2013

Forecasting the Number and Locations of Machine Installs Serviced by IBM in the U.S.

Zheng Shi
Lehigh University

Forecasting the Number and Locations of Machine Installs Serviced by IBM in the U.S.

by

Zheng Shi

A Thesis

Presented to the Graduate and Research Committee

of Lehigh University

in Candidacy for the Degree of

Master of Science

in

Department of Industrial and System Engineering

Lehigh University

Jan 2013

Thesis is accepted and approved in partial fulfillment of the requirements for the Master of Science in Industrial and System Engineering in Dept. of Industrial and Systems Engineering.

Forecasting the Number and Locations of Machine Installs Serviced by IBM in the U.S.

Zheng Shi

Date Approved

Nov.30, 2012

Thesis Director:

Dr. George R. Wilson

Department Chair:

Dr. Tamás Terlaky

ACKNOWLEDGMENTS

I’d like to express my deep appreciation to my supervisor Prof. George R. Wilson. Ever since I began my master study at Lehigh University, I have been learning plenty of knowledge from you. You are a respectable teacher, a good friend, and a great mentor! I’d like to thank you for offering me an opportunity to take part in the IBM forecast project and guiding me to finish the project.

I’d like to thank Mr. Mark Booth who helps me quite a lot in daily work and makes me realize the meaning of Operations Research. I’d like to thank Prof. Louis Plebani for helping me learn PostgreSQL and process the raw data. Also, I want to thank Mr. Terry Hammaker for offering me an opportunity to be part of his group in IBM. And I want to thank Mr. Rod V. Amburgh, Mr. Eugene Kelton, and Ms. Christina Ma for their diligent efforts to support my work.

I’d like to thank Prof. Tamás Terlaky for taking time out of his busy schedule to review my material and providing valuable suggestions that can further improve this thesis. And I want to thank all the professors in Dept. of Industrial and Systems Engineering for their great lectures. I also want to thank Ms. Rita Frey for all her kind and warm assistance.

I want to thank my best friends, Zhongyuan Wei, Jingbo Li, Xi Bai, Ben Niu, and Zi Wang for all their help and support. I cannot imagine how my life would be if I didn’t know them.

Finally, it would be impossible to say enough about my parents, Junjie Shi and Hailing Cheng, and my girlfriend, Sally Xu. All the effort in my academic career would be meaningless without you.

Table of Contents

List of Tables

List of Figures

Abstract

Forecasting the Number and Locations of Machine Installs Serviced by IBM in the U.S.

Zheng Shi

Lehigh University, 2013

Supervisor: **Prof. George R. Wilson**

This thesis presents two strategies to forecast the number of machines installed (installed machine count) serviced by IBM at the National Level, the Sub Region Level, and the Zip code Level in the U.S. Based on the available data, the first effort is a Poisson forecast strategy. The Poisson forecast strategy combines a 96% significant Two-Sided Hypothesis Test on Poisson Population Mean (2-HTPPM) and an Optimal Reallocation Strategy (ORS). This thesis uses Integer-Nonlinear-Constrained (INLC) Optimization model to realize the ORS, and then implements a Dynamic Programming Algorithm (DPA) to solve the INLC Optimization model. The econometric forecast strategy is also developed which contains elements of Missing Data Treatment (MDT), Feature Selection (FS), and Two-Stage Econometric (TSE) Strategy. In the future, if there will be more available data, the econometric forecast strategy can be applied to improve the forecast accuracy at the Sub Region Level and the Zip code Level.

Key words: installed machine count; National Level; Sub Region Level; Zip code Level; Poisson forecast strategy; 2-HTPPM; ORS; INLC Optimization model; DPA; econometric forecast strategy; MDT; FS; TSE Strategy.

Introduction

This master thesis is an academic accomplishment based on the IBM forecast project initiated at the beginning of the year of 2011. The IBM forecast project aims to develop a strategy to forecast the number and locations of all types of machines installed in the U.S. and subject to maintenance agreements with IBM. Since April of 2011, the forecast project has being worked on, and by the time of this thesis published, the forecast strategy will have been accomplished and presented to IBM.

In this chapter, the background and research work on the forecast project will be introduced.

Background of the IBM Forecast Project

Founded in the year of 1911, through a century of successful operations, International Business Machine Corporation (IBM) has built an international business empire. In recent decades, IBM has been providing products to customers at every corner of the world, and has achieved an increasing market share in the world with respect to its business offerings. However, due to its increasing market share, IBM has an increasing expenditure in providing post-sale maintenance service to its worldwide customers. In order to save money for its shareholders, IBM must try every means to reduce that expenditure.

Ever since IBM aimed to reduce the expenditure of its post-sale maintenance service, it has been pursuing a globally efficient reconfiguration of post-sale maintenance service resources and materials (the resources and materials mainly consist of part inventories, labor force, and other related investments). IBM is working on an efficient reconfiguration strategy in the U.S, first, which then can be applied to the rest of the world. The final, realized reconfiguration would allow IBM to efficiently control the level of part inventory, level of labor force, logistic network, and so on, such that IBM can reduce its overall service contract expenditure. To realize the efficient reconfiguration of the post-sale maintenance service in the U.S., there are two prerequisites that must be satisfied: (1) a proper maintenance service system to assign the resources and materials to each level of geography in the U.S.; (2) awareness of the post-sale maintenance service demand at each level of geography in the U.S.

The IBM post-sale maintenance service system is a system which can support the flow of resources and materials at various levels of geographic granularity. This system consists of service centers built upon a geographically hierarchical schema (shown in Fig.1). We take the U.S. system as an example to introduce the operating mechanism of the system. In the U.S., IBM defines the whole country as the top level, the National Level, and divides the whole country into Western Region and Eastern Region at the Region Level, under which there are 15 sub regions at the Sub Region Level. Then, there are 50 states at the State Level and the counties of these 50 states at the County Level. At the bottom of the schema, IBM defines the Zip code level containing all the active zip code areas in the U.S. ("active" zip code areas are those areas at which IBM sold products to customers). Following this schema, IBM has service centers at all levels and assigns the maintenance service resources and materials to the service centers at various levels of aggregation from the National Level down to the Zip code Level. Obviously, this system satisfies the first prerequisite.

So far as the post-sale maintenance service demand is concerned, the level of demand at a certain geographic level depends on the number of machine installs serviced by IBM at that level (IBM prefers the term of "machine" to "product", so the term of "machine" will be used in the following of the thesis.). In other words, once IBM obtains the information of how many machines are installed at a certain level, it would be able to forecast the maintenance service demand at that level. At present, among all levels of the system, IBM has the relatively best estimation of the number of machine installed at the National Level. However, the estimation becomes more and more inaccurate as geographic areas become smaller. This situation keeps IBM from correctly forecasting the maintenance service demand at each level of the system so that the second prerequisite is not satisfied.

Therefore, in order to realize the efficient reconfiguration in the U.S., the priority is to forecast the number of all types of machines installed at each level of the system in the U.S. Ultimately, this forecast strategy can be applied to the rest of the world to help IBM realize its globally efficient reconfiguration.

National Level
↓
Region Level
↓
Sub Region Level
↓
State Level
↓
County Level
↓
Zip code Level

Fig. 1: The Geographically Hierarchical Schema of the IBM Maintenance Service System

For simplicity, “installed machine count” will be used to represent “the number of machines installed” in the next section.

Detailed Description of the IBM Forecast Project

The objective of the research work is to develop the strategy for IBM to forecast the installed machine count for all types of machines under service contracts at each Level of the post-sale maintenance service system in the U.S. At IBM's request, it is necessary to forecast the installed machine count at the National Level, the Sub Region Level, and the Zip code Level. To realize the forecast strategy, there must have available data to conduct the research. Ever since the April of 2011, we have been working on data preparation together with IBM. Through one year's effort, we finally had three data sets available for our research:

- the Number of Observed Machine Failures (OFN)
- Engineering Machine Failure Rate (EFR)
- Estimate of Installed Machine Count (EIMC)

The above data sets are for all types of machines within a certain period of time at the Zip code Level in the U.S. The OFN is the number of the observed machine failures over the time period; the EFR is the failure rate per machine, and it is a constant for each type of machine over the U.S.; the EIMC is an estimation of the real installed machine count at the Zip code Level. This data is collected through IBM's daily operations, such as responding to the machine failures, customer visits, and regular machine maintenance. We can aggregate OFN and EIMC to the Sub Region Level and the National Level to get the data at those levels. And IBM considers OFN and EFR are more reliable than EIMC.

To capture the relationships between these three data sets, we made an important assumption that the process of machine failures in each area at each level is a Poisson process, or the number of failures of each type of machine in each area at each level has a Poisson distribution. In future work, this assumption can be verified by, for example, using data of the time between machine failures to test if the inter-arrival time of machine failures has an exponential distribution.

At this time, there are not enough data to prove the assumption; however, there is a justification to believe the above assumption is valid. Based on our knowledge of failures being rare events scattered among multiple customers, the machine failures are almost certainly mutually independent; hence, if sort the OFN based on the time line, we would find that OFN in different time intervals mutually independent and the OFN would only depend on the length of the time interval. We can reasonably conclude that the OFN has an independent and stationary increment, and we can assume the time between machine failures has an exponential distribution. According to the definition of the Poisson process (Ross, 2010), it can be assumed that the OFN for each type of machine in each area at each level has a Poisson distribution.

Based on the above assumption and the available data, this thesis presents a Poisson forecast strategy, and we use this strategy to forecast the installed machine count at the National Level, the Sub Region Level, and the Zip code Level in the U.S. The Poisson forecast strategy contains two parts: Two-Sided Hypothesis Test on Poisson Population Mean (2-HTPPM) and Optimal Reallocation Strategy (ORS) which is implemented by Dynamic Programming Algorithm (DPA).

To improve the accuracy of the forecast results at the Sub Region Level and the Zip code Level, this thesis also presents an econometric forecast strategy, and this strategy would be implemented if there can be more available data in the future. The econometric strategy combines Missing Data Treatment (MDT), Feature Selection (FS), and Two-Stage

Econometric (TSE) Strategy to forecast the installed machine count at the Sub Region and the Zip code Levels. The thesis proposes a copula random number generator to generate random numbers to test the performance of the econometric forecast strategy.

Poisson Forecast Strategy

Based on the assumption that OFN has a Poisson distribution (see ***Introduction***), the idea of the research work is to develop a forecast strategy such that the forecasted installed machine counts support the viewed occurrences of the machine failures. In this chapter, the Poisson forecast strategy is presented, and it contains two parts: the 96% significant Two-Sided Hypothesis Test on Poisson Population Mean (2-HTPPM) and the Optimal Reallocation Strategy (ORS).

At the National Level, the installed machine count is forecasted by building the 2-HTPPM to test if EIMC supports the OFN, and “fix” the EIMC if it fails the 2-HTPPM. By doing 2-HTPPM, the forecast results at the National Level can be obtained, and we denote the forecast results as the Forecasted Installed Machine Count (FIMC) at the National Level.

At the Sub Region Level and the Zip code Level, the Optimal Reallocation Strategy (ORS) is implemented to reallocate the FIMC at the National Level to the Sub Region Level, and then reallocate the FIMC at the Sub Region Level to the Zip code Level. Here, we construct the ORS as an Integer-Nonlinear-Constrained (INLC) Optimization model, and then apply the Dynamic Programming Algorithm (DPA) to solve the INLC problem.

As an illustration, the Poisson forecast strategy is implemented to forecast the installed machine count of one particular type of machine at the three levels.

Two-Sided Hypothesis Test on Poisson Population Mean at the National Level

Hypothesis test theory defines the hypothesis test on a Poisson population mean as a way to determine if, given a certain significance level, the occurrence of events supports the claimed

Poisson population means (Johnson, 2005). Here, the two-sided hypothesis test can be stated as the following:

$$\begin{gathered} H_0 : Poisson\ population\ mean = the\ claimed\ mean \\ H_1 : Poisson\ population\ mean \neq the\ claimed\ mean \\ Significance\ Level : \alpha \end{gathered}$$

$$\begin{gathered} If\ \ P(x \geq the\ occurence\ of\ \ events) \geq \frac{\alpha}{2}\ \ and\ \ P(x \leq the\ occurence\ of\ \ events) \geq \frac{\alpha}{2}, \\ the\ claimed\ mean\ is\ valid; \\ otherwise, reject\ the\ claimed\ mean \end{gathered}$$

At the National Level, according to the above theory, MATLAB is implemented to build the 96% significant Hypothesis Test on Poisson Population Mean (2-HTPPM) to determine if the EIMC supports the OFN. Here, the occurrence of events is the OFR; since EFR is the failure rate per machine, the Poisson mean or the average failure rate should be the product of EFR and the installed machine count. Then the 2-HTPPM can be expressed as:

$$\begin{aligned} &H_0 : Poisson\ population\ mean = EFR \times Installed\ \ Machine\ Count \\ &H_1 : Poisson\ population\ mean \neq EFR \times Installed\ \ Machine\ Count \end{aligned}$$

$$\begin{aligned} &if\ \ P(x \geq OFN) \geq 0.02\ \ \ and\ \ P(x \leq OFN) \geq 0.02 \\ &the\ claimed\ mean\ is\ valid; otherwise, reject\ the\ claimed\ mean \end{aligned}$$

Instead of inserting the EIMC into the equation of $EFR \times Installed\ Machine\ Count$, the *Installed Machine Count* is set as a variable for each type of machine. Then by assigning different values to the variables of the *Installed Machine Count*, the ranges of the variables which lead to $P(x \geq OFN) \geq 0.02$ and $P(x \leq OFN) \geq 0.02$ can be obtained. For each type of machine, the range can be seen as the 96% significant confidence interval of the *Installed Machine Count* making the 2-HTPPM a positive result.

For each type of machine, if the EIMC falls into the 96% significant confidence interval, there is no evidence to reject it as an accurate estimation. Besides, in the previous chapter, we mentioned that IBM has a relatively accurate estimation of installed machine count at the National Level (see ***Introduction***), and hence we must trust the EIMC that can fall into the confidence interval. However, if the EIMC does not fall into the confidence interval, we would have two situations:

- the EIMC is greater than the upper bound of the confidence interval
- the EIMC is less than the lower bound of the confidence interval

Both situations suggest an inaccurate EIMC at the National Level. Since IBM has confidence in its estimation at the National Level, we would like to make the smallest effort to fix the inaccurate EIMC. The smallest effort means to make the installed machine count equal to the upper bound, if the first situation happens, and equal to the lower bound, if the second situation. Then, combine the accurate EIMC and the fixed EIMC as the forecast results at the National Level, leading to a revised result for the Forecasted Installed Machine Count (FIMC) at the National Level.

Optimal Reallocation Strategies at the Sub Region and the Zip code Levels

It is a fact that the installed machine count at a certain level (excluding the Zip code Level) must be the sum of the installed machine count of the geographic partition at the lower level. In the last section, we obtained the FIMC at the National Level (see **Two-Sided Hypothesis Test on Poisson Population Mean at the National Level** in this chapter), and hence we can use the Optimal Reallocation Strategy (ORS) to reallocate the FIMC at the

National Level to the Sub Region Level and the Zip code Level. Before starting the forecast operations based on the ORS, let us look at the two alternative plans to carry out the ORS at the Sub Region and the Zip code Levels.

First plan:

Step 1

Reallocate FIMC at the National level over all sub regions, and make sure each of them can pass the 2-HTPPM after reallocation; denote reallocated installed machine count as FIMC at the Sub Region level.

Step 2

Reallocate FIMC at the Sub Region level over all zip code areas, and make sure each of them can pass the 2-HTPPM after reallocation; denote reallocated installed machine count as FIMC at the Zip code level.

For each type of machine, the first plan is to reallocate the FIMC at the National Level to all the machine's active sub regions ("active" sub regions are sub regions in which the machine are installed) to get the FIMC at the Sub Region Level. And then, to reallocate the FIMC in each sub region to the corresponding active zip code areas to determine the FIMC at the Zip code Level. The FIMC at each level must pass the 2-HTPPM.

Second plan:

Step 1

A: Do the 2-HTPPM for EIMC at the Sub Region Level, first, and denote the sub region as a "pass region" if the EIMC can pass 2-HTPPM and denote the installed machine count as N_{pass} at the Sub Region level.

B: Reallocate $FIMC - \sum N_{pass}$ *(where FIMC is from the National level) to those sub regions which fail the 2-HTPPM at Step 1, and make sure each of them can pass the 2-HTPPM after reallocation; combine* N_{pass} *and the reallocated installed machine count of these sub regions as the FIMC at the Sub Region level.*

Step 2

A: Do the 2-HTPPM for EIMC at the Zip code Level, first, and denote the zip code area as a "pass area" if the EIMC can pass 2-HTPPM and denote the installed machine count as N_{pass} *at the Zip code level.*

B: Reallocate $FIMC - \sum N_{pass}$ *(where FIMC is from the Sub Region level) to those zip code areas which fail the 2-HTPPM at Step 1, and make sure each of them can pass the 2-HTPPM after reallocation; combine* N_{pass} *and the reallocated installed machine count of these zip code areas as the FIMC at the Zip code level.*

The second plan shares the same reallocation method logic with the first plan, but it requires a 2-HTPPM before reallocation at each level. In the second plan, the EIMC is kept in each sub region or zip code area unchanged if the EIMC can pass the 2-HTPPM, and then reallocate the ones which cannot pass the test.

While, to some extent, IBM has confidence in the accuracy of the data of EIMC as long as the data for the Zip code Level aggregates to the National Level, since aggregation to the National Level has a tendency to neutralize the errors of estimation, the data of EIMC at the Sub Region Level and the Zip code Level is assumed to be untrustworthy by IBM. Those EIMC which can pass the 2-HTPPM at these two levels cannot be treated as being as accurate as those at the National Level. Therefore, at this time, the first plan is chosen to execute the ORS. In the future, if the evidence would be provided by IBM to show that some parts of the data of EMIC at the Sub Region and the Zip code Levels are reliable, we can choose the second plan and revise the ORS illustrated in this thesis.

In the following sections, the thesis takes the reallocation of installed machine count at the Sub Region level as an example to illustrate the ORS, and the same logic can be applied in the reallocation of installed machine count at the Zip code Level.

Integer-Nonlinear-Constrained Optimization Model

Having chosen the first plan to continue the ORS, we need to build a model to realize the strategy. In addition, the reallocated installed machine count (or FIMC) must have "Legitimacy", which means *(for each type of machine)*:

- The reallocated installed machine count in sub regions (zip code areas) should be, in some sense, the "most likely numbers" installed in the sub regions (zip code areas).
- The sum of the reallocated installed machine count in sub regions (zip code areas) should be equal to the FIMC at the National Level (Sub Region Level).
- The reallocated installed machine count must pass the 2-HTPPM.

For example, in order to execute ORS at the Sub Region Level, it is necessary to build an optimization model to both reallocate the FIMC at the National Level to the Sub Region Level and fulfill the requirement of "Legitimacy". Let us first look at the data and variables we have for the reallocation model at the Sub Region Level: *(for one type of machine)*

$Data:$

$R =$ # of active sub regions at the Sub Region Level

$n_r =$ # of observed machine failures in sub region r at the Sub Region Level over a specific period of time

$N =$ installed machine count to reallocate (N is also the FIMC at National Level)

$\lambda =$ failure rate per machine over a specifice period of time

$Variable:$

$N_r =$ reallocated installed machine count in sub region r

$N_r \cdot \lambda =$ Poisson mean or average failure rate in sub region r over a specific period of time

A question arises, and this question leads us to the objective function of the reallocation model. The question is "how can we make sure the machine is most likely installed in that sub region?" Given n_r has a Poisson distribution in each sub region r with $N_r \cdot \lambda$ as the Poisson population mean, the Poisson mass function (shown as *Equation 1.1*) can be used to measure "how likely" n_r will occur when N_r is allocated to the sub region *r,* providing a mean of $N_r \cdot \lambda$.

$$p(x = n_r) = \frac{(N_r \cdot \lambda)^{n_r} e^{-N_r \lambda}}{n_r!} \tag{1.1}$$

Then, according to *Equation 1.1*, *Equation 1.2* can be formed to measure "how likely" all the N_r allocated to all sub regions. And *Equation 1.2* can be rewritten as *Equation 1.3*.

$$\sum_r^R p(x_r = n_r) \quad = \quad \sum_r^R \frac{(N_r \cdot \lambda)^{n_r} e^{-N_r \lambda}}{n_r!} \tag{1.2}$$

$$\sum_r^R p(x_r = n_r) \quad = \quad \sum_r^R p(n_r \mid N_r \cdot \lambda) \tag{1.3}$$

Maximizing *Equation 1.3* can make sure all the reallocated installed machine counts are the most likely numbers to be located at the sub regions. Furthermore, we set up constraints so that we can fulfill the other requirements:

1. To make sure the total number of reallocated installed machine count is equal to the FIMC at the National Level, a constraint shown as 1.4 is set up: *(We name this constraint as "Conservation Constraint" for future use)*

$$\sum_{r}^{R} N_r = N \tag{1.4}$$

2. To guarantee the positive 2-HTPPM results, the following constraints shown as 1.5 are added:

$$\begin{aligned} &p(x \geq n_r) \geq 0.02 \quad \forall r \\ &p(x \leq n_r) \geq 0.02 \quad \forall r \end{aligned} \tag{1.5}$$

3. In the forecast problem, since any single machine cannot be split into several parts of machine, an integrality constraint shown as 1.6 must be introduced:

$$N_r \text{ is integer} \quad \forall r \tag{1.6}$$

Finally, the model may be written as: *(for one type of machine)*

$$\begin{aligned} \max \ & \sum_{r}^{R} p(n_r \mid N_r \cdot \lambda) \\ \text{s.t.} \ & \sum_{r}^{R} N_r = N \\ & p(x \geq n_r) \geq 0.02 \quad \forall r \\ & p(x \leq n_r) \geq 0.02 \quad \forall r \\ & N_r \text{ is integer} \qquad \forall r \end{aligned}$$

This model describes a problem which is an Integer-Nonlinear-Constrained (INLC) Optimization problem. It is well known that the difficulties in solving integer optimization problem and nonlinear-unconstrained optimization problem are greater than solving linear or continuous optimization. And our case is a hybrid constrained optimization problem combining integer and nonlinear optimization problems so that the difficulty is even much higher than each pure kind. Furthermore, the difficulty in dealing with INLC Optimization problem is growing as the size of the problem is increasing. Since there are around three thousands types of machines, of which each has tens of active sub regions and hundreds of active zip code areas need to be reallocated, the problem has a very big size leading to a high level of difficulty, and, hence, we cannot expect a high level of efficiency. Therefore, we need to find a good optimization tool to solve the reallocation problem by capturing the idea in the INLC Optimization model and reducing the difficulty.

Dynamic Programming Algorithm

Dynamic Programming Algorithm (DPA) in mathematical optimization is famous in making certain complex problems easier, and it allows the control of more details at each step during the optimizing process. Over the history of Operations Research, there have been many scholars that have developed lots of applications of DPA. In other words, DPA can be used to solve all kinds of optimization problems when the problems are decomposable into stages by variables or groups of variables (Denardo, 1982; Kleywegt and Shapiro, 2001; Winston and Venkataramanan, 2003). As an illustration, the “Knapsack Problem” or, generally, the resource allocation problem mentioned by E. V. Denardo exploits this type of decomposition. The production of various commodities can be modeled as having different stages no matter, in reality, if they are simultaneously assigned resources, or if they are receiving resources one by one. Then this problem can be solved stage by stage (Denardo, 1982). For any dynamic programming problems, there are several necessary elements: Stage and Stage Variable (SGV), State and State Variables (STV), Transition Function (TF), and Sub-Objective Function (SOF) (Denardo, 1982; Kleywegt and Shapiro, 2001; Winston and Venkataramanan, 2003).

Back to the original INLC Optimization model, the INLC Optimization model aims to reallocate the FIMC of one type of machine at the National Level into several sub regions at the Sub Region Level. Therefore, according to the literature, we can see the INLC Optimization problem as a nonlinear version of the “Knapsack problem”. Through implementing a DPA, we can use its backward recursion to put integer variables into the DPA and store the feasible solutions state by state and stage by stage, and this process also allows us to track the change of variables. To realize the objective function contained in the original INLC Optimization model, we can simply calculate the feasible values of the objective function and store them state by state and stage by stage. And for the constraints in INLC Optimization model, we can check if the variables or objective function values violate the constraints by direct computation at each state and stage. There are many programming environments which can realize a DPA (Benavides *et al*., 2007; Zietz, 2007; Sundström and Guzzella, 2009). In this thesis, we use MATLAB to implement the DPA.

Before we develop the DPA, Stage and SGV, State and STV, SOF, and TF in the DPA need to be defined. Recall there are *R* regions and *N* installed machine count to reallocate (*N* is the FIMC at the National Level), and there are n_r in both the INLC objective function and 2-HTPPM constraints.

We define each sub region as one Stage where $r = 1, 2, 3, ..., R$, and each N_r is a SGV. Here, we define $N_r = 0, 1, 2, ..., N \ \forall r$. At each Stage, we define each possible allocation of installed machine count as one State, and N_r' is a STV where $N_r' = 0, 1, 2, ..., N_r$. Combining definition of SGV and STV, we can make sure that every possible reallocation throughout the whole DPA can be stored; then, we define the TF as $(N_r - N_r')$ which indicates that if we allocate N_r to Stage *r*, then $(N_r - N_r')$ installed machine count will be allocated to the Stages from $r+1$ to R. Using backward recursion, the TF can be defined:

Last Stage TF:

At last Stage, Stage *R*, we will have no future stages, and hence we have no reallocation possible after the last Stage. Then, we have $N_r - N_r'$ for last Stage, and the TF is 0.

TFs from Second Last Stage to First Stage:

At each Stage $r, where 1 \le r < R$, we will have N_r' of allocated installed machine count and leave $(N_r - N_r')$ for future stages. Therefore, the TFs from second last Stage to first Stage can be written as *Equation 2.1*:

$$\begin{gathered}\mathrm{TF}_{(r|1 \le r < R)} = (N_r - N_r') \\ \\ \text{where } N_r = 0, 1, 2, ..., N \text{ and} \\ \text{where } N_r' = 0, 1, 2, ..., N_r\end{gathered} \tag{2.1}$$

The definitions of SGV, STV, and TF guarantee that the total number of installed machine count assigned to each Stage will be equal to *N* so that we can satisfy the *Conservation Constraint* (see *Equation 1.4*) in INLC Optimization model.

We can now define SOF in our DPA. Recall we have objective function in INLC Optimization model as $\max \sum_{r}^{R} p(n_r \mid N_r \cdot \lambda)$. Then we can use DPA's backward recursion to define SOF for each Stage.

Last Stage SOF:

Assume that there are N_R available to the last Stage, and there will be no stages after the last Stage. Then we will have each SGV is also a STV, and the TF equals to 0. Therefore, the SOF at last Stage can be expressed as *Equation 2.2*:

$$\begin{aligned} &\mathrm{SOF}_R(N_R) = p(n_R \mid N'_R \cdot \lambda) \\ &\text{where } N_R = 0, 1, 2, ..., N \text{ and} \\ &\text{where } N_R = N'_R \end{aligned} \tag{2.2}$$

SOFs from Second Last Stage to First Stage:

Assume that there are N_r available to the Stage *r*, $\forall r \neq R$, and then we will assign N'_r to Stage *r*, where $N'_r = 0, 1, 2, ..., N_r$, and leave $(N_r - N'_r)$ to the stages from $r+1$ to R, and thus the TF at Stage *r* is $(N_r - N'_r)$. Therefore, the SOF at Stage *r* can be expressed as *Equation 2.3* according to *Equation 2.1*:

$$\begin{aligned} \mathrm{SOF}_{1 \le r < R}(N_r) &= \max_i \{ p(n_r \mid N'_r \cdot \lambda) + \mathrm{SOF}_{r+1}(N_r - N'_r) \} \\ &= \max_i \{ p(n_r \mid N'_r \cdot \lambda) + p(n_{r+1} \mid (N_r - N'_r) \cdot \lambda) \} \\ &\text{where } N_r = 0, 1, 2, ..., N \text{ and} \\ &\text{where } N'_r = 0, 1, 2, ..., N_r \text{ and} \\ &\quad N_1 = N \end{aligned} \tag{2.3}$$

After defining the SOFs for all stages, there is also a need to embed the 2-HTPPM constraints into the DPA's backward recursion. If one possible allocation of the installed machine count at one Stage fails the 2-HTPPM, we would make the value of SOF of that possible allocation equal to negative infinity so that we can eliminate this possible allocation since we are maximizing SOF. The complete DPA is shown as the following: *(for one type of machine)*

Last Stage:

$$\begin{aligned}
&\mathrm{SOF}_R(N_R) = p(n_R \mid N'_R \cdot \lambda) \\
&\text{s.t. if } p(x \geq n_R) < 0.02 \text{ or } p(x \leq n_R) < 0.02 \\
&\text{then } \mathrm{SOF}_{r=R}(N_R) = \text{-}\inf \\
\\
&\text{where } N_R = N'_R \qquad\qquad \forall\, N_R \\
&\qquad\quad N_R = 0, 1, 2, ..., N
\end{aligned}$$

From Second Last Stage to Frist Stage:

$$\begin{aligned}
&\mathrm{SOF}_{1\leq r<R}(N_r) = \max_i\{p(n_r \mid N'_r \cdot \lambda) + \mathrm{SOF}_{r+1}(N_r - N'_r)\} \\
&\text{s.t. if } p(x \geq n_r) < 0.02 \text{ or } p(x \leq n_r) < 0.02 \\
&\text{then } \mathrm{SOF}_{1<r<R}(N_r) = -\inf \\
\\
&\text{where } N'_r = 0, 1, 2, ..., N_r \quad \forall 1 \leq r < R \\
&\qquad\quad N_r = 0, 1, 2, ..., N \quad \forall 1 \leq r < R \\
&\qquad\quad N_1 = N
\end{aligned}$$

At the first Stage, we can evaluate the feasible solution which maximizes the SOF of the first Stage, and this feasible solution is the optimal solution, and, hence, the reallocated installed machine count or FIMC.

In future sections of this chapter, we will use the Poisson forecast strategy to forecast the installed machine count utilizing data for one particular type of machine.

Data

The data of OFN, EFR, and EIMC are collected from the IBM database, which is information at the Zip code Level within the past 6 years. The EFR is the monthly failure rate per machine, and it is a constant for each type of machine over the U.S.; to obtain the data of OFN and EIMC at the National and the Sub Region Levels, we can aggregate them to the National and the Sub Region Levels. In this thesis, we only forecast the installed machine count of one type of machine, so we choose the data of one type of machine, Machine A, from the IBM database.

Here, due to confidentiality, we cannot use the real names of the machine type, sub regions, and zip code areas, therefore we use “Machine A” to stand for machine’s real name, use Roman numerals to stand for the machine’s active sub regions, and use Arabic numerals to stand for the active zip code areas of each sub region. The following tables show the data for Machine A.

Table 1: The Data for the National Level

OFN	EIMC	No. of Sub Regions	EFR (Per 6 Years)
176	1217	15	0.19799992575

Table 2: The Data for the Sub Region Level

Sub Region	OFN	EIMC	No. of Zip code Areas
I	2	29	15

II	2	18	10
III	7	54	15
IV	7	68	22
V	7	63	29
VI	7	41	12
VII	8	79	14
VIII	10	59	15
IX	12	272	15
X	12	58	18
XI	14	57	23
XII	17	81	21
XIII	21	60	15
XIV	23	133	33
XV	27	145	27
TOTAL	176	1217	284

Concerning the data for the Zip code Level, we list the data of zip code areas under *Sub Region III and IX* in **Table 3** and **Table 4,** as examples, and more data for the Zip code Level can be found in the tables of data and results in the Appendix (see ***Appendix: Data and FIMC at the Zip code Level***).

Table 3: The Data for the Zip code Level: the Data for the zip code areas under *Sub Region III*

Zip code Area	OFN	EIMC
1	0	6
2	0	1
3	1	1
4	1	14

5	1	1
6	0	1
7	0	1
8	1	1
9	0	1
10	1	2
11	2	6
12	0	1
13	0	1
14	0	1
15	0	16
TOTAL	7	54

Table 4: The Data for the Zip code Level: the Data for the zip code areas under *Sub Region IX*

Zip code Area	OFN	EIMC
1	0	1
2	2	2
3	2	2
4	0	1
5	0	1
6	1	1
7	1	1
8	0	1
9	3	7
10	0	16
11	0	2
12	1	1

13	0	3
14	2	232
15	0	1
TOTAL	12	272

Results

At the National Level, by constructing the 2-HTPPM, the 96% significant confidence interval of Machine A’s installed machine count is obtained, which is $[757,\ 1037]$. Since the EIMC is 1217 which is greater than the upper bound of the confidence interval, we make the FIMC at the National Level equal to the upper bound, 1037. Results are shown in **Table 5**.

Table 5: Results at the National Level

OFN	Confidence Interval		Significance Level	EIMC	FIMC
	Lower Bound	Upper Bound			
176	757	1037	96%	1217	1037

In **Table 5**, the FIMC is 180 less than the EIMC. And there are several reasons that can result in the difference: (1) the machines which were previously on the service contracts have been moved to another country or retired; (2) a data issue: the EIMC is obtained by aggregation of EIMC at the Zip code Level, so there might be a small errors of estimations at data entry (although the aggregation to the National Level tends to neutralize the errors), and the difference between EIMC and FIMC suggests the error; (3) the time periods of EIMC and OFN are not the same, but the FIMC only reflects the time period of OFN.

In the next step of the process, the ORS is implemented to reallocate 1037 Machine A to 15 sub regions, and the results of FIMC at the Sub Region Level shown in **Table 6**.

Table 6: Results at the Sub Region Level

Sub Region	OFN	EIMC	Reallocated Installed Machine Count (or FIMC)
I	2	29	10
II	2	18	10
III	7	54	36
IV	7	68	36
V	7	63	36
VI	7	41	36
VII	8	79	42
VIII	10	59	52
IX	12	272	63
X	12	58	63
XI	14	57	74
XII	17	81	90
XIII	21	60	111
XIV	23	133	177
XV	27	145	201
TOTAL	176	1217	1037

In **Table 6**, some sub regions have big differences between EIMC and FIMC. For example, the data (EIMC) shows there should be 272 machines installed in *Sub Region IX*, however, the forecast result of installed machine count, 63, is way less than 272. The opposite example is that the EIMC suggests 60 machines installed in *Sub Region XIII*, however, the FIMC shows the installed machine count should be almost doubled. One explanation is that there were 272 (or 60) machines on the service contracts, but the customers in *Sub Region IX* (*Sub Region XIII*) moved out (in) some of the machines to (from) other sub regions or some of the machines were just retired (recently purchased). Records were not updated in a timely fashion. The other explanation is that the ORS reallocates the FIMC at the National Level to the Sub Region Level based on OFN of each sub region. In order to make the allocated machine count "most likely" installed in all sub regions, the sub region which has a greater OFN is most likely allocated a larger installed machine count. Besides, the data of EIMC at the Sub Region Level is the aggregation of data for the Zip code Level. The level of aggregation is not high enough to neutralize the errors of estimations in the data at the Zip code Level. Therefore, there are big differences between EIMC and FIMC in some sub regions.

Also in **Table 6**, we can see that the sub regions which have the same OFN usually have the same FIMC (for example, *Sub Region IX* and *Sub Region X* have the same OFN and FIMC), and we will discuss this finding in the next section (see **Discussion of the Poisson Forecast Strategy** in this chapter).

After the FIMC at the Sub Region Level is obtained, we can do the ORS at the Zip code Level by reallocating the FIMC (at the Sub Region Level) to the active zip code areas. Here, we list the results for the zip code areas under *Sub Region III and IX* in **Table 7** and **Table 8,** as examples, and more results can be found in the Appendix (see ***Appendix: Data and FIMC at the Zip code Level***).

Table 7: Results at the Zip code Level: Results for the zip code areas under *Sub Region III*

Zip code Area	OFN	EIMC	Reallocated Installed Machine Count (or FIMC)
1	0	6	0
2	0	1	0

3	1	1	5
4	1	14	5
5	1	1	5
6	0	1	0
7	0	1	0
8	1	1	5
9	0	1	0
10	1	2	5
11	2	6	11
12	0	1	0
13	0	1	0
14	0	1	0
15	0	16	0
TOTAL	7	54	36

Table 8: Results at the Zip code Level: Results for the zip code areas under *Sub Region IX*

Zip code Area	OFN	EIMC	Reallocated Installed Machine Count (or FIMC)
1	0	1	0
2	2	2	10
3	2	2	11
4	0	1	0
5	0	1	0
6	1	1	5
7	1	1	5
8	0	1	0
9	3	7	16
10	0	16	0

11	0	2	0
12	1	1	5
13	0	3	0
14	2	232	11
15	0	1	0
TOTAL	12	272	63

In **Table 7-8**, we can find that some zip code areas have obvious differences between EIMC and FIMC. For examples, in **Table 7**, No.4 zip code area has 14 machines in the data of EIMC, but the FIMC is only 5 machines; No.15 zip code area has 16 machines in the data of EIMC, but its FIMC is 0; the EIMC for No.3 zip code area is 1, while the FIMC is 5; the EIMC for No. 11 zip code area is 6, while the FIMC is 11. First of all, the inaccurate data of EIMC at the Zip code Level should be the most important reason for the big differences between EIMC and FIMC in some zip code areas. The other reason is that the customers in these zip code areas moved out or in the machines to or from other zip code areas, or the machines were just retired or recently purchased with a lag in time in recording these changes. Besides, the ORS is based on OFN, and, hence, more OFN probably lead to more FIMC, or vice versa.

Furthermore, in **Table 7-8**, we find that the zip code areas which have the same OFN are most likely allocated the same FIMC. Also, we find that the zip code areas, at which there is no record of OFN, have no reallocated installed machine count (or FMIC). We will discuss these two findings in the next section.

Discussion of the Poisson Forecast Strategy

Given the three available data sets: OFN, EIMC, and EFR, we are working with three assumptions throughout the research on the Poisson forecast strategy:

- The first assumption: the OFN has a Poisson distribution for each type of machine at each level.
- The second assumption: the data of OFN is considered reliable data.
- The third assumption: the data of EIMC at the National Level is the relatively best estimation of the real installed machine count, while that at the Sub Region Level or the Zip code Level is not considered trustworthy.

Based on the first and the second assumptions, we embed the concept of Poisson distribution into the forecast strategy. At all levels, the idea of the Poisson forecast strategy is to make the FIMC best support the OFN. Based on the third assumption, we make the smallest effort (see **Two-Sided Hypothesis Test on Poisson Population Mean at the National Level** in this chapter) to obtain the FIMC at the National Level, while at the Sub Region Level and the Zip code Level, we use the Optimal Reallocation Strategy (ORS) to get the FIMC.

By looking at the forecast results, we can find some issues which are probably the problems. From **Table 6-8** (see **Results** in this chapter), we find that if two sub regions (several zip code areas) have the same OFN, they would have the same FIMC. And from **Table 7-8** (see **Results** in this chapter) and other tables of results (see ***Appendix: Data and FIMC at the Zip code Level***), we see that if the OFN in a zip code area is equal to zero, the ORS would not assign installed machine count to that zip code area. In reality, although there is a probability of the above situations occurring, two sub regions (several zip code areas) having the same OFN may not have the same number of machines installed, and the zip code areas having no machine failures may still have machines installed.

However, based on the idea that is to make the FIMC best support the OFN, the objective function of the INLC Optimization model is to maximize the total probability of the OFN happening in all sub regions or zip code areas. In the process of running the DPA, if one feasible solution can make the objective function have the optimal value, it must be treated as the optimal solution. And there is no evidence to show that we should add constraints in the model to avoid the problems discussed above.

In the next chapter, an econometric strategy will be introduced, and it can be used to forecast the installed machine count at the Sub Region Level and the Zip code Level, and this econometric forecast strategy would allow us to worry less about the issues mentioned above. However, the strategy would only be realized if and only if there would be more data available to us, and given the limited available data (OFN, EFR, and EIMC), the Poisson forecast strategy is the best strategy we can develop to forecast the installed machine count in the U.S.

Econometric Forecast Strategy

Recall, we used ORS to obtain the FIMC at the Sub Region and the Zip code Levels (see chapter, ***Poisson Forecast Strategy***). Since the operational mechanism of ORS is to allocate the installed machine count based on the data or the parameter, OFN, this mechanism may induce the possible problems mentioned in **Discussion of the Poisson Forecast Strategy**. We would like to find a way to improve the performance of forecasting installed machine count at the Sub Region Level and the Zip code Level. According to the literature (Tessier and Armstrong, 1977; Fomby *et al*., 1984; Hendry and Clements, 1994), econometric analysis can reduce the forecast uncertainty resulted from model structure and parameter uncertainty, and, hence, improve forecasting at the Sub Region Level and the Zip code Level.

In this chapter, we present an econometric forecast strategy to forecast the installed machine count at the Sub Region Level and the Zip code Level such that we can improve the accuracy of forecast results at those levels. The econometric forecast strategy is constituted of three parts:

- Missing Data Treatment (MDT)
- Feature Selection (FS)
- Two-Stage Econometric (TSE) Strategy

According to the literature (Gujarati, 2003; Lewis-Beck *et al*., 2003), the predictors or variables which will be used in the econometric forecast strategy can be categorized as endogenous predictors and exogenous predictors. The endogenous predictors are predictors whose values depend on the installed machine count, and in the forecast problem, the OFN is the only endogenous predictor. Exogenous predictors are predictors whose values are independent of, but, to some extent, can help determine the installed machine count. The exogenous predictors in the forecast problem:

- Predictors directly determining the installed machine count, such as the number of IBM customers.
- Predictors indirectly determining the installed machine count
    - Economic situations, such as GDP per capita, business output value, etc.
    - Social situations, such as population, volume of IT human resources, etc.

We already have the data of OFN at the Sub Region Level and the Zip code Level. However, at this time, the econometric forecast strategy is only a recommendation since we do not have enough data of the exogenous predictors to apply it to forecast the installed machine count at the Sub Region Level and the Zip code Level.

To our knowledge, it is quite probable that the data of the exogenous predictors available in the future would have missing items (for example: IBM has the data of populations of 2,000,000 urban areas over the world, but it does not know the populations of rural areas.). Therefore, we present the Missing Data Treatment (MDT), in this chapter, in case the future data of exogenous predictors would have missing items.

After having the complete data set of each exogenous predictor by utilizing MDT, we would need to decide which exogenous predictor should be built into the TSE Strategy, and, hence, we need to apply Feature Selection (FS) to make the decision. In this chapter, we will present three alternative approaches to do FS.

The TSE Strategy contains two stages. At the first stage, we build the Multiple Regression (MR) model to estimate the regression coefficients and constant. Here, we use the data of machines which have accurate EIMC at the National Level (Recall we used 2-HTPPM to find the accurate EIMC at the National Level in the chapter, ***Poisson Forecast Strategy***) as the sample data of the dependent variable to obtain estimates of the regression coefficients. At the second stage, we build the Constrained Least Square Regression (CLSR) models at the Sub Region Level and the Zip code Level to get the FIMC at those levels. Here, the constraint in the CLSR model is the *Conservation Constraint* (see chapter, ***Poisson Forecast Strategy***), and this constraint allows us to make sure the sum of the FIMC of each type of machine at each level (except at the National Level) is the FIMC at the higher level.

Missing Data Treatment

Missing Data Treatments (MDT) plays an important role in dealing with the data of the predictors having missing items. According to research done by Roderick Little and Donald Rubin, and Paul Allison (Little and Rubin, 1987; Alllison, 2001), there are many kinds of treatments available; however, two modern approaches, Maximum Likelihood (ML) and Multiple Imputation (MI), perform better in avoiding biased results than any other method, and, hence, would be candidates in the future research. According to research experiences stated in the literature (Raghunathan, 2004; Howell, 2009), SPSS Statistic is good at dealing with ML, while SAS is good at dealing with MI. In the future research, we would like to combine these two approaches to take advantage of each approach.

Missing Data Mechanism

According to Roderick Little, Donald Rubin, and Paul Allison (Little and Rubin, 1987; Alllison, 2001), there is an important prerequisite to use either ML or MI: the missing data must be missing at random (MAR). The missing data which is MAR suggests that the "missingness" does not depend on the value itself. For example, the data of the populations of rural areas are missing, but the "missingness" does not depend on the values of the populations in these areas. Recall, we would like to have the data of exogenous predictors, such as population, number of customers, and GDP per capita. Since it is common sense that the missing data of these exogenous predictors least likely depends on the value itself, we can reasonably expect the "missingness" of data of these exogenous predictors is MAR. (One opposite example illustrated by David C. Howell (Howell, 2009) is that "if we are studying mental health and people who have been diagnosed as depressed are less likely than others to report their mental status, the data are not missing at random.").

Missing Data Strategy in Future Research

According to David C. Howell (Howell, 2009), there are many ways to use ML to process the missing data, but the most common and efficient approach is the Expectation-Maximization Algorithm (EMA). The EMA includes iterative expectation steps and maximization steps. In the expectation steps, we can use the known data to estimate the parameters, such as mean, variance, covariance, and so on; then, we can build the regression equations to impute the missing data based on the estimated parameters, and since this step aims to make a good match of estimated parameters by imputing data into missing slots, it is named as maximization step. We will do the expectation step, where we use the imputed values with already known values as a complete data to estimate the parameters, and then we will do the maximization step again. The EMA is going to stop when the estimated parameters obtained from the two steps converge.

The alternative approach to ML is MI, which is basically imputing the missing values based on the currently existing values. According to Roderick Little and Donald Rubin, Paul Allison, Trivellore E. Raghunathan, and Ting Hsiang Lin (Little and Rubin, 1987; Alllison, 2001; Raghunathan, 2004; Lin, 2010), the most significant difference between MI and ML is that MI is not an iterative method but is generating multiple sets of complete data at the same time and combining all data sets to process the missing data.

In our future research, we can combine these two approaches to process the missing data: first, we will use the one iteration step of the EMA to generate multiple complete data sets. Here, we are going to use a sequence of regression equations to obtain the imputed data (we can bring auxiliary variables into the sequence), and randomly add errors to the imputed data; then, we can combine all of the sets to get the final estimated parameters. In a final step, we are going to use the final parameters to get the final complete data sets of the exogenous predictors. Based on our knowledge, both SPSS Statistics and SAS can finish the task.

Feature Selection

According to the literature (Derksen and Keselman, 1992; Bernstein *et al*., 1996; Harrell, 2010), we have three candidates of approaches coming from two well-known categories: Feature Ranking and Subset Selection. Stepwise Regression (SR) and Hierarchical Regression (HR) are popular representatives of Feature Ranking; Best Subset Selection (BSS) is a widely used approach belonging to the Subset Selection.

Stepwise Regression (SR) is the most classic approach in the field of feature selection. It allows us to check the necessity of exogenous predictors one by one in the first stage model of a TSE Strategy. However, many scholars have expressed negative opinions in using SR, not only because it may cause biased *p*-value and R^2, but also because the default model together with the design of the selection process can hurt the result of selection. In other words, the selection process is very likely damaged by the designer. Besides, if we only have limited exogenous predictors available in the first stage model of a TSE Strategy, and the results of the SR tells us to delete most of them, we would not be able to forecast the installed machine count.

Hierarchical Regression (HR) is often treated as an ideal substitute of SR. Basically, HR has two steps. In the first step, we can build the default model (including only significant exogenous predictors) of the first stage model of a TSE Strategy, where we can choose the exogenous predictors having strong correlation with the installed machine count to stay in the model. In the second step, we will check other predictors excluded in the first step to decide whether or not they should be kept in the model. Since we have a limited number of exogenous predictors, according to the logic of HR, we can exclude the predictor which is least significant in the model, and then retain others in the model.

Best Subset Selection (BSS) is an approach which allows us to divide the predictors into different sets, and use BSS to check their importance in the model. By selecting the "best set" of the exogenous predictors in the first stage model of a TSE Strategy, we can take into account the intercorrelation of exogenous predictors so that we can mostly avoid the bias resulting from "one by one checking". However, the applicability of BSS depends on the number of exogenous predictors. If we only have the data of a very limited number of exogenous predictors, we may not be able to divide them into very many sets, and therefore, we may not be able to use this approach.

In our future research, we can implement SAS, SPSS Statistics, and other stand-alone software to compare each of these three approaches, and choose the most appropriate approach to do the feature selection.

Two-Stage Econometric Strategy

In the future research, once we finish MDT and FS, we will enter into the most important step of the econometric forecast strategy, Two-Stage Econometric (TSE) Strategy. TSE Strategy has two stages: Multiple Regression (MR) model stage and Constrained Least Square Regression (CLSR) model stage.

Multiple Regression Model at the National Level

Recall, we used 2-HTPPM to obtain the accurate EIMC and the fixed EIMC at the National Level (see chapter, ***Poisson Forecast Strategy***). At the first stage of a TSE Strategy, we can use the accurate EIMC at the National Level as the dependent variables to estimate the regression coefficients and constant in MR model. Below is the first stage model of a TSE Strategy: (Suppose: there are I machines having the accurate EIMC at the National Level; after FS, we have 2 exogenous predictors that survive.)

$$Y_i = \alpha_0 + \alpha_1 \cdot A_i + \alpha_2 \cdot B_i + \beta \cdot OFN_i + \varepsilon_i \qquad \forall i$$

In the above model, Y_i is the machine *i*'s EIMC; ε_i , $i = 1, 2, ..., I$ is the forecast error; α_0 is the constant in a MR model, while $\beta, \alpha_1, and\ \alpha_2$ are coefficients for predictors; $A_i\ and\ B_i$ are exogenous predictors of machine *i*, while OFN_i is the number of machine *i's* failures at the National Level. Here, we plug the EIMC, OFN, and exogenous predictors of machines having accurate EIMC at the National Level into a MR model. Then, we can implement a regression package to obtain estimates of the coefficients and the constant:

$\hat{\alpha}_0$: estimate of the constant

$\hat{\beta}$: estimate of the coefficient of OFN

$\hat{\alpha}_1$ and $\hat{\alpha}_2$: estimates of the coefficients of the exogenous predictors

Constrained Least Square Regression Models at the Sub Region Level and the Zip code Level

The idea of using Constrained Least Square Regression (CLSR) model to forecast the FIMC is quite straightforward: in our forecast problem, we must have the FIMC at each level (except at the National Level) satisfy the *Conservation Constraint*, and with more structural information in the future, we would need to have more constraints in the forecast problem. According to the literature (Hendry and Clements, 1994; Golub and Van Loan, 1996), CLSR model allows us to build constraints into the traditional least square regression model. `

After finishing the first stage, we can have the following MR models with the estimates of regression coefficients and constant at the Sub Region Level and the Zip code Level:

At the Sub Region Level:

$$Y_{i,s} = \hat{\alpha}_0 + \hat{\alpha}_1 \cdot A_{i,s} + \hat{\alpha}_2 \cdot B_{i,s} + \hat{\beta} \cdot OFN_{i,s} \qquad \forall i, s$$

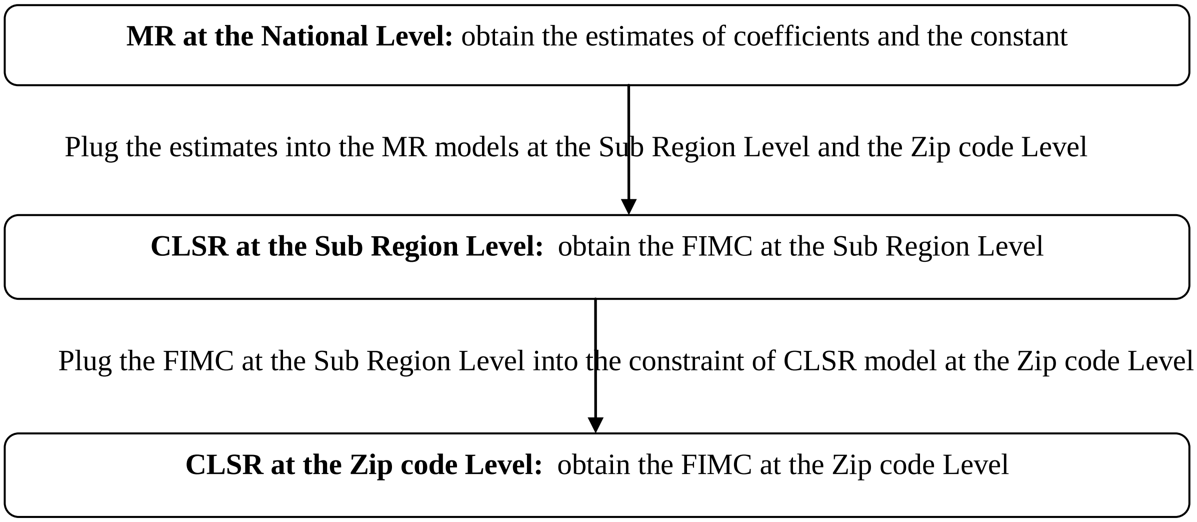
Here, $Y_{i,s}$ is the FIMC of machine *i* in sub region *s*; $A_{i,s}$ and $B_{i,s}$ are exogenous predictors of machine *i* in sub region *s*; $OFN_{i,s}$ is the number of observed machine *i*'s failures in sub region *s*.

At the Zip code Level:

$$Y_{i,z} = \hat{\alpha}_0 + \hat{\alpha}_1 \cdot A_{i,z} + \hat{\alpha}_2 \cdot B_{i,z} + \hat{\beta} \cdot OFN_{i,z} \qquad \forall i, z$$

Here, $Y_{i,z}$ is the FIMC of machine *i* in zip code area *z*; $A_{i,z}$ and $B_{i,z}$ are exogenous predictors of machine *i* in zip code area *z*; $OFN_{i,z}$ is the number of observed machine *i*'s failures in zip code area *z*.

Then we can build the CLSR models to forecast the installed machine count, and the forecast procedure is shown in **Fig.2**.



MR at the National Level: obtain the estimates of coefficients and the constant

Plug the estimates into the MR models at the Sub Region Level and the Zip code Level

CLSR at the Sub Region Level: obtain the FIMC at the Sub Region Level

Plug the FIMC at the Sub Region Level into the constraint of CLSR model at the Zip code Level

CLSR at the Zip code Level: obtain the FIMC at the Zip code Level

Fig. 2: The Forecast Procedure for the Econometric Forecast Strategy

First, we will use the CLSR model to forecast the installed machine count at the Sub Region Level, and the CLSR model at the Sub Region Level is: *(for each type of machine)*

$$\min \sum_{s=1}^{S} \left[Y_{i,s} - (\hat{\alpha}_0 + \hat{\alpha}_1 \cdot A_{i,s} + \hat{\alpha}_2 \cdot B_{i,s} + \hat{\beta} \cdot OFN_{i,s}) \right]^2$$

$$\text{s.t.} \quad \sum_{s=1}^{S} Y_{i,s} = N_i$$

We can see from the above model that the *Conservation Constraint* (see chapter, ***Poisson Forecast Strategy***) is built into the CLSR model as $\sum_{s=1}^{S} Y_{i,s} = N_i$, where N_i is the FIMC of machine *i* at the National Level.

Then, according to the same logic, we can build the CLSR model at the Zip code Level to obtain the FIMC at the Zip code Level: *(for each type of machine)*

$$\min \sum_{z=1}^{Z} \left[Y_{i,z} - (\hat{\alpha}_0 + \hat{\alpha}_1 \cdot A_{i,z} + \hat{\alpha}_2 \cdot B_{i,z} + \hat{\beta} \cdot OFN_{i,z}) \right]^2$$

$$\text{s.t.} \ \sum_{z=1}^{Z} Y_{i,z} = N_{i,s}$$

In the CLSR model at the Zip code Level, $N_{i,s}$ is the FIMC of machine *i* in the sub region *s* which contains the zip code areas from 1 to *Z*.

Discussion of the Econometric Forecast Strategy

The econometric forecast strategy is developed to improve the forecast accuracy at the Sub Region Level and the Zip code Level. However, the realization of implementing the strategy needs more available data which we do not have by the time of the thesis preparation, and, hence, we cannot know the performance of the econometric forecast strategy.

There is a reasonable method to test if the econometric forecast strategy can afford the accurate forecast at the Sub Region Level and the Zip code Level. And this method is to use random numbers instead of real data. The problem in generating random numbers is that we cannot know the distributions of the exogenous predictors, so the traditional random number generator may not be applicable in our case. After making a great effort in literature review, we find a good random number generator, the copula random number generator.

According to the literature (Hu *et al*., 2007; Strelen and Nassaj, 2007; Yan, 2007; Danaher and Smith, 2011), the copula generator has been widely used in academic research in finance, marketing, and other business sectors. And the great advantage of copula generator is that it does not need the distributions of the predictors. The copula generator can use the intercorrelations between the predictors and the dependent variable to generate the random numbers of the predictors. In our research, we can assume certain intercorrelations between the exogenous predictors and the real installed machine count. And then, we can use a copula generator function in MATLAB to generate the random numbers of the exogenous predictors so that we can test the performance of the econometric forecast strategy.

Summary

In this thesis, we present two forecast strategies to forecast the installed machine count of all type of machines at the National Level, the Sub Region Level, and the Zip code Level in the U.S.

Based on the available data, we develop a Poisson forecast strategy. And this strategy can be divided into two parts:

- 96% significant Two-Sided Hypothesis Test on Poisson Population Mean (2-HTPPM)
- Optimal Reallocation Strategy (ORS)

At the National Level, we use 2-HTPPM to test if the EIMC at the National Level is accurate. Then, we make the smallest effort to fix the inaccurate EIMC by checking whether the EIMC is greater than the upper bound or less than the lower bound of the 96% significant confidence interval of the installed machine count. Finally, we obtain the forecast result, the Forecasted Installed Machine Count (FIMC) at the National Level by combining the accurate EIMC and fixed ones.

At the Sub Region Level and the Zip code Level, we carry out an ORS. In this thesis, we take the ORS of one type of machine at the Sub Region Level as an example. First, we build an Integer-Nonlinear-Constrained (INLC) Optimization model to realize the ORS. However, the level of difficulty in solving the problem combining integer optimization and nonlinear optimization is too high to realize the ORS efficiently. Therefore, we use a Dynamic Programming Algorithm (DPA) to solve the INLC Optimization model. Then, we use the data of OFN, EFR, and EIMC of one type of machine to obtain the forecast results, FIMC, at the Sub Region Level and the Zip code Level.

However, there are some data-related and structural problems which may cause inaccurate forecast results when applying a Poisson forecast strategy to forecast the installed machine count at the Sub Region Level and the Zip code Level. In the future research, in an attempt to avoid these problems, we present an econometric forecast strategy. However, this strategy can be realized if and only if we can have more available data in the future. The econometric forecast strategy has three parts:

- Missing Data Treatment (MDT)
- Feature Selection (FS)
- Two-Stage Econometric (TSE) Strategy

Here, we define the data in the future research as endogenous predictors and exogenous predictors. The OFN is the only endogenous predictor in the forecast problem, while we still have exogenous predictors waiting for more available data.

To cope with the possible situation that the data sets might have missing items, we first present a Missing Data Treatment (MDT). In this section, we discuss the missing data mechanism of the data available in future. And we present a method to combine Maximum Likelihood (ML) and Multiple Imputation (MI) approaches, using sequential regression equations to process the possible missing data.

In Feature Selection (FS), we present three possible approaches. Stepwise Regression (SR) is the classic approach which is the most economical one to realize. However, it has some obvious disadvantages which may induce an error in choosing the significant exogenous predictors in our future research. Hierarchical Regression (HR) is widely known as the ideal substitute of SR. HR allows us to keep the best predictors in the model and check the significance of the exogenous predictors more effectively. Best Subset Selection (BBS) is the best approach available. By using BBS, we can divide the exogenous predictors into several groups, and check the significance by groups. However, both HR and BBS are difficult to realize if there is not enough data available in the future, and the available software to solve HR and BBS is limited. In the future research, we can compare the three approaches, and use the one which can best fit the forecast problem.

In TSE Strategy, we first build a Multiple Regression (MR) model for each type of machine at the National Level, and plug the accurate EIMC, OFN, and exogenous predictors into the model to obtain the estimates of coefficients and constant. Then, in the second stage, we build Constrained Least Square Regression (CLSR) models and embed the *Conservation Constraint* (see chapter, ***Poisson Forecast Strategy***) into the models. We first build a CLSR model at the Sub Region Level. After we obtain the FIMC at the Sub Region Level, we can build a CLSR model to get the FIMC at the Zip code Level. We propose a good method to test the performance of the econometric forecast strategy, generation of random numbers to substitute for real data. Due to the situation that we cannot know the distributions of the exogenous predictors, we propose a copula random number generator to accomplish the generation.

Conclusion

A significant contribution made through this thesis is to demonstrate how significance-based confidence intervals can provide viable constraints to an allocation process of an entity whose amount and location are uncertain. A hierarchical approach is taken, disaggregating an entity for which there is higher certainty to finer grain geographic regions for which the amount and the geographic positioning of that entity have larger uncertainty. A suitable alternative to the methodology outlined in this thesis is not found in the applicable literature.

Adopting an approach quite different from traditional statistical forecast strategies, such as regression, this thesis presents a Poisson forecast strategy which is a combination of statistical theories and optimization methodologies. With limited data (observed machine failures, engineering machine failure rate, and estimation of the number and locations of machines installed), the Poisson forecast strategy can accomplish forecasting the number and locations of machines installed that most strongly support the occurrence of observed machine failures. At the National Level, a hypothesis test on Poisson population mean is applied to fix the estimation of the number and locations of machines installed into a confidence interval which can support the occurrence of observed machine failures. At the Sub Region and the Zip code Levels, a forecast is accomplished through finding optimal number and locations of machines installed which maximize the probability of occurrence of machine failures. The reallocation of machines is accomplished using a Dynamic Programming Algorithm, applied to decompose the problem into steps such that a difficult Integer-Nonlinear-Constrained Optimization problem can be solved very efficiently.

This thesis also presents an econometric forecast strategy combining a Missing Data Treatment, a Feature Selection, and a Two-Stage Econometric Strategy to improve the accuracy of forecasting at the Sub Region and the Zip code Levels. However, to realize this econometric forecast strategy, there must be available data of exogenous predictors in future research. A treatment of missing data can be applied if the future data has missing items. Unlike traditional missing data techniques, a Missing Data Treatment combines a Multiple Imputation technique and a Maximum Likelihood technique and utilizes sequential regression equations to process missing items. This Missing Data Treatment can avoid a biased result as much as possible. A Feature Selection can be implemented to select significant predictors among all exogenous predictors whose data may be available in future research. A Feature Selection provides three alternative approaches, Stepwise Regression, Hierarchical Regression, and Best Subset Selection. Each approach has its merits and disadvantages, and the Feature Selection can choose the one which fits the research in future. A Two-Stage Econometric Strategy has two stages of models. The first stage model is a Multiple Regression model which uses data at the National Level to obtain estimates of regression coefficients and constant, and, by building the first stage model, a biased result caused by inaccurate estimates of parameters can be avoided as much as possible. The second stage model proposed is a Constrained Least Square Regression model. This model combines traditional Least Square Regression and a Conservation Constraint, and, hence, can make sure forecasted number and locations of machines installed would be optimal.

It is envisioned that this econometrics-based set of techniques would provide a “tuning” mechanism for the Poisson confidence interval constrained allocation model in this thesis.

Suitably chosen exogenous predictors would help stabilize the forecasting method that is otherwise at the mercy of inaccurate internal (exogenous) data.

Reference

Allison, P. D. (2001). **Missing Data**. Thousand Oaks, CA, Sage Publications, Inc.

Benavides, N. L., Carr, R. D. and Hart, W. E. (2007). **Python Optimization Modeling Objects (Pyomo)**. from https://projects.coin-or.org/Coopr/export/1013/coopr/branches/coopr_dev/doc/pyomo-sand07.pdf.

Bernstein, I., Conroy, R. and Harrell, F. (1996). **Ira Bernstein, Ronan Conroy, and Frank Harrell Comments**. from http://groups.google.com/group/sci.stat.math/tree/browse_frm/month/1997-06/5b52e8f6f2bb1c25?rnum=11&lnk=nl.

Danaher, P. J. and Smith, M. S. (2011). **Modelling Multivariate Distributions Using Copulas: Applications in Marketing**. Marketing Science **30**(1): 4-21.

Denardo, E. V. (1982). **Dynamic Programming: Models and Applications**. Engelwood Cliffs, NJ, Prentice-Hall, Inc.

Derksen, S. and Keselman, H. J. (1992). **Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables**. British Journal of Mathematical and Statistical Psychology **45**(2): 265-282.

Fomby, T. B., Hill, R. C. and Johnson, S. R. (1984). **Advanced Econometric Methods**. New York, NY, Springer-Verlag.

Golub, G. H. and Van Loan, F. V. (1996). **Matrix Computation**. Baltimore, MD, The Johns Hopkins University Press.

Gujarati, D. N. (2003). **Basic Econometrics**. New York, NY, McGraw-Hill Companies, Inc.

Harrell, F. E. Jr. (2010). **Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis**. New York, NY, Springer-Verlag New York, Inc.

Hendry, D. F. and Clements, M. P. (1994). **Can Econometrics Improve Economic Forecasting?** Swiss Journal of Economics and Statistics (SJES) **130**(III): 267-298.

Howell, D. C. (2009). **Treatment of Missing Data**. from http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Missing.html.

Hu, X., He, J. and Ly, H. (2007). **Generating Multivariate Nonnormal Distribution Random Numbers Based on Copula Function.** Journal of Information and Computing Science **2**(3): 191-196.

Johnson, R. A. (2005). **Miller and Freund's Probability and Statistics for Engineers**. Upper Saddle River, NJ, Pearson Prentice Hall.

Kleywegt, A. J. and Shapiro, A. (2001). **Stochastic Optimization**. Chapter 102 in Handbook of Industrial Engineering: Technology and Operations Management, editor: Salvendy, G. New York, NY, John Wiley & Sons: 2625-2650.

Lewis-Beck, M. S., Bryman, A. E. and Liao, T. F. (2003). **The SAGE Encyclopedia of Social Science Research Methods**, Thousand Oaks, CA, Sage Publications, Inc.

Lin, T. H. (2010). **A comparison of multiple imputation with EM algorithm and MCMC method for quality of life missing data**. Quality & Quantity **44**(2): 277-287.

Little, R. J. and Rubin, D. B. (1987). **Statistical Analysis with Missing Data**. New York, NY, John Wiley & Sons.

Raghunathan, T. E. (2004). **What do we do with missing data? Some options for anaysis of imcomplete data**. Annual Review of Public Health **25**: 99-117.

Ross, S. M. (2010). **Introduction to Probability Models**. New Delhi, India, Academic Press: An Imprint of Elsevier.

Strelen, J. C. and Nassaj, F. (2007). **Analysis and Generation of Random Vectors with Copula**. Proceedings of the 2007 Winter Simulation Conference: 488-496.

Sundström, O. and Guzzella, L. (2009). **A Generic Dynamic Programming Matlab Function**. Proceedings of the 18th IEEE International Conference on Control Applications & Intelligent Control: 1625-1630.

Tessier, T. H. and Armstrong, J. S. (1977). **Improving Current Sales Estimates with Econometric Models.** from http://www.forecastingprinciples.com/paperpdf/improvingsalesestimates.pdf

Winston, W. L. and Venkataramanan, M. (2003). **Introduction to Mathematical Programming Operations Research: Volume One**. Pacific Grove, CA, Brooks/Cole-Thomson Learning.

Yan, J. (2007). **Enjoy the Joy of Copula: With a Package Copula**. Journal of Statistical Software **21**(4): 1-21.

Zietz, J. (2007). **Dynamic Programming: An Introduction by Example**. The Journal Of Economic Education **38**(2): 165-186.

Appendix: Data and FIMC at the Zip code Level

Table 9: Data and FIMC for the zip code areas under *Sub Region I*

Zip code Area	OFN	EIMC	Reallocated Installed Machine Count (or FIMC)
1	0	2	0
2	1	1	5
3	0	2	0
4	0	1	0
5	0	11	0
6	1	1	5
7	0	1	0
8	0	1	0
9	0	1	0
10	0	1	0
11	0	1	0
12	0	1	0
13	0	2	0
14	0	2	0
15	0	1	0
TOTAL	2	29	10

Table 10: Data and FIMC for the zip code areas under *Sub Region II*

Zip code Area	OFN	EIMC	Reallocated Installed Machine Count (or FIMC)
1	0	1	0

2	0	6	0
3	0	1	0
4	0	3	0
5	1	2	5
6	0	1	0
7	0	1	0
8	1	1	5
9	0	1	0
10	0	1	0
TOTAL	2	18	10

Table 11: Data and FIMC for the zip code areas under *Sub Region IV*

Zip code Area	OFN	EIMC	Reallocated Installed Machine Count (or FIMC)
1	1	1	5
2	0	4	0
3	0	1	0
4	0	3	0
5	0	26	0
6	0	5	0
7	0	1	0
8	0	1	0
9	0	1	0
10	1	3	5
11	0	1	0
12	0	1	0
13	0	1	0
14	2	1	10

15	0	1	0
16	2	2	11
17	1	2	5
18	0	1	0
19	0	7	0
20	0	2	0
21	0	1	0
22	0	2	0
TOTAL	7	68	36

Table 12: Data and FIMC for the zip code areas under *Sub Region V*

Zip code Area	OFN	EIMC	Reallocated Installed Machine Count (or FIMC)
1	0	1	0
2	1	1	5
3	0	2	0
4	0	1	0
5	1	1	5
6	0	2	0
7	0	3	0
8	0	1	0
9	0	2	0
10	0	2	0
11	0	1	0
12	2	14	11
13	0	2	0
14	0	10	0
15	0	1	0

16	0	1	0
17	0	1	0
18	0	1	0
19	0	3	0
20	0	1	0
21	0	1	0
22	0	2	0
23	1	2	5
24	0	1	0
25	0	2	0
26	1	1	5
27	0	1	0
28	0	1	0
29	1	1	5
TOTAL	7	63	36

Table 13: Data and FIMC for the zip code areas under *Sub Region VI*

Zip code Area	OFN	EIMC	Reallocated Installed Machine Count (or FIMC)
1	0	1	0
2	0	3	0
3	0	1	0
4	2	1	10
5	2	1	10
6	2	1	11
7	0	2	0
8	0	13	0

9	0	13	0
10	0	1	0
11	1	3	5
12	0	1	0
TOTAL	7	41	36

Table 14: Data and FIMC for the zip code areas under *Sub Region VII*

Zip code Area	OFN	EIMC	Reallocated Installed Machine Count (or FIMC)
1	0	1	0
2	0	1	0
3	1	1	5
4	0	2	0
5	1	7	6
6	0	1	0
7	0	1	0
8	1	1	5
9	1	32	5
10	0	1	0
11	2	28	11
12	0	1	0
13	1	1	5
14	1	1	5
TOTAL	8	79	42

Table 15: Data and FIMC for the zip code areas under *Sub Region VIII*

Zip code Area	OFN	EIMC	Reallocated Installed Machine Count (or FIMC)
1	2	13	10
2	0	4	0
3	0	4	0
4	0	2	0
5	1	1	5
6	1	3	5
7	0	1	0
8	0	3	0
9	0	1	0
10	0	1	0
11	0	1	0
12	0	1	0
13	3	19	16
14	3	4	16
15	0	1	0
TOTAL	10	59	52

Table 16: Data and FIMC for the zip code areas under *Sub Region X*

Zip code Area	OFN	EIMC	Reallocated Installed Machine Count (or FIMC)
1	1	1	5
2	3	3	16
3	0	1	0
4	0	1	0

5	0	1	0
6	0	1	0
7	0	1	0
8	1	1	5
9	0	2	0
10	1	6	5
11	0	6	0
12	0	5	0
13	5	8	27
14	1	7	5
15	0	1	0
16	0	1	0
17	0	11	0
18	0	1	0
TOTAL	12	58	63

Table 17: Data and FIMC for the zip code areas under *Sub Region XI*

Zip code Area	OFN	EIMC	Reallocated Installed Machine Count (or FIMC)
1	1	1	5
2	2	2	11
3	1	1	5
4	0	2	0
5	0	2	0
6	0	1	0
7	1	1	5
8	0	1	0
9	1	1	5

10	1	1	5
11	2	1	11
12	1	1	5
13	3	13	17
14	0	2	0
15	0	1	0
16	0	1	0
17	0	16	0
18	0	2	0
19	0	2	0
20	1	2	5
21	0	1	0
22	0	1	0
23	0	1	0
TOTAL	14	57	74

Table 18: Data and FIMC for the zip code areas under *Sub Region XII*

Zip code Area	OFN	EIMC	Reallocated Installed Machine Count (or FIMC)
1	0	15	0
2	8	31	44
3	0	2	0
4	0	4	0
5	0	1	0
6	0	2	0
7	1	2	5
8	0	1	0
9	1	1	5

10	1	1	5
11	0	1	0
12	1	1	5
13	0	5	0
14	0	1	0
15	0	4	0
16	0	2	0
17	1	1	5
18	0	2	0
19	2	1	10
20	2	2	11
21	0	1	0
TOTAL	17	81	90

Table 19: Data and FIMC for the zip code areas under *Sub Region XIII*

Zip code Area	OFN	EIMC	Reallocated Installed Machine Count (or FIMC)
1	3	2	16
2	0	12	0
3	0	2	0
4	0	3	0
5	0	9	0
6	2	2	10
7	2	1	10
8	9	12	49
9	1	1	5
10	0	2	0
11	2	2	11

12	1	7	5
13	1	1	5
14	0	1	0
15	0	3	0
TOTAL	21	60	111

Table 20: Data and FIMC for the zip code areas under *Sub Region XIV*

Zip code Area	OFN	EIMC	Reallocated Installed Machine Count (or FIMC)
1	0	1	0
2	0	1	0
3	0	1	0
4	2	5	16
5	1	7	6
6	0	2	0
7	0	3	0
8	0	7	0
9	0	10	0
10	1	2	6
11	0	2	0
12	1	1	6
13	0	1	0
14	6	16	67
15	1	34	6
16	1	2	6
17	0	1	0
18	0	1	0
19	1	1	6

20	1	6	6
21	0	1	0
22	1	1	6
23	0	1	0
24	1	1	6
25	0	2	0
26	0	5	0
27	1	1	6
28	1	2	6
29	2	2	16
30	1	1	6
31	0	1	0
32	0	2	0
33	1	9	6
TOTAL	23	133	177

Table 21: Data and FIMC for the zip code areas under *Sub Region XV*

Zip code Area	OFN	EIMC	Reallocated Installed Machine Count (or FIMC)
1	0	2	0
2	0	1	0
3	0	1	0
4	6	7	43
5	1	2	6
6	0	8	0
7	6	26	67
8	0	6	0

9	2	11	11
10	0	1	0
11	0	6	0
12	0	1	0
13	0	4	0
14	0	1	0
15	0	4	0
16	0	2	0
17	3	5	18
18	0	4	0
19	0	2	0
20	0	1	0
21	4	24	24
22	0	14	0
23	0	2	0
24	5	3	32
25	0	4	0
26	0	1	0
27	0	2	0
TOTAL	27	145	201

Vita

Mr. Zheng Shi was born on Oct 8, 1986 in Shanxi province, P. R. China. In the year of 2009, he obtained a Bachelor of Economics in Finance at Nankai University and a Bachelor of Arts in English Literature at Tianjin University.

In the year of 2010, Mr. Zheng Shi began his master study in the Dept. of Industrial and Systems Engineering, Lehigh University, and pursued the degree of Master of Science in Industrial and System Engineering.

Ever since the April of 2011, Mr. Zheng Shi has been working with Prof. George R. Wilson on the IBM forecast project, and on July 1, 2012, he was employed by IBM.