CHAPTER 3

RESEARCH METHODOLOGY

The details of the research methodology is divided into four major parts. The first part addresses the site surveying; the second deals with the data collection and pre-analysis; the third part focuses on the data analysis and modeling using two main analytical approaches, namely, Factor Analysis (FA) and Multiple Regression Analysis (MRA) and the fourth part examines management applications based on the results of the first three parts.

3.1 Site surveying:

The site survey process was conducted by visiting three industrial paper production sites in Thailand. The criteria for site selection are as below.

- All input variables, both material input and utility consumption, in the production process, can be measured in the daily operation.
- All output variables, both wastewater quality and quantity of SS, TDS, COD and BOD, can be measured in the same operational day.
- Records of Data from both input and output variables are available and are of good quality in terms of the accuracy of the data.

3.2 Data collection and pre-analysis:

Data collection was carried out at the site selected by considering input and output variables in the daily operations from each of major processing steps (Figure 3.1). The input variables, fibrous materials are collected and measured at the step of stock preparation in the unit of kg/day. Water and electricity are collected from all steps of process related to the operation of paper machine. Water is measured in terms of mill water flow (m^3/day), electricity is measured in the unit of kWhr/day. Wet end chemicals (alum, emulsifier, clay, defoamer, cato, starch, wet strength) are measured at the first step on paper machine called wet end operations. Coating chemicals are measured at the second step on paper machine called dry end

operations. The output variables are measured in terms of water quality (mg/l) and mill water flow (m^3/day) and calculated as wastewater loadings.

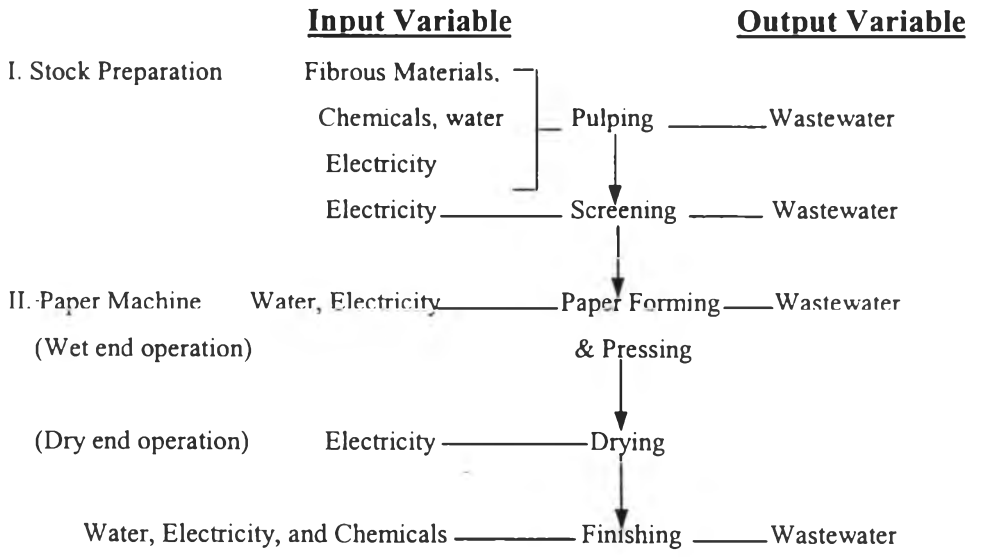

Figure 3.1 Input and output variables from papermaking process

Data pre-analysis is performed by data compilation, conversion, and conditioning as follows.

3.2.1 Data compilation is conducted by recording daily operational data regarding input variables: type and amount of fibrous materials, type and amount of chemicals, water use, electricity consumption and wastewater load variables (SS load, TDS load, COD load, and BOD load). These data were entered into a separate spreadsheet for each kind industrial paper: Gypsum Back, Gypsum Face and Duplex coated board.

3.2.2 Data conversion is carried out by converting the unit of all measured variables from time base (day) to production base (kWhr/ton for electricity, m^3/ton for water and kg/ton for other input variables).

3.2.3 Data conditioning is performed by identifying and removing invalid data. Because this industrial papermaking facility has a single production line for three kinds of industrial paper, the data recorded during the change of paper grades are removed because they are not representative of any single paper grade. These data are organized into the spreadsheet in the form of a matrix with observations as rows and input variables as columns.

3.3 Data Analysis and Modeling:

3.3.1 Factor Analysis (FA):

The purpose of FA is to transform large portions of the entire set of variables into a smaller and more interpretable sets, so that only the essential information remains. FA derives a mathematical model from which factors are estimated. In this study, FA model is constructed for the production conditions of industrial papermaking assisted by SPSS Programs through the data reduction technique. The process of FA follows six major steps that are described below [11, 43-45]:

Step I: Preparation of the Original Input Data Matrix for all variables,

Step II: Calculation of the Correlation Matrix between variables,

Step III: Factor Extraction or Factorization,

Step IV: Rotation of the Factor Matrix (optional),

Step V: Calculation of the Factor Scores.

Step VI: Validation of FA model.

The details of each step are presented in the following discussion (Figure 3.2):

Step I: Preparation of the Original Input Data Matrix for all variables,

The data matrix contains the complete set of data for every observation or case, providing all of the valid input variables to be used for calculation of the

$$\mathbf{X}_{(nxp)} = \begin{bmatrix} x_{11} & x_{21} & .. & x_{p1} \\ x_{12} & x_{22} & .. & x_{p2} \\ \vdots & \vdots & .. & \vdots \\ x_{1n} & x_{2n} & .. & x_{pn} \end{bmatrix}$$

correlation coefficient in the next step. In this matrix, all of the types of input variables that are converted into the same units of product produced (kg/ton) are put into the columns of the matrix while the numbers of the cases are put into the rows of the matrix. The size of the matrix depends on the number of all of the variables (p) and the number of all cases, or the sample size (n) obtained from daily operations. The original data matrix is obtained in the form shown above. Each column represents one input variable, such as water, electricity, and etc., while its values for all observations or cases are given in row.

Through step I, VI original data input matrix with different input variables (material supplies and utility consumption) and different cases (n) for both gypsum

liner board and duplex coated board are obtained. These data matrices are used for calculation of the correlation matrices.

Step II: Calculation of the Correlation Matrix between variables,

In this step, each pairwise of input variables in the original data matrix is calculated to find a correlation coefficient (r) by this equation;

$$r_{ij} = \sum_{k=1}^{n}(x_{ik} - \bar{x}_i)-(x_{jk} - \bar{x}_j)/\sqrt{\sum_{k=1}^{n}(x_{jk} - x_j)^2 \sum_{k=1}^{n}(x_{jk} - x_i)^2 / n}$$

where $i, j = 1, 2, \ldots\ldots\ldots, p$

The value of the correlation coefficient is a measure of the degree of dependency between two variables (x_i, x_j). Then the correlation matrix is presented in the form of a symmetric ($n \times n$) matrix and the diagonal value ($r_{11}, r_{22}, .., r_{pp}$) of **R** is 1. The correlation value (r) is between −1 and +1. The meaning of correlation coefficients in the correlation matrix is as follows.

1) A coefficient of +1 indicates that the two variables are perfectly positively correlated or have correlation in the same direction, so as one variable increases, the other increases by a proportionate change.
2) A coefficient of −1 indicates a perfect negative relationship or correlation in the opposite direction. If one variable increases the other decreases by a proportionate change.
3) A coefficient of zero indicates no linear relationship at all. For example, when one variable changes, the other may stay the same.

$$\mathbf{R}_{(pxp)} = \begin{bmatrix} r_{11} & r_{12} & .. & r_{1p} \\ r_{21} & r_{22} & .. & r_{2p} \\ \vdots & \vdots & .. & \vdots \\ r_{p1} & r_{p2} & .. & r_{pp} \end{bmatrix}$$

For any variables to be grouped into the same factor, a magnitude of correlation coefficient (r) of each pair of these variables must be greater than 0.3 in absolute terms. Based upon the value of correlation coefficients, the variables are grouped into the same particular unobserved variables called factors by FA in the next step. Thus, the correlation matrix is fundamental matrix used for factor extraction.

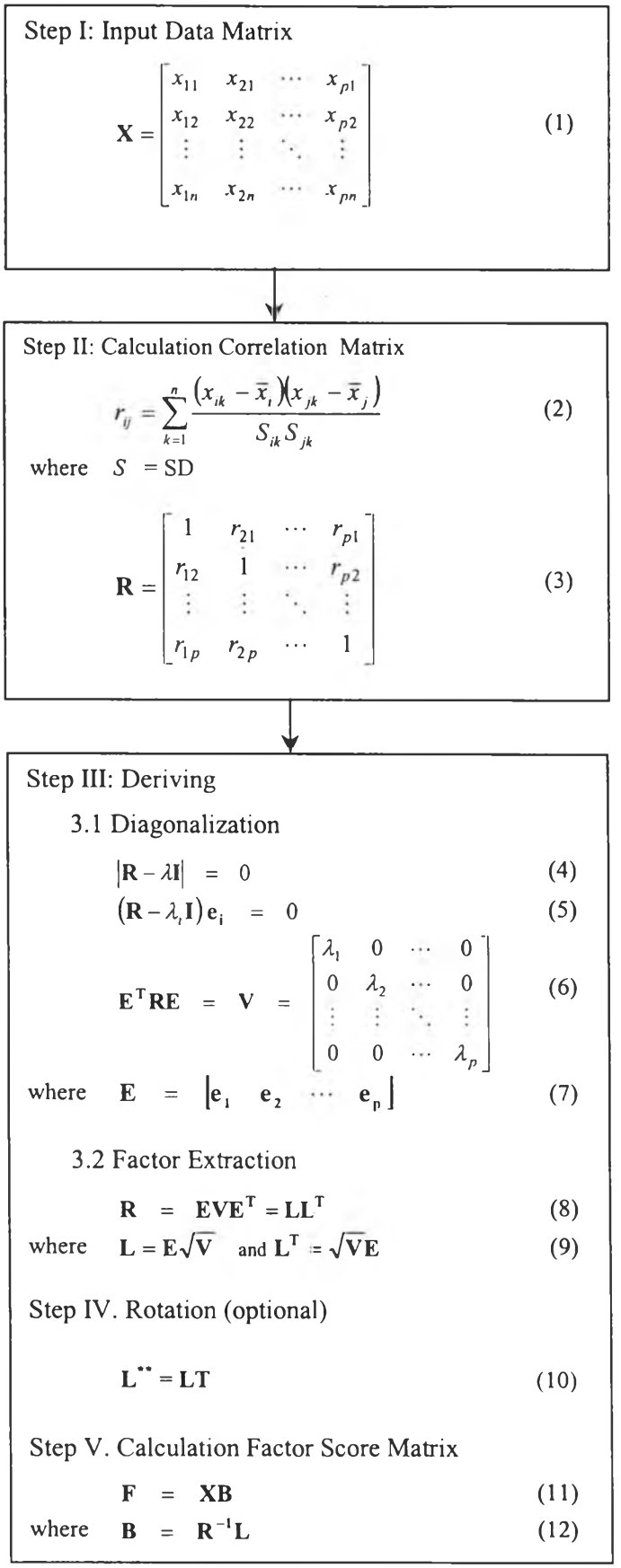

Figure 3.2 Factor analysis procedure for Gypsum liner board production

Step III: Factor Extraction or Factorization,

The goal of factor extraction or factorization is to determine the factors that can be extracted from a large set of data through the correlation matrix. According to the theory of descriptive statistical analysis of multivariate variables, some data exhibit multivariate non-normal distribution due to the non-normality of some variables. The method called " principle component" is, therefore, appropriate for factorization in such situations because there is no restriction requirement for normal distribution.

The linear combinations of the variables are found in the principle component method using a correlation matrix (**R**) which is a symmetric matrix. Then the eigenvalues and eigenvectors ; $\lambda, \mathbf{e}$ are obtained from the roots of characteristic equation below.

$$\mathbf{(R-\lambda I)X = 0} \tag{3.1}$$

According to the important properties of the characteristic roots and vectors of every real symmetric matrix, there exists an orthogonal matrix **P** such that $\mathbf{P^T R\ P = D}$ where **D** is the diagonal matrix of the characteristic roots of **R**.

Through the diagonalization of **R** by post-and pre-multiplying it by the matrix of eigenvectors (**E**) and its transpose ($\mathbf{E^T}$), the diagonal matrix of eigenvalues (**V**) is obtained as shown in the following equation;

$$\mathbf{E^T RE = V} \tag{3.2}$$

where

$$\mathbf{V} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix}$$

where

$$\mathbf{E} = \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \cdots & \mathbf{e}_p \end{bmatrix}$$

The columns in **E** are called eigenvectors, and the values in the main diagonal of **V** are called eigenvalues. The first eigenvector corresponds to the first eigenvalue, and so forth. From this equation, the matrix of eigenvectors (**E**) pre-multiplied by its

transpose ($\mathbf{E}^T$) produces the identify matrix (**I**) with ones in the positive diagonal and zeros elsewhere. This equation can be reorganized in this following form.

$$\mathbf{E}^T\mathbf{E} = \mathbf{I} \tag{3.3}$$

$$\mathbf{R} = \mathbf{E}\mathbf{V}\mathbf{E}^T \tag{3.4}$$

Then the correlation matrix (**R**) can be decomposed into the matrices of eigenvalues and corresponding eigenvectors. Through the scaling of the principle component (factor) the various magnitudes of factor vectors are obtained as shown below.

$$\mathbf{R} = (\mathbf{E}\sqrt{\mathbf{V}})(\sqrt{\mathbf{V}}\mathbf{E}^T) \tag{3.5}$$

where $$\sqrt{\mathbf{V}} = \begin{bmatrix} \sqrt{\lambda_1} & 0 & .... & 0 \\ 0 & \sqrt{\lambda_2} & .... & 0 \\ : & : & .... & : \\ 0 & 0 & .... & \sqrt{\lambda_p} \end{bmatrix}$$

and $\mathbf{E}\sqrt{\mathbf{V}}$ is **L** and $\sqrt{\mathbf{V}}\mathbf{E}^T$ is $\mathbf{L}^T$

$$\mathbf{R} = \mathbf{L}\mathbf{L}^T \tag{3.6}$$

where **L** is the un-rotated factor matrix of factor loadings and $\mathbf{L}^T$ is the transpose of **L**

The form of correlation matrix in equation 3.5 and 3.6 shows the achievement of diagonalization in the form of each combination of eigenvectors and the square root of eigenvalues. The result means that all variables are dispersed from the correlation matrix with different magnitudes. These magnitudes explain the dispersion of variables called "variance". This equation is called the fundamental equation of FA, because it represents the decomposition of the correlation matrix into the product of the factor matrix of factor loadings; **L** and its transpose ($\mathbf{L}^T$).

Through step III, the eigenvalues are obtained and the un-rotated factor matrix (**L**) can be found by direct matrix multiplication ($\mathbf{L} = \mathbf{E}\sqrt{\mathbf{V}}$) as below.

$$\mathbf{L} = \begin{bmatrix} l_{11} & l_{21} & \cdots & l_{p1} \\ l_{12} & l_{22} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ l_{1p} & l_{2p} & \cdots & l_{pp} \end{bmatrix}$$

where l = factor loading

This factor matrix of factor loading (**L**) describes the pattern variation of variables (p variables) with common factors (F_m factors) in different magnitudes, but in the same direction for each common factor. These magnitudes, called factor loadings (l) or correlation coefficients between variables and factors, explain how each variable relates to each factor based on the same eigenvalue or variance. The first factor will account for the largest amount of variance in the data sample. The second factor accounts for the next largest amount of variance remaining in the sample and will be uncorrelated with the first factor. The successive factors account for smaller proportions of the total sample variance, and all factors will be uncorrelated with each other. Based upon the significance of factor loading (> 0.5) in this factor matrix, the highly correlated variables are grouped into the same particular factors with different percentage of total variance explained. Then, the number of significant factors is obtained.

Due to the scaling of factor in the eigenvectors matrix as unit vector, the proportion of variance that each common factor accounts for is one for all variables. That is the communality of a variable is also one for all the variables.

The communality for a variable is the sum of the squared loadings (SSL) within a variable across factors. The proportion of variance in the original variables accounted for by a factor is represented by the SSL for the factor divided by the number of variables for the orthogonal rotation.

Then, the eigenvalues or latent roots will be used as criteria in selecting the number of significant factors. The factor that has eigenvalue greater than 1 is considered to be significant and the factors with eigenvalue less than 1 will be neglected. The reason for the above criteria is that factors with a variance less than 1 are no better than a single variable, since each variable has a variance of 1.

Normally, existing relationship is found as a percentage of total variation or total variance for each factor. Based upon the eigenvalue >1 that is plotted between the eigenvalue and number of factors called Screeplot, all significant factors are

obtained. The percentage of total variance for each factor measures the amount of data variation in the original matrix that can be reproduced by a factor. It is equal to [sum square of loading (SSL) / Number of factors (m)] x 100. This value measures a factor's comprehensiveness and strength. Note that in FA using principle component method, number of factors (m) is equal to number of variables (p).

In this step, if the factor loadings are clearly shown, rotation is not necessary. If they are not clearly shown, then rotation is performed to find the most meaningful interpretation of extracted factors. This means that high (> 0.5) or low (< 0.5) values of factor loading can be classified into factors, but moderate values of factor loading is difficult to classify into any factor. In this case, rotation should be performed.

Step IV: Rotation of Factor Matrix (optional),

The purpose of rotating the factor matrix is to redistribute the variance from earlier factors to later ones in order to achieve the most meaningful pattern of factors. Normally, it is used after extraction to maximize high correlations and minimize low ones. The orthogonal rotation is conducted by rotating a pair of factor axes in a perpendicular line that are passed through the points of all variables, as much as possible. Thus, the angle between these axes are always maintained at 90-degrees. This will make as many values of factor loading in each column of the factor matrix, as close to zero as possible or will make the number of high loading as small as possible. The rotation method called varimax will be used because its goal is to maximize the sum of variances of the required loading in the columns of the factor matrix. This orthogonal rotation method implies that the factors are mathematically independent. In this step, the rotated factor matrix (**L****) is obtained by multiplying un-rotated factor matrix (**L**) with the transformation matrix (**T**) of sines and cosines of an angle ϕ,

$$\mathbf{L^{**}} = \mathbf{LT} \tag{3.7}$$

where

$$\mathbf{T} = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}$$

After rotation is performed, other relationships already mentioned in Step III can also be found, namely, communality and proportion of variance.

If rotation is not required, the naming of factors may either be carried out or not. Basically, high loading of the contributing variables in each factor are utilized to give a descriptive name to the factors and the label is then communicated to those who would apply the results.

Step V: Calculation of Factor Scores.

Through the factor matrix, the values of factor scores are estimated. These scores measure some definable property of the object that has characteristics or variables with an individual value or score on that factor and represent estimation of the underlying factor value for each case or observation;

$$\hat{F}_i = \text{fn}(\hat{F}_{i1}, \hat{F}_{i2}, \ldots\ldots, \hat{F}_{in}), \quad i = 1, 2, 3, \ldots, n \tag{3.8}$$

Such a value must be based on the individual' values or scores obtained from the directly measured variables, the x_j in the standardized form (Z) called "z score" ($z = (x_{ij} - \bar{x}_j)s_j$). The first step involves calculation of regression coefficients for weighting variable in the standardized form to produce factor scores. Because these factors are extracted using the principle component method, there are no differences in using any method (Bartlett and Anderson-Rubin methods) to calculate the factor score coefficients. However, according to the best estimation of factor score coefficient [11], the linear relationship based on the use of the regression method is the most appropriate approach for estimating these scores as detailed in the following equations:

$$\hat{F}_{ij} = \hat{\beta}_{11}\mathbf{Z}_{11} + \hat{\beta}_{12}\mathbf{Z}_{12} + \ldots\ldots + \hat{\beta}_{1p}\mathbf{Z}_{1}\mathbf{p} \tag{3.9}$$

This equation can be represented in matrix form as the following.

$$\hat{\mathbf{F}} = \mathbf{Z}\hat{\mathbf{B}} \tag{3.10}$$

where $\hat{F}$ is an $n \times m$ matrix of m factor scores for the n observations,

$\hat{B}$ is a $p \times n$ matrix of estimated factor scores coefficients, and

$\mathbf{Z}$ is an $n \times p$ matrix of observed variables

Equation 3.10 can be written as $1/n\,\mathbf{Z}^T \hat{F} = 1/n\,\mathbf{Z}^T\mathbf{Z}\hat{B}$ (3.11)

or $\mathbf{L} = \mathbf{R}\hat{B}$

as $1/n\,\mathbf{Z}^T \hat{F} = \mathbf{L}$ and $1/n\,\mathbf{Z}^T\mathbf{Z} = \mathbf{R}$

Therefore, the factor score coefficient matrix is given below.

$$\mathbf{B} = \mathbf{R}^{-1}\mathbf{L} \tag{3.12}$$

where **B** is a factor score coefficient matrix of variable scores

Factor scores, then, can be generated on the first factor and so forth. Thus, they are a product of matrices of standardized scores on variables (**Z**) and factor score coefficient (**B**) as indicated in the following equation.

$$\mathbf{F} = \mathbf{ZB} \tag{3.13}$$

Moreover, predicting scores on variables from factor scores is also possible as shown in the equation below.

$$\mathbf{Z} = \mathbf{FL}^T \tag{3.14}$$

This equation will be used when there is only an un-rotated factor matrix.

$$\mathbf{Z} = \mathbf{F}(\mathbf{L}^{**})^T \tag{3.15}$$

If a rotated factor matrix is available, equation 3.16 will be used.

Therefore, from equation 3.13 and 3.14, FA provides two form of the equations for calculation of these common factors (F_m) as shown below.

$$F_1 = b_{11}Z_1 + b_{12}Z_2 + \ldots\ldots\ldots\ldots + b_{1m}Zp \quad (I.1)$$

$$F_m = b_{p1}Z_1 + b_{p2}Z_2 + \ldots\ldots\ldots\ldots + b_{pm}Zp \quad (I.p)$$

where b_{ij} = factor score coefficient

Therefore, FA is shown to be useful for isolating, identifying, and discovering the hidden factors that are common to a large number of interrelated variables. These hidden factors or common factors can, then, be used to explain the interrelations of a large set of observable variables.

Through Step V, factor scores matrix of significant factors for all products of industrial paper are obtained and used as the new independent variables or predictor variables for MRA.

Step VI: Validation of FA model.

FA model validation is performed by moving case-by-case. In this method, the most recent n data values for 7 months is used to validate the FA model. The term moving case-by-case is based on the time that as a new observation is used, it replaces the oldest observation in the original data set and a new FA model is computed. The result from step VI provides the stability of the FA model in terms of physical meaning for industrial papermaking.

3.3.2 Multiple Regression Analysis (MRA):

In MRA, a statistical model is derived from a set of independent variables or predictor variables (x) in order to predict the dependent variables or response variables (y). In this work, the MRA model is constructed for prediction of the wastewater load of industrial paper production for the two main products. Development of the MRA model involves five major steps assisted by SPSS program for regression analysis (Figure 3.3) [11,14, 46-47]:

Step I: Data Collection and Preparation of Predictor Variables,

Step II: Model Investigation,

Step III: Model Testing,

Step IV: Estimation of Model Parameters,

Step V: Evaluation and Interpretation of the Model,

Step VI : Validation of the MRA model.

Step I: Data Collection and Preparation of Predictor Variable

In this study, data collection and preparation consist of the same set of data that are used for FA. Data for independent variables or predictor variables are reduced from several variables to a smaller number through FA. These variables are in the form of significant factor scores. Data for dependent variables or response variables include the wastewater loads, namely, SS load, TDS load, COD load, and BOD load. These variables are also in the same case or observation set of significant factor scores. Both predictor variables (x) and response variables (y) are organized into the same spreadsheet for determination of their relationships; $y = f(x)$ in the later steps.

Step II: Model Investigation,

This step provides a measure of the goodness of fit of the estimated regression equation to the data through the least squares method (LS). LS is a technique for finding the value of the regression coefficient that is to minimize the sum of the squared deviations between the observed values of the response variable (y) and the predicted values of the response variable ($\hat{y}$).

Least Squares involves two major processes. First, a response variable is selected and the best fit of the relationship between each predictor variable and the selected response variable is investigated using 11 different basic models (Figure 3.3). The best curve fitting is performed on the basis of the following statistical significances:

1) high coefficient of determination (R^2) and 2) low significance level (α) for overall prediction.

- R^2, the coefficient of determination, is the proportion of the total variance of y explained by the model or accounted for by x; $x_1 \ldots x_p$.

$$R^2 = SSM/SST = 1 - SSR/SST$$

where $SST = \sum(y_i - \bar{y})^2$ and is called the total sum of squared deviations or the total sum of squares before regression. The greater the variation among the y_i observations, the larger is *SST*.

where $SSM = \sum(\hat{y}_i - \bar{y})^2$ and is the sum of squares accounted for by the regression model. The larger *SSM* is in relation to SST, the greater is the effect of the regression relation in accounting for the total variation in the y_i observations.

where $SSR = \sum(y_i - \hat{y}_i)^2$ and is the sum of the squared residual or error., it is based on the difference between the values of the observed variable (y_i) and the prediction value ($\hat{y}_i$) for each case. The greater the variation of the y_i observations around the fitted regression line, the larger is *SSR*.

The relationship among *SST*, *SSM*, and *SSR* is as shown below.

Total Sum of Squares (*SST*) = Explained Sum of Square of Regression Model (*SSM*)

$$\sum(y_i - \bar{y})^2 \qquad \sum(\hat{y}_i - \bar{y})^2$$

+ Residual or Error Sum of Squares (*SSR*)

$$\sum(y_i - \hat{y}_i)^2$$

The values of R^2 are between 0 to +1. If x can account for all the variance of y then *SSM* is equal to *SST* and R^2 is equal to one. Since R^2 is the proportionate reduction of total variation associated with the use of predictor variable (x), the larger R^2 is, the more the total variation of y_i is reduced by introducing the predictor variable (x).

Thus, in the second process, the best curve fit is checked for its significance. In this process, *F-statistic* is performed to test the significance of R^2, while *SE* (Standard Error of the Estimate) is used to measure the dispersion of the observed points around the computed points.

- *F-statistic* = *Explained Mean Square/Residual Mean Square,* or

  $F = MSM/MSR$

where $MSM = SSM/k$ and $MSR = SSR/n-k-1$

k = number of samples of predictor variables

n = size of each samples of predictor variables

- Standard Error of the Estimate; $SE = SSR/(n-(k+1))$

All determinations of statistical significance of the two major processes (*F-statistic and SE*) can be obtained through the analysis of variance.

- Analysis of Variance (ANOVA): This statistical method involves a two - stage process. The first stage is used to investigate whether there is a difference in the means among the variables, and then to locate where these differences may be. The second stage is conducted depending on the nature of the hypothesis (Table 3.1).

Table 3.1 ANOVA Table for Linear Regression

Source of Variation	Sum of Square (*SS*)	Degree of Freedom (*df*)	Mean of Square (*MS*)
Regression or Model	*SSM*	*k*	*MSM=SSM/k*
Residual or Error	*SSR*	*n-(k+1)*	*MSR=SSR/(n-k-1)*
Total	*SST*	*n-1*	*MST=SST/(n-1)*

From Table 3.1, it can be seen that the relationship that holds among the sums of squares (that is, *SST =SSM + SSR*) also holds for the degrees of freedom (*df*); *Total df = Regression model df + Residual or error df.*

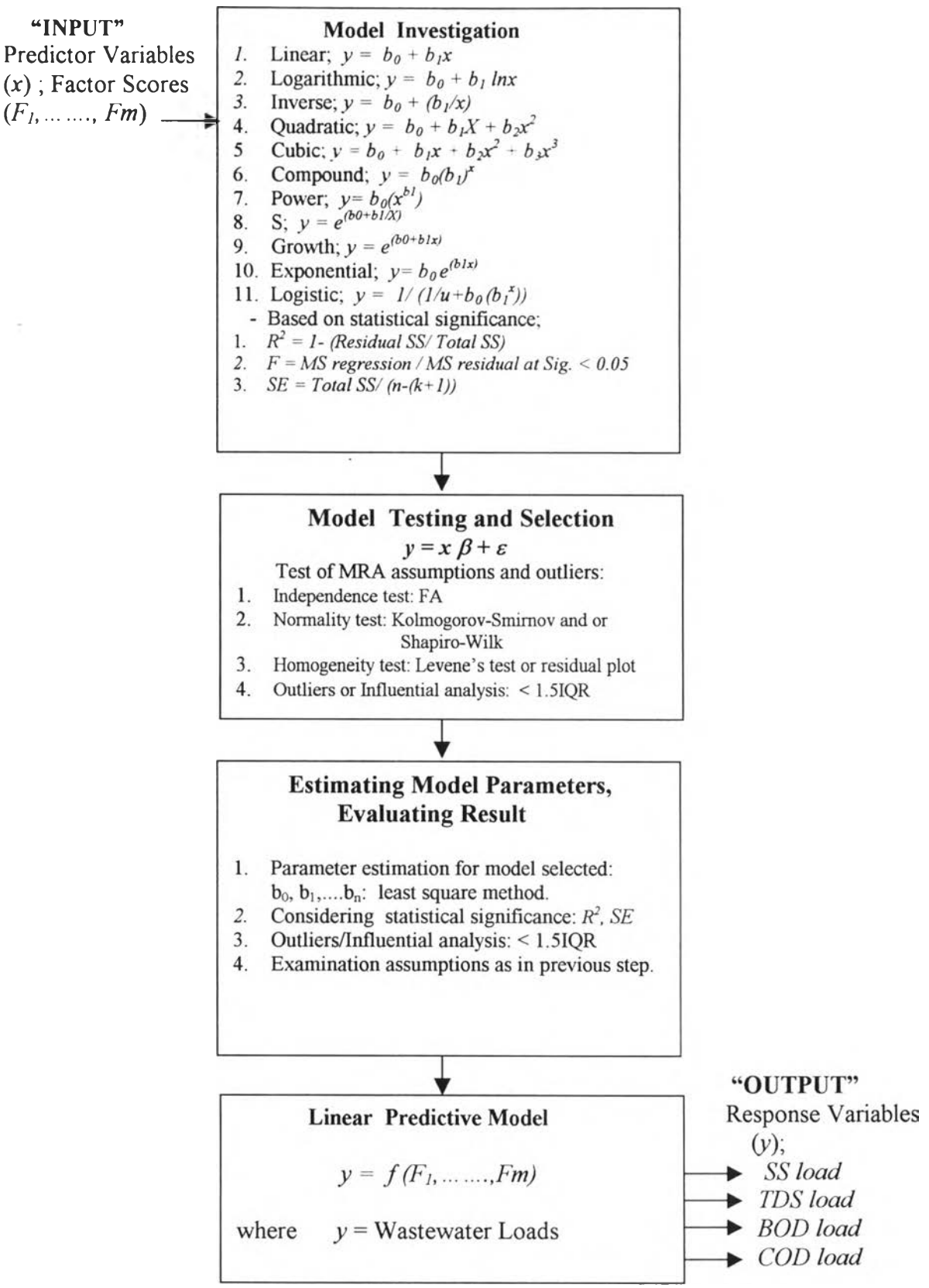

Figure 3.3 Multiple Regression Analysis Procedure for Industrial Paper Production

Then, the hypothesis are tested using the *F-statistic* that checks whether the hypothesis can be either rejected or not. The procedure of hypothesis testing concerns these standard terms.

- Null hypothesis (denoted by H_0):

  H_0: $\beta_i = \beta_1 = \beta_2 = \ldots = \beta_k = 0$ or

  All predictor variables in the model could not be used for prediction.

  where $\beta_i, \ldots, \beta_k$ are the model parameters that have values of 0.

In this hypothesis, the statement of a zero or null difference is directly tested in the sense that the final conclusion will be either rejection of H_0 or failure to reject H_0.

- Alternative hypothesis (denoted by H_a):

  H_a: $\beta_i = \beta_1 = \beta_{i2} = \ldots = \beta_k \neq 0$ or

  Some predictor variables in the model could be used for prediction.

  where $\beta_i, \ldots, \beta_k$ are the model parameters that have not values of 0.

In this hypothesis, the statement must be true if the H_0 is false.

The conclusion involved a decision either to reject or to fail to reject the H_0 is determined by a comparison of the *F-statistic* and the critical value.

- Critical value: The value that separates the critical region from the values of the *F-statistic* that would not lead to rejection of the H_0.
- Significance level (α): The probability of rejecting the H_0 when it is true. Typical value in this study is 0.05 that is the value of $\alpha = 0.05$.

If the *F-statistic* calculated is less then or equal the critical value at $\alpha = 0.05$, H_0 is accepted. This means that all predictor variables (y) do not influence the response variable (x) or they are not appropriated for prediction.

If the *F-statistic* calculated is greater then the critical value at $\alpha = 0.05$, H_a is rejected. This means that at least one predictor variable (y) influences the response variable (x) or that it can be used for prediction. In this case, the *t-statistic* is used for testing any model parameter (β_i) $\neq 0$ or which x is in relation to y by the following hypothesis.

H_0: $\beta_i = 0$

H_a: $\beta_i \neq 0$, $i = 1, \ldots, k$

$$t = \frac{b_i - 0}{S_{b_i}}$$

where S_{b_i} = Standard deviation of b_i

If the *t-statistic* calculated is less then or equal the critical value at $\alpha = 0.05$, H_0 is accepted.

If the *t-statistic* calculated is greater then the critical value at $\alpha = 0.05$, H_a is rejected.

Thus, at any given α level, either the *F-statistic* or *t-statistic* can use for testing $\beta_i = 0$ or $\beta_i \neq 0$.

Usually, when the *F-statistic* is calculated by the computer program for statistical analysis such as SPSS and SAS, the probability value (*P-value*) is considered. Since in a regression model building, the variables involved are assumed in the standard normal distribution (Z) that is a correspondence between area and probability, thus, the *P-value* is equal to twice the area to the right of the *F-statistic*. The *P-value* is the probability of getting a value at least as extreme as the observed one such as the mean value.

If the *P-value* is less than or equal to the significance level (α), that is $p \leq \alpha = 0.05$, the H_0 is rejected.

If the *P-value* is greater than the significance level (α), that is $p > \alpha = 0.05$, H_0 is accepted or failed to reject.

Through the steps of the model investigation, the model parameters and their error terms of the predictor variables (x) for the regression models that meet the above statistical significances are obtained and proposed for testing the assumptions underlying MRA in the next step. However, in any case where the investigation does not succeed, it means that the match of the estimated regression equation to the data does not fit well under 11 different basic models. It may have to do with whether or not there are other more complicated relationships within the data.

Step III: Model Testing,

The proposed model resulting from Step II is tested using the three assumptions underlying MRA: Independence, Normality, and Homogeneity in order to obtain the linear relationship between the predictor and response variables that

does not affect the statistical procedure used for MRA. These assumptions are tested by the following methods.

1. The independence of the error terms can generally be considered through the residual plot that exhibits an association between the errors and a sequencing observation or time. However, in this study, FA is performed before for predictor variables (x; $F_1, \ldots . F_m$). The variables resulted from FA as the predictor variables for MRA are unrelated and independent. Thus, only assumptions 2 and 3 are considered for this application.
2. The normality of the error distribution can be considered through the Kolmogorov-Smirnov approach for sample size (n) > 50 and or Shapiro-Wilks approach for $n < 50$ at a significance level of 0.05 under the following hypothesis.

   H_0: The error of the sampled observation is normally distributed

   H_a: The error of the sampled observation is not normally distributed

   where H_0 is null hypothesis and H_1 is alternative hypothesis.

   The test statistic is $D = \sup_x |S(x) - F_0(x)|$,

   where D equals the supremum, overall x, of the absolute value of the difference $S(x) - F_0(x)$.

   where $S(x)$ is the observed cumulative frequency probability function computed from the sample data,

   where $F_0(x)$ is the expected relative cumulative frequency probability function.

$F_0(x)$ can also be computed from the standard normal curve area. Thus, the value of x in each observation is converted to the form of standardized variables. Then the probability of each standardized variables for each

observation is obtained from the table. This probability value will be subtracted from the value in the table, then $F_0(x)$ is obtained.

The largest amount of the differing frequency: D is compared with the value in the table. If the critical value of D is not greater than the value in the standard table, then, H_0 is accepted.

However, when this value is calculated by the computer program, the p-value of D at a significance level of 0.05 is considered. If the calculated p-value of $D \geq \alpha = 0.05$, H_0 is accepted.

As for the Shapiro-Wilk approach, the original data will be converted in the form of logarithm $_{10}$, and the test of normality is also in the form of logarithm $_{10}$. If the p-value of Shapiro-Wilk from the calculation of the computer program is $\geq \alpha = 0.05$, H_0 is accepted.

3. The constant variance or homogeneity of the error term can be considered either by the residual plot between residuals or errors values and predicted values, or Levene statistic that calculate the probability under different conditions. For residual plot, if the data is distributed near zero, it means that the error is constant. For Levene statistic, it is usually calculated based on the mean or the median in order to test the variation of data under the following hypothesis.

H_0: The tested data is not differently distributed.
H_a: The tested data is differently distributed.

If the significance value of this statistic is greater than 0.05, H_0 is accepted.

In this study, if the tested data is in the form of factors, the Levene statistic may not succeed due to the sum of case weight being less then the number of data groups. When this happens, the residual plot is carried out.

Moreover, the outliers as influential observation of errors are considered from the 1.5 Inter-quartile Range (IQR = Q3-Q1) through the Box-plot diagram (see Figure 3.4). When the data have the series of data values from the lowest to the highest, the diagram of Box-plot can describe the data distribution.

Step IV: Estimation of Model Parameters,

The regression coefficients or model parameters of the composite model that is obtained from all of the proposed models for each response variable (y) are used as starting value for estimation of the real model parameters of the final predictive equation through the least square (LS) method.

The principle of the LS method is to minimize *SSR* (sum of squared residuals) or $\sum e_i^2$ (errors).

$$SSR = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_i)^2$$

where y_i is the observed value of the dependent variable or response variable for the i^{th} observation , and $\hat{y}_i$ is the computed value of the dependent variable or predictor variable (x) for the i^{th} observation.

The procedure for finding the values of the parameters using the LS method involves differential calculus that is not essential to understanding the principles of regression analysis.

$$\frac{\partial sse}{\partial b_0} = \frac{\partial \sum (y_i - b_0 - b_1 x_i)^2}{\partial b_0} = -2\sum (y_i - b_0 - b_1 x_i) = 0$$

$$\frac{\partial sse}{\partial b_1} = \frac{\partial \sum (y_i - b_0 - b_1 x_i)^2}{\partial b_1} = -2\sum (y_i - b_0 - b_1 x_i) = 0$$

These equations are rewritten and normal equations are obtained as below.

$$nb_0 + \sum x_i b_1 = \sum y_i$$

$$\sum x_i b_0 + \sum x_i^2 b_1 = \sum x_i y_i$$

The normal equations are solved to determine b_0 and b_1 .

$$b_1 = \frac{ss_{xy}}{ss_{xx}}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

b_0 and b_1 are estimators of model parameters obtaining through the LS method. Once the model parameters are determined, the equation for the best fitting line is obtained due to their good properties such as unbiased and minimized variability.

For the non-linear relationship such as the polynomial regression model; $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$, this step is also conducted by an iteration search process with the method of Levenberg-Marquardt [24], that is designed to capitalize on the best features of the linearization. This method deals with the problem of minimizing a function in the absence of any restrictions for the constructed model. It involves the rate of convergence and the partial derivatives of the model with respect to each of the parameters [25-26].

In this step, the predictive model is obtained and is evaluated in the next step.

Step V: Evaluation and Interpretation the Model.

The purpose of evaluation of the predictive model is to obtain the best fitted model. It is carried out by examining the statistical significance of the predictive model obtained from step IV as described in step II and testing the appropriateness of the predictive model as described in step III.

In addition, one sample T test is performed to determine whether the value of the mean is equal to zero through the significance value of t under the following hypothesis;

H_0: $\mu_0 = 0$ or

The mean of the residual in the model is equal to zero.

H_a: $\mu_0 \neq 0$ or

The mean of the residual in the model is not equal to zero.

The value of *t-test;* $t = \bar{x} - \mu_0 / S\sqrt{n}$ for $n \leq 30$, and $t = \bar{x} - \mu_0 / \sigma\sqrt{n}$ for $n > 30$, if t calculated is greater than critical value at $\alpha = 0.05$, H_0 is accepted, while H_a is rejected. This means that the mean of the residual is equal to zero, and also indicates that the model has a normal distribution.

Through step V, if the result of the model meets the statistical significance and MRA assumptions, this predictive model is obtained and used for validation in the later part. If the model does not meet the statistical significance test, the outlier or influential observations must be removed (detailed in step III, Figure 3.4) and Step II will be repeated.

Step VI: Validation of the MRA model

The purpose for validating the MRA model is to assess its generalizability and its predictive ability of MRA model. All complete observations for each type of industrial paper products obtained from operational days within a 14 month period are used to validate the model. The percentage of relation between each type of wastewater load and related input factors is determined from calculating the coefficient of determination (R^2), just as in the model building step and multiplying it by 100 (% relation $= R^2 \times 100$).

The prediction ability of a predictive environmental model for the wastewater load is then determined by calculating the percentage of relation of the model building (MB), multiplying it by 100 and dividing it by the percentage of the relation of the model validation (MV), (% prediction accuracy = % relation of MV x 100 / % relation of MB). As a rule of thumb for fluctuated data, if the prediction accuracy > 40%, the model is applicable.

The results from the predictive models can be applied in wastewater management due to the ability of the models in identification of the root causes of wastewater generation. The suggested actions derived from the model results are aimed at improving industrial paper production, particularly in unusual cases that may greatly affect environmental quality due to wastewater generation.