROM WAVEFORM PATTERN

Many phenomena are measurable as waveform patterns. With oscillation, all measured values that change through time can be treated as waveforms. Economic indicators such as stock prices presented in waveform illustrations are all too familiar to us. The technique of extracting features inherent to waveforms as numerical values is in fact a key technique providing solutions to issues associated with monitoring.

3.4.1 DEFINITIONS OF VARIATION AND ABUNDANCE AS THEY RELATE TO WAVEFORM PATTERNS

3.4.1.1 Past Waveform Features

The features of waveforms include frequency and amplitude. As shown in Figure 3.4, categories such as average frequency, magnitude of oscillation, and maximum magnitude of oscillation have been commonly used. Frequency analysis (FFT: Fast Fourier Transform), wavelets, etc., which will be discussed in Section 3.6, have also been common. Frequency analysis is a technique of expressing the characteristics of waveforms in terms of a frequency axis and an energy axis. A wavelet is a technique of expressing the characteristics of waveforms in terms of a time axis, in addition to the frequency axis and the energy axis.

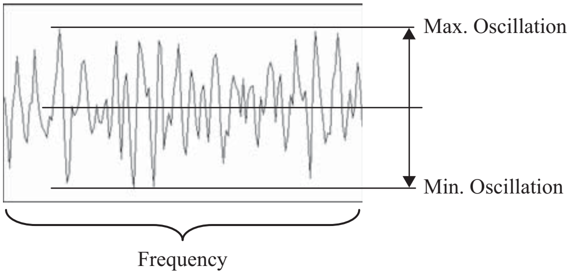


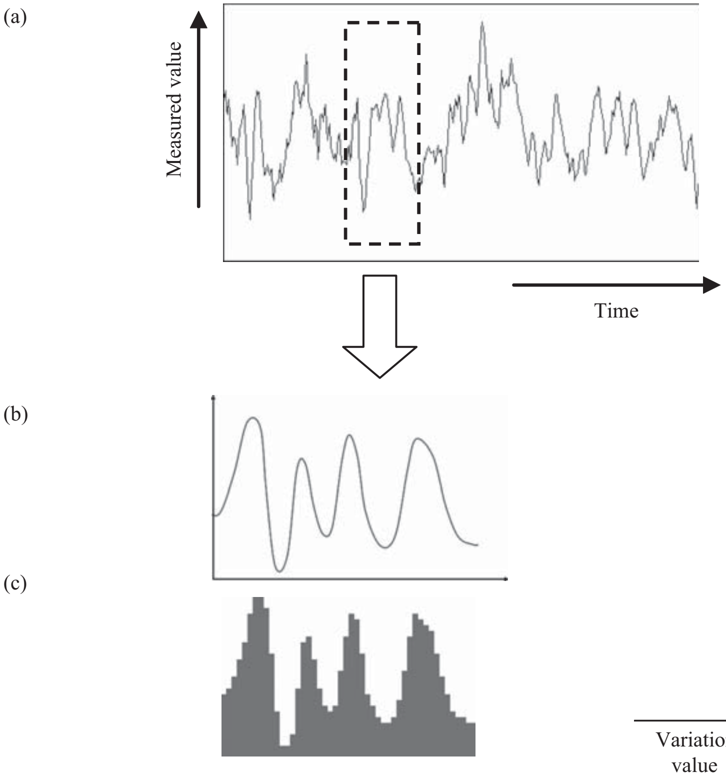
Figure 3.4. Original waveform, oscillation, and frequency.

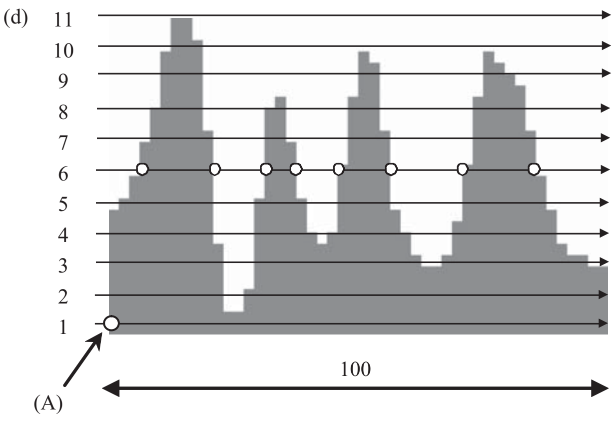
Frequency analysis and wavelets are excellent methods of explaining the characteristics of waveforms. Nonetheless, it cannot quite be said that both convey sufficient information on the characteristics of the given waveform pattern. For instance, frequency analysis is a processing method dealing with waveforms over a relatively long period of time and is therefore not adept at capturing waveform changes that may be occurring within short time spans. Furthermore, both frequency analysis and wavelets presuppose the human operator's evaluation and judgment of the results they produce. In other words, what is produced as a result of characterization is a two-dimensional or three-dimensional graph, beyond which, in many cases, all relevant work is deferred to human judgment.

3.4.1.2 Extraction of Variation and Abundance Information from Waveform Patterns

The variation and abundance information extraction method was proposed as a means of expressing the characteristics of waveform patterns in more accurately quantified terms, and it was based on exactly the same concept as character pattern processing. The method is explained below.

As in Figure 3.5 (a), a domain with an appropriate width and height is defined within the bounds of the waveform. An enlarged view of the same domain is image (b). If we fill in the body of the





Variation value	Abundance value
0	0
2	6
6	20
8	28
8	38
8	47
8	65
8	78
6	88
3	95
1	100

Figure 3.5. Waveform pattern; variation and abundance values.

waveform with gray, we obtain a graph looking like (c). When we process waveform data on the computer, the data will be converted into a digital version, and, to make that point easier to understand, (c) is represented in a staircase-shaped graph configuration.

Next, we define, as in (d), multiple, horizontal straight lines as reference lines. In the example at hand, 11 reference lines are defined and numbered 1 to 11 in ascending order, starting from the bottom. We then count the points at which each of the reference lines cross the contour of the waveform. Focusing on the sixth reference line (line No. 6), we encircle the points at which the graph waveform intersects it. We count eight circles. We will refer to this number of intersections as the "variation value." The sum of those segments of a given reference line that are covered by the shaded areas of the graph will be referred to as the *abundance value*. Note in passing that, in this example, the width of the domain is set at 100 so that the abundance value for the first reference line is 100.

Also, as in the example of numbers, a "waveform-free, blank domain" is assumed to exist to the left of the waveform for the purpose of counting the variation values so that the number of variation values on reference line No. 1, for example, will be 1. In other words, one episode of intersection occurs at the point marked A. On the other reference lines, as well, if a waveform already exists at the point of departure, then one instance of change has already occurred at the leftmost end of the domain.

Reading the variation values and abundance values from all the reference lines, we tabulate the result as shown in the table on the right side, under (d). The numbers in the rows in the table reflect their corresponding reference line readings; waveform and numbers can be easily checked. There are 11 reference lines, and with two feature values, the variation value and the abundance value is collected for each and a total of 22 quantified feature values are acquired. These 22 numerical values together provide the feature value set of the waveform.

Since a waveform pattern is a continual data string, in this case, 22 pieces of numerical information can be sequentially acquired by sliding the domain defined in Figure 3.5 (a) in the direction of the horizontal axis. If the waveform shows a regular oscillation, all the numerical information obtained by moving the domain in this manner can be defined as Unit Data.

3.4.2 MEANING OF VARIATION VALUE AND ABUNDANCE VALUE IN A WAVEFORM PATTERN

A variation value can be considered a further developed version of the **zero cross** concept that has been in use in such fields as signal processing. A zero cross refers to the point on the $y=0$ axis at which the waveform intersects it, as shown in Figure 3.6, and the number of intersections approximately corresponds to the same meaning as frequency. The variation value is a definition of the zero cross concept expanded in the direction of the oscillation of the waveform.

Next, we study the meaning of the abundance value. Figure 3.7 shows a sine waveform and a triangle waveform. Since they are both expressions of one and the same amplitude and periodic cycle, they represent the same variation value. But they differ from each other in wave shape; and so the abundance w in the sine waveform differs from the abundance v in the triangle waveform. Thus we gain an understanding of why it is said that abundance is another way of saying "a feature value that captures the differences in waveform configuration."

Furthermore, because the reference lines are drawn in the direction of the oscillation, if there is zero variation value or abundance value on the reference lines, it means that the waveform has not swung far enough to reach any reference line. Thus it is understood that both variation value and abundance value carry oscillation amplitude information.

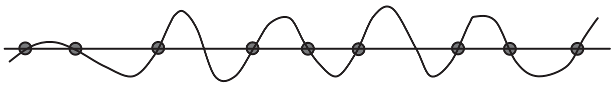


Figure 3.6. Zero cross concept.

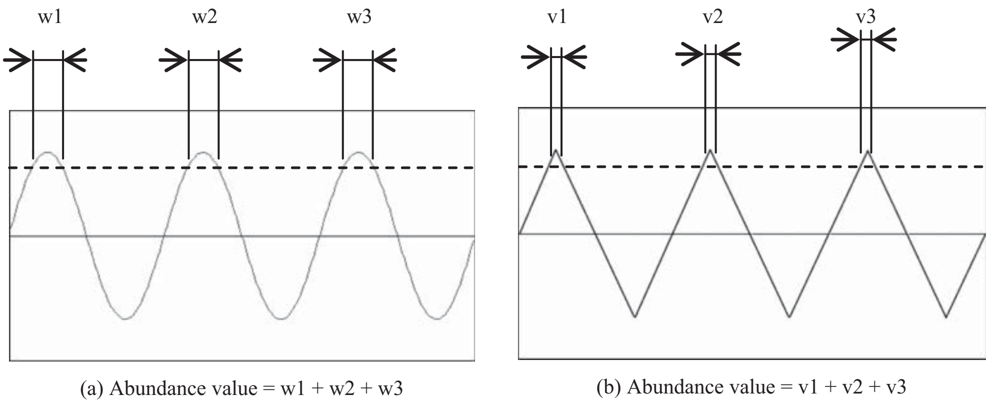


Figure 3.7. Difference in shape between the sine waveform and the triangle waveform.

One further point about the two feature values, variation value and abundance value, is that, in addition to the merit of easy-to-perform quantification of waveforms, they can be defined with respect to any given time width. For instance, if information on ever-changing measured values pertaining to the ongoing operation of a set of equipment is to be extracted, it is possible with the feature values to define domain width at any time interval that may suit the need.

The numbers of variation values and abundance values will increase in proportion to the increasing number of reference lines. But because, with the MT System, many variables can be processed in a short span of time, the compatibility of the two is quite high.

3.5 DIFFERENCES BETWEEN OTHER WAVEFORM FEATURES AND VARIATION VALUES/ABUNDANCE VALUES

In this section, we will review the main points of frequency analysis and wavelets, which have been in common use historically as a means of extracting waveform features, and examine how they differ from methods based on variation values and abundance values.

3.5.1 FREQUENCY ANALYSIS

Frequency analysis is a method that consists of replacing waveform data with the sum of the trigonometric functions of many periodic cycles that express the intensity of the degree of impact (energy) of each periodic cycle. This method can represent the properties of an entire waveform over a relatively extended span of time and is widely employed in such fields as industrial production and economics. An upgraded version of it with enhanced computation processing speed is the “Fast Fourier Transform (FFT).”

A sample display of frequency analysis is shown in Figure 3.8. In this display, the horizontal axis traces the frequency and the vertical axis the energy. The spikes on the graph show that the object waveform packs a high level of energy at the spiked frequencies. As this example indicates, frequency analysis is employed for examining the properties of object waveforms, for evaluating multiple waveforms to determine similarity and for other purposes.

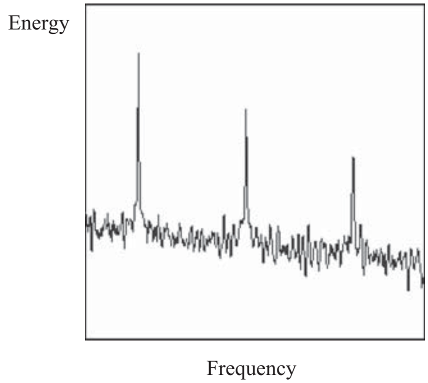


Figure 3.8. Sample waveform display by FFT.

3.5.2 WAVELET

A wavelet can be regarded as a frequency analysis with a temporal variation (change-through-time) element added to it. Thus, the result of analysis by wavelets comes with the frequency axis and time axis built in, outputting an image that has energy expressed in degrees of density of shades or in color, or with the use of isobars.

A sample wavelet display is shown in Figure 3.9. The horizontal axis shows time, and the vertical axis, frequency. Energy is described in terms of isobars. From this image, we can confirm how the waveform frequency component varies with the passage of time.

3.5.3 HOW FREQUENCY ANALYSIS AND WAVELETS DIFFER FROM VARIATION AND ABUNDANCE VALUES

Frequency analysis and wavelets are both commonly used waveform characterization techniques. We enumerate below the important points of difference between these techniques and the variation and abundance components of the MT System.

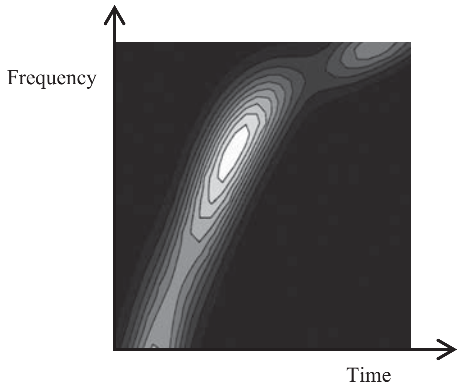


Figure 3.9. Wavelet projection of signal size in two-dimensional isobaric projection.

Frequency analysis and wavelets display properties and the homogeneous qualities of a waveform in the form of a graph or an image, which are expressions that humans see and can easily understand.

Variation and abundance values constitute the sort of information about waveform characteristics that is expressed in numerical values and does not, when visually perused, lend itself to facile comprehension by the human operator as to the properties of the waveform. Since the information is intended for processing directly by a computer, it can be utilized as-is for pattern recognition.

Frequency analysis shows averaged-out properties of a waveform over a sufficiently extended period of time, making it difficult for the method to adequately capture short-duration events (variations). On the other hand, feature extraction by the variation-and-abundance method extracts feature values at a fixed interval, allowing pattern discrepancies, if any, to be detected for a fixed duration of time on a unit-by-unit basis.

CHAPTER 4

MT Method Application Procedure and Important Points to Heed

This chapter explains the procedures for the application of the MT Method using two examples. The first example concerns pattern recognition as applied to numerals. A Unit Space will be created for the numeral "5," and a determination will be made as to whether the target pattern can be considered a member of the "set of 5s." The second example concerns weather prediction. With past meteorological data up to and including today's on hand, we will predict if it will rain the following day. Both are simplified to the extreme, but a basic processing methodology is showcased as it relates to pattern recognition and prediction-related issues.

4.1 EXAMPLE OF CHARACTER RECOGNITION

In this section, we will create a Unit Space following the definition of several patterns readable as the numeral "5." We will then determine whether the object pattern can be considered a member of the set of 5s. And we will look at and conduct cause diagnostics if it proves unreadable as a 5.

4.1.1 DEFINITION OF A UNIT SPACE

We have prepared, for this section, a "group of 16 character patterns readable as 5," shown in Figure 4.1. Some appear rather strange, but each of them can only be read as 5 if the constraint is given that each is a numeral character pattern.

The MT System defines a unique Unit Space for each numeral. If the ten different character patterns for 0 to 9 are to be recognized, ten different Unit Spaces will be defined.

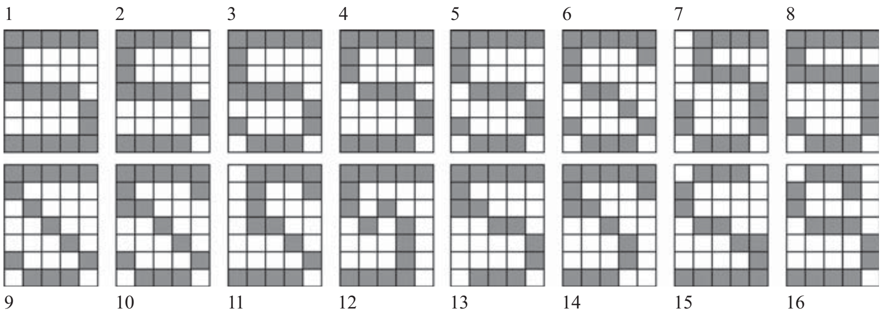


Figure 4.1. Unit space for pattern “5”.

4.1.2 FEATURE EXTRACTION

For feature extraction from a character pattern, we apply the method discussed in Section 3.3. Variation values (differential characteristics) and abundance values (integral characteristics) are extracted from the rows of each character pattern. Figure 4.2 shows the 14 feature values extracted from each of the numeral character patterns. Table 4.1 sorts the feature values extracted from the 16 characters in Figure 4.1.

In Table 4.1, character type data is arranged in the vertical direction, and feature value data in the horizontal direction. Of the feature values, variation values are under Columns 1 to 7, and

Table 4.1. Data that make up the features of 16 numerals readable as 5

	Variation							Abundance						
	Feature value number													
Sample number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
1	1	2	2	2	1	1	1	5	1	1	4	1	1	5
2	2	2	2	2	1	1	2	4	1	1	4	1	1	4
3	1	2	2	2	1	3	2	5	1	1	4	1	2	3
4	1	3	2	2	1	1	2	5	2	1	3	1	1	4
5	1	2	2	2	1	3	2	5	1	1	3	1	2	3
6	1	3	2	2	2	3	2	5	2	1	2	1	2	3
7	1	2	2	1	3	3	2	4	1	3	1	2	2	3
8	1	2	1	1	1	3	2	5	1	5	1	1	2	3
9	1	2	2	2	2	3	2	5	1	1	1	1	2	3
10	1	3	2	2	2	3	2	5	2	2	1	1	2	3

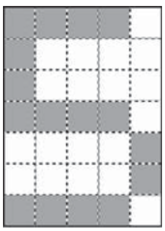
Feature values in Figure 4.2

(*Continued on following page*)

Table 4.1. (*Continued*)

	Variation							Abundance						
	Feature value number													
Sample number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
11	1	3	2	2	2	1	2	4	2	1	2	1	1	4
12	1	2	4	4	2	2	2	5	1	2	2	1	1	4
13	1	2	2	2	1	1	2	5	1	2	2	1	1	3
14	1	3	2	2	2	2	2	5	2	2	1	1	1	3
15	2	2	2	2	1	1	1	3	1	1	2	2	1	5
16	2	4	2	2	1	1	2	3	2	1	3	1	1	4

Feature values in Figure 4.2



Variation	Abundance
2	4
2	1
2	1
2	4
1	1
1	1
2	4

(a) Pattern No. 2

Variation	Abundance
1	5
2	1
2	1
2	4
1	1
3	2
2	3

(b) Pattern No. 3

Figure 4.2. Example of feature extraction.

abundance values under Columns 8 to 14. That means that data expressed in numeric values for the characters in Figure 4.1 are tabulated in 14 columns × 16 rows. This table, then, represents the Unit Data for the character patterns for "5."

Under the MT Method, the following condition is required, so, the number of samples is set at 16.

(number of samples) > (number of variables).

4.1.3 COMPUTATION OF THE MAHALANOBIS DISTANCE OF FEATURE SPACE AND UNIT DATA

Conducting computations of the Unit Data shown in Table 4.1 according to the procedure discussed in Section 2.5, we obtain, with respect to each variable, the average value and standard deviation as well as the normalized value, correlation matrix, and inverse matrix of the correlation matrix, as shown in Table 4.2.

Table 4.2. Computation result

	Feature value No.													
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Average	1.19	2.44	2.06	2.00	1.50	2.00	1.88	4.56	1.38	1.63	2.25	1.13	1.44	3.56
Standard Deviation	0.39	0.61	0.56	0.61	0.61	0.94	0.33	0.70	0.48	1.05	1.09	0.33	0.50	0.70
	Normalized values													
Data No.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
1	−0.48	−0.72	−0.11	0.00	−0.82	−1.07	−2.65	0.62	−0.77	−0.59	1.61	−0.38	−0.88	2.04
2	2.08	−0.72	−0.11	0.00	−0.82	−1.07	0.38	−0.80	−0.77	−0.59	1.61	−0.38	−0.88	0.62
3	−0.48	−0.72	−0.11	0.00	−0.82	1.07	0.38	0.62	−0.77	−0.59	1.61	−0.38	1.13	−0.80
4	−0.48	0.92	−0.11	0.00	−0.82	−1.07	0.38	0.62	1.29	−0.59	0.69	−0.38	−0.88	0.62
5	−0.48	−0.72	−0.11	0.00	−0.82	1.07	0.38	0.62	−0.77	−0.59	0.69	−0.38	1.13	−0.80
6	−0.48	0.92	−0.11	0.00	0.82	1.07	0.38	0.62	1.29	−0.59	−0.23	−0.38	1.13	−0.80
7	−0.48	−0.72	−0.11	−1.63	2.45	1.07	0.38	−0.80	−0.77	1.31	−1.15	2.65	1.13	−0.80
8	−0.48	−0.72	−1.91	−1.63	−0.82	1.07	0.38	0.62	−0.77	3.20	−1.15	−0.38	1.13	−0.80
9	−0.48	−0.72	−0.11	0.00	0.82	1.07	0.38	0.62	−0.77	−0.59	−1.15	−0.38	1.13	−0.80
10	−0.48	0.92	−0.11	0.00	0.82	1.07	0.38	0.62	1.29	0.36	−1.15	−0.38	1.13	−0.80
11	−0.48	0.92	−0.11	0.00	0.82	−1.07	0.38	−0.80	1.29	−0.59	−0.23	−0.38	−0.88	0.62
12	−0.48	−0.72	3.49	3.27	0.82	0.00	0.38	0.62	−0.77	0.36	−0.23	−0.38	−0.88	0.62
13	−0.48	−0.72	−0.11	0.00	−0.82	−1.07	0.38	0.62	−0.77	0.36	−0.23	−0.38	−0.88	−0.80
14	−0.48	0.92	−0.11	0.00	0.82	0.00	0.38	0.62	1.29	0.36	−1.15	−0.38	−0.88	−0.80
15	2.08	−0.72	−0.11	0.00	−0.82	−1.07	−2.65	−2.22	−0.77	−0.59	−0.23	2.65	−0.88	2.04
16	2.08	2.56	−0.11	0.00	−0.82	−1.07	0.38	−2.22	1.29	−0.59	0.69	−0.38	−0.88	0.62

Correlation matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1.00	0.18	−0.05	0.00	−0.39	−0.51	−0.30	−0.84	−0.04	−0.29	0.33	0.30	−0.42	0.53
2	0.18	1.00	−0.08	0.00	0.08	−0.22	0.27	−0.28	0.93	−0.23	−0.07	−0.27	−0.22	0.01
3	−0.05	−0.08	1.00	0.92	0.28	−0.12	0.04	0.07	−0.09	−0.28	0.08	−0.04	−0.33	0.23
4	0.00	0.00	0.92	1.00	0.00	−0.22	0.00	0.14	0.00	−0.39	0.19	−0.31	−0.41	0.29
5	−0.39	0.08	0.28	0.00	1.00	0.44	0.31	0.07	0.21	0.19	−0.66	0.31	0.31	−0.36
6	−0.51	−0.22	−0.12	−0.22	0.44	1.00	0.40	0.47	−0.14	0.38	−0.43	0.00	0.94	−0.76
7	−0.30	0.27	0.04	0.00	0.31	0.40	1.00	0.30	0.29	0.22	−0.26	−0.43	0.33	−0.77
8	−0.84	−0.28	0.07	0.14	0.07	0.47	0.30	1.00	−0.07	0.20	−0.10	−0.57	0.37	−0.51
9	−0.04	0.93	−0.09	0.00	0.21	−0.14	0.29	−0.07	1.00	−0.21	−0.18	−0.29	−0.16	−0.07
10	−0.29	−0.23	−0.28	−0.39	0.19	0.38	0.22	0.20	−0.21	1.00	−0.57	0.13	0.31	−0.39
11	0.33	−0.07	0.08	0.19	−0.66	−0.43	−0.26	−0.10	−0.18	−0.57	1.00	−0.26	−0.32	0.47
12	0.30	−0.27	−0.04	−0.31	0.31	0.00	−0.43	−0.57	−0.29	0.13	−0.26	1.00	0.05	0.23
13	−0.42	−0.22	−0.33	−0.41	0.31	0.94	0.33	0.37	−0.16	0.31	−0.32	0.05	1.00	−0.70
14	0.53	0.01	0.23	0.29	−0.36	−0.76	−0.77	−0.51	−0.07	−0.39	0.47	0.23	−0.70	1.00

Table 4.2. (*Continued*)

	Inverse matrix													
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
1	7.0	−1.7	8.2	−7.9	−0.2	−2.9	−1.4	5.5	2.6	0.0	−1.7	−1.8	3.1	−0.3
2	−1.7	134.9	−246.5	235.9	82.8	−41.8	42.5	75.8	−131.0	17.1	49.1	102.4	14.5	18.5
3	8.2	−246.5	534.8	−517.0	−178.3	88.2	−84.9	−137.6	243.1	−42.6	−112.2	−210.1	−30.9	−28.4
4	−7.9	235.9	−517.0	505.3	173.3	−95.5	78.0	130.8	−231.1	44.7	110.8	203.3	38.9	21.1
5	−0.2	82.8	−178.3	173.3	64.0	−35.4	25.4	47.3	−81.5	16.1	39.6	68.6	14.6	6.5
6	−2.9	−41.8	88.2	−95.5	−35.4	52.9	−6.7	−26.9	38.2	−14.0	−21.7	−36.0	−37.1	7.3
7	−1.4	42.5	−84.9	78.0	25.4	−6.7	22.6	26.5	−43.0	3.1	14.2	36.4	0.0	15.4
8	5.5	75.8	−137.6	130.8	47.3	−26.9	26.5	52.5	−72.6	9.2	26.6	60.8	11.2	13.8
9	2.6	−131.0	243.1	−231.1	−81.5	38.2	−43.0	−72.6	129.9	−15.4	−47.1	−99.0	−11.5	−19.7
10	0.0	17.1	−42.6	44.7	16.1	−14.0	3.1	9.2	−15.4	7.9	12.2	16.1	8.4	−2.6
11	−1.7	49.1	−112.2	110.8	39.6	−21.7	14.2	26.6	−47.1	12.2	28.5	43.7	8.6	0.5
12	−1.8	102.4	−210.1	203.3	68.6	−36.0	36.4	60.8	−99.0	16.1	43.7	88.6	13.1	13.5
13	3.1	14.5	−30.9	38.9	14.6	−37.1	0.0	11.2	−11.5	8.4	8.6	13.1	31.5	−5.6
14	−0.3	18.5	−28.4	21.1	6.5	7.3	15.4	13.8	−19.7	−2.6	0.5	13.5	−5.6	18.5

The total of combinations of the 14 variables will be ${}_{14}C_2 = 91$, which means that 91 correlative coefficients will be found. If we are to display the correlative coefficients in ellipses, we will have 91 ellipses drawn, as shown in Figure 4.3. In the MT System, it is understood that the computations can be performed based on these 91 ellipses.

We will next compute the Mahalanobis Distance (MD) of the 16 items of Unit Data themselves. For instance, the feature value of the third character is as shown in Table 4.3. The table shows the normalized values as well. These 14 pieces of data are computed using the following equation. Specifically, if we express the inverse matrices of the correlation matrices in Table 4.2 as **A** and the normalized vectors for the 14 items of data in Table 4.3 as **Y**, then the MD works out as follows:

$$D^2 = \mathbf{Y}\,\mathbf{A}\,\mathbf{Y}^T/k$$
$$= 0.88$$

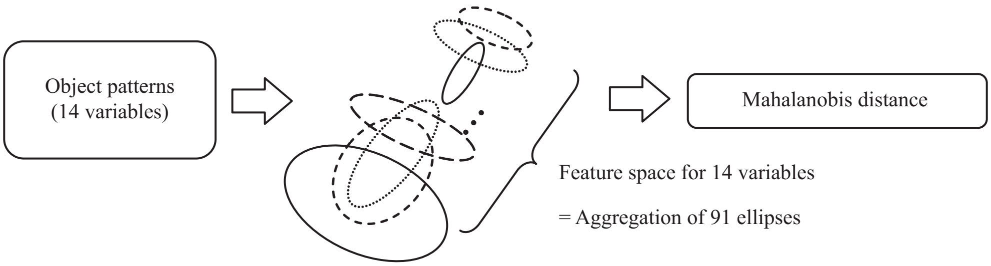


Figure 4.3. The Mahalanobis distance is computed from the aggregation of ellipses.

Table 4.3. Feature values for third character

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Feature value	1	2	2	2	1	3	2	5	1	1	4	1	2	3
Normalized value	−0.5	−0.7	−0.1	0.0	−0.8	1.1	0.4	0.6	−0.8	−0.6	1.6	−0.4	1.1	−0.8

MD computations performed on other Unit Data yield the results that are shown in Table 4.4; the histogram figures are shown in Figure 4.4.

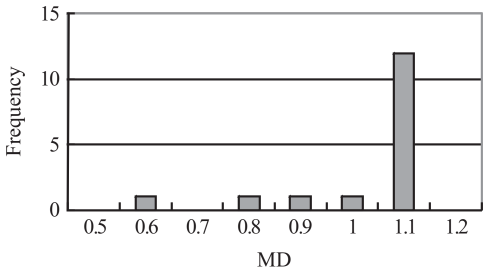


Figure 4.4. Histogram of MD.

Table 4.4. Mahalanobis distance of unit data

Data No.	MD(D^2)
1	1.07
2	1.07
3	0.88
4	1.07
5	0.97
6	0.59
7	1.07
8	1.06
9	1.06
10	0.75
11	1.07
12	1.07
13	1.07
14	1.07
15	1.07
16	1.07

From Table 4.4, we understand that, overall, the MDs are expressed in small values, averaging 1.0. This is how it always works out regardless of the scale of the Unit Data. This is one important feature of the MT Method, and its implications are as follows:

1. **Unit quantities** expressed in terms of 1 cm or 1 g (gram), which are units used in our everyday lives, can be defined as Unit Spaces as well.
2. When the MD of target data is computed and the solution turns out to be 1 (one) or thereabouts, the pattern of the unknown (target) data can be said to be close to the pattern of the Unit Space.

When the MT Method was first proposed, the Unit Space was referred to as "**reference space**". This renaming occurred in consideration of the feature summarized above in (1). Namely, the new name conveys the notion that it is a space that subsumes "unit" volume. Figure 1.13 in Chapter 1 is intended as an illustration of that notion.

4.1.4 COMPUTATION OF THE MAHALANOBIS DISTANCE OF UNKNOWN (TARGET) DATA

Let us suppose now that the four patterns shown in Figure 4.5 are our determination unknown data (target data). These patterns do not exist among the 16 patterns we looked at in Table 4.1. In Table 4.5, we see the results of the feature extraction to which the target data was submitted. It is necessary that the number of characteristics (number of variables) for the target data be the same as for the Unit Data. That means that in this particular case, the number of variables is 14.

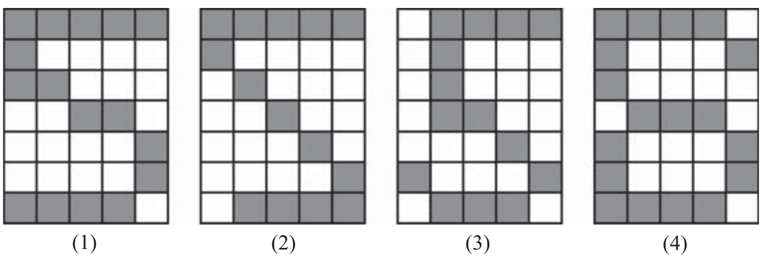


Figure 4.5. Patterns of target data.

Table 4.5. Feature values of target data

	Feature value No.													
Data No.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
(1)	1	2	2	2	1	1	2	5	1	2	2	1	1	4
(2)	1	2	2	2	2	1	1	5	1	1	1	1	1	4
(3)	1	2	2	2	2	3	2	4	1	1	2	1	2	3
(4)	2	3	2	2	3	3	2	4	2	1	3	2	2	4

Attention must now be directed to the following point: Of the four patterns, one of them, (4), is not readable as a "5," but, in terms of feature values, can we identify in it some distinguishing numerical data not shared by the other patterns? The numerical data for this pattern is given under sample (4) in Table 4.5. Its feature values in Column 1 and the Column 12 are given as "2," differing from those values in samples (1)–(3). These are readily noticeable differences. But patterns with "2" for their value in the first column exist also in the Unit Data in Table 4.1. The same can be said of its value in the twelfth column. Namely, none of the individual feature values themselves comes with any particularly singular value.

For pattern (4), which seems hardly readable as "5," let's see what the MD computation results look like. Will (1)–(3) be recognized as members of the group of 5s?

Based on the MD computation formula, the MDs for these four target data items work out as displayed in Table 4.6.

Table 4.6. Mahalanobis distance of target data

Data No.	**MD(D^2)**
(1)	1.82
(2)	4.06
(3)	3.21
(4)	110.18

The MDs for patterns (1)–(3) are relatively modest, but pattern (4) stands out with a high figure: MD = 110.2. With the MT Method, the threshold for determining the admissibility of a given variable to membership in the Unit Data is, as generally accepted, set at 4. Accordingly, pattern (4) cannot be regarded as a 5. As for pattern (2), which has yielded MD = 4.1, it is a borderline case which defies a clear-cut determination of it as a 5.

The following summarizes the above discussion:

Pattern (1) → Can be called a member of the 5 group.
Pattern (2) → Is a borderline case that may or may not be part of the 5 group.
Pattern (3) → Can be called a member of the 5 group.
Pattern (4) → Cannot be called a member of the 5 group.

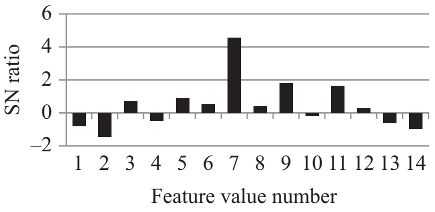
Caution should be taken as follows: The MDs acquired here do not represent distances to (or from) some ideal pattern that exists somewhere, but are no more or no less than distances to (or from) the centers of the 16 patterns listed in Figure 4.1. Mathematically speaking, they are distances from an average of the Unit Data shown in Table 4.1. It is thus important to define Unit Data impartially. As with people, if knowledge harbors a bias, the result of any judgment made by such knowledge will prove lacking in impartiality.

4.1.5 CAUSE DIAGNOSTICS

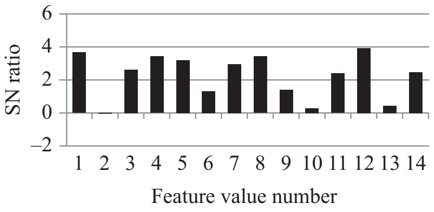
Of the target data we looked at in Table 4.6, items (2) and (4) show substantial MD values, and so cause diagnostics to explain these occurrences are conducted. In this case, there are 14 variables, and so we use an orthogonal array L_{16}. The results of the computations are shown in Table 4.7, as well as in graphs in Figure 4.6. The numerical values in the table show the Signal to Noise (SN) ratio differences as computed according to the rules governing the usage of the orthogonal array, namely "Level 1 – The variables are used" and "Level 2 – The variables are not used."

Table 4.7. Results of cause diagnostics (SN ratio differences between level 1 and level 2)

	Feature value No.													
Data No.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
(2)	−0.67	−1.34	0.65	−0.41	0.87	0.42	4.47	0.33	1.73	−0.07	1.55	0.18	−0.53	−0.84
(4)	3.29	−0.09	2.28	3.04	2.87	1.09	2.63	3.08	1.14	0.09	2.08	3.52	0.23	2.12



(a) Pattern (2)



(b) Pattern (4)

Figure 4.6. Cause diagnostics results (graph display of Table 4.8).

A look at pattern (2) in Figure 4.6 makes it clear that variable 7 shows the greatest difference, followed by variables 9 and 11. Based on these results, it is estimated that the greatest contribution to making the MD as great as it appears comes from variable 7 singly or from variables 9 and 11 combined. We now make a comparison of the variables in pattern (2) with the variables of the group of Unit Data.

Table 4.8. Feature values for pattern (2)

Data No.	Feature value No. 1	2	3	4	5	6	7	8	9	10	11	12	13	14
(2)	1	2	2	2	2	1	1	5	1	1	1	1	1	4

The variables of pattern (2) are extracted and displayed in Table 4.8.

In Table 4.8, the values of variables (feature values) 7, 9, and 11 are circled. We see, with respect to variable 7, that the value is 1. Note in this connection that, in the Unit Data in Table 4.1, there exist two items among the 16 samples that display "1" for this value. This means that, although there are few, such cases with the same value do exist in the Unit Data, indicating that it can therefore be determined that this variable alone would not single-handedly be responsible for the abnormality.

We now turn to variables 9 and 11. The values of these variables are also "1," and a comparison with the Unit Data shows that they are not singly responsible for any abnormality. In the Unit Data in Table 4.1 there is no data with these three variables with the value "1." Therefore, it is estimated that these three variables combined to cause the distance from pattern (2) to grow somewhat large.

As for pattern (4), as we see in Figure 4.6 (b), most of the variables contribute to the enhancement of the MD, but a look at the individual variables themselves makes it clear that there is nothing abnormal about their values per se. It can therefore be judged that it is the balance of (the correlation between) the patterns as a whole that is out of order.

The example we have studied here concerns character recognition, and so the cause diagnostics have run their course at this point. In the case of equipment monitoring or other cases, cause diagnostics will shed light on which measured value, or values, might have been involved in the generation of a given abnormality. This will open the way for further pursuit of the fundamental cause of abnormalities precisely where such measured values are involved.

4.1.6 SETTING A THRESHOLD

Whatever the pattern recognition method that is employed, normal/abnormal decisions are made based on a comparison of numerical values acquired through computation against a *threshold*. In Item 4.1.4, we gave a general outline of the conditions for determining whether an object pattern is readable as a "5." At this time, we will discuss a few additional topics such as the concept of the setting of a threshold, the setting of a "gray zone" and other key matters of practical interest.

4.1.6.1 "4" as a Working Criterion for the Threshold

The judgment threshold with the MT Method is generally said to boil down to 4 or thereabouts. This is because, for all practical statistical mathematics purposes, if the MD exceeds 4, the probability of unknown (target) data being a member of the Unit Space shrinks to a small possibility. The reasonableness of this figure (the value of 4) is clear from the result of the character pattern recognition we discuss in this chapter. At the same time, after having gone through the sequence of procedures with the example used, we will gain an understanding that, while statistical mathematics ought to be duly respected, a threshold cannot be decided solely based on mathematical theory.

What, then, are the nonmathematical items of consideration to bear in mind in deciding on a threshold? Here are two points to consider:

HOW TO DEFINE A UNIT SPACE

As a factor with an impact on the threshold, we must first consider *what patterns decide the Unit Space.* With the example used in this chapter, 16 patterns readable as "5" were defined. They were not necessarily *neatly* formed patterns, and some were more than somewhat malformed. It would have been possible, nonetheless, to add to the Unit Space even more seriously malformed patterns that were nonetheless readable as a "5." That is all to say that the mathematical setting is subject to change as a function of how the Unit Space is defined. As a consequence, even if the very same object of judgment was in question, the MD could measure differently.

The concept involved here is illustrated in the schematic drawing in Figure 4.7. It shows that, depending on the definition of the Unit Space, distance measurement scale changes so that the same object of judgment will yield differing distances. In other words, according to the scale for Unit Space "a," the distance comes out as 3.3, but for Unit Space "b," 4.5 will show. Thus, the same target data is seen yielding differing distances depending on the Unit Space.

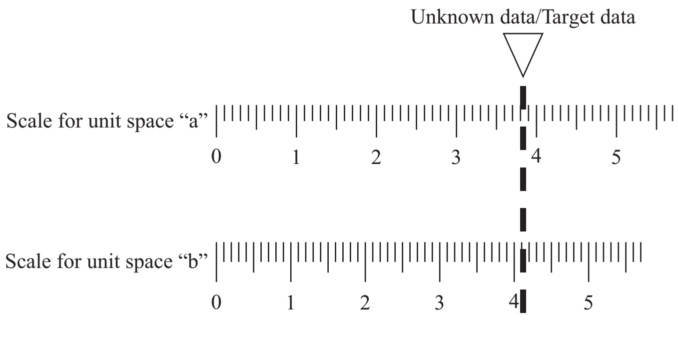


Figure 4.7. Concept of measurements yielding differing MDs depending upon unit space definition.

DIFFERENT VARIETIES AND SCOPES OF VARIABLES

The next issue concerns the type of variable. For example, if the variation and abundance values extracted from a numeral pattern are composed of a set of integers within the range of 0 to 5, it means that the values are available only within certain bounds. The same principle will hold when category data is adopted for the variables. In cases like this, the threshold value cannot be said to reflect the probability of attribution to the Unit Space accurately.

4.1.6.2 Reasonable Method of Setting a Threshold

As already discussed, 4 is the expedient criterion set as the threshold. But in reality there arises a need to set an appropriate threshold only after finding the MDs of several target data. And, in principle, the threshold will be set in such a way that *the sum of losses incurred by mistaking normal for abnormal (error of the first kind)* and *the sum of losses incurred by mistaking abnormal for normal (error of the second kind)* will be at its smallest. For instance, with a postal code reader system, the threshold will be set so that it will produce the smallest loss after the sum of the losses incurred by failure to recognize the code and the sum of losses incurred by misreading of the code have been added.

In the evaluation of any technique, it is necessary to have a numerical means of expression; in quality engineering, the SN ratio is utilized (see Notes 2.3 and 2.4). The SN ratio provides a suitable means of judgment in the treatment of the sum of losses, as seen in the example above, because the SN ratio is directly related to the treatment of losses.

Note 4.1 What is an ideal, average character pattern?

When assigning various patterns to the Unit Data, one will develop a curiosity as to what sort of pattern will be considered ideal. But, in the case of characters, “readability as such-and-such a character” is the condition for belonging to the Unit Space, so it is not that ideal patterns are designated and exist as such.

It is possible, nonetheless, to define an *average pattern*. Average values of Unit Data are as displayed in Table 4.2. Note, however, that the values that can be assumed by variables are, in the case discussed, only integers. Because the average values come with values below the decimal point in this case, there exists no expressible character pattern. If the average values of the variables are rounded off to the nearest numbers to obtain integers, the pattern appears as shown in Figure 4.8.

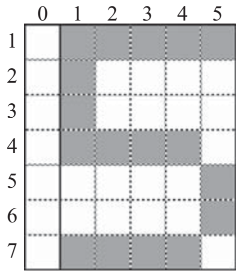
Variation	Abundance
1	5
2	1
2	2
2	2
2	1
2	1
2	4

Figure 4.8. The pattern of an average 5.

Note 4.2 Variation and abundance values in the vertical direction

In the example used in this section, each pattern is expressed in 14 sets of variables to obtain the variation and abundance data. But the resulting information does not suffice for the purposes of character recognition for 0–9. Some combinations of variables come out the same for “2” and “5.” More specifically, 5 and 2 are of broadly similar shapes in line symmetry, and difficulty of discrimination could occur if based solely on variation and abundance information in the horizontal direction.

To resolve this problem, it is necessary, as shown in Figure 4.9, to make use of variation and abundance data in the vertical direction as well. If the variables were to be characterized in the vertical direction as well, we would have 5 columns × 2 varieties = 10 variables when added together, resulting in a total of 24 variables. Note, however, that the condition, “number of samples > number of variables” must be satisfied so that 25 or more samples become necessary.



Variation	Abundance
1	5
2	1
2	1
2	4
1	1
1	1
2	4

Variation	3	5	5	5	4
Abundance	5	3	3	3	3

Figure 4.9. Example of additional characterization in the vertical direction.

4.1.6.3 Setting a Gray Zone

There are, in actual pattern recognition work, many instances of the practical inconvenience of having only one threshold to work with. Instead of classifying all objects that exceed the threshold as abnormal, then, as shown in Figure 4.10, a method may be used that sets a secondary threshold defined at a higher value so that cases that fall short of it can be judged as falling into the gray zone. Objects that have been determined to be gray zone cases will be subjected to other measures—human operators, for example—for a higher-level judgment. In particular, at the initial stage of operating a pattern recognition system, it is common practice to accord gray zone limits with ample latitude and narrow the limits down as the body of data accumulates.

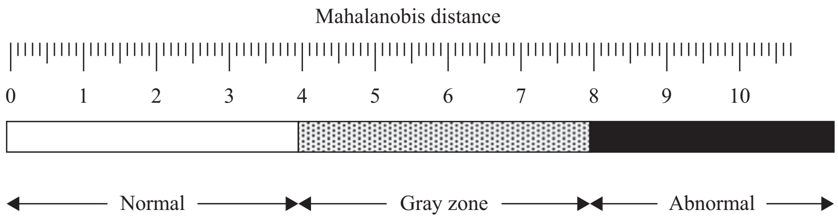


Figure 4.10. Setting a gray zone.

4.1.7 UNIT SPACE RENEWAL

If the Mahalanobis Distance exceeds the threshold in a given case but the target/unknown data is judged to be normal using an alternative method (for example in the judgment of an experienced operator), then the data can be added to the Unit Space and updated. Especially at the initial stage of operation of a pattern recognition system, the threshold is in many cases set on the tighter (lower) side to avoid overlooking abnormalities. For this reason, the Unit Space should be updated when data has accumulated.

When working a pattern recognition system, the operator will be helping improve the accuracy of judgment by updating the Unit Space in step with the increasing body of data (Table 4.9). Updating the Unit Space is an idea common to all pattern recognition systems, but with the MT Method as well as the T Method.

Table 4.9. Unit space renewal

	Sample number	Item/variable 1	2	3	…	*k*
	1	x_{11}	x_{12}	x_{13}	…	x_{1k}
	2	x_{21}	x_{22}	x_{23}	…	x_{2k}
	3	x_{31}	x_{32}	x_{23}	…	x_{3k}
	…	…	…	…	…	…
	n	x_{n1}	x_{n2}	x_{n3}	…	x_{nk}
Added Unit Data	n+1	$x_{n+1,1}$	$x_{n+1,2}$	$x_{n+1,3}$	…	$x_{n+1,k}$
	n+2	$x_{n+2,1}$	$x_{n+2,2}$	$x_{n+2,3}$	…	$x_{n+2,k}$
	…	…	…	…	…	…
	n+m	$x_{n+m,1}$	$x_{n+m,2}$	$x_{n+m,3}$	…	$x_{n+m,k}$

Note 4.3 Direct utilization of pixel data and recognition accuracy

In Section 3.8, we stated that, because image data amounts to a great numbers of pixels, mosaicing, or characterization, is performed; the character patterns chosen as examples in this chapter are represented by five horizontal grids × seven vertical grids of information, which is carried by a modest number of pixels. Accordingly, the pixel information itself can be utilized directly. In other words, as Figure 4.11 (a) illustrates, a white grid will be defined as 0 (zero) and a black grid as 1 so that a set of 35 one-or-zero variables are obtained as (b). Examples of the application of this approach are shown in Section 5.2 for reference.

1	1	1	1	1
1	0	0	0	0
1	0	0	0	0
1	1	1	1	0
0	0	0	0	1
0	0	0	0	1
1	1	1	1	1

(a)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1	1	1	1	1	0	0	0	0	1	0	0	0	0	1	1	1

19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
1	0	0	0	0	0	1	0	0	0	0	1	1	1	1	1	1

(b)

Figure 4.11. Direct conversion of image information into variables.

4.2 EXAMPLE OF WEATHER PREDICTION

In this section, we look at an example of "Tomorrow's Weather Prediction" and study the procedure for MT Method application. In the example of character recognition discussed in the previous section, the variables (items) were feature values extracted from the characters. But in the example we take up here, the variables are measured values such as atmospheric pressure, temperature, etc. as well as category data such as direction of the wind (the data was sourced from meteorological statistics released by the Japanese Meteorological Agency).

The example is very much simplified, but it will provide a reference case in the area of product or plant accident/incident prediction.

4.2.1 CONCEPTS UNDERLYING WEATHER PREDICTION

Actual weather forecasting is based on an enormous body of observed data coupled with atmospheric movement theory and other studies and disciplines, but the example we describe here, as shown below, is prediction in simplified form. A conceptual scheme is shown in Figure 4.12.

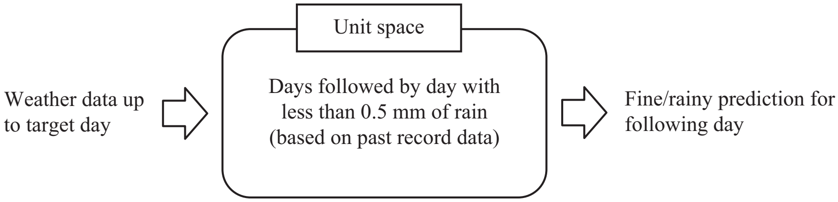


Figure 4.12. Concept underlying weather predication.

Days followed by a rainless day (a day with no rainfall or less than 0.5 mm of rainfall) are picked out of the meteorological database for a given city for a given year, and a Unit Space is formed from them. Since this city is characterized between April and October by rain- and snow-free weather, the data for that period will be used.

A "weather forecast for the following day," namely, a rainy/fine prediction, is made for the same period of the following year. Since the result is already known, it is possible to evaluate the degree of prediction accuracy.

Part of the data is shown in Table 4.10. Here we are to predict the weather for May 5 on the basis of the data for May 4 and earlier.

Table 4.10. Part of data used for weather predication

Date	Atmospheric pressure (hPa)	Rainfall (mm)	Temperature (°C)	Humidity (%)	Wind velocity (m/s)	Wind direction
5/1	995.5	4	8.7	78	4.3	NNW
5/2	1009.2	0	5.5	60	6.1	NNW
5/3	1014.4	0	7.4	60	4.0	NNW
5/4	1014.2	0	10.9	53	2.5	SE
5/5	1007.4	0	11.7	58	3.8	SSE
5/6	1012.3	0	13.3	49	2.7	NNW
5/7	1016.3	0	13.8	31	2.2	SSE
5/8	1019.0	0	11.7	44	2.7	NNW
5/9	1018.0	0	10.9	52	3.2	SE
5/10	1010.4	9.5	10.9	73	7.7	SSE
5/11	1004.5	0.5	12.5	56	6.0	SSE
5/12	1014.9	0	13.9	31	4.3	WNW

4.2.2 METEOROLOGICAL DATA AND PREPROCESSING

4.2.2.1 Differential Computation Processing for Temperature, etc.

As shown in Figure 4.13, atmospheric temperature values are higher in the summer. While it is presumed that temperature plays a role in weather prediction, information on variation must surely be important as well. So, we decide to add to the variable any **difference** between the temperature for the day under consideration and temperatures for the previous day and the day before that. Figure 4.14 is a graph showing the difference from the previous day. Differences were likewise acquired for the atmospheric pressure and other measured items.

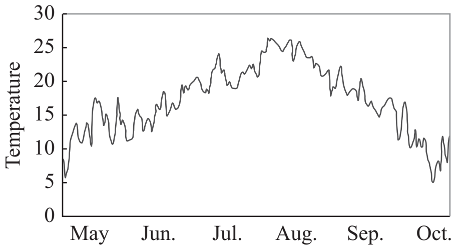


Figure 4.13. Temperature variation.

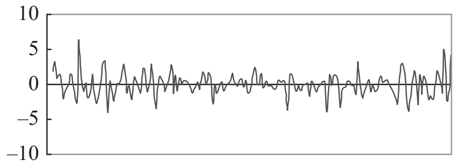


Figure 4.14. Temperature differential graph.

4.2.2.2 Binary Expression of Wind Direction

Direction of the wind indicates the "northerly, southerly, easterly, westerly" type of information, which is category data with no bearing on vertical relation or sequential order. There arises therefore a need to express this information in computer-readable one-or-zero binary data. There are several ways of expression in binary data, but for these purposes we will use three sets of one-or-zero data, as defined in Figure 4.15. There are 16 directions if we include directions such as NNW, but, once again, for our purposes, we will condense it and use only eight directions.

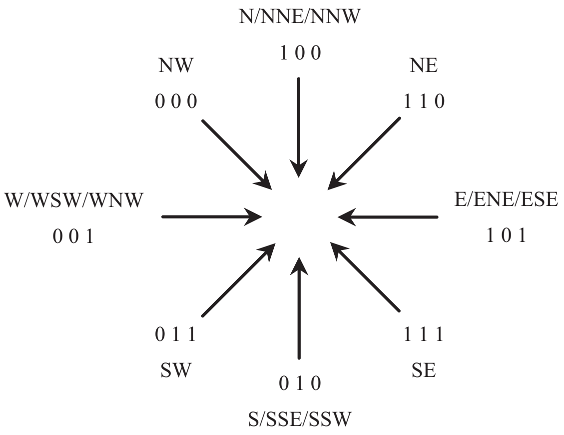


Figure 4.15. Binary expression of wind direction.

4.2.2.3 Example of Meteorological Data

Table 4.11 is a tabulation of "variables for the prediction of the weather for the following day." The measured data for atmospheric pressure and other items is broken into six types, but the number of variables comes to 24.

4.2.3 DEFINITION OF UNIT SPACE

From the body of data for May–October, "days followed by "a day with less than 0.5 mm of rainfall" were picked and used to populate Unit Space. In Table 4.11, the entries for May 3–8 and May 10–11 match the specifications. May 9, the day on which the rainfall was 0 mm, but followed by a day with 9.5 mm of rain, is not included in the Unit Space. On the other hand, if the rainfall was heavy for the target day but less than 0.5 mm for the following day, such a day will count as a member of the Unit Space. Data collected from Table 4.11 that meet such a criterion will form the Unit Space.

4.2.4 RAINFALL PREDICTION BASED ON TARGET DATA

Mahalanobis Distance (MD) measurements were taken with the weather data for May to October of the year following the year for which the Unit Space was defined and set as the object of

Table 4.11. Weather data tabulated for analysis purposes

	Atmospheric Pressure (hPa)w			Rainfall (mm)			Temp. (°C)			Humidity (%)			Wind velocity (m/sec)			Direction of wind (Pattern)								
Day	Day focused on (DF)	Difference from day before (DFB)	Difference from two days ago (DFT)	DF	DFB	DFT	DF	DFB	DFT	DF	DFB	DFT	DF	DFB	DFT	DF			DFB			DFT		
5/03	1,014	5.2	18.9	0	0	−4	7.4	1.9	−1.3	60	0	−18	7.6	−4.2	−0.9	1	0	0	1	0	0	1	0	0
5/04	1,014	−0.2	5	0	0	0	10.9	3.5	5.4	53	−7	−7	5.5	−2.1	−6.3	1	1	1	1	0	0	1	0	0
5/05	1,007	−6.8	−7	0	0	0	11.7	0.8	4.3	58	5	−2	9.2	3.7	1.6	0	1	0	1	1	1	1	0	0
5/06	1,012	4.9	−1.9	0	0	0	13.3	1.6	2.4	49	−9	−4	7.8	−1.4	2.3	1	0	0	0	1	0	1	1	1
5/07	1,016	4	8.9	0	0	0	13.8	0.5	2.1	31	−18	−27	5.6	−2.2	−3.6	0	1	0	1	0	0	0	1	0
5/08	1,019	2.7	6.7	0	0	0	11.7	−2.1	−1.6	44	13	−5	8.6	3	0.8	1	0	0	0	1	0	1	0	0
5/09	1,018	−1	1.7	0	0	0	10.9	−0.8	−2.9	52	8	21	7.5	−1.1	1.9	1	1	1	1	0	0	0	1	0
5/10	1,010	−7.6	−8.6	9.5	9.5	9.5	10.9	0	−0.8	73	21	29	13.9	6.4	5.3	0	1	0	1	1	1	1	0	0
5/11	1,005	−5.9	−13.5	0.5	−9	0.5	12.5	1.6	1.6	56	−17	4	9.9	−4	2.4	0	1	0	0	1	0	1	1	1
5/12	1,015	10.4	4.5	0	−0.5	−9.5	13.9	1.4	3	31	−25	−42	10.4	0.5	−3.5	0	0	1	0	1	0	0	1	0

observation. One part of the results acquired is shown in Table 4.12. The column under the "No." heading is a sequentially arranged list of events with No. 1 representing the starting date of May 1. If the rainfall amount for the day following a day with a small MD figure is small, and, conversely, if the rainfall amount for the day following a day with a large MD figure is large, then that means that the prediction was right on the mark. The table also lists rainfall amounts for following days where we see that such a trend with regard to the relation between the MD and the rainfall amount holds true for the days before and after No. 129.

Table 4.12. MD and amount of rainfall for the following day

Data No.	MD(D^2)	Amount of rainfall for the following day (mm)
43	1.55	0
44	5.95	0
45	3.85	26.5
…	…	…
125	2.25	0
126	1.47	0
127	1.68	2
128	27.32	4
129	23.35	60
130	15.13	2.5
131	4.90	0
132	3.12	15
133	3.35	0
…	…	…
169	2.25	3
170	2.02	0
171	3.74	0.5
172	5.64	9.5
173	4.17	4
174	4.95	3.5

Table 4.13. MD-Based prediction and actual weather

		MD-based prediction	
		No rainfall	**Actual rain**
Actual weather as it occurred	No rainfall	63	42
	Actual rain	28	47

Table 4.13 shows the contrast between prediction and actual rainfall. While the match rate is not very high, we can say that it is appreciably higher than if the prediction had relied on baseless forecasting. Accuracy in prediction will improve if use is made of weather data from other points of observation, and if a more appropriate measurement sampling interval can be identified, or other such improvements are realized.

CHAPTER 5

T METHOD APPLICATION PROCEDURES AND KEY POINTS

In this Chapter, application procedures for T Method (1) and RT Method (=T Method (3)) are explained.

T Method (1) is discussed using "yield prediction in the context of manufacturing production processes" as an illustrative example. It concerns the prediction of a production yield on the basis of process data acquired from day-to-day manufacturing routines such as temperature, pressure, and other readings.

RT Method is discussed using a numeral pattern as an illustrative example. Unlike the case given in Chapter 4, density information (1,0) of a pixel for use *in raw form* is utilized. Consequently, a Unit Space is created for the numeral "7" as part of the procedure for determining whether or not the object pattern can be regarded as belonging in the "5" group.

5.1 YIELD PREDICTION FOR MANUFACTURING-PRODUCTION USING T METHOD-1

In most cases in an ongoing manufacturing process, it is difficult to interrupt and insert tests into intervals of production runs. If daily manufacturing data or data pertaining to manufacturing under altered conditions is available, use of T Method-1 may make it possible to identify the conditions for improvement.

Process data for the case given here is shown in Table 5.1. In this table, the yields, which are the output, are rearranged in ascending order of the value of output. There are six items: B temperature (°C), C temperature (°C), Pressure 1 ($10 \times$ MPa), Pressure 2 ($10 \times$ MPa), preheating time (hours), and manufacturing time (minutes). These will all turn out as measured values. Since the yield, which comes out as an output value, is data that can assume any value between 0.0 (%) and 100.0 (%), that is to say, any percentage value, the data should be analyzed upon its omega conversion (logit transformation). However, given that the yield has proven to be in this case less than 90%, the output value will be analyzed in raw data form. (For a detailed discussion on omega transformation, see Note 5.1.)

Table 5.1. Process data

Data No.	"B" temp	"C" temp	P 1	P 2	Pre-ht time	Manuf time	Yield
1	575.0	274.0	205.5	210.5	5.0	180.0	81.55%
2	575.0	304.0	207.5	218.5	6.0	120.0	82.99%
3	570.0	279.0	199.5	207.5	7.0	120.0	83.03%
4	577.0	205.0	173.0	164.0	7.0	120.0	84.56%
5	573.0	254.0	160.0	164.0	7.0	120.0	84.60%
6	577.0	216.0	176.0	168.0	6.0	120.0	85.52%
7	582.0	222.5	171.0	167.0	5.0	180.0	89.47%

5.1.1 DEFINITION OF THE UNIT SPACE

Figure 5.1 shows the distribution of outputs, that is, yields, displayed in Table 5.1. The figure allows us to judge that a concentration of high yields occurs within the 84.0–85.0% yield range. The total number of data entries is small (seven), and even at the spot of highest data concentration the yield is not 100%. For this reason, it is appropriate for the purposes of yield prediction to use T Method-1.

Given the yield distribution, 4 and 5, two samples found in the 84.0–85.0% yield range, are selected from Table 5.1 as samples of the Unit Space. This data is, again, shown in Table 5.2.

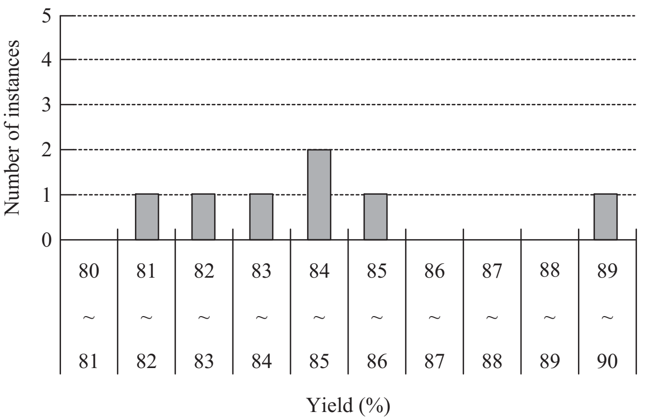


Figure 5.1. Distribution of outputs.

Table 5.2. Data in the unit space and the average values

Sample No.	"B" temp	"C" temp	P 1	P 2	Pre-ht time	Manuf time	Yield
4	577.0	205.0	173.0	164.0	7.0	120.0	84.56%
5	573.0	254.0	160.0	164.0	7.0	120.0	84.60%
Average	575.0	229.5	166.5	164.0	7.0	120.0	84.58%

We will next find from the two samples of the Unit Space the average value $\underline{x}_j$ ($j = 1, 2, \ldots, 6$) of each item and the average value M_0 of the output value (yield). The average values are as shown in the bottom row of Table 5.2, in T Method-1; it is these average values that form the center of the Unit Space.

5.1.2 DEFINITION OF SIGNAL DATA

Of the process data tabulated in Table 5.1, the five data entries, Nos. 1–3 and Nos. 6 and 7, that were not selected for membership in the Unit Space as samples, are treated as Signal Data, as shown in Table 5.3.

Table 5.3. Signal data

Data No.	"B" temp	"C" temp	P 1	P 2	Pre-ht time	Manuf time	Yield
1	575.0	274.0	205.5	210.5	5.0	180.0	81.55%
2	575.0	304.0	207.5	218.5	6.0	120.0	82.99%
3	570.0	279.0	199.5	207.5	7.0	120.0	83.03%
6	577.0	216.0	176.0	168.0	6.0	120.0	85.52%
7	582.0	222.5	171.0	167.0	5.0	180.0	89.47%

5.1.3 NORMALIZATION OF SIGNAL DATA

The average values obtained from the Unit Space samples are used for Signal Data normalization. Normalization is performed by subtracting the averages of the items and yield of the Unit Space from the yield value from each item in the Signal Data. Below is a demonstration of how "B" temperature X_{11} and output M_1 for sample No. 1 are computed:

$$X_{11} = 575.0 - 575.0 = 0.0 \tag{5.1}$$

$$M_1 = 81.55 - 84.58 = -3.03 \tag{5.2}$$

All the pieces of normalized Signal Data are tabulated in Table 5.4. In a statistics-based approach normalization is generally performed by dividing what remains after subtracting the average values from the standard deviation of the item, but T Method-1 dispenses with this computation. This is because in many cases items that have zero standard deviation in the Unit Space, which would make the computation impossible, are of key importance for prediction and estimation purposes.

Table 5.4. Normalized signal data

Data No.	"B" temp	"C" temp	P 1	P 2	Pre-ht time	Manuf time	Yield
1	0.0	44.5	39.0	46.5	−2.0	60.0	−3.03%
2	0.0	74.5	41.0	54.5	−1.0	0.0	−1.59%
3	−5.0	49.5	33.0	43.5	0.0	0.0	−1.55%
6	2.0	−13.5	9.5	4.0	−1.0	0.0	0.94%
7	7.0	−7.0	4.5	3.0	−2.0	60.0	4.89%

5.1.4 COMPUTATION OF PROPORTIONAL COEFFICIENT AND SN RATIO OF EACH ITEM OF SIGNAL DATA

To find out which of the items will be useful for prediction and estimation, an item-by-item computation of the proportional coefficient β and SN ratio η is performed. Figure 5.2 is a graph illustration of the relationships between temperature or pressure, etc., on the one hand, and the output value (yield). The horizontal axis represents the output value. For both the horizontal and the vertical axis, the zero point constitutes the center of the Unit Space.

In T Method-1, SN ratios η and proportional coefficients β are calculated from the relationship between the output value and the item value. The larger the SN ratio η here, the closer to a straight line is the relationship between the output value (yield) and the other items. In other words, such items will act as more significant contributors when a general estimation of the yield is to be made. Furthermore, proportional coefficient β will provide a measure of the steepness of the incline of the straight line for each of the items in Figure 5.2.

The following matters can be explained by the six graphs in Figure 5.2.

(a) Yield and "B" Temperature

Given the right upward incline of the line, the proportional coefficient β is positive. Also, the graph shows that the data is arranged in relatively neat formation along the regression line that passes the zero point and outbound scatter is limited. That is to say, the SN ratio η is large, leading to the assumption that "B" temperature is well-suited to the purpose of general yield estimation.

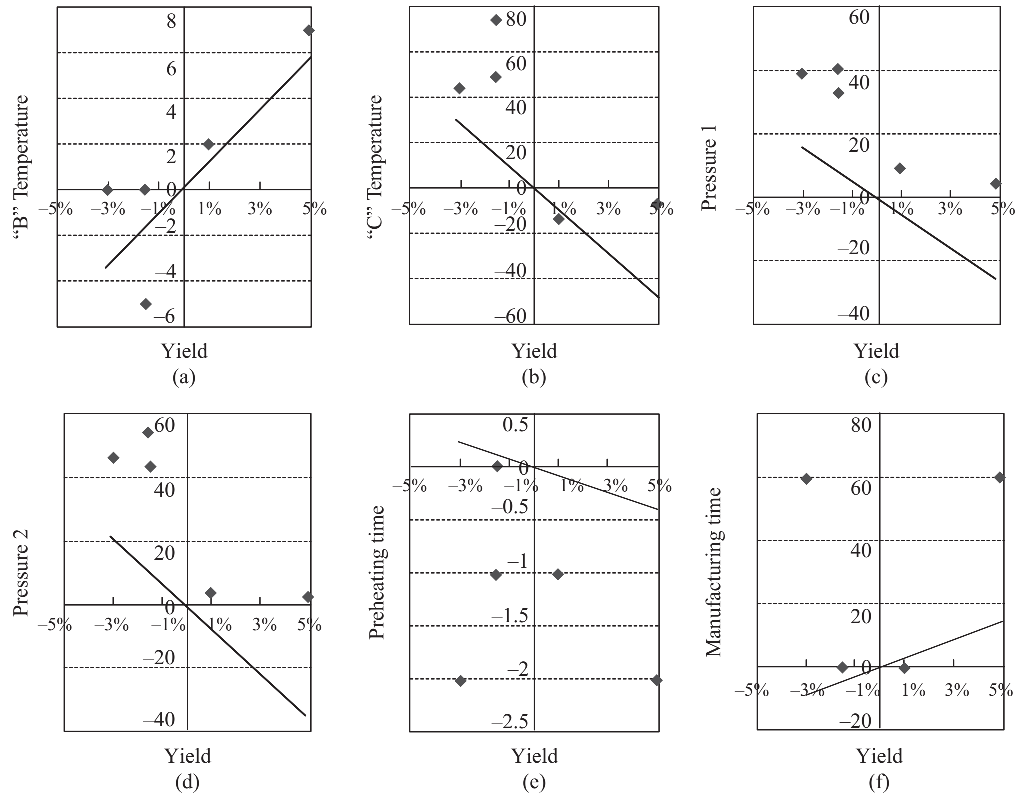


Figure 5.2. Scatter of output value yields and items (normalized data).

(b) Yield and “C” Temperature

Given the right downward incline of the line, the proportional coefficient β turns out negative. Also, the graph shows that the data is arranged in relatively neat formation along the regression line that passes the zero point and outbound scatter is limited. That is to say, the SN ratio η is large, giving credence to the assumption that “C” temperature is also well-suited to the purpose of general yield estimation.

(c) Yield and Pressure 1

The scatter of the data away from the regression line that passes the zero point is considerable. That is to say, the SN ratio η is somewhat small, so it appears that Pressure 1 is not very useful for general yield estimation.

(d) Yield and Pressure 2

The scatter of the data away from the regression line that passes the zero point is considerable. That is to say, the SN ratio η is somewhat small, so it seems that Pressure 2 is not very useful for general yield estimation.

(e) Yield and Preheating Time

The scatter of the data away from the regression line that passes the zero point is substantial. That is to say, the SN ratio η is small, so it appears that preheating time is not useful for general yield estimation.

(f) Yield and Manufacturing Time

The scatter of the data away from the regression line that passes the zero point is substantial. That is to say, the SN ratio η is small, so it seems that manufacturing time is not useful for general yield estimation.

Next, we compare the points explained in (a) to (f) above with the results of the item-by-item computation of the proportional coefficient β and the SN ratio η. The computation of the proportional coefficient β and the SN ratio η is performed item by item with the use of normalized data X_{ij} and normalized output value M_i. How to compute the first item No. 1, "B" temperature, is explained below.

Effective divider

$$\begin{aligned} r &= M_1^2 + M_2^2 + \cdots + M_l^2 \\ &= (-0.0303)^2 + (-0.0159)^2 + \cdots + 0.0489^2 = 0.00389 \end{aligned} \tag{5.3}$$

Total variation

$$\begin{aligned} S_{T1} &= X_{11}^2 + X_{21}^2 + \cdots + X_{l1}^2 \\ &= 0.0^2 + 0.0^2 + \cdots + 7.0^2 = 78.0 \quad (\mathrm{f}=5) \end{aligned} \tag{5.4}$$

Variation of proportional term

$$\begin{aligned} S_{\beta 1} &= \frac{(M_1 X_{11} + M_2 X_{21} + \cdots + M_l X_{l1})^2}{r} \\ &= \frac{\{(-0.0303)\times 0.0 + (-0.0159)\times 0.0 + \cdots + 0.0489\times 7.0\}^2}{0.00389} \\ &= 49.4433 \qquad (\mathrm{f}=1) \end{aligned} \tag{5.5}$$

Error variation

$$S_{e1} = S_{T1} - S_{\beta 1} = 78.0 - 49.4433 = 28.5567 \qquad (f = 4) \tag{5.6}$$

Error variance

$$V_{e1} = \frac{S_{e1}}{l-1} = \frac{28.5567}{4} = 7.1392 \tag{5.7}$$

Therefore, the proportional coefficient β and the SN ratio η can be found as follows:
Proportional Coefficient:

$$\begin{aligned}\beta_1 &= \frac{M_1 X_{11} + M_2 X_{21} + \cdots + M_l X_{l1}}{r} \\ &= \frac{(-0.0303) \times 0.0 + (-0.0159) \times 0.0 + \cdots + 0.0489 \times 7.0}{0.00389} \\ &= 112.7296 = 112.73\end{aligned} \tag{5.8}$$

Given the SN ratio η_1 $S_{\beta 1} > V_{e1}$, the equation works out as follows. The SN ratio in this case is a duplicate ratio.

$$\eta_1 = r\frac{\frac{1}{r}\left(S_{\beta 1} - V_{e1}\right)}{V_{e1}} = \frac{\frac{1}{0.00389}(49.4433 - 7.1392)}{7.1392} = 1523.01 \tag{5.9}$$

The proportional coefficient β and the SN ratio η are found for the other items as well. If the SN ratio η turns out to be negative, the value will be treated as a zero in accordance with the definition given in equation (2.22) in Chapter 2, and the item will not be used in the computation of the general estimation value M. The proportional coefficients β and the SN ratio values η, found item by item, are displayed in Table 5.5.

Table 5.5. The proportional coefficients β and the SN ratios η

β, η	**"B" temp**	**"C" temp**	**P 1**	**P 2**	**Pre-ht time**	**Manuf time**
β	112.73	−968.81	–523.23	–710.78	−7.89	286.84
η	1523.01	315.26	71.21	140.46	0.00	0.00

5.1.5 COMPUTATION OF SIGNAL DATA INTEGRATED ESTIMATE VALUE

The integrated estimate yield value for Signal Data is found using the proportional coefficient β and SN ratio η (duplicate ratio), item by item. Integrated estimate yield value $\hat{M}_i$ for the i-th Signal Data entry is found, based on the equation (5.10), by dividing value X_{ij} of the given item by incline β_j and adding it after it has been weighted with the SN ratio η_j (see Note 5.2).

$$\hat{M}_i = \frac{\eta_1 \times \frac{X_{i1}}{\beta_1} + \eta_2 \times \frac{X_{i2}}{\beta_2} + \eta_3 \times \frac{X_{i3}}{\beta_3} + \eta_4 \times \frac{X_{i4}}{\beta_4} + \eta_5 \times \frac{X_{i5}}{\beta_5} + \eta_6 \times \frac{X_{i6}}{\beta_6}}{\eta_1 + \eta_2 + \eta_3 + \eta_4 + \eta_5 + \eta_6}$$

$$= \frac{1523.01 \times \frac{X_{i1}}{112.73} + 315.26 \times \frac{X_{i2}}{(-968.81)} + 71.21 \times \frac{X_{i3}}{(-523.23\ )} + 140.46 \times \frac{X_{i4}}{(-710.78)} + 0 + 0}{1523.01 + 315.26 + 71.21 + 140.46 + 0 + 0} \tag{5.10}$$

Note that $X_{j1}, X_{j2}, \ldots, X_{j6}$ are normalized "B" temperature, "C" temperature, Pressure 1, Pressure 2, preheating time, and manufacturing time of the i-th Signal Data. From the integrated estimate value $\hat{M}_i$ computation equation, we can see that the larger the SN ratio η of an item, the greater the degree of its contribution to general yield estimation. Note, again, in this regard that the SN ratios η of the preheating time and manufacturing time are negative and therefore treated as zero; they will then not be used for the estimation.

Integrated estimate value $\hat{M}_1$ is computed below for data No. 1 of the normalized signal from Table 5.4.

$$\hat{M}_1 = \frac{1523.01 \times \frac{0.0}{112.73} + 315.26 \times \frac{44.5}{(-968.81)} + 71.21 \times \frac{39.0}{(-523.23)} + 140.46 \times \frac{46.5}{(-710.78)} + 0 + 0}{1523.01 + 315.26 + 71.21 + 140.46 + 0 + 0}$$

$$= \frac{(-28.98)}{2049.95} = -0.0141 \tag{5.11}$$

Table 5.6. Actual and integrated estimate values of yields of signal data

Data No.	Actual value M	Estimated value M
1	−3.03%	−1.41%
2	−1.59%	−1.98%
3	−1.55%	−4.72%
6	0.94%	1.43%
7	4.89%	4.67%

$\hat{M}_i$ is found in the same way for other Signal Data as well. True value (actual value) M and the $\hat{M}$ of the yield are shown in Table 5.6.

5.1.6 COMPUTATION OF THE SN RATIO (db) FOR INTEGRATED ESTIMATE VALUE

Procedures for computing SN ratio η (db) for integrated estimate value are shown below, using data taken from Table 5.6. The computation results will be used in the following Subsection, 5.1.7, where the evaluation of an item's relative importance will be discussed.

Linear equation

$$\begin{aligned} L &= M_1\hat{M}_1 + M_2\hat{M}_2 + \cdots + M_l\hat{M}_l \\ &= (-0.0303)\times(-0.0141) + (-0.0159)\times(-0.0198) + \cdots + 0.0489\times 0.0467 \\ &= 0.00389 \end{aligned} \tag{5.12}$$

Effective divider

$$\begin{aligned} r &= M_1^2 + M_2^2 + \cdots + M_l^2 \\ &= (-0.0303)^2 + (-0.0159)^2 + \cdots + 0.0489^2 = 0.00389 \end{aligned} \tag{5.13}$$

Total variation

$$\begin{aligned} S_T &= \hat{M}_1^2 + \hat{M}_2^2 + \cdots + \hat{M}_l^2 \\ &= (-0.0141)^2 + (-0.0198)^2 + \cdots + 0.0467^2 = 0.00520 \quad (\mathrm{f}=5) \end{aligned} \tag{5.14}$$

Variation S_β of a proportional term

$$S_\beta = \frac{L^2}{r} = \frac{0.00389^2}{0.00389} = 0.00389 \quad (\mathrm{f}=1) \tag{5.15}$$

Error variation

$$S_e = S_T - S_\beta = 0.00520 - 0.00389 = 0.00131 \quad (\mathrm{f}=4) \tag{5.16}$$

Error variance

$$V_e = \frac{S_e}{l-1} = \frac{0.00131}{4} = 0.00033 \tag{5.17}$$

Using the above computation results, the SN ratio η for general estimation can be found as follows:

$$\eta = 10\log\left(\frac{\frac{1}{r}(S_\beta - V_e)}{V_e}\right) = 10\log\left(\frac{\frac{1}{0.00389}\left(0.00389 - 0.00033\right)}{0.00033}\right)$$
$$= 10\log\left(2795.97853\right) = 34.47\,(db) \tag{5.18}$$

5.1.7 EVALUATION OF THE IMPORTANCE OF AN ITEM

Thus far, we have been performing general estimation using all six items. But the six items include some that are effective for integrated estimation purposes as well as others that are not very effective. Therefore, item importance evaluation is performed with the use of an orthogonal array.

5.1.7.1 Item Layout for Orthogonal Array L_{12}

We have six items to deal with here, so an L_{12} type orthogonal array is chosen for item allotment. The reason for choosing an orthogonal array L_{12}, which is a $4\times$prime-type array, is that it will enable the minimization of the impact of any interaction between the items, should any occur (see Note 5.3).

5.1.7.2 Computation of the SN Ratio (db) of the Orthogonal Array by Row

The six items, from "B" temperature to manufacturing time, are allotted to Columns 1 through 6 (Table 5.7). Numerals "1" and "2" in the Table indicate the level at which a given item is allotted to the column, and thus:

Level 1: The item will be used.
Level 2: The item will *not* be used.

In Test No. 1, the data in Columns 1 through 6 are all at Level 1, which signifies that all six items will be used. For the SN ratio of integrated estimation, the SN ratio of η (= 34.47 db), which is found in the previous Subsection, 5.1.6, will be used.

In Test No. 2, because Columns 1 to 5 are at Level 1, five items, "B" temperature through preheating time are used; and it is shown that the integrated estimate SN ratio is 34.47 (db). Likewise, in Test No. 12, because Columns 3, 4, and 6 are at Level 1, three items, Pressure 1, Pressure 2, and manufacturing time, are used, and it is shown that the SN ratio is 20.65 (db).

Table 5.7. Orthogonal array L_{12} and the layout of items

No.	"B" temp	"C" temp	P 1	P 2	Pre-heat time	Manuf time	*e*	*e*	*e*	*e*	*e*	Integrated estimate SN ratio (db)
1	1	1	1	1	1	1	1	1	1	1	1	34.47
2	1	1	1	1	1	2	2	2	2	2	2	34.47
3	1	1	2	2	2	1	1	1	2	2	2	33.87
4	1	2	1	2	2	1	2	2	1	1	2	32.64
5	1	2	2	1	2	2	1	2	1	2	1	33.16
6	1	2	2	2	1	2	2	1	2	1	1	31.83
7	2	1	2	2	1	1	2	2	1	2	1	24.99
8	2	1	2	1	2	2	2	1	1	1	2	24.16
9	2	1	1	2	2	2	1	2	2	1	1	24.29
10	2	2	2	1	1	1	1	2	2	1	2	21.48
11	2	2	1	2	1	2	1	1	1	2	2	18.53
12	2	2	1	1	2	1	2	1	2	2	1	20.65

(Note: Level 1: "Item will be used." Level 2: "Item will *not* be used." Here, items treated as SN ratio $\eta = 0$ in Table 5.5 are allotted as well.)

5.1.7.3 Creation of an Integrated Estimate SN Ratio (db) Factorial Effect Graph

On the basis of Table 5.7, an auxiliary table to the integrated estimate SN ratio (db) is sought and the resulting data is displayed in Table 5.8.

Table 5.8. Integrated estimate SN ratio (db) auxiliary table (averages by level)

Item	Level 1	Level 2
"B" temperature	33.41	22.35
"C" temperature	29.37	26.38
Pressure 1	27.51	28.25
Pressure 2	28.06	27.69
Preheating time	27.62	28.13
Manufacturing time	28.02	27.74

Going further, based on this auxiliary table, Figure 5.3, a factorial effect graph is created. From this factorial effect graph we learn how these “B” and “C” temperatures, with inclines dropping steeply from Level 1 to Level 2, play significant roles in yield prediction.

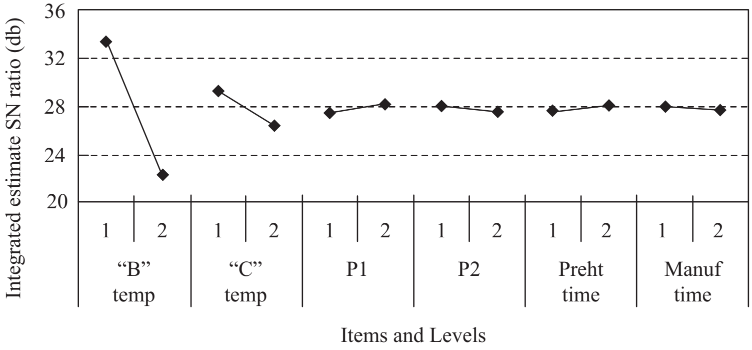


Figure 5.3. Factorial effects for the six items.

5.1.7.4 Computation of Integrated Estimate Value $\hat{M}$ for Signal Data Under Optimum Conditions

A comparison between the SN ratio using only “B” and “C” temperatures, which are assumed to be significant contributors to yield prediction based on the factorial effect graph, versus all of the items is shown in Table 5.9. It is understood that there is not a substantial difference between them.

Table 5.9. Comparison of integrated estimate SN ratios (db)

Case	Items used	Integrated estimate SN ratio (db)
1	All items	34.47
2	“B” temperature, “C” temperature	33.87

5.1.8 INTEGRATED YIELD ESTIMATION FOR UNKNOWN DATA

From the subsequent manufacturing/processing we have acquired some unknown data, as shown in Table 5.10. On the basis of the data, the yield is submitted to integrated estimation.

For the unknown data in Table 5.11, as well, normalization is executed by subtracting from it the average value of each of the items in the Unit Space. Normalized unknown data is shown in Table 5.11.

Table 5.10. Unknown data

Data No.	"B" temp	"C" temp	P 1	P 2	Pre-ht time	Manuf time	Yield
<1>	563.0	306.5	185.5	183.5	2.8	60.0	unknown

Table 5.11. Normalized unknown data

Data No.	"B" temp	"C" temp	P 1	P 2	Pre-ht time	Manuf time	Yield
<1>	–12.0	77.0	19.0	19.5	–4.2	–60.0	unknown

With normalized unknown data <1>, as in other cases we have described, integrated estimate yield value $\hat{M}_{<1>}$ is computed using equation (5.10).

$$\hat{M}_{<1>} = \frac{1523.01 \times \dfrac{(-12.0)}{112.73} + 315.26 \times \dfrac{77.0}{(-968.81)} + 71.21 \times \dfrac{19.0}{(-523.23\ )} + 140.46 \times \dfrac{19.5}{(-710.78)} + 0 + 0}{1523.01 + 315.26 + 71.21 + 140.46 + 0 + 0}$$

$$= \frac{-193.62}{2049.95} = -0.0945 \tag{5.19}$$

5.1.9 COMPUTATION OF INTEGRATED ESTIMATE VALUES BEFORE NORMALIZATION

Through the computations performed so far, yields (output values) have been normalized in terms of the average Unit Space values. Therefore, in order to find the integrated estimate value $\hat{y}$ of the actual yield, the average value M_0 (=84.58%) of the yield of the Unit Space is added to the normalized value $\hat{M}$. For instance, the integrated estimate value $\hat{y}_1$ for the yield of Signal Data No. 1 is found as follows:

$$\hat{y}_1 = \hat{M}_1 + M_0 = -1.41 + 84.58 = 83.17\ (\%) \tag{5.20}$$

For the unknown data as well, the integrated estimate value $\hat{y}$ of the yield is found by adding the yield average value M_0 (= 84.58) of the Unit Space to the likewise normalized value $\hat{M}$. Therefore, the yield's integrated estimate value $\hat{y}_{<1>}$ for unknown data <1> will be:

$$\hat{y}_{<1>} = \hat{M}_{<1>} + \mathrm{M}_0 = -9.45 + 84.58 = 75.13\ (\%) \tag{5.21}$$

Subsequently we have learned that the actual yield value $y_{<1>}$ acquired for unknown data is 73.30%. The actual value y and the generally estimated value $\hat{y}$ of the yields for Signal Data and unknown data are shown in Table 5.12.

Table 5.12. The Actual value y and the integrated estimate value $\hat{y}$ of the yields for signal data and unknown data

No.		Actual value y	Integrated estimate value $\hat{y}$
Signal Data	1	81.55%	83.17%
	2	82.99%	82.60%
	3	83.03%	79.86%
	6	85.52%	86.01%
	7	89.47%	89.25%
Unknown data	<1>	73.30%	75.13%

Figure 5.4 shows a scatter diagram of the actual values and integrated estimate values of the yields for the five Signal Data entries and one piece of unknown data listed in Table 5.12. The horizontal axis represents the actual values of the yield, y, and the vertical axis the integrated estimate values of the yield, $\hat{y}$.

We learn from the figure that the plotted dots overlap rather closely, almost a 45° straight line, indicating that the Signal Data as well as the unknown data are fairly good estimations.

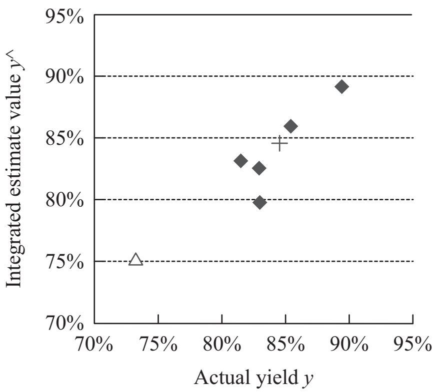


Figure 5.4. Actual yield value y and integrated estimate value $\hat{y}$.

Note 5.1 Omega transformation (Logit transformation)

Dr. Taguchi's work in tile quality improvement assistance through applied orthogonal array tests in the 1950s is today legendary. In the context of the same work, he once encountered a case in which he added up the improvement effects upon the percent defective, which was attributable to various factors, only to find to his surprise that the percent defective figure had come out to a negative value. Realizing that percentage figures could not be added up, he proposed and used a transformation formula that would make it possible to perform percentage figure addition and named it "omega [Ω] transformation" after Fischer's Z conversion of correlative coefficients. Omega transformation is the same thing as the logit transformation used in logistic regression in statistics. This is also the same clarity conversion formula used in telephony, as defined in 1952 by the CCIF (International Telephone Advisory Committee). Dr. Taguchi remarked:

"In a situation where data is accepted only within the limits of 0 to 1 (or 0% to 100%), there is an alternative way of finding the SN ratio by first performing an omega transformation (rather than by using raw data). … Omega transformation is a method of converting the variable domain from (0, 1) to (–∞,∞) through processing, using the following equation in which all ratios take a value within the parameters of 0 < p < 1, as follows:

$$y = -10\log\left(\frac{1}{p} - 1\right) \tag{5.22}$$

"The purpose of the transformation is to allow arithmetical addition to be performed of the effects of various causes with impact on percentage rate data."

In other words, where, as with yield, the permitted scope is restricted to 0% to 100%, it is advisable for the data first to be omega-transformed before being submitted to analysis.[6] The relationship between yield and omega transformation is illustrated in Figure 5.5.

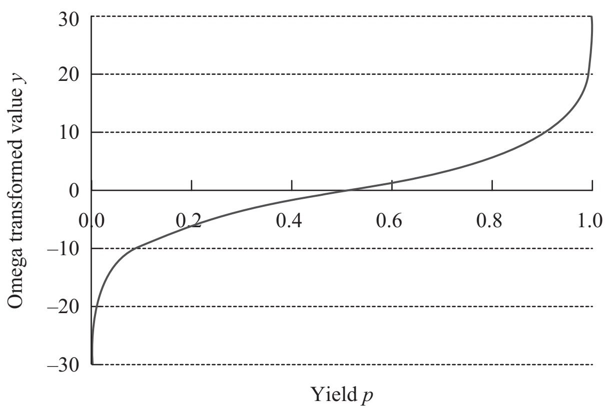


Figure 5.5. Relationship of yield p to omega-transformed value y.

Note 5.1 Omega transformation (Logit transformation) (Cont.)

If addibility is to be established as a valid factor, all percentage values, strictly speaking, should be omega-transformed, but if operational transformations are an issue, a generally followed practice is to omit the conversion altogether if the percentage value in question is within 30–70%; transformation is a sine qua non if less than 10% or more than 90% of data is included. Appropriate judgment on a case-by-case basis is necessary if the data is within 20–80% (see Figure 5.5).

Kodak has defined the photographic and reprographic image density scale with Macbeth density $D = \log 1/p$ (p: permeability or reflectance ratio), but the given formula is p [white 1, black 0] $\rightarrow D$ [white 0, black $+\infty$] with no room for addibility. For this reason, the Taguchi method uses the omega-converted value [white $-\infty$, black $+\infty$] of the absorption coefficient q $(= 1 - p)$ [white 0, black 1].

Dr. Yoshiko Yokoyama, addressing this issue, observed: "Think of an eigenvalue called *percent defective*. Suppose that the percent defective of a process is 2%. It would be quite a challenge to bring it down into the vicinity of 1%. But if the percent defective stands at, say, 30%, it will not be so difficult to reduce it by 1%. What this means is, it can be said, that an entirely different sort of difficulty is involved when reduction of 2% to 1% is the issue and when reduction of 30% to 29% is the issue, even though both cases are about a reduction of 1%."

In dealing with properties with a 0–100% constraint, then, one would do well to first submit the data to omega transformation before performing the evaluation.

We now examine what relevance omega transformation may have with respect to yield prediction for a manufacturing process, which is the theme of Section 5.1. When omega transformation is applied to the yield of 89.47% for Signal Data 7 in Table 5.12 and the yield of 73.30% for unknown data <1>, we obtain the results displayed in Table 5.13.

Table 5.13. Omega-transformed yield values

	Yield p	Omega-Transformed Value y
(a) Yield of 89.47% Improved by 1%		
After improvement	0.9047	9.77
Current	0.8947	9.29
Difference	0.0100	0.48
(b) Yield of 73.30% Improved by 1%		
After improvement	0.7430	4.61
Current	0.7330	4.39
Difference	0.0100	0.22

The omega transformation makes it clear that improving an 89.47% yield by 1% requires more than two times as much improvement efforts as improving a 73.30% yield by 1%.

Note 5.2 Treatment of η and β in integrated estimate equations

Some readers may wonder: "Whereas the integrated estimate equation has η_j and β_j of all items left intact, why is it that they are not expressed as coefficients item by item in the form of $\hat{M}= a_1x_{i1}+a_2x_{i2}+\cdots+a_kx_{ik}$, as would be the case in the regression equation?"

The integrated estimate equation is expressed as follows, as explained in Subsection 2.5.4:

$$\hat{M}_i = \frac{\eta_1\times\dfrac{X_{i1}}{\beta_1}+\eta_2\times\dfrac{X_{i2}}{\beta_2}+\cdots+\eta_k\times\dfrac{X_{ik}}{\beta_k}}{\eta_1+\eta_2+\cdots+\eta_k} \qquad (i=1, 2,\ldots, l) \qquad (5.23)$$

Here, given that ($\hat{M}_{ij}$ =) X_{ij}/β_j is the estimated output value $\hat{M}_{ij}$ found from zero point proportional equation $X_{ij} = \beta_j M_{ij}$ with respect to item j, equation (5.23) can also be expressed as follows:

$$\hat{M}_i = \frac{\eta_1\times\hat{M}_{i1}+\eta_2\times\hat{M}_{i2}+\cdots+\eta_k\times\hat{M}_{ik}}{\eta_1+\eta_2+\cdots+\eta_k} \qquad (i=1, 2,\ldots, l) \qquad (5.24)$$

From equation (5.24), we see that the integrated estimate value, $\hat{M}_i$, is arrived at by applying estimated value $\hat{M}_{i2}$ *for item 2,…*, and estimated value $\hat{M}_{ik}$ for item k for weighted averages, $\eta_1, \eta_2, \ldots, \eta_k$, which indicate the relative degrees of importance of the items, to estimated value $\hat{M}_{i1}$ for item 1.

Furthermore, integrated estimate equation (5.23) can also be developed as follows:

$$\begin{aligned}\hat{M}_i &= \frac{\eta_1\times\dfrac{X_{i1}}{\beta_1}+\eta_2\times\dfrac{X_{i2}}{\beta_2}+\cdots+\eta_k\times\dfrac{X_{ik}}{\beta_k}}{\eta_1+\eta_2+\cdots+\eta_k} \\ &= \left(\frac{\eta_1}{\beta_1\sum_{j=1}^{k}\eta_j}\right)\times X_{i1}+\left(\frac{\eta_2}{\beta_2\sum_{j=1}^{k}\eta_j}\right)\times X_{i2}+\cdots+\left(\frac{\eta_k}{\beta_k\sum_{j=1}^{k}\eta_j}\right)\times X_{ik} \qquad (i=1, 2, \ldots, l) \\ &= a_1X_{i1}+a_2X_{i2}+\cdots+a_kX_{ik} \end{aligned} \qquad (5.25)$$

In view of the above, we understand that the general estimation formula can be expressed as follows, as in the case of multiple regression analysis:

$$\hat{M} = a_1X_{i1}+a_2X_{i2}+\cdots+a_kX_{ik} \qquad (5.26)$$

Note 5.2 Treatment of η and β in integrated estimate equation (Cont.)

If so, why is it that instead of the integrated estimate equation being expressed in the same way the regression equation is, the η_j and β_j of each of its items are left intact? The answer is: this is for the sake of making it easier to see in detail how the degree of impact of each item is taken into consideration by weighting the estimated value of each item with η_j. Another reason is that the digits that follow the decimal point of coefficient a_i will assume increasingly smaller values such that rounding off will produce effect-distorting errors unless many significant digits are assigned.

Note 5.3 What is a 4 × prime two-level orthogonal array?

"The importance of a given item," says Dr. Taguchi, "is evaluated in terms of the degree to which deterioration occurs when the item in question is done without, not in its SN ratio, but in the SN ratio η of integrated estimation. It is thus advisable to use a two-level orthogonal array (even more advisably, a 4 × prime two-level type, specifically). … The purpose of using an orthogonal array for making comparisons of SN ratios η (in the context of integrated estimation) is that, in finding out to what degree each of the items is relevant toward maximizing the accuracy of prediction and estimation. If the evaluation is to have any reliability, each such item needs to be comparatively evaluated under various conditions."

In short, we recommend the use of a two-level-type orthogonal array in the evaluation of the relative importance of items.

A "two-level orthogonal array with a 4 × prime number" refers to any of the following series:

$$L_{12}\,(4\times 3), L_{20}\,(4\times 5), L_{28}\,(4\times 7), L_{44}\,(4\times 11), L_{52}\,(4\times 13), L_{68}\,(4\times 17), \text{ etc.}$$

The cyclic orthogonal array of the 4 × prime number type created using Paley's method, in particular, is known for the almost completely even scatter of interactions between two columns. For more on how to create these orthogonal arrays, refer to Chapter 7.

5.2 CHARACTER PATTERN RECOGNITION USING THE RT METHOD

With the RT Method, no matter how great the number of items is, they are condensed into two variables. In this Section, we use patterns drawn on boards with 5 × 7 grids, each featuring a numeral character pattern consisting of a total of 35 picture elements. In Chapter 4, Section 4.1, we did some exercises in character pattern characterization. For the purposes of the work in this Section, however, we define the white grid as zero (0), and the black grid as one (1). That means that we will have, per character, a set of 35 pieces of data, each either 0 or 1. We will perform character recognition based on the two variables into which the 35 items have been consolidated.

5.2.1 DEFINITION OF THE UNIT SPACE

The Unit Space here is defined as a *group of characters readable as the numeral* "7"; and 16 such character patterns have been prepared. Of the 16 character patterns, character patterns Nos.1 and 15 are displayed below in Figure 5.6.

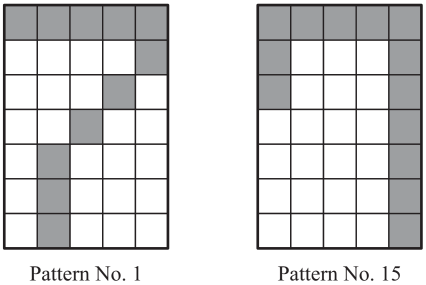


Figure 5.6 Sample patterns of "7" unit space.

The 16 Unit Space characters readable as "7" are each represented, in Table 5.14, as a set of 35 pieces of data, each as either 0 or 1. In the table below, the samples appear in the vertical direction, and the items in the horizontal direction.

"Average value" refers to the Unit Space average, the items of which have all been accounted for. Item No. 35, for example, examined in greater detail here, is as follows:

$$\bar{x}_{35} = \frac{1}{16}\left(0 + \cdots + 0 + 1 + 1 + 1 + 0\right) = \frac{3}{16} = 0.1875 \tag{5.27}$$

Note that, due to lack of space, the average values in Table 5.14 are rounded off at two digits after the decimal point.

As for linear equation L and effective divider r, they are found from equations (5.29) and (5.30) in Subsection 5.2.2, and retain digits as far down from the decimal point as the fourth digit. As demonstrated in Chapter 2, Table 2.13, linear equation L is displayed sample by sample in the (L, r) column, and effective divider r is displayed in the bottom row for average values in the (L, r) column because, in the case of the Unit Space data, the values displayed are average values for linear equations $L_1, L_2, \ldots, L_n$.

Table 5.14. Unit Space character "7" data; linear equation; average values of data

Data No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	L, r
1	1	1	1	1	1	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	7.6875
2	1	1	1	1	1	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	8.4375
3	1	1	1	1	1	1	0	0	0	1	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	8.9375
4	1	1	1	1	1	1	0	0	0	1	1	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	8.7500
5	1	1	1	1	1	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	8.5625
6	1	1	1	1	1	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	9.3125
7	1	1	1	1	1	1	0	0	0	1	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	9.8125
8	1	1	1	1	1	1	0	0	0	1	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	9.5000
9	1	1	1	1	1	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	8.8125
10	1	1	1	1	1	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	9.5625
11	1	1	1	1	1	1	0	0	0	1	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	10.0625
12	1	1	1	1	1	1	0	0	0	1	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	9.5000
13	1	1	1	1	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	7.5625
14	1	1	1	1	1	1	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	8.3125
15	1	1	1	1	1	1	0	0	0	1	1	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	8.8125
16	1	1	1	1	1	1	0	0	0	1	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	9.0000
Average	1	1	1	1	1	0.8	0	0	0	1	0.5	0	0	0.3	0.8	0	0	0.3	0.5	0.3	0	0.2	0.6	0.1	0.2	0	0.6	0.3	0	0.2	0.2	0.4	0.2	0	0.2	8.9141

5.2.2 COMPUTATION OF SENSITIVITY β AND STANDARD SN RATIO OF UNIT SPACE SAMPLES

Using the RT Method computation formula discussed in Subsection 2.5.6, we will find sensitivity β and standard SN ratio η (in the form of a duplicate ratio) for each Unit Space sample. How to find sensitivity β and standard SN ratio η (in the form of a duplicate ratio) for Unit Space sample No. 1 is explained here:

$$\text{Sensitivity } \beta_1 = \frac{L_1}{r} = \frac{7.6875}{8.9141} = 0.8624 \tag{5.28}$$

Where for the linear equation, we have:

$$L_1 = 1\times 1 + 1\times 1 + \cdots + 0\times 0 + 0.1875\times 0 = 7.6875 \tag{5.29}$$

and, for the effective divider,

$$r = 1^2 + 1^2 + \cdots + 0^2 + 0.1875^2 = 8.9141 \tag{5.30}$$

Standard SN ratio η_1 (duplicate ratio) is computed after all fluctuations of S_{T1}, and the proportional item variation $S\beta_1$, etc. have first been found.

Total variation

$$S_{T1} = 1^2 + 1^2 + \cdots + 0^2 + 0^2 = 11 \quad (f = 35) \tag{5.31}$$

Variation of proportional term $S\beta_1$

$$S_{\beta 1} = \frac{L_1^2}{r} = \frac{7.6875^2}{8.9141} = 6.6297 \quad (f = 1) \tag{5.32}$$

Error variation

$$S_{e1} = S_{T1} - S_{\beta 1} = 11 - 6.6297 = 4.3703 \quad (f = 34) \tag{5.33}$$

Error variance

$$V_{e1} = \frac{S_{e1}}{k1} = \frac{4.3703}{34} = 0.1285 \tag{5.34}$$

Accordingly, standard SN ratio η_1 (duplicate ratio) is given as the following equation:

$$\eta_1 = \frac{1}{V_{e1}} = \frac{1}{0.1285} = 7.7798 \tag{5.35}$$

In the same manner, sensitivity β and standard SN ratio η (duplicate ratio) have been found for each individual sample of the Unit Space separately and are tabulated in Table 5.15.

Table 5.15. Sensitivity β and standard SN ratio η (duplicate ratio) for each individual unit space sample

Sample No.	Sensitivity β	SN ratio η
1	0.8624	7.7798
2	0.9465	8.4712
3	1.0026	8.4179
4	0.9816	7.7079
5	0.9606	12.2514
6	1.0447	14.9697
7	1.1008	15.4650
8	1.0657	11.8238
9	0.9886	14.8608
10	1.0727	19.5190
11	1.1288	20.7178
12	1.0657	11.8238
13	0.8484	7.4169
14	0.9325	8.0029
15	0.9886	7.9293
16	1.0096	8.6885

5.2.3 COMPUTATION OF TWO VARIABLES Y_1 AND Y_2, UNIT SPACE SAMPLE BY SAMPLE

We compute the two variables Y_1 and Y_2 using sensitivity β and SN ratio η (duplicate ratio) in Table 5.15. The new variable Y_1 will retain sensitivity β in intact form, but Y_2 will not retain the intact form of SN ratio η, which will be converted to enable the evaluation of the state of scatter from the standard condition (average value of the Unit Space character "7" in Table 5.14). The computation of Y_1 and Y_2 for Unit Space sample No. 1 in Table 5.15 works out as follows:

$$Y_1 = \beta_1 = 0.8624 \tag{5.36}$$

$$Y_2 = \frac{1}{\sqrt{\eta_1}} = \frac{1}{\sqrt{7.7798}} = 0.3585 \tag{5.37}$$

In the same way, we compute Y_1 and Y_2 for all the Unit Space samples; the average values for both as found are shown in Table 5.16.

Table 5.16. Y_1 and Y_2 for the unit space samples

Sample No.	Y_1	Y_2
1	0.8624	0.3585
2	0.9465	0.3436
3	1.0026	0.3447
4	0.9816	0.3602
5	0.9606	0.2857
6	1.0447	0.2585
7	1.1008	0.2543
8	1.0657	0.2908
9	0.9886	0.2594
10	1.0727	0.2263
11	1.1288	0.2197
12	1.0657	0.2908
13	0.8484	0.3672
14	0.9325	0.3535
15	0.9886	0.3551
16	1.0096	0.3393
Average	1.0000	0.3067

5.2.4 COMPUTATION OF DISTANCES OF UNIT SPACE SAMPLES

After going through the preceding procedures, we have before us two variables, Y_1 and Y_2, which consolidate the 35 feature values of each of the Unit space samples. Next, using the MT Method, the Mahalanobis Distance, *MD,* of each of the Unit Space samples from the center of the Unit Space is found. The result of this procedure is displayed in Table 5.17.

For a discussion of computations revolving around the MT Method, refer to Subsections 2.3.2 and 2.5.1. If the RT Method is chosen, the computation process will be identical to that which is discussed in 2.3.2, since computation of two-dimensional Mahalanobis Distances will be involved. The correlation coefficient between Y_1 and Y_2 in this case will be –0.779. Note also that the average value of the MD for Unit Space data using the MT Method is 1.0.

Table 5.17. Distances of unit space samples

Sample No.	**MD(D^2)**
1	1.770
2	0.291
3	0.788
4	1.023
5	0.990
6	0.504
7	0.863
8	0.521
9	1.449
10	1.415
11	1.650
12	0.521
13	2.087
14	0.467
15	0.938
16	0.722
Average	1.000

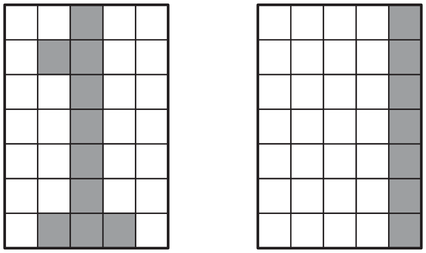
5.2.5 SIGNAL DATA

In connection with performing an evaluation of discrimination ability, we prepared as Signal Data two groups of 12 characters, one a group of characters readable as the numeral "1," and another a group of characters readable as the numeral "9," which resemble the Unit Space character "7." If discrimination is possible with characters that are similar, it will be easy to discriminate dissimilar characters. Of the 12 character patterns in the Signal Data set of characters readable as "1," character patterns Nos. 1 and 4 are shown in Figure 5.7.

As for the 12 patterns for the Signal Data "1," the 35 data sets (0, 1) are as shown in Table 5.18. The samples appear in the vertical direction, and the items in the horizontal direction.

As for Signal Data "9," out of the 12 character patterns prepared for it, patterns Nos. 1 and 5 are shown below as examples in Figure 5.8.

In the same manner as described earlier, the 35 data sets (0, 1) for the 12 patterns for Signal Data for "9," are shown in Table 5.19. The samples appear in the vertical direction, and the items in the horizontal direction.



(a) Pattern No. 1 (b) Pattern No. 4

Figure 5.7. Sample signal data "1" patterns.

5.2.6 Sample-by-Sample Computation of Sensitivity and Standard SN Ratio for Signal Data

As is the case for Unit Space, sensitivity β and standard SR ratio η (duplicate ratio) are found for each individual sample of the signal. We will explain below how to find sensitivity β_1 and standard SN ratio η_1 (duplicate ratio) for sample No. 1 of signal "1."

$$\text{Sensitivity } \beta_1 = \frac{L_1'}{\text{r}} = \frac{2.6875}{8.9141} = 0.3015 \tag{5.38}$$

where:
Linear equation

$$L_1' = 1 \times 0 + 1 \times 0 + \cdots + 0 \times 1 + 0.1875 \times 0 = 2.6875 \tag{5.39}$$

Total variation

$$S_{T1} = 0^2 + 0^2 + \cdots + 1^2 + 0^2 = 10 \quad (\text{f} = 35) \tag{5.40}$$

Variation of proportional term $S_{\beta 1}$

$$S_{\beta 1} = \frac{L_1'^2}{r} = \frac{2.6875^2}{8.9141} = 0.8103 \quad (\text{f} = 1) \tag{5.41}$$

Error variation

$$S_{e1} = S_{T1} - S_{\beta 1} = 10 - 0.8103 = 9.1897 \quad (\text{f} = 34) \tag{5.42}$$