A Parametric Model to Estimate Design Effort in Product Development

Adil Muhammad Salam

A Thesis

in

The Department

of

Mechanical and Industrial Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of Masters of Applied Science (Industrial Engineering) at

Concordia, University

Montreal, Quebec, Canada

August 2007

Your file Votre référence
ISBN: 978-0-494-34771-3
Our file Notre référence
ISBN: 978-0-494-34771-3

ABSTRACT

The design and development of a product is a complex process, which requires many resources and various types of expertise. Because the process is complex, it is essential to estimate the design effort required to complete a product development project. The estimation of design effort is, in turn, a prime factor to predicting lead-time, cost, and effort requirements of a project. In this thesis, parametric models for estimating design effort are proposed. A case study involving engineering departments at Pratt & Whitney Canada (PWC) is presented. First, research is conducted on each of the design, aerodynamics, analytical and drafting departments at PWC to identify factors that could be best utilized in estimating design effort. Four factors are identified for parametric modeling. These factors are type of design, degree of change, concurrency, and experience of departmental personnel. The parametric model applied to each department uses least square regression. Furthermore, the jackknife technique is utilized to ameliorate the bias in regression equations coefficients. This research also uses a data masking technique in order to protect confidential data information of PWC. The masking technique enables to calibrate the impact of each factor considered in the parametric modeling, while not being affected by the masking. Data analysis is first utilized to establish regression based parametric models. Later, the regression equations are tested for their validation. It is found that the proposed parametric models provide good estimate of design effort when compared to the original estimates, with maximum relative errors of less than 10%. Furthermore, in each parametric model, the factors that significantly affect the design effort are identified using ANOVA table. Based on the outcomes reported in ANOVA, the number of factors is reduced and new models are

developed with the reduced number of factors. Lastly, the application of the models, limitations and possible future studies are also discussed in this thesis.

ACKNOWLEDGMENTS

First and foremost I would like to thank God.

I would like to thank my thesis supervisors; Drs. Bhuiyan and Gouw, who supported me financially, and whom I consulted throughout this project.

I would like to thank the personnel at Pratt & Whitney Canada, especially Yvan Beauregard, my industry supervisor, who was the main supporter of this Master's project.

I would like to thank my friend Dr. Raza, whom I discussed my ideas with, and who helped me in structuring this thesis.

I would like to thank my father, Abdul Salam for carefully reading and editing my thesis.

I would like to thank my mother, Sajeela Salam for her moral support.

I would like to thank my wife and son, Aisha Sharif and Omar Salam, who patiently waited for me the long nights that "I burned the midnight oil" working on this thesis.

Finally, I would like to thank Natural Sciences and Engineering Research Council of Canada (NSERC) for providing funding in collaboration with Pratt & Whitney Canada, making this project a reality.

Dedicated to my parents, Abdul and Sajeela Salam, as well as my wife and son, Aisha Sharif and Omar Salam

TABLE OF CONTENTS

LIST OF APPENDICES

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND SYMBOLS

A	Design job A
ACT	Actual design effort
a_i (i=1,…,n)	Regression coefficients in the parametric model
B	Design job B
BCADP	Bias-Corrected Correlated-Noise Additive Data Perturbation
C	Design job C
CADP	Correlation-noise additive data perturbation
CD	Investments
CE	Concurrent engineering
CF	Compressor fan
C-GADP	Copula-based GAPD
C-GADP	Copula-based GADP
c_j	Individual concurrency values for period j
COEF	Coefficient of regression
Con	Concurrency
Con_j	Cumulative concurrency values for design job j
D	Design job D
DC	Degree of change
DE	Experience of departmental personnel
DJ	Design job
D_m	Effort driver (m)
DM	Data masking
DP	Data perturbation
DSM	Design structure matrix
E	Design job E
Ê	Estimated design effort in hours
EC	Engineering change
F	Design job F
F_j	Number of functions at level j
GADP	General additive data perturbation
GE	General Electric
H_0	Null hypothesis
H_1	Alternative hypothesis
HE	Home equity
i	Index for person i
IBR LPC	Integrated blade-rotor low-pressure compressor
Jack	Jackknife
L	Liabilities
LA	Liquid assets
LCL	Lower control limit
LR	Linear regression

MD	Masked data
MDP	Multiplicative data perturbation
MLRM	Multiple linear regression method
MMRE	Mean magnitude relative error
MMSE	Mean magnitude square error
n	Sample of size n
No Con	No concurrency
ns	Number of sub samples
NW	Net worth
OD	Original data
p	Number of confidential attributes
PC	Product complexity
PGM	Performance generation model
Pr	Probability
PRED	Predicted
Ps_i	Pseudo-value for the entire sample, omitting sub-sample i
PWC	Pratt & Whitney Canada
q	Number of non-confidential attributes
r	Correlation coefficient
R^2	Coefficient of determination
r_L	Critical value of the coefficient of correlation
S	Matrix containing non-confidential attributes
S/B	Stocks/Bonds
S/C	Savings/Checking
SADP	Simple additive data perturbation
SAS	Single attribute security
SPC	Statistical process control
SPLC	Security provided for linear combinations
SR	Severity of requirements
SSR	Sum of squared regression
SSTO	Sum of squared total
t	t-test value
TD	Type of design
U	Matrix {X,S}
UCL	Upper control limit
V	Matrix {S, Y}
X	Matrix containing confidential attributes
X_i	i^{th} observation of independent variable in a simple linear regression
y	Dependent variable in a simple linear regression
y'	Manpower utilized in each time period
Z_{ε_i}	Standardized residual of observation i
α	Shape parameter

$\hat{\beta}$	Estimated coefficient of regression
$\tilde{\beta}$	Jackknife estimator
β_i	Regression coefficient of ith independent variable in MLRM
$\hat{\beta}_{-i}$	Least-squares estimator omitting sub-sample i
ε	Error component
$\overline{\varepsilon}$	Mean of residuals
ε_i	Residual value of observation i
θ	Canonical correlation coefficient
μ_U	Vector mean of observations in matrix U
μ_X	Vector mean of observations in matrix X
σ	Standard deviation
σ_ε	Standard deviation of residuals
Σ_{SS}	Covariance matrix for S
Σ_{SY}	Correlation matrix for S and Y
Σ_{UU}	Covariance matrix for U
Σ_{UY}	Correlation matrix for U and Y
Σ_{VV}	Covariance matrix for V
Σ_{XS}	Correlation matrix for X and S
Σ_{XV}	Correlation matrix for X and V
Σ_{XX}	Covariance matrix for X
Σ_{XY}	Correlation matrix for X and Y
Σ_{YY}	Covariance matrix for Y

1. INTRODUCTION

The fundamental concerns of a project manager are to determine how long a project will take, and how much it is going to cost. To accurately estimate the time required to complete a project would resolve a lot of problems related to forecasting, scheduling, bidding and reputation. Even though, there have been studies to control a project according to plan (Thamhain and Wilemon, 1986; Grant *et al.*, 2006), there exists a need to further investigate this matter. A case study conducted by Bounds (1998), stated that only 26% percent of the projects completed in the US were on time and within budget. Frimpong (2000) found that only 25 % of construction projects were within budget and completed on time. Assaf and Al-Hejji (2006) conducted a survey and found that only 30% of the construction projects were completed on time. The delays for the projects ranged from 10 –30%. Moreover, the research of Norris (1971) and of Murmann (1994) pointed out that the unexpected or underestimated cost of projects was between 97-151 % more than the original estimate. It is even more drastic in schedule, running from 41–258 % later than originally estimated. According to Bronikowski (1986) these inaccurate estimates would sometimes lead to the termination of projects resulting in the company incurring huge costs and waste of effort of their resources. If a new product is being launched, time to market is critical. According to Ulrich and Eppinger (2003), missing the target schedule could result in failing to launch the new product with competitors taking control of the market. There are several studies reported in the literature that show that the underestimation of design effort is a major cause of delays and budget cost errors (Colmer *et al.*, 2001; Bashir and Thomson, 2004).

1.2 Design Effort

Design effort is defined as the amount of time, in terms of person hours, required to complete a task or project. Understanding the design effort required in completing a task or project is essential. When an approximation on design effort is the same as the actual effort, within a certain acceptable degree of error, then the ability to schedule, forecast, and conduct a feasibility study will become much easier.

There have been studies on understanding the factors that affect the design effort and lead-time. Zirger and Hartley (1994) developed a list of such factors, which includes market and component complexity, technical uncertainty, and control. Much research was conducted on developing different types of parametric models to understand the product development time and effort. Parametric models including probabilistic models, (Kara *et al.,* 1999) were created to estimate the lead-time. Iteration models were explored to predict iteration times (Smith and Eppinger 1997), while non-linear models were also investigated to predict the lead-time (Mehrez and David, 1999; Jun *et al.*, 2005). Linear models were developed to estimate the design effort (Bashir and Thomson, 1999, 2001, 2004). Design effort is a critical component of the product development lead-time, and little research has been conducted in this area. Therefore, the research presented in this thesis concentrates on estimating design effort.

1.3 Thesis Objective

The objective of this thesis is to estimate the design effort in a product development project. In order to achieve this objective, a parametric model was developed and applied to the design of a compressor fan at Pratt and Whitney Canada

(PWC), a major aerospace company primarily focused on the design, manufacture, and maintenance of engines.

1.4 Organization of Thesis

This thesis is organized as follows. A review of existing work on estimating design effort and tools and concepts such as concurrent engineering and lead-time estimation are described in Chapter 2. Chapter 3 discusses the parametric model used to estimate design effort. It also discusses a multiple linear regression model to be used for design effort estimation, its assumptions, and its suitability. Furthermore, since the data used in this thesis is confidential, data masking was necessary; thus data masking techniques are also reviewed. Chapter 4 discusses the data analysis, results and discussion of the results. Finally, Chapter 5 presents conclusions and limitations of the thesis, and discusses avenues for future research.

2. LITERATURE REVIEW

With the advent of factors such as globalization, time-based competition, and changing consumer tastes, the importance of introducing quality products as quick as possible to the marketplace is emphasized. While market trends are forcing shorter product development times in order to meet their lead-time goals, companies are trying to develop tools and techniques to streamline their product development processes (Bhuiyan, 2001). Several studies have been conducted on reducing lead-time using methodologies such as concurrent engineering (CE) (Winner *et al.*, 1988; Clark and Fujimoto, 1991), and having accurate estimates of lead-time and design effort to find areas of improvement. Winner *et al.* (1988) define CE as the integration of inter-related functions at the outset of the product development process in order to minimize risk and reduce effort downstream in the process, and to better meet customers' needs. CE is a methodology that is highly valued and used in industry; PWC is an example of a company that implements CE practices. Thus, CE needs to be further investigated.

The research of Winner *et al.* (1988) and of Clark and Fujimoto (1991) point out that CE, as compared to the typical sequential engineering (SE) reduces the overall lead-time. SE is the practice of using sequential dependence which refers to a one-way transfer of information between upstream and downstream activities (Bhuiyan, 2001). They conducted case studies to point out how the lead-time is affected under different cases of evolution (slow and fast) as well as varied cases of sensitivity of information (low and high).

Smith and Eppinger (1998) made a study to choose between SE and CE. Their research pointed out that even though CE generally results in reducing lead-time; it might result in an increase in the total amount of engineering (design) effort, development cost and in some cases even lead-time. A two-stage design process was used to compare and contrast both SE and CE, while keeping task interdependencies into account. The model they developed was based on the Work Transformation Matrix (WTM) model of product development, which is an extension of the Design Structure Matrix (DSM), a matrix-based method that is used to determine which activities can be conducted concurrently and which must be conducted sequentially.

Loch and Terwiesch (1998) showed through an analytical model, that when upstream and downstream activities are overlapped (CE), lead-time is minimized. They highlighted that when the downstream activities were conducted based on upstream information, due to communications problems, it would result in downstream rework and iterations or engineering changes (ECs). This would mean that the design effort would increase.

Yassine *et al.* (1999) developed a theoretical framework to evaluate the effect of CE. A stochastic model was developed to evaluate the possibility of iterations or redesign. It was found that applying their model to the design of an automotive cylinder block suggested a potential for reducing the lead-time by 18%. Once again, the use of CE results in more iterations, and though lead-time is reduced, design effort increases.

Joglekar *et al.* (2001) developed a Performance Generation Model (PGM). The model compares and contrasts the sequential, overlapped and CE strategies typically applied. It underlined under which conditions applying CE is the best strategy, in the context of lead-time and not design effort. The main findings they came up were optimal scheduling strategies.

Yassine and Braha (2003) discussed how many companies face challenges while implementing CE and focus on this problem. Having to consider iterations, overlapping, decomposition, and convergence, they suggest a unified and modeling approach based on the design structure matrix (DSM) method. The DSM is a model that allows the exchange of information, which allows decision makers, or managers to understand the relationship of tasks enabling them to better understand and plan for CE initiatives, would result in fewer ECs, which would not drastically impact the effort required.

Bhuiyan *et al.* (2004) used simulation to model the new product development (NPD) process. By developing a stochastic computer model they determined the effect on lead-time of overlapping activities as well as functional relationships under varying uncertainty conditions. They came up with four conclusions. The first conclusion being that whether or not overlapping occurs; increasing functional relationship will lead to a trade-off between development time and effort time. Second, they found that starting early without complete information (early start in the dark) is inferior to a sequential (over the wall) approach even with no functional relationship. Third, if there are some functional relationships then increasing the overlapping of activities (concurrent) will be beneficial in low uncertainty conditions and will be harmful in conditions of high uncertainty. Finally, they found that sequential engineering is better is high uncertainty conditions where as CE is better is low uncertainty conditions. The model was validated against five case studies.

Wang and Yan (2005) focused on determining an optimal overlap (concurrency) between upstream and downstream activities. They developed a model named the design activity group model to model the concurrent product development process. An optimization model was developed and solved, whose objective was to minimize the total costs for design delay and revisions. They consider both the lead-time and design effort in the objective function. Proving the cost function to be convex, they applied a one-dimensional search algorithm to solve the problem.

Bogus *et al.* (2005) have also conducted research on the effect that CE has on lead-time. Similar to the work of Bhuiyan et al. (2004), they determined that the evolution of data would have an impact on sensitivity, which would directly affect the amount of rework required, which would increase the required effort. Their main finding was that the less sensitive the information is, the less risky it will be to commence a downstream activity before the upstream activity has finished, resulting in the reduction of the overall lead-time.

To properly understand the amount of lead-time is another very important factor in product development. The lead-time to complete the project would be the total amount of time required to complete the project. Understanding lead-time is quite essential, since lead-time and design effort are linked together, as pointed out in the studies of Yassine *et al.* (1999), Yassine and Braha (2003), Wang and Yan (2005). Bhuiyan *et al.* (2004) pointed out by increasing functional interaction (increasing information exchange amongst team members) will lead to a trade-off between lead-time and the amount of effort required. Lead-time is related to many aspects such as project management, and if managed properly, would not only reduce the lead-time but also the cost of the project (Phan *et al.* 1988).

Zirger and Hartley (1994) developed a conceptual model to determine the factors that affect lead-time, shown in Table 1. Their main findings were that the first three factors would increase the lead-time as they were increased (+). The remaining seven factors would reduce the lead-time as they were increased (-). They pointed out that a limitation of this model would be that it can only be used for a particular or similar industry, and thus cannot be widely used.

Table 1: Zirger and Hartley's (1994) findings

Intermediaries	Lead- time
Market and technical uncertainty	+
Coordinative complexity	+
Component complexity	+
Extent of information sharing	-
Timeliness of information	-
Speed of decision-making	-
Explicitness of goals	-
Goal congruence	-
Linking rewards to goals	-
Control	-

Smith and Eppinger (1997) developed a sequential iteration model to predict the lead-time of an iterative design process. The work could be achieved by using the DSM, while assuming that each design activity had a deterministic duration with a probability of repetition.

Kara *et al.* (1999) came up with a two-stage probabilistic model to estimate the lead-time of a project. They included uncertainty in their model including the simulation of randomness in the process. The first stage was to break down a product into its tasks to find the precedence relationships. These relationships are then modeled in the second stage. The model while estimating completion time emphasizes on the precedence relationship effects, which is real-life CE projects, are complex and hard to understand.

Mehrez and David (1999) developed a two-phase simple non-linear model to determine the lead-time and timing of research and development (R&D) projects. The paper illustrated an example that shows that solutions of both concurrently completing and sequentially completing tasks can occur in different scenarios.

Jun *et al.* (2005) developed a model to estimate the lead-time in design while accounting for uncertainties, iterations and evolution of the design process. The research illustrates the concept of rework and redesign will occur when tasks are completed concurrently. The model was validated against a case study to show how the development process will progress over time and how their model effectively models lead-time. They also determined that rework is a function of the sensitivity of information. The main findings were that they highlighted how the lead- time progresses over time and their model, which was validated with a case study, can accurately predict the lead- time of products taking concurrency and risk into account. They pointed out that a limitation of this model would be that it can only be used for a particular or similar industry, and thus cannot be widely used.

Design effort in the context of product development is essential to understand. Having an approximation of design effort, within a certain acceptable degree of error, will facilitate the ability to schedule, forecast, and conduct a feasibility study. Understanding the cost of design effort in product development would alleviate many of the management decisions that are made under conditions of uncertainty. While much attention has been given to determining overall effort for a product development process, estimating design effort has been given limited attention in the literature.

Colmer *et al.* (1999) studied design effort and costing improvement issues in a new product development project. There paper presented a discussion on why these new projects are vulnerable to cost overruns. They discussed how the limitations of cost estimates lead to poor decision making, which contributes to cost and schedule overruns. They discussed some issues regarding project and risk management, which will help projects meet their own needs and the need of the customers.

Bashir and Thomson (1999) developed a model to estimate design effort based on the product complexity (PC). PC was assumed to be a metric that depends on the number of functions and the depth of their functional trees. Based on this assumption the formula used to calculate PC is presented below, and an example is shown in Figure 1.

$$PC = \sum_{j=1}^{l} F_j j \tag{2.1}$$

where

F_j, Number of functions at level j

l, Number of levels

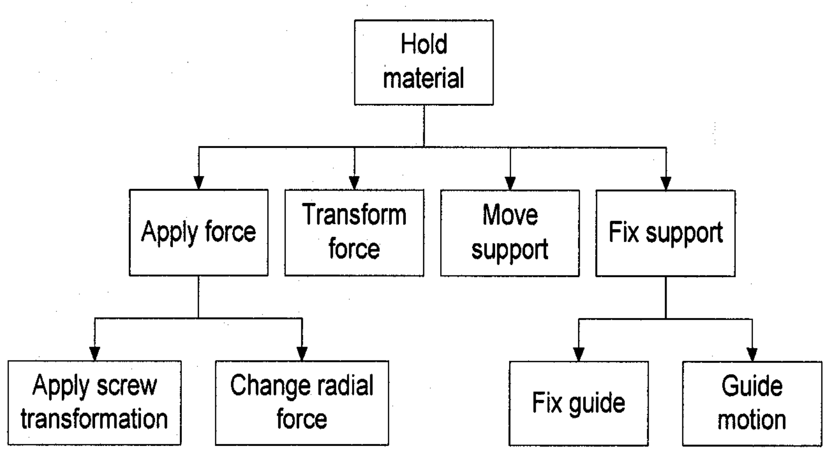

Figure 1: PC hierarchy

From Figure 1, the PC would be:

$$PC = (1 \times 1) + (2 \times 4) + (3 \times 4) = 21$$

They conducted a study of two companies and estimated the design effort, $\hat{E}$ for various projects. They determined that the impact of PC on design effort was that as the

PC went up, the required design effort also increased. The resulting function was $\hat{E} = aPC^b$, where a and b were constants determined from the data.

Bashir and Thomson (2001) built on their previous work to include another company specific factor, namely severity of requirements (SR). The SR values with discussion with the company's management were determined as follows:

- Design requirements were not too difficult to meet, SR = 1
- Design requirements were difficult to meet, SR = 2
- Design requirements were extremely difficult to meet, SR = 3

Based on these attributes, values for SR were assigned to each of the projects. They determined that the multivariate model (including SR) performed better than the univariate model (considering only PC). The multivariate model had a mean magnitude square error (MMSE) of 15%, versus the univariate model having a mean magnitude relative error (MMRE) of 30%. They applied the model by Norden (1964) to estimate the project lead-time. His model to estimate duration is as follows:

$$y' = 2\hat{E}\alpha t e^{-\alpha t^2} \tag{2.2}$$

where

y', Manpower utilized in each time period (t)

$\hat{E}$, Total design effort in hours

α, Shape parameter

Bashir and Thomson (2004) developed a parametric model to estimate the time required to design hydroelectric generators for General Electrics (GE). They analyzed fifteen designs made between the years 1985-1989. They applied an already existing equation given as equation 2.3. This equation was validated by empirical evidence supporting it with the work of Walston and Felix (1977), Boehm (1981), and Jeffrey (1987).

$$\hat{\mathrm{E}} = \mathrm{aPC^b D_1^{c_1} D_2^{c_2} \cdots D_m^{c_n}} \tag{2.3}$$

where

$\hat{E}$, Estimated design effort in hours

PC, Product complexity

D_m, Effort driver (factor m)

a, b, c_n, Constants (weights) that are estimated from historical data

The factors selected were the following:

- Product complexity (PC)
- Difficulty to expertise (DE)
- Type of drawing to be submitted to customer (TD)
- Involvement of design parameters

They applied the jackknife estimator to calculate the values of the constants and the resulting equation can be seen below:

$$\hat{E} = 0.12PC^2 DE^{0.41} TD^{0.35} DP^{0.69} \tag{2.4}$$

Having applied their equation to 15 jobs, they found that their model was a better tool for estimation for the company reducing the MMRE of the prediction from 27% to 13%.

2.1 Summary

In this day and age, much effort is being focused on product development. There are various tools and techniques being used to quantify and reduce the development time. Amongst these tools, CE involves the simultaneous execution of upstream and downstream activities, in an effort to reduce the lead-time. However, studies have shown that while lead-time is reduced, this typically involves an increase in effort. Being able to estimate the design effort in a product development project could help to schedule, allocate resources, and understand the cost.

While existing work done on design effort estimation shows promising results, there has been limited research conducted in this area. Since design is a critical part of product development, being able to estimate its effort can add value to the product development. A parametric approach that uses a multivariate linear regression model is discussed in the following chapter.

3. PARAMETRIC MODEL OF DESIGN EFFORT ESTIMATION

The following chapter discusses three main aspects. The first section discusses the component being designed at PWC, as well as the departments being analyzed. The second section describes the design effort estimation model. It will also discuss the jackknife technique, which is used to eliminate the biases due to small sample sizes, as in this case. Finally, the last section discusses the data masking technique to be used to mask the confidential data provided by PWC.

3.1 Compressor Fan Design

The estimation model is applied to the design of a component in the compressor module of a certain class of turbo fan engines. The compressor module, sucks in the air and sends it through a compression stage in which, in some cases, the air is compressed up to a factor of 30 (PilotFriend, 2000). The compressed air is sent to the combustion chamber, where it is mixed with fuel and burned. Then, it passes into the turbines and its heat is converted into thrust, allowing the aircraft to move. Figure 2 below shows a typical engine highlighting the compressor and turbine model, consisting of the low and high pressure compressor and turbine.

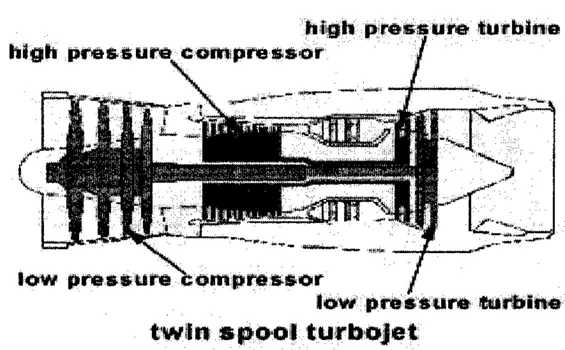

Figure 2: Compressor and turbine module (PilotFriend, 2000)

The component that is being studied is the integrated blade-rotor low-pressure compressor (IBR LPC) fan, which will be abbreviated to compressor fan (CF). The CF is responsible for guiding the air into the compressor module. A typical compressor fan is shown in Figure 3.

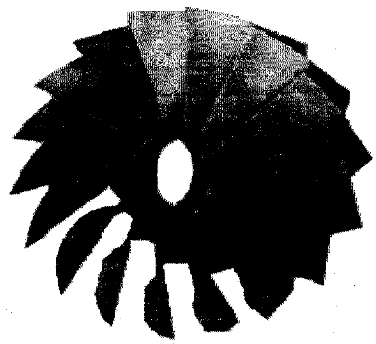

Figure 3: Compressor fan (Air Force Research Laboratory, 2006)

The design of the compressor fan is a complex process, which involves expertise from several departments. Even though there are several departments involved, there are four key departments that together carry out the great majority of the work for the design of the CF. Thus, parametric models are developed for these four departments.

3.1.1 Compressor design departments

The four departments for which parametric models are developed are the following:

1. Compressor design department
2. Compressor aerodynamics department
3. Compressor structures (analytical) department
4. Compressor drafting department

The design department is responsible for the overall design of the component. This department ensures that all components mesh well with the rest of the engine. They use

CAD/CAM software (CATIA) for designing the components. The aerodynamics department is responsible for the profile of the blades. They run some types of analysis to make sure that the airflow follows the path most favourable to the performance of the engine, while maintaining a high level of safety. The analytical department takes input from the previous two departments and is responsible for ensuring that the part is durable and is able to pass all the various tests, such as stress analysis, foreign object damage, blade off, etc. Finally, the drafting department is responsible for developing detailed drawings for the component, while ensuring that the design conforms to manufacturing standards.

3.2 Parametric Model

As mentioned earlier, Bashir and Thomson (2004) suggested a parametric model built upon using the data for design time recorded for hydro projects at GE. This research presents a modified version of this model. While the earlier model uses product complexity (PC) as a driver of effort, it is omitted in our work since our model is applied to a specific type of component in the same engine family for PWC; hence PC is not a variable. The proposed design effort estimation equation is shown below:

$$\hat{E} = a_o D_1^{a_1} D_2^{a_2} \cdots D_m^{a_m} \tag{3.1}$$

It should be noted that Bashir and Thomson (2004) estimated the overall effort of the projects. Even though their findings highlighted many things for the companies involved, it was felt that having more granularity would be more useful from a resource allocation standpoint. In this regard, one estimation model was created and applied to each of the four above-mentioned key departments involved in designing the compressor fan.

3.2.1 Factors selected for the model for PWC

The design effort estimation model was developed using data provided on seven specific design jobs (DJs) by PWC for the IBR LPC fan for a certain class of turbo fan engines. It was essential to determine the principal factors that may have significant effect on the design effort estimation. After extensive interviews and discussions with managers, designers, and project engineers at PWC, the following four factors were selected as the effort drivers to be used for this model:

1. Type of design (TD)
2. Degree of change (DC)
3. Concurrency (Con)
4. Experience of departmental personnel (DE)

3.2.1.1 Type of design

When designing a component, the effort will naturally depend on the type of design (TD). An initial design will likely not require the same amount of time as a redesign. In this regard, each design was assigned one of the following attributes.

- Initial Design: 1
- Redesign: 2

3.2.1.2 Degree of change

The next factor considered was the degree of change (DC). This factor is used to attribute a value to the level of rework created from the initial design to a redesign, or from a redesign to a second redesign. If there is a major change to the initial design, the amount of rework generated would be expected to be greater than if were a minor change.

In fact, in some cases, a major change resulted in more effort than the initial design. Thus, different values were attributed to the designs as shown below:

- Initial design: 1
- Redesign with minor modifications: 2
- Redesign with major modifications: 3

3.2.1.3 Concurrency

At PWC, product development teams practice concurrent engineering (CE). It has been shown by various researchers such as Loch and Terwiesch (1998), Yassine et al. (1999), Joglekar et al. (2001), Yassine and Braha (2003), Bhuiyan et al. (2004) that CE reduces the overall lead-time to design components, However, it is important to understand the level of concurrency involved in the design, because it has also been shown by several researchers such as Bhuiyan et al. (2004), Wang and Yan (2005), and Jun et al. (2005) that CE also results in rework which will add to the amount of effort required to design the component. Thus concurrency of activities was considered an important factor to include in the model. The concurrency values of each department for each of the i design jobs (A to G) are calculated. To do this, a concurrency matrix was first created as shown in, for each of the j periods (nine in this case). A weight from zero to one was given for the concurrency value, in a given period (c_j), that the compressor design department (D1) had with the other departments (D2-D9), designing the component. The net concurrency value for i^{th} DJ is calculated using the following equation:

$$Con_i = \frac{\sum_{j=1}^{n} c_j}{\text{Number of Periods}} \qquad (3.2)$$

Table 2: Concurrency matrix for DJ A

Dept/Period	1	2	3	4	5	6	7	8	9
D1	1	1	1	1	1	1	1	1	1
D2	1	1	1	1	1	1	1	1	1
D3	1	1	1	1	1	1	1	1	1
D4	0	0	1	1	1	1	1	1	1
D5	0	0	0	1	1	1	1	0	1
D6	0	0	0	0	1	1	1	0	1
D7	0	0	1	1	1	0	1	1	0
D8	0	1	1	0	0	0	1	1	1
D9	1	1	1	1	1	1	1	1	1
CI	0.33	0.50	0.75	0.75	0.88	0.75	1.00	0.75	0.88

As can be seen from the Table 2, the net concurrency value for DJ A would be:

$$\text{Con}_{A} = \frac{0.33 + 0.50 + 0.75 + 0.88 + 0.75 + 1.00 + 0.75 + 0.88}{9} = 0.73$$

This shows that there is a high level of concurrency (73%) with DJ A and all the other DJ's. The net concurrency values of all of the seven DJs for D1 are shown in Table 3.

Table 3: Concurrency values for D1

DJ	Con
A	0.73
B	0.65
C	0.66
D	0.83
E	0.69
F	0.71
G	0.70

3.2.1.4 Experience of department personnel

The experience of the person working on the job plays a major role when determining the amount of time s/he will require to complete the task. Naturally, a person having several years of experience working on the design or analysis of the component will be expected to complete the task more quickly then a person with very little or no experience at all. The attributes for design experience (DE) were assigned for different ranges of years of experience as can be seen below:

- 0-2 years of experience: 1
- 3-4 years of experience: 2
- 5 + years of experience: 3

It is also assumed that if there is more than one person in a department having different levels of experience, then the experience level for a particular job will be calculated from the weighted average of experience from all the n persons working on the design job. This can be seen from the following equation:

$$DE = \sum_{i=1}^{n} (\%\text{ hours of i})*(\text{experience of i}) \tag{3.3}$$

In the case study carried out at PWC, seven observations of DE contain the actual design effort available for the IBR LPC fan for the analysis as well as a quantitative measures of the aforementioned factors; TD, DC, and DE. Table 4 shows the values collected for the seven DJs for D1.

Table 4: Values for D1

DJ	TD	DC	Con	DE
A	1	1	0.73	3.00
B	2	2	0.65	3.00
C	2	3	0.66	3.00
D	1	1	0.83	3.00
E	2	2	0.69	2.25
F	2	3	0.71	2.65
G	1	1	0.70	2.44

3.3 Linear Regression

Linear regression (LR) will be used to solve this problem of estimating design effort. Its suitability will be explained shortly. The simple LR model can be denoted as follows.

$$y = \beta_0 + \beta_1 X_1 + \varepsilon \tag{3.4}$$

where

y, Response

β_0, Intercept

β_1, Slope

ε, Error

In the case, as in the problem under study, when y depends on several variables, the multiple linear regression model (MLRM) will be used. The MLRM looks as follows:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon \tag{3.5}$$

where

y, Response dependant on k predictor values

β_j , Regression coefficients

ε, Error

In order to calculate the regression coefficients, the method of least squares was used.

3.3.1 Least Squares Estimation for MLRM

The least squares estimation generates an equation that will minimize the sum of the square errors (Kutner *et al.*, 2004). The coefficients $\hat{\beta}$ will be calculated from the following equation. These values are readily available from statistical software used to analyze the data, in this case Microsoft Excel.

$$\hat{\beta} = (X'X)^{-1} X'y \tag{3.6}$$

3.3.2 MLRM Assumptions

In order to use the MLRM the following two assumptions need to be satisfied.

1. Linearity assumption

2. Normality assumption

3.3.2.1 Linearity assumption

In order to determine if the function is linear for a given case of the MLRM, scatter plots or residual plots can be made. In the case of the scattered plot, the standardized residual could be plotted against the non-standardized predicted value. For a linear function, the scatter plot should not have any curvilinear patterns.

Another graphical manner in statistically proving the validity of the MLRM is by creating statistical process control (SPC) charts. In the context of this project, the predicted values against the actual values of design effort will be plotted. The errors or residuals will be the deviation from the line ln (Predicted) = ln (Actual) + ε. The expected error, assuming a normal distribution is zero, i.e. E (ε) =0. Thus, the mean of the function, f(x) will simply be f(x). The equations for the upper control limits (UCL) and lower control limits (LCL) are shown below.

$$UCL = f(x) + 3\sigma \tag{3.7}$$

$$LCL = f(x) - 3\sigma \tag{3.8}$$

where

f(x), ln (Actual) = ln (Predicted) + ε

σ, Standard deviation of the residuals

According to Montgomery (1985), if the residuals are within 3σ of the expected value of the function, then the function is considered to be statistically in control. In other words, the assumption of linearity will hold statistically.

3.3.2.2 Normality assumption

The third assumption that must be validated is that the error values are following a normal behaviour. The test to determine error normality requires the coefficient of correlation, r. The value of r is calculated from the following equation.

$$r = \pm\sqrt{\mathrm{R}^2} \tag{3.9}$$

where

R^2, Coefficient of determination

The value of R is calculated from the following equation.

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SSTO}} \tag{3.10}$$

where

SSR, Regression sum of squares

SSTO, Total sum of squares

This will also involve a hypothesis test in which the critical values, r_L prepared by Looney and Gulledge (1985) are to be compared against the calculated r. The H_o assumes the error has a normal behaviour. The H_1 assumes the error not to have a normal behaviour. The outcome of the test would be as follows:

$$\text{If } \mathrm{r} \geq r_L(1-\alpha), \text{conclude} \mathrm{H}_0$$

$$\text{If } r < r_L(1-\alpha), \text{conclude} \mathrm{H}_1$$

3.3.3 Suitability of the MLRM

In order to use the MLRM it is essential that the equation which will be analyzed and estimated to be in the form of an MLRM. For that reason, let us recall Eq. 3.1 that will be used to estimate the design effort:

$$\hat{E} = a_o D_1^{a_1} D_2^{a_2} \cdots D_m^{a_m}$$

This equation, in its original form, is not in the form of an MLRM. However, if the natural log (ln) is taken of both sides the equation will have the following form:

$$\ln\hat{E} = \ln(a_0) + a_1 \ln(D_1) + a_2 \ln(D_2) + \cdots + a_m \ln(D_m)$$

If this equation were to represent the specific parameters of our problem, it would yield:

$$\ln\hat{E} = \ln a_0 + a_1 \ln TD + a_2 \ln DC + a_3 \ln Con + a_4 \ln DE \quad (3.11)$$

If we recall equation 3.5 we can see that the modified equation has the form of an MLRM.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

Having this form, we can use it to estimate its parameters using the least squares estimator.

3.4 Jackknife Technique

The jackknife technique was used to determine the regression coefficients of each of the model parameters. This technique was originally a computer-based method for estimating biases and the standard errors. According to Efron and Tibshirani (1993), this technique is commonly used not only to improve the problem of biased estimation due to small sample size, but also in situations where the distribution of the data is hard to analyze. In this technique, the data are divided into sub-samples, and the sub-samples are obtained by deleting one observation at a time. The calculations are carried out for each sub sample. Given a data set $x = (x_1, x_2, x_3, \ldots, x_n)$, the i^{th} jackknife sample x_i is defined to be x with the i^{th} data point removed. The pseudo-values, Ps_i, are determined using the following equation:

$$Ps_i = ns\hat{\beta} - (ns-1)\hat{\beta}_{-i} \tag{3.12}$$

where

Ps_i, Pseudo-value for the entire sample, omitting sub-sample i.

ns, Number of sub-samples

$\hat{\beta}$, Least-squares estimator of the whole sample

$\hat{\beta}_{-i}$, Least-squares estimator for the entire sample, omitting sub sample i

The jackknife estimator $\tilde{\beta}$ is determined as follows:

$$\tilde{\beta} = \frac{\sum_{i=1}^{ns} Ps_i}{ns} \tag{3.13}$$

The technique is used to estimate regression coefficients in a regression analysis. The regression analysis is used for parametric modeling of design effort. In the proposed regression based analysis, the regression coefficients of each of the factors are jackknifed, by creating sub samples by deleting one observation at a time.

In the case study carried out at PWC, seven observations were available for each of the four departments considered for the analysis. With each department, the jackknife technique is used to determine the regression coefficients of the each the parameters consider in the regression based parametric modeling. As mentioned earlier, the GADP model of Muralidhar *et al.* (1999) was used to mask the confidential data.

3.5 Data Masking

The data provided by PWC in the application of the design effort estimation model is confidential and proprietary, thus it is essential to properly mask the data in such a manner that the findings will also be similar when applied to the original unmasked data. In this section, a review of various data masking techniques is presented.

Muralidhar *et al.* (1999) conducted research on data perturbation (DP), also known as data masking (DM). They briefly outlined the characteristics required in order for DP to be effective. When using DP there are four different types of bias that can occur. They are as follows.

- Type A: Change in summary measures of individual confidential attributes due to a change in variance.
- Type B: Relationships between the confidential attributes are not the same with the original data (OD) compared to the masked data (MD).
- Type C: Relationships between confidential and non-confidential variables are not the same of the OD compared to the MD.
- Type D: The distributions of the confidential attributes are not the same of the OD and the MD.

They pointed out that theses biases should not occur with the addition of the following two security requirements for meaningful DP:

- The single attribute security (SAS) provided should be at least 1.00 for each individual attribute.
- The security provided for linear combinations (SPLC) should be at least 0.50. DM techniques are used in order to protect confidential data while preserving the characteristics of the data.

They discussed the current DP techniques that already exist. These are Simple Additive Data Perturbation (SADP), Correlation-Noise Additive Data Perturbation

(CADP), Bias-Corrected Correlated-Noise Additive Data Perturbation (BCADP), Multiplicative Data Perturbation (MDP), and their new proposed method, General Additive Data Perturbation (GADP). A summary of the techniques meeting the requirements in a multivariate normal distribution can be seen in Table 5.

Table 5: Muralidhar et al.'s findings

DM method	Type A	Type B	Type C	Type D	SAS ≥ 1	SPLC ≥ 0.5
SADP	YES	YES	YES	NO	YES	NO
CADP	YES	NO	YES	NO	YES	NO
BCADP	NO	NO	YES	NO	NO	NO
MDP	YES	YES	YES	YES	YES	NO
GADP	NO	NO	NO	NO	YES	YES

As can be seen the GADP method what they proposed would be the best since it is free from all the biases and it meets the security requirements. After discussing the methodology and how to implement their model, they pointed out the limitations of their work by conducting a comparative study of non-normal data using both the MDP and GADP techniques. They found when non-normal data is present; the results between both methods would be comparable, indicating that the MD generated by the GADP would also have bias.

Sarathy *et al.* (2002) used the work of Muralidhar et al. (1999) to further this research. They pointed out the limitation that the previous model does not preserve the non-linear relationships in non-normal data, hence, providing "biased" data. Their objective was to develop a new method of DP for non-normal data that would provide more meaningful MD, while ensuring that the security requirements are upheld. They coupled the existing GADP technique, and leveraged the copula approach based on Abe Sklar's theorem (1959) to overcome this obstacle. The copula approach takes into account the effect of the marginal distributions and the dependence between them. With the GADP and the copula approach they developed a new model, named the copula-based GADP (C-GADP). The model was validated with the use of a bank database. The C-GADP and GADP were compared to the original data in the case when responses to queries involving multiples variables were recorded. Table 6 presents the results on responses of queries involving multi variables highlights the superiority of the new C-GADP method as compared to the GADP method.

Table 6: Sarathy *et al.*'s comparison of DP techniques for queries

	Response from		
Query	Original	GADP	C-GADP
What is the average balance in savings/checking for individuals whose home equity is greater than $150K	$27,000	$27,000	$27,000
What is the average value of stocks/bonds for individuals whose home equity is greater than $150K	$80,000	$67,000	$79,000
What is the average liabilities for individuals whose home equity is greater than $150K	$145,000	$123,000	$146,000

It is concluded by pointing out the limitations of their model. The model is limited by Sklar's theorem assuming that the dependency of the attributes can be identified by rank order correlations. However, they point out that rank order correlation is only able to capture the monotonic relationships between attributes and not the non-monotonic ones. Monotonic relationships are relationships that preserve the given order.

Muralidhar and Sarathy (2006) developed a new DM procedure, named data shuffling. A heuristic to implement the data shuffling presented, which was to be based on the perturbation method the approach Sarathy *et al.* (2002) developed in which the data was to be swapped. The research highlighted that the method provides a nonparametric method for DM and established the following three advantages that are non-existing in other data swapping techniques. First, the marginal distribution of the original data is preserved in the masked data. Second, the pair-wise monotonic relationships are upheld in the masked data. Third, access to the masked data does not increase the risk of disclosure. As is the case of the research of Sarathy *et al.* (2002), this model is also limited to data having monotonic relationships. Data with complex (non-monotonic) relationships will not maintain all the desired attributes of the original data. Another limitation of their research is that it would be difficult to implement their procedure with a dynamic (changing) database.

As can be seen from the Table 5, the GADP method is the most suitable of the methods that are used when linear combination is involved. It is free from the biases and it preserves the characteristics of the original data. Thus, the GADP method will be used and is discussed in the following sub-section

3.5.1 Review of the GADP model

Muralidhar *et al.* (1999) explained a 5-step process, which will be used below, on how to mask data.

Step 0:

In order to mask data, the values for the non-confidential and confidential terms are required. Assuming there are sample sizes of n, let matrix X represent the numerical values for the p confidential attributes, and let matrix S represent the numerical values of the q non-confidential attributes, both having a multivariate normal distribution. Matrix U as seen below will represent the joint matrices of X and Y, that is U = {X,S}.

$$U = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} & s_{11} & \cdots & s_{1q} \\ x_{21} & x_{22} & \cdots & x_{2p} & s_{21} & \cdots & s_{2q} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} & s_{n1} & \cdots & s_{n2} \end{bmatrix}$$

Step 1:

Calculate the mean and covariance matrices for U. The means vector of U, μ_U will be as follows:

$$\mu_U = [\mu_X \ \mu_S]$$

The covariance matrix of U, Σ_{UU} will be represented as:

$$\Sigma_{UU} = \begin{bmatrix} \Sigma_{XX} & \\ \Sigma_{XS} & \Sigma_{SS} \end{bmatrix}$$

where

Σ_{XX}, Covariance matrix of X

Σ_{XS}, Covariancematrix of (X and S)

Σ_{SS}, Covariance matrix of S

$$\Sigma_{UY} = [\Sigma_{XY} \ \Sigma_{SY}]$$

If Y represents the p perturbed attributes, then the joint covariance matrix of X, S, and Y will be denoted as Σ^G as is shown in matrix form below:

$$\Sigma^G = \begin{bmatrix} \Sigma_{XX} & & \\ \Sigma_{XS} & \Sigma_{SS} & \\ \Sigma_{XY} & \Sigma_{SY} & \Sigma_{YY} \end{bmatrix}$$

Step 2:

Choose the covariance matrices $\Sigma_{XY}, \Sigma_{SY,}$ *and* Σ_{YY} for security and bias considerations. The calculation of Σ_{XY} according to Parsa *et al.* (1998) to achieve a maximum-security level is as follows:

$$\Sigma_{XY} = \theta^2 \, \Sigma_{XX} \tag{3.14}$$

where

θ, Canonical correlation

The canonical correlation θ is equal to largest root of the eigen-value from the following matrix.

$$\Sigma_{XX}^{-1}\,\Sigma_{XV}\Sigma_{VV}^{-1}\Sigma_{VX}$$

where

V = {S, Y}

$$\Sigma_{VV} = \begin{bmatrix} \Sigma_{SS} & \Sigma_{SY} \\ \Sigma_{YS} & \Sigma_{YY} \end{bmatrix}$$

Therefore, the term used for the calculation, θ^2 will simply be the largest eigen-value from the matrix mentioned above.

Step 3:

Given U= c_i, calculate the expected value and variance.

If we let U= c_i, it would represent the vector in the i^{th} observation set from U.

$$c_i = \begin{bmatrix} x_{i1} & x_{i2} & \cdots & x_{ip} & s_{i1} & \cdots & s_{iq} \end{bmatrix}$$

From Graybill (1976) the distribution of Y|U = c_i has the following expected value and variance.

$$E(Y \mid \mathrm{U} = \mathrm{c_i}) = \mu_X + (\Sigma_{\mathrm{YU}})(\Sigma_{\mathrm{UU}})^{-1}(ci - \mu_U) \qquad (3.15)$$

$$Var(Y \mid \mathrm{U} = \mathrm{c_i}) = \Sigma_{\mathrm{YY}} - (\Sigma_{\mathrm{YU}})(\Sigma_{\mathrm{UU}})^{-1}\,(\Sigma_{\mathrm{UY}}) \qquad (3.16)$$

Step 4:

In order to mask the value, one must generate a random observation from a multivariate normal distribution, with the expected value (mean) and variance from equations 3.15 and 3.16, respectively. This can be accomplished by using the random number generator functions for multivariate normal distributions with mathematical software packages such as MATLAB or MAPLE.

Step 5:

Repeat steps 3 and 4 for every i^{th} observation, $\forall$ i = 1, 2, ..., n.

3.5.2 Example of DM using the GADP Model

In this section, a fictitious example is presented where a bank has a database containing information about 30 customers. Some of the information is confidential and some is not. Table 7 presents the bank's database.

Table 7: Bank's database

Customer	Home Equity	Stocks/ Bonds	Liabilities	Savings/ Checking	CDs, etc.
1	142.68	70.07	122.30	28.59	70.59
2	99.94	47.96	80.50	25.46	53.14
3	98.21	39.20	63.52	15.87	31.50
4	94.90	51.51	78.49	26.80	50.80
5	82.52	51.68	59.93	20.54	57.10
6	108.46	45.50	101.73	16.56	47.27
7	97.33	28.76	60.96	25.57	61.78
8	110.79	60.90	86.06	25.53	60.74
9	117.50	56.67	83.75	22.38	67.82
10	74.98	27.91	59.67	13.76	49.45
11	117.37	51.70	95.45	18.66	30.32
12	83.90	52.72	82.91	13.36	52.49
13	84.95	40.62	76.44	19.12	60.76
14	85.08	42.89	65.53	18.18	44.31
15	93.81	47.87	91.71	5.48	27.55
16	69.56	39.07	57.27	15.31	52.58
17	116.53	52.18	99.18	22.95	50.24
18	87.74	37.44	45.49	23.79	51.63
19	119.19	61.79	101.01	21.70	62.83
20	139.46	69.01	110.24	23.61	55.93
21	105.90	58.26	81.46	23.66	68.76
22	92.15	64.62	63.73	21.00	42.79
23	111.52	54.82	70.33	26.64	53.03
24	77.17	42.20	61.32	9.37	50.54
25	101.22	57.36	85.21	22.73	47.64
26	100.25	50.05	67.37	28.33	48.58
27	96.64	58.89	76.70	24.20	48.13
28	86.25	57.74	88.54	19.65	50.20
29	80.19	49.39	66.89	14.85	46.16
30	99.00	57.80	84.56	18.21	48.54

In this case, Home Equity (HE), Stocks/Bonds (S/B), and Liabilities (L) are considered to be confidential attributes. The Savings/Checking (S/C) and CDs, etc. (CD) are not considered to be confidential.

Step 1:

The resulting mean matrix μ_U will be as follows.

$$\mu_U = [101.27\ 48.49\ 79.10\ 20.59\ 49.04]$$

The covariance matrix, Σ_{UU} of this data set is shown below.

$$\Sigma_{UU} = \begin{bmatrix} 440.63 & 142.68 & 362.66 & 68.03 & 114.09 \\ 142.68 & 86.44 & 170.54 & 19.19 & 34.41 \\ 362.66 & 170.54 & 469.22 & 42.70 & 75.79 \\ 68.03 & 19.19 & 42.70 & 20.36 & 28.63 \\ 114.09 & 34.41 & 75.79 & 28.63 & 98.16 \end{bmatrix}$$

Step 2:

The covariance matrices, Σ_{YY}, Σ_{YS}, and Σ_{YX} are calculated and shown below.

$$\Sigma_{YY} = \begin{bmatrix} 440.63 & 142.68 & 362.66 \\ 142.68 & 86.44 & 170.54 \\ 362.66 & 170.54 & 469.22 \end{bmatrix}$$

$$\Sigma_{YS} = \begin{bmatrix} 42.70 & 75.79 \\ 20.36 & 28.63 \\ 28.63 & 98.16 \end{bmatrix}$$

$$\Sigma_{YX} = 0.35 * \Sigma_{XX} = \begin{bmatrix} 154.22 & 49.94 & 126.93 \\ 49.94 & 30.25 & 59.69 \\ 126.93 & 59.69 & 164.23 \end{bmatrix}$$

The value of 0.35 is the largest eigen-value of the matrix multiplication shown in step 2 of section 3.5.1.

Step 3:

The expected value and variance is to be calculated as shown from equation 3.15 and equation 3.16 respectively. This will be done for every customer, c. However, the sample calculation for the expected value and variance matrix will be shown given the c value is of the last customer (i.e. = c_{30}).

$$c_{30} = \begin{bmatrix} 108.82 & 47.81 & 62.66 & 18.84 & 49.17 \end{bmatrix}$$

$$E(Y \mid U = c_{i}|_{i=30}) = \begin{bmatrix} 93.50 \\ 40.64 \\ 66.23 \end{bmatrix}$$

$$Var(Y \mid U = c_{i}|_{i=30}) = \begin{bmatrix} 341.62 & 95.24 & 273.00 \\ 95.24 & 54.52 & 126.40 \\ 273.00 & 126.40 & 332.69 \end{bmatrix}$$

Step 4:

The random number generated for customer 30, can be seen below.

$$(Y \mid U = c_{i}|_{i=30}) = \begin{bmatrix} 50.02 & 31.00 & 30.63 \end{bmatrix}$$

Step 5:

Steps 3 and 4 were repeated for every customer (c_1 to c_{30}). However, the calculation of every customer is not shown, rather it is summarized in

Table 8. It shows the masked values of the confidential attributes; HE, S/B, and L. The non-confidential attributes S/C and CD are not masked and are presented in their original form.

Table 8: Masked values of bank's database

Customer	Home Equity	Stocks/ Bonds	Liabilities	Savings/ Checking	CDs, etc.
1	140.44	61.86	103.65	25.31	52.71
2	121.55	59.05	104.99	25.18	62.73
3	116.60	46.65	74.97	24.43	54.89
4	128.98	61.23	127.03	21.89	62.35
5	115.98	52.27	89.41	22.71	47.87
6	124.00	63.76	100.14	27.74	56.16
7	124.17	62.77	99.17	19.50	35.43
8	75.01	30.09	36.34	19.46	41.36
9	80.98	35.10	78.40	15.41	54.70
10	109.94	56.29	93.37	30.11	65.39
11	73.52	45.28	67.29	19.33	45.70
12	120.22	52.55	98.87	22.51	55.74
13	98.96	49.95	82.69	19.47	45.29
14	82.94	39.05	74.05	16.70	50.41
15	98.75	38.72	48.74	20.02	36.16
16	89.56	43.74	88.79	17.14	55.89
17	133.09	61.90	106.40	23.21	54.52
18	121.29	60.79	102.66	26.17	62.45
19	98.90	44.42	63.75	19.64	44.00
20	108.71	51.00	58.33	27.56	48.76
21	88.28	41.20	58.67	12.10	29.45
22	101.10	42.20	62.34	14.53	28.96
23	95.78	47.91	77.96	22.03	53.41
24	71.92	45.09	70.56	14.70	31.82
25	90.93	50.03	84.98	16.90	54.97
26	112.92	41.66	70.78	13.88	37.70
27	112.12	51.34	102.71	17.02	56.18
28	115.84	67.73	107.55	20.09	49.04
29	94.26	61.62	91.26	24.25	47.96
30	50.02	31.00	30.63	18.84	49.17

3.5.2.1 Discussion of the bank example

Even though the data has been masked, it is important to ensure that the characteristics of the masked data mimic the characteristics of the original data, i.e., if the data masking was effective. For this reason, to show the effectiveness of this technique, comparing a linear combination of these terms will be conducted. The correlation of Net worth (NW) to Liquid assets (LA) will be compared in both cases. The calculation of NW and LA for this example is shown below:

$$NW = HE + S/B - L + S/C + CD \qquad (3.17)$$

$$LA = S/B + S/C + CD \qquad (3.18)$$

The correlation calculated for the original database of NW to LA was equal to 0.84. The correlation of NW to LA in the masked database was equal to 0.76. There is less than a 10% difference in the correlations of both databases. Having a larger sample size will decrease the difference as Muralidhar et al. (1999) presented a similar example having 10,000 customers and the correlations of NW to LA were the same with the original data base and the masked one.

It should be noted that the security measures will not be discussed, since the application of this method is not to determine how secure it will be from hackers, rather to show that the masked values mimic the behaviour of the unmasked values.

3.6 Summary

A parametric modelling approach similar to the one of Bashir and Thomson (2004) is suggested to estimate the design effort of the compressor fan for four departments at PWC. The factors considered in the model are type of design, degree of change, concurrency, and experience of departmental personnel. In order to solve the problem, the model is transformed into a multivariate linear regression model. The assumptions made and its suitability is discussed.

The research was conducted at PWC and the sensitive data provided by the company needed to be concealed. To achieve this, various data masking techniques are explored and the GADP model was selected. The technique is explained in detail, and an example is presented showing how the technique is able to maintain the characteristics of the original data. The following chapter will present the analysis of the departments studied.

4. DATA ANALYSIS

The following chapter presents the analysis of the four departments studied at PWC. The statistical tests are presented showing the linearity and normal assumptions being fulfilled. The estimation of design effort, $\hat{E}$ is shown and the residual and relative errors are presented. Thereafter, for each department, the factor analysis to statistically determine the significant factors is described. The factors deemed insignificant are removed, and the analysis is repeated again until only the significant factors remain. A sensitivity analysis showing the impact of each significant factor is also discussed. The chapter ends by summarizing the findings for all of the departments.

4.1 Analysis Department 1

The masked data for the seven design jobs are shown in Table 9. It should be recalled that only the actual design effort, ACT, is considered confidential. Hence, only the column of ACT is masked. The remaining factors, type of design (TD), degree of change (DC), concurrency (Con), and experience of departmental personnel (DE), are not confidential, and thus are not masked.

Table 9: Data of D1

DJ	TD	DC	Con	DE	ACT
A	1	1	0.73	3.00	259.07
B	2	2	0.65	3.00	121.39
C	2	3	0.66	3.00	288.73
D	1	1	0.83	3.00	249.98
E	2	2	0.69	2.25	462.65
F	2	3	0.71	2.65	480.04
G	1	1	0.70	2.44	734.14

As mentioned in Section 3.3, in order to estimate the coefficients, a multiple linear regression is used. The MLRM is used with the natural log (ln) of every value, as per the required format discussed in the section mentioned above. The ln value of each term is presented in Table 10.

Table 10: ln of data of D1

DJ	ln (TD)	ln (DC)	ln (Con)	ln (DE)	ln (ACT)
A	0	0	-0.315	1.099	5.557
B	0.693	0.693	-0.431	1.099	4.799
C	0.693	1.099	-0.416	1.099	5.665
D	0	0	-0.186	1.099	5.521
E	0.693	0.693	-0.371	0.811	6.137
F	0.693	1.099	-0.343	0.975	6.174
G	0	0	-0.357	0.892	6.599

Since the sample size is very small, the jackknife technique is used to eliminate biases. The jackknife regression coefficients ($COEF_{JACK}$) generated is presented in Table 11.

Table 11: Regression coefficients of D1

Constants	All	$Jack_A$	$Jack_B$	$Jack_C$	$Jack_D$	$Jack_E$	$Jack_F$	$Jack_G$	$COEF_{JACK}$
ln (a_o)	10.63	10.59	10.61	10.70	10.53	10.61	10.70	10.57	10.62
a_1	-3.33	-3.35	-3.31	-3.25	-3.42	-3.34	-3.39	-3.24	-3.328
a_2	2.11	2.10	2.09	2.04	2.15	2.11	2.17	2.08	2.108
a_3	-0.57	-0.64	-0.59	-0.48	-1.11	-0.59	-0.48	-0.40	-0.6107
a_4	-4.76	-4.72	-4.74	-4.80	-4.85	-4.74	-4.80	-4.67	-4.762

From the jackknife values above the predicted value, PRED of the ln (ACT) is:

$$\text{PRED ln (ACT)} = 10.62 - 3.328 * \ln(TD) + 2.108 * \ln(DC) - 0.6107 * \ln(Con) - 4.762 * \ln(DE)$$

The residuals (errors) for this regression using the jackknife coefficients are tabulated in Table 12.

Table 12: Residuals of D1

DJ	PRED ln(ACT)	ln(ACT)	Residual (Error)
A	5.5763	5.5571	-0.0192
B	4.8012	4.7990	-0.0022
C	5.6466	5.6655	0.0189
D	5.4979	5.5214	0.0235
E	6.1348	6.1370	0.0022
F	6.1928	6.1739	-0.0189
G	6.5859	6.5987	0.0128

The residuals are small, indicating a good model.

The MLRM can only be used if the linearity assumptions are fulfilled. In this regard, the first test of linearity is conducted by generating a scatter plot of the standardized residuals against the predicted values for the whole regression. The standardized residuals should be within a value of ± 1. The standardized residual is calculated with the following equation.

$$Z_{\varepsilon_i} = \frac{\varepsilon_i - \bar{\varepsilon}}{\sigma_\varepsilon}, \ \forall \ i = 1, 2, \ldots, n \qquad (4.1)$$

where

Z_{ε_i}, Standardized residual of observation i

ε_i, Residual value of observation i

$\bar{\varepsilon}$, Mean of residuals

σ_ε, Standard deviation of residuals

The residual plot is displayed in Figure 4.

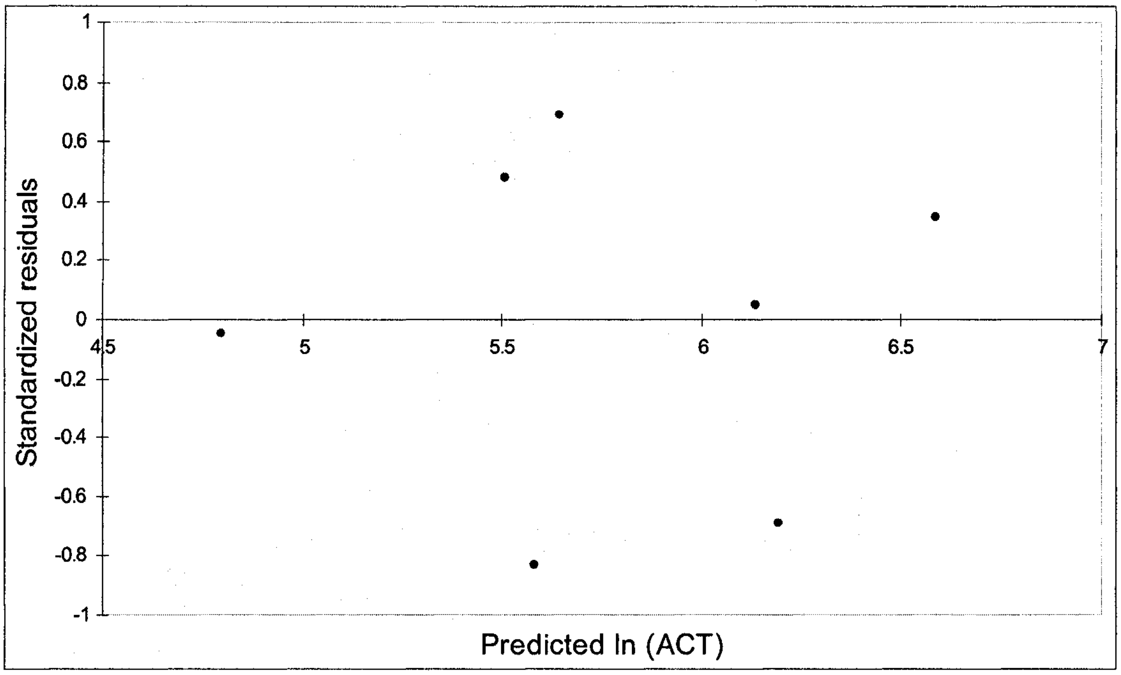

Figure 4: Residual plot of D1

As can be seen from Figure 4, no curvilinear patterns are observed. The residual plots for all the jackknife samples of D1 can be seen in Figures 11 - 17 in Appendix A. There are no curvilinear patterns in the jackknife samples as well, indicating that the requirements of the first linearity test are fulfilled.

The second linearity test is another visual test, where the statistical process control (SPC) charts are generated for the samples. The SPC chart of D1 is portrayed in Figure 5.

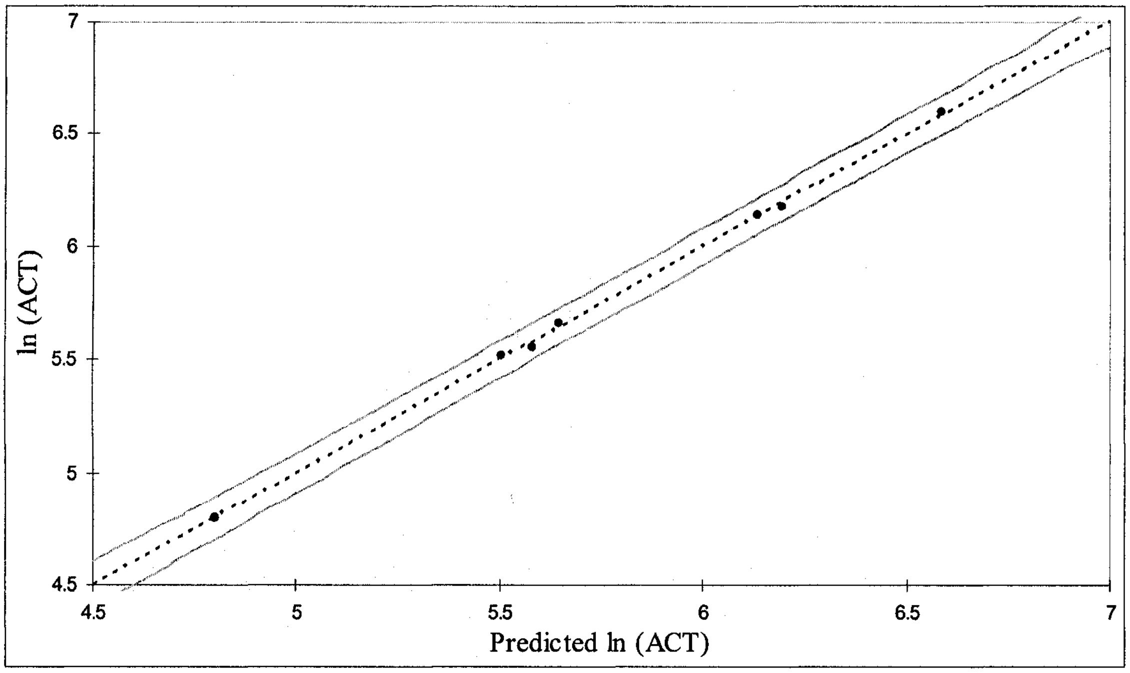

Figure 5: SPC chart of D1

As can be seen from the figure above, the points fall within the upper and lower control limits, therefore, the function is statistically in control. The SPC charts for all the jackknife samples of D1 can be seen in Figures 18-24 in Appendix B. All of the jackknife samples fall within their respective control limits. Thus, the linearity assumption is fulfilled for this department.

The next test to be fulfilled is the normality of errors test. This is done by calculating the coefficient of correlation, r. Recall equation 3.8, where $r = \pm\sqrt{R^2}$ .The r-value has to be grater than the critical value r_L, in this case $r_L = 0.898$, in order to pass the test, having a normal behaviour (Looney and Gulledge, 1985). Since the jackknife technique was used, the r-value was calculated from the minimum R^2 value of the entire sample and all jackknife samples. The values of R^2 can be seen in Table 13. The reason for calculating it from the minimum R^2 value is because if it passes the test in the worst case, it will pass in all other cases.

Table 13: R^2 values of D1

Sample	R^2
All	0.999
$Jack_A$	1.000
$Jack_B$	0.998
$Jack_C$	1.000
$Jack_D$	1.000
$Jack_E$	0.999
$Jack_F$	1.000
$Jack_G$	0.999
R^2min	0.998
r_{R^2min}	0.999

It is clear from Table 13 that the r-value calculated is greater than r_L. Hence the data has a normal behaviour for error.

Now as it is ascertained that the model created fulfills the MLRM, it is known that it is reliable and can be used. The model is now transformed back into the original form to estimate the design effort, $\hat{E}$. In order to transform the equation to its original form, the exponential of both sides is taken. The transformation is shown below:

$e^{PRED\ \ln(ACT)} = \exp[10.62 - 3.328*\ln(TD) + 2.108*\ln(DC) - 0.6107*\ln(Con) - 4.762*\ln(DE)]$

Estimated Design Effort = $e^{10.62} * TD^{-3.328} * DC^{2.108} * Con^{-0.6107} * DE^{-4.762}$

$\hat{E}_{D1} = 4.078x10^4\ TD^{-3.328}\ DC^{2.108}\ Con^{-0.6107}\ DE^{-4.762}$

As it is important to see how accurate the model is, the relative error for each of the DJs is calculated by the following equation.

$$\text{Relative Error}_i = 100 * \frac{|\hat{E}_i - ACT_i|}{ACT_i},\ \forall\ i = A, B, \ldots, G \qquad (4.2)$$

where

$\hat{E}_i$, Expected design effort for DJ_i

ACT_i, Actual design effort for DJ_i

For each of the DJs in D1, Table 14 shows the actual (masked) hours, the predicted hours from the model, and their respective relative errors.

Table 14: Relative errors of D1

DJ	ACT (hours)	PRED (hours)	Relative Error (%)
A	259.07	264.09	1.94
B	121.39	121.66	0.22
C	288.73	283.32	1.87
D	249.98	244.18	2.32
E	462.65	461.63	0.22
F	480.04	489.20	1.91
G	734.14	724.80	1.27

It can be seen that the model built is accurate, with a maximum relative error of only 2.32%. Table 15 shows the correlation between all factors.

Table 15: Correlation matrix of D1

Variables	ln (TD)	ln (DC)	ln (Con)	ln (DE)	ln (ACT)
ln (TD)	1.000	0.945	-0.688	-0.152	-0.182
ln (DC)	0.945	1.000	-0.614	-0.053	-0.069
ln (Con)	-0.688	-0.614	1.000	0.161	0.094
ln (DE)	-0.152	-0.053	0.161	1.000	-0.765
ln (ACT)	-0.182	-0.069	0.094	-0.765	1.000

Since the original data cannot be disclosed, it should be noted that the correlation of the masked data has a similar behaviour to the unmasked one, as would be expected, and the ranked-order correlation discussed in Section 2.4 is maintained.

Even though the model is accurate, it is important to see that all the factors are statistically significant. The following criterion defined by Kutner *et al.* (2004) is used to determine if the factors are deemed significant.

If

$(\Pr > |t|) \leq 0.05$, the factor is considered statistically significant

Else

Factor is considered statistically insignificant

The summary of statistics, showing the Pr > |t| for the entire sample, and for all jackknife samples can be seen in Table 16.

Table 16: Significant factors of D1

Factor	All	$Jack_A$	$Jack_B$	$Jack_C$	$Jack_D$	$Jack_E$	$Jack_F$	$Jack_G$
ln (a_o)	0.000	0.007	0.017	0.007	0.004	0.017	0.007	0.010
ln (TD)	0.001	0.019	0.050	0.022	0.010	0.044	0.020	0.037
ln (DC)	0.001	0.020	0.055	0.025	0.010	0.042	0.024	0.030
ln (Con)	0.107	0.182	0.324	0.243	0.103	0.324	0.243	0.464
ln (DE)	0.000	0.013	0.031	0.013	0.007	0.031	0.013	0.025

From Table 16, it can be seen that all of the factors are considered to be significant with the exception of concurrency. Therefore, the analysis will now be repeated without concurrency as a factor.

4.1.1 Analysis of department D1 without concurrency

If the concurrency values are omitted from Table 9, the new regression coefficients generated, using the jackknife technique can be seen in Table 17.

Table 17: Regression coefficients of D1 without concurrency

Constants	All	$Jack_A$	$Jack_B$	$Jack_C$	$Jack_D$	$Jack_E$	$Jack_F$	$Jack_G$	$COEF_{JACK}$
ln (a_o)	10.805	10.799	10.947	10.882	10.765	10.947	10.882	10.553	10.825
a_1	-3.205	-3.207	-3.393	-3.114	-3.219	-3.103	-3.342	-3.085	-3.209
a_2	2.077	2.075	2.196	1.977	2.064	2.013	2.206	2.035	2.081
a_4	-4.771	-4.761	-4.909	-4.845	-4.709	-4.909	-4.845	-4.564	-4.792

From the jackknife values above, the PRED of the ln (ACT) without concurrency would be as follows:

$$\text{PRED ln (ACT)}_{\text{No Con}} = 10.825 - 3.209*\ln(\text{TD}) + 2.081*\ln(\text{DC}) - 4.792*\ln(\text{DE})$$

The residuals for this regression using the jackknife coefficients, without concurrency, are presented in Table 18.

Table 18: Residuals of D1 without concurrency

DJ	PRED ln(ACT)	ln(ACT)	Residual (Error)
A	5.5718	5.5571	-0.0147
B	4.7646	4.799	0.0344
C	5.6656	5.6655	-0.0001
D	5.5379	5.5214	-0.0165
E	6.1714	6.137	-0.0344
F	6.1737	6.1739	0.0001
G	6.564	6.5987	0.0347

As can be seen from the table above, the residuals are small, indicating a good model.

The scatter plot of the standardized residuals, which is used for the first linearity test, is shown in Figure 6.

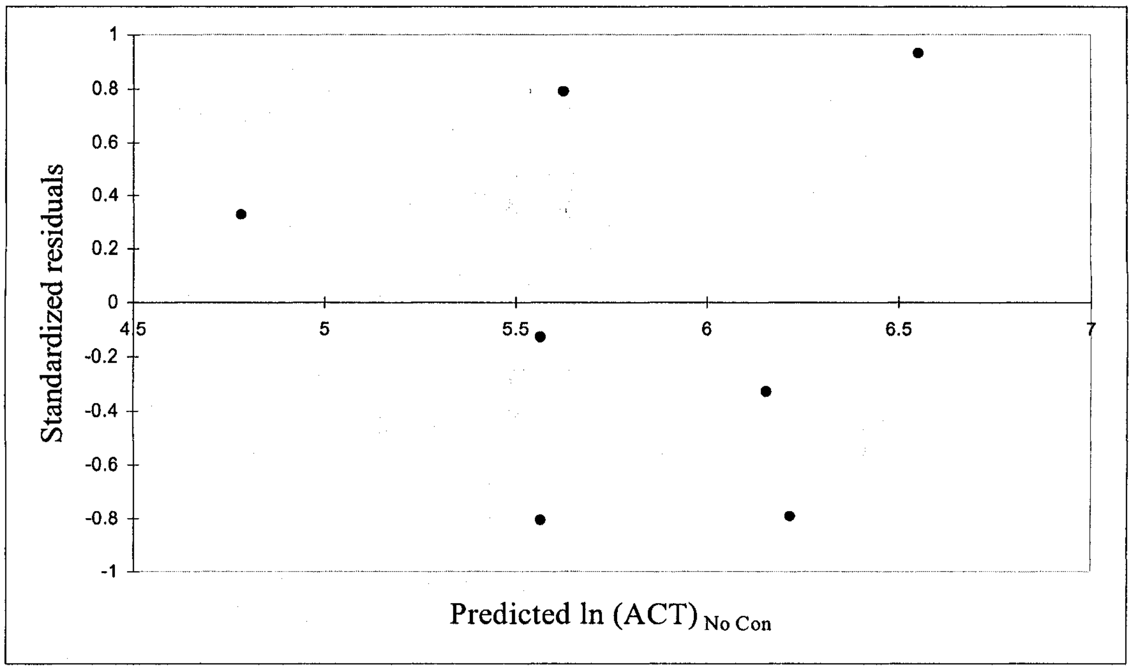

Figure 6: Residual plot of D1 without concurrency

As can be seen from Figure 6, no curvilinear patterns are observed. The residual plots for all the jackknife samples can be seen in Figures 25-31 in Appendix C. There are no curvilinear patterns in the jackknife samples as well, indicating the model passes the first test of linearity. The SPC chart, for the second linearity test is shown in Figure 7.

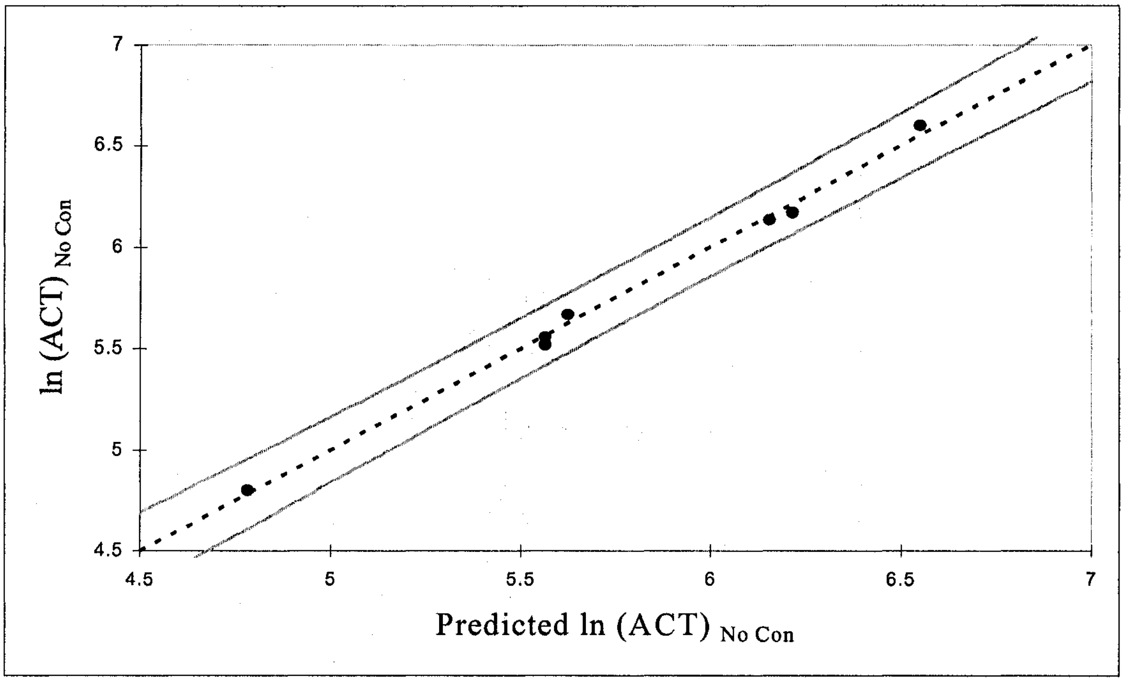

Figure 7: SPC chart of D1 without concurrency

As can be from the figure above, all the points fall within the upper and lower control limits and the function is statistically in control. The SPC charts for all the jackknife samples can be seen in Figures 32-38 in Appendix D. All of the jackknife samples fall within the control limits. Thus, the linearity assumption is fulfilled for this department.

The r-value calculated from the minimum R^2 values for the normality of errors test can be seen in Table 19.

Table 19: R^2 values of D1 without concurrency

Sample	R^2
All	0.996
$Jack_A$	0.996
$Jack_B$	0.992
$Jack_C$	0.998
$Jack_D$	0.997
$Jack_E$	0.996
$Jack_F$	0.998
$Jack_G$	0.998
R^2min	0.992
r_{R^2min}	0.996

It is clear from Table 19 that the r-value calculated is greater than r_L value of 0.898. Hence, the data has a normal behaviour for error.

Now that it is confirmed that the model created fulfills the MLRM, it is known that it is reliable for use. The model is transformed back into the original form to estimate design effort, $\hat{E}$. The transformation is shown below:

$$e^{\text{PRED ln (ACT)}}{}_{\text{No Con}} = \exp[10.825 - 3.209*\ln(\text{TD}) + 2.081*\ln(\text{DC}) - 4.792*\ln(\text{DE})]$$

$$\text{Estimated Design Effort}_{\text{No Con}} = e^{10.825} * \text{TD}^{-3.209} * \text{DC}^{2.081} * \text{DE}^{-4.792}$$

$$\hat{E}_{D1_{\text{No Con}}} = 5.026\text{x}10^4\ \text{TD}^{-3.209}\ \text{DC}^{2.081}\ \text{DE}^{-4.792}$$

It is important to see how accurate the model is. Thus the relative error for each of the DJs is calculated. For each of the DJs in the department, Table 20 shows the actual (masked) hours, the estimated hours and their respective relative errors.

Table 20: Relative errors of D1 without concurrency

DJ	ACT (hours)	PRED (hours)	Relative Error (%)
A	259.07	260.01	0.36
B	121.39	118.93	2.03
C	288.73	276.57	4.21
D	249.98	260.01	4.01
E	462.65	472.12	2.05
F	480.04	501.14	4.40
G	734.14	699.75	4.68

It can be seen that the model built was quite accurate with a maximum relative error of only 4.68%. Table 21 shows the correlation between all the factors.

Table 21: Correlation matrix of D1 without concurrency

Variables	ln (TD)	ln (DC)	ln (DE)	ln (ACT)
ln (TD)	1.000	0.945	-0.152	-0.182
ln (DC)	0.945	1.000	-0.053	-0.069
ln (DE)	-0.152	-0.053	1.000	-0.765
ln (ACT)	-0.182	-0.069	-0.765	1.000

Since the original data cannot be disclosed, it should be noted that the correlation of the masked data has a similar behaviour to the unmasked one, as would be expected, and the ranked-order correlation is the same.

Once again, the remaining factors are verified for statistical significance. The criterion defined in Section 4.1 will be used to determine if the factors are deemed significant. The summary of statistics, showing the Pr > |t| for the entire sample, and for all jackknife samples for D1 without concurrency can be seen in Table 22.

Table 22: Significant factors of D1 without concurrency

Factor	All	$Jack_A$	$Jack_B$	$Jack_C$	$Jack_D$	$Jack_E$	$Jack_F$	$Jack_G$
ln (a_o)	0.000	0.001	0.001	0.000	0.000	0.001	0.000	0.000
ln (TD)	0.000	0.005	0.012	0.003	0.003	0.008	0.003	0.002
ln (DC)	0.001	0.006	0.013	0.005	0.004	0.008	0.005	0.002
ln (DE)	0.000	0.003	0.004	0.001	0.002	0.004	0.001	0.001

From Table 22 it can be seen that all the factors are considered to be significant. Thus, the analysis for D1 is complete. Sensitivity analyses in the form of graphs, showing the main effect that each significant factor has on the design effort are depicted in Figures 8-10.

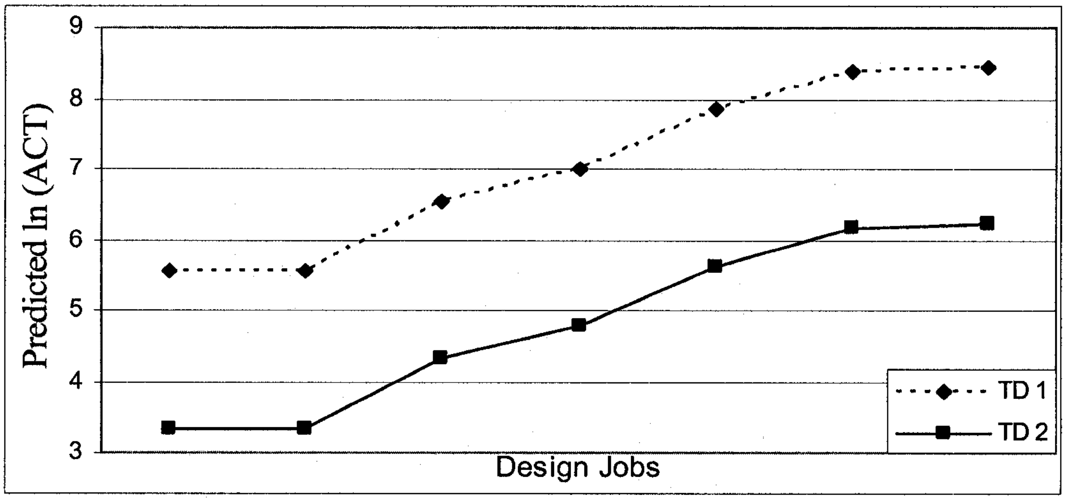

Figure 8: Impact of the type of design on design for D1

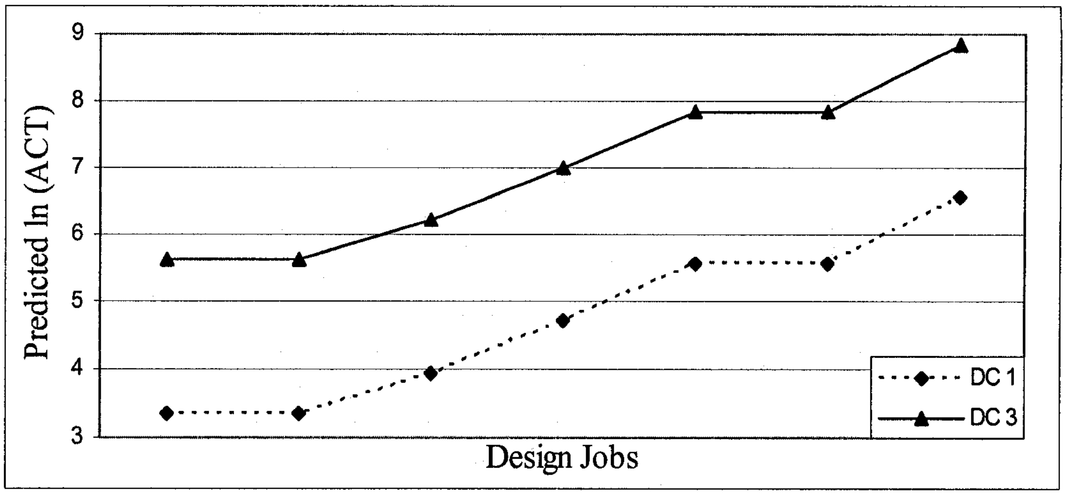

Figure 9: Impact of the degree of change on effort for D1

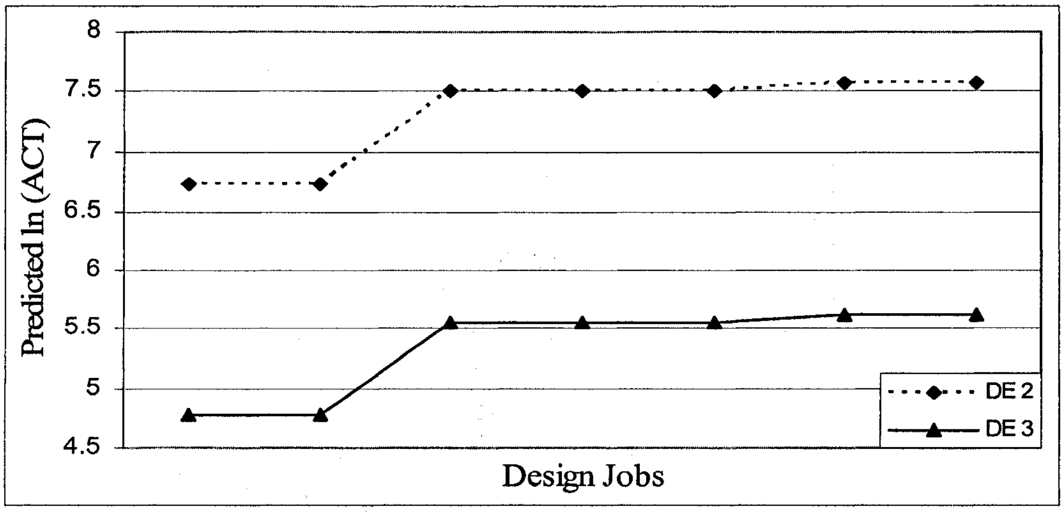

Figure 10: Impact of the experience of departmental personnel on effort for D1

Based on the results, tests conducted, and comparisons of the actual project data, it can be seen that the model generates accurate results. It meets the requirements of the MLRM and has a maximum relative error of only 2.32%, and when concurrency is not considered, the maximum relative error slightly increases to 4.68%. From Figure 8, it can be seen that if the design were an initial one, it would take longer than that of a redesign. Figure 9 points out that the degree of change being major will require more time than a minor one. Finally, Figure 10 identifies, as would be expected, that the higher the net level of experience of the departmental personnel working on the design job, the lesser the effort is required.

4.2 Analysis of Other Departments

A similar thorough analysis was carried out for the remaining departments (D2-D4). The summary of the analysis of the remaining departments are presented in the following sections. However, all the intermediary steps including the figures and tables can be found in Appendices E through G.

4.2.1 Analysis of Department 2

The main finding in terms of the generated model, the maximum relative error, and the impact of the significant factors are highlighted for Department 2 below.

- $\hat{E}_{D2} = 2.63\text{x}10^{2}\,\mathrm{TD}^{1.75}\,\mathrm{DC}^{0.221}\,\mathrm{Con}^{-0.131}\,\mathrm{DE}^{-0.345}$
- Maximum relative error of 2.99%
- Concurrency not significant
- $\hat{E}_{D2_{\text{No Con}}} = 2.69\text{x}10^{2}\,\mathrm{TD}^{1.76}\,\mathrm{DC}^{0.225}\,\mathrm{DE}^{-0.325}$
- Maximum relative error of 3.30%

A sensitivity analysis was conducted and the findings were as follows. From Figure 43, it can be seen that if the design were an initial one, it would take less time than that of a redesign. Figure 44 points out that the degree of change being major will require more time than a minor one. Finally, Figure 45 identifies, as would be expected, that the higher the net level of experience of the departmental personnel working on the design job, the less effort is required.

4.2.2 Analysis of Department 3

Similar to Department 2, an overview of the analysis for the third department is presented below:

- $\hat{E}_{D3} = 6.34\text{x}10^{2}\,\mathrm{TD}^{1.23}\,\mathrm{DC}^{-0.222}\,\mathrm{Con}^{0.229}\,\mathrm{DE}^{-0.906}$
- Maximum relative error of 2.29%

The following impacts (main effects) of the significant factors were determined. If the type of design is an initial one, it will require less time than that of a redesign (Figure 48). As the degree of change increases, the required effort decreases (Figure 49). As the level of concurrency increases, the required design effort decreases (Figure 50). Finally, the higher the net experience the team members working on the design have, the less effort they will require to complete the design (Figure 51).

4.2.3 Analysis of Department 4

A review of the last department analyzed (Department 4) is summarized below:

- $\hat{E}_{D4} = 5.26\text{x}10^2\,\text{TD}^{1.34}\,\text{DC}^{-0.253}\,\text{Con}^{-0.446}\,\text{DE}^{-0.411}$
- Maximum relative error of 6.06 %
- Concurrency and degree of change not significant
- $\hat{E}_{D4_{\text{No Con}}} = 7.28\text{x}10^2\,\text{TD}^{1.14}\,\text{DC}^{-0.068}\,\text{DE}^{-0.562}$
- Maximum relative error of 8.46%
- Degree of change still not significant
- $\hat{E}_{D4_{\text{No Con, No DC}}} = 7.36\text{x}10^2\,\text{TD}^{1.05}\,\text{DE}^{-0.570}$
- Maximum relative error of 9.88 %

As can be seen from the sensitivity analysis, that Figure 58 points if the type of design was an initial one, it would require less effort than that of a redesign. Finally, as the case in the previous three departments, Figure 59 highlights that as would be expected, as the experience level increases, the required time to complete the design decreases.

4.3 Summary of Findings

A summary of all the design effort estimation equations along with their maximum relative errors for each of the departments are summarized in Table 23.

Table 23: Summary of findings of all departments

Department	Design effort ($\hat{E}$)	Maximum relative error (%)
D1	$4.078\text{x}10^4\ TD^{-3.328}\ DC^{2.108}\ Con^{-0.6107}\ DE^{-4.762}$	2.32
D1 $_{\text{No Con}}$	$5.026\text{x}10^4\ TD^{-3.209}\ DC^{2.081}\ DE^{-4.792}$	4.68
D2	$2.63\text{x}10^2\ TD^{1.75}\ DC^{0.221}\ Con^{-0.131}\ DE^{-0.345}$	2.99
D2 $_{\text{No Con}}$	$2.69\text{x}10^2\ TD^{1.76}\ DC^{0.225}\ DE^{-0.325}$	3.30
D3	$6.09\text{x}10^2\ TD^{1.18}\ DC^{-0.184}\ Con^{0.229}\ DE^{-0.906}$	2.29
D4	$5.26\text{x}10^2\ TD^{1.34}\ DC^{-0.253}\ Con^{-0.446}\ DE^{-0.411}$	6.06
D4 $_{\text{No Con}}$	$7.28\text{x}10^2\ TD^{1.14}\ DC^{-0.068}\ DE^{-0.562}$	8.46
D4 $_{\text{No Con, No DC}}$	$7.36\text{x}10^2\ TD^{1.05}\ DE^{-0.570}$	9.88

It should be recalled that the values for the attributes can be the following:

- Type of design, (TD) is either 1 or 2
- Degree of change, (DC) is either 1, 2 or 3
- Concurrency (Con) is $0 \leq Con \leq 1$
- Experience of departmental personnel, (DE) is $1 \leq DE \leq 3$

As can be seen from Table 23, the model built and applied to the fours departments was quite precise with a maximum relative error of less than 10%. Thus, the model would be a good tool to be used to accurately estimate design effort with minimal error.

However, it is clear that the models differ from department to department. Only in department D1 does the type of design, being an initial one, require less time than that of a redesign. This is probably because the other departments are more sensitive to change. Thus, a redesign would result in a lot more rework, even more than the initial design.

For the first two departments (D1 and D2), the higher the level of change, the more time will be required. For the third department (D3), it shows the opposite effect.

However, the value for the exponent is very small (near zero). Having a very small sample size could have offset the results showing this reversal. In the fourth department, it is found to be significant.

The factor of concurrency is significant in the third department (D3), and the higher the level of concurrency the less time of effort is required. This could be explained by them being the analytical department and the more information that is shared with them throughout the process, the less time they would require. Concurrency is not significant in all departments. Even though studies of Joglekar *et al.* (2001), Bhuiyan *et al.* (2004), Jun *et al.* (2005) amongst many others, show that concurrency does affect design effort, the fact was that the DJs are very similar in their concurrency values, thus resulting in concurrency being treated as a constant and rendering it a statistically insignificant factor, as reported in the ANOVA.

The finding for the experience of departmental personnel is consistent in all four cases. The analysis points out, as would be expected, that as the net level of the experience of the members of the department working on the design job is higher, the less effort would be required.

5. CONCLUSIONS AND FUTURE APPLICATIONS

The importance of determining design effort is essential to estimate the lead-time, cost, and manpower needed. The product development process is a very complex process and it is dependent on many factors involved in the design process. Even though it is complicated to estimate the design effort needed for the development of a product, it is essential for the concerned personnel to know it precisely.

Depending upon the product complexity and development strategies, a list of potential factors can be identified that can help to improve such estimations. In this thesis, a case study focusing on four departments (design, aerodynamics, analytical and drafting) at PWC is presented to estimate the design effort of the integrated blade-rotor low- pressure compressor fans.

The study utilizes a parametric modeling technique to estimate the design effort for each of the four departments. Four design factors are considered: type of design, degree of change, concurrency, and experience of departmental personnel. An analysis of each department initially considers all the above-mentioned factors and reduces to keep only the statistically significant factors. Since the data used in this thesis is confidential, coming from a leading aerospace company, an appropriate method for data masking was required.

The model performed well according to a number of accuracy tests suggested in the thesis. Comparisons of the design effort estimation determined by the models are compared with the actual design efforts reported by PWC for the each of the seven design jobs for the four departments. The models are able to estimate the design effort with a minimal error. The maximum error observed in any study with a department is less than 10%.

Sensitivity analyses were conducted for all of the significant factors for each of the department. The analysis showed the impact that each individual factor has on design effort. Moreover, the findings for each of the departments were compared and contrasted. The first two departments each had three significant factors, where all the selected factors were significant for Department 3, and only two factors were deemed significant for Department 4. For Department 1, it is found that the change from initial design to redesign and/or an increase in the type of design and the net experience of departmental personnel had a positive effect in reducing the design effort; whereas the increase in the type of design had an adverse effect on the required effort. Department 2 had similar findings with the change in the increase of the type of design having a negative effect on reducing the design effort. The findings for Department 3 were similar to those of Department 2, with the change in the increase in the degree of change having a positive effect in reducing the required effort. Finally, for Department 4, it was found that the increase in the type of design, as the case of the previous two departments, had a negative effect in reducing the required effort. Similarly, all the case in all the previous departments, the increase in the net experience of the departmental personnel had a positive effect in reducing the design effort.

Even though the model has promising results there are limitations to this model. Even though the methodology can be used for other major components such as the turbine or combustor, the model is specific to the IBR LPC fans for a certain class of engines. Using this approach models will have to be developed for each type of component. Also, factors will have to be selected and validated for each type of component. Even though the factors are general, there may be other factors that will be relevant to other components being studied.

There can be several future applications of this thesis. As mentioned in the limitations, the model is specific to the component, other models can be developed which can be applied not limiting itself to a particular type of component or class of engine, rather the model can be more general. Another possible application could be to study the length of time (lead-time) to design the components. Knowing the lead-time and phasing of hours will significantly help management in scheduling their tasks and in assigning priorities.

6. REFERENCES

Air Force Research Laboratory, "Fan and Compressor Branch," www.pr.afrl.af.mil/divisions/prt/images/fan.gif, 2006.

S. A. Assaf and S. Al-Hejji, "Causes of delay in large construction projects," International Journal of Project Management, vol. 24, no.4, pp. 349-356, 2006.

H. A. Bashir and V. Thomson, "Estimating design complexity" Journal of Engineering Design, vol. 10, no. 3, pp. 244-257, 1999.

H. A. Bashir and V. Thomson, "Models for estimating design effort and time," Design Studies, vol. 22, no. 2, pp. 141-155, 2001.

H. A. Bashir and V. Thomson, "Estimating design effort for GE hydro projects," Computers & Industrial Engineering, vol. 46, no. 2, pp. 195-204, 2004.

N. Bhuiyan, "Dynamic Models of Concurrent Engineering Processes and Performance," PhD Research Project, McGill University, Montreal, Canada, Dec. 2001.

N. Bhuiyan, D. Gerwin, and V. Thomson, "Simulation of the new product development process for performance improvement," Management Science, vol. 50, no. 12, pp. 1690-1703, Dec. 2004.

B. W. Boehm, Software engineering economy, NJ: Prentice-Hall Inc., 1981.

S. M. Bogus, K. R. Molenaar, and J. E. Diekmann, "Concurrent engineering approach to reducing design delivery time," Journal of Construction Engineering and Management-Asce, vol. 131, no. 11, pp. 1179-1185, 2005.

G. Bounds, "The Last Word on Project Management," IIE Solutions, pp. 41-43, Nov. 1998.

R. J. Bronikowski, Managing the Engineering Design Function, NY: Van Nostrand Reinhold Company, 1986.

K. B. Clark and T. Fujimoto, "Product Development Performance," Harvard Business School Press, Boston, Massachusetts, 1991.

G. Colmer, M. Dunkley, K. Gray, P. Pugh, and A. Williamson, "Estimating the cost of new technology products," International Journal of Technology Management, vol. 17, nos. 7-8, pp. 840-846, 1999.

M. L. Cook, "Software Metrics: an Introduction and Annotated Bibliography," Software Engineering Notes (ACM SIGSOFT), vol. 7, no. 2, pp. 41-60, 1982.

B. Efron and R. Tibshirani, An Introduction to the Bootstrap (Monographs on Statistics and Applied Probability), FL: Chapman and Hall, 1994.

Y. Frimpong, "Project management in developing countries: causes of delay and cost overruns in construction of groundwater projects," Masters Research Project, University of Technology, Sydney, Australia, Nov. 2000.

K. P. Grant, W. M. Cashman, and D. S. Christensen, "Delivering Projects on Time," Research Technology Management, vol. 49, no. 6, pp. 52-58, 2006.

F. A. Graybill, Theory and Application of the Linear Model, MA: Duxbury Press, 1976.

D. R. Jeffery, "The relationship between team size, experience, and attitudes and software development productivity," IEEE Computer Society Press, COMPSAC, pp. 2-8, 1987.

N. R. Joglekar, A. A. Yassine, S.D, Eppinger, and D. E. Whitney, "Performance of coupled product development activities with a deadline," Management Science, vol. 47, no. 12, pp. 1605-1620, Dec. 2001.

H. B. Jun, H. S. Ahn, and H. W. Suh, "On identifying and estimating the cycle time of product development process," IEEE Transactions on Engineering Management, vol. 52, no. 3, pp. 336-349, Aug. 2005.

S. Kara, B. Kayis, and H. Kaebernick, "Modeling concurrent engineering projects under uncertainty," Concurrent Engineering-Research and Applications, vol. 7, no. 3, pp. 269-274, 1999.

M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li, Applied Linear Statistical Models, NY: McGraw-Hill, 2004.

C. H. Loch, and C. Terwiesch, "Communication and uncertainty in concurrent engineering," Management Science, vol. 44, no. 8, pp. 1032-1048, 1998.

S. Looney and T. Gulledge Jr., "Use of the correlation coefficient with normal probability pots," The American Statistician, vol. 39, pp. 75-79, Feb. 1985.

A. Mehrez and I. David, "A two-phase simple model to determine the timing and duration of R & D project tasks," Production Planning & Control, vol. 10, no. 1, pp. 48-53, 1999.

D. C. Montgomery, Introduction to Statistical Quality Control, NY: John Wiley, 1985.

K. Muralidhar, R. Parsa, and R. Sarathy, "A General Additive Data Perturbation Method for Database Security," Management Science, vol. 45, no. 10, pp. 1399-1415, Oct. 1999.

K. Muralidhar and R. Sarathy, "Data shuffling - A new masking approach for numerical data," Management Science, vol. 52, no. 5, pp. 658-670, 2006.

P. A. Murmann, "Expected Development Time Reductions in the German Mechanical Engineering Industry," Journal of Product Innovation Management, vol. 11, no. 3, pp. 236-252, June 1994.

K. P. Norris, "The Accuracy of Project Cost and Duration Estimates in Industrial R&D," R&D Management, vol. 2, no. 1, pp. 25-36, Oct. 1971.

R. A. Parsa, K. Muralidhar, and R. Sarathy, "Conditions for maximum security in linear data perturbation," ASA Proceedings of the Section on Government Statistics and Section on Social Statistics, pp. 70-72, 1998.

D. Phan, D. Vogel, and J. Nunamaker, "The search for perfect project management," Computer world, vol. 22, no. 39, pp. 95-100, 1998.

PilotFriend, "Aero Engines: Aircraft Engine History- Jet Engines," www.pilotfriend.com/aero_engines/aero_jet.htm, 2000.

R. Sarathy, K. Muralidhar, and R. Parsa "Perturbing non-normal confidential attributes: The copula approach," Management Science, vol. 48, no. 12, pp. 1613-1627, 2002.

A. Sklar, "Fonctions de Répartition a n dimensions et Leurs Mages," Pub. Inst. Statist. Univ. Paris, vol. 8, pp. 229-231, 1959.

R. P. Smith and S. D. Eppinger, "A predictive model of sequential iteration in engineering design," Management Science, vol. 43, no. 8, pp. 1104-1120, 1997.

R. P. Smith and S. D. Eppinger, "Deciding between sequential and concurrent tasks in engineering design," Concurrent Engineering-Research and Applications, vol. 6, no. 1, pp. 15-25, 1998.

H. J. Thamhain and D. L. Wilemon, "Criteria for controlling projects according to plan," Project Management Journal, vol. 17, no. 2, pp. 75-81, 1986.

K. T. Ulrich and S. D. Eppinger, Product Design and Development, NY: McGraw-Hill, 1995.

C. E. Walston and C. P. Felix, "A Method of Programming Measurement and Estimation," IBM System Journal, vol. 16, no. 1, pp. 54-73, 1977

Z. Wang and H. S. Yan, "Optimizing the concurrency for a group of design activities," IEEE Transactions on Engineering Management, vol. 52, no. 1, pp. 102-118, 2005.

R. I. Winner, J. P. Pennell, H. E. Bertrand, and M. M. G. Slusarezuk, "The Role of Concurrent Engineering in Weapons System Acquisition," Institute for Defense Analyses, Alexandria, VA, USA, IDA Report R-338, (1988).

A. Yassine and D. Braha, "Complex concurrent engineering and the design structure matrix method," Concurrent Engineering-Research and Applications, vol. 11, no. 3, pp. 165-176, 2003.

A. A. Yassine, K. R. Chelst, and D. R. Falkenburg, "A decision analytic framework for evaluating concurrent engineering," IEEE Transactions on Engineering Management, vol. 46, no. 2, pp. 144-157, 1999.

B. J. Zirger and J. L. Hartley, "A Conceptual-Model of Product Development Cycle Time," Journal of Engineering and Technology Management, vol. 11, nos. 3-4, pp. 229-251, 1994.

APPENDIX A: RESIDUAL PLOTS OF D1

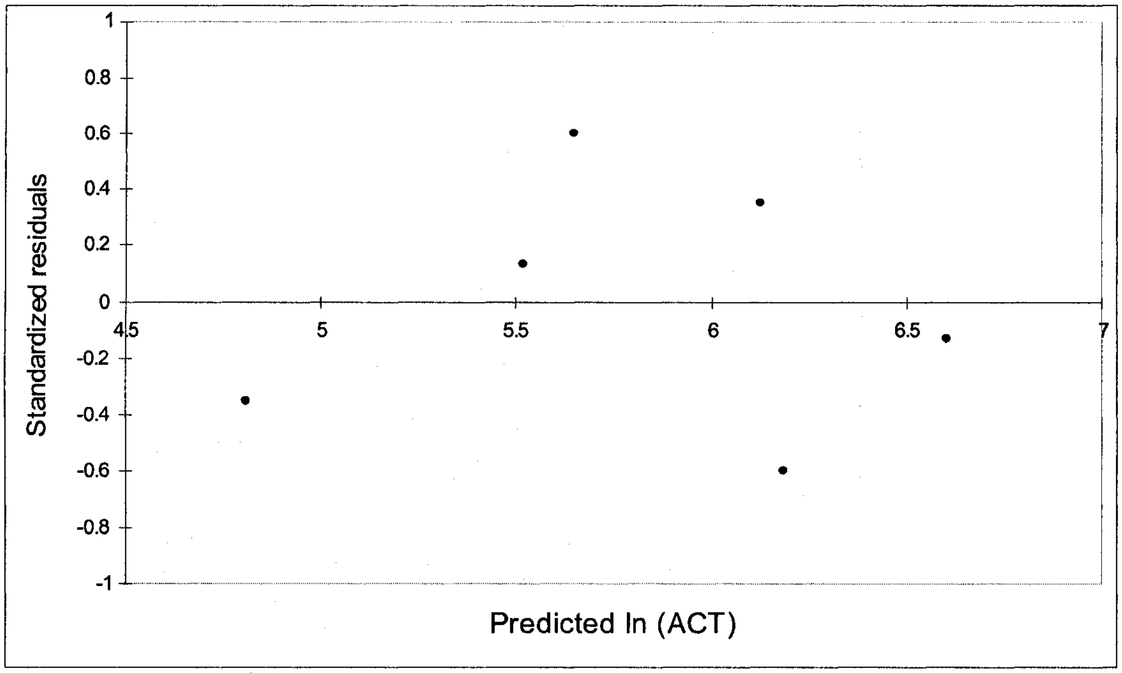

Figure 11: Residual plot of jackknife sample A

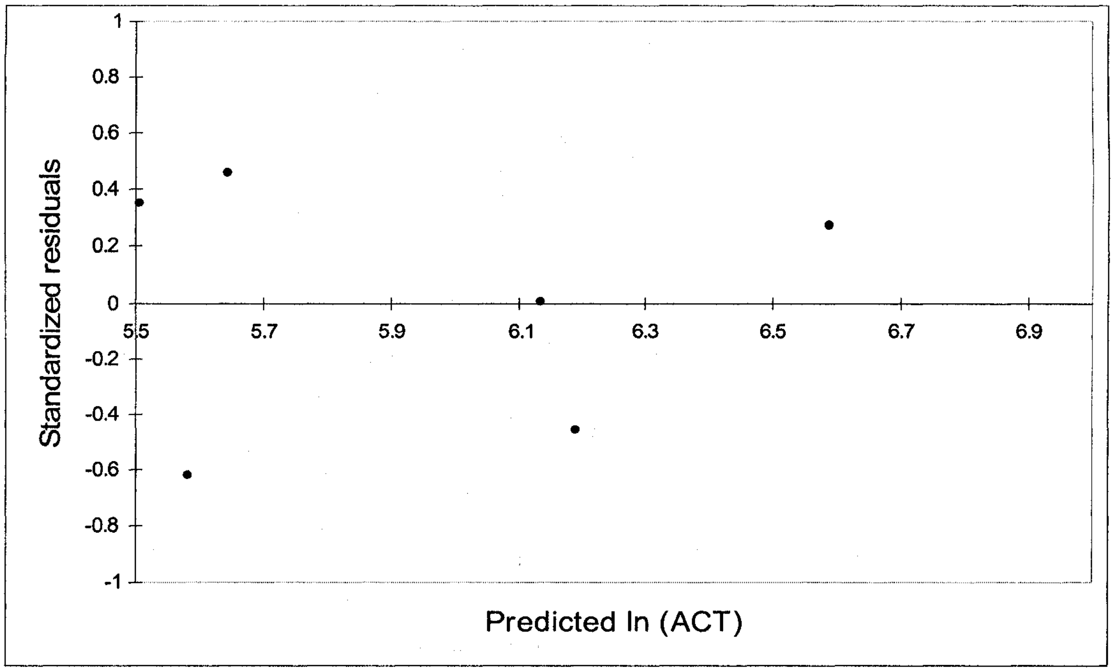

Figure 12: Residual plot of jackknife sample B

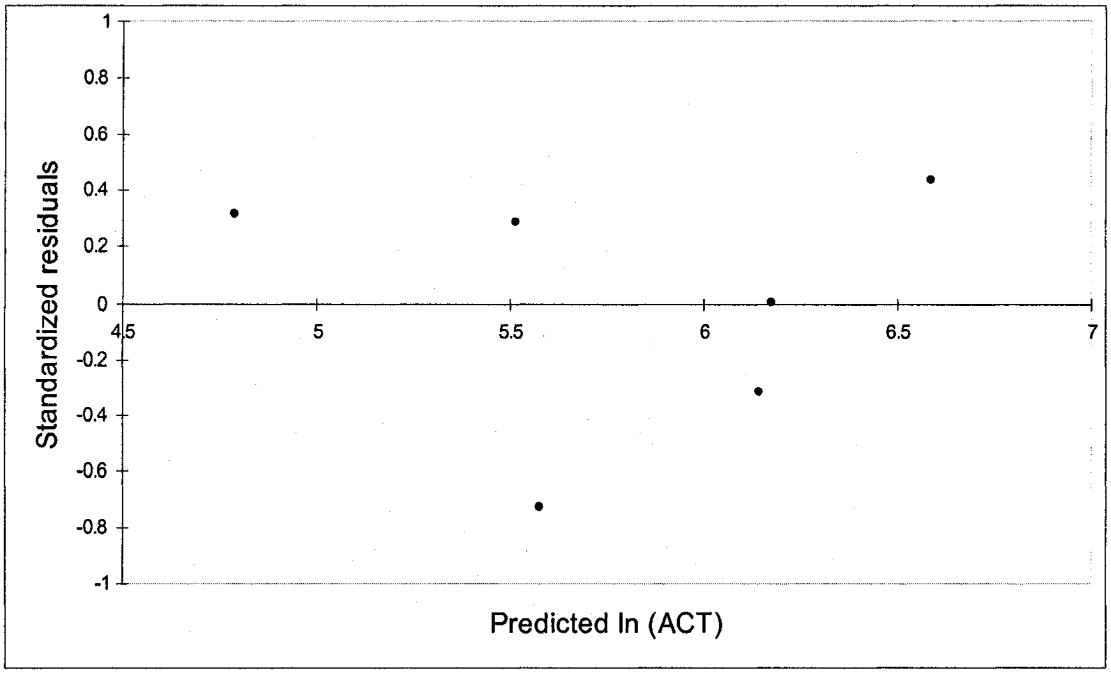

Figure 13: Residual plot of jackknife sample C

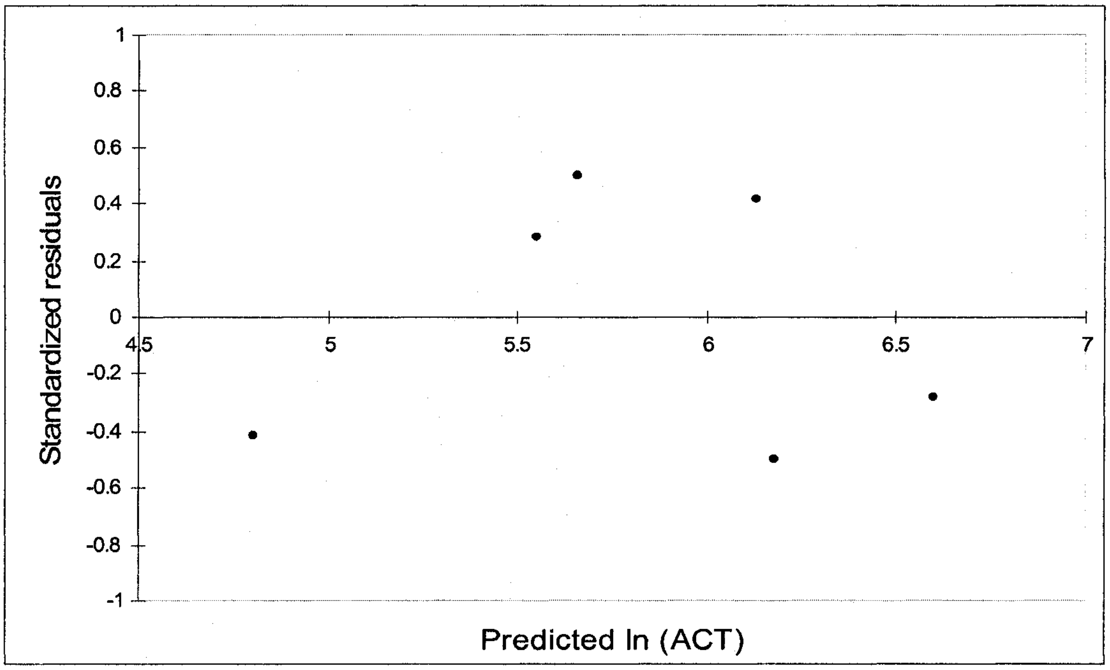

Figure 14: Residual plot of jackknife sample D

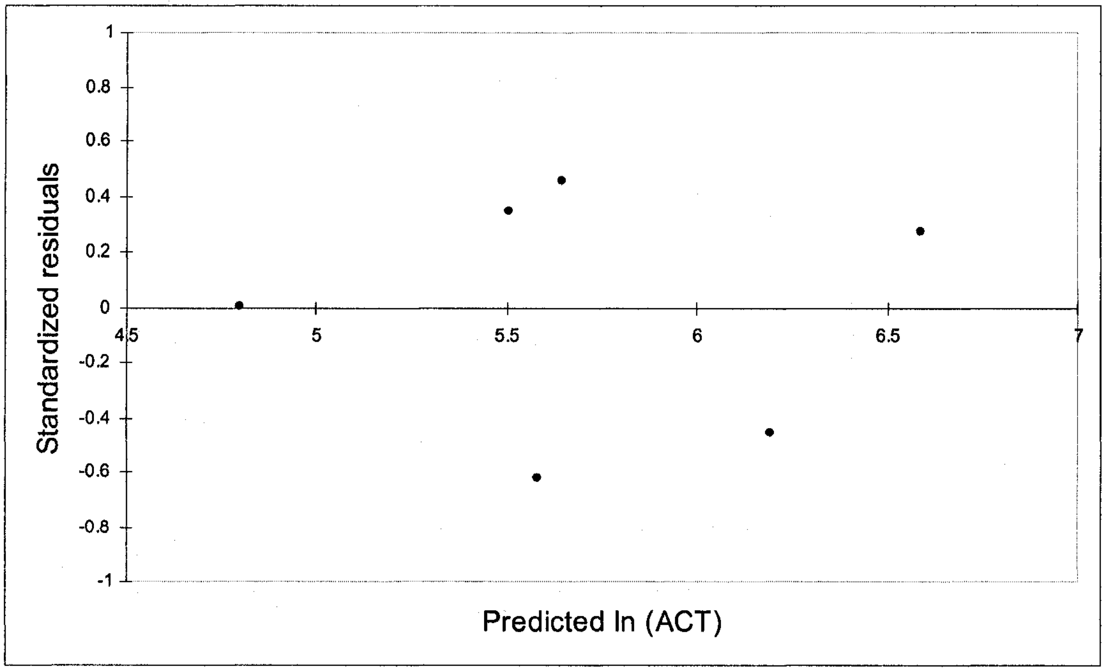

Figure 15: Residual plot of jackknife sample E

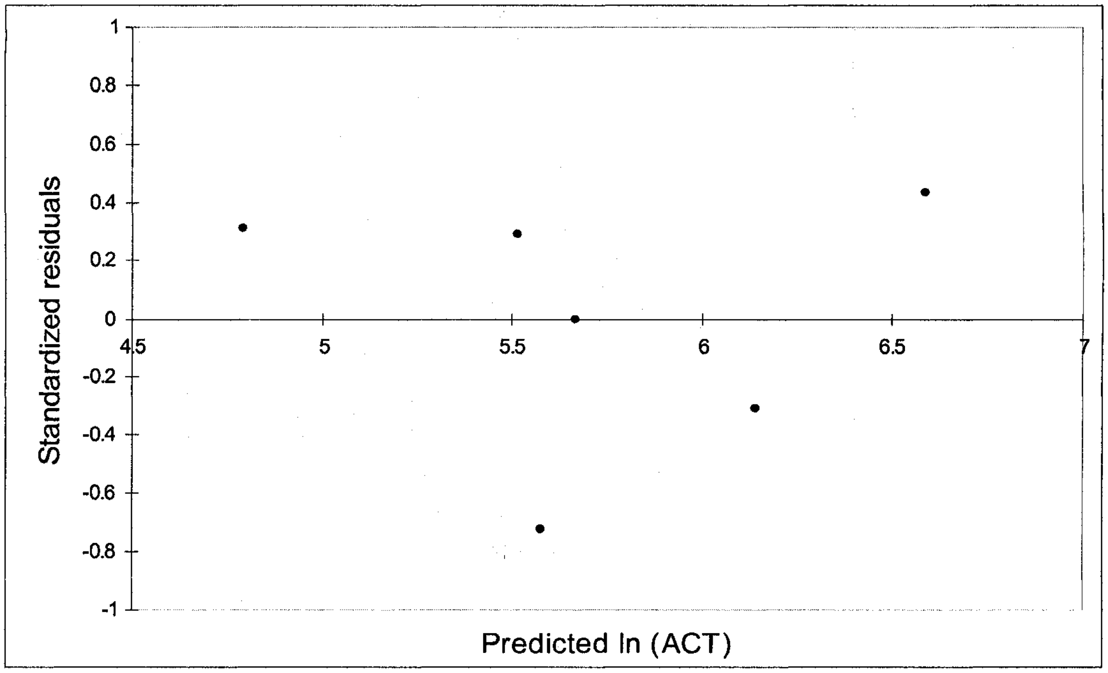

Figure 16: Residual plot of jackknife sample F

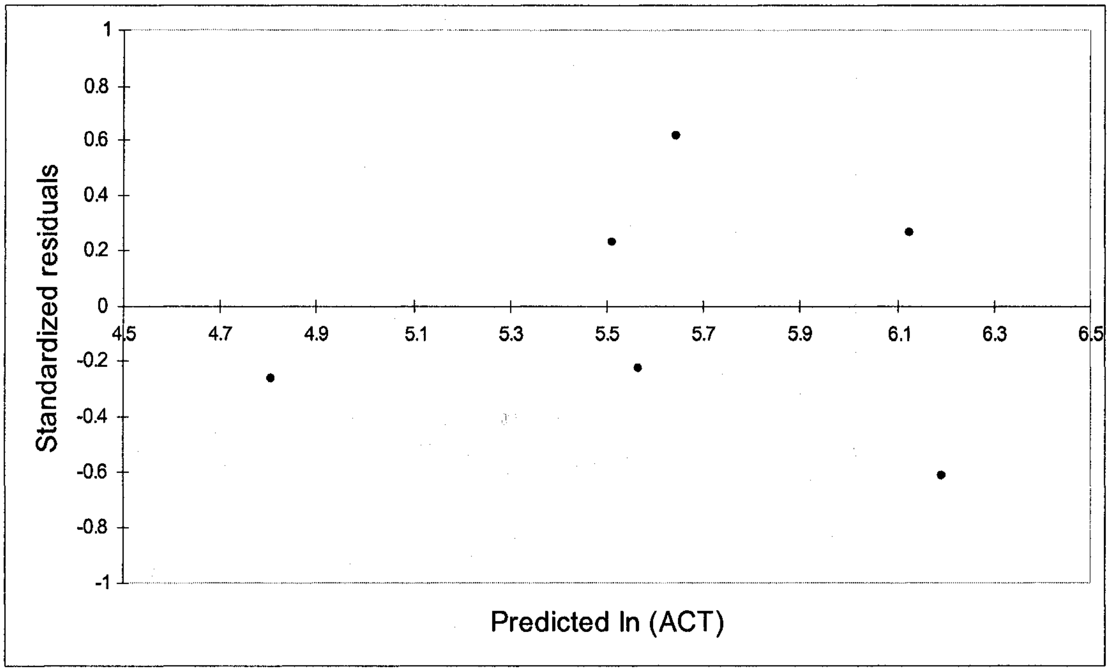

Figure 17: Residual plot of jackknife sample G

APPENDIX B: SPC CHARTS OF D1

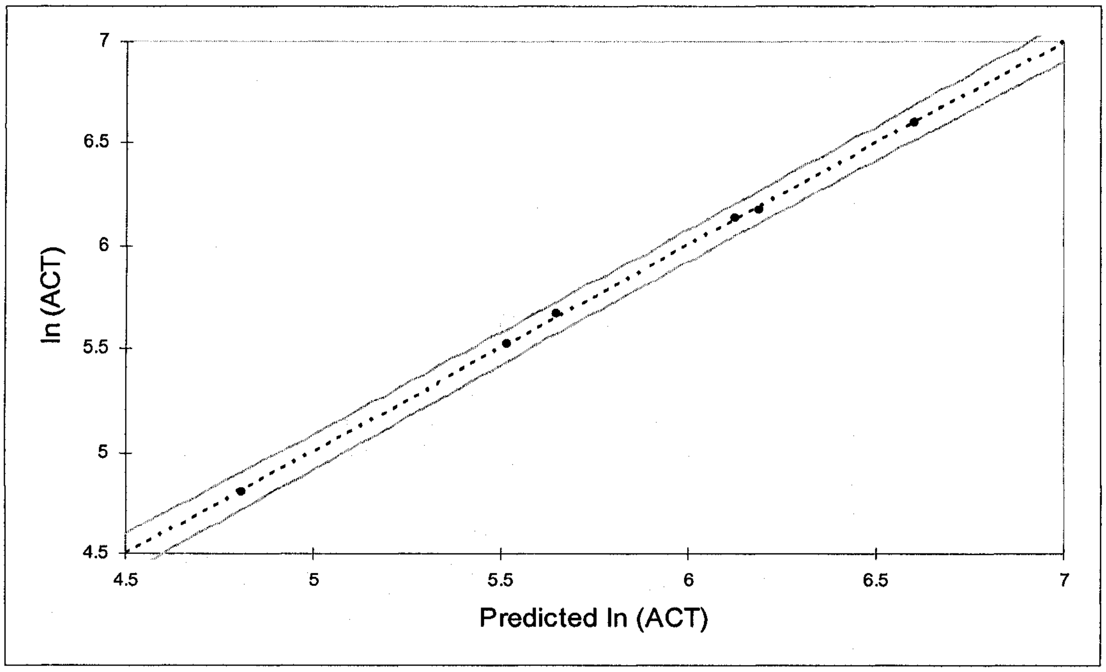

Figure 18: SPC chart of jackknife sample A

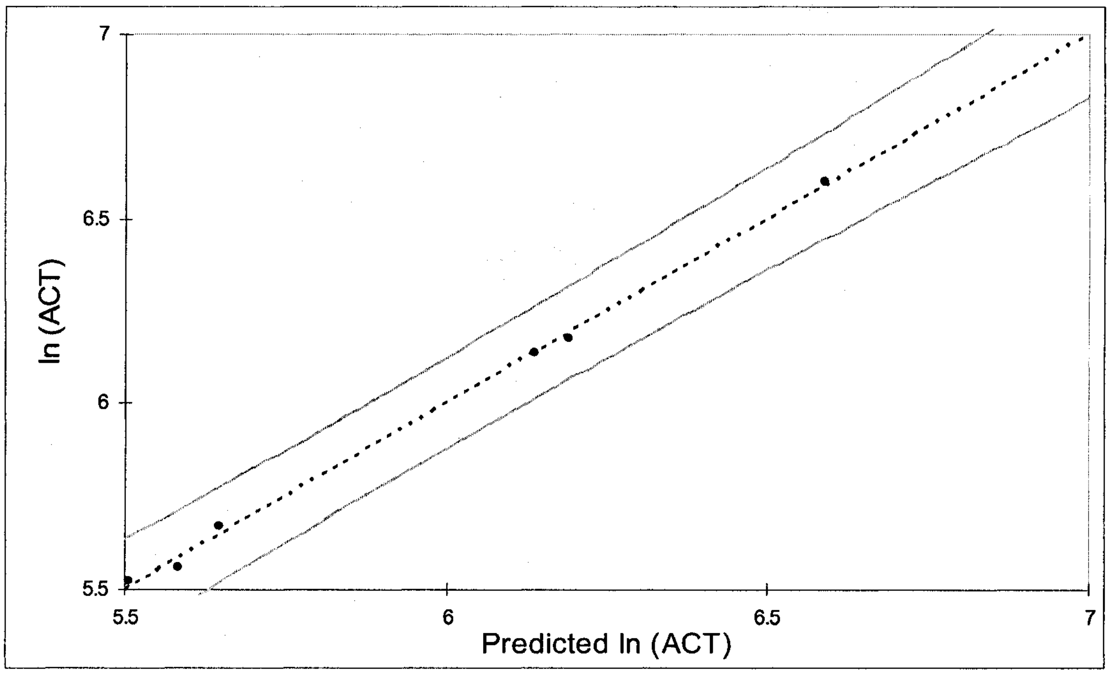

Figure 19: SPC chart of jackknife sample B

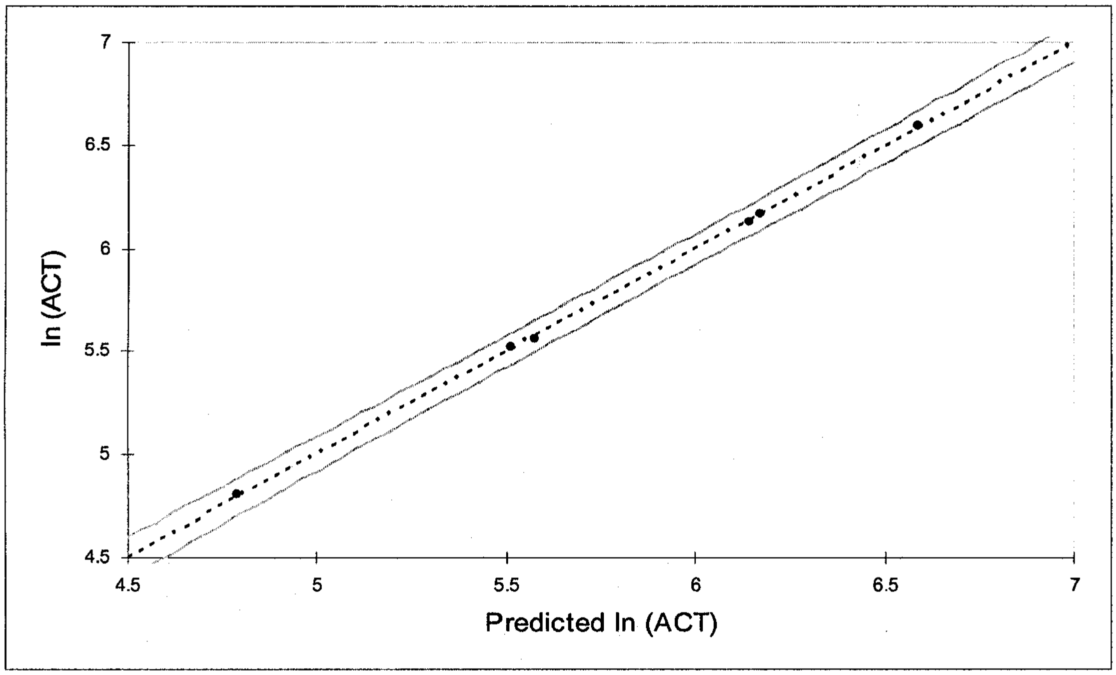

Figure 20: SPC chart of jackknife sample C

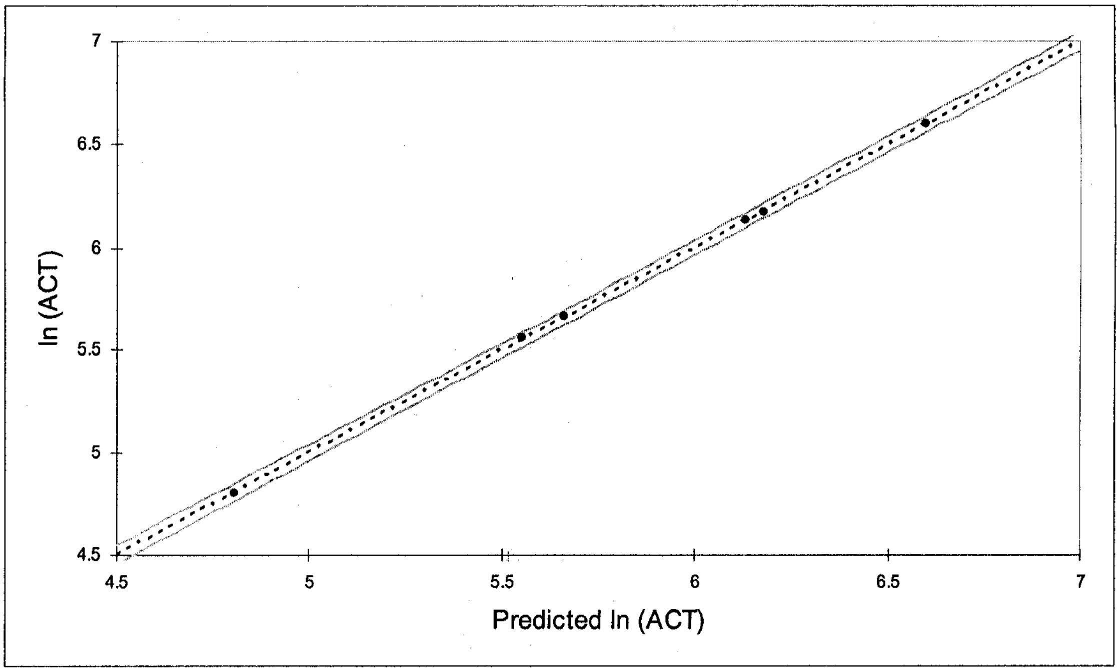

Figure 21: SPC chart of jackknife sample D

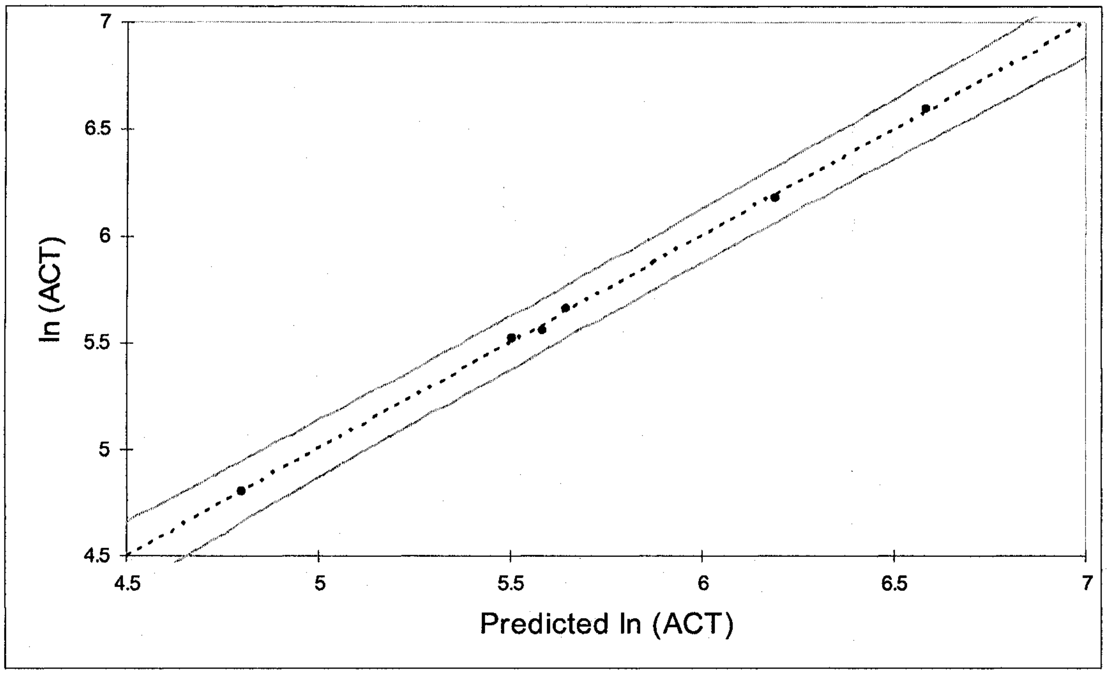

Figure 22: SPC chart of jackknife sample E

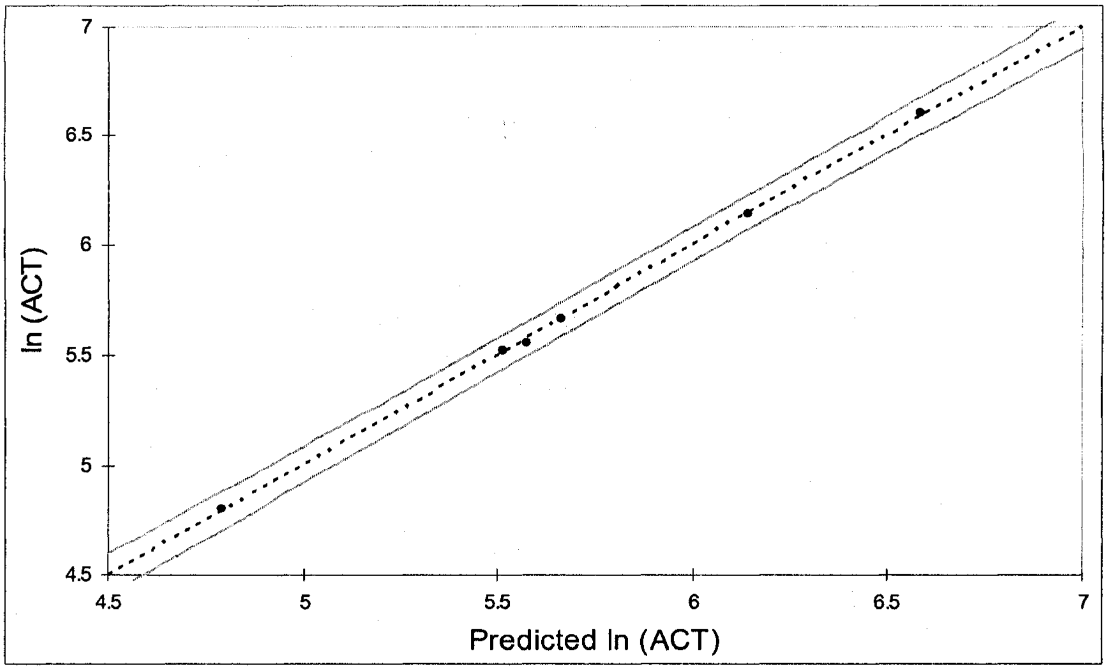

Figure 23: SPC chart of jackknife sample F

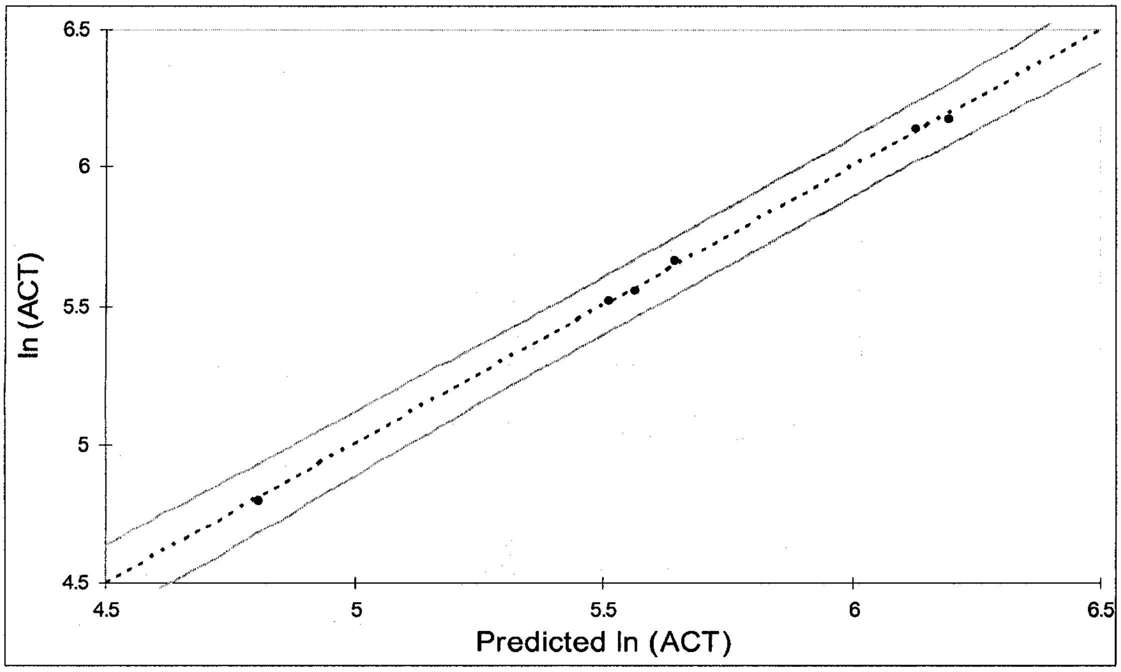

Figure 24: SPC chart of jackknife sample G

APPENDIX C: RESIDUAL PLOTS OF D1 (NO CONCURRENCY)

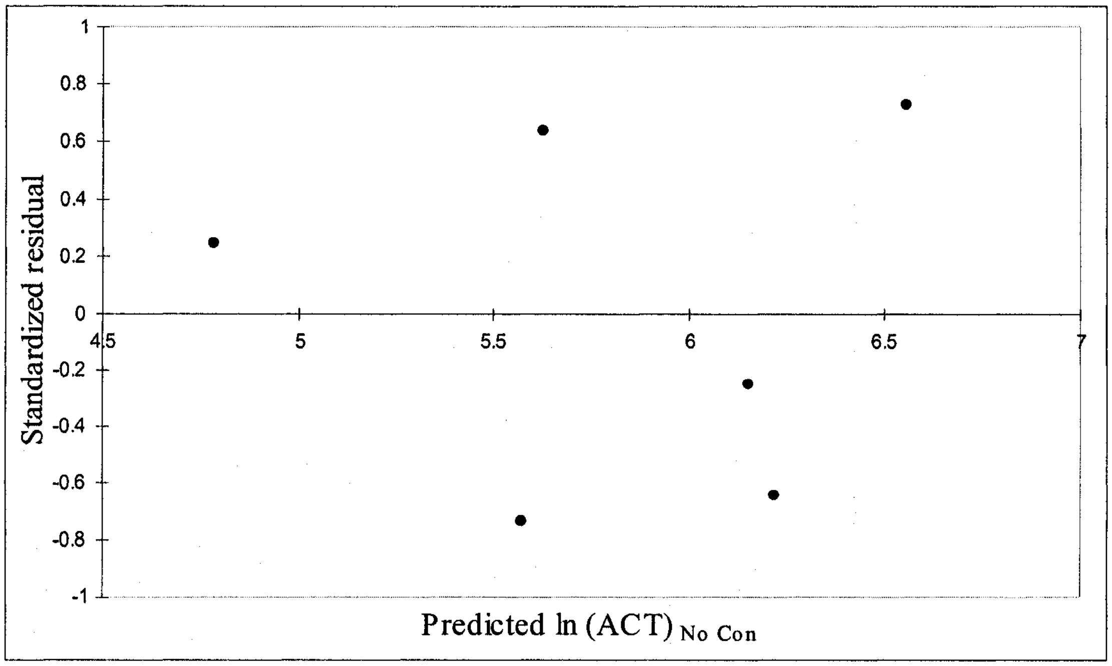

Figure 25: Residual plot of jackknife sample A (no concurrency)

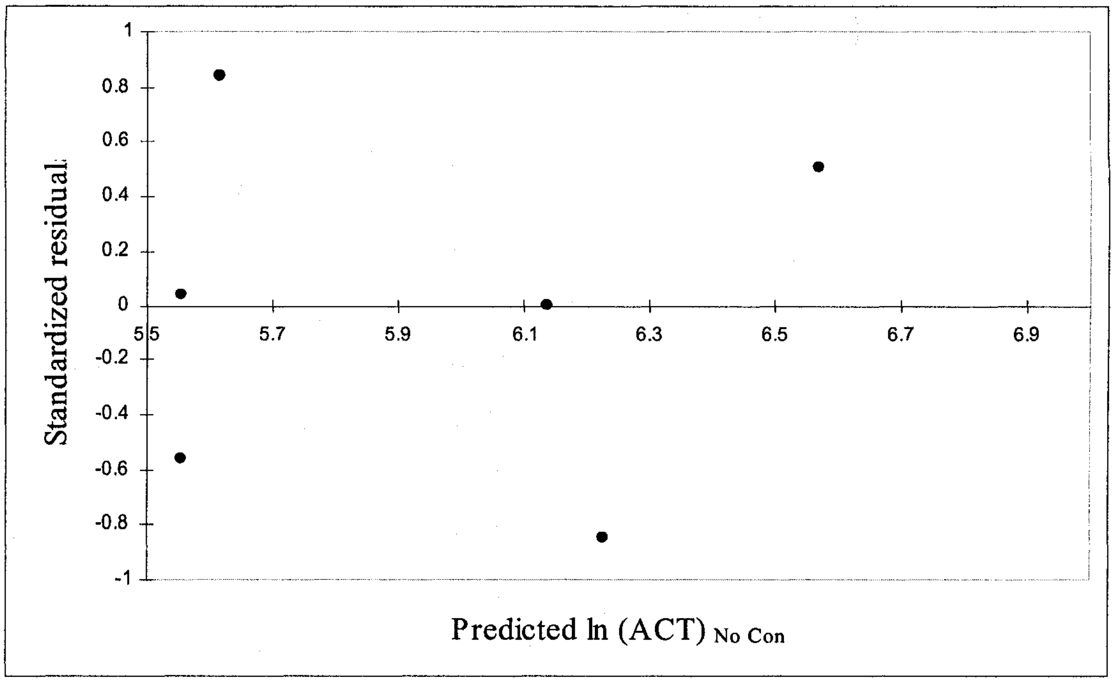

Figure 26: Residual plot of jackknife sample B (no concurrency)

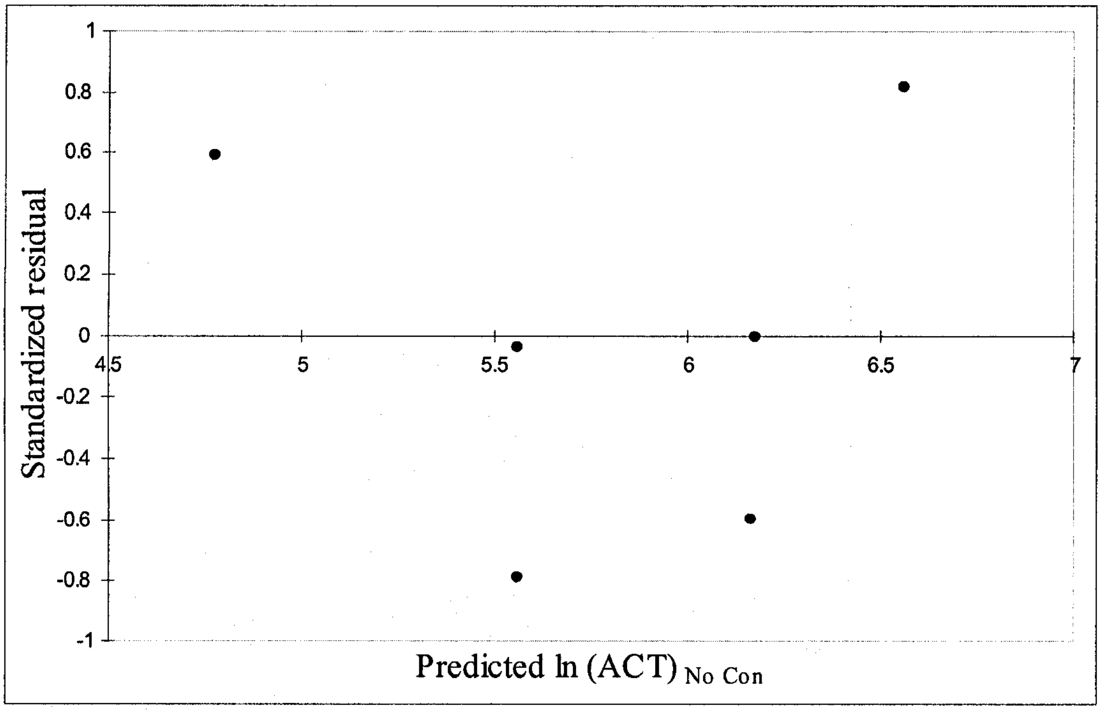

Figure 27: Residual plot of jackknife sample C (no concurrency)

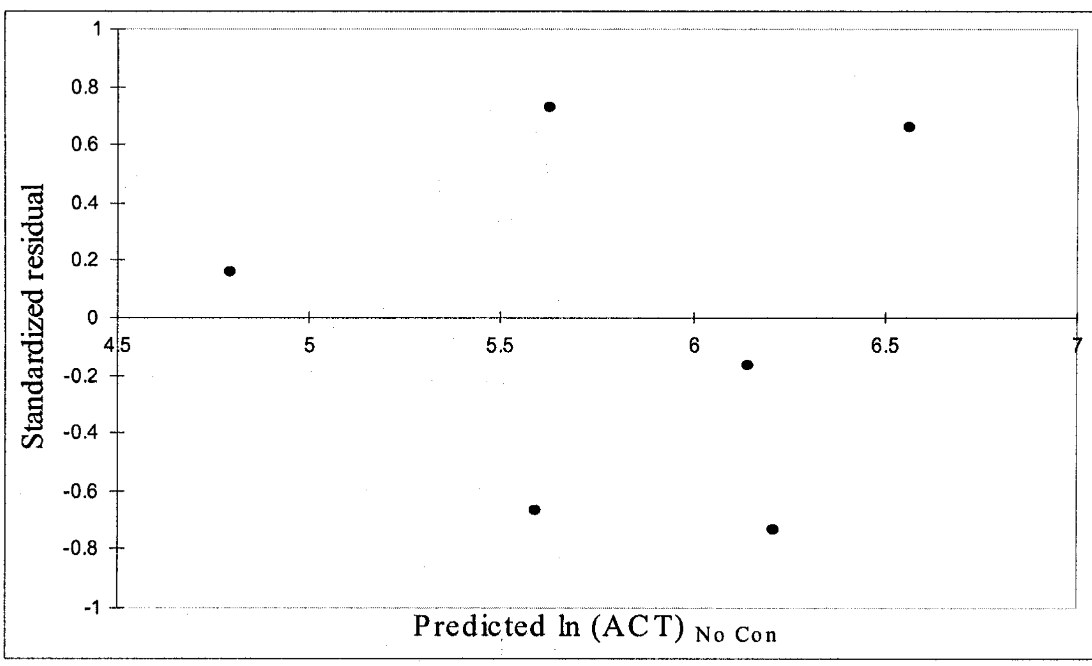

Figure 28: Residual plot of jackknife sample D (no concurrency)

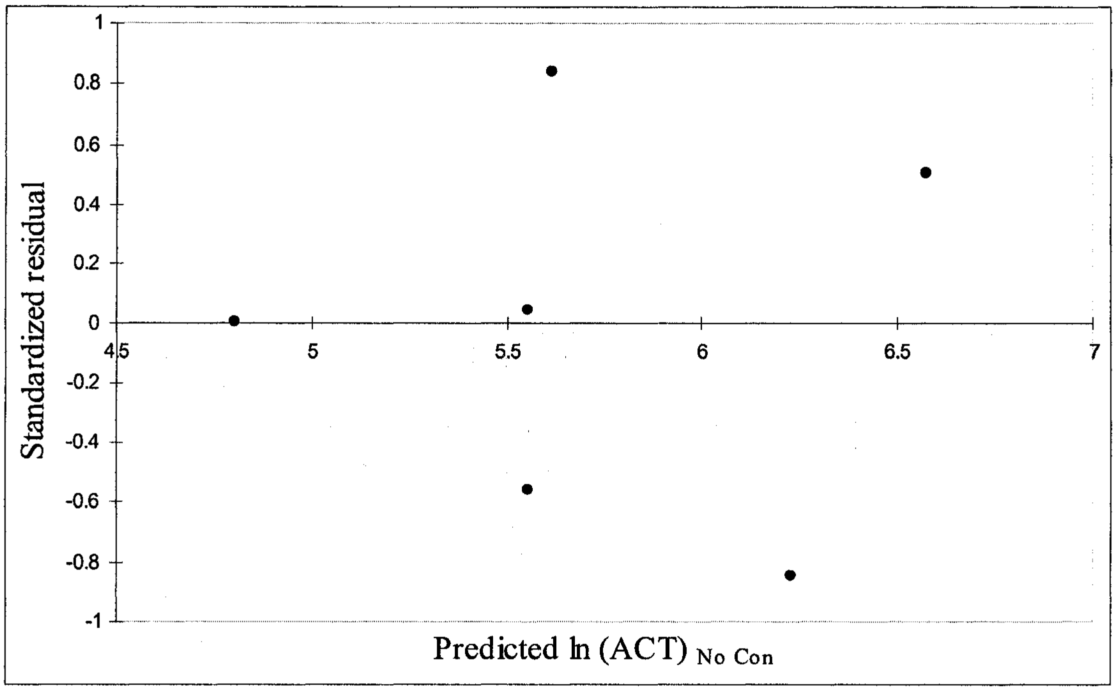

Figure 29: Residual plot of jackknife sample E (no concurrency)

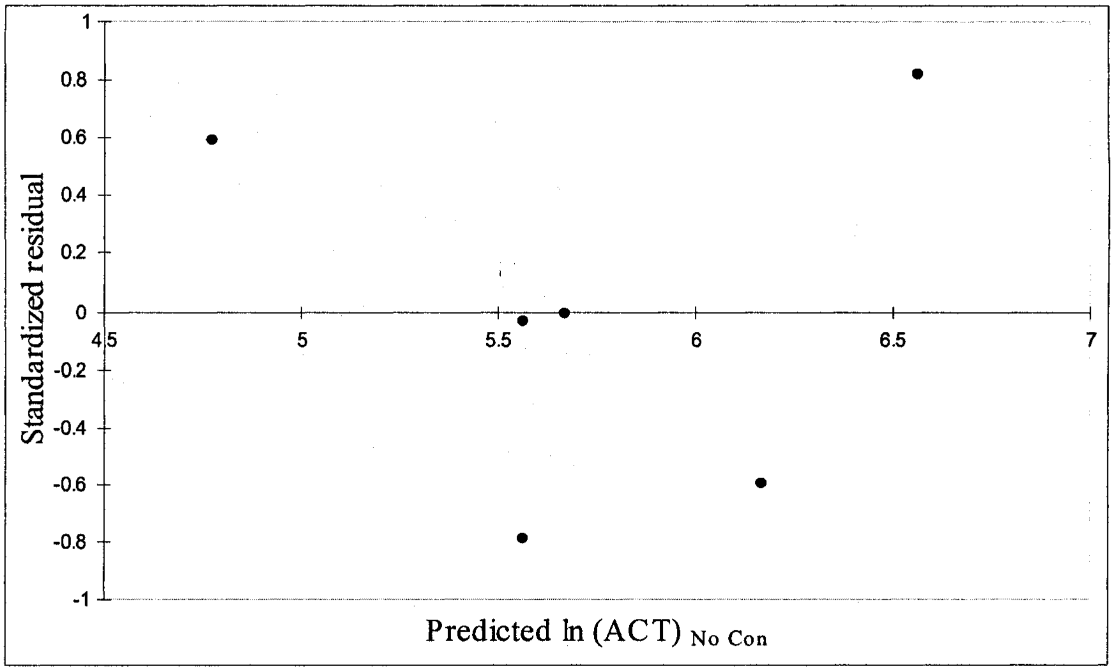

Figure 30: Residual plot of jackknife sample F (no concurrency)

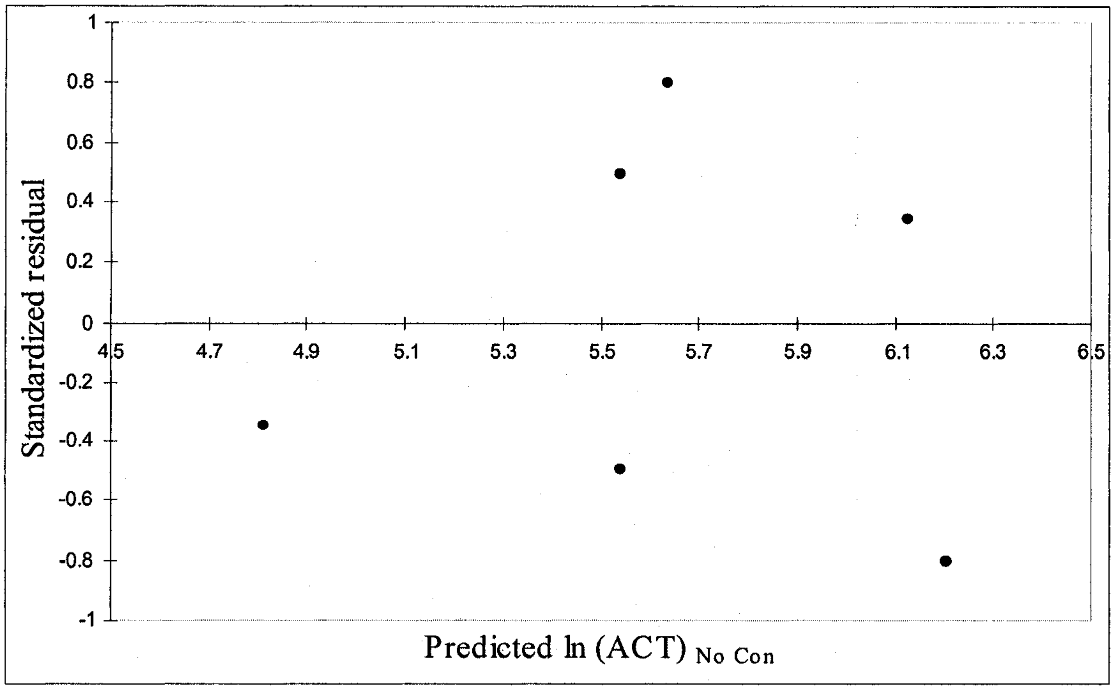

Figure 31: Residual plot of jackknife sample G (no concurrency)

APPENDIX D: SPC CHARTS OF D1 (NO CONCURRENCY)

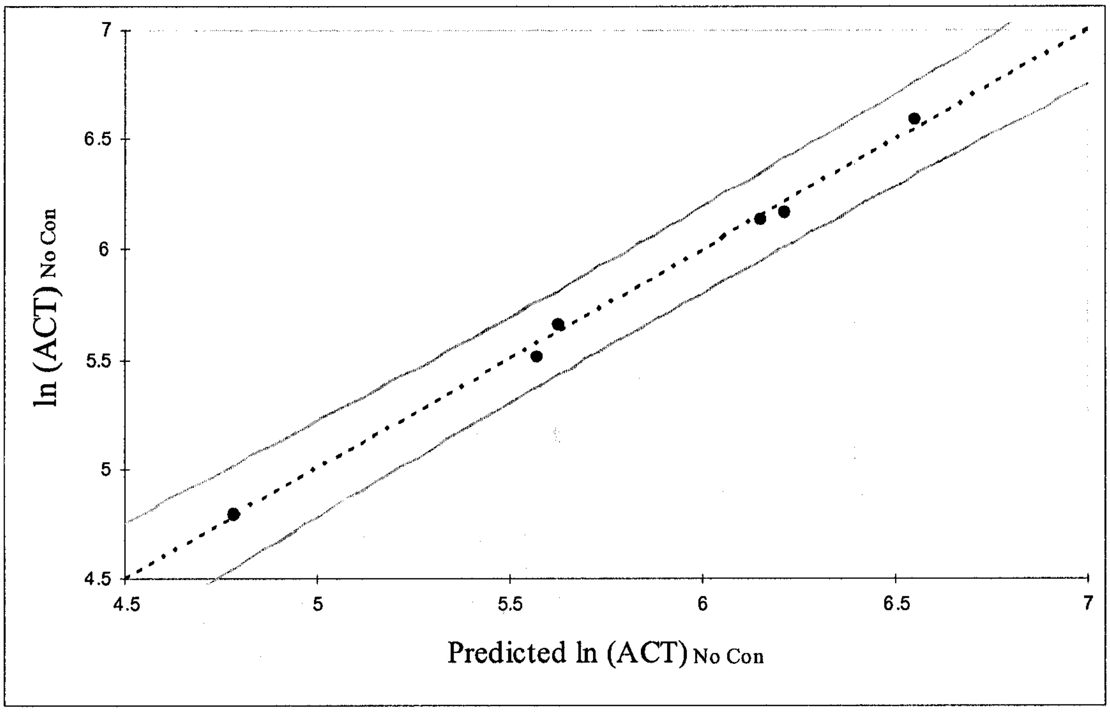

Figure 32: SPC chart of jackknife sample A (no concurrency)

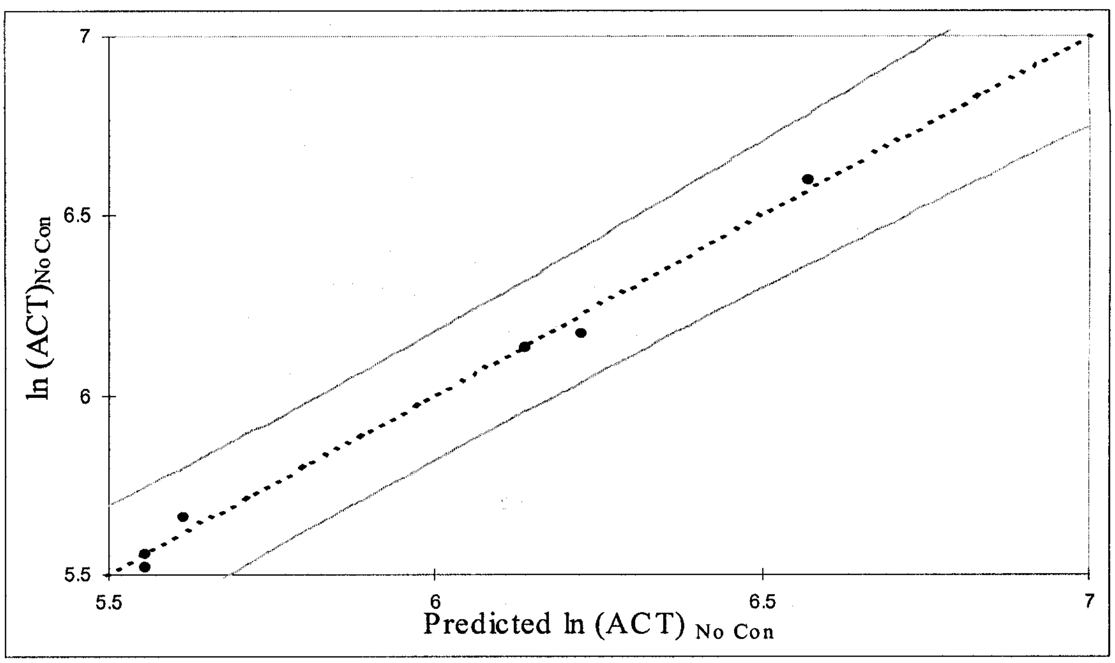

Figure 33: SPC chart of jackknife sample B (no concurrency)

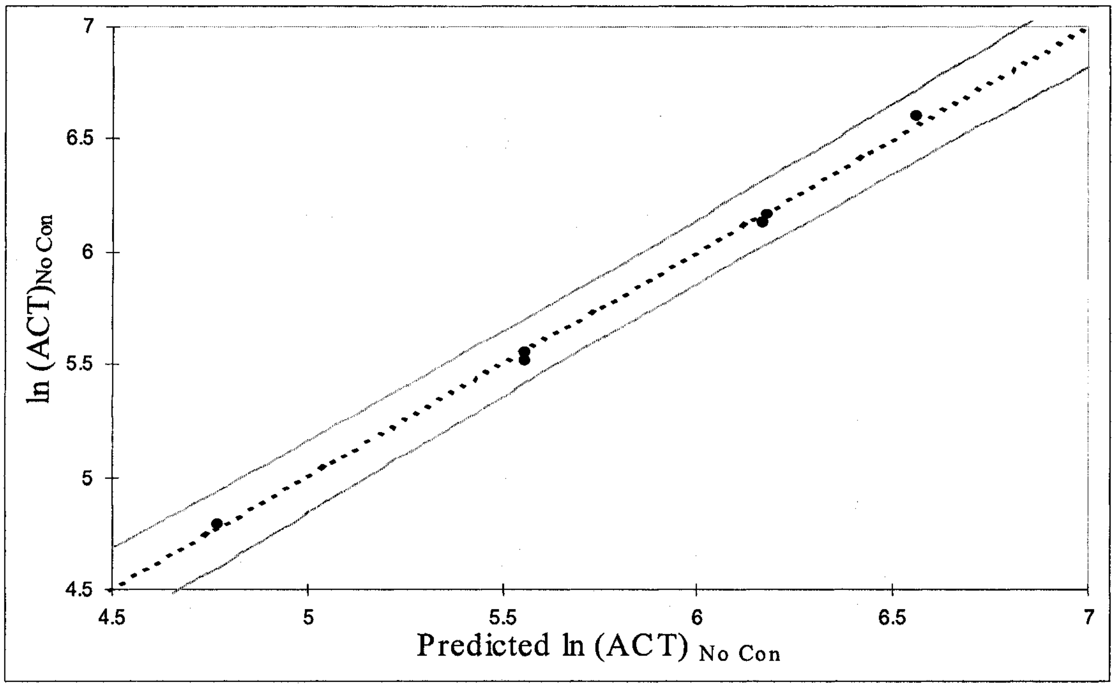

Figure 34: SPC chart of jackknife sample C (no concurrency)

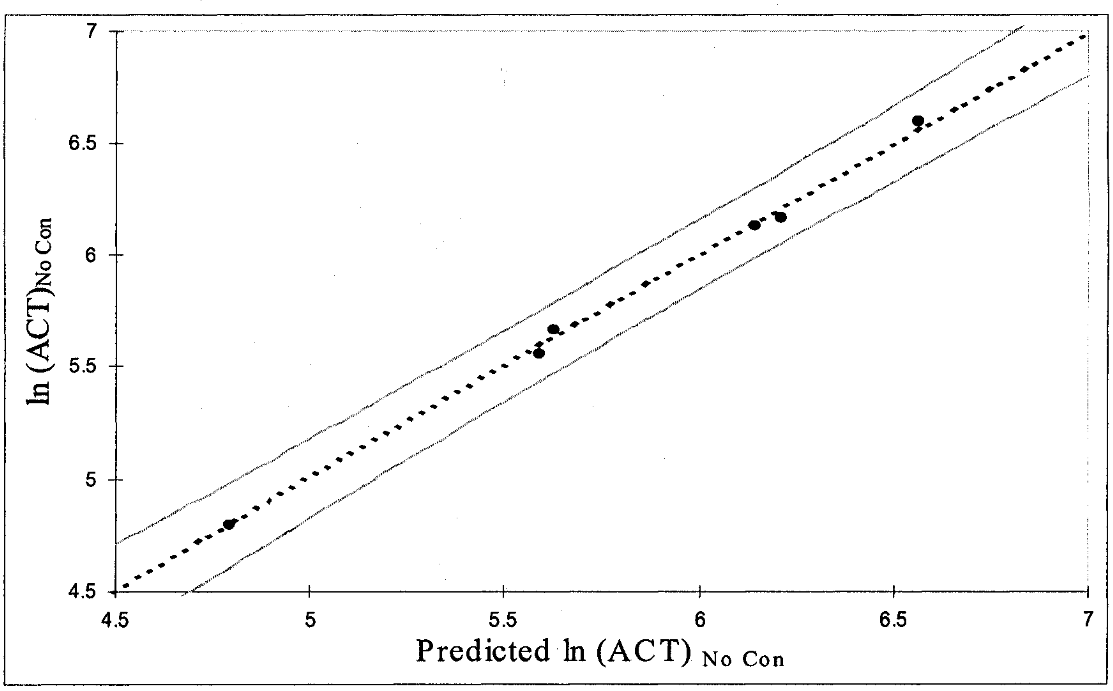

Figure 35: SPC chart of jackknife sample D (no concurrency)

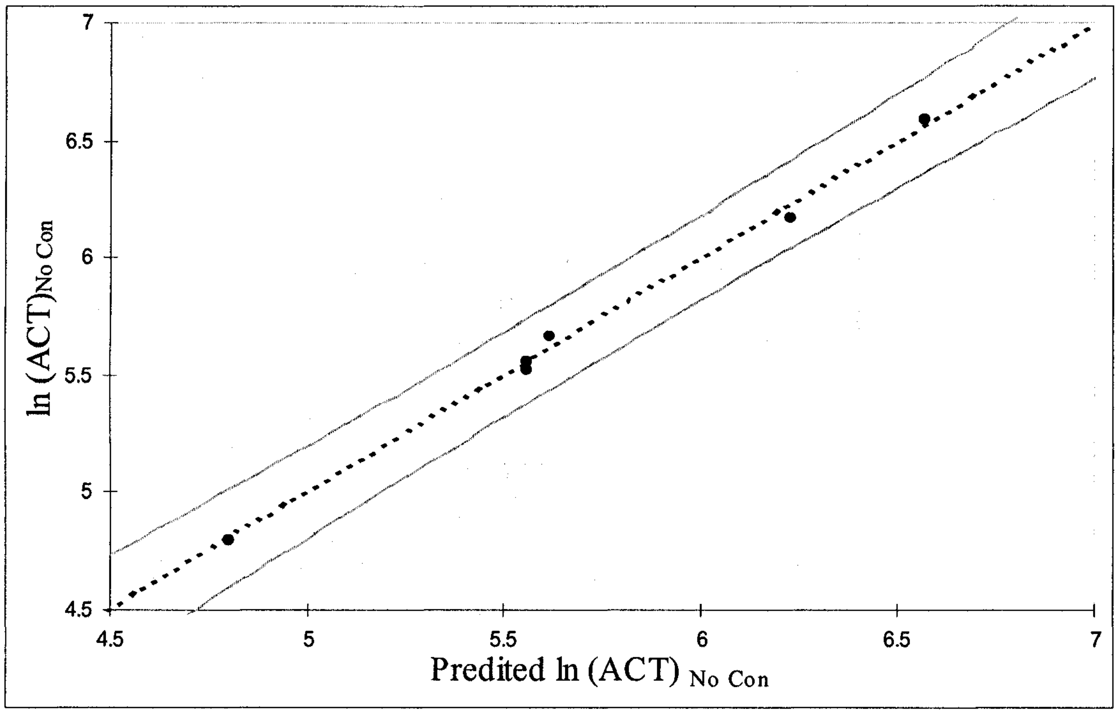

Figure 36: SPC chart of jackknife sample E (no concurrency)

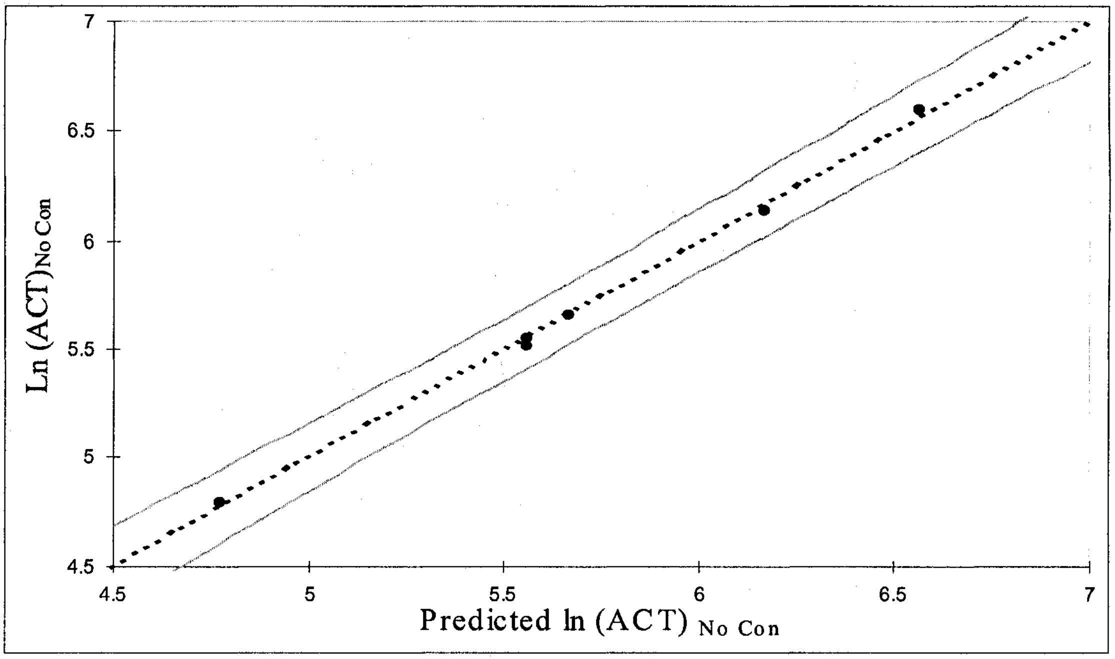

Figure 37: SPC chart of jackknife sample F (no concurrency)

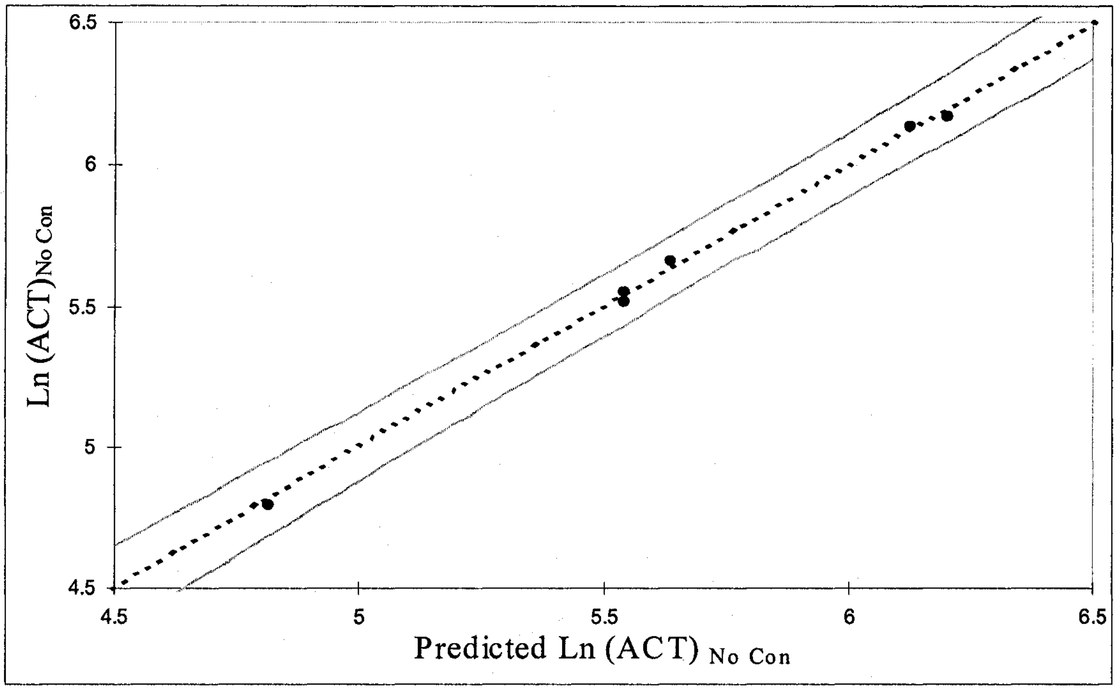

Figure 38: SPC chart of jackknife sample G (no concurrency)

APPENDIX E: ANALYSIS OF DEPARTMENT 2

Table 24: Data of D2

DJ	TD	DC	Con	DE	ACT
A	1	1	0.73	3.00	182.45
B	2	2	0.67	2.80	784.87
C	2	3	0.64	3.00	825.71
D	1	1	0.83	2.00	218.56
E	2	2	0.73	2.07	816.78
F	2	3	0.71	2.43	864.88
G	1	1	0.66	1.70	228.07

Table 25: ln of data of D2

DJ	ln (TD)	ln (DC)	ln (Con)	ln (DE)	ln (ACT)
A	0	0	-0.313	1.099	5.206
B	0.693	0.693	-0.405	1.030	6.666
C	0.693	1.099	-0.439	1.099	6.716
D	0	0	-0.182	0.693	5.387
E	0.693	0.693	-0.313	0.728	6.705
F	0.693	1.099	-0.341	0.888	6.763
G	0	0	-0.418	0.531	5.430

Table 26: Regression coefficients of D2

Constants	All	$Jack_A$	$Jack_B$	$Jack_C$	$Jack_D$	$Jack_E$	$Jack_F$	$Jack_G$	$COEF_{JACK}$
$\ln(a_o)$	5.583	5.504	5.644	5.598	5.428	5.644	5.598	5.605	5.575
a_1	1.762	1.727	1.604	1.789	1.814	1.908	1.732	1.690	1.752
a_2	0.223	0.180	0.339	0.200	0.191	0.147	0.257	0.231	0.221
a_3	-0.063	-0.050	0.023	-0.035	-0.401	0.023	-0.035	-0.441	-0.131
a_4	-0.337	-0.180	-0.383	-0.346	-0.315	-0.383	-0.346	-0.466	-0.345

Table 27: Residuals of D2

DJ	PRED ln(ACT)	ln(ACT)	Residual (Error)
A	5.2359	5.2065	-0.0295
B	6.6393	6.6655	0.0262
C	6.7094	6.7162	0.0068
D	5.3589	5.3871	0.0281
E	6.7315	6.7054	-0.0262
F	6.7694	6.7626	-0.0068
G	5.4459	5.4297	-0.0162

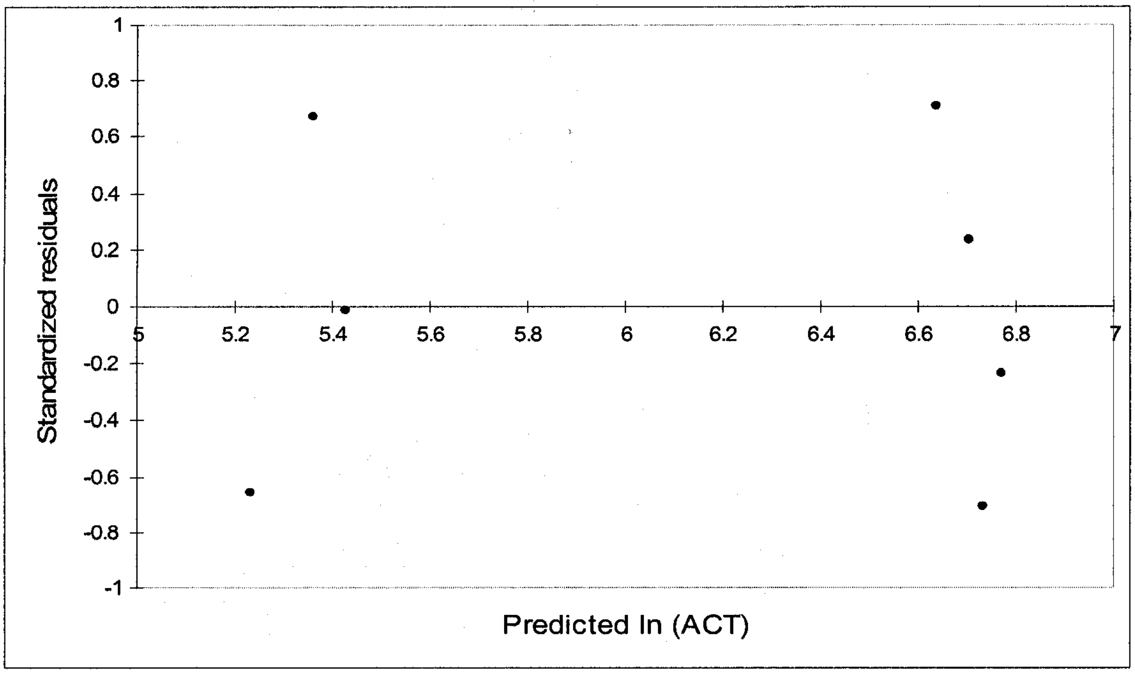

Figure 39: Residual plot of D2

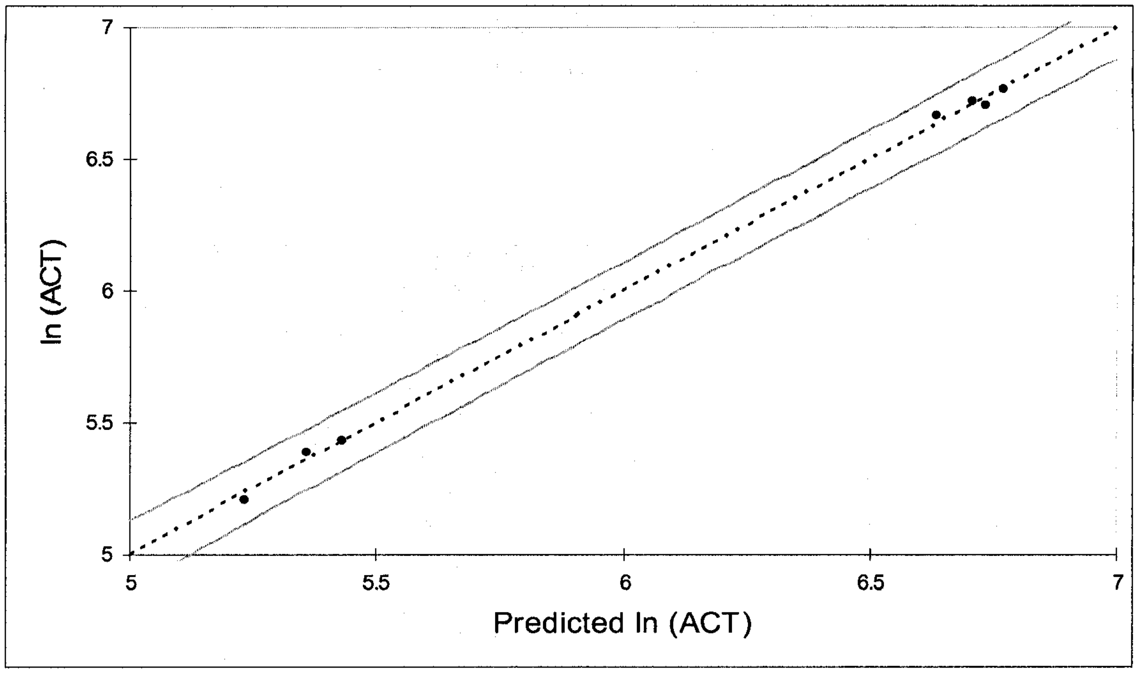

Figure 40: SPC chart of D2

Table 28: R^2 values of D2

Sample	R^2
All	0.999
$Jack_A$	1.000
$Jack_B$	1.000
$Jack_C$	0.999
$Jack_D$	1.000
$Jack_E$	1.000
$Jack_F$	0.999
$Jack_G$	0.999
R^2min	0.999
r_{R^2min}	0.999

$$\hat{E}_{D2} = 2.63x10^2\, TD^{1.75}\, DC^{0.221}\, Con^{-0.131}\, DE^{-0.345}$$

Table 29: Relative errors of D2

DJ	ACT (hours)	PRED (hours)	Relative Error (%)
A	182.45	187.91	2.99
B	784.87	764.59	2.58
C	825.71	820.12	0.68
D	218.56	212.50	2.77
E	816.78	838.44	2.65
F	864.88	870.78	0.68
G	228.07	231.79	1.63

Table 30: Correlation matrix of D2

Variables	ln (TD)	ln (DC)	ln (Con)	ln (DE)	ln (ACT)
ln (TD)	1.000	0.945	-0.427	0.389	0.995
ln (DC)	0.945	1.000	-0.451	0.436	0.950
ln (Con)	-0.427	-0.451	1.000	-0.276	-0.427
ln (DE)	0.389	0.436	-0.276	1.000	0.312
ln (ACT)	0.995	0.950	-0.427	0.312	1.000

Table 31: Significant factors of D2

Factor	All	$Jack_A$	$Jack_B$	$Jack_C$	$Jack_D$	$Jack_E$	$Jack_F$	$Jack_G$
ln (a_o)	0.000	0.004	0.009	0.015	0.008	0.009	0.015	0.023
ln (TD)	0.006	0.015	0.058	0.074	0.023	0.047	0.078	0.196
ln (DC)	0.157	0.110	0.197	0.436	0.161	0.357	0.396	0.368
ln (Con)	0.788	0.568	0.913	0.927	0.236	0.913	0.927	0.895
ln (DE)	0.053	0.144	0.114	0.205	0.079	0.114	0.205	0.699

Table 32: Regression coefficients of D2 without concurrency

Constants	All	$Jack_A$	$Jack_B$	$Jack_C$	$Jack_D$	$Jack_E$	$Jack_F$	$Jack_G$	$COEF_{JACK}$
$\ln(a_o)$	5.600	5.517	5.637	5.609	5.582	5.637	5.609	5.580	5.596
a_1	1.762	1.727	1.608	1.793	1.778	1.904	1.728	1.773	1.759
a_2	0.228	0.183	0.334	0.198	0.225	0.148	0.264	0.223	0.225
a_4	-0.334	-0.177	-0.382	-0.346	-0.324	-0.382	-0.346	-0.317	-0.325

Table 33: Residuals of D2 without concurrency

DJ	PRED ln(ACT)	ln(ACT)	Residual (Error)
A	5.2389	5.2065	-0.0325
B	6.6364	6.6655	0.0291
C	6.7052	6.7162	0.0111
D	5.3706	5.3871	0.0164
E	6.7345	6.7054	-0.0291
F	6.7736	6.7626	-0.0111
G	5.4234	5.4297	0.0062

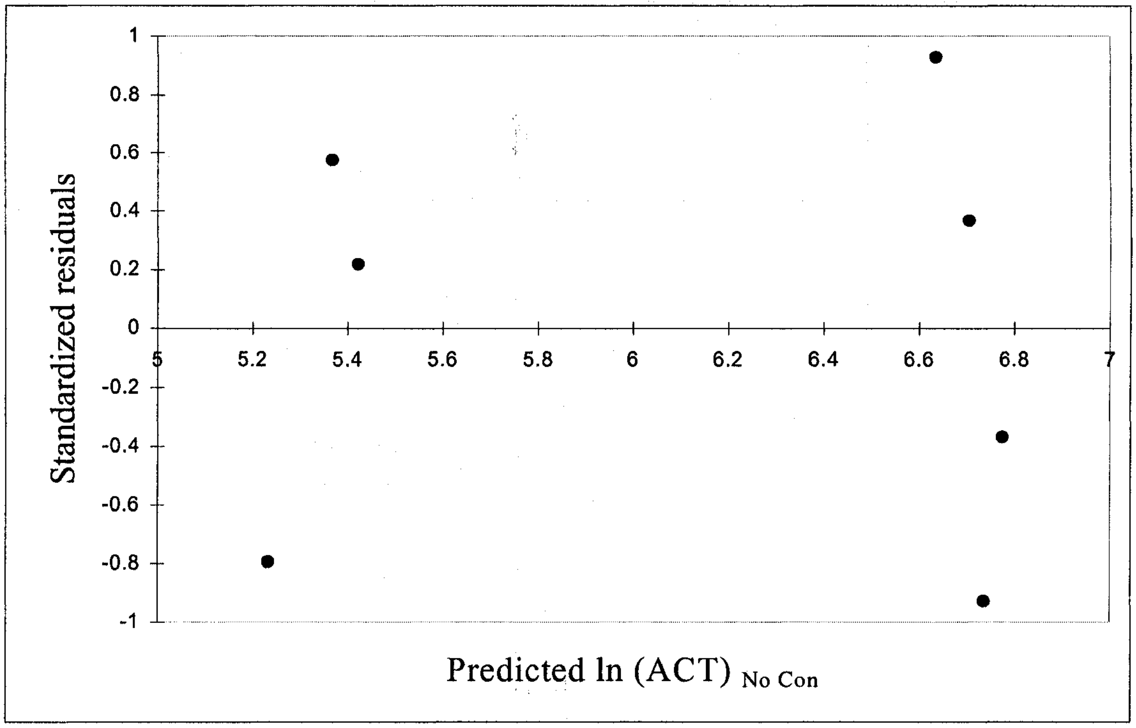

Figure 41: Residuals of D2 without concurrency

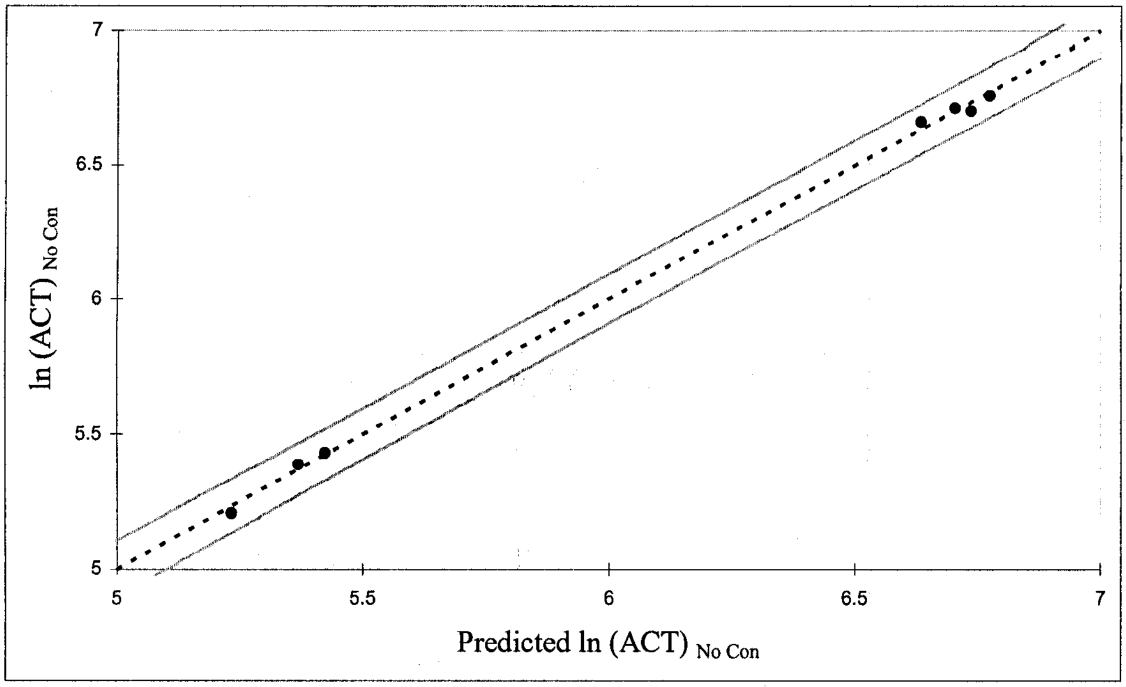

Figure 42: SPC chart of D2 without concurrency

Table 34: R^2 values of D2 without concurrency

Sample	R^2
All	0.999
$Jack_A$	1.000
$Jack_B$	1.000
$Jack_C$	0.999
$Jack_D$	0.999
$Jack_E$	1.000
$Jack_F$	0.999
$Jack_G$	0.999
R^2min	0.999
r_{R^2min}	0.999

$$\hat{E}_{D2_{\text{No Con}}} = 2.69\text{x}10^2\ TD^{1.76}\ DC^{0.225}\ DE^{-0.325}$$

Table 35: Relative errors of D2 without concurrency

DJ	ACT (hours)	PRED (hours)	Relative Error (%)
A	182.45	188.47	3.30
B	784.87	762.33	2.87
C	825.71	816.63	1.10
D	218.56	215.00	1.63
E	816.78	840.93	2.96
F	864.88	874.49	1.11
G	228.07	226.66	0.62

Table 36: Correlation matrix of D2 without concurrency

Variables	ln (TD)	ln (DC)	ln (DE)	ln (ACT)
Ln (TD)	1.000	0.945	0.389	0.995
Ln (DC)	0.945	1.000	0.436	0.950
Ln (DE)	0.389	0.436	1.000	0.312
Ln (ACT)	0.995	0.950	0.312	1.000

Table 37: Significant factors of D2 without concurrency

Factor	All	$Jack_A$	$Jack_B$	$Jack_C$	$Jack_D$	$Jack_E$	$Jack_F$	$Jack_G$
ln (a_o)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ln (TD)	0.001	0.001	0.004	0.007	0.005	0.003	0.007	0.006
ln (DC)	0.072	0.023	0.045	0.229	0.137	0.153	0.167	0.160
ln (DE)	0.016	0.042	0.016	0.052	0.052	0.016	0.052	0.090

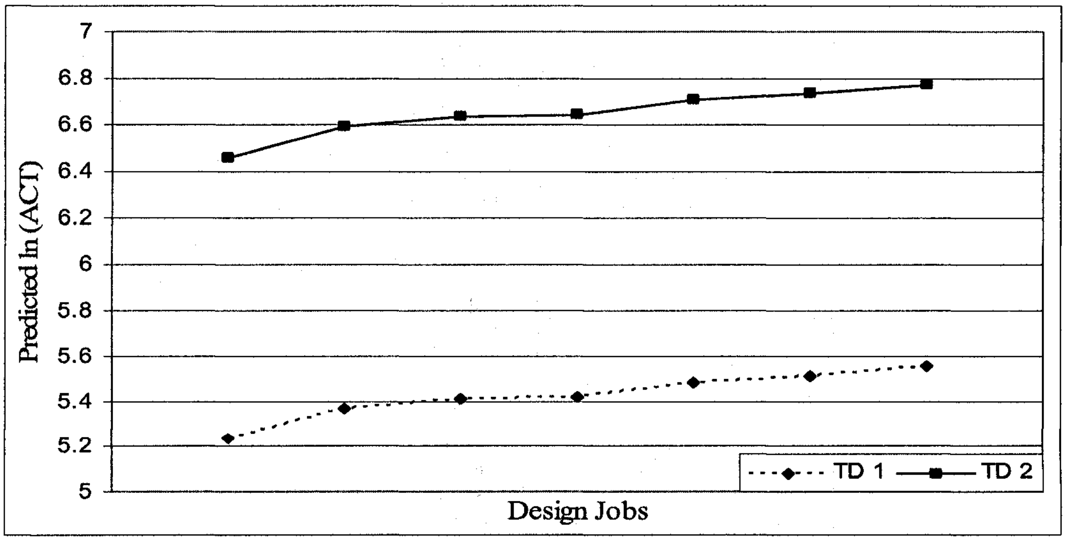

Figure 43: Impact of the type of design on effort for D2

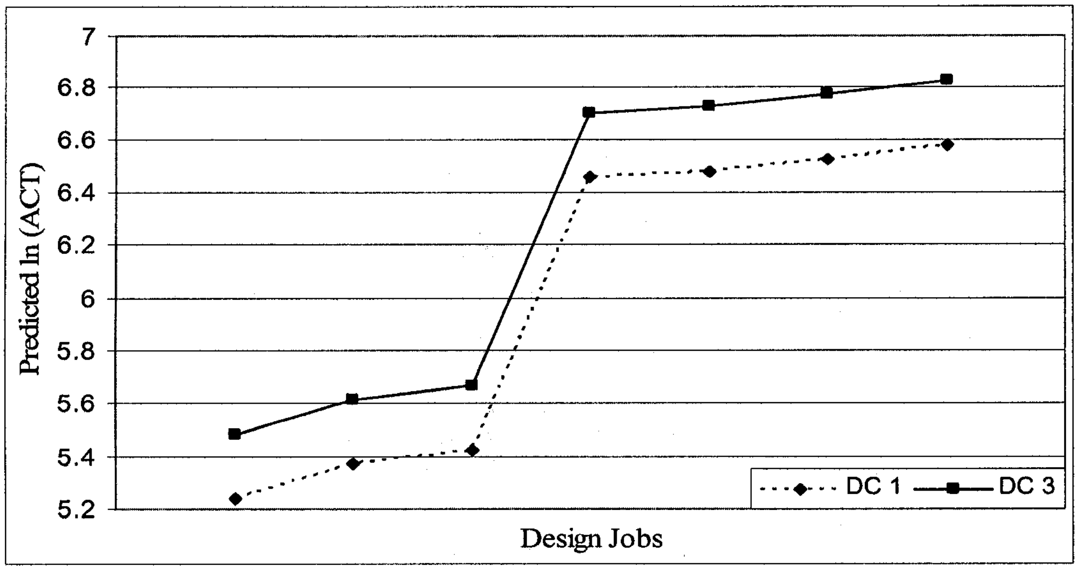

Figure 44: Impact of the degree of change on effort for D2

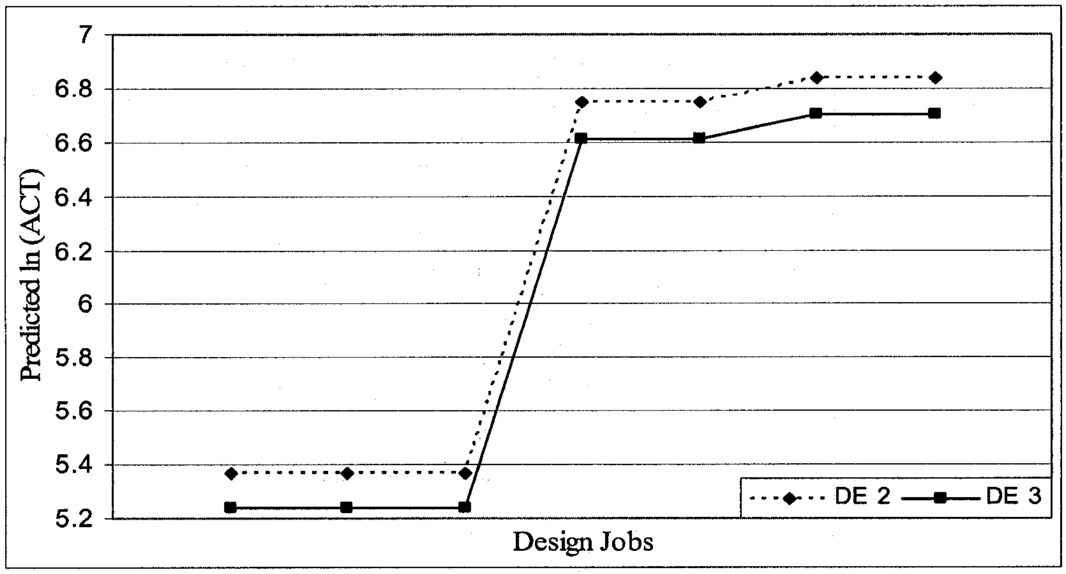

Figure 45: Impact of the experience of departmental personnel on effort for D2

APPENDIX F: ANALYSIS OF DEPARTMENT 3

Table 38: Data of D3

DJ	TD	DC	Con	DE	ACT
A	1	1	0.73	2.05	356.94
B	2	2	0.71	2.24	662.15
C	2	3	0.68	3.00	455.02
D	1	1	0.77	3.00	224.34
E	2	2	0.82	1.85	712.91
F	2	3	0.72	2.65	493.73
G	1	1	0.63	2.70	294.12

Table 39: ln of data of D3

DJ	ln (TD)	ln (DC)	ln (Con)	ln (DE)	ln (ACT)
A	0	0	-0.313	0.718	5.878
B	0.693	0.693	-0.345	0.806	6.495
C	0.693	1.099	-0.389	1.099	6.120
D	0	0	-0.266	1.099	5.413
E	0.693	0.693	-0.199	0.615	6.569
F	0.693	1.099	-0.326	0.975	6.202
G	0	0	-0.466	0.993	5.684

Table 40: Regression coefficients of D3

Constants	All	$Jack_A$	$Jack_B$	$Jack_C$	$Jack_D$	$Jack_E$	$Jack_F$	$Jack_G$	$COEF_{JACK}$
ln (ao)	6.41	6.32	6.44	6.41	6.48	6.44	6.41	6.38	6.411
a1	1.20	1.30	1.11	1.20	1.06	1.21	1.19	1.21	1.182
a2	-0.203	-0.264	-0.144	-0.206	-0.039	-0.213	-0.196	-0.224	-0.184
a3	-0.820	-0.815	-0.771	-0.818	-1.393	-0.771	-0.818	-0.934	-0.903
a4	-1.10	-1.02	-1.12	-1.10	-1.46	-1.12	-1.10	-1.11	-1.148

Table 41: Residuals of D3

DJ	PRED ln(ACT)	ln(ACT)	Residual (Error)
A	5.8695	5.8776	0.0081
B	6.4884	6.4955	0.0071
C	6.1184	6.1203	0.0019
D	5.3900	5.4132	0.0231
E	6.5764	6.5694	-0.0071
F	6.2039	6.2020	-0.0019
G	5.6920	5.6840	-0.0080

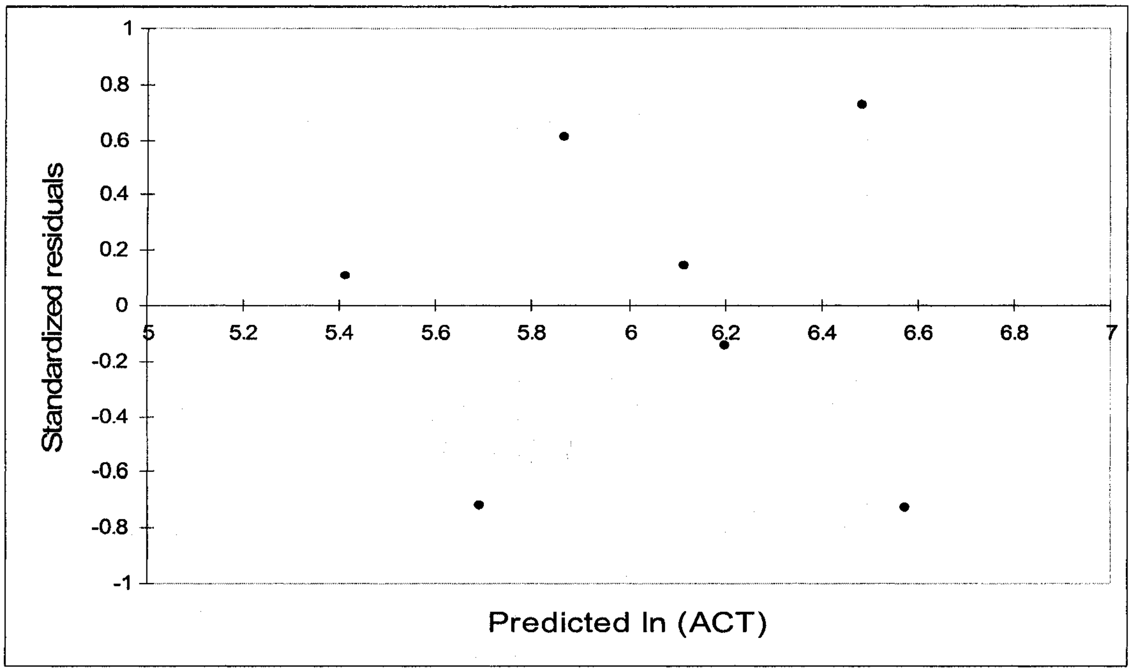

Figure 46: Residual plot of D3

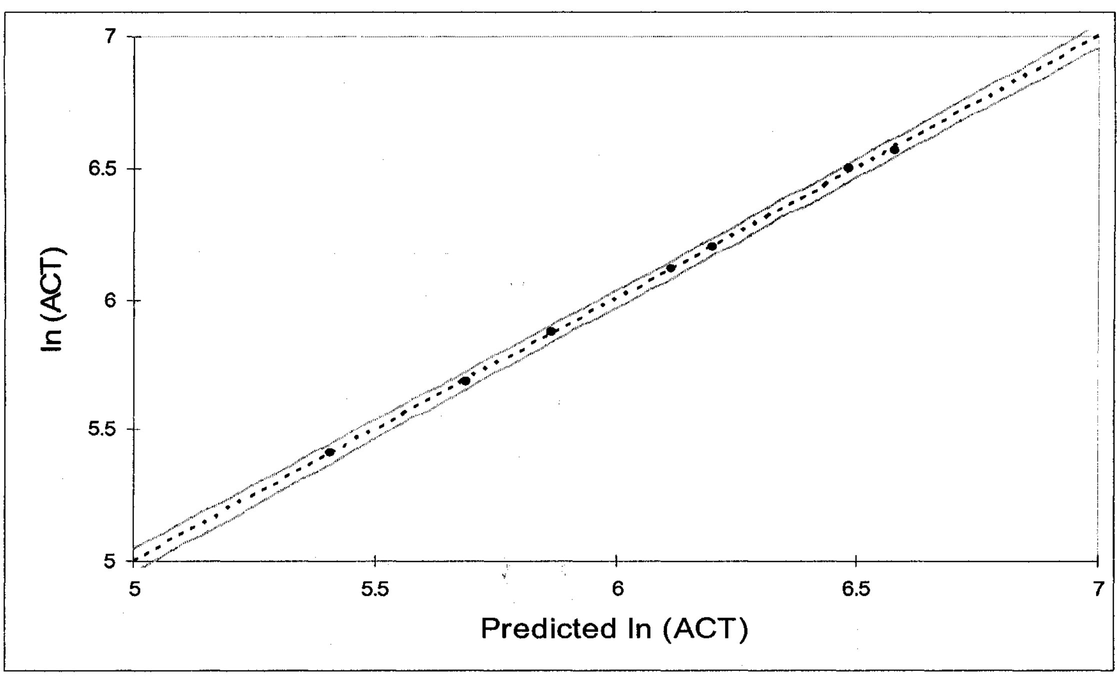

Figure 47: SPC chart of D3

Table 42: R^2 values of D3

Sample	R^2
All	0.9997
$Jack_A$	1.0000
$Jack_B$	0.9999
$Jack_C$	0.9997
$Jack_D$	0.9999
$Jack_E$	0.9999
$Jack_F$	0.9997
$Jack_G$	1.0000
R^2min	0.9997
r_{R^2min}	0.9999

$$\hat{E}_{D3} = 6.09\text{x}10^2\ \text{TD}^{1.18}\,\text{DC}^{-0.184}\ \text{Con}^{-0.903}\,\text{DE}^{-1.15}$$

Table 43: Relative errors of D3

DJ	ACT (hours)	PRED (hours)	Relative Error (%)
A	356.94	354.06	0.81
B	662.15	657.48	0.70
C	455.02	454.14	0.19
D	224.34	219.21	2.29
E	712.91	717.97	0.71
F	493.73	494.69	0.19
G	294.12	296.49	0.81

Table 44: Correlation matrix of D3

Variables	ln (TD)	ln (DC)	ln (DE)	ln (Con)	ln (ACT)
Ln (TD)	1.000	0.945	0.210	-0.177	0.873
Ln (DC)	0.945	1.000	0.065	0.062	0.708
Ln (DE)	0.210	0.065	1.000	-0.518	0.295
Ln (Con)	-0.177	0.062	-0.518	1.000	-0.612
Ln (ACT)	0.873	0.708	0.295	-0.612	1.000

Table 45: Significant factors of D3

Factor	All	$Jack_A$	$Jack_B$	$Jack_C$	$Jack_D$	$Jack_E$	$Jack_F$	$Jack_G$
ln (a_o)	0.000	0.003	0.002	0.005	0.006	0.002	0.005	0.001
ln (TD)	0.002	0.019	0.029	0.044	0.061	0.018	0.052	0.007
ln (DC)	0.041	0.060	0.148	0.186	0.787	0.071	0.229	0.028
ln (Con)	0.007	0.021	0.036	0.075	0.168	0.036	0.075	0.016
ln (DE)	0.001	0.017	0.013	0.032	0.101	0.013	0.032	0.005

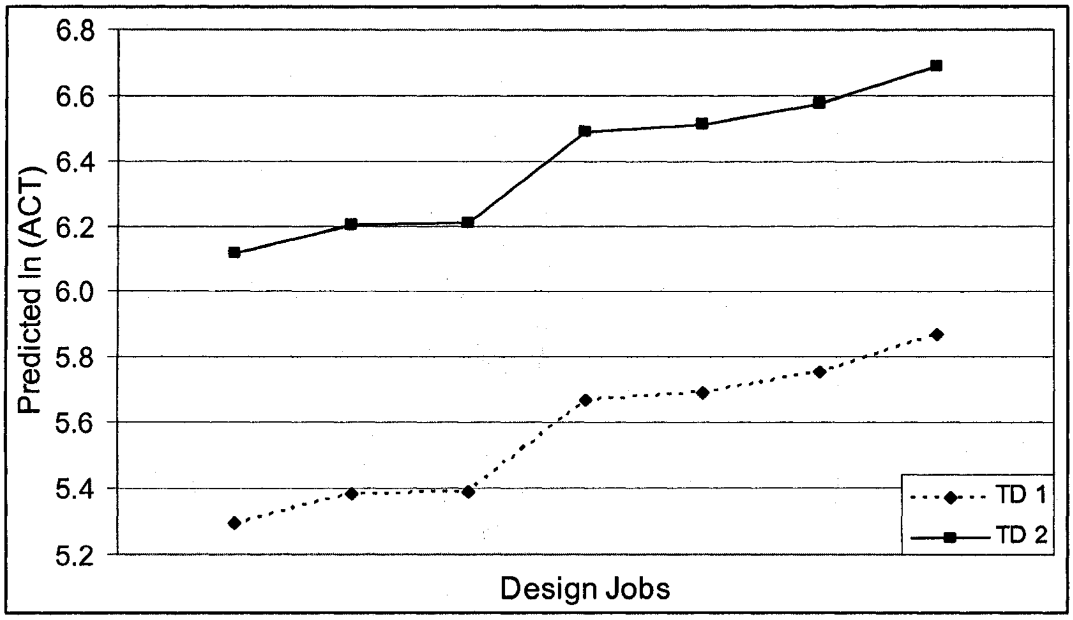

Figure 48: Impact of the type of design on effort for D3

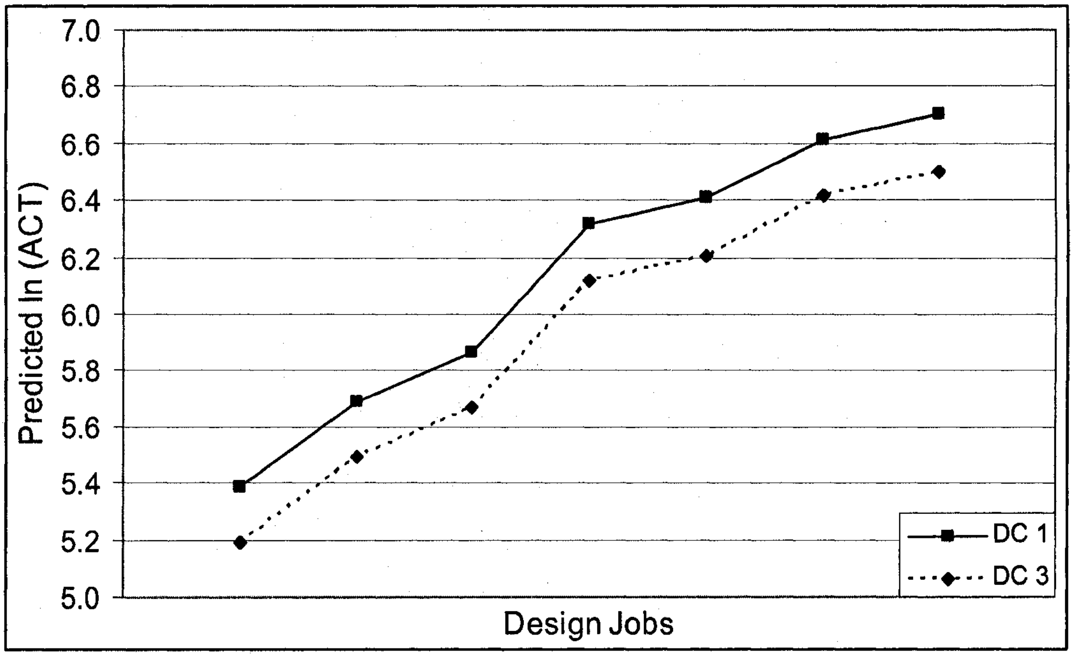

Figure 49: Impact of the degree of change on effort for D3

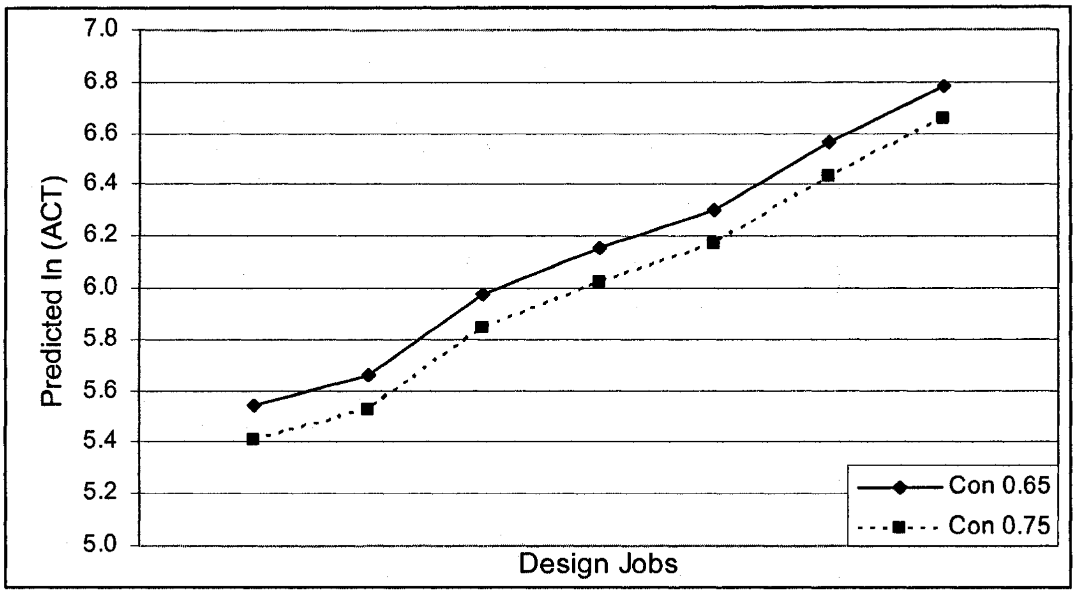

Figure 50: Impact of concurrency on effort for D3

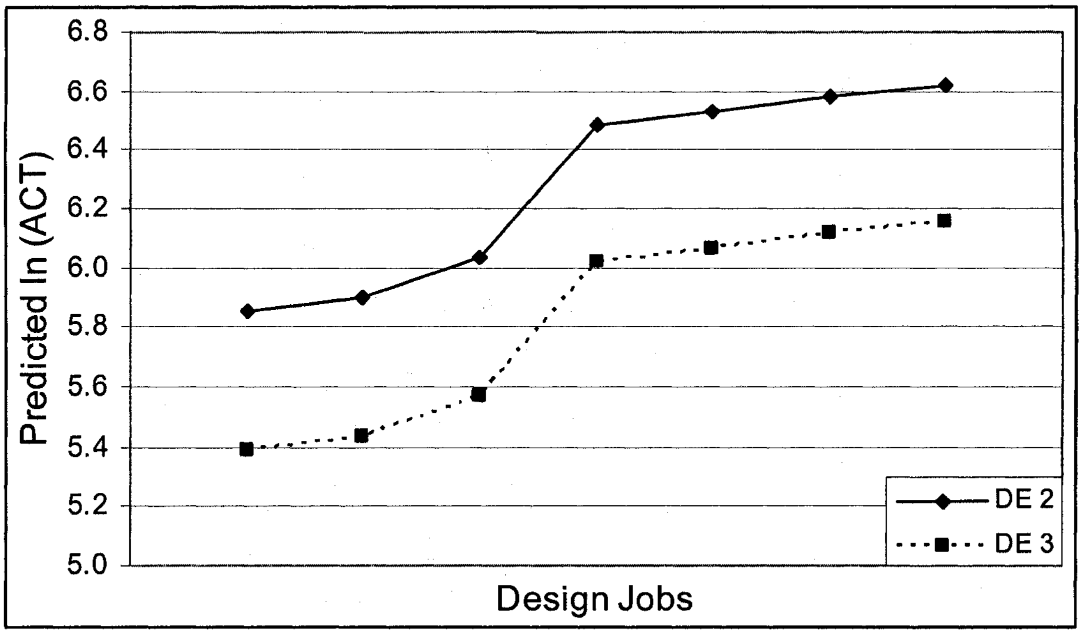

Figure 51: Impact of the experience of departmental personnel on effort for D3

APPENDIX G: ANALYSIS OF DEPARTMENT 4

Table 46: Data of D4

DJ	TD	DC	Con	DE	ACT
A	1	1	0.64	3.00	419.97
B	2	2	0.64	3.00	829.04
C	2	3	0.51	2.45	977.45
D	1	1	0.77	2.76	388.89
E	2	2	0.69	2.00	1035.54
F	2	3	0.65	2.80	770.83
G	1	1	0.49	1.45	584.48

Table 47: ln of data of D4

DJ	ln (TD)	ln (DC)	ln (Con)	ln (DE)	ln (ACT)
A	0	0	-0.448	1.099	6.040
B	0.693	0.693	-0.446	1.099	6.720
C	0.693	1.099	-0.671	0.896	6.885
D	0	0	-0.266	1.015	5.963
E	0.693	0.693	-0.365	0.693	6.943
F	0.693	1.099	-0.430	1.030	6.647
G	0	0	-0.712	0.372	6.371

Table 48: Regression coefficients of D4

Constants	All	$Jack_A$	$Jack_B$	$Jack_C$	$Jack_D$	$Jack_E$	$Jack_F$	$Jack_G$	$COEF_{JACK}$
$\ln(a_o)$	6.29	6.38	5.95	6.39	6.31	5.95	6.39	6.47	6.26
a_1	1.33	1.32	1.88	1.40	1.30	1.09	1.10	1.28	1.34
a_2	-0.229	-0.181	-0.627	-0.319	-0.217	-0.130	-0.012	-0.285	-0.253
a_3	-0.379	-0.289	-0.695	-0.221	-0.349	-0.695	-0.221	-0.652	-0.446
a_4	-0.412	-0.505	-0.192	-0.451	-0.416	-0.192	-0.451	-0.666	-0.411

Table 49: Residuals of D4

DJ	PRED ln(ACT)	ln(ACT)	Residual (Error)
A	6.0134	6.0402	0.02683
B	6.7664	6.7203	-0.04616
C	6.8473	6.8849	0.03766
D	5.9663	5.9633	-0.00300
E	6.8965	6.9427	0.04616
F	6.6851	6.6475	-0.03766
G	6.4296	6.3707	-0.05886

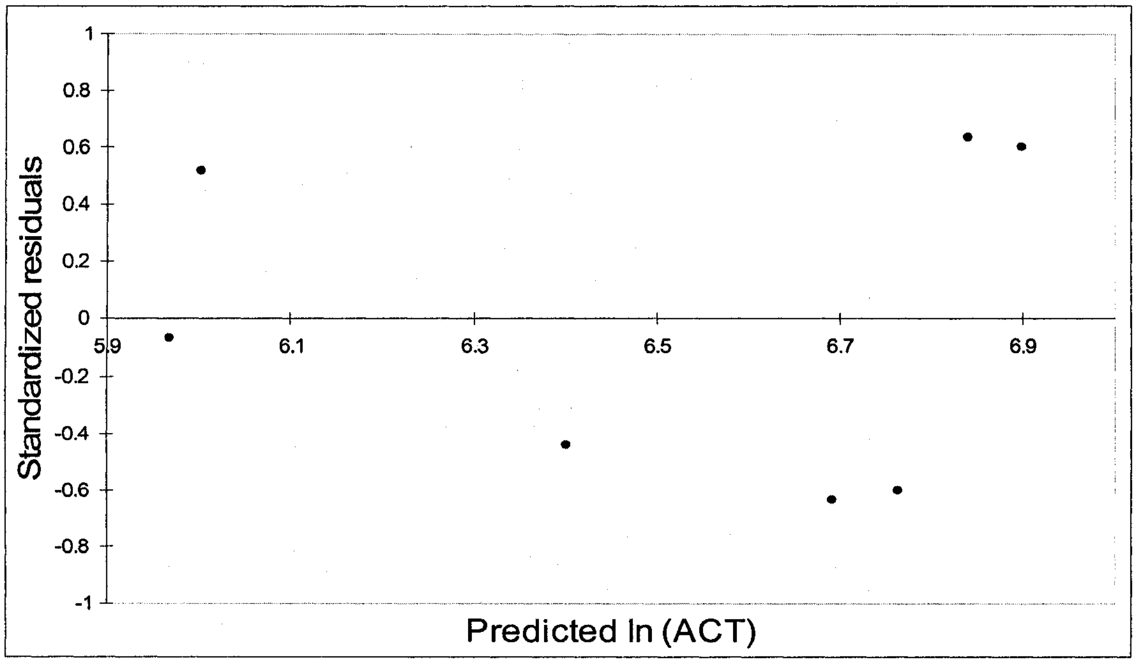

Figure 52: Residual plot of D4

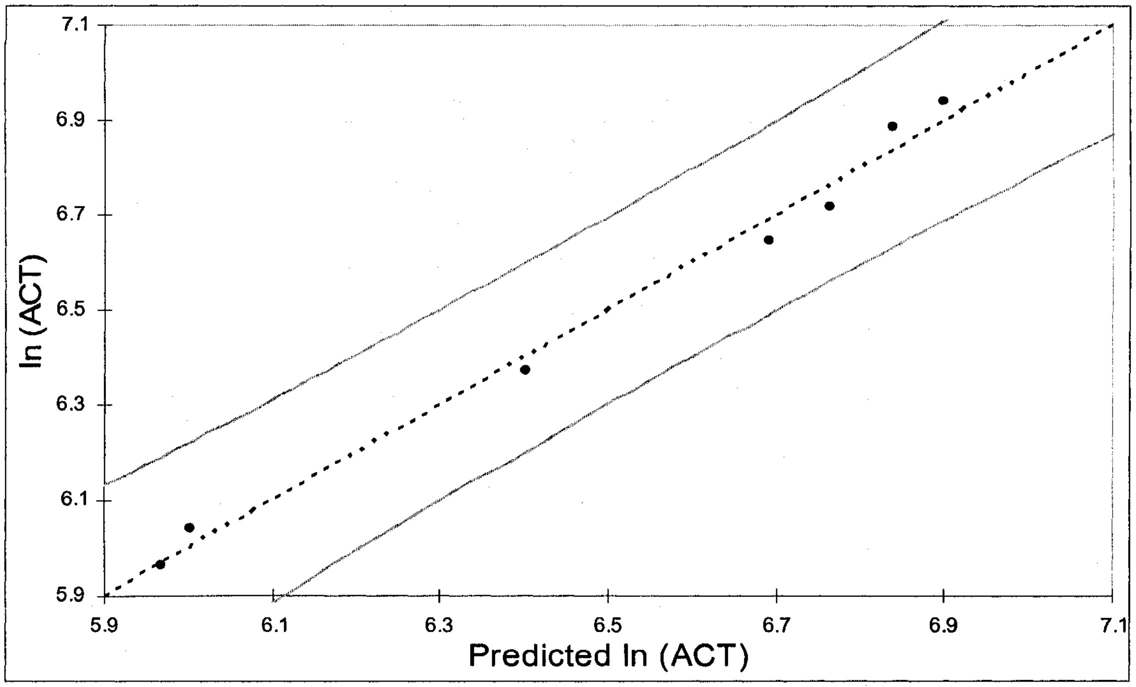

Figure 53: SPC chart of D4

Table 50: R^2 values of D4

Sample	R^2
All	0.989
$Jack_A$	0.990
$Jack_B$	0.998
$Jack_C$	0.994
$Jack_D$	0.982
$Jack_E$	0.998
$Jack_F$	0.995
$Jack_G$	1.000
R^2min	0.982
r_{R^2min}	0.991

$$\hat{E}_{D4} = 5.26 \text{x} 10^2 \, TD^{1.34} \, DC^{-0.253} \, Con^{-0.446} \, DE^{-0.411}$$

Table 51: Relative errors of D4

DJ	ACT (hours)	PRED (hours)	Relative Error (%)
A	419.97	393.73	2.65
B	829.04	827.24	4.72
C	977.45	904.69	3.70
D	388.89	421.59	0.30
E	1035.5	1037.79	4.51
F	770.83	832.82	3.84
G	584.48	595.7	6.06

Table 52: Correlation matrix of D4

Variables	ln (TD)	ln (DC)	ln (DE)	ln (Con)	ln (ACT)
ln (TD)	1.000	0.945	-0.010	0.202	0.915
ln (DC)	0.945	1.000	-0.130	0.224	0.843
ln (DE)	-0.010	-0.130	1.000	0.559	-0.285
ln (Con)	0.202	0.224	0.559	1.000	-0.181
ln (ACT)	0.915	0.843	-0.285	-0.181	1.000

Table 53: Significant factors of D4

Factor	All	$Jack_A$	$Jack_B$	$Jack_C$	$Jack_D$	$Jack_E$	$Jack_F$	$Jack_G$
ln (a_o)	0.001	0.030	0.021	0.024	0.041	0.021	0.024	0.004
ln (TD)	0.040	0.151	0.096	0.121	0.213	0.106	0.187	0.023
ln (DC)	0.389	0.605	0.204	0.378	0.621	0.487	0.971	0.079
ln (Con)	0.286	0.545	0.179	0.583	0.610	0.179	0.583	0.058
ln (DE)	0.111	0.262	0.371	0.200	0.305	0.371	0.200	0.037

Table 54: Regression coefficients of D4 without concurrency

Constants	All	$Jack_A$	$Jack_B$	$Jack_C$	$Jack_D$	$Jack_E$	$Jack_F$	$Jack_G$	$COEF_{JACK}$
$\ln(a_o)$	6.585	6.614	6.586	6.562	6.580	6.586	6.562	6.647	6.591
a_1	1.143	1.201	1.136	1.351	1.098	1.149	0.939	1.115	1.141
a_2	-0.069	-0.055	-0.065	-0.280	-0.077	-0.073	0.132	-0.060	-0.068
a_4	-0.555	-0.644	-0.557	-0.528	-0.510	-0.557	-0.528	-0.610	-0.562

Table 55: Residuals for D4 without concurrency

DJ	PRED ln (ACT)	ln (ACT)	Residual (Error)
A	5.9737	5.8776	-0.0961
B	6.7176	6.4955	-0.2221
C	6.8037	6.1203	-0.6834
D	6.0205	5.4132	-0.6073
E	6.9454	6.5694	-0.3760
F	6.7287	6.2020	-0.5267
G	6.3821	5.6840	-0.6981

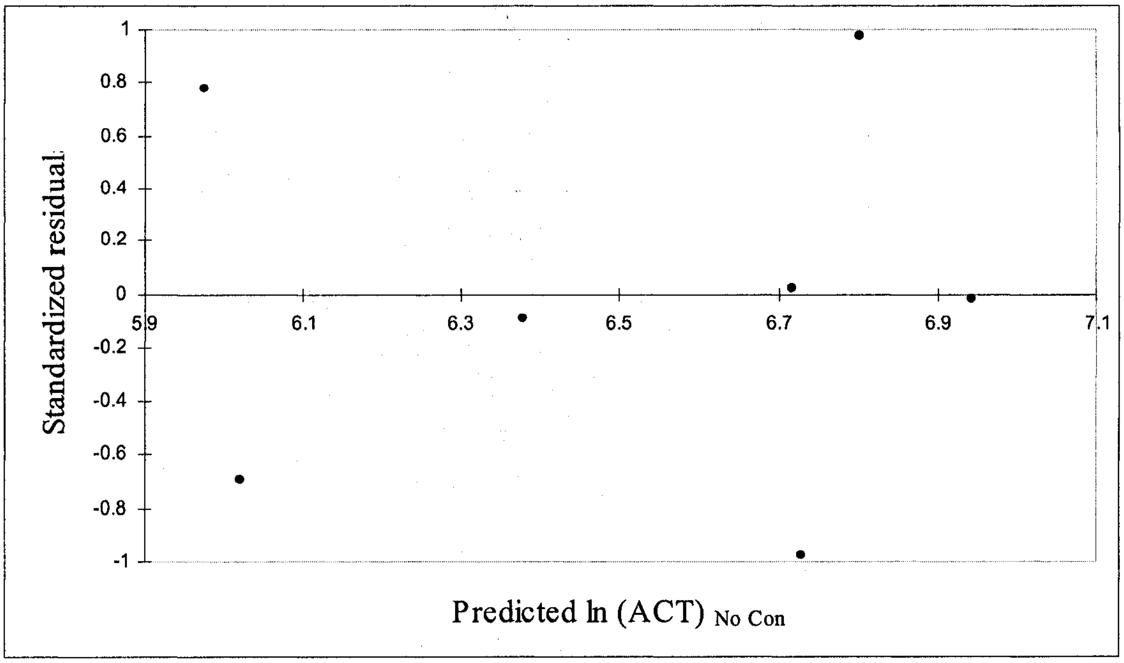

Figure 54: Residual plot of D4 without concurrency

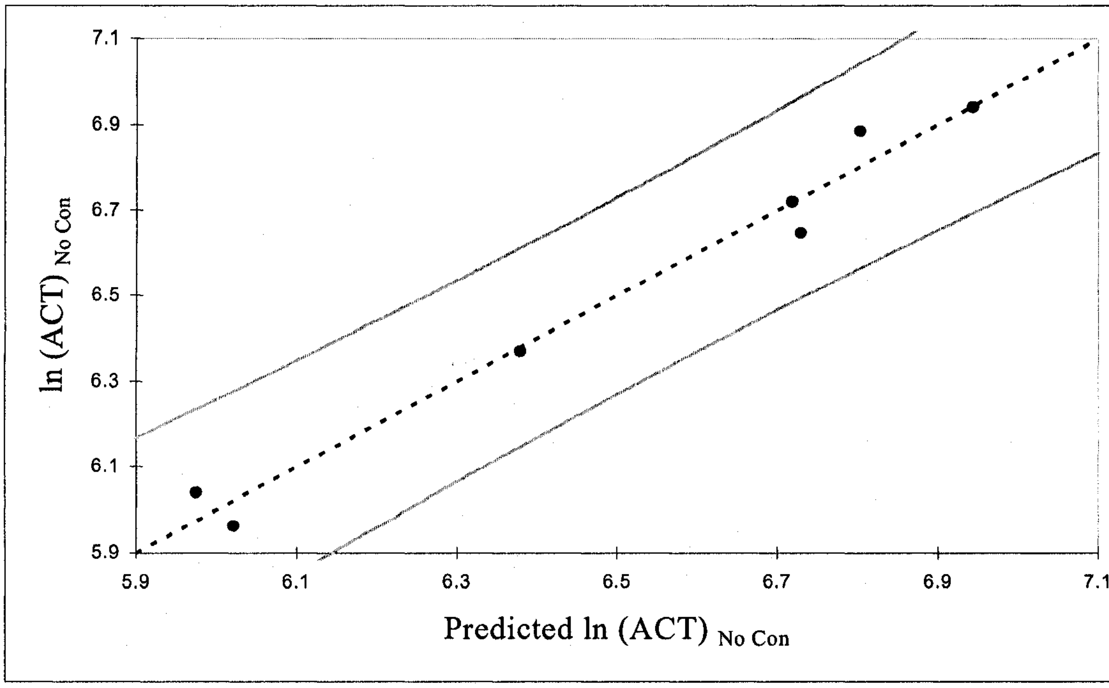

Figure 55: SPC chart of D4 without concurrency

Table 56: R^2 values of D4 without concurrency

Sample	R^2
All	0.977
$Jack_A$	0.982
$Jack_B$	0.976
$Jack_C$	0.990
$Jack_D$	0.974
$Jack_E$	0.970
$Jack_F$	0.992
$Jack_G$	0.977
R^2min	0.970
r_{R^2min}	0.985

$$\hat{E}_{D4_{\text{No Con}}} = 7.28\text{x}10^2\ TD^{1.14}\ DC^{-0.0680}\ DE^{-0.562}$$

Table 57: Relative errors of D4 without concurrency

DJ	ACT (hours)	PRED (hours)	Relative Error (%)
A	419.97	392.94	6.44
B	829.04	826.81	0.27
C	977.45	901.19	7.80
D	388.89	411.79	5.89
E	1035.54	1038.34	0.27
F	770.83	836.06	8.46
G	584.48	591.19	1.15

Table 58: Correlation matrix of D4 without concurrency

Variables	ln (TD)	Ln (DC)	ln (DE)	ln (ACT)
ln (TD)	1.000	0.945	0.202	0.915
ln (DC)	0.945	1.000	0.224	0.843
ln (DE)	0.202	0.224	1.000	-0.181
ln (ACT)	0.915	0.843	-0.181	1.000

Table 59: Significant factors of D4 without concurrency

Factor	All	$Jack_A$	$Jack_B$	$Jack_C$	$Jack_D$	$Jack_E$	$Jack_F$	$Jack_G$
ln (a_o)	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.003
ln (TD)	0.027	0.047	0.138	0.028	0.067	0.132	0.055	0.095
ln (DC)	0.761	0.805	0.863	0.272	0.757	0.838	0.546	0.836
ln (DE)	0.024	0.046	0.090	0.032	0.074	0.090	0.032	0.206

Table 60: Regression coefficients of D4 without concurrency and DC

Constants	All	$Jack_A$	$Jack_B$	$Jack_C$	$Jack_D$	$Jack_E$	$Jack_F$	$Jack_G$	$COEF_{JACK}$
ln (a_o)	6.588	6.617	6.596	6.582	6.584	6.579	6.562	6.664	6.598
a_1	1.054	1.132	1.047	1.020	1.001	1.046	1.097	1.035	1.054
a_4	-0.560	-0.649	-0.569	-0.552	-0.515	-0.549	-0.528	-0.627	-0.570

Table 61: Residuals of D4 without concurrency and DC

DJ	PRED ln(ACT)	ln(ACT)	Residual (Error)
A	5.9718	6.0402	0.0683
B	6.7023	6.7203	0.0179
C	6.8177	6.8849	0.0672
D	6.0194	5.9633	-0.0561
E	6.9334	6.9427	0.0093
F	6.7416	6.6475	-0.0942
G	6.3862	6.3707	-0.0155

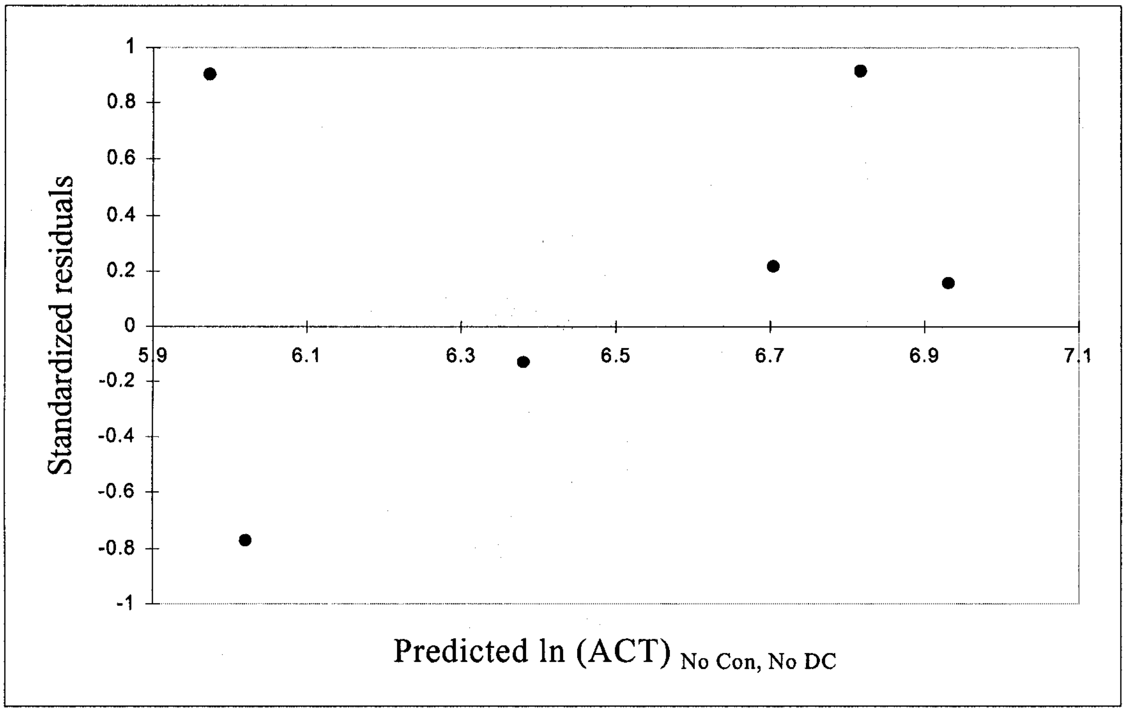

Figure 56: Residual plot of D4 without concurrency and DC

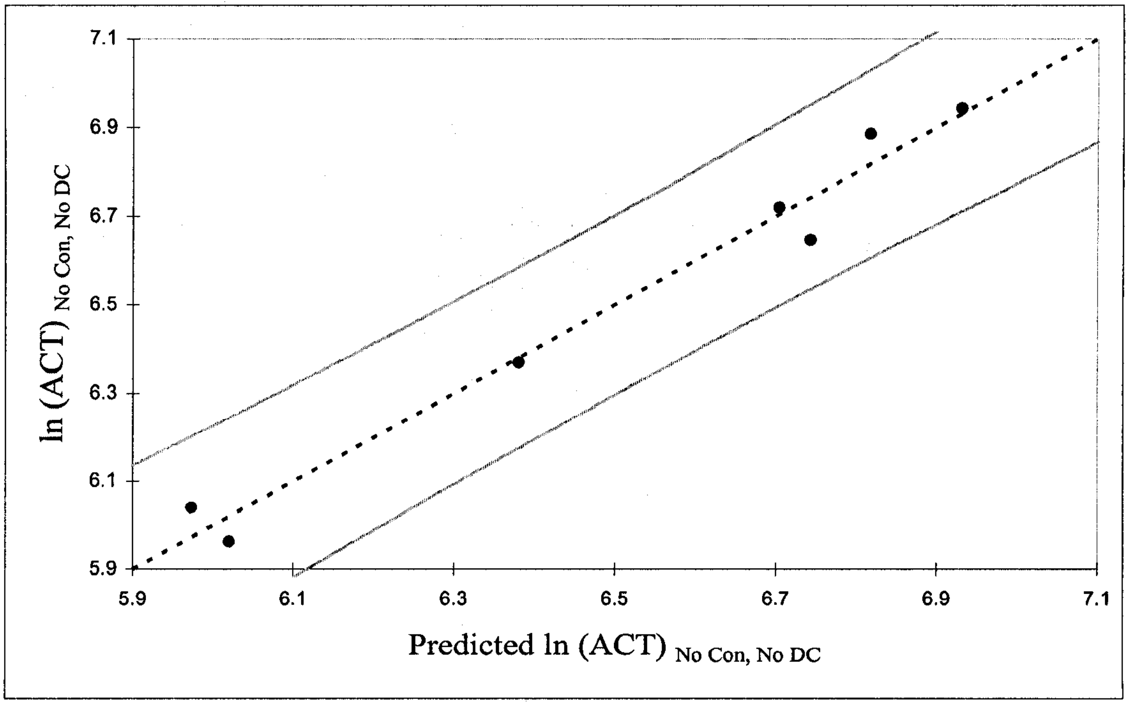

Figure 57: SPC chart of D4 without concurrency and DC

Table 62: R^2 values of D4 without concurrency and DC

Sample	R^2
All	0.977
$Jack_A$	0.981
$Jack_B$	0.976
$Jack_C$	0.979
$Jack_D$	0.972
$Jack_E$	0.970
$Jack_F$	0.990
$Jack_G$	0.977
R^2min	0.970
r_{R^2min}	0.985

$$\hat{E}_{D4_{\text{No Con, No DC}}} = 7.36\text{x}10^2\,TD^{1.05}\,DE^{-0.570}$$

Table 63: Relative errors of D4 without concurrency and DC

DJ	ACT (hours)	PRED (hours)	Relative Error (%)
A	419.97	392.23	6.61
B	829.04	814.3	1.78
C	977.45	913.92	6.50
D	388.89	411.32	5.77
E	1035.54	1025.98	0.92
F	770.83	846.95	9.88
G	584.48	593.59	1.56

Table 64: Correlation matrix of D4 without concurrency and DC

Variables	ln (TD)	ln (DE)	ln (ACT)
Ln (TD)	1.000	0.202	0.915
Ln (DE)	0.202	1.000	-0.181
Ln (ACT)	0.915	-0.181	1.000

Table 65: Significant factors of D4 without concurrency and DC

Factor	All	$Jack_A$	$Jack_B$	$Jack_C$	$Jack_D$	$Jack_E$	$Jack_F$	$Jack_G$
ln (a_o)	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.003
ln (TD)	0.027	0.047	0.138	0.028	0.067	0.132	0.055	0.095
ln (DE)	0.024	0.046	0.090	0.032	0.074	0.090	0.032	0.206

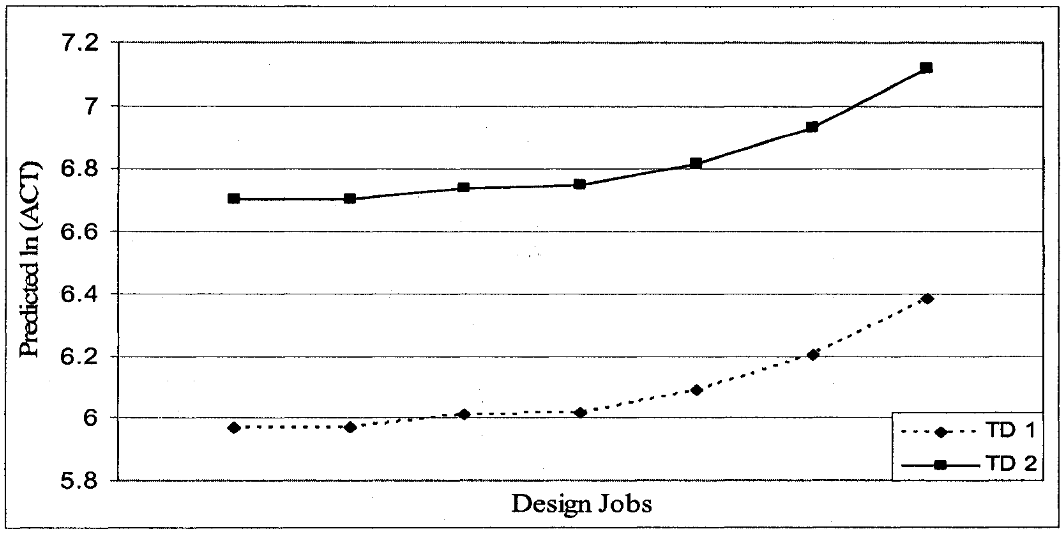

Figure 58: Impact of the type of design on effort for D4

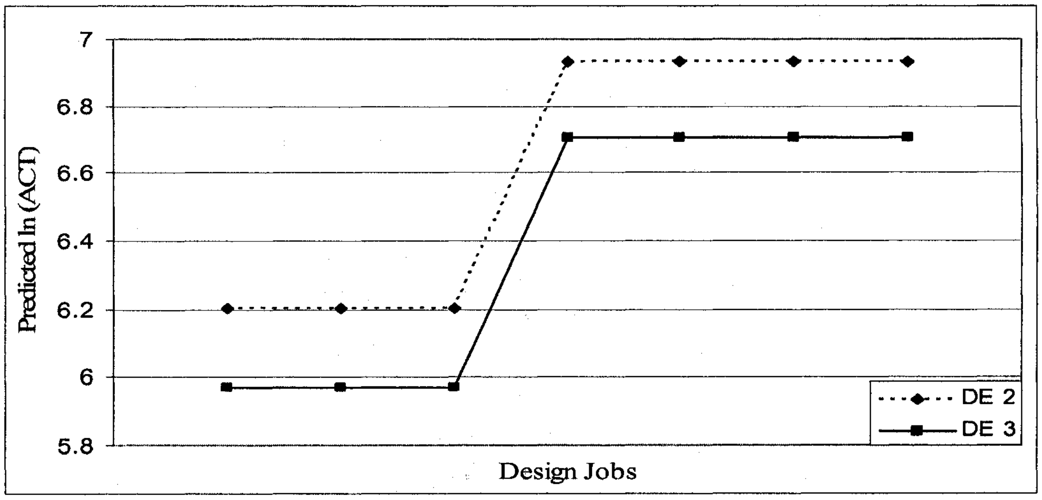

Figure 59: Impact of the experience of departmental personnel on effort for D4