Analytic framework for TRL-based cost and schedule models

by

Bernard El-Khoury

B.S. in Engineering, Ecole Centrale Paris (2009)

Submitted to the Engineering Systems Division
in partial fulfillment of the requirements for the degree of

Master of Science in Technology and Policy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2012

Author...

Technology and Policy Program, Engineering Systems Division
August 22, 2012

Certified by...

C. Robert Kenley
Research Associate,
Lean Advancement Initiative
Thesis Supervisor

Certified by...

Deborah Nightingale
Professor of the Practice of Aeronautics and Astronautics and Engineering Systems
Director, Sociotechnical Systems Research Center
Thesis Supervisor

Accepted by...

Joel P. Clark
Professor of Materials Systems and Engineering Systems
Acting Director, Technology and Policy Program

Analytic framework for TRL-based cost and schedule models

by

Bernard El-Khoury

Submitted to the Engineering Systems Division
on August 22, 2012, in partial fulfillment of the
requirements for the degree of
Master of Science in Technology and Policy

Abstract

Many government agencies have adopted the Technology Readiness Level (TRL) scale to help improve technology development management under ever increasing cost, schedule, and complexity constraints. Many TRL-based cost and schedule models have been developed to monitor technology maturation, mitigate program risk, characterize TRL transition times, or model schedule and cost risk for individual technologies as well technology systems and portfolios. In this thesis, we develop a 4-level classification of TRL models based on the often-implicit assumptions they make. For each level, we clarify the assumption, we list all supporting theoretical and empirical evidence, and then we use the same assumption to propose alternative or improved models whenever possible. Our results include a justification of the GAO's recommendations on TRL, two new methodologies for robust estimation of transition variable medians and for forecasting TRL transition variables using historical data, and a set of recommendations for TRL-based regression models.

Thesis Supervisor: C. Robert Kenley
Title: Research Associate, Lean Advancement Initiative

Thesis Supervisor: Deborah Nightingale
Title: Professor of the Practice of Aeronautics and Astronautics and Engineering Systems
Director, Sociotechnical Systems Research Center

Acknowledgments

I would like to express my gratitude to my supervisor, Bob Kenley, for his valuable comments, guidance, patience, and encouragement throughout my study. Without him constantly reminding me to focus on the core topic and leave the other ideas for later research, this thesis would have never been finished on time.

I also appreciate the support, sacrifice, care and the unconditional love from my family.

List of Abbreviations and Acronyms

AD^2: Advancement Degree of Difficulty

AHP: Analytic Hierarchical Process

ANOVA: Analysis Of Variance

ATO: Acquisition Technology Objective

CI: Confidence Interval

CTE: Critical Technology Element

CSP trade space: Cost, Schedule, Performance trade space

DAG: Defense Acquisition Guidebook

DoD: Department of Defense

DoE: Department of Energy

EA: Enterprise Architecting

EMRL: Engineering and Manufacturing Readiness Level

FY: Fiscal Year

GAO: Government Accounting Office

HRL: Human Readiness Level

ID: Influence Diagram

ITAM: Integrated Technology Analysis Methodology

ITI: Integrated Technology Index

IRL: Integration Readiness Level

IRT: Independent Review Team

LRL: Logistics Readiness Level

MAE: Mean Absolute Error

MDA: Milestone Decision Authority

MDAP: Major Defense Acquisition Program

MRL: Manufacturing Readiness Level

NASA: National Aeronautics and Space Administration

NATO: North Atlantic Treaty Organization

NDI: Non-Developmental Item

OFE: Objective Function of Error

PBTs TRL: Process-Based Technologies TRL

PRL: Programmatic Readiness Levels

R^2: statistical coefficient of determination

R&D^3: Research and Development Degree of Difficulty

RMSE: Root Mean Squares Error

SRL: System Readiness Level

S&T: Science and Technology

TM: Technology Maturity

TML: Technological Maturity Level

TNV: Technology Need Value

TPRI: Technology Performance Risk Index

TRA: Technology Readiness Assessment

TRL: Technology Readiness Level

TTRL: Technology Transfer Readiness Level

UDF: User-Defined Function (in Microsoft Excel)

X_{i-j} : Transition time from TRLi to TRLj

WBS: Work Breakdown Structure

WSARA: Weapon Systems Acquisition Reform Act

WTRL: cost-Weighted TRL

Table of Contents

List of Tables

List of Figures

Chapter 1

Chapter 1. Introduction and Research Motivation

More than ever, innovation and technology are key elements to the competitive edge in today's world. In order to sustain growth, companies - as well as state agencies - are forced to develop new technologies faster, cheaper, with less tolerance for risk.

However, technology development is highly unpredictable: not only is a project manager faced with "known unknowns" i.e. uncertainties that can be roughly estimated and controlled, he also has to deal with "unknown unknowns" which are completely unforeseeable uncertainties due to the very nature of developing a new technology (no historical data, analogous data, or reliable expert opinions). Industry has adopted many technology management frameworks to address this challenge, such as technology roadmapping, technology benchmarking, technology watches, and technology risk management (Foden and Berends, 2010). The aim is to control key factors such as cost, schedule, technology maturity, and manufacturability.

For US governmental agencies (DoD, DoE, NASA, Army), the challenges are bigger and the stakes are higher. For instance, the Department of Defense (DoD) Acquisition Program (1) develops a very large portfolio of technologies, (2) develops a lot of high complexity system technologies (3) manages a budget of a several hundred billion dollars (GAO, 2009) in a monopsonistic contracting environment with little market competition, (4) suffers from frequent design changes due to modifications in requirements, (5) and is under constant pressure to accelerate the development of technologies required for pressing national security issues.

DoD has poorly addressed its challenges for many technologies. The above constraints often result in cost overruns. DoD has endured a cost growth of $296 billion on its 96 major acquisition programs in fiscal year (FY) 2008 (GAO, 2009). More than 2 out of 3 programs suffered a cost growth. In another performance audit of 72 DoD programs, the Government Accounting Office (GAO) reported more than 25% cost growth for 44% of the programs in FY07 with an average schedule delay of 21 months (GAO, 2008). Other GAO reports (GAO, 1999, 2006, 2008) also noted failure to meet capability and performance requirements. DoD also faces problems with increasing systems complexity and integration. The F-35 is such an example: not only did it face serious cost problems (Pentagon officials disclosed a cost overrun of more than 50 percent (Shalal-Esa, 2011)), a Pentagon study identified 13 areas of concern (some of which are critical) due to system complexity and concurrency (DoD, 2011).

As awareness of the problem increased and as defense budget continued to be cut, the defense acquisition community was under greater political pressure to improve the management of new technologies. In May 2009, President Obama signed the Weapon Systems Acquisition Reform Act (WSARA) to end "waste and inefficiency". Some of the Act's key provisions included (WSAR Act, 2009):

- Appointment of a Director of Cost Assessment and Program Evaluation (CAPE), who will communicate directly with the Secretary of Defense and Deputy Secretary of Defense, and issue policies and establish guidance on cost estimating and developing confidence levels for such cost estimates; (section 101)

- Requirement that the Director of Defense Research and Engineering periodically assess the technological maturity of Major Defense Acquisition Programs (MDAPs) and annually report his/her findings to Congress; (section 104)
- Requirement that DoD revise guidelines and tighten regulations pertaining to conflicts of interest of contractors working on MDAPs. (section 207)

The Government Accountability Office also weighed in on the issue. In a now famous report, GAO concluded that "Maturing new technology before it is included on a product is perhaps the most important determinant of the success of the eventual product—or weapon system", GAO went further by encouraging the use of "a disciplined and knowledge-based approach of assessing technology maturity, such as TRLs, DoD-wide" (GAO,1999). Technology Readiness Levels (TRLs) are a 1-to-9 scale developed by NASA (and used only by NASA at the time) that describes the maturity of a technology with respect to a particular use.

GAO mainly suggested the use of TRLs to make sure that technologies are mature enough before integrating them in the acquisition cycle (GAO, 1999). DoD later required the use of TRLs as criteria to pass Milestones B and C in the Acquisition cycle (DoD TRA Deskbook, 2009).

Although those practices advocated by WSARA and the GAO reports already have a positive effect on managing technology maturation and on controlling cost growth and schedule delays, they make minimal use of TRL measurements for cost and schedule modeling. TRLs are practically only used by Science and Technology Organizations (STOs) to have an idea about the technology's riskiness by looking at its maturity (Graettinger et al, 2003), or they are used by the DoD acquisition in Technology Readiness Assessments (TRAs) to make sure the technology has matured enough to pass certain milestone (DoD TRA Deskbook, 2009). Furthermore, Azizian et al. (2009), Cornford and Sarsfield (2004), Nolte (2008), and Fernandez (2010) point out that TRL is not well integrated into cost, schedule, and risk modeling tools. The general aim of this research is to explore the total potential benefit of integrating TRLs, cost, and schedule into a single technology management framework. In the spirit of the GAO and WSARA recommendations of keeping tighter control of technology maturity and of improving cost/schedule estimation and management, this thesis takes a more fundamental theoretical approach to study all TRL-based cost and schedule models.

Many models have already been proposed to use TRLs for cost and schedule modeling for individual technologies (e.g., GAO, 1999; Smoker and Smith, 2007; Dubos and Saleh, 2008; and Conrow, 2011) as well as technology systems and portfolios (e,g., Dubos and Saleh, 2010; Lee and Thomas, 2003; and Sauser at al., 2008). However, those models are based on different assumptions (which can lead to different results). Some models are based on a theoretical foundation, while other ones make implicit assumptions and don't justify the approach as long as it generates robust or useful results.

This research builds on an earlier paper by El-khoury and Kenley (2012), and goes into detail in constructing a unifying framework for all TRL models based on the assumptions that they make. The framework (seen below in figure 1.1) divides all TRL cost/schedule models into 4 categories based on how strong the assumptions about TRL they make. For each level in the framework, we will (1) state the assumption, (2) list the available literature relevant to that level of assumption, (3) look at theoretical evidence supporting this assumption, (4) look at empirical evidence supporting this assumption, and finally we will (5) propose new methodologies that make better use of the assumptions whenever possible.

Such a framework has many benefits: first it puts all TRL-based models in one clear theory. Second, it makes explicit the assumption used by a model, which helps to tell what model has stronger theoretical or empirical foundations. Third, once we know the assumptions being made, we can propose modifications to improve the model, or alternative models that make full use of the assumptions. Finally, the model goes beyond the statement and the clarification of the assumptions, to backing those assumptions by theory, and more importantly by empirical data whenever possible. Similarly, we will propose new methodologies at Levels 2,3, and 4 of the model, and the improvement brought by those methodologies will be quantitatively tested.

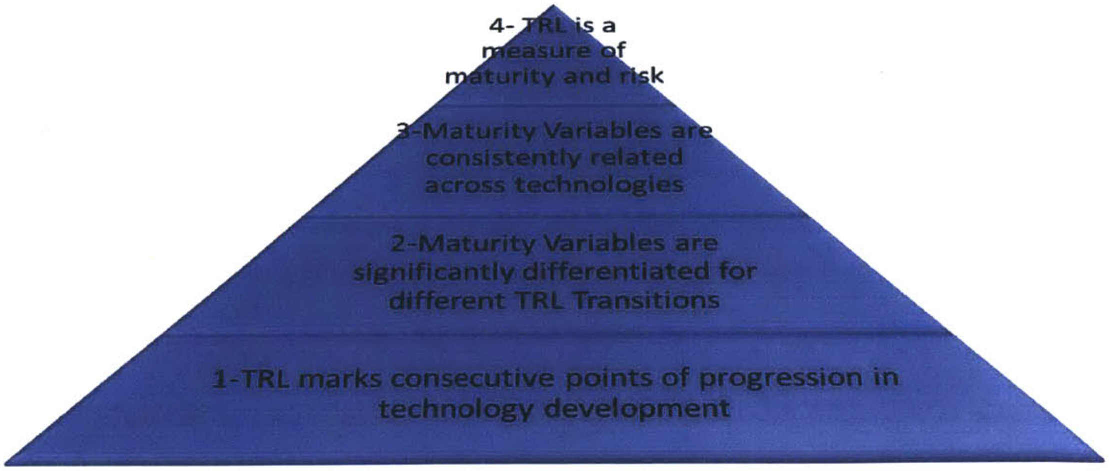

4- TRL is a measure of maturity and risk
3-Maturity Variables are consistently related across technologies
2-Maturity Variables are significantly differentiated for different TRL Transitions
1-TRL marks consecutive points of progression in technology development

Figure 1:1 The four-level framework for TRL-based cost and schedule models

In addition to the above benefits of the framework, the models developed at different levels will allow us to answer practical research questions at each level of the framework.

- With a level-1 model, we will answer the research question:
    - *Does available evidence support or contradict GAO's recommendation on pre-production maturation?*
- At level 2, we will answer the questions:
    - *Does it make statistical sense to look at the distribution of each TRL transition time separately?*
    - *If yes, is there a way to improve the accuracy and fidelity of the estimates and confidence intervals?*
- At level 3, we will develop a model that answers:
    - *If we use historical data of a technology's development, can we significantly improve the accuracy of the technology schedule forecast over level-2 models?*
- And at level 4, we will answer:
    - *Out of the available methodologies, what is the best one to use in performing regression of cost/schedule against TRL?*

Answering those questions would have direct practical implications for project managers. In general, improving cost and schedule modeling can lead to (1) reduction in cost and schedule uncertainty, which itself can lead to (2) reduction of overall project cost and duration or (3) reduction of cost and schedule overruns. Other benefits include (4) a better view of integrated cost, schedule, and maturity instead of a non-integrated view of cost and schedule separately, and (5) increased control over the project by having a high number of intermediate reference points, which would allow a better cost/schedule arbitrage within a single project, or across different projects.

In opposition to this approach, Cornford and Sarsfield (2004) claim that TRL cannot be integrated in cost/risk models because of the high uncertainty in the TRL scale that could propagate into major errors in the cost estimations. Although the TRL scale has many inherent weaknesses (discussed in more detail in section 3.3.3.2), this definitely should not prevent us from putting our full effort into looking for any potential significant benefit of this approach, however minor it may be. After all, even a small percentage of cost reduction will be multiplied by DoD's annual $300 billion in cost overruns. This is especially true when "Every dollar spent on inefficiencies in acquiring one weapon system is less money available for other opportunities" (GAO, 2006).

In Chapter 2, we scope the models by locating them within the risk management context, and then we justify our choice of the adopted class of models. In chapter 3, we explore the relation between technology maturity and TRL: we start by defining the multidimensional concept of technology maturity and relate it to the maturity measures found in the literature. Then we focus on the TRL scale itself by defining it, explaining how it is used in the acquisition lifecycle, and by looking at its most relevant characteristics. In chapter 4, we introduce the two datasets used in the thesis and we evaluate their quality. In chapter 5, we introduce the TRL models assumption-based framework and explain the logic behind it. In chapters 6 to 9, we present each of the 4 levels by stating the assumptions and listing the corresponding literature, then by looking for theoretical and empirical evidence supporting the assumption, and then by proposing new methodologies (especially in chapters 7 and 8). Finally in chapter 10, we conclude by summarizing our major results and by proposing directions for future research.

Chapter 2

Chapter 2. Research Scope

2.1 Placing this research in the larger context of risk management

While some project variables are beyond the control of the project managers such as budget cuts, others are within their reach (for example, Browning and Eppinger (2002) model how the choice of product development processes impacts project cost and schedule). Therefore, having better information means that the project managers can make better decisions to address risks in technology development. For instance, the manager can decide to drop a very high risk project in order to allocate more resources to the highest identified risk in a Major Defense Acquisition Program (MDAP), he can also engage in a risk reduction procedure, like rapid prototyping, or he can compare different system requirements, configurations, or acquisition strategies and select a baseline set of requirements, system configuration, and acquisition strategy that has an appropriate level of risk. More generally, proper risk management requires the best risk analysis tools. A wrong identification/analysis of risk means that risk management resources are not optimally allocated (Conrow, 2003).

We have already mentioned that the aim of this thesis is to propose a classification and improvement of TRL models. By doing so, we are improving the risk analysis part of the risk management process. The following diagram locates this research within the risk management context.

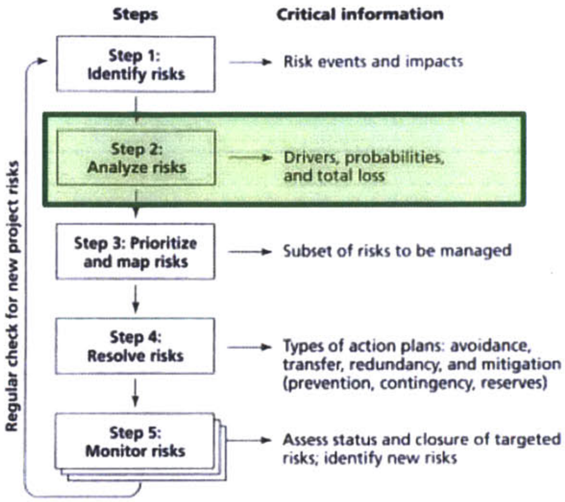

Steps
Critical information
Step 1: Identify risks
Risk events and impacts
Step 2: Analyze risks
Drivers, probabilities, and total loss
Step 3: Prioritize and map risks
Subset of risks to be managed
Step 4: Resolve risks
Types of action plans: avoidance, transfer, redundancy, and mitigation (prevention, contingency, reserves)
Step 5: Monitor risks
Assess status and closure of targeted risks; identify new risks
Regular check for new project risks

Figure 2:1 Simplified 5-step process of risk management (Smith and Merrit, 2002)

The below diagram summarizes how this research contributes to better technology management.

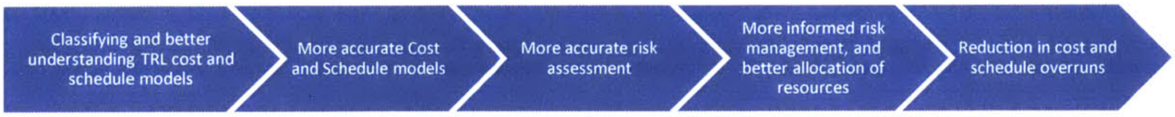

Classifying and better understanding TRL cost and schedule models
More accurate Cost and Schedule models
More accurate risk assessment
More informed risk management, and better allocation of resources
Reduction in cost and schedule overruns

Figure 2:2 Contribution of the research to better technology management

2.2 The choice of TRL-based cost and schedule models

It should be noted that acquisition programs entail many types of risks (for example design/engineering, manufacturing, support, technology, and threat, as defined in DoD5000.1 (2007)) and that there are multiple ways and approaches to model those risks. In our case however, we only look at cost and schedule risk types and we model them by using only TRL as an input (figure 2.3). Combined cost-schedule modeling (grey area in figure 2.3) is recommended for future research when data becomes available.

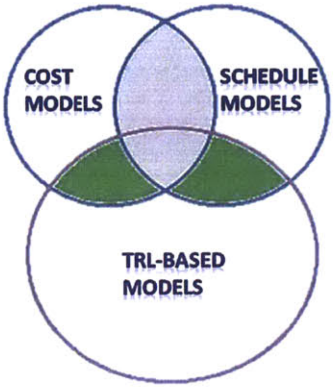


Figure 2:3 Types of models considered in the thesis

2.2.1 Reasons for choosing only Cost and Schedule risk types

Many risk types can be considered in risk management. However, we limit those to only cost and schedule for the following reasons:

- Cost and Schedule are traditionally the most important factors and have the highest weight both practically and politically (this is why they are directly addressed in the GAO reports and in WSARA).

-Those two are the most studied variables in the literature. We found that models in the literature either deal with schedule-related variables such as total development time (Dubos and Saleh, 2010), time to market (Kenley and Creque, 1999), schedule slippage (Dubos and

Saleh, 2008), risk of schedule slippage (Dubos and Saleh, 2008), or with cost-related variables such as total cost (Smoker and Smith, 2007), absolute cost growth (Lee & Thomas, 2003), and relative cost growth (Lee and Thomas, 2003).

-The program management literature traditionally reduces the major program variables to a Cost-Schedule-Performance (CSP) "trade space" (Defense Acquisition Guidebook, 2010), meaning that the project manager has a space of feasible (C,S,P) triplet solutions, and he manages the program by trading between those 3 variables to achieve the best possible (C,S,P) combination. However, microeconomic theory stipulates that P will get fixed to a high value early on, and that (C,S) will be the only variables to be traded throughout the project. The contracting environment will lead to an initial high P value that will be met, and as a consequence, C and S will tend to surpass the original estimates. In fact the Government is interested in a low C, low S, and high P, while the contractor is interested in a high C, high S, and high P. Since both parties only agree on a high Performance, and since there is little initial knowledge about the technical potential of the technology, P will get set to an unrealistically high value (Conrow, 2003). Now that the contractor is having cost and schedule overruns because of the high P, he has little incentive to go cheap or fast because the government usually assumes those risks. Conrow (1997) empirically confirmed this conclusion. Using data from 3 different studies (Perry 1971, Dews 1979, and Conrow 1995), he found that major military development programs met their Performance targets on average, while Cost and Schedule were often adjusted upwards to meet performance. Based on Conrow's conclusions, we can eliminate the Performance variable and limit our study to Cost and Schedule.

-Finally note that we decided to dismiss the bivariate Cost-Schedule models (i.e. models in the grey area in figure 2.3) that model Cost and Schedule together at the same time. Those represent a completely different class of models with a different set of assumptions. The tendency in those models would be to extend classical one-variable models to some bivariate normal (or lognormal) distribution of Cost and Schedule. However, Garvey (2000) warns that those distributions are not appropriate in representing Cost and Schedule. Such a simple distribution would reduce the interaction between cost and schedule to a simple linear coefficient of correlation that only says "high costs correspond to delayed schedules", or the opposite. It fails to capture the causation in cost-schedule decision making. It creates confusion between two types of cost/schedule correlation: (1) a post-decision-making positive correlation across projects (e.g. MDAPs and systems projects always consume more time and money than smaller projects), and (2) a pre-decision-making negative correlation within a project (i.e. the arbitrage done by the project manager between cost and schedule). On a more empirical level, Conrow (2003) notes that there are no bivariate distributions in the literature due to the lack of data. His own study shows that the bivariate lognormal distribution is a bad fit for Cost-Schedule, and that Cost-Schedule correlations are low (which is expected, since although they are related, the relation is NOT linear). For all the above reasons, we dropped joint cost-schedule models from our framework. The reader can refer to El-Khoury and Kenley (2012) on a method for bivariate cost-schedule modeling that takes the decision-making process into consideration.

2.2.2 Reasons for choosing TRL-based models only

In addition to the reasons that led to the choice of cost and schedule models, many practical reasons led to the choice of TRL-based models over other maturation measures:

-The TRL measurements are often already available. For all major DoD acquisition projects, the assessments have to be performed for the milestones (The undersecretary of defense, 2011), so there is no extra cost to perform the measurement. Furthermore, some projects (especially Major Defense Acquisition Programs) get more frequent Technology Readiness Assessments, which means more available TRL measurements.

-TRL scales have a widespread use across many agencies (e.g. DoD, DoE, NASA, FAA, and the Army) and countries (US, UK, Canada, EU), and the procedures of determining the scores are already established.

-There are available tools to determine TRL scores (Technology Readiness Calculator, Nolte, 2005) that further streamline and encourage its use. Hence the cost of integrating it into technology cost and schedule models will be minimal, while the benefit for the program manager is potentially high.

-Finally, and as a consequence of the above, TRL is the only maturity variable that has available data. As we mentioned earlier, it is important to note that this thesis is not only limited to the theoretical framework, it also uses a data-based approach to empirically validate the proposed models at each level in our framework. In fact, the literature points out a lot of weaknesses of using TRL in technology management (refer to section 3.3.3.2), and proposes many more developed alternative frameworks such as AD^2, RD^3, SRL, ITI (all discussed in detail in section 3.2.4), or combinations of those (Mankins, 1998; Nolte, 2008; Bilbro, 2008; and Sauser et al, 2008). However, there is no way of quantitatively evaluating those alternative approaches due to absence of data, as they are still not used by any of the major agencies. As a result, we look at models that use TRL as the only input. Future research can look into augmenting our models with those other measures once data becomes available.

Chapter 3

Chapter 3. Technology Maturity and the TRL scale

Before going into TRL-based technology forecasting models, it is important to understand technology maturity along with the different measures (especially TRL) of technology maturity.

The main purpose of this chapter is to define the "maturity" scope of this thesis by putting the TRL scale in the right context. This section can be also considered a literature review on the TRL scale and the measurements of technology maturity.

In subsection 3.1, we define the multidimensional concept of technology maturity, and then we locate the technology development phase (phase captured by the TRL scale) within the larger technology lifecycle.

In subsection 3.2, we introduce the concept of technology maturity measurement. We classify all the maturity measurements found in the literature (based on their relation with the NASA TRL), and see how those measures fit in the context of multidimensional maturity measurement.

Finally in subsection 3.3, we focus on the TRL scale. First, we propose a (more intuitive) definition of the TRL scale. Then, we describe how TRL is officially used in DoD's acquisition cycle, and we finish by looking at the TRL scale's properties and characteristics that will be relevant for the models developed in subsequent chapters.

3.1 Technology maturity

The purpose of this section is to locate TRL within the larger context of technology maturation. First, we look at the technology lifecycle that defines the technology evolution process and delimit the section captured by the TRL scale. Then we go deeper into the definition of technology maturity, its measurement, and its determining factors.

3.1.1 TRL's location in the technology lifecycle

In his book, Nolte (2008) presents a technology's "whale chart" biological lifecycle from initial idea to obsolescence.

Conception is when the initial thought or supposition appears. It is the sometimes long process of incubating the idea while unsure if it might work or not.

Birth is when the inventor realizes that the idea might work. It is when all the earlier concepts are developed enough so that an early application can be determined.

Childhood consists mainly of the extensive laboratory research whose main goal is to prove that the technology can actually work (even if it is only in a controlled environment in a laboratory)

Adolescence is when the developer is ready to release the first version of the technology product. This marks the transition from an environment focused on the science and technology behind the product, to an environment focused on the engineering challenges to make the product.

Adulthood is when the product has been successfully debugged and marketed, and that product adoption reaches its peak.

Maturity is when the market for the new product is already saturated. Suppliers can now only offer incremental innovations and product differentiations to keep market shares.

Old age is when the technology is old enough that it has no new adopters. A few suppliers keep serving the diminishing demand, while looking for new opportunities.

Senility is when the technology is old enough that there is no more room for incremental innovation and that it is starting to lose its utility. The market is reduced to a niche one as most suppliers have now turned to other more promising technologies

Death is when the technology dies because it has reached full obsolescence. Suppliers, materials, and know-how start to disappear. The technology can no longer compete with newer substitutes, so it either disappears entirely or survives in a museum.

When we map the utility of the technology as it goes through all the above stages, we get Nolte's Whale-shaped chart (figure 3.1) illustrating the slow increase of a technology's utility as it is developed, and the steep decrease of its utility when it becomes obsolete.

To put this cycle into perspective, Nolte superimposes his chart with Moore's technology adoption cycle. Goeffry Moore (1991) describes how a technology is adopted from a market perspective. Moore distinguishes 4 groups of technology adopters:

First, the Innovators are the technology enthusiasts who buy the technology just because it is new.

Second, Early Adopters are the ones who recognize the potential of the technology, and they help debug the technology and make it more user-friendly.

Then, Pragmatists form the bulk of the technology adopters. They encompass those who use the technology because they've seen it already work, and those who are forced to use it in order to keep up with the competition.

Finally, traditionalists or laggards do not like high technology products, and will only adopt the technology at a late stage when no alternatives are left and when newer substitutes are starting to become available.

Moore (1991) stresses the importance of targeting each of those different segments in technology marketing, especially when "crossing the chasm" between the early adopters and the pragmatists.

Of course, this cycle is from a customer's adoption perspective, and a customer cannot adopt a technology unless it was developed enough to get to a usable level. This means that Moore's adoption cycle comes after the technology development phase, which we place in superposition to the technology lifecycle whale chart in figure 3.1.

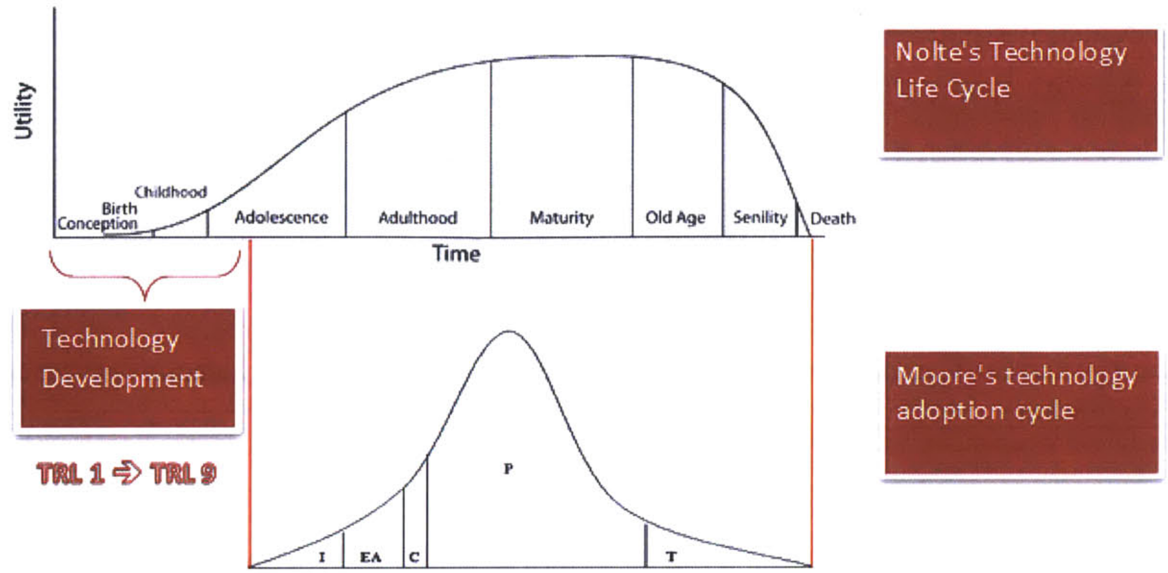


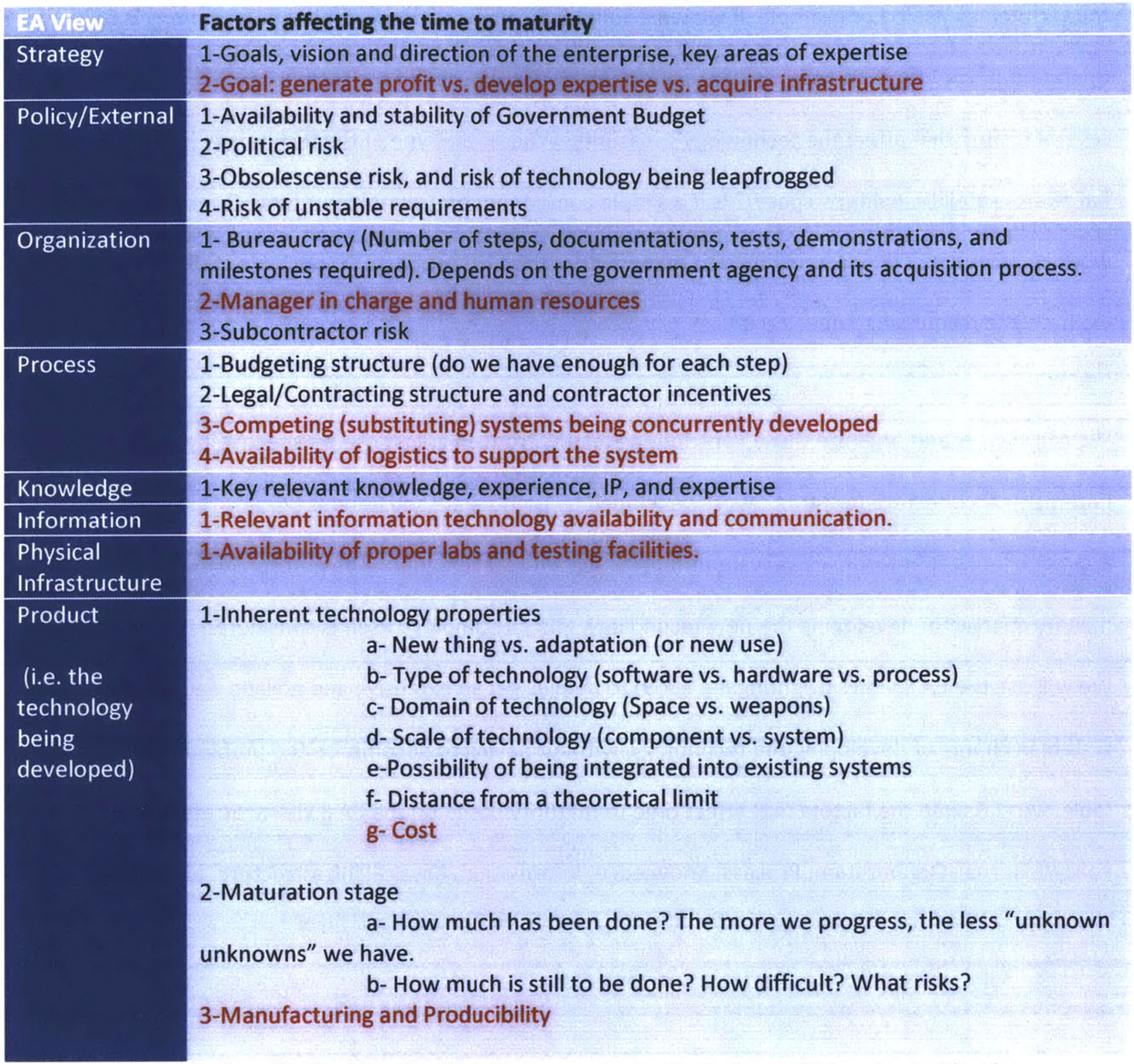
Figure 3:1 Nolte's technology life cycle superimposed with Moore's adoption cycle and the technology development phase

Now that we have located the technology development phase within the technology lifecycle, we will next define technology maturity and establish a framework for measuring it.

3.1.2 The multidimensional concept of maturity

In theory, technology maturity could be defined as "the quantity of information we know about the technology and how good/efficient we are at applying it". In practice however, product managers need a more operational definition of technology maturity. For them, a technology's maturity is measured on different practical dimensions such as "How long until I can market a functioning product? How much investment do I still need? What is the risk of not having the technology work as intended?" Viewed from this perspective, we can see that there is no straight answer to the question "how mature is a developing technology?" since maturity is defined on many levels. Furthermore, if we pick one of those definitions, measuring maturity operationally means estimating uncertain future outcomes, and those outcomes are causally related to numerous aspects of the technology development environment, and of the technology itself. For example, if we want to look at "time remaining until the technology is fully operational" (or "Time-To-Maturity") as a measure of the maturity of the technology, then there are several factors that affect the technology's maturity: What is the type of the technology (software, hardware, satellite, military, space)? Is it a simple component or a complex system? Do we have a working prototype? Do we know how to manufacture the final product? Do we have stable funding? Do we have the resources to develop it?

We can classify and organize those contributing factors from an enterprise perspective by using the Enterprise Architecture (EA) framework developed by Nightingale and Rhodes (2004). This holistic approach looks at enterprises as systems interacting with stakeholders. The enterprise in this case is a military contractor developing the new technology, and a technology itself is considered the product. We will use the EA 8 views (Nightingale, 2009) to help us get an extensive and holistic view of the system in charge of developing the technology, without having to decompose the problem. In the below table, we classified the factors that affect time to maturity based on the EA 8 views: Strategy, Policy/External, Organization, Process, Knowledge, Information, Physical Infrastructure, and Product.



EA View	Factors affecting the time to maturity
Strategy	1-Goals, vision and direction of the enterprise, key areas of expertise 2-Goal: generate profit vs. develop expertise vs. acquire infrastructure
Policy/External	1-Availability and stability of Government Budget 2-Political risk 3-Obsolescense risk, and risk of technology being leapfrogged 4-Risk of unstable requirements
Organization	1- Bureaucracy (Number of steps, documentations, tests, demonstrations, and milestones required). Depends on the government agency and its acquisition process. 2-Manager in charge and human resources 3-Subcontractor risk
Process	1-Budgeting structure (do we have enough for each step) 2-Legal/Contracting structure and contractor incentives 3-Competing (substituting) systems being concurrently developed 4-Availability of logistics to support the system
Knowledge	1-Key relevant knowledge, experience, IP, and expertise
Information	1-Relevant information technology availability and communication.
Physical Infrastructure	1-Availability of proper labs and testing facilities.
Product (i.e. the technology being developed)	1-Inherent technology properties a- New thing vs. adaptation (or new use) b- Type of technology (software vs. hardware vs. process) c- Domain of technology (Space vs. weapons) d- Scale of technology (component vs. system) e-Possibility of being integrated into existing systems f- Distance from a theoretical limit g- Cost 2-Maturation stage a- How much has been done? The more we progress, the less "unknown unknowns" we have. b- How much is still to be done? How difficult? What risks? 3-Manufacturing and Producibility

Table 3.1 Factors affecting Time-To-Maturity, grouped using the 8 views of Enterprise Architecting

While all those factors ultimately determine how long a technology will take to mature, a lot of them are beyond the control of the contractor. This is why we distinguished the controllable variables from the uncontrollable ones by coloring them in red. Although the black and red variables all affect how long a technology needs to mature, the contractor's risk management and decision making can only affect "Time-To-Maturity" via the red variables.

In summary, we saw that measures of maturity correspond to early stages in the technology lifecycle, during Nolte's phases of conception, birth, childhood, and adolescence, and before the start of the product adoption cycle. Furthermore, we saw that maturity is defined on multiple dimensions, and that it is causally affected by a multitude of factors.

3.2 Measures of technology maturity

Since technology maturity (in the broad sense defined above) is a complex notion that depends on many factors, it cannot be reduced to one dimension, and hence cannot be fully captured by only one measure. This explains why several measures were created to try to capture different aspects of the maturation progress. Although TRL is the most commonly used of those maturity scales, the aim of this section is to briefly present the TRL scale, and then compare it to the other scales so that it can be put in the context of the different measurements of technology maturity.

3.2.1 The NASA TRL scale

TRLs capture the maturity of the technology by looking at major milestones in the technology development process. At low TRLs, those milestones are concept development and basic laboratory tests. At high TRLs, they correspond to prototype validation and field testing. The TRL scale is a systematic high-level measure of maturity that keeps track of important stages in the development of a technology. Below is the formal definition of the scale as used by NASA (Mankins, 1995):

TRL	NASA TRL Definition
1	Basic principles observed and reported
2	Technology concept and/or application formulated
3	Analytical and experimental critical function and/or characteristic proof of concept
4	Component and/or breadboard validation in laboratory environment
5	Component and/or breadboard validation in relevant environment
6	System/subsystem model or prototype demonstration in a relevant environment (ground or space)
7	System prototype demonstration in a space environment
8	Actual system completed and "flight qualified" through test and demonstration (ground or space)
9	Actual system "flight proven" through successful mission operations

Table 3.2 NASA TRL scale definitions

We can see that the TRL scale is intended to be consistent across technologies (Mankins, 1995), and that it tries to encompass the major steps common to all technology development. As a result, it loses specificity and fails to capture all the technology-specific factors mentioned in table 3.1. This is why a plethora of alternative TRL scales and complementary maturity measures have emerged. Depending on how they are a substitute for or a complement to the NASA TRL scale, those other maturity scales can be classified into 2 main groups:

1. **Alternative scales** that substitute for TRL by adapting TRL to other organizations, domains, or technologies, making it more tailored or suited to the specific attributes of those domains.
2. **Supporting scales** that complement TRL either by (1) measuring other factors of technology maturity risk, or by (2) measuring the readiness of factors or processes that support product development, but that are not directly related to the technology itself. Those scales are intended to be used with TRL to get a complete picture of technology development risk.

Below is the detailed proposed classification of those scales, followed by a brief description of each:

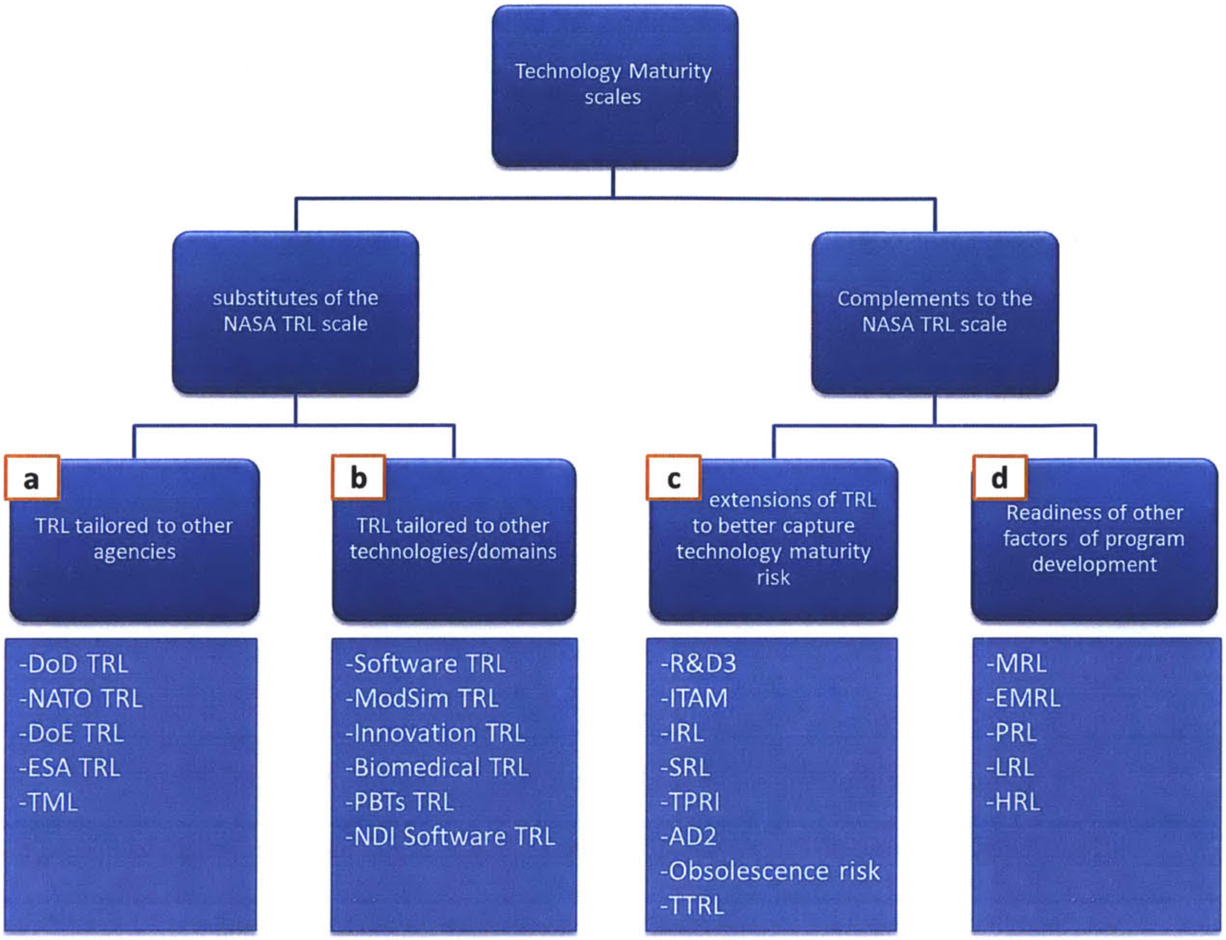

Technology Maturity scales
substitutes of the NASA TRL scale
Complements to the NASA TRL scale
a
TRL tailored to other agencies
b
TRL tailored to other technologies/domains
c
extensions of TRL to better capture technology maturity risk
d
Readiness of other factors of program development
-DoD TRL
-NATO TRL
-DoE TRL
-ESA TRL
-TML
-Software TRL
-ModSim TRL
-Innovation TRL
-Biomedical TRL
-PBTs TRL
-NDI Software TRL
-R&D3
-ITAM
-IRL
-SRL
-TPRI
-AD2
-Obsolescence risk
-TTRL
-MRL
-EMRL
-PRL
-LRL
-HRL

Figure 3:2 Classification of TRL-related technology maturity scales

3.2.2 TRL tailored to other agencies

- **DoD TRL**: After the 1999 GAO report recommending the use of TRLs, the Deputy Under Secretary of Defense for Science and Technology issued a memorandum recommending the use of TRLs in all new major programs. The TRL was later included in the Technology Readiness Assessment (TRA) Deskbook (DoD, 2009) with a 9-level definition very similar to that of NASA, but with more emphasis on the supporting information that needs to be produced at each level (For more information about the DoD Acuisition cycle and Technology Readiness Assessment, please refer to section 3.3.2).

- **NATO TRL:** Following the formal adoption of TRL by DoD, the Undersea Research Centre (NURC) at the North Atlantic Treaty Organization (NATO) developed its own TRL scale. Although this 10-level scale is based on the DoD TRL definitions, the NURC scale has more detailed descriptions in an attempt to make it more compatible with NATO needs (NATO, 2010).
- **DoE TRL**: The Department of Energy (DoE) has also adopted a tailored 9-level version of the NASA TRL scale for its own Technology Readiness Assessment process. For example, the level descriptions go more into detail on the use of simulants, the scale of pilot projects, and the transition from cold to hot commissioning. The DoE TRA guide (2009) also maps 3 dimensions of testing recommendations relative to TRL: Scale of Testing (Going from Lab, to Engineering/Pilot, to Full scale), Fidelity (Going from Paper, to Pieces, to Similar, to Identical), and Environment (going from Simulated, to Relevant, to Operational). However, before that in the late nineties, DoE's Nuclear Materials Stabilization Task Group (NMSTG) used another scale called "Technology Maturity" (TM). Unlike TRL, the TM score was a continuous scale with values from 0 to 10 since it was computed as a weighted average between seven different maturity scales (requirements, processes, equipment, facilities, schedule, personnel, and safety). Kenley and Creque (1999) later used regression analysis to identify Hardware equipment, Facility, Operational safety, Process Maturity as the four most relevant maturity parameters when predicting schedule.
- **ESA TRL:** is a TRL scale used by the European Space Agency. Since it is geared towards space technologies, it is very similar to the NASA TRL scale. The Technology Readiness Levels Handbook for space applications (TEC-SHS, 2008), contains relatively detailed definitions of each TRL level, with examples, key questions to ask for transitions, and appropriate evidence required. (TEC-SHS, 2008)

- **TML**: Technological Maturity Levels. It was developed for the Canadian Department of National Defense (DND). It is based on the NATO TRL definition combined with an Interface Maturity Level, a Design Maturity Level , a System Readiness Level (all defined by the UK MoD, 2006), and the Manufacturing Readiness Level (MRL) scale. Although the intention is to capture the multidimensional aspect of technology maturity, those 5 scales end up being merged into one TML number. (Hobson, 2006)

3.2.3 TRL tailored to other technologies/domains

- **Software TRL:** While the original TRL scale was intended for hardware, this 9-level scale was developed to track Software development. It was first developed by NASA in 1999 (Nolte, 2008), and it now has different versions for different agencies (DoD, NASA, Army, Missile Defense Agency, Air Force Research Laboratory). The levels go from scientific feasibility, to application of software engineering principles, to final implementation/integration steps.
- **ModSim TRL:** It is a modified 9-level version of the NASA TRL by mainly adding the Predictive Capability Maturity Model (PCMM) attributes. It is tailored for Modeling and Simulation applications. The methodology consists of first giving maturity scores on 9 dimensions (Capability Maturity, Verification, Validation, User Qualification, Code Readiness, Models, Geometry, Quantification of Margins and Uncertainties, and System), and then assigning the ModSim TRL score as the minimum of those scores. (Clay et al, 2007)
- **Innovation TRL:** The innovation readiness level is a 6-level scale intended to depict the development of innovation. This is why the 6 levels of IRL correspond to 6 phases of technology adoption. Each of those levels (phases) is measured along 5 dimensions: Technology, Market, Organization, Partnership, and Risk. (Tao, 2008)

- **Biomedical TRL:** was developed by the U.S. Army Medical Research and Materiel Command. It consists of four 9-level TRL scales developed for four different categories: Pharmaceutical (Drugs) TRL, Pharmaceutical (Biologics, Vaccines) TRL, Medical Devices TRL, and Medical IM/IT and Medical Informatics TRL. While it is similar in the overall structure to the NASA TRL, most descriptions of the TRL levels tend to have more specific category-related descriptions. The levels also have detailed descriptions of the required supporting information to pass a certain TRL (U.S. Army Medical Research and Materiel Command, 2003)
- **PBTs TRL:** Process-based Technologies TRL is a 9-level scale developed by Graettinger et al. (2003). It is intended for practices, processes, methods, approaches, (and frameworks for those), as opposed to technologies like hardware, software, embedded systems, or biomedical devices. PBTs mature on two levels: the environment (which becomes more representative as the community of users expands from initial risk takers to mainstream users), and the completeness of the technology (as it goes from basic properties, to a defined core, to implementation mechanisms, to best practices, to a body of knowledge).
- **NDI Software TRL**: Smith (2005) proposes to address the weaknesses of software TRLs by adopting a multi-criteria measure of maturity for Non-Developmental Item (NDI) Software. He simultaneously measures 5 relevant attributes on 5 different scales: Requirements Satisfaction, Environmental Fidelity, Criticality, Product Availability, and Product Maturity (Smith, 2005).

3.2.4 Extensions of TRL to better capture technology maturity risk

- **R&D^3**: The Research and Development Degree of Difficulty scale was proposed by Mankins (1998) to address TRL's weak predictive power. R&D^3 complements TRL as a measure of how much difficulty is expected to be encountered in the maturation of a technology. It consists of 5 levels simply corresponding to the predicted probability of success in maturing the technology.

Level I corresponds to 99% or almost certain success, while Level V corresponds to 10%-20% or almost certain failure (Mankins 1998).

- **ITAM:** After introducing the R&D^3 measure, Mankins (2002) proposed the Integrated Technology Analysis Methodology (ITAM) as a framework to measure both technology development progress as well as anticipated future research and technology uncertainties, while taking into account the importance of critical subsystem technologies. The methodology is based on calculating, then ranking the Integrated Technology Indexes (ITIs) of different competing candidate technology systems. The ITIs are calculated using 3 numbers: ΔTRL (Delta TRL, which is simply the difference between current and desired TRL), R&D^3 (Research & Development Degree of Difficulty, explained earlier), and Technology Need Value (TNV, measures how critical a technology is: TNV=3 if the technology is "enabling", 2 if it is "very significant", and 1 if it is just "enhancing"). The ITI for a candidate system is defined as:

$$ITI = \frac{\sum_{subsystem\ technologies}(\Delta TRL \times R\&D^3 \times TNV)}{Total\ \#\ of\ subsystem\ technoloogies}$$

The ITI is a measure of the cumulative maturation that must be achieved for each technology, amplified by the anticipated difficulty of its maturation, and the project importance of the technology within the system, all normalized by the total number of technologies in the system. The ITI number is intended to be later used in comparing different technology system solutions as an indicator of required R&D risk (the lower the ITI, the better). Although the ITI was a first step in developing a more integrated/complete predictive measure, Mankins does not explain the logic behind the way he aggregated the factors: why multiplication and not addition? Are the scales calibrated in the right way so that the multiplication of a 1-to-8 number by a 1-to-5 number by a 1-to-3 number makes sense? Or should those 3 factors somehow be weighted?

Another weakness is that computing ΔTRL makes no practical sense because TRL is an ordinal scale: the TRL numbers are simply placeholders and the values have no inherent meaning, hence there is no reason to believe that the difference TRL3-TRL1= 2 is equal to the maturity difference of TRL9-TRL7=2 (this ordinality problem is discussed in details in sections 3.3.3.2 and 9.4).

- **IRL:** The Integration Readiness Level scale is the first step towards extending TRL from component technologies to systems of technologies. An IRL measures how well two technologies are integrated, and is defined only with respect to those two technologies. It is a 9-level scale that goes from the identification of an interface between the two technologies, to creating a common interaction language, to enhancing quality, then the control over the interaction, to exchanging structured information for the communication, and finally to testing and validating the integration (Sauser et al, 2010).
- **SRL:** The System Readiness Level scale uses IRL to measure the readiness of a system. It is an improvement over the ITI in that it does consider the system structure of the technology (through all the pairs of IRLs between components). If $[TRL]_{n.1}$ is the vector of TRLs of technology components, and $[IRL]_{n.n}$ the symmetric matrix of Integration Levels of all pairs of component technologies. Then Sauser et al.(2008) defines the SRL vector as $[SRL]_{n.1} = [IRL]_{n.n} \cdot [TRL]_{n.1}$, and then defines the System Readiness Level as the normalized sum of the elements of $[SRL]_{n.1}$. The obtained SRL is a number between 0 and 1 that is correlated to phases in the DoD acquisition lifecycle: low values correspond to early phases in the acquisition cycle, while higher values correspond to production and operations and support. In other words, SRL is the (normalized) sum of products of each technology's TRL by the IRLs of all the technologies it interfaces with. One weakness is that it reduces the maturity of the whole system to only one number; another is the fact that SRL is just a linear sum such that it does not

take into account the criticality of some components in the WBS. Tan et al.(2009) extend SRL to the stochastic domain by allowing TRL and IRL to be variable (distributions generated by the different expert opinions for TRL and IRL measures) , then repeating the above calculation in a Monte-Carlo simulation to generate a distribution for SRL.

- **TPRI:** the Technology Performance Risk Index is an extension to R&D^3 with respect to a particular performance measure. If A is the percentage of achieved performance (relative to a predefined threshold), then the TPRI is defined as $TPRI = 1 - \frac{A}{1+(1-A)*R\&D^3}$ (the formula corresponds to a closed-loop feedback mechanism). The resulting TPRI is a number between 0 and 1 where low values correspond to low performance risk, and where high values correspond to high chances of failing to meet performance goals. Mahafza (2005) does not provide evidence in support of the closed-loop mechanism functional form.
- **AD^2:** The Advancement Degree of Difficulty is similar to Mankins' R&D^3 in that it also tries to determine the remaining effort/risk in maturing the technology. The main addition is that it provides a methodology and application framework for R&D^3 (whereas R&D^3 scale by itself is just a list of risk levels with high-level definitions, with no accompanying methodology on how to determine those risk levels). A team of experts determines the steps necessary for maturing the technology; they then estimate the expected degree of difficulty to advance to the R&D goals, and design tests that help determine those degrees of difficulty. They give R&D^3 scores in 5 specific areas: Design and Analysis, Manufacturing, Software Development, Test, and Operations (Bilbro, 2007). AD^2 is intended to be completely complementary to TRL by measuring everything we could know about the difficulty of advancement. Nolte (2008) says that AD^2 is the best predictive method for determining difficulty of advancement because of its comprehensiveness and available procedure. As a result, it would be very useful in developing cost and schedule risk

models since it actually is the best idea we have of what is left to be done to mature the technology.

- **Obsolescence/Leapfrogging risk:** Valerdi and Kohl (2004) consider the risk of a technology becoming obsolete or being leapfrogged by another one. They develop a 5-level scale of obsolescence risk and integrate it via a cost multiplier under technology risk in the Constructive Systems Engineering Cost Model (COSYSMO).
- **TTRL:** stands for Technology Transfer Readiness Level. This scale deals with the issue that the TRL scale measures maturity only with respect to a certain application. For example, many mature military technologies needed time and development before getting mature enough for civilian use. The TTRL was developed by the European Space Agency to evaluate and monitor the transfer of technologies from space application to commercial non-space applications (Holt, 2007).

3.2.5 Readiness of other factors of program development

- **MRL:** Manufacturing Readiness Level addresses the gap between developing the technology and manufacturing the product. It is the only scale other than TRL used in DoD TRA assessments. MRLs provide a common understanding of the relative maturity and risks associated with manufacturing technologies, products, and processes being considered to meet DoD requirements (Morgan, 2008). It is a 10-level scale that follows manufacturing from early stages of proof of concept of manufacturability, to refining the manufacturing strategy with demonstration of proper enabling/critical technologies, components, prototype materials, tooling and test equipment, and personnel skills, to the full-rate reliable production with all engineering, performance, quality, material, human, and cost requirements under control (OSD

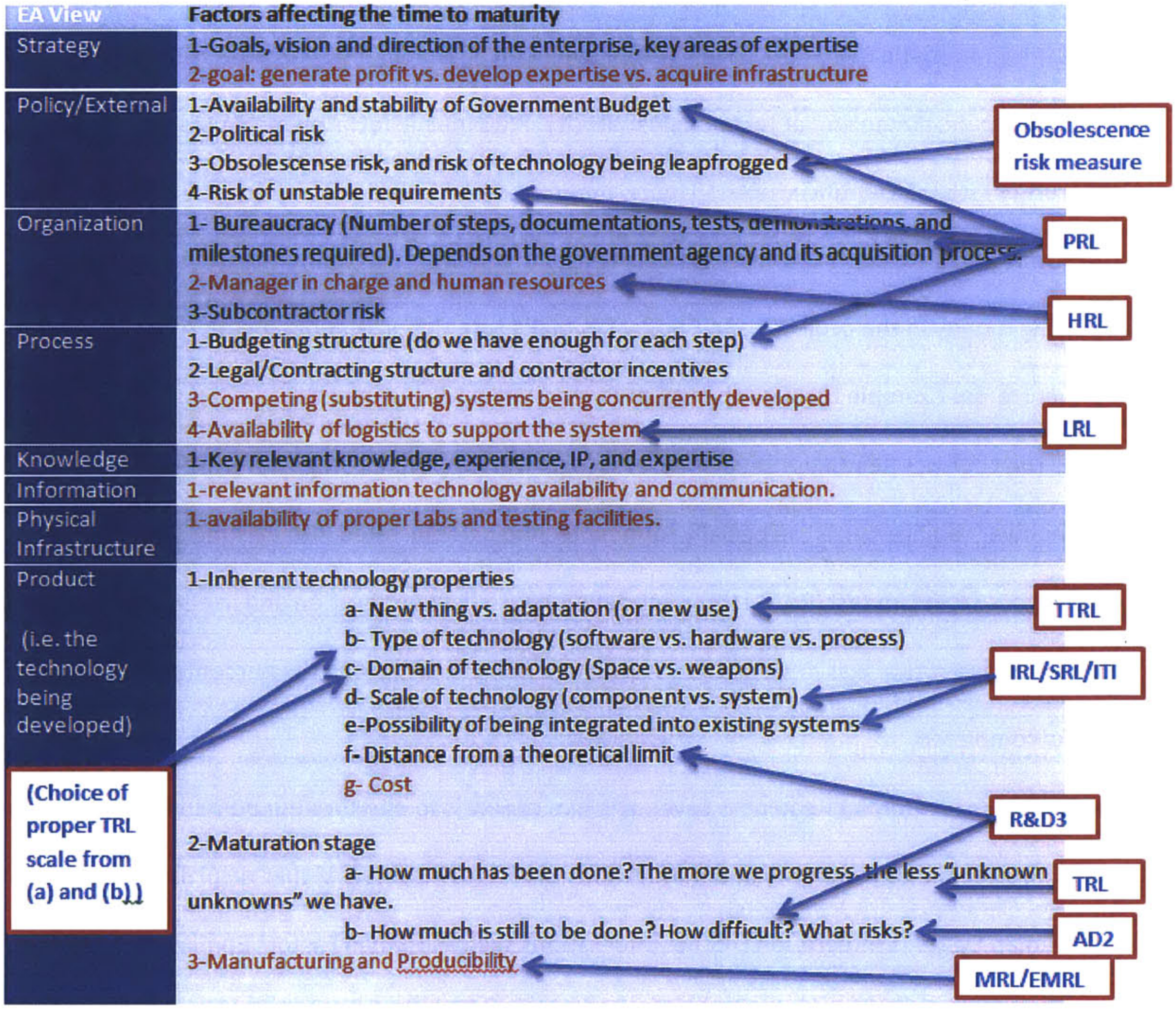
Manufacturing Technology Program, 2011). Figure 3.5 shows how MRLs correlate with TRLs within the DoD Acquisition lifecycle.

- **EMRL:** stands for Engineering and Manufacturing Readiness Level. The Missile Defense Agency uses this 5-level maturity scale to capture design and manufacturing knowledge for product development, demonstration, production, and deployment (Fiorino, 2003).
- **PRL:** Programmatic Readiness Levels were developed by Nolte (2008) to measure the programmatic dimension of technology maturity. PRLs mainly focus on measures of Documentation (scientific articles/papers, progress reports, technical reports), Customer Focus (to track the customer's changing expectations and requirements, until he settles finally by putting money in the project), and Budget (to make sure the project has an appropriate budget structure, for example through the use of Earned Value Management).
- **LRL:** The Logistics Readiness Level measures logistical support (e.g. logistics workload, manpower requirements), especially for new technologies inserted into existing systems. The scale goes from Lab test/R&D, to project definition, to fleet verification and use (Broadus, 2006). An excel-based tool was developed to calculate a numeric LRL giving the percentage of required tasks completed.
- **HRL:** stands for Human Readiness Level. The aim of HRL is to measure the human factors affecting technology maturity and maturation risk. It is a 9-level scale that records the evolution of Human Systems Integration (HSI) considerations such as Human Factors Engineering, System Safety, Health Hazards, Personnel Survivability, Manpower, Personnel, Training, and Habitability (Phillips, 2010).

Other scales within that category that are mentioned in the literature, but for which we did not find the definitions of the scales are Design Readiness Level, Operational Readiness Level, and Capability Readiness Level (Bilbro, 2008).

3.2.6 Relating the maturity scales to the multidimensional measurement of technology maturity

Now that we have laid out the different maturity measures, Table 3.3 shows how they relate to the factors influencing Time-To-Maturity that were defined earlier:



EA View	Factors affecting the time to maturity
Strategy	1-Goals, vision and direction of the enterprise, key areas of expertise 2-goal: generate profit vs. develop expertise vs. acquire infrastructure
Policy/External	1-Availability and stability of Government Budget 2-Political risk 3-Obsolescense risk, and risk of technology being leapfrogged 4-Risk of unstable requirements
Organization	1- Bureaucracy (Number of steps, documentations, tests, demonstrations, and milestones required). Depends on the government agency and its acquisition process. 2-Manager in charge and human resources 3-Subcontractor risk
Process	1-Budgeting structure (do we have enough for each step) 2-Legal/Contracting structure and contractor incentives 3-Competing (substituting) systems being concurrently developed 4-Availability of logistics to support the system
Knowledge	1-Key relevant knowledge, experience, IP, and expertise
Information	1-relevant information technology availability and communication.
Physical Infrastructure	1-availability of proper Labs and testing facilities.
Product (i.e. the technology being developed)	1-Inherent technology properties a- New thing vs. adaptation (or new use) b- Type of technology (software vs. hardware vs. process) c- Domain of technology (Space vs. weapons) d- Scale of technology (component vs. system) e-Possibility of being integrated into existing systems f- Distance from a theoretical limit g- Cost 2-Maturation stage a- How much has been done? The more we progress, the less "unknown unknowns" we have. b- How much is still to be done? How difficult? What risks? 3-Manufacturing and Producibility

Obsolescence risk measure
PRL
HRL
LRL
TTRL
IRL/SRL/ITI
(Choice of proper TRL scale from (a) and (b))
R&D3
TRL
AD2
MRL/EMRL

Table 3.3 Correspondence between maturity measures and factors influencing "time to maturity"

Those multiple scales allow for a more tailored measurement for every case, which means a more complete picture of technology maturity. Hence, for a good maturity assessment, one should start by determining the appropriate scale from classes (a) and (b) in figure 3.2, and then complement it by as many scales from classes (c) and (d) in figure 3.2 as needed. For example, in a software development program, the program manager should use software TRL, then add IRL and SRL if there is complexity in the project, then add obsolescence risk if the software faces obsolescence, then add PRL to better track documentation/customer/budget, and add AD^2 and RD^3 for more accurate assessment of risk due to progression uncertainty. In the same spirit, GAO (2006) recommended that DoD expand its use of metrics to cover aspects of technology maturity not explicitly attended to by the TRL.

Scales other than TRL have no historical data as they were developed only recently or were not adopted. Hence, we limit our maturity measurements to TRL only. After all, TRL does capture a major part of maturity, and a measurement does not have to be fully exact. Because of this data limitation, we adopt a definition of "measurement" as used in information theory: a measurement simply means a reduction of uncertainty in the quantity of interest (Hubbard, 2007). In this case, TRL data constitutes a measurement of maturity as long as it significantly reduces the uncertainty of maturity variables (i.e. cost and schedule).

In summary, maturity is a complex multidimensional concept. This led to the development of many maturity scales other than the NASA TRL. While a lot of those scales are simple adaptations of the NASA TRL to other fields/technologies, others were intended as an extension to TRL to provide a more complete measure of technology risk. For now however, we can only do TRL-based maturity measurements because it is the only maturity scale with historical data. Although TRL alone is not a complete measure of maturity, it is still possible to build TRL-only-based models because TRL captures essential information in technology maturation.

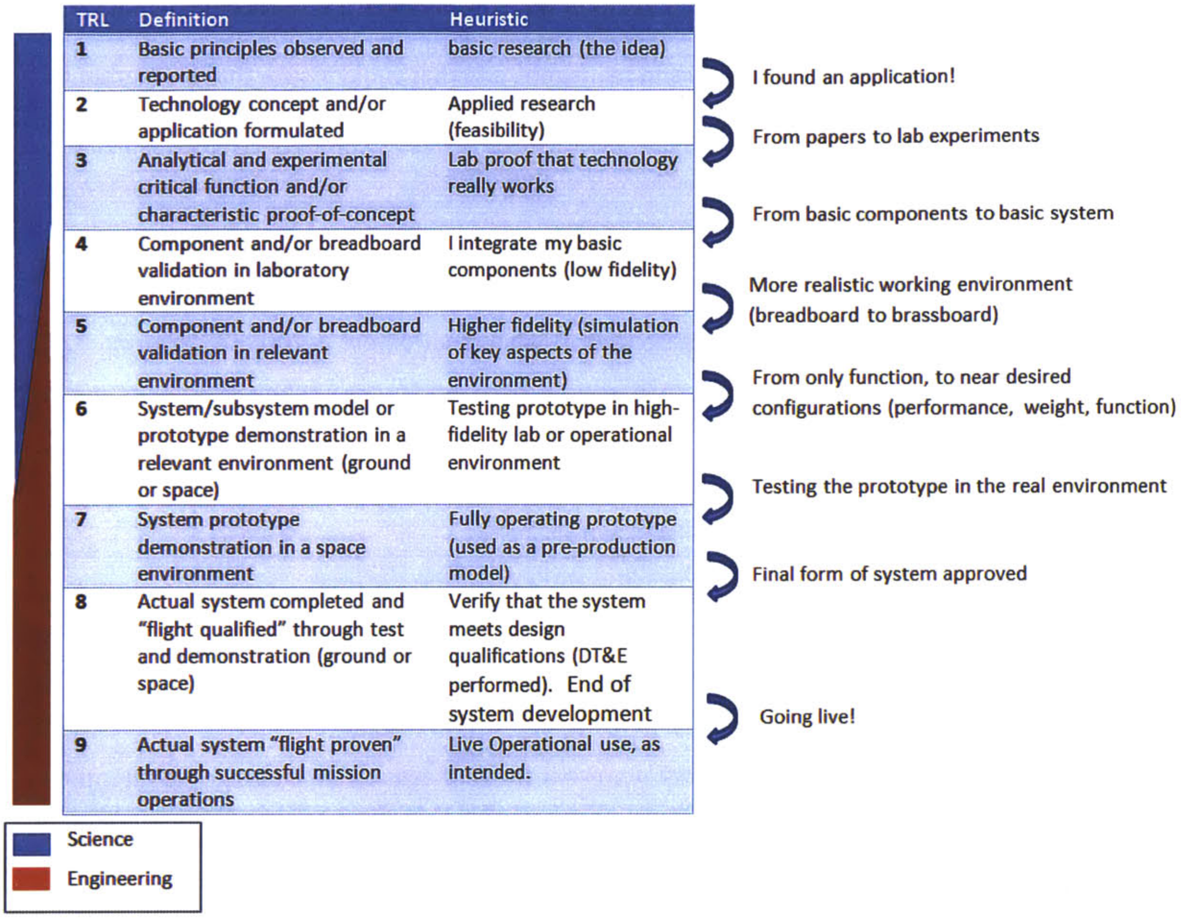
3.3 The TRL scale

Now that we have defined the technology development and measurement context around TRL, this section will focus on the TRL scale itself prior to presenting TRL-based models in later sections. This section consists of 3 subsections. First, we will go further into the definition of the TRL scale, then we will explain how TRL is used in the DoD Technology Readiness Assessment process, and finally we will look in more details at the properties and characteristics of the TRL scale.

3.3.1 Definition

Technology Readiness Level (TRL) is a 1-to-9 discrete scale developed by NASA to systematically assess the maturity of technologies (materials, components, devices, etc.) that are being developed for incorporation into a system. A low TRL (1-2) indicates a technology that is still at a basic research level, while a high TRL (8-9) indicates a technology that has been incorporated into a system that is in use. The first three levels can be considered as pure scientific research, while the last three ones are pure engineering development (Nolte, 2008). TRLs 4 to 6 form the transition phase from science to applied technology.

To make the scale more intuitive and understandable, we developed a parallel "heuristic" scale that helps understand the definitions of the TRL levels in simple terms. We also indicate the transitions between those levels, to outline in simple terms the event that triggers the transition from one TRL to another.



TRL	Definition	Heuristic
1	Basic principles observed and reported	basic research (the idea)
2	Technology concept and/or application formulated	Applied research (feasibility)
3	Analytical and experimental critical function and/or characteristic proof-of-concept	Lab proof that technology really works
4	Component and/or breadboard validation in laboratory environment	I integrate my basic components (low fidelity)
5	Component and/or breadboard validation in relevant environment	Higher fidelity (simulation of key aspects of the environment)
6	System/subsystem model or prototype demonstration in a relevant environment (ground or space)	Testing prototype in high-fidelity lab or operational environment
7	System prototype demonstration in a space environment	Fully operating prototype (used as a pre-production model)
8	Actual system completed and "flight qualified" through test and demonstration (ground or space)	Verify that the system meets design qualifications (DT&E performed). End of system development
9	Actual system "flight proven" through successful mission operations	Live Operational use, as intended.

I found an application!
From papers to lab experiments
From basic components to basic system
More realistic working environment (breadboard to brassboard)
From only function, to near desired configurations (performance, weight, function)
Testing the prototype in the real environment
Final form of system approved
Going live!
🟦 Science
🟥 Engineering

Figure 3:3 The TRL scale with simplified definitions of levels and transitions

3.3.2 TRL within DoD

This section will briefly explain how TRL is integrated in the DoD acquisition lifecycle.

The TRL scale in its current form was first outlined in 1995 paper by NASA scientist John C. Mankins who described TRL as a "systematic metric/measurement system that supports assessments of the maturity of a particular technology and the consistent comparison of maturity between different types of technology." However the concept had been used by NASA since the 1970's (Nolte, 2008). A 1989 NASA paper by Stanley Sadin outlines a seven-level readiness scale (which matches with the first 7 levels of the later 9-level scale) without using the "Technology Readiness Level" terminology. The papers by

Mankins and Sadin showed the advantages of a technology maturity measure that could eliminate ambiguities between researchers, engineers, and managers, and that could assure that a program is mature enough before entering production.

It was the influential 1999 GAO report that institutionalized the use of TRLs, as it led to a spread of TRL use from a few Air Force programs in the 1990s to a DoD-wide requirement in 2001 to perform Technology Readiness Assessments (TRAs) at both milestones B and C (Defense Acquisition Guidebook, 2010). Later in 2003, DoD detailed the use of TRLs in its Technology Readiness Assessment Deskbook (2009).

3.3.2.1 The DoD Acquisition Cycle

Before explaining the TRA process, we will quickly present the DoD acquisition lifecycle, which is the process that all programs acquired by DoD go through. The cycle is a stage gate process formed of 5 phases: Material Solution Analysis (refine the technology concept and prepare a technology development strategy), Technology Development (reduce risk by selecting best technology alternatives and maturing them), Engineering and Manufacturing Development (system integration and system demonstration), Production and Deployment (low-rate initial production, then full-rate production, then deployment), and Operations and Support (sustainment and disposal). The first 4 phases are separated by 3 Milestones (A, B, and C) where the Milestone Decision Authority (MDA) approves the entry of an acquisition program into the next phase (DAG, 2010). The Pre-Systems acquisition period (pre-Milestone B) is usually handled by Science and Technology (S&T) organizations, while post-Milestone B phases are handled by DoD Acquisition programs.

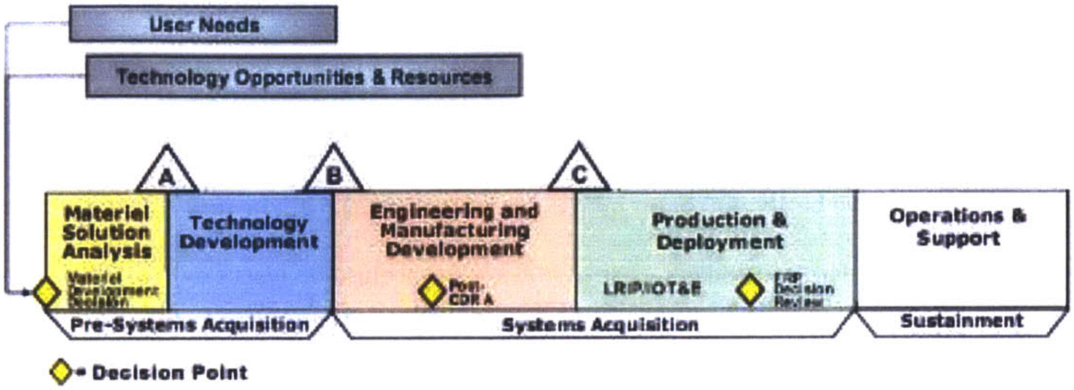


Figure 3:4 The DoD acquisition cycle.

3.3.2.2 The DoD Technology Readiness Assessment

Technology Readiness Assessments are required steps that prove a technology is mature enough before passing acquisition milestones B and C (TRA Deskbook, 2009). The Defense Acquisition Guidebook (2010) describes TRA as a "systematic, metrics-based process that assesses the maturity of Critical Technology Elements (CTEs), including sustainment drivers". To avoid having to assess the maturity of all components in the WBS, the TRA focuses only on the important technology elements of the system (i.e. the CTEs). An element that is new or being used in a new way is considered "critical" if it is necessary to achieve the successful development of a system, its acquisition, or its operational utility.

The TRA is conducted by an Independent Review Team (IRT) of experts who determine the system's CTEs, and then measure their maturity. All CTEs need to be at least at TRL 6 prior to Milestone B, and at least at TRL 7 prior to Milestone C (DAG, 2010). When the Critical Technology Elements do not all meet the TRL requirement, the program can be delayed, its requirements can change, or it can be restructured to use only the mature elements (Dion-Schwarz, 2008).

While TRL is used to keep track of technical maturity, DoD encourages the use of a parallel Manufacturing Readiness Level (MRL) scale that complements TRL by giving an indicator of the progress and the risks associated with the technology's manufacturing and production processes. A 2010 GAO report titled "DoD can achieve better outcomes by standardizing the way manufacturing risks are managed" pushed for a standardized and early use of MRLs in the acquisition lifecycle (GAO, 2010). The below graph ties all those cycles together by superposing TRL levels with the MRL scale, Nolte's technology lifecycle, GAO transition risk, and major DoD acquisition cycle phases.

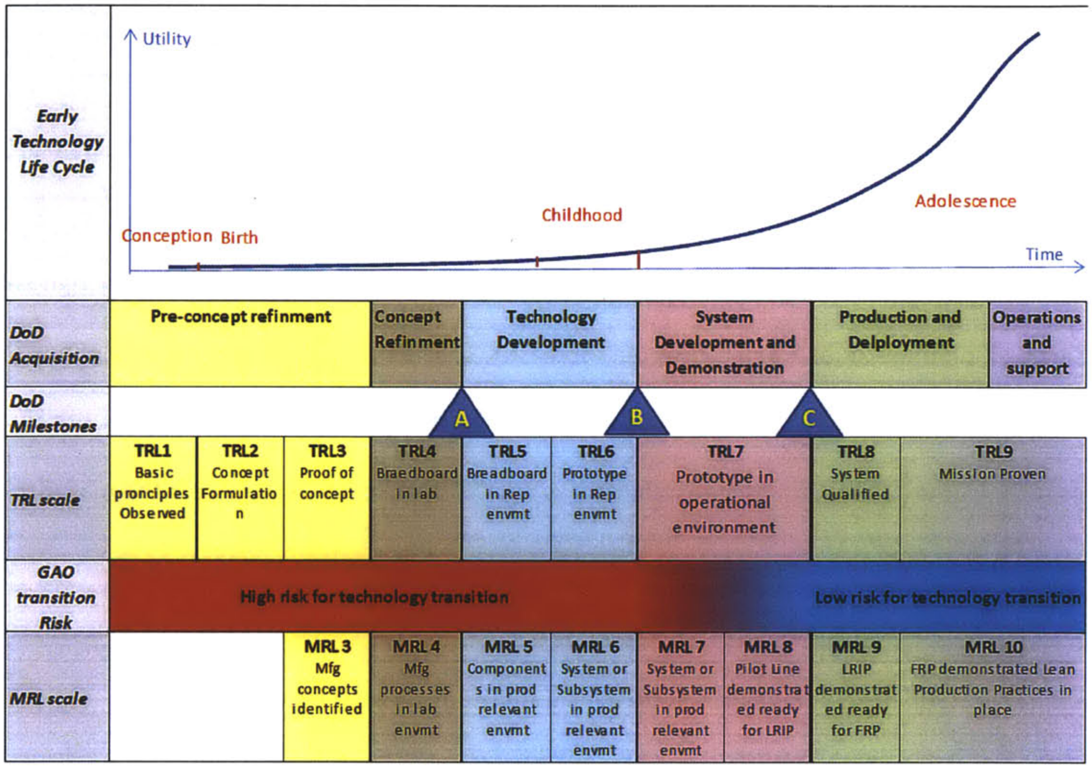

Early Technology Life Cycle
Utility
Conception
Birth
Childhood
Adolescence
Time
DoD Acquisition
Pre-concept refinment
Concept Refinment
Technology Development
System Development and Demonstration
Production and Delployment
Operations and support
DoD Milestones
A
B
C
TRL scale
TRL1 Basic pronciples Observed
TRL2 Concept Formulatio n
TRL3 Proof of concept
TRL4 Braedboard in lab
TRL5 Breadboard in Rep envmt
TRL6 Prototype in Rep envmt
TRL7 Prototype in operational environment
TRL8 System Qualified
TRL9 Mission Proven
GAO transition Risk
High risk for technology transition
Low risk for technology transition
MRL scale
MRL 3 Mfg concepts identified
MRL 4 Mfg processes in lab envmt
MRL 5 Component s in prod relevant envmt
MRL 6 System or Subsystem in prod relevant envmt
MRL 7 System or Subsystem in prod relevant envmt
MRL 8 Pilot Line demonstrat ed ready for LRIP
MRL 9 LRIP demonstrat ed ready for FRP
MRL 10 FRP demonstrated Lean Production Practices in place

Figure 3:5 Correspondence between Nolte's technology lifecycle, the GAO transition risk, the DoD Acquisition cycle, the TRL scale, and the MRL scale. Sources: adapted from DoD TRA Deskbook (2009), Nolte (2008), Sharif et al. (2012) , Azizian (2009), Morgan(2008), Graben (2009), Dion-Schwarz (2008), and GAO (1999)

This whole technology development phase can be put back in the larger contexts of technology lifecycle and technology adoption lifecycle defined earlier in section 3.1.

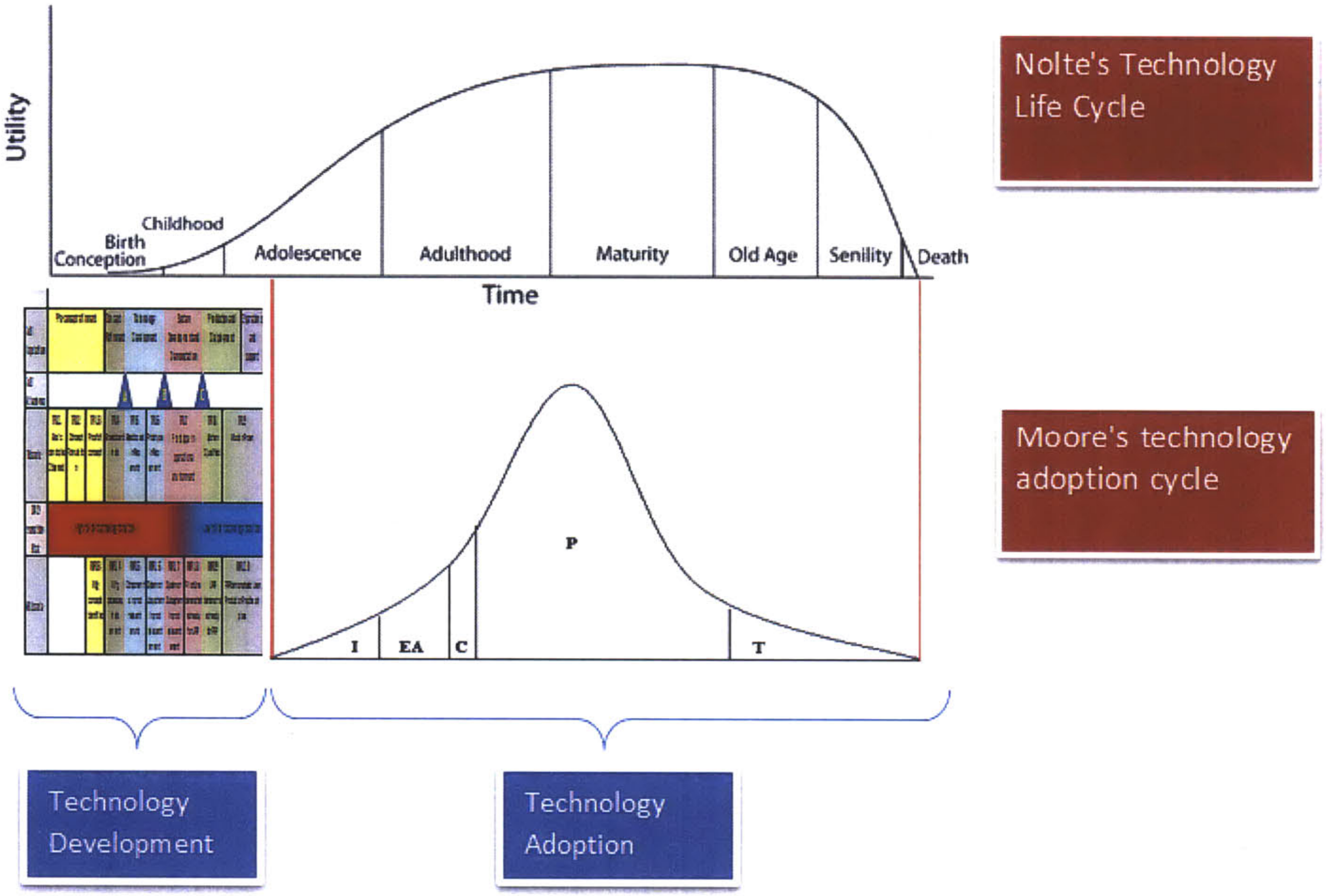


Figure 3:6 Localization of the correspondence table within the larger technology cycles.

3.3.2.3 The use of TRL in the TRA

It is important to note that the TRA makes a minimal use of the TRL measurements. The TRA is only a one-time measurement performed to give a Yes/No answer on whether the technology is ready to proceed in the acquisition cycle. The TRL measurement is not used for any risk management or cost/schedule modeling purposes.

However there is currently a tendency to extend the number/frequency of TRL measurements and their uses. For example, although TRL measurements are only required for milestones B and C, DoD's Best Practices encourages "early evaluation of project's maturity before Milestone A" (Guidance and Best Practices for Assessing Technology Maturity, TRA deskbook, 2009). The document also notes that "CTE identification should be a continuing element of every program. An initial determination of potential CTEs should be completed during Milestone A" (Guidance and Best Practices for Identifying CTEs, TRA Deskbook, 2009). Furthermore, the fact that DoD TRL definitions come with a list of "supporting information" required for each transition (for both hardware and software) indicates a move towards continuous monitoring of TRL scores. Finally, Dion-Scwarz (2008) lists several uses of TRL as follows:

-Provides a common understanding of technology status (maturity)

-Conveys what has been accomplished (demonstrated) with the technology

–Used as a factor in technical risk management

–Used to make decisions concerning technology funding

–Used to make decisions concerning transition of technology

–Used to scope acquisition programs and their requirements

–Used as a basis for certification under statute

In this thesis we go beyond those uses. We are mainly interested in models that use TRL for risk management. Instead of limiting the models to only measuring technical risk, we will use them for cost/schedule modeling in order to integrate the key management variables – maturity, cost, schedule - into a single technology management framework.

3.3.3 Properties/characteristics of the TRL scale

Now that we have a clear definition of TRL in the technology context, the measurement context, and the DoD Aqcuisition context, we will look at properties of the scale itself. Such properties are important because the models we will present in the framework will be built on those properties of the scale. We will divide those properties into strengths and weaknesses:

3.3.3.1 Strengths of the TRL scale

The TRL scale is simple (has only 9 points, it is clear and can be easily understood), well-structured (covers all relevant phases of technology maturation, and the defined stages do not overlap), stable (the definition of the scale does not change in time or with a new technology), adaptable (can be adapted to different domains as explained in section 3.2), and it can be used systematically and across technologies. The TRL calculator Excel tools (Nolte, 2005) are a good example of how the above properties are used to make TRL measurement standardized and streamlined.

On a more theoretical level, the "stable" and "well-structured" properties can be translated into "complete" and "monotonic" meaning that every technology has to go through all 9 levels, and in order. This property will be important for the level 1 assumption of our framework (section 6.2).

3.3.3.2 Weaknesses of the TRL scale

Most of the TRL's weaknesses appear when managers try to use TRL beyond its context or original intended use. They mostly shed light on what TRL does NOT measure (which we have already made clear in table 3.3). Below is a list of those weaknesses:

- The scale's definitions leave room for confusion (the levels are not specific and accurate enough). Cornford and Sarsfield (2004) and Minning et al. (2003) point out that some terms in TRL definitions such as "relevant environment" can be open for interpretation. Furthermore, although a TRL measurement should be consistent across different measurers, it is still a subjective measure. Not only are the measurements based on personal judgment and belief (Garvey, 2000), but they can also carry personal biases and interests. In a series of Army technology readiness assessment reports, we found that Independent Review Teams (IRTs) often give a lower TRL score than project managers in the assessments (a result of a conflict of interest, because the project manager needs high TRL values to pass the milestones). However, we found that different IRTs can also disagree among each other (not a result of a conflict of interest in this case).
- The TRL scale is an ordinal scale: while we know that TRL 3 is more mature than TRL 1, we don't know how much more mature it is. Is the difference between TRLs 1 and 3, the same as the difference between TRLs 5 and 8? Is TRL 6 twice as mature as TRL 3? The answer is No, because the values in an ordinal scale are mere placeholders to indicate an order, and have no numerical meaning (they could be easily replaced by letters A, B, C, etc., without losing information). The ordinality of the scale is especially problematic when mathematical operations are performed on TRL scores in risk management applications. In those cases, extracting risk factors out of an

ordinal scale can lead to large errors (Conrow, 2003). To address this issue, Conrow (2009) proposed a calibration of the TRL scale using an Analytic Hierarchy Process (AHP).

- TRL combines many dimensions of technology readiness into one metric. We earlier saw that there are many dimensions to technology maturity. TRL reduces those to one integer number, which means that information is being lost in the measurement process. Smith (2005) expresses this concern of "blurring contributions to readiness" for software TRLs. He also quotes Nolte on that issue "The TRL scale measures maturity along a single axis, the axis of technology capability demonstration. A full measure of technology maturity, or in the commercial-world product maturity, would be a multi-dimensional metric. It's not uncommon to find references to 12 or more dimensions of product or technology maturity. One writer speaks of 16 different dimensions of maturity. The TRL measures only one of the 16."
- TRL Lacks granularity: it has only 9 steps to describe the whole technology development process. This can be considered too few for using the scale in some applications.
- TRL does not account for the use of the technology. A system can be fully mature for one use, but may still need development and maturation for another application (refer to TTRL in section 3.2.4 for more on this issue).
- As a result of the TRL scale's monotonicity, it does not take technology obsolescence into account. Although TRL is mainly intended to be used for the technology development phase, some technologies do have rapid aging; and obsolescence becomes a concern. Should a technology remain at TRL 9 when it starts to lose utility? Nolte (2008) says that TRLs become useless once the technology matures. For DoD, obsolescence corresponds to the period of diminishing manufacturing sources when spare parts become no-longer available, while for software, Smith (2005) describes the many ways how a software ages when it undergoes

maintenance. Valerdi and Kohl (2004) address the problems of leapfrogging risk and technology obsolescence by integrating measures of those in their technology risk factor.

- TRL is unsuitable to measure the maturity of a system: Suppose a technology necessitates two components A and B. If A is at TRL 8 and B is at TRL2, what would the total Readiness level be? Now imagine that both components A and B are at TRL level 8, but we're still far from integrating them together and making a final product, what would the total Readiness level be in this case? Finally, what if component A is very critical for the final product, while there is a substitute for component B? We see that the definition of TRL does not allow taking complex systems and integration issues into account (IRL, SRL, and ITI in section 3.2.4 were developed to address this issue).

Finally, there are two weaknesses that show up often in the literature. While they are not weaknesses in the scale itself, they emerge when the scale is used as a measure of technology risk.

- The first weakness is that TRL is a bad measure for risk since it measures achieved developments only, without measuring the likelihood of a successful transition in the future. For instance, one technology can be at TRL 6 and still have a costly and long development ahead of it, while another technology can be at TRL 2 and have easy low-risk development ahead of it. Azizian et al. (2009), Cornford and Sarsfield (2004), Nolte (2008), Mahafza (2005), Sharif et al. (2012), and Fernandez (2010) all point out this problem in the TRL scale. In response to this issue, Mankins (1998) and Bilbro (2007) respectively developed the AD^2 and the RD^3 measures. Those measures (already defined in section 3.2.4), aim to quantify the difficulty and risk of what is left to be done, as opposed to what has already been done.

- The second problem is that risk managers would like to use TRL as a measure of risk. However TRL does not fully represent risk since it is only related to the probability of occurrence term while it misses the other factor in risk, which is the consequence term. Hence, caution is required as to how TRL should be integrated in risk analysis. However, even if we make sure to use TRL to measure only the probability of occurrence, it is still a weak measure (especially when compared with other measures like RD^3 and AD^2, for the reasons explained above). This is why Conrow (2003) recommends never using TRL alone in risk assessments.

As a result of the above weaknesses, Cornford and Sarsfield (2004) claim that TRL cannot be integrated in cost/risk models because uncertainty in the inputs of such a low resolution scale could lead to major errors in cost estimations. Although those issues weaken the results generated from TRL-based models, we believe that useful measurements can still be generated, especially if we carefully identify and evaluate the assumptions we use in each model. The study of those assumptions will be the basis of our TRL models framework in chapters 5-9.

Chapter 4

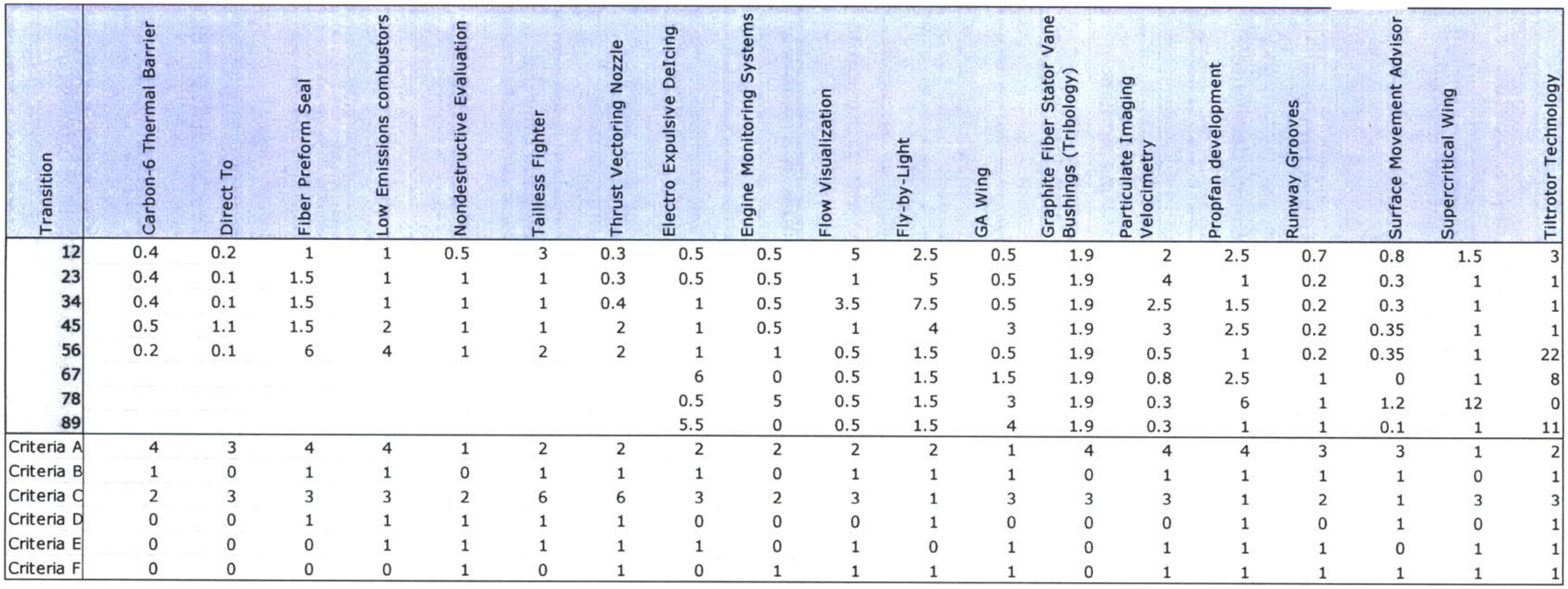
Chapter 4. The Datasets

Since this thesis adopts a data-based approach in supporting the assumptions and in testing the models, it is important to reserve a chapter describing and discussing the datasets that will be used. After all, the quality of the models and of the results strongly depends on the quality of the data they are based on. We used two datasets: the NASA dataset is a relatively high quality dataset, but it has only 19 data points; and the Army dataset is larger (582 data points), but with lower quality.

4.1 The NASA Dataset

4.1.1 Presenting the dataset

The first dataset is from a case study done by the Systems Analysis Branch at NASA's Langley research center looking at typical times that aeronautical technologies take to mature (Peisen and Schulz, 1999). The data was collected through interviews with NASA personnel. The following table shows the full data set:



Transition	Carbon-6 Thermal Barrier	Direct To	Fiber Preform Seal	Low Emissions combustors	Nondestructive Evaluation	Tailless Fighter	Thrust Vectoring Nozzle	Electro Expulsive DeIcing	Engine Monitoring Systems	Flow Visualization	Fly-by-Light	GA Wing	Graphite Fiber Stator Vane Bushings (Tribology)	Particulate Imaging Velocimetry	Propfan development	Runway Grooves	Surface Movement Advisor	Supercritical Wing	Tiltrotor Technology
12	0.4	0.2	1	1	0.5	3	0.3	0.5	0.5	5	2.5	0.5	1.9	2	2.5	0.7	0.8	1.5	3
23	0.4	0.1	1.5	1	1	1	0.3	0.5	0.5	1	5	0.5	1.9	4	1	0.2	0.3	1	1
34	0.4	0.1	1.5	1	1	1	0.4	1	0.5	3.5	7.5	0.5	1.9	2.5	1.5	0.2	0.3	1	1
45	0.5	1.1	1.5	2	1	1	2	1	0.5	1	4	3	1.9	3	2.5	0.2	0.35	1	1
56	0.2	0.1	6	4	1	2	2	1	1	0.5	1.5	0.5	1.9	0.5	1	0.2	0.35	1	22
67								6	0	0.5	1.5	1.5	1.9	0.8	2.5	1	0	1	8
78								0.5	5	0.5	1.5	3	1.9	0.3	6	1	1.2	12	0
89								5.5	0	0.5	1.5	4	1.9	0.3	1	1	0.1	1	11
Criteria A	4	3	4	4	1	2	2	2	2	2	2	1	4	4	4	3	3	1	2
Criteria B	1	0	1	1	0	1	1	1	0	1	1	1	0	1	1	1	1	0	1
Criteria C	2	3	3	3	2	6	6	3	2	3	1	3	3	3	1	2	1	3	3
Criteria D	0	0	1	1	1	1	1	0	0	0	1	0	0	0	1	0	1	0	1
Criteria E	0	0	0	1	1	1	1	1	0	1	0	1	0	1	1	1	0	1	1
Criteria F	0	0	0	0	1	0	1	0	1	1	1	1	0	1	1	1	1	1	1

Criteria A, Type of Technology: 1 = Airframe, 2 = Flight Systems, 3 = Ground Systems, 4 = Propulsion
Criteria B, Enabling Technology Needed: 0 = No, 1 = Yes
Criteria C, Primary Goal/Benefit: 1 = Cost Reduction, 2 = Safety, 3 = Performance
Criteria D, Focused Program: 0 = No, 1 = Yes
Criteria E, New Product Needed: 0 = No, 1 = Yes
Criteria F, NASA Tesgint: 0 = No, 1 = Yes

Table 4.1 The NASA dataset

The columns constitute different NASA technologies that were selected for the study. The first eight rows correspond to the TRL transitions (the values of the variables are the durations, in years, of the technology TRL transition times). The last 6 rows are additional criteria characterizing the technology being developed; they are explained just after the table in Peisen and Schulz (1999).

Except for the regression performed for demonstration purposes only in section 7.3.3, the Criteria A,B,C,D,E, and F will not be used in the analysis. First, the report does not explain in enough detail the meaning of those variables; second, the dataset is already too small to be segmented along those criteria; and finally, we have the intention of generalizing our results to other agencies using similar technology measures, and those agencies probably do not use the same criteria.

4.1.2 Sample quality

We consider this dataset our high quality sample since it had a technical report written just to present it, it identifies the technologies used for the study, and it has (almost) all the transition times at a relatively high resolution. However, the dataset does suffer some drawbacks:

1. The major problem for this thesis is the rather small number of technologies (only 19 data points available). Furthermore, seven of those data points are incomplete (they have only the first 5 transitions). As a result special precaution was taken throughout this thesis while handling the dataset.
2. There is a lack of information about the collection of the data points or any additional description of a technology's specific case (this is especially important for outlier analysis).
3. The sample suffers from selection bias since it only contains the technologies that made it to TRL 9 or 6. Many projects usually are abandoned at early TRLs. The negative effect of this phenomenon is that it reduces the representativeness of our work so that it cannot forecast

program termination risk, and is limited to programs that make it to TRL 9 or 6. Its positive effect however, is that it eliminates one source of uncertainty, so that we will not have to worry about modeling highly unpredictable external variables (such as budget cuts, program termination, requirement changes, etc.) that can lead a program to stop.

4. The data comes from retrospective interviews based on descriptions of TRLs (as opposed to rigorous TRL calculators), which means it could contain measurement errors. Furthermore, the authors say that they had to do some smoothing when the interviewees did not remember the exact transition time; and that "some" points were at a 1-year resolution level but did not indicate which points were at this resolution.
5. Finally, although the authors did not directly address the issue of sample representativeness, they did mention that the list of technologies was formed "with input from NASA Headquarters and three NASA Research Centers" and that the data covered all major types of technologies. Hence, we can assume than the sample is representative of the NASA technologies population.

4.1.3 Sample transformation

We started by performing a minor data transformation: there were three transitions that had a duration of zero. We performed a simple smoothing by assuming that this transition happened so quickly because of extra development effort in the precedent transition. So we changed the precedent step to 80% of its original value, and gave the remaining 20% to the "zero" step. We had two justifications for this modification. A theoretical one: a phase cannot be finished instantaneously; and practical one: it is necessary to eliminate zero for the log-transformation later performed on the data.

Below is a visualization of the data showing separately the evolution of each of the technologies:

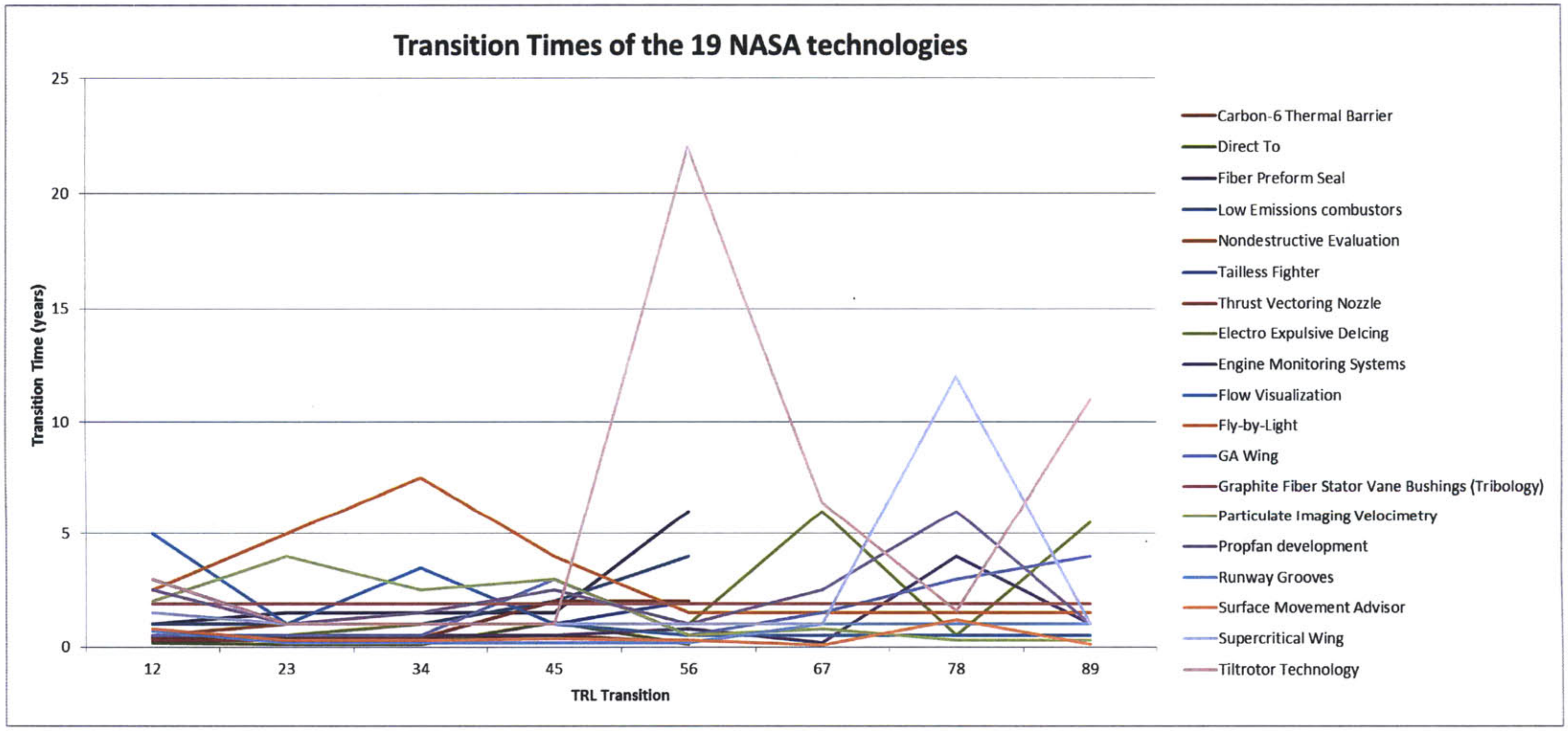


Figure 4:1 Transition times of the 19 NASA technologies

In the above graph, each colored line represents one technology, and the graph shows the transition time (in years) at each of the eight TRL transitions. One outlier is clearly visible: Titltrotor Technology (which is the technology used in the V22 osprey) took 22 years to transition from TRL 5 to TRL 6, which is significantly larger than the other transitions in the dataset.

Another visible property is the positive skewness of the data: while most values are very small, a few transitions take relatively larger times, thus clustering the bulk of the data around zero. The log-transformation was natural in such cases of positive data with clear positive skewness. It reduces the larger values and separates the smaller ones, making the data more normal. It also increases the randomness of residuals in regressions performed with the data set, and increases the correlation between the different transition times (both of which are beneficial effects).

After performing the two above corrections (removing the outlier technology, and taking the log), we get the new graph for technology log-transition times shown in figure 4.2. This transformed dataset will be used in the rest of this thesis.

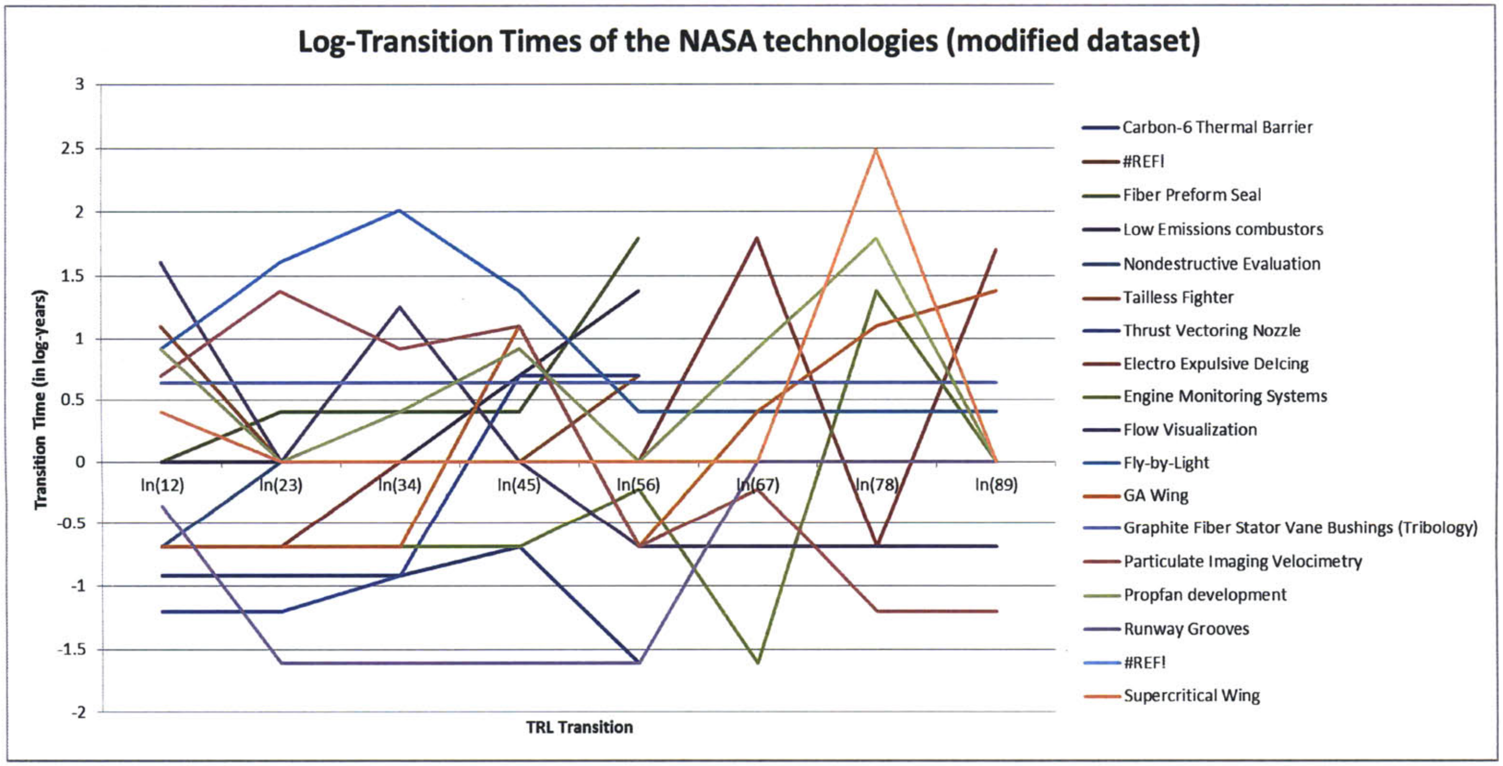


Figure 4:2 Log-Transition Times of the NASA technologies (outliers removed)

4.2 The Army Dataset

The Army dataset is a larger one, but of lesser quality and lesser importance to this thesis. It will be used in only one application where its poor properties have little effect on the result (section 7.3.1). It contains partial TRL transition times of 582 technologies/components that were commissioned between FY 2005 and FY 2009. The data were extracted from Army Acquisition Technology Objectives (ATOs) PowerPoint briefings. The briefings contained charts with the projected milestone and TRL dates for the critical components. We manually extracted the TRL transition times from those charts and stored them in a database.

As a result, the dataset has two major quality problems:

1. It is based on contracts/promises and projected schedules (as opposed to historical ones). And we already know that the final cost/schedule in DoD acquisition programs is often different from the initial estimates. Hence, instead of having a set of technology transition times, we only have a set of initial estimates of transition times.
2. Since the data was recorded visually from low-resolution graphs (values rounded to quarters for financial/accountability reasons), it also suffers from low resolution/precision. Given the manner in which the TRL steps were placed on the graphs; we could not extract the time data with a resolution below 0.25 years. This means that the error can be as high as 0.5 years.

The dataset also suffers from a minor issue. Although it contains transition information for a high number of technologies/components, the dataset does not contain a large number of transitions for a single technology. The dataset mostly contains one or two transition variables per technology (typically transitions from TRL 3 to TRL 6). This makes it harder to use the data to look for trends in the development of a single technology. However, it definitely allows looking at the distribution of a specific transition across technologies, which is how it will be used in section 7.3.1.

In summary, although the NASA dataset does have some minor issues, it is the more reliable one, and all the important results in this thesis will be based on it, or its modified version that dismisses an outlier and uses the log of the original transition times.

Chapter 5

Chapter 5. Overview of the TRL framework

We propose the below framework to classify and study the TRL-based models. The framework classifies TRL models into levels based on the assumptions they make. The pyramid depicts the increasingly strong assumptions: while the level-1 assumption is as trivial as "TRL marks consecutive points of progression in technology development", the level-4 assumption goes far in confirming that "TRL is a measure of maturity and risk".

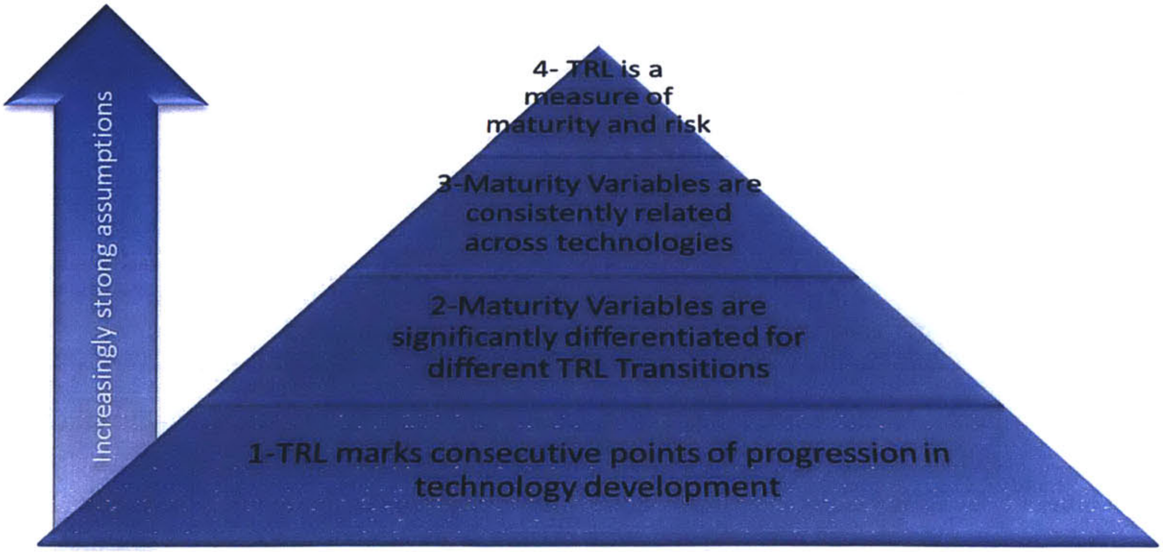

Increasingly strong assumptions
4- TRL is a measure of maturity and risk
3-Maturity Variables are consistently related across technologies
2-Maturity Variables are significantly differentiated for different TRL Transitions
1-TRL marks consecutive points of progression in technology development

Figure 5:1 The TRL framework indicating the increasing strength of the assumptions at each level

Lower levels use weak assumptions to build simple, but robust models. Higher levels require stronger assumptions to build more advanced/useful models, but that are less robust due to the strong assumptions.

Another way of classifying those four levels is by looking at the type of the uncertainty present in the models. According to Alexander (1981), there are five classes of model uncertainty ranging from the certain to the chaotic:

-Class I: Certainty

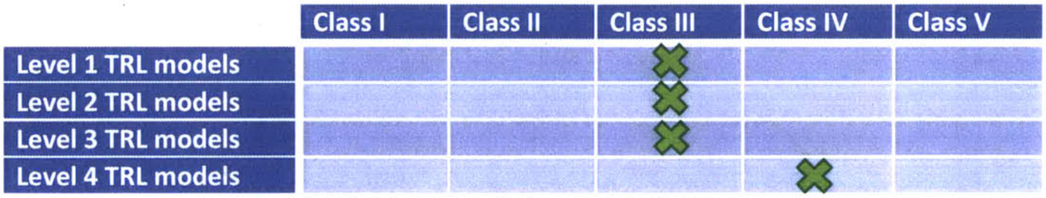
-Class II: Probability distributions of known form embedded in known models, covering known possible states

-Class III: Probability distributions of unknown form embedded in known models

-Class IV: Uncertain models (strong uncertainties)

-Class V: Chaos

The below table shows where the framework models are located with respect to this classification:



	Class I	Class II	Class III	Class IV	Class V
Level 1 TRL models			✖		
Level 2 TRL models			✖		
Level 3 TRL models			✖		
Level 4 TRL models				✖	

Table 5.1 Classification of the framework's models with respect to classes of uncertainty

Our study of the literature naturally led to the development of the framework. When we were earlier trying to develop TRL-based cost and schedule models, we realized that there was a rich literature on the approaches of using TRL to model cost and schedule risk. When we tried to improve a certain model, we discovered another model that had more useful results, but at the price of stronger (often unstated) assumptions. We found that the models in the literature can be grouped into four levels of increasingly strong assumptions. Hence the assumptions-based framework for models that use TRL emerged naturally as a way to classify those models, and to compare, develop, and improve them using the most relevant characteristic: the assumptions that they make.

The framework allows us to perform a more informed comparison or development of TRL-based models. For example, take two models A and B such that A is more precise and more robust than B. If A and B are models at the same level in our framework using the same assumptions, then the comparison is on even terms, and A is indeed "better" than B. If A however is at a higher level in our framework using a stronger assumption than B, we might want to reconsider this conclusion or we might want to consider a more nuanced one. Similarly, assume we are developing an improved version of model A called A' (A' has a lower bias, and smaller margins of error for example). A' is an improvement over A only as long as it performs better using the same assumptions. Furthermore, since the framework states the assumptions explicitly and clearly, we can use the assumption itself at each level as a guide to develop A', thus making sure that the assumptions are used to their full potential.

In order to preserve the generality of this work, we define "transition variable" as a management-relevant variable between two TRL levels for the remainder of this thesis. The presented models will look at cost and schedule transition variables (although they could also be extended to other maturity variables such as performance or technology risk). The schedule transition variables will be noted X_{i-j} and the Cost transition variables C_{i-j} (both are considered random variables). For example X_{3-4} corresponds to the time the technology takes to transition from TRL3 to TRL4, and C_{3-8} is the cost incurred in transitioning the technology from TRL3 to TRL8.

Similarly, in order to preserve generality with respect to the maturity measure, the models are also compatible with all the "TRL substitute measures" defined in section 3.2. Since those are simple adaptations of TRL to specific domains/technologies, all the essential properties of the TRL scale are preserved, and as a result, the developed models and results would still hold.

The framework is based on an earlier paper by El-Khoury and Kenley (2012), and we will adopt in this thesis a more thorough and detailed approach with respect to the definitions, the literature review, and the evidence supporting the assumptions at each level. Then we will explain in details the proposed TRL-based methodologies that make better use of the assumptions. More precisely, for each level in the framework, we will in order (1) state the assumption, (2) list the available literature relevant to that level, (3) look at theoretical evidence supporting this assumption, (4) look at empirical evidence supporting this assumption, and finally we will (5) propose new methodologies that make better use of the assumptions whenever possible (mostly at levels 2 and 3).

Furthermore, at the end of each level, we will be able to answer a practical research question:

- With a level-1 model, we will answer the research question:
  - *Does available evidence support or contradict GAO's recommendation on pre-production maturation?*
- At level 2, we will answer the questions:
  - *Does it make statistical sense to look at the distribution of each TRL transition time separately?*
  - *If yes, is there a way to improve the accuracy and fidelity of the estimates and confidence intervals?*

- At level 3, we will develop a model that answers:
    - *If we use historical data of a technology's development, can we significantly improve the accuracy of the technology schedule forecast over level-2 models?*
- And at level 4, we will answer:
    - *Out of the available methodologies, what is the best one to use in performing regression of cost/schedule against TRL?*

The next four chapters will be dedicated to the respective study of the assumptions that define the four levels of the framework.

Chapter 6

Chapter 6. Level-1 assumption

6.1 Definition and literature review

This first-level assumption is that "TRL marks consecutive points of progression in technology development". While this assumption might sound very basic and trivial, it can already have important managerial implications since it directly relates to GAO's recommendation on technology transition risk. GAO's 1999 influential report advocating for the use of TRLs can be considered as the major example in the literature using this assumption. GAO (1999) recommended maturing technologies to at least TRL7 to reduce risks and unknowns adequately before proceeding with engineering and manufacturing development. Other sources that mention this basic reverse relation between TRL scores and technology risk include LeGresley (2000) and Nolte (2008).

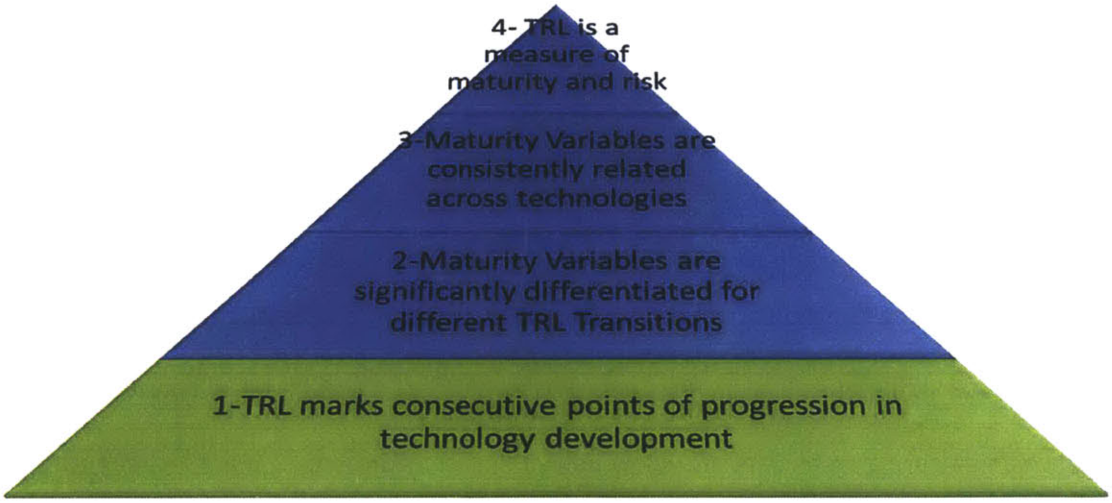

4-TRL is a measure of maturity and risk
3-Maturity Variables are consistently related across technologies
2-Maturity Variables are significantly differentiated for different TRL Transitions
1-TRL marks consecutive points of progression in technology development

Figure 6:1 The TRL framework, the first-level assumption

By the end of this chapter, we will have answered the following research question:

- *Does available level-1 empirical evidence support or contradict GAO's recommendation on pre-production maturation?*

6.2 Theoretical evidence

The first-level assumption states that "TRL marks consecutive points of progression in technology development". This assumption appears trivial and very hard to contest. It actually comes from two basic properties of the TRL scale mentioned in 3.3.3.1: the TRL scale is "complete" and "monotonic".

The scale is "complete" in the sense that it covers the entire technology development space: every technology has to be at a defined TRL at any point in time; and, throughout its development, it has to go through all of the 8 TRL transitions. The scale is "monotonic" because TRL always goes through those transitions in the same increasing order.

As a direct result, TRL will always mark consecutive points of progression (in an increasing order) in technology development.

One way this assumption could be violated is when technology obsolescence is considered. Considering obsolescence could lead to the problem of decreasing TRLs (i.e. loss of monotonicity) due to the loss of technology manufacturing capabilities and know-how. However, obsolescence is not an issue in this case since we are only interested in the initial technology development phase of the technology life cycle (refer to figure 3.1). Hence, all the framework models follow maturity variables until TRL reaches level 9 (or until the level at which the technology stops maturing, in case the development is halted prior to TRL 9).

6.3 Empirical evidence

The assumption is too basic to be supported by empirical evidence, it is obvious that all technologies are at a certain TRL, and all technologies have historically progressed on the TRL scale in an increasing manner. This is why we will look at empirical evidence supporting the GAO recommendations built on the assumption.

In simple terms, the GAO risk levels come from the fact that the higher the TRLs, the smaller the remaining overall uncertainty. A project at TRL2 is subject to risks (cost, schedule, technology) on transitions from TRL2 to TRL9, while a project at TRL6 is only subject to risks on transitions from TRL6 to

TRL9. This is a direct consequence of the fact that all technologies have to go through all TRL transitions (completeness), and they have to go through them in order (monotonicity).

GAO illustrates this concept by showing how the risks (or unknowns) are reduced as the TRL increases and as the product moves towards meeting the requirements. In particular, the report (GAO, 1999) identifies the TRL 6-7 transition as the threshold for going from high risk to low risk (also refer to figure 3.5).

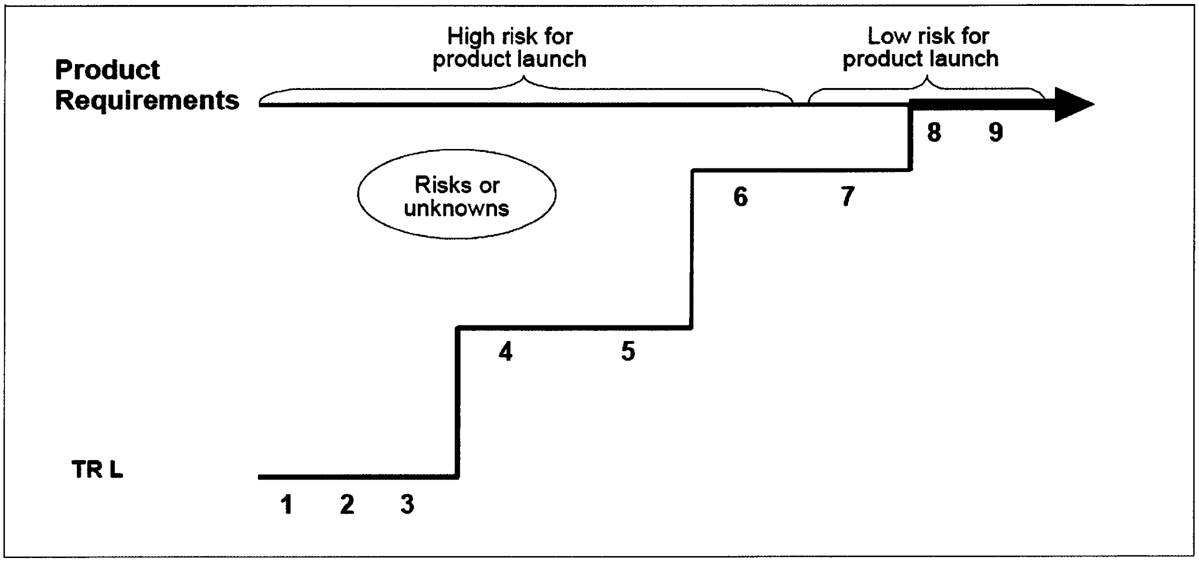


Figure 6:2 GAO technology transition risk

The above diagram illustrates the conceptual distance in maturity between the product's current state and the requirements. However, it is possible to quantify this uncertainty. Instead of using the abstract multi-dimensional concept of "maturity", we will choose a single managerial variable. In this case, we will measure schedule risk by making use of the NASA dataset.

We will use the standard deviation of the "time to maturity" as a proxy for uncertainty, or the remaining schedule risk. In the below graph, we plotted the standard deviation of the time to maturity (i.e. the remaining development time) for each TRL. For example the value indicated at TRL 3 is the standard deviation of X_{3-9} (the transition time from TRL 3 to TRL 9), while the value indicated at TRL 4 is the is the standard deviation of X_{4-9} (the transition time from TRL 4 to TRL 9), etc.

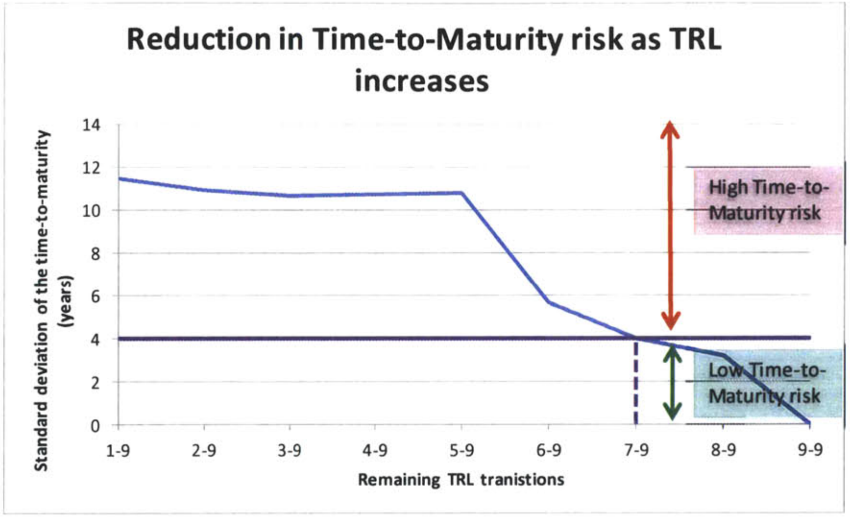


Figure 6:3 Reduction in Time-to-Maturity risk as TRL increases

We can see that the time-to-maturity risk has very high values (Standard deviation larger than 10 years) up until TRL 5. Once the project passes TRL 6, the project manager has higher control over the schedule as the standard deviation drops sharply to 5.6 years. This risk continues dropping such that GAO defines the threshold of low risk as that beyond TRL7. This graph confirms GAO's recommendations because the transition to TRL 7 corresponds indeed to a significant drop in schedule risk when compared to earlier stages (the transition from TRL 7 to TRL 8 does not correspond to a substantial drop in schedule risk, hence it makes sense to set the risk reduction threshold at TRL 7).

This analysis allows us to answer the level-1 research question: while the theoretical evidence supports the GAO recommendations in a general sense, the empirical evidence strongly supports those recommendations. Our quantitative analysis of schedule risk showed that TRL 7 in particular is well-placed to be a cut-off value from high to low schedule risk, which is perfectly in accordance with the GAO model.

Chapter 7

Chapter 7. Level-2 assumption

7.1 Definition and literature review

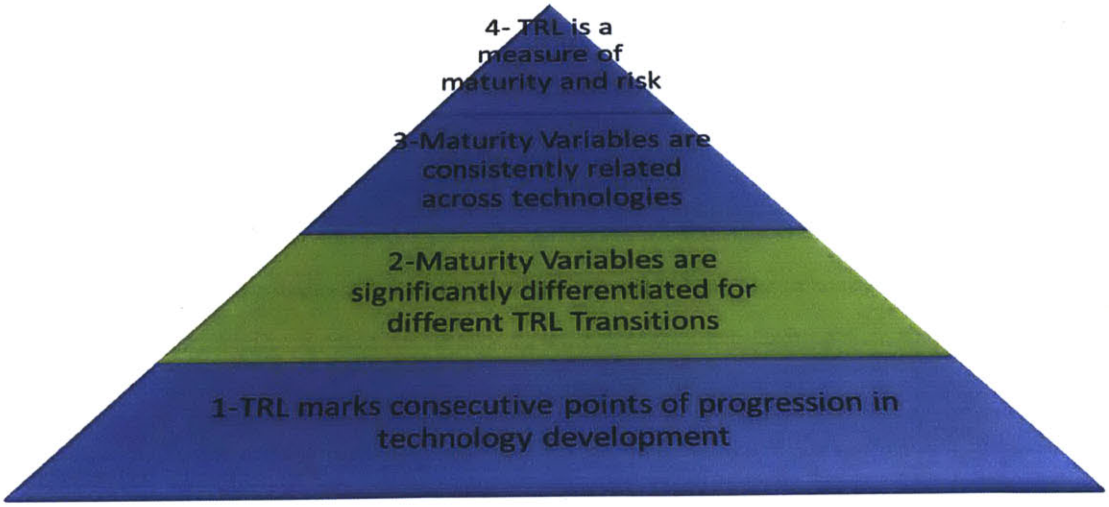

4- TRL is a measure of maturity and risk
3-Maturity Variables are consistently related across technologies
2-Maturity Variables are significantly differentiated for different TRL Transitions
1-TRL marks consecutive points of progression in technology development

Figure 7:1 The TRL framework, the second-level assumption

The level-2 assumption states "maturity variables are significantly differentiated for different TRL transitions." This means that when we look at one technology transition, the maturity variables have a probability distribution different enough from other TRL transitions and with a low enough variance. For instance, Peisen and Schulz (1999) noted that there is "considerable variability" in the time that technologies take to mature. This level-2 assumption stipulates that this variability is low enough for statistical forecasting to be applied on the technology maturation time.

Many papers in the literature fall within this category of models, starting with the NASA SAIC paper itself (Peisen and Schulz, 1999). The authors characterize technology transition times analyzing their distributions and by comparing averages and standard deviations of different subgroups of the sample. Similarly Dubos and Saleh (2010) evaluated the distribution of every TRL transition time. They found that TRL transition times have lognormal distributions and used that to propose average estimators and confidence intervals. Finally, Lee and Thomas (2003) analyzed the distributions of absolute and relative cost growth at each TRL, and fitted the Johnson's distributions to the data.

In this chapter, we will clarify and validate the assumptions that allow us to use those methodologies. Then we will examine alternative methodologies that are more suitable to the typically small datasets of transition variables.

By the end of this chapter, we will have answered the two following research questions:

- *Does it make statistical sense to look at the distribution of each TRL transition time separately?*
- *If yes, is there a way to improve the accuracy and fidelity of the estimates and confidence intervals?*

7.2 Theoretical evidence

This second-level assumption is still a weak one. It is theoretically supported by the very design of the TRL scale. Since each transition in the scale corresponds to a well-defined action common to any technology development, we should expect each of those transitions to share common properties across technologies, and thus we should expect them to be significantly differentiated from other transitions.

For example, the TRL1-2 transition that happens when "an application of the basic principles is found" is different from the TRL2-3 transition that corresponds to "going from paper to lab experiments", which itself is different from the TRL6-7 transition that happens when "the prototype is tested in the real environment". The descriptions of those transition processes are clear enough to expect their properties to be different from each other while being coherent across projects.

On a practical level, this is equivalent to making an assertion of the type: "it generally costs less (in a statistically significant sense) to go from theoretical papers to feasibility papers than to go from a prototype in a lab environment to a prototype in the operational environment", or "TRL transitions 1-2 and 2-3 are characterized well enough (i.e. they have low enough variances) that we gain information by studying the distributions of transition variables X_{1-2} and X_{2-3} as opposed to only looking at the distribution of X_{1-3}."

7.3 Empirical evidence

7.3.1 Analysis Of Variance

To validate this assumption empirically, we will look at the distribution of different technology transition variables and test if the transition times are precise enough to be statistically distinguishable.

First, taking the historical data to generate a representative empirical distribution is justified by two facts:

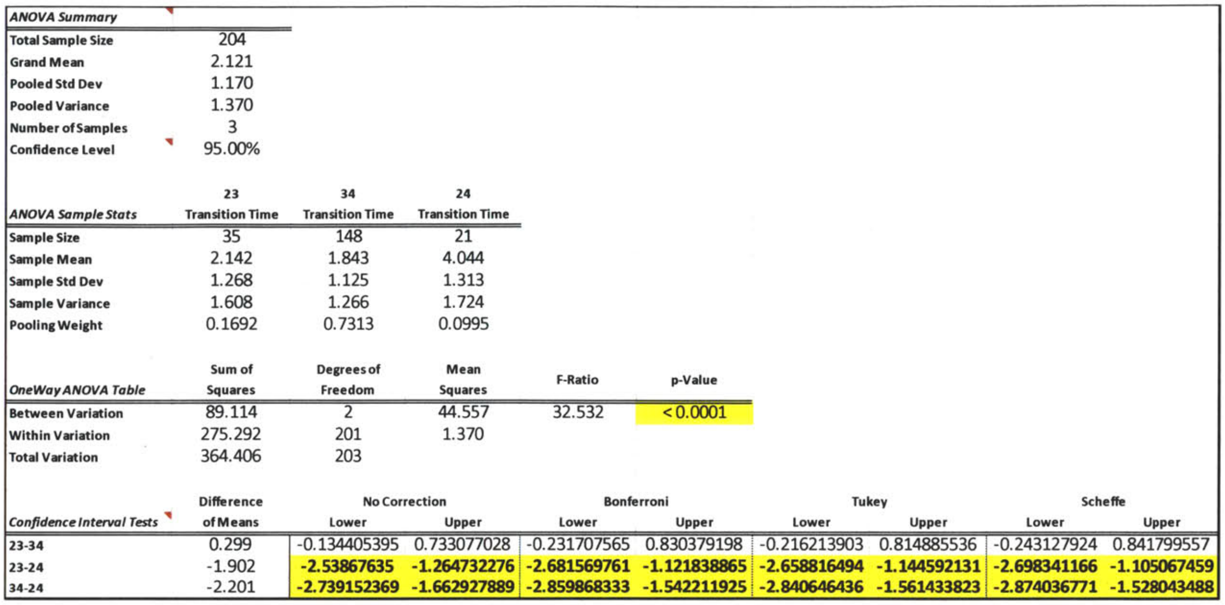
1. The technologies are independent. Hence historical frequency does represent probability.
2. The sample is representative of the population. Hence we assume no discrepancies due to sampling issues.

The way we make sure that each transition has a low enough variance so that it is distinguishable within the scale is by using an Analysis Of Variance (ANOVA). We perform the ANOVA test on transition times X_{1-2}, X_{2-3}, and X_{1-3}. Transitions X_{1-2} , X_{2-3} are "distinguishable" if their variances are small enough so that they are both different from transition X_{1-3} (Otherwise, TRL2 would be an unnecessary step, because introducing it would not add any statistically significant information over just using TRLs 1 and 3). In statistical terms, we are testing if the means of X_{1-2} and X_{2-3} are equal to that of X_{1-3} at a 95% significance level (Albright at al. 2006). If the equality hypothesis is rejected (i.e. the confidence interval of the difference of means does not contain the value zero), then X_{1-2}, X_{2-3} are statistically distinguishable from X_{1-3}, hence TRL 2 does improve the scale's resolution in measuring transition times.

We performed this test for all couples of TRL transitions where data was available (all transitions between TRL 2 and TRL7). We used the Army dataset because the NASA dataset was not large enough to establish small 95% confidence intervals.

Before performing the ANOVA test, we were concerned about the two major weaknesses of the Army dataset and their effect on the results. The fact that the dataset is low resolution (low precision) does not fundamentally alter the results because the same imprecision in X_{1-2} and X_{2-3} get transferred to the sum in X_{1-3}. The imprecision gets cancelled out later when X_{1-2} and X_{2-3} are subtracted from X_{1-3} to compare the difference to zero in the test. The fact that the Army data consists mainly of promises as opposed to historical values does weaken the test's result however. We know that promises are usually biased downwards, and there is a high chance that actual contract transition times will have higher variances than the ones agreed upon in the contract. Both of those conditions could falsely lead us to the rejection of the mean equality hypothesis in the ANOVA test. To counter this problem, we added random normal noise to the transition times (with a mean of 25% of the average transition time, and a standard deviation of 50% of the average transition time, in line with historical observations).

Once those data quality issues were resolved, we performed a total of eight statistical comparisons. In the results below, we see that all the equality hypotheses are rejected with very low p-values indicating that the transition times are indeed significantly differentiated.

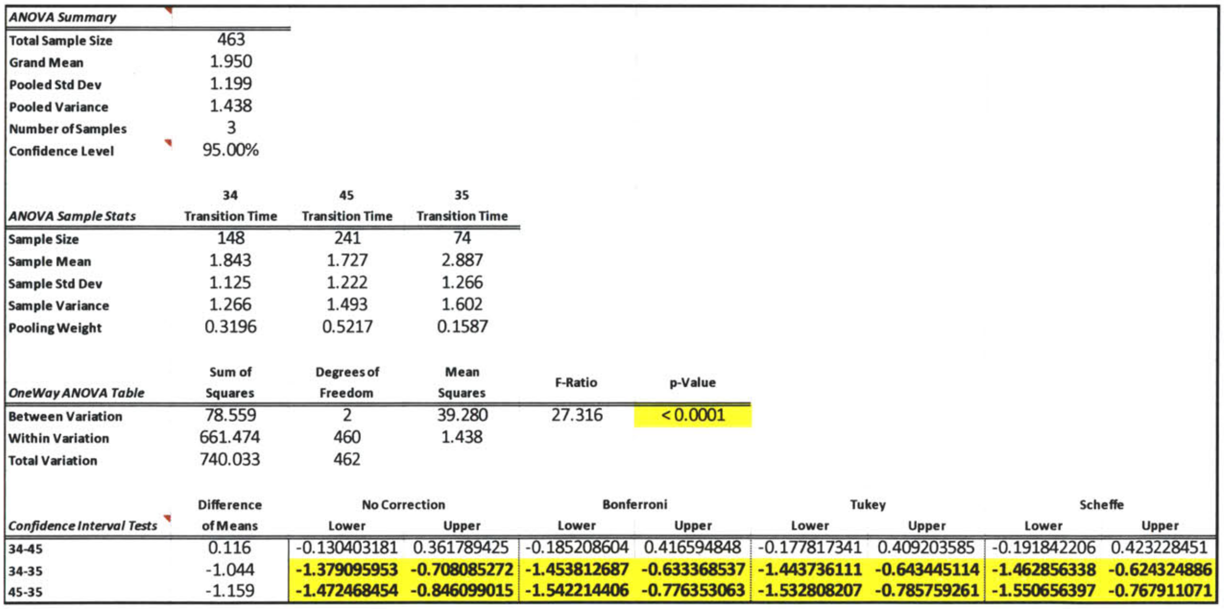


ANOVA Summary	
Total Sample Size	204
Grand Mean	2.121
Pooled Std Dev	1.170
Pooled Variance	1.370
Number of Samples	3
Confidence Level	95.00%

ANOVA Sample Stats	23 Transition Time	34 Transition Time	24 Transition Time
Sample Size	35	148	21
Sample Mean	2.142	1.843	4.044
Sample Std Dev	1.268	1.125	1.313
Sample Variance	1.608	1.266	1.724
Pooling Weight	0.1692	0.7313	0.0995

OneWay ANOVA Table	Sum of Squares	Degrees of Freedom	Mean Squares	F-Ratio	p-Value
Between Variation	89.114	2	44.557	32.532	< 0.0001
Within Variation	275.292	201	1.370		
Total Variation	364.406	203			

	Difference	No Correction		Bonferroni		Tukey		Scheffe	
Confidence Interval Tests	of Means	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
23-34	0.299	-0.134405395	0.733077028	-0.231707565	0.830379198	-0.216213903	0.814885536	-0.243127924	0.841799557
23-24	-1.902	**-2.53867635**	**-1.264732276**	**-2.681569761**	**-1.121838865**	**-2.658816494**	**-1.144592131**	**-2.698341166**	**-1.105067459**
34-24	-2.201	**-2.739152369**	**-1.662927889**	**-2.859868333**	**-1.542211925**	**-2.840646436**	**-1.561433823**	**-2.874036771**	**-1.528043488**

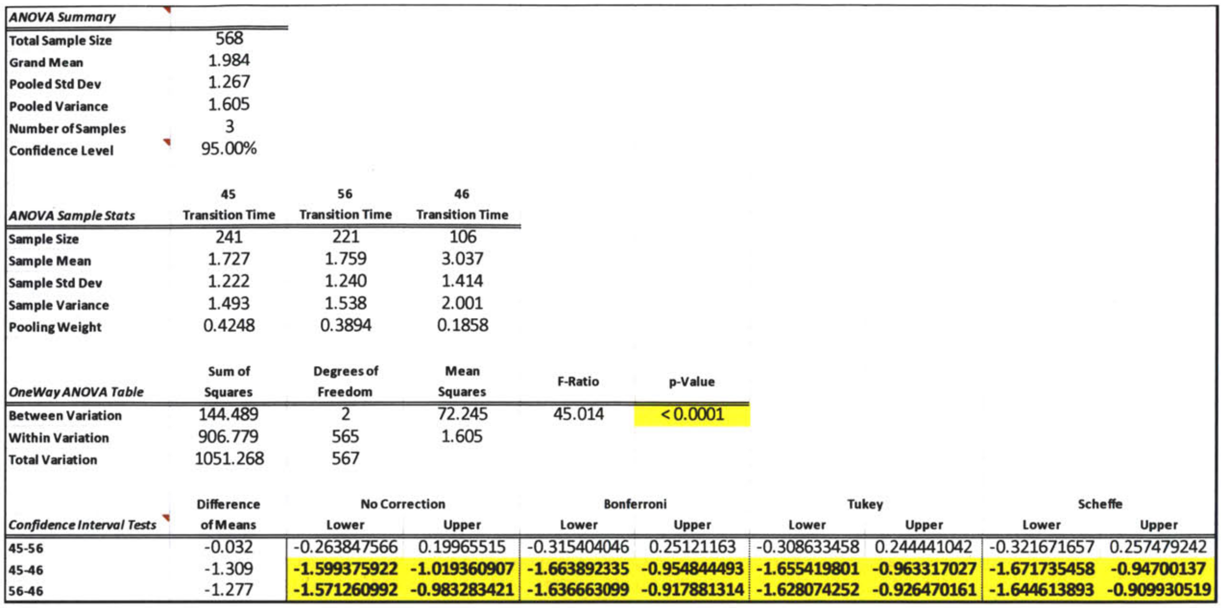


ANOVA Summary	
Total Sample Size	463
Grand Mean	1.950
Pooled Std Dev	1.199
Pooled Variance	1.438
Number of Samples	3
Confidence Level	95.00%

ANOVA Sample Stats	34 Transition Time	45 Transition Time	35 Transition Time
Sample Size	148	241	74
Sample Mean	1.843	1.727	2.887
Sample Std Dev	1.125	1.222	1.266
Sample Variance	1.266	1.493	1.602
Pooling Weight	0.3196	0.5217	0.1587

OneWay ANOVA Table	Sum of Squares	Degrees of Freedom	Mean Squares	F-Ratio	p-Value
Between Variation	78.559	2	39.280	27.316	< 0.0001
Within Variation	661.474	460	1.438		
Total Variation	740.033	462			

	Difference	No Correction		Bonferroni		Tukey		Scheffe	
Confidence Interval Tests	of Means	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
34-45	0.116	-0.130403181	0.361789425	-0.185208604	0.416594848	-0.177817341	0.409203585	-0.191842206	0.423228451
34-35	-1.044	**-1.379095953**	**-0.708085272**	**-1.453812687**	**-0.633368537**	**-1.443736111**	**-0.643445114**	**-1.462856338**	**-0.624324886**
45-35	-1.159	**-1.472468454**	**-0.846099015**	**-1.542214406**	**-0.776353063**	**-1.532808207**	**-0.785759261**	**-1.550656397**	**-0.767911071**

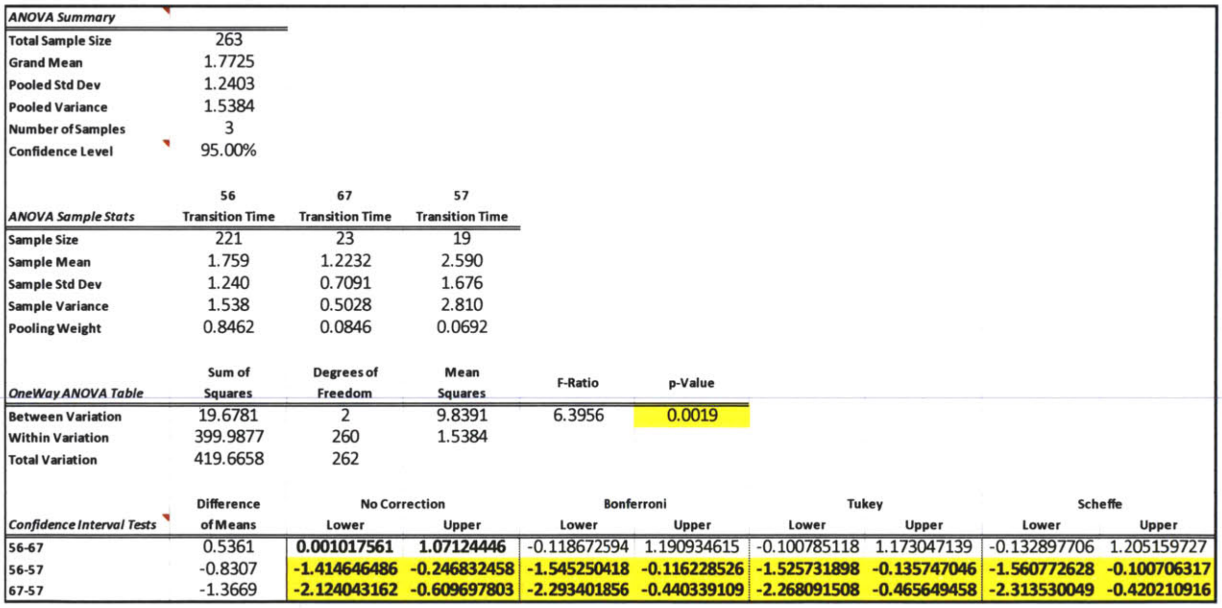


ANOVA Summary	
Total Sample Size	568
Grand Mean	1.984
Pooled Std Dev	1.267
Pooled Variance	1.605
Number of Samples	3
Confidence Level	95.00%

	45	56	46
ANOVA Sample Stats	Transition Time	Transition Time	Transition Time
Sample Size	241	221	106
Sample Mean	1.727	1.759	3.037
Sample Std Dev	1.222	1.240	1.414
Sample Variance	1.493	1.538	2.001
Pooling Weight	0.4248	0.3894	0.1858

OneWay ANOVA Table	Sum of Squares	Degrees of Freedom	Mean Squares	F-Ratio	p-Value
Between Variation	144.489	2	72.245	45.014	< 0.0001
Within Variation	906.779	565	1.605		
Total Variation	1051.268	567			

	Difference	No Correction		Bonferroni		Tukey		Scheffe	
Confidence Interval Tests	of Means	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
45-56	-0.032	-0.263847566	0.19965515	-0.315404046	0.25121163	-0.308633458	0.244441042	-0.321671657	0.257479242
45-46	-1.309	**-1.599375922**	**-1.019360907**	**-1.663892335**	**-0.954844493**	**-1.655419801**	**-0.963317027**	**-1.671735458**	**-0.94700137**
56-46	-1.277	**-1.571260992**	**-0.983283421**	**-1.636663099**	**-0.917881314**	**-1.628074252**	**-0.926470161**	**-1.644613893**	**-0.909930519**



ANOVA Summary	
Total Sample Size	263
Grand Mean	1.7725
Pooled Std Dev	1.2403
Pooled Variance	1.5384
Number of Samples	3
Confidence Level	95.00%

	56	67	57
ANOVA Sample Stats	Transition Time	Transition Time	Transition Time
Sample Size	221	23	19
Sample Mean	1.759	1.2232	2.590
Sample Std Dev	1.240	0.7091	1.676
Sample Variance	1.538	0.5028	2.810
Pooling Weight	0.8462	0.0846	0.0692

OneWay ANOVA Table	Sum of Squares	Degrees of Freedom	Mean Squares	F-Ratio	p-Value
Between Variation	19.6781	2	9.8391	6.3956	0.0019
Within Variation	399.9877	260	1.5384		
Total Variation	419.6658	262			

	Difference	No Correction		Bonferroni		Tukey		Scheffe	
Confidence Interval Tests	of Means	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
56-67	0.5361	**0.001017561**	**1.07124446**	-0.118672594	1.190934615	-0.100785118	1.173047139	-0.132897706	1.205159727
56-57	-0.8307	**-1.414646486**	**-0.246832458**	**-1.545250418**	**-0.116228526**	**-1.525731898**	**-0.135747046**	**-1.560772628**	**-0.100706317**
67-57	-1.3669	**-2.124043162**	**-0.609697803**	**-2.293401856**	**-0.440339109**	**-2.268091508**	**-0.465649458**	**-2.313530049**	**-0.420210916**

Figure 7:2 Results of the eight ANOVA analyses. The relevant confidence intervals and the p-values are highlighted.

Although the conditions for the ANOVA test were not perfectly respected (we had quasi-normality of a couple of distributions, and the variances were almost equal), we believe the results still hold because the confidence intervals were far from zero by comfortable margins (very low p-values) and the results held for all three types of corrections of the confidence interval. The Box-whisker plots (Annex 1) visually confirm this same conclusion.

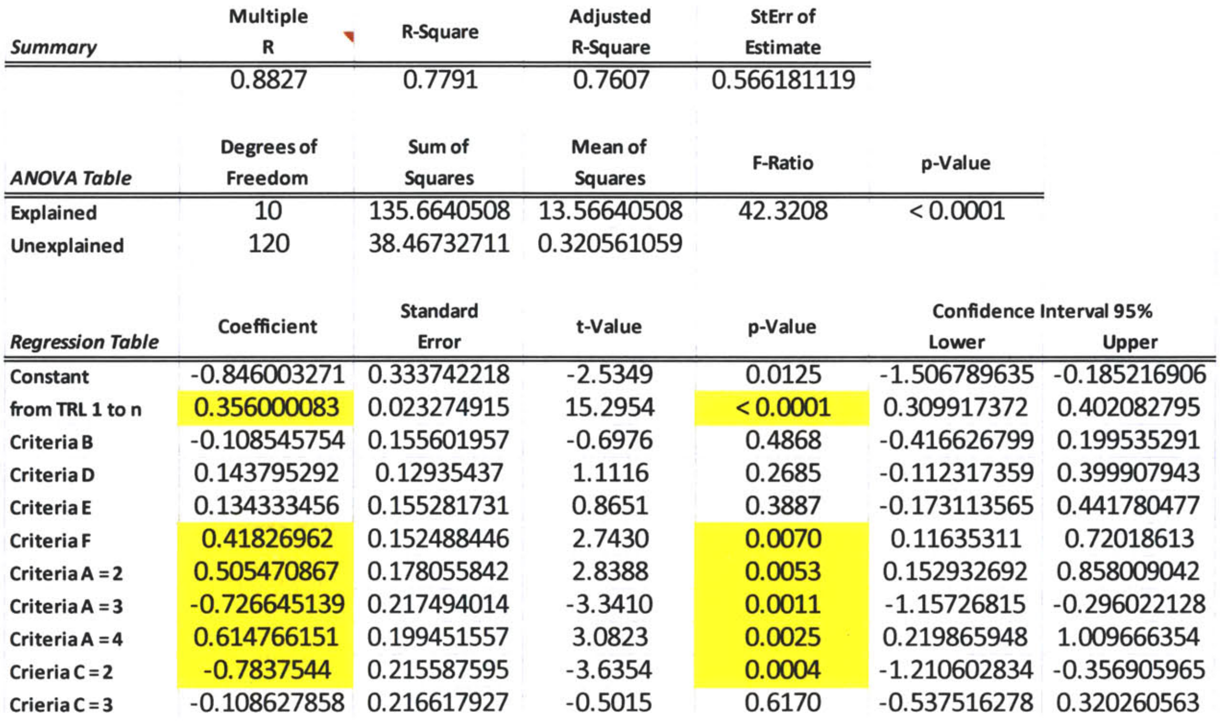
As a result, we have gone beyond the ordinality property that we used in the level 1 assumption. We have showed that the distance (in terms of maturation time) between the TRLs is statistically significant. This justifies the TRL definitions with respect to the "resolution" of the scale: not only do the TRLs convey information about order on the scale, their definitions divide the scale in a representative manner such that we would lose information/precision if a TRL step was omitted from the scale.

7.3.2 Validity of the results across agencies

The above evidence supports that the TRL transition times are statistically distinguishable within the Army dataset. This property seems to hold across technologies and agencies. In a study of a dataset from DoE's Nuclear Materials Stabilization Task Group, Crepin, El-Khoury, and Kenley (2012) developed an algorithm to transform Technology Maturity (TM) scores into TRLs. They used multiple statistical tests to show that DoE TRL transition times were similar to NASA's TRL transition times (for example, X_{4-5} values for NASA are not statistically different from DoE's X_{4-5} values). This correspondence also holds when we compare Army TRL transition times to NASA TRL transition times. This is a powerful result that generalizes the level-2 assumption across technologies and agencies.

7.3.3 Variance reduction through sample segmentation

For management purposes however, we have to reduce variance to generate better estimates. One such way of reducing variance is by performing a relevant segmentation of the population. By creating more homogenous samples, we should be able to have smaller variances for the transition times, and more accurate estimates. We performed such an analysis on the NASA dataset by performing a regression of the log-transition times against TRL and dummies of the A-F criteria accompanying the dataset. Below are the results of the regression:



Summary	Multiple R	R-Square	Adjusted R-Square	StErr of Estimate
	0.8827	0.7791	0.7607	0.566181119

ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F-Ratio	p-Value
Explained	10	135.6640508	13.56640508	42.3208	< 0.0001
Unexplained	120	38.46732711	0.320561059		

Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Upper
Constant	-0.846003271	0.333742218	-2.5349	0.0125	-1.506789635	-0.185216906
from TRL 1 to n	0.356000083	0.023274915	15.2954	< 0.0001	0.309917372	0.402082795
Criteria B	-0.108545754	0.155601957	-0.6976	0.4868	-0.416626799	0.199535291
Criteria D	0.143795292	0.12935437	1.1116	0.2685	-0.112317359	0.399907943
Criteria E	0.134333456	0.155281731	0.8651	0.3887	-0.173113565	0.441780477
Criteria F	0.41826962	0.152488446	2.7430	0.0070	0.11635311	0.72018613
Criteria A = 2	0.505470867	0.178055842	2.8388	0.0053	0.152932692	0.858009042
Criteria A = 3	-0.726645139	0.217494014	-3.3410	0.0011	-1.15726815	-0.296022128
Criteria A = 4	0.614766151	0.199451557	3.0823	0.0025	0.219865948	1.009666354
Crieria C = 2	-0.7837544	0.215587595	-3.6354	0.0004	-1.210602834	-0.356905965
Crieria C = 3	-0.108627858	0.216617927	-0.5015	0.6170	-0.537516278	0.320260563

Figure 7:3 Regression of Log-transition times against TRL score (time taken from TRL 1 to reach that TRL score) and NASA criteria A-F.

The overall regression is satisfying with an adjusted R-square value of 76% (meaning that more than 76% of the variation in transition times is explained by the independent variables, after adjusting for the number of those independent variables). Out of the input variables, TRL was, as expected, the most explanatory variable: the higher TRL reached by the technology, the more time is needed. However, we can notice that criteria A, C, and F are also explanatory, all remarkably having p-values inferior to 1%.

This means that, when compared to Airframe technologies, Flight systems take 40% more time on average to reach the next TRL, Propulsion technologies take an extra 50% on average, while ground systems finish 70% earlier on average. Similarly, a program whose main benefit is safety finished 80% earlier than a program whose main benefit is cost reduction. Finally, a technology that requires NASA testing takes 40% more time to mature than a technology that does not.

This segmentation of technologies along those important characteristics makes intuitive sense, and the regression confirms this by showing rather large differences in maturation time across categories. Although such a segmentation approach appears empirically promising and useful, we chose not to continue in this direction because we are already handling a small dataset. An 18-point dataset is too small to segment, and is already a small dataset in itself for typical parametric statistical analysis. Since data scarcity is typical in technology development projects, we will focus on statistical estimation techniques that are well-suited to small datasets. When larger datasets become available, program managers will have interest in segmenting the data to reduce variance using the larger datasets. This segmentation will eventually lead to small datasets, and once again to the need for small-sample statistical estimation, further justifying the proposed methodology in the next section.

Section 7.3 can be summarized into 4 important results:

1. The TRL scale does not have an overly high number of steps.
2. The TRL levels are rather regularly placed on the maturity scale. With respect to schedule risk, the TRLs are placed at points in the maturation process such as no two TRLs are overly close or overly far.

3. 1 and 2 provide a justification to study the empirical distributions of TRL transition times for estimation purposes. Furthermore, those distributions seem to be coherent across different agencies and domains.
4. Finally, while it is quantitatively attractive to segment the data into coherent subsamples, this leads to the problem of very small datasets that can no longer be studied with classical parametric statistics. Thus the need for non-parametric estimation emerges.

7.4 Proposed methodology

7.4.1 The need for non-parametric statistics

We saw that level-2 models consist mainly of building an empirical distribution of a transition variable, and then using it to do estimations of important statistics.

Classical parametric estimation uses Student's T as the distribution of the population mean. This is true when one of the below assumptions is verified (Garvey, 2000):

-The sample size is larger than 30 (by using the Central Limit Theorem)

or

-The population itself follows a normal distribution

or

-The sample size is between 15 and 30, *and* the variable's distribution has limited skewness

For the NASA dataset, we are in neither of the above cases. The sample is less than 30 in size, and it is skewed, which also means that the population does not follow a normal distribution. As a result, we would have to make strong parametric assumptions about the distribution of the transition variable, assigning it to a specific known distribution with fixed parameters (e.g. "TRL transition time $X_{4\text{-}6}$ follows a lognormal distribution with mean μ and standard deviation σ).

In consequence, we are faced with the two following choices in building our estimates:

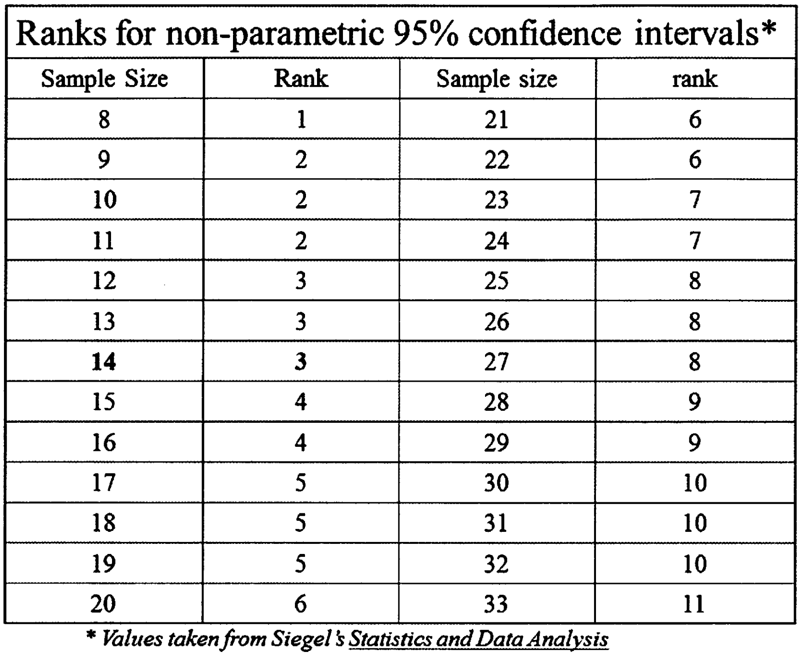
1. Still apply the classic analysis while acknowledging that the result will not be perfectly correct.
2. Apply other (non-parametric) techniques that recognize that we have little data, while trying to extract the maximum of information out of the little data we have.

In other terms, the tradeoff is between a small yet maybe incorrect confidence interval, or a larger yet correct one. We opted for the latter through the use of non-parametric estimators.

Since TRL data is typically scarce, and because of its high skewness, we substituted a more robust measure such as the median for the average (the average is very sensitive to outliers and skewness, especially in small datasets). However, it is harder to generate median estimators and confidence intervals.

7.4.2 Different non-parametric techniques to generate median confidence intervals

There are multiple ways of non-parametrically generating median confidence intervals. One simple way is to use the confidence interval quantile test that is based on the binomial distribution. The confidence interval generation uses Z-values of the T-distribution (Salter, 2006). However, Conover (1980) recommends using a different approach for small datasets by ranking the values in the sample, then using the below table to determine the confidence interval.



Ranks for non-parametric 95% confidence intervals*

Sample Size	Rank	Sample size	rank
8	1	21	6
9	2	22	6
10	2	23	7
11	2	24	7
12	3	25	8
13	3	26	8
14	3	27	8
15	4	28	9
16	4	29	9
17	5	30	10
18	5	31	10
19	5	32	10
20	6	33	11

* *Values taken from Siegel's Statistics and Data Analysis*

Table 7.1 Table to generate 95% median confidence intervals for small datasets (Conover, 1980)

In our case, after ranking the values of our sample of size 19, the 95% median confidence interval would simply be defined by the 5th value and the 15th one (or the 3rd and 10th values for a 13 data-point sample). While this method is accurate, robust to the data distribution, and easy to apply, it has a relatively large interval size.

Another technique is proposed by Olive (2005) as a modification of large-sample confidence interval formulation to improve the performance of the interval for small samples. It consists of two simple calculations that generate the median confidence interval's upper and lower bounds as a function of sample size and level of confidence.

7.4.3 The bootstrap

One of the best non-parametric techniques however is the bootstrap. It is a simple yet computationally intensive technique. It uses resampling with replacement to generate the best possible empirical distribution of the statistic of interest (Efron and Tibshirani, 1993). The bootstrap is recommended by Mooney and Duval (1993) especially in cases of asymmetrical data, even more so when the data is truncated and skewed (which is the case for all maturity variables). The bootstrap is mostly useful when we want to make inference on small-sized samples, to best characterize the little information contained in the sample without resorting to parametric assumptions.

For example, if we wanted to calculate the 95% confidence interval for the median of a dataset of size 20, we would generate a subsample of 20 elements (with replacement) and determine its median. This operation is repeated 1000 times so that we now have 1000 values for the median (500 is usually considered the minimum number of bootstrap samples for reliable results). The resulting histogram of the 1000 calculated medians is an empirical distribution of the median (Efron and Tibshirani, 1993) and is used to generate confidence intervals. To get the 95% median confidence interval, the 1000 medians are sorted in an increasing order, then the value of the 25th median would be the lower confidence limit, while the value of the 975th median would constitute the upper confidence limit.

The below graph compares the 95% median confidence intervals generated with the techniques above:

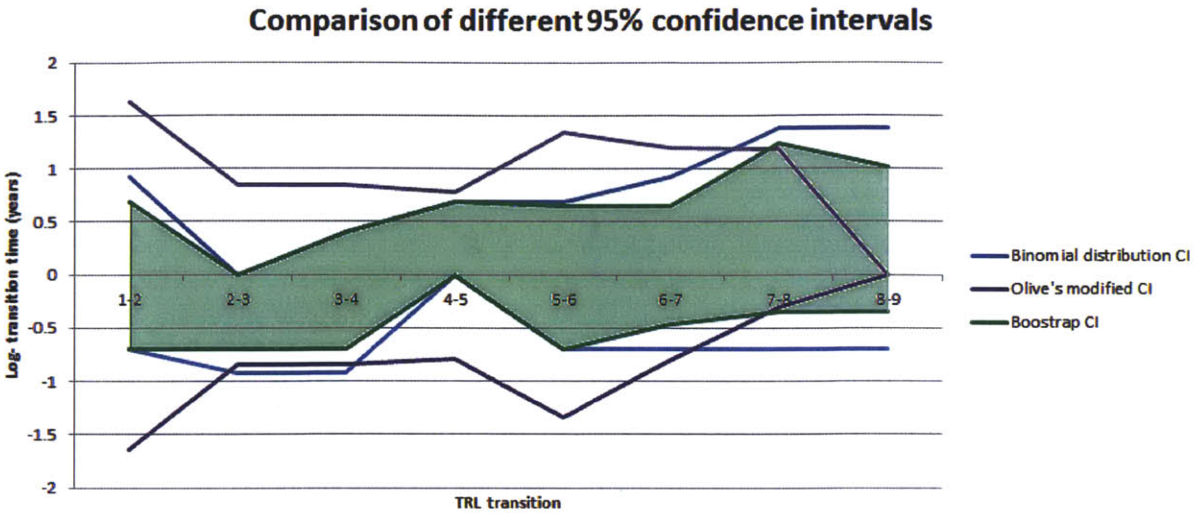
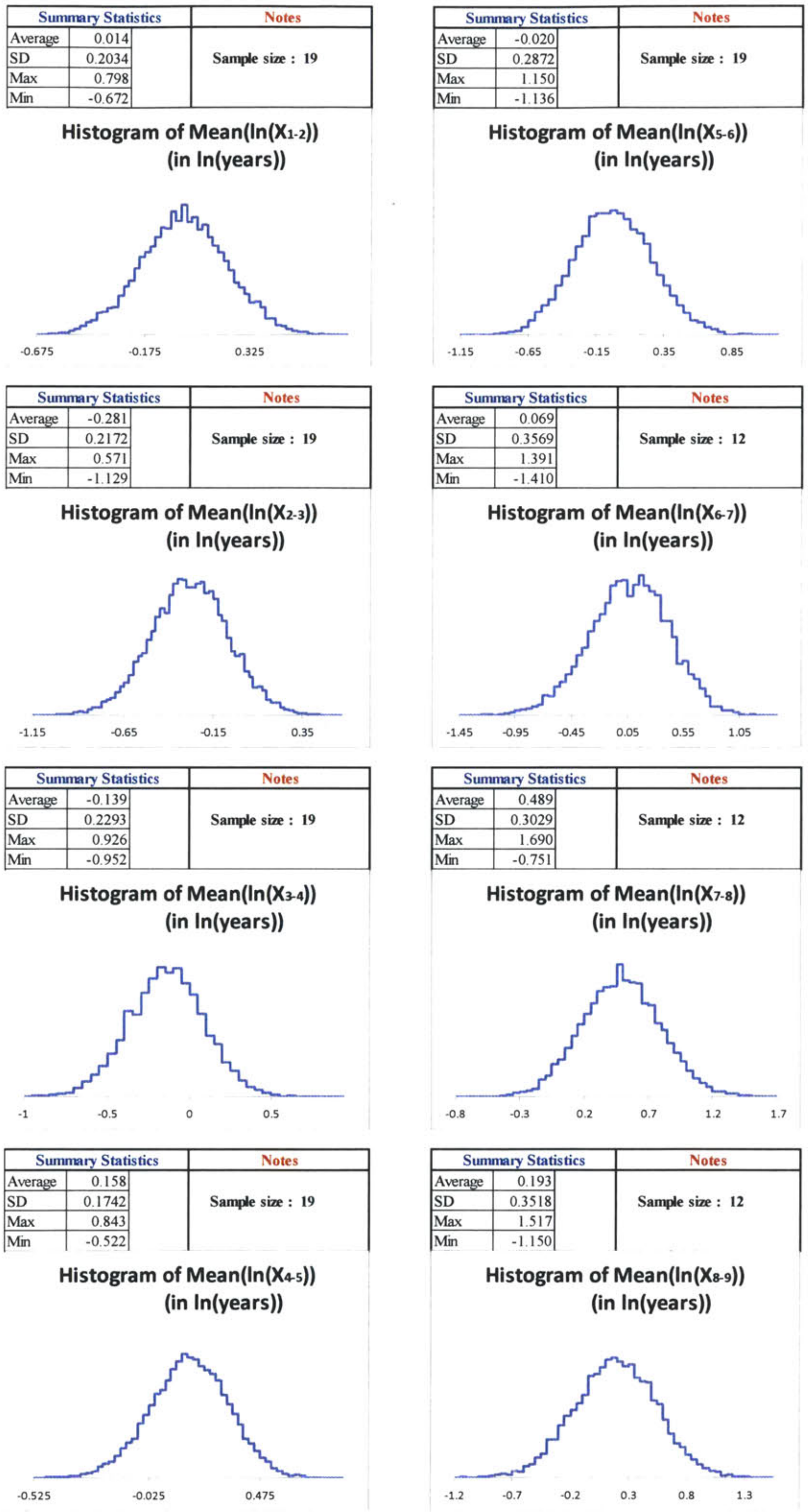


Figure 7:4 Comparison of 95% confidence intervals (of the median of log-transition time) generated using three different techniques

We can see clearly that the bootstrap confidence intervals (area highlighted in green) are the narrowest (except for the last transition where the CI generated by Olive's method was unrealistically reduced to the number zero because this value appears many times in the X_{8-9} sample). This reduced bootstrap interval size is not at the expense of CI accuracy. This performance is just a result of the method making better use of the available information.

We now present the bootstrap-generated empirical distributions of the mean (figure 7.5) and median (figure 7.6) that we obtained for the NASA log- transition times, along with some summary statistics:



Summary Statistics		Notes
Average	0.014	Sample size : 19
SD	0.2034	
Max	0.798	
Min	-0.672	

Summary Statistics		Notes
Average	-0.281	Sample size : 19
SD	0.2172	
Max	0.571	
Min	-1.129	

Summary Statistics		Notes
Average	-0.139	Sample size : 19
SD	0.2293	
Max	0.926	
Min	-0.952	

Summary Statistics		Notes
Average	0.158	Sample size : 19
SD	0.1742	
Max	0.843	
Min	-0.522	

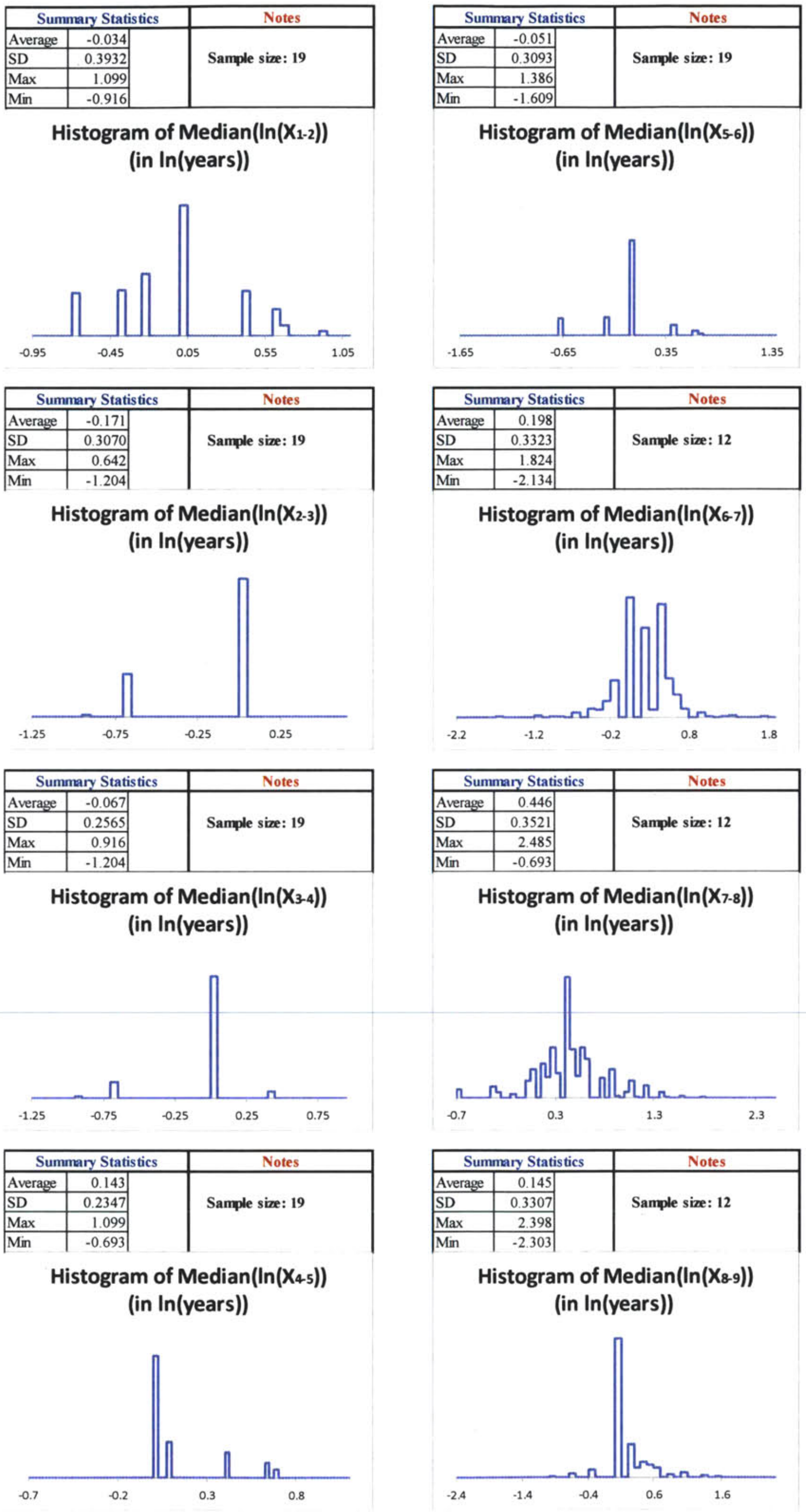
Summary Statistics		Notes
Average	-0.020	Sample size : 19
SD	0.2872	
Max	1.150	
Min	-1.136	

Summary Statistics		Notes
Average	0.069	Sample size : 12
SD	0.3569	
Max	1.391	
Min	-1.410	

Summary Statistics		Notes
Average	0.489	Sample size : 12
SD	0.3029	
Max	1.690	
Min	-0.751	

Summary Statistics		Notes
Average	0.193	Sample size : 12
SD	0.3518	
Max	1.517	
Min	-1.150	

Figure 7:5 Bootstrap-generated histograms for the means of NASA log- transition times



Summary Statistics		Notes
Average	-0.034	Sample size: 19
SD	0.3932	
Max	1.099	
Min	-0.916	

Histogram of Median(ln(X_{1-2}))
(in ln(years))

Summary Statistics		Notes
Average	-0.051	Sample size: 19
SD	0.3093	
Max	1.386	
Min	-1.609	

Histogram of Median(ln(X_{5-6}))
(in ln(years))

Summary Statistics		Notes
Average	-0.171	Sample size: 19
SD	0.3070	
Max	0.642	
Min	-1.204	

Histogram of Median(ln(X_{2-3}))
(in ln(years))

Summary Statistics		Notes
Average	0.198	Sample size: 12
SD	0.3323	
Max	1.824	
Min	-2.134	

Histogram of Median(ln(X_{6-7}))
(in ln(years))

Summary Statistics		Notes
Average	-0.067	Sample size: 19
SD	0.2565	
Max	0.916	
Min	-1.204	

Histogram of Median(ln(X_{3-4}))
(in ln(years))

Summary Statistics		Notes
Average	0.446	Sample size: 12
SD	0.3521	
Max	2.485	
Min	-0.693	

Histogram of Median(ln(X_{7-8}))
(in ln(years))

Summary Statistics		Notes
Average	0.143	Sample size: 19
SD	0.2347	
Max	1.099	
Min	-0.693	

Histogram of Median(ln(X_{4-5}))
(in ln(years))

Summary Statistics		Notes
Average	0.145	Sample size: 12
SD	0.3307	
Max	2.398	
Min	-2.303	

Histogram of Median(ln(X_{8-9}))
(in ln(years))

Figure 7:6 Bootstrap-generated histograms for the medians of NASA log- transition times

The mean histograms were included (1) to show how they compare to median histograms, and (2) as a test of the parametric normality assumption. Although the graphs approach a normal shape, they do present some irregularities and skewness.

As for the median histograms, we can make two observations. First, we can see that the histograms are very discontinuous and peak only at certain values (especially when compared to the average histograms). This comes from the fact that the median itself is a discrete statistic (it can only take a couple of predefined values that appear in the sample). Second, the last three transition variables appear to have a smoother median distribution. This can be attributed to two factors: (1) the sample size is smaller (12 data points), hence there are more chances of the bootstrap sample taking extreme values, and (2) the size of the sample is even, meaning that averages were sometimes taken to generate the median, which gives rise to new intermediary values.

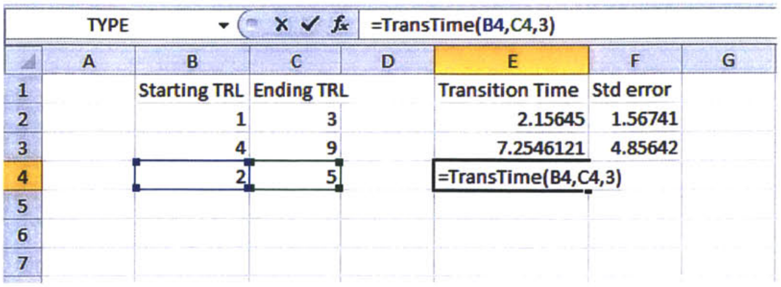
Nevertheless, though they might look unnatural (for example, a 95% confidence interval could be exactly the same as a 50% confidence interval), those distributions contain the most information that we can get out of the dataset. One way of overcoming this discrepancy in shape is by using the smoothed bootstrap (De Angelis and Young, 1992), which is a procedure that might also improve some second order properties of the estimator. It consists of adding a random noise to each of the bootstrap samples. This would eliminate the discreteness of the median leading to smoother distributions that a project manager would be more familiar with.

Another modification of the bootstrap that we can use is the iterated bootstrap - or "bootstrap within the bootstrap". To illustrate how it works using the earlier example, for each of the 1000 bootstrap samples, we would take another 500 bootstrap resamples. Those second order samples are used to find the mean and variance of each of the 1000 medians generated at the first level in order to "Studentize" them. This T-bootstrap technique would need much more computing resources, but leads to better and less biased confidence intervals according to DiCiccio and Efron (1991).

Ideally, we would have preferred to use an iterated and smoothed version of the bootstrap. This was not possible however, due to limited computing resources. As a result, the regular bootstrap (with 10,000 repetitions) was performed to generate empirical distributions for the mean and median of NASA transition times.

7.4.4 Implementing the bootstrap in Excel User Defined Functions

We saved the mean distributions in two Excel User-Defined Functions (UDFs). The user inputs the starting TRL and the ending TRL. If he is using the "TransTime" function, he would enter the percentile rank and the function returns the corresponding percentile of the TRL transition time. If he is using the "TransTimeInv" function, he would enter the time value and get the corresponding percentile rank. Thus the functions can be used to get any information about the distribution of the particular NASA transition time, including estimations and confidence intervals. Additionally, the functions are used in a fashion similar to the operation of other Excel-defined distributions (Gaussian, T, F, Chi-squared, etc.), so the manager would quickly get familiar to the way they are used.



=TransTime(B4,C4,3)

	A	B	C	D	E	F	G
1		Starting TRL	Ending TRL		Transition Time	Std error	
2		1	3		2.15645	1.56741	
3		4	9		7.2546121	4.85642	
4		2	5		=TransTime(B4,C4,3)		
5							
6							
7							

Figure 7:7 Snapshot of the Transition Time user defined function in excel

Finally, if the project manager has his own dataset and would like to determine percentile values or a confidence interval for the median for his particular dataset (as opposed to using the NASA dataset), we developed a "Bootstrap" Excel function. The user simply selects the data range, specifies the number of iterations, and selects the percentile rank (or the percentile value when using the "BootstrapInv" function). The function performs the Bootstrap procedure and returns the percentile value (resp. the percentile rank) from the empirical distribution of the median.

The VBA code of those four Excel functions can be found in Annex 2.

By now we have a clear answer to the second research question:

- *If it makes statistical sense to look at TRL transition distributions separately, is there a way to improve the accuracy and fidelity of the estimates and confidence intervals?*

Not only did we answer this question by introducing and explaining the best non-parametrical approach for small datasets (the bootstrap), we also implemented the solution in a practical Excel tool for program managers.

In summary, the level-2 assumption is made whenever we want to study the statistics of maturity variables on each of the TRL transitions.

The ANOVA test was in line with the theoretical evidence in supporting the assumption that the TRL transition variables are significantly differentiated.

While classical parametric estimators could be used, we recommend using the median bootstrap as a robust and accurate estimator better suited to small and skewed datasets (the datasets are small either from the start, or after undergoing segmentation for variance reduction).

Although the iterated and smoothed version of the bootstrap would have been most suitable in our case, limited computational resources prevented us from implementing it. We then programmed the bootstrap results in two static ("TransTime", and "TransTimeInv") and two dynamic ("Bootstrap", and "BootstrapInv") Excel User-Defined Functions.

As a final note, we recommend complementing those models with other techniques from the classical statistical analysis toolbox, such as distribution fitting, or careful outlier analysis (especially for small datasets).

Chapter 8

Chapter 8. Level-3 assumption

8.1 Definition and Literature review

This level builds on the previous one: not only are maturity variables well-defined and differentiated, they are also related in a consistent (predictable) manner. It is clearly a stronger assumption because it means that if we pick any technology, we should expect some form of stable relation between separate transitions in the same technology. For management purposes, this means that for any project, we can use early transition information to make predictions on later transition variables (while level-2 models do not use any past information).

This assumption is made whenever we make statements such as: (1) "if a technology is already maturing quickly or cheaply relative to the average, then it is more likely to continue doing so for the remainder of its development" or (2) "high research and development costs are related to high development and engineering costs".

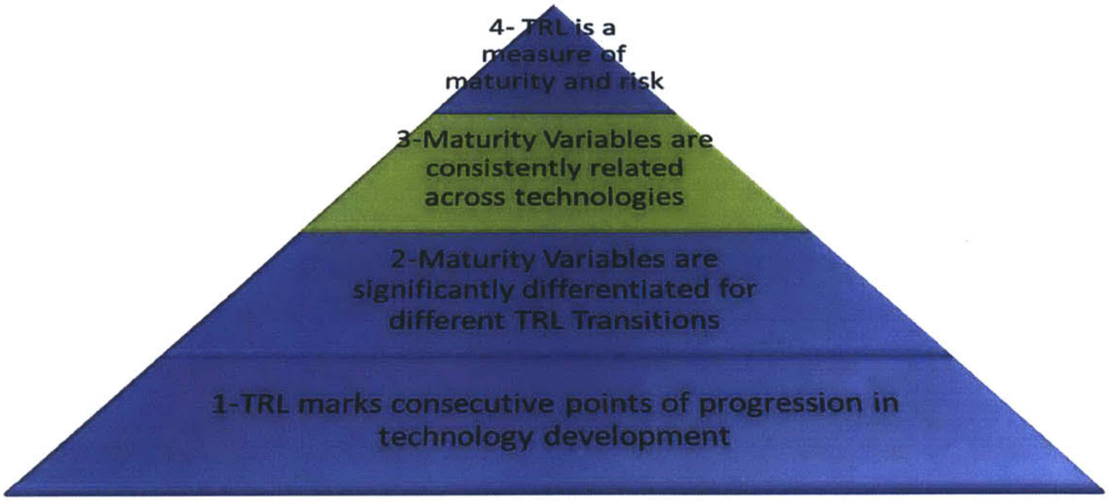

4-TRL is a measure of maturity and risk
3-Maturity Variables are consistently related across technologies
2-Maturity Variables are significantly differentiated for different TRL Transitions
1-TRL marks consecutive points of progression in technology development

Figure 8:1The TRL framework, the third-level assumption

This level is absent in the literature. It is only briefly described in El-Khoury and Kenley (2012). However, here we will explain the details of the methodology, and evaluate results from applying it to the NASA dataset.

By the end of this chapter, we will have answered the question:

- *If we use historical data of a technology's development, can we significantly improve the accuracy of the technology schedule forecast over level-2 models?*

8.2 Theoretical evidence

Many theoretical reasons partially support this assumption. We identified three lines of reasoning each supporting a different kind of relation between transition variables:

The first line of reasoning argues for a positive correlation between transition variables within a single project. The argument is that some factors make technologies take longer or cost more than average, irrespective of the TRL transition. Are the performance requirements set too high that the technology will need more than average time or cost to develop? Do the contract terms encourage the contractor to deliver early? Is the contractor more costly than average? Is there a relationship between system complexity and the total time necessary for TRL transitions? We can see that in all those cases, a third factor would impact schedule (or cost) in the same way throughout many transitions in the project.

The second line of reasoning argues for a negative correlation between transition variables. One argument could be that a project that has spent an overly long period to achieve one transition might have already done some work for the next transition making it shorter. Another argument could be that if a project was having high cost or schedule problems, then this would alert officials in the agency to put another more efficient manager in charge and make sure he compensates the earlier higher-than-average time or cost with lower-than-average time or cost later.

The third class of arguments is that cost (or schedule) has similar evolution patterns, independently of the technology. For example, Smoker and Smith (2007) present a model where cost grows exponentially relative to the initial estimate. As a result, if a project is having cost overruns in early TRLs, we could adjust our estimates for future overruns based on this assumed stable evolution pattern. Similarly, if we know for example that, for technical reasons, TRL transition 7-8 always takes 50% more time than transition 6-7, then this property can help use past data for more informed forecasts.

Those three arguments might be partially true, and the reality might well be some mix of the 3 cases. The question however is "how do all those effects and relations add up? Can we detect any trends in the data?"

8.3 Empirical evidence

Ideally, we would empirically test each of the three above hypotheses by controlling all the variables that were mentioned. But since we are far from having all the required data, we cannot look at the causes, and we can only look at the consequences. So we will try to identify relevant relations in the NASA dataset without looking for root causes.

"Relation" between random variables can be defined in many ways. We will first try to graphically detect trends or relations, and then we will do so using the correlation matrix.

Although the NASA dataset might not look prone to forecasting, we can still detect some minor trends:

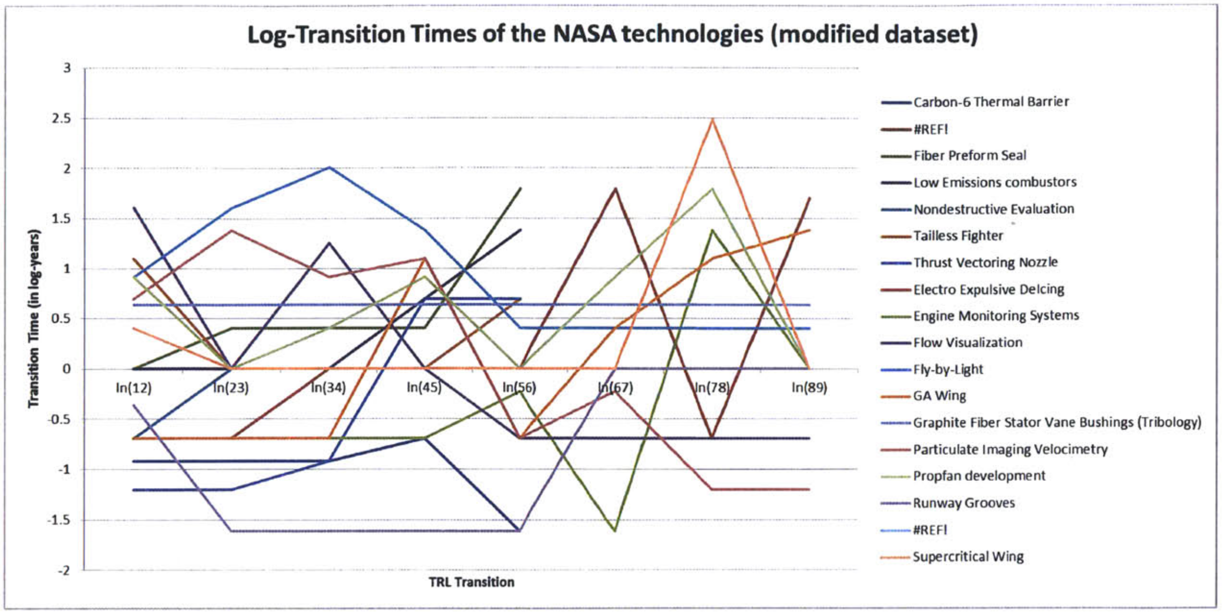
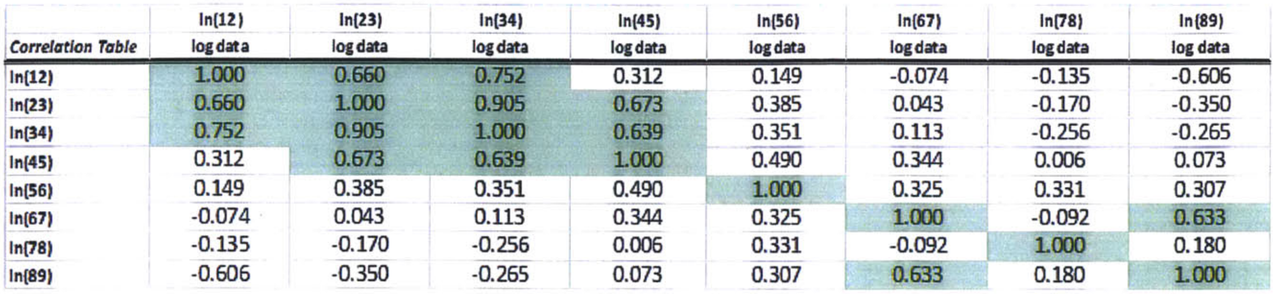


Figure 8:2 Log-transition times of NASA technologies

By looking at each curve (i.e. each technology), we can identify some trends before TRL 6. There is only a minor variation from transition 1-2 to 2-3, and almost all the technologies that have a decreasing transition time decrease in their log-transition time to 0. Afterwards, all the technologies tend to increase in transition time after 2-3 or 3-4, and then they mostly decrease again after 4-5. The technologies become less predictable after transition 5-6 since some technologies evolve at a constant pace while others show extreme oscillation. Finally, transition 8-9 appears to be more predictable as most technologies converge towards a log-transition time near zero.

Correlation analysis (table 8.1) confirms our intuition of a well-behaving group before TRL6. In the below correlation table, we highlighted correlation factors larger than 0.5. We can clearly see a cluster of highly correlated transition times 1-2, 2-3, 3-4, and 4-5, as expected. The rest of the transitions do not show any correlations except for a (minor) positive correlation between 6-7 and 8-9.



Correlation Table	ln(12) log data	ln(23) log data	ln(34) log data	ln(45) log data	ln(56) log data	ln(67) log data	ln(78) log data	ln(89) log data
ln(12)	1.000	0.660	0.752	0.312	0.149	-0.074	-0.135	-0.606
ln(23)	0.660	1.000	0.905	0.673	0.385	0.043	-0.170	-0.350
ln(34)	0.752	0.905	1.000	0.639	0.351	0.113	-0.256	-0.265
ln(45)	0.312	0.673	0.639	1.000	0.490	0.344	0.006	0.073
ln(56)	0.149	0.385	0.351	0.490	1.000	0.325	0.331	0.307
ln(67)	-0.074	0.043	0.113	0.344	0.325	1.000	-0.092	0.633
ln(78)	-0.135	-0.170	-0.256	0.006	0.331	-0.092	1.000	0.180
ln(89)	-0.606	-0.350	-0.265	0.073	0.307	0.633	0.180	1.000

Table 8.1 Correlation table of the NASA log-transition times (modified dataset)

In simple terms, this initial cluster of positive correlation means that if a technology is maturing fast (resp. slow) in early TRL stages, then it is likely to keep maturing fast (resp. slow) in later stages. This phenomenon is true up to TRL 5.

For later TRLs however, we can see that transitions 6-7, 7-8, and 8-9 are very uncorrelated with all the early TRL transition times, as if they were independent of them. This might be due to the fact that a technology changes hands after TRL 6, going from NASA to industry (Peisen and Schulz, 1999).

In summary, both the visual inspection and the correlation table show that there is indeed a relation (of positive correlation) between the TRL transition times. This correlation however, is limited to TRLs 1-to-5. As for later TRLs, they do not have a linear correlation with earlier variables, and we doubt the presence of other types of relations due to the fact that those transitions are handled separately by a different organization.

We further test those relations in the next section. In it, we develop several models to try to detect any kind of relation that would improve forecasts for transition times.

8.4 Proposed methodology

We try to exploit the effects identified above by developing several forecasting methods, and then comparing their performance. After a brief overview of the general forecasting approaches, we will present the Excel-based comparison methodology that we used to comparatively evaluate the accuracy of the forecasting techniques. Then, we will present the twelve forecasting techniques that were tested. And finally, we will discuss the major results of the comparison and make recommendations.

8.3.1 Overview of the forecasting approaches

This section briefly introduces the proposed forecasting techniques. The below graph summarizes all those forecasting approaches:

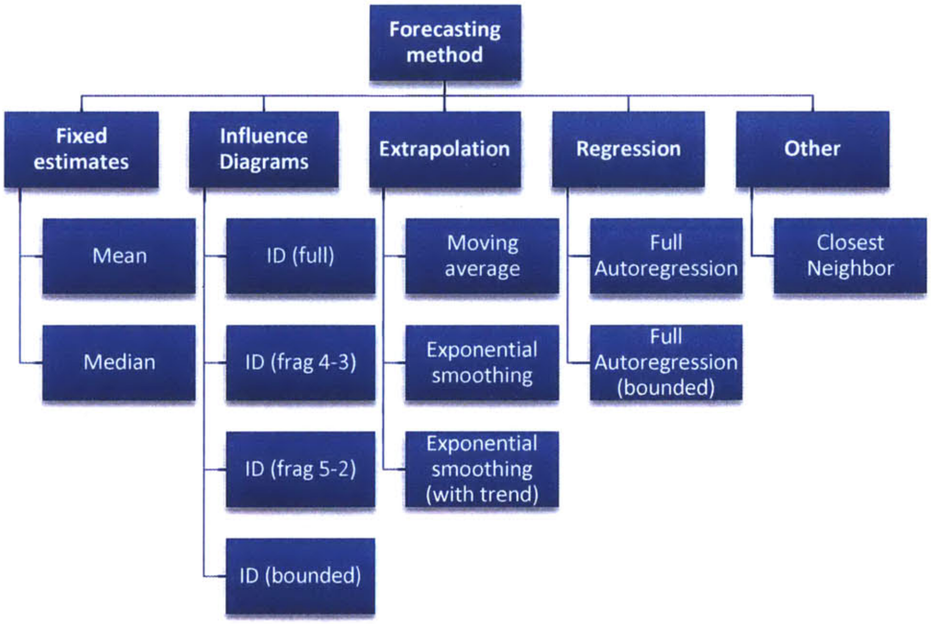


Figure 8:3 The forecasting techniques proposed for level 3

Fixed estimates methods give the same forecast for all the technologies, they do not use past transitions to forecast future ones, they constitute the reference measures. As a result, they are models that use the level-2 assumption, and they do not make use of the level-3 assumption. They will be used only as a reference to evaluate the performance of the other "smarter" level-3 techniques.

On the other end of the spectrum, extrapolation techniques do not use training data; they are only based on the past transitions of the forecasted technology itself.

Influence diagram methods use a probabilistic approach by assuming a multivariate normal distribution of the variables, and mapping the relations between them.

Regression techniques are related to the influence diagram and extrapolation techniques. The forecasted transition is regressed against the known steps (using the training data), and the results are applied to the known past transitions in order to generate the forecast.

Finally, the closest neighbor technique tries to forecast the transitions by imitating the variations of the technology in the training set that correlates the most with the technology being forecasted.

8.3.2 The comparison technique

Since a method's performance cannot be easily summarized by one number, we were faced with many comparison issues:

1- We would like to measure the forecasting error at all stages of the technology development. For example, at an early stage, when we only know the 1-2 transition, the method will have to predict all 7 future transitions; while when we're at a later stage, we'll have to forecast only one or two transitions. Classical error measures such as the Root Mean Squares Error (RMSE) and the Mean Absolute Error (MAE) are normally used to compare different methods for the same forecasted span. Averaging on the number of forecasted values is not enough to make the MAEs of predicting 7 future steps and 1 future step comparable. Hence the comparison of each "forecasted span" (number of forecasted values) should be done independently.

2- Furthermore, the error on one technology is not comparable to another one. One technology might be easy to forecast with many methods, while another one can be highly variable or random, making it hard to predict.

3- It would be helpful to have an idea about the robustness of the method with respect to the training and validation data, by having an idea of the sensitivity of the results with respect to the different forecasted technologies, but also to the different training subsets.

4- If we wanted to create some "aggregated error score" for every method; we run into the problem of considering the early forecasts (some methods cannot produce any forecast with only one or two known transitions), the problem of the importance of far forecasts vs. closer ones, and the problem of the inclusion or not of the outliers in the comparison.

We took the following measures to address those problems:

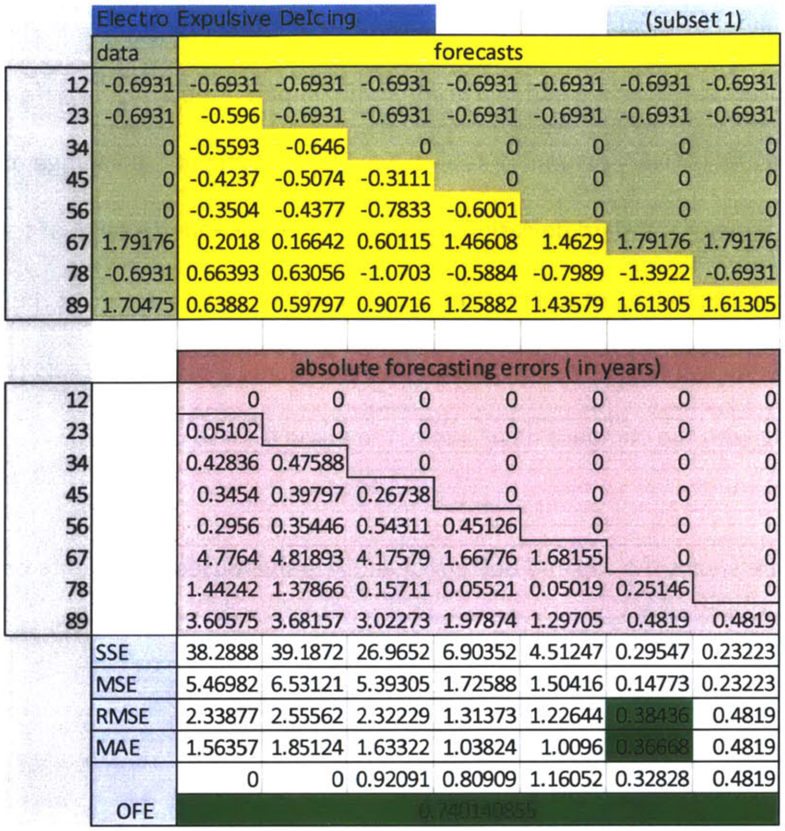
First of all, because the dataset size was a major issue, we took extra care by evaluating all methods 5 times with 5 different (random) 12-point training sets. Then we evaluated the forecasting errors (of each of the 5 forecasts) on the full dataset.

Although it is somehow optimistic (in error terms) to include the training data in the validation set, (1) it was necessary due to the small number of complete data points, (2) it was impossible to compare forecasts between the complete and the incomplete data points, (3) it helped solve the problem of cross-technology comparison, and finally (4) it did not prevent us from achieving the main goal that is more comparative than purely predictive.

For every method, we created an Excel sheet for a thorough evaluation of the technique (refer to figure 8.5). Every sheet contained 90 forecast evaluation tables (1 set of forecasts for each of the 18 technologies, repeated 5 times for each of the 5 training data samples).

The two tables in figure 8.4 show how a single forecast was performed. On top (in blue), we have the name of the technology and the training subset, while the left column (in grey) shows the 8 variables (TRL transitions). The upper table is the forecast table: the known values (i.e. past values) are in green, and the forecasted ones are in yellow. The different columns correspond to different known transitions: the first column makes forecasts for 7 transitions knowing only the 1st one, while the last column makes forecasts for the last transition knowing all the previous ones.

The lower red table shows the absolute errors of every forecast. The errors are in years (we took the difference of exponentials of the above forecasted and actual values). We aggregated those errors in 3 different ways: RMSE, MAE, and OFE (Objective Function of Error, which we will explain later).



Electro Expulsive DeIcing (subset 1)

	data	forecasts						
12	-0.6931	-0.6931	-0.6931	-0.6931	-0.6931	-0.6931	-0.6931	-0.6931
23	-0.6931	-0.596	-0.6931	-0.6931	-0.6931	-0.6931	-0.6931	-0.6931
34	0	-0.5593	-0.646	0	0	0	0	0
45	0	-0.4237	-0.5074	-0.3111	0	0	0	0
56	0	-0.3504	-0.4377	-0.7833	-0.6001	0	0	0
67	1.79176	0.2018	0.16642	0.60115	1.46608	1.4629	1.79176	1.79176
78	-0.6931	0.66393	0.63056	-1.0703	-0.5884	-0.7989	-1.3922	-0.6931
89	1.70475	0.63882	0.59797	0.90716	1.25882	1.43579	1.61305	1.61305

		absolute forecasting errors (in years)						
12		0	0	0	0	0	0	0
23		0.05102	0	0	0	0	0	0
34		0.42836	0.47588	0	0	0	0	0
45		0.3454	0.39797	0.26738	0	0	0	0
56		0.2956	0.35446	0.54311	0.45126	0	0	0
67		4.7764	4.81893	4.17579	1.66776	1.68155	0	0
78		1.44242	1.37866	0.15711	0.05521	0.05019	0.25146	0
89		3.60575	3.68157	3.02273	1.97874	1.29705	0.4819	0.4819
	SSE	38.2888	39.1872	26.9652	6.90352	4.51247	0.29547	0.23223
	MSE	5.46982	6.53121	5.39305	1.72588	1.50416	0.14773	0.23223
	RMSE	2.33877	2.55562	2.32229	1.31373	1.22644	0.38436	0.4819
	MAE	1.56357	1.85124	1.63322	1.03824	1.0096	0.36668	0.4819
		0	0	0.92091	0.80909	1.16052	0.32828	0.4819
	OFE	0.740140855						

Figure 8:4 Excel table example, showing the series of forecasts generated for one technology (for one forecasting method, for one of the training sets), and different measures of error of those forecasts

We can see that problem 1 is now addressed: the forecasts were classified by "forecasted span" (i.e. the table columns), thus the numbers obtained in MAE, RMSE, and OFE are not to be compared with each other, but should be compared across different training sets and different technologies, which means that they can now be averaged across those so that we can compare their values between different forecasting methods.

We addressed problem 2 by averaging those error measures across all technologies for every method and for every training subset (see "average of errors across subset" figure 8.5).

We addressed problem 3 by creating different horizontal and (more importantly) vertical comparisons. On every "line" of forecast tables, we added a graph that compared the MAEs of different technologies at every step. At the end of every column of tables, we added a table of the average of errors (across technologies) and a table of standard deviations of errors. Then we added a table of total OFE per transition per training set (which are summarized in the final comparison graph, at the bottom of figure 8.5). Finally, we included a grand Total OFE table with the best and median forecasts (with respect to the 5 training sets), along with the standard deviation of the grand total OFE.

This annotated picture shows the top and bottom of an excel sheet used to evaluate one method:

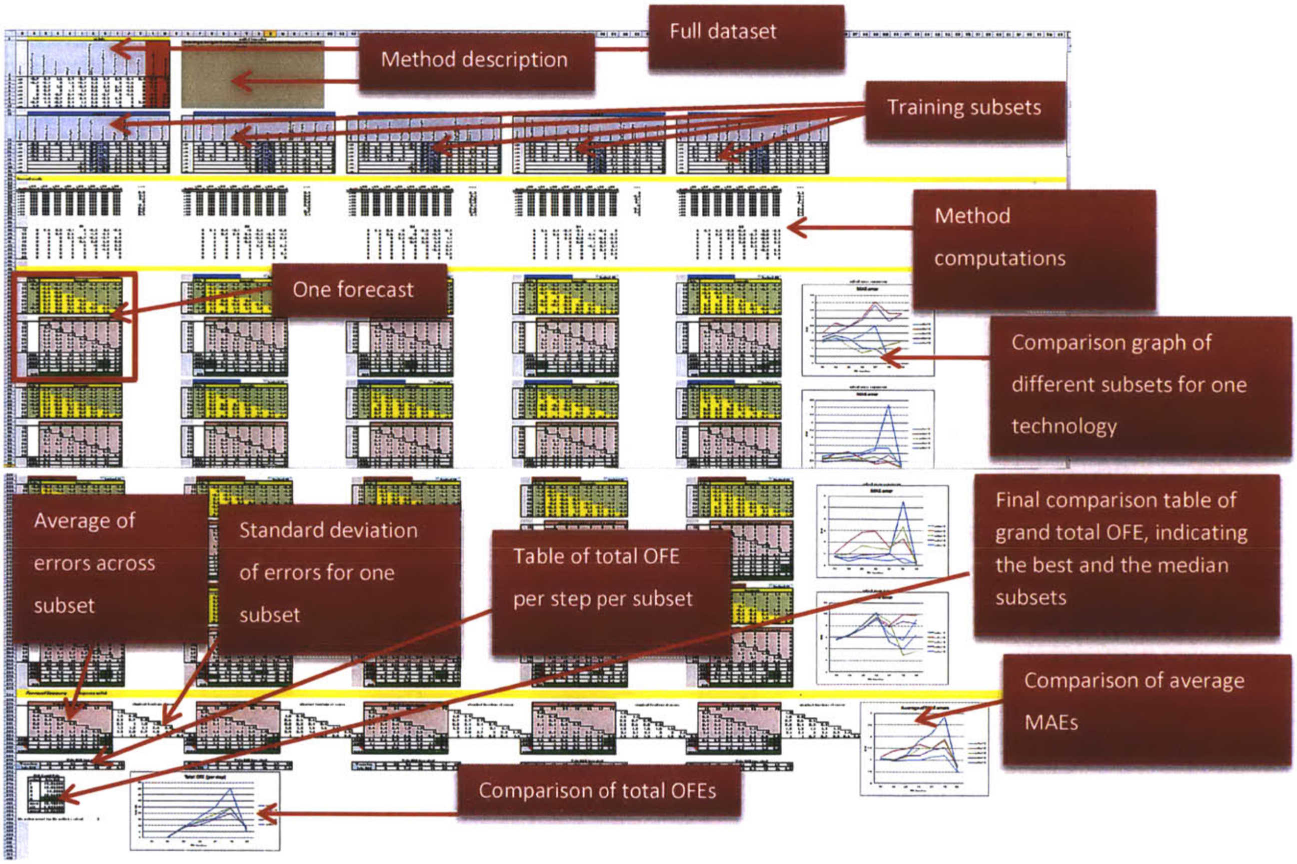
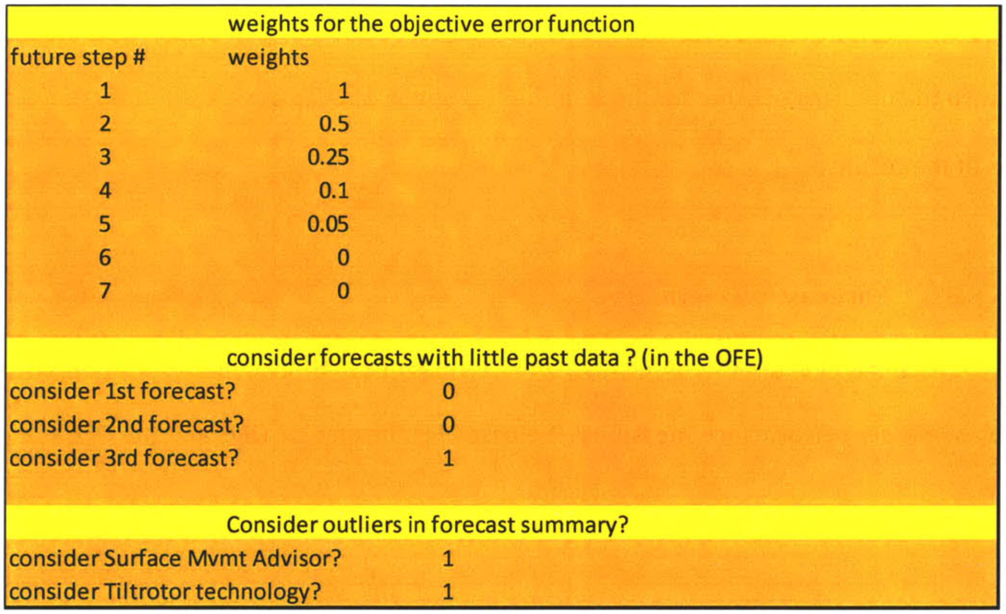

Full dataset
Method description
Training subsets
Method computations
One forecast
Comparison graph of different subsets for one technology
Average of errors across subset
Standard deviation of errors for one subset
Table of total OFE per step per subset
Final comparison table of grand total OFE, indicating the best and the median subsets
Comparison of average MAEs
Comparison of total OFEs

Figure 8:5 Annotated extract of one Excel sheet used to evaluate one forecasting method (only the top and bottom of the Excel sheet are visible)

As to how we addressed problem 4, it was mainly through a "customizable" Objective Function of Error (OFE). The "raw" error data corresponds to the 28 error values that appear in the red tables of the forecasts. The OFE's purpose is to aggregate this data in a way that reflects the user's needs. We added a worksheet of "control parameters" (figure 8.6) where the user can input the relative weight of forecast spans, he can choose whether to consider or not a particular column, and it also allows dropping a technology (mainly the outliers) out of the final total OFE.



	weights for the objective error function	
future step #	weights	
1	1	
2	0.5	
3	0.25	
4	0.1	
5	0.05	
6	0	
7	0	
	consider forecasts with little past data ? (in the OFE)	
consider 1st forecast?		0
consider 2nd forecast?		0
consider 3rd forecast?		1
	Consider outliers in forecast summary?	
consider Surface Mvmt Advisor?		1
consider Tiltrotor technology?		1

Figure 8:6 The Objective Function of Error (OFE) control parameters Excel sheet

The final "comparison of total OFEs" graph (figure 8.5) measures how "wrong" the forecasted values are at each step. Hence the value of the graph for point 23, means "now that we knew X_{2-3} (TRL 3 has just started), how wrong were the method's forecasts for the rest of the steps? ". "Wrong" is defined in the sense of the Objective Function of Error, i.e. the weighted sum of the errors (in years), by giving more weight to the next (close) steps than the further ones.

For example, in this case, we gave a weight of 1 for the forecast of the next transition, of 0.5 for the one after, then 0.1 and 0.05, and 0 for forecasts further in the future. We also decided to drop the first two forecasted columns because they did not have enough past data for a reliable forecast. And finally, we considered the extreme outlier a very rare case, and we took it out of the OFE to prevent it from having a disproportional weight in the final comparison results. The advantage of this approach is that all those parameters can be changed dynamically to test the best methods for different user objectives.

Finally, for each method, the subset that had the median grand total OFE score was taken out to be compared with the medians of other methods in the final graph of OFEs per step (see section 8.4.4 for the analysis of the results).

8.3.3 Forecasting methods

In this section, we present in greater detail the twelve forecasting methods that were used. We also present their summary performance measures (the best OFE, the median OFE, and the OFE standard deviation)

Mean

We simply took the mean of each subset for each transition and used it as a (fixed) forecast. This is equivalent to using the assumption at level 2 of our framework since we do not take past transitions into account (i.e. the forecasts do not change as the technology matures). We got a median OFE of 13.2 years, and a very small standard deviation of 0.14 years.

Median

It is the same as the previous method, but with medians instead of means. We got a median OFE of 12.91 years, and also a low standard deviation of 0.17

Influence diagram

For a set of correlated random variables, an influence diagram is defined by the vector of means m, the vector of conditional variances v, and a triangular matrix B containing the arc coefficients that indicate the influence a preceding variable has on the succeeding one(s) (Shachter and Kenley, 1989). Whenever a value is realized (known transitions), the means get updated and propagated through the influence diagram (through the arc coefficients) resulting in new means (and normal distributions), and thus new forecasts of the remaining variables. This process is equivalent to Bayesian updating in Bayesian nets. The Matlab algorithms that generate the B matrix (from the covariance matrix) and that update the means are included in Annex 3.

The method tended to have some good results (better than the median method) for the first three transitions. Then at transitions 56 and 67, those results either stayed reasonably good for most subsets, or completely "exploded" by diverging to OFE's of more than 8000 years in one of the subsets.

In summary, the method produced a median score of 14.1 and a best score of 9.72. However it had a very large standard deviation due to the explosive behavior of subset 1. This behavior came from wrong forecasts for only 1 or 2 technologies, and was facilitated by the fact that the errors were differences of exponentials.

In the other influence diagram methods, we tried to correct this behavior either by (1) cutting the data in two parts (in methods ID(frag4-3) and ID (frag5-2)) thus making the influence diagram "forget" the past data at TRL 5 or TRL 6, and looking back only as far as those steps when forecasting for further transitions, or by (2) simply bounding the full Influence diagram making sure it does not predict values higher or lower than some pre-fixed limits (the code is in Annex 3).

None of those techniques was better than the fixed estimates in terms of the median subset.

ID (frag4-3) had a median of 13.99 and reduced the "explosion" of subset 1 to a few hundred (instead of a few thousands).

ID (frag 5-2) had a median of 14.85 and did not reduce the "explosion".

ID (bounded) had the best forecast: it had a median of 13.86, the same best score as ID (full), and significantly reduced the explosion. The standard deviation, however, was still 3.03 (because of the subset 1 score that was still relatively high).

Note however that all those scores go down by more than 2.5 if we add the early forecasts (i.e the first two columns) to the OFE, and become better than the fixed forecasts for the median subset (and much better for the best subset).

Moving average

This method does not need a training set. We used it with a past span of two and included the parameter alpha in the "control parameters" worksheet; it got scores of around 13.7 for typical values of alpha. By using Excel's data table to vary the parameter, we were able to go down to 12.99 for the optimal alpha value (0.125). Theoretically, we should have used training sets to produce an optimized alpha value for each set, and then test the scores of those subsets (but such a procedure was not programmable in Excel, and it would not have improved the already bad OFE of 12.99 years)

Exponential smoothing

This method (Albright et al, 2006) also performed poorly since even the optimized alpha value (for the full dataset), gave a score of 14.00.

Exponential smoothing with trend (Holt-Winters method)

This method (Albright et al, 2006) also did not perform very well, the best results were obtained for high alpha and beta values (very reactive case) and achieved best scores of around 50, which indicated no need for further testing. It probably did not perform very well because we had to use the cumulative transition times (so that there is a trend to follow), which reduces the "visibility" of the variations.

Autoregression

This method consisted of regressing the variable that is being forecasted against all the already known transitions in the training data, then generating the forecast by applying the resulting linear function to the known transitions of the particular technology (Matlab code available in Annex 3).

For example, if we already know transitions 1-2, 2-3, and 3-4, and we want to forecast transition 7-8; we would use the training set to regress 7-8 against 1-2, 2-3, and 3-4, then we would get the forecast by multiplying the regression coefficients by the (known) values of 1-2, 2-3 and 3-4 transitions. Hence for each subset of the training data, we had to run 28 regressions, and apply them to all of the 18 technologies.

Note that there are very few data points to perform the regression (especially for training subset 4 that had very few post-step-6 points). We still performed the regression with very few degrees of freedom, because we were mainly interested in the result.

The regressions had good R^2 values, but often very high coefficient p-values (we still ran the test to see the results). Ideally, if enough data is available, a stepwise *partial autoregression* should be run (meaning the regression should be made only against the significant past known variables with low p-values).

As for the results, the method had a best score of 8.78, but a median score of 22.64 because two of the subsets had an “explosive” behavior at transition 7-8 (One possible explanation can be the collinearity introduced by the high correlation between the coefficients of the first 5 variables).

We addressed this issue by creating a *bounded full autoregression* method. The result was the best of all the forecasting techniques. It had the same best score as the unbounded full autoregression, and the median score went down to 11.24, with a relatively small standard deviation of 2.70.

Closest neighbor

This method consisted of generating the forecasts progressively by assuming an evolution similar to the technology in the subset that, so far, correlates the best with the technology being forecasted (code in Annex 3).

For example, if we already have transitions 1-2, 2-3, and 3-4 for technology t_i and we want to forecast 4-5, we would calculate the correlation between those 3 variables and the rest of technologies in the subset. If technology t_j was the one with highest correlation, the forecast for technology t_j would be $X_{4-5,i}=X_{3-4,i} * (X_{4-5,j}/X_{3-4,j})$ (i.e. same growth rate), or $X_{4-5,i}=X_{3-4,i} + (X_{4-5,j}-X_{3-4,j})$ (i.e. same absolute growth) if $X_{3-4,j}$ is equal to 0.

This method did not perform very well; it had a median of 21.53 and had a large OFE for transition 7-8 (where the different technologies oscillate the most).

8.3.4 Analysis of the results

Although the Excel file generates a rich set of graphs that allows to compare and evaluate the methods thoroughly and under different settings of the OFE, we will limit the comparison to the general results.

Figure 8.7 compares the total OFE of all the methods using the median subsets (for the OFE settings visible in figure 8.6). Each curve corresponds to one forecasting technique. The graph shows the total OFE (weighted average of forecast error over a few next steps, averaged across technologies, and with the median result taken with respect to the training datasets), at transitions 3-4 to 8-9.

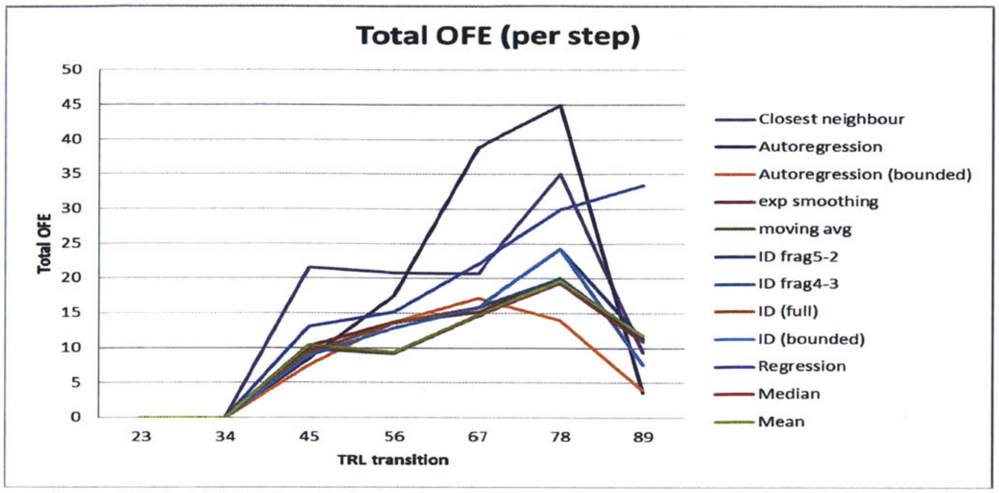


Figure 8:7 Total OFE for all the 12 forecasting methods (with the original OFE settings)

In figure 8.8, we removed the worst techniques ("closest neighbor", "autoregression", and "regression") for a better visibility, and included the 2-3 and 3-4 transition forecasts in the OFE.

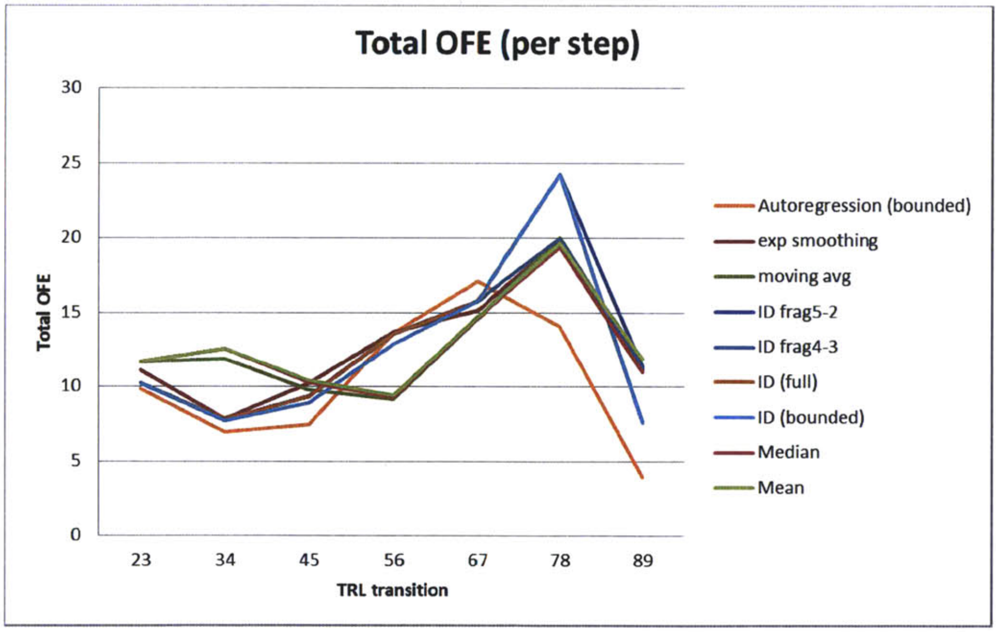


Figure 8:8 Total OFE for the 9 best forecasting methods (with the 2-3 transition OFE included)

The bounded Autoregression (orange curve) method seems to be the best overall forecasting method. As for the rest of the methods they perform better than fixed estimates in early stages and worse in the later stages.

This conclusion on the other methods holds if we add the "extreme" outlier to the comparison (figure 8.9), but not if we change the weighting function by giving equal importance to all future forecasts making far forecasts weigh as much as closer ones (figure 8.10 and figure 8.11). Meanwhile, the bounded Autoregression always has the smallest error for most transitions.

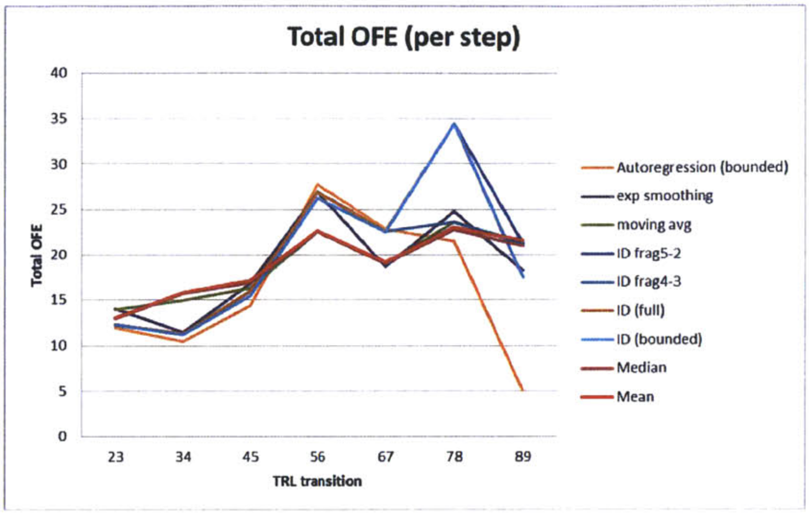
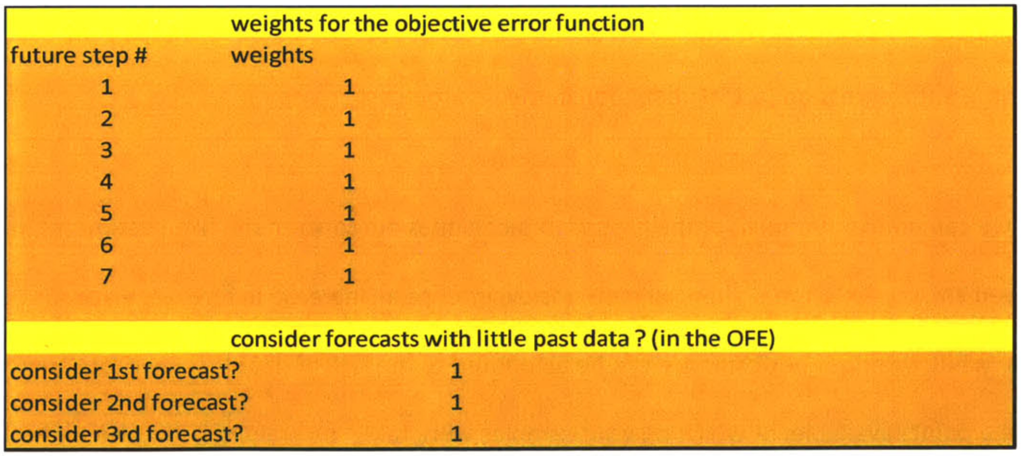


Figure 8:9 Total OFE for the 9 best forecasting methods (with the extreme outlier included in the validation dataset)



	weights for the objective error function
future step #	weights
1	1
2	1
3	1
4	1
5	1
6	1
7	1

	consider forecasts with little past data ? (in the OFE)
consider 1st forecast?	1
consider 2nd forecast?	1
consider 3rd forecast?	1

Figure 8:10 OFE control parameters for figure 8.11

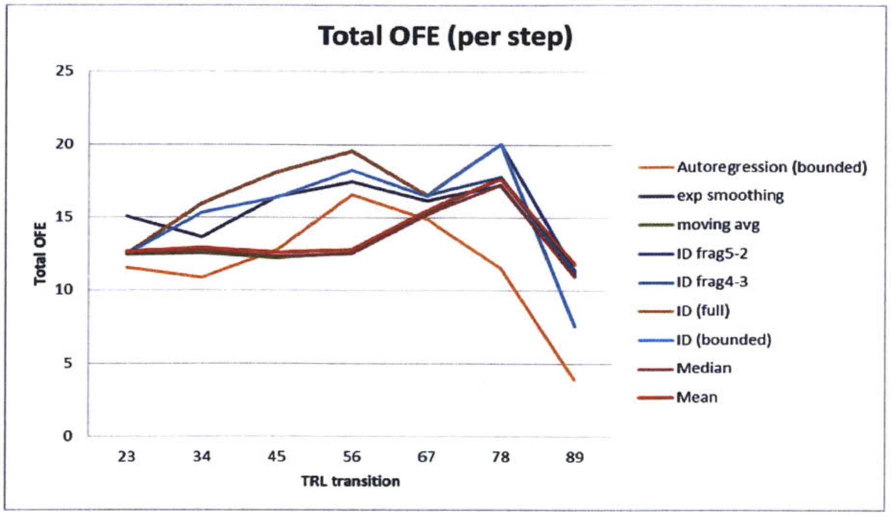


Figure 8:11 Total OFE for the 9 best forecasting methods (with equal weight given to close and far forecast errors)

Overall, the autoregression method can be considered the best method. Not only does it consistently outperform the other techniques for different OFE settings, its results also have a lower standard deviation.

We can now answer the research question: yes, we can improve the accuracy of forecasting models by using past data, in this case we went from an OFE of 13.2 for the best fixed estimate (the median) down to 11.2 for the autoregression (a 15% reduction in median forecast error).

However, we can notice that most of the proposed techniques outperform the fixed estimates in the well-behaved TRL1 to TRL5 area. Then, all methods experience an increase in forecast error for the last transitions. While this poor performance can be attributed to the lack of data for those late transitions (only 11 data points available, of which only subsamples were taken for training), it is also a result of the high oscillation in transition times seen at those stages, or simply the higher variance of X_{6-7} , X_{7-8}, and X_{8-9}.

The below standard deviations graph shows how the standard deviation increases after transition 4-5, making it harder to do forecasting. This shape of the variance also supports the GAO recommendations because getting to TRL 7 means getting past the 6-7 transition which is the transition with the highest standard deviation (i.e. highest uncertainty). Hence it would make sense to wait for that uncertainty to pass before committing agency money to the project.

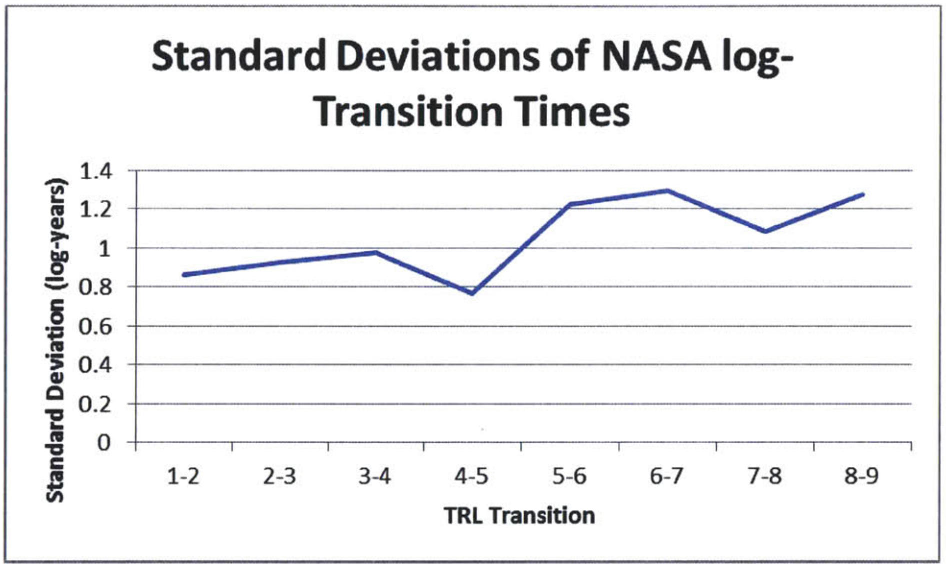


Figure 8:12 Standard deviations of the log-transition times of the NASA TRL data.

We finish this chapter with word of caution on the robustness of those methods. Although extra care was taken by resampling the dataset many times and looking only at median performances of the techniques, it is still possible that the methods have overlearned the dataset or got "lucky" for this specific sample. More data is needed to properly validate the approach to make sure that the improvement remains as the dataset gets larger. As a result, it might be better to use the robust median bootstrap technique of level 2 until level-3 forecasting techniques are validated for larger datasets.

Chapter 9

Chapter 9. Level-4 assumption

9.1 Definition and Literature review

So far, we have not used what is measured by the TRL numbers; we have just used the fact that TRLs are ordered placeholders, defining distinguishable intervals, and that those intervals can be related. The fourth level in our framework gives meaning to the TRL numbers by stating that those numbers are a measure of maturity and risk.

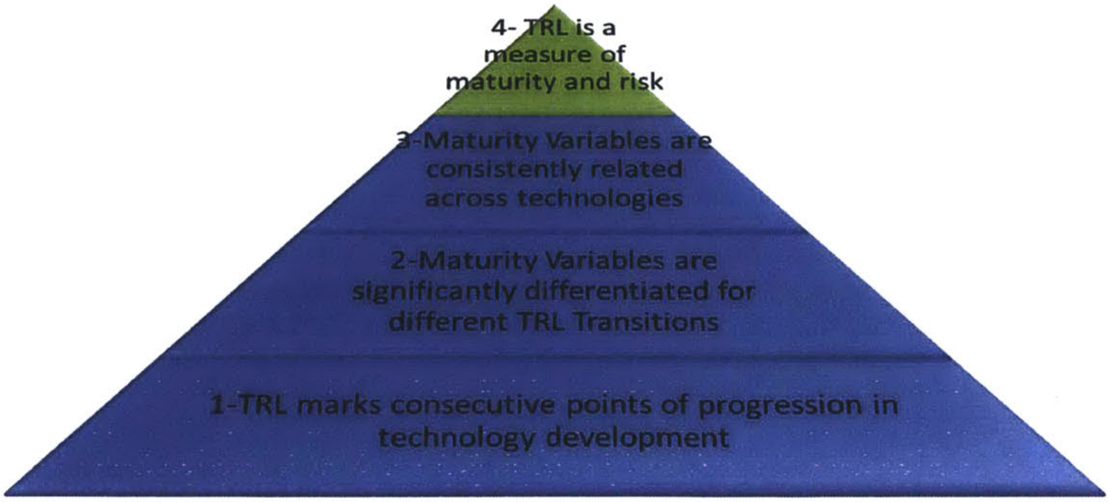

4- TRL is a measure of maturity and risk
3-Maturity Variables are consistently related across technologies
2-Maturity Variables are significantly differentiated for different TRL Transitions
1-TRL marks consecutive points of progression in technology development

Figure 9:1 The TRL framework, the fourth-level assumption

This assumption can be decomposed into two parts. The first part is that TRL measures maturity. This part alone is of little use to managers since they are interested in the operational dimensions of maturity (cost risk, schedule risk, performance risk). So the second part of the assumption extends the first part to claim that the TRL numbers are also measures of those operational dimensions of maturity

More specifically, risk analysts would like to use a direct relation between TRL and risk in the form of Risk = f(TRL). Many such models (for both cost and schedule) are available in the literature. Lee and Thomas (2001) regressed cost growth measures against a system TRL (average of the components weighted by their cost), while Dubos and Saleh (2008) regressed Relative Schedule Slippage against a truncated value of the cost-weighted TRL. We found similar regressions against other maturity scales. Kenley and Creque (1999) performed a regression of Time-To-Maturity against a maturity scale called Technology Maturity (TM), while Hoy and Hudak (1994) regressed cost uncertainty against a cost-weighted risk score.

In 2009, Boeing filed a patent on "systems, methods, and computer products for modeling a monetary measure for a good based upon technology maturity levels". The patent has a broad scope, as it concerns any association of a TRL level with a component cost uncertainty, that is used to create component cost distributions, which are then added to generate the good's (i.e. the system's) cost distribution (Mathews et al, 2009).

The below table summarizes the literature's regression models in greater detail.

Paper	**Lee & Thomas (2001)**	**Dubos & Saleh (2008)**	**Hoy & Hudak (1994)**	**Kenley & Creque (1999)**
Data Used	28 NASA programs (SAIC)	28 NASA programs (SAIC)	45 programs of DOD Selected Acquisition Reports. Of which, only 20 were used	20 Technology Maturity (TM) assessments for 14 stabilization technologies for nuclear materials.
Dependent variables	-Annual cost growth (ACG) -Annual relative cost growth (RCG) (**cost estimation**, and **cost uncertainty**)	-Relative Schedule Slippage: RSS= (total duration-initial estimation)/initial estimation -RSS' upper bound (**schedule slippage**, and **schedule slippage uncertainty**)	-Cost Uncertainty factor: Y (for both Development and Production, the Y_{mean}, Y_{min} and Y_{max} were calculated) (**cost estimation**, and **cost uncertainty**)	-Years until operational technology (Y) (**schedule estimation, schedule risk,** and **programmatic risk**)
Independent variables	-WTRL (TRL weighted by component costs) -Initial Cost Estimation (ICE)	-Floor(WRTL)	-Cost-weighted risk score X. Which is composed of 19 risk scales grouped into Hardware, Software, and Integration/Schedule Technical forms	-Technology maturity : TM (weighted average of 7 maturity parameters
Probability model, and justification	Johnson's four parameter families (bounded, unbounded, lognormal) at each TRL. Because of asymmetry and skewness.	Normal distribution. Because not enough points, it is just to illustrate the concept	The output Y (cost uncertainty factor) was modeled as a triangular random variable. In order to avoid negative cost uncertainties with a normal distribution.	Years until operational technology (Y) variable is assumed to be normal
Regression type and result	Linear ACG=5.9%*ICE (28 points, R^2=0.933) RCG=1.6%*WTRL (28 points, R^2=0.617)	Exponential RSS=8.29*exp(-0.56*WTRL). (6 data points only, R=0.94)	Linear $Y_{mean,development}$ =1+0.057*X (R^2=0.99) $Y_{mean,production}$=1+0.032*X (R^2=0.94) (Y_{mean} was considered heteroscedastic)	Linear Y= 0.3825*TM (R^2=0.8893)

Table 9.1 Summary of regressions against maturity scales found in the literature

While those models cannot be more valid than the assumptions they make, there is still room for improvement by correcting for the ordinality of the TRL scale, and by being more careful with the relationship between TRL and risk. Those other considerations will be looked at in section 9.4, which will allow us to answer the research question:

- *Out of the available methodologies, what is the best one to use in performing regression of cost or schedule against TRL?*

For now, we will focus on the assumptions themselves and look for the supporting evidence in sections 9.2 and 9.3.

9.2 Theoretical evidence

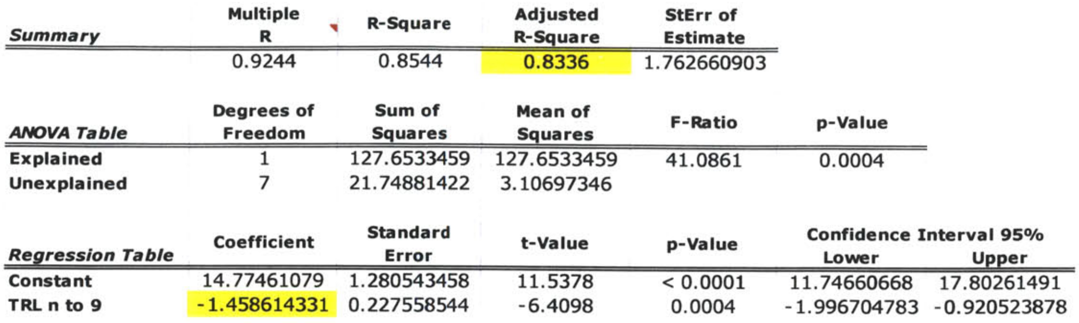
We know that the first part of the assumption is true: TRL is indeed a major measure of uncertainty, and it was developed to be so. However, it is obviously not a complete measure of maturity. We saw in section 3.1 that maturity is influenced by multiple factors, and table 3.3 shows clearly that TRL is a measure of only one of those factors. The question that naturally follows is "how much of maturity does TRL measure?" This question will be addressed in the following empirical evidence section through regression analysis.

As for the second part of the assumption, we already know from section 3.3.3.2 that TRL is a weak measure of maturity risk. Many measures like RD^3 and AD^2 were specifically developed to compensate for TRL's failure to measure risk. They are intended to be complements to TRL in that they measure what TRL misses: future maturation risk. As a result, this assumption does not have theoretical support. Conrow (2009) notes that TRL is only weakly correlated with risk. This assumption is partially true however, in the weak sense defined by the assumption for level 1. The risk decreases as TRL increases because the risk now is on a fewer number of remaining steps, and not because the TRL number itself is a measure of risk.

9.3 Empirical evidence

We saw that the level-4 assumption is only partially true. The aim of this section is to empirically quantify how much variability in maturity variables is captured by TRL. The literature already gives an idea with mostly very high R^2 values at around 90% or above, indicating that TRL surprisingly explains a very large percentage of the variation in maturity variables. However, Conrow (2011) contests those results by performing his own regression of Schedule Change on the data from Lee and Thomas (2001), and getting a R^2 of 0.26. Conrow (2009) also mentions that R^2 should be adjusted for the degrees of freedom; and he obtains adjusted R^2 values of 0.52 and 0.02 when repeating the calculations of Dubos and Saleh (2008).

We performed our own measurement by regressing the standard deviation of the Time-To-Maturity against current TRL. As expected we got an inverse relation (the higher the current TRL, the less risk in the time to maturity). The regression results in figure 9.2 show that TRL explained more than 83% of variation of Time-To-Maturity risk (adjusted R^2). The β coefficient is equal to -1.46, which means that on average, by maturing one extra TRL level, the schedule uncertainty (the standard deviation) is reduced by 1.46 years.



Summary	Multiple R	R-Square	Adjusted R-Square	StErr of Estimate
	0.9244	0.8544	0.8336	1.762660903

ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F-Ratio	p-Value
Explained	1	127.6533459	127.6533459	41.0861	0.0004
Unexplained	7	21.74881422	3.10697346		

Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Upper
Constant	14.77461079	1.280543458	11.5378	< 0.0001	11.74660668	17.80261491
TRL n to 9	-1.458614331	0.227558544	-6.4098	0.0004	-1.996704783	-0.920523878

Figure 9:2 Regression of Time-to-Maturity standard deviation against current TRL

In summary, TRL is a partial measure of maturity risk. It is important to take the definition of the regression's dependent variable into consideration (e.g. total cost vs. absolute cost growth vs. relative cost growth vs. probability of cost growth). We do not have sufficient data to confirm a high correlation in all those cases, but it appears there is already enough evidence in the literature to confirm a useful correlation between the level of TRL and the maturity variable risk.

9.4 Some considerations on regressing risk against TRL

Although the above evidence provides partial justification for TRL-based regressions of maturity variables, there are two important considerations that most of those models ignore. Those two problems (already mentioned in section 3.3.3.2) are the fact that TRL only measures the probability of occurrence term of risk, and the fact that TRL is an ordinal scale.

First, Conrow (2003, 2009, 2011) points out that TRL risk is decomposed into a probability of occurrence term and a consequence of occurrence term. While we have already explained how TRL can be a partial measure of the probability of occurrence, it provides little information on the consequence of occurrence. If a technology has a low TRL when product development begins, we can say that there are high chances of cost overruns, but it is harder to estimate by how much the overruns will be. So if we look at cost risk for example as our relevant maturity parameter, then there are many cost-related variables we could look at. "Probability of cost growth" would be a variable that can be modeled well by using TRL (because it is a probability of occurrence term). On the other hand "Absolute cost growth" is a variable that would be poorly modeled by TRL (because it is a consequence of occurrence term). "Relative cost growth" might have a chance of being well-modeled by TRL since it is a normalization of the consequence of occurrence.

The second problem is also pointed out by Conrow (2003, 2009, 2011): it is the ordinality of the TRL scale. The numbers in the TRL scale have no particular meaning, they are simple placeholders that just indicate the order, and could well be replaced by letters A, B, C, D, etc. with no loss of information. Mathematical operations (sums, averages, differences, multiplications) on TRL scores are not defined; they have no mathematical meaning, and can lead to wrong results. Such operations are nevertheless found in the literature.

Dubos and Saleh (2008), and Lee and Thomas (2001) both perform TRL averaging when creating the cost-weighted TRL (WTRL). First, averaging ordinal numbers does not have a mathematical meaning. Second, if the scale were cardinal, there would be no reason to weigh the average with component costs (cost does not necessarily represent the component's criticality in the system). And third, if the scale were cardinal, there would be no reason for averaging to be the appropriate operation to aggregate TRLs in a system; the minimum might be a better operation, and more generally, the Work Breakdown Structure (WBS) should be taken into account to detect parallelisms and critical paths.

This issue of operating on ordinal values can also be found in other TRL-based maturity measures. For example, we presented in 3.2.4 the Integrated Technology Index (ITI):

$$ITI = \frac{\sum_{subsystem\ technologies}(\Delta TRL \times R\&D^3 \times TNV)}{Total\ \#\ of\ subsystem\ technoloogies}$$

The ITI disregards the ordinality of TRL twice: first when performing differences of TRLs, and second when performing a weighted average of those numbers.

Furthermore, regressions on the TRL scale are over-constrained by the fact that TRL is treated as cardinal. A linear regression for example would be looking for a constant increment of the maturity variable between any two consecutive TRLs. This constraint has no reason to be true.

We are not implying that all results involving operations on TRL are necessarily wrong and misleading. The application might be small enough or defined on a small range of the TRL scale that the operation would not lead to major mistakes. However, such uses if they were to exist, should be very cautious in defining the range of TRL values, and then testing them extensively to show that the range of results could carry useful information for the user.

As a solution to this problem, Conrow (2009) proposes a calibration of the TRL scale in an attempt to give a maturity meaning to the TRL numbers. Conrow asked experts to compare the maturity of the 36 pairs of TRL values, and then aggregated and normalized the results using an Analytic Hierarchical Process (AHP). He then scaled the results so that the calibrated TRL9 had a value of 9, and then fitted the results to a 3rd degree polynomial. The final result was:

$$TRL_{adjusted} = 0.346 + 0.012 * TRL^3$$

While this calibration provides a simple technique to get a better idea of the real maturity distances between the TRL numbers and for averages to have more meaning, it still suffers a couple of drawbacks.

First, AHP is based on human assessments that are often imprecise and subjective, especially when using not-so-well defined words such as "maturity". Second, there is a conceptual problem in calibrating the scale with respect to "maturity". What does it mean for an expert to answer "TRL 9 is 2.5 times more mature than TRL 6"? If we want to make an ordinal scale cardinal, then the calibration has to be done with respect to an underlying cardinal space. We have already pointed out that the maturity concept is a multidimensional concept, and we cannot say that A is more mature than B by using only one number. For the comparison to make sense it should pick one of the operational dimensions of maturity. For example "how much more chances of meeting schedule does TRL 9 have over TRL 6?" or "what is the ratio of cost overrun risk between TRL5 and TRL4?" While Conrow's intention was to have one TRL scale calibration for all risk models involving TRL, we recommend calibrating TRL separately for each of the operational dimensions of maturity.

Nevertheless, the calibration still captures the general idea that the distance between maturity variables becomes larger at higher TRLs. When we repeated our earlier regression using Conrow's calibration, the regression quality improved as the adjusted R^2 went from 83% up to 92%.

One alternative approach to avoid the problem of ordinality in regressions is to treat the TRLs as categorical variables. Instead of embedding in advance a certain distance between the TRL levels in the regression, we can simply consider each TRL a different category and let the regression compare them by pairs.

In summary, while the problem of probability vs. consequence of occurrence and the problem of ordinality might be addressed by using specific methodologies, we still cannot solve the issue of TRL not being a measure of maturity, or it not being a measure of risk (this would necessitate a multiple regression against several maturity scales, which is not in the scope of this thesis). To answer our earlier question on the best methodology in performing regression against TRL, this methodology should:

-Avoid using any kind of averaging of TRLs. A WBS-based approach or SRLs (refer to section 3.2.4) can be used instead, if system TRLs need to be computed.

-Have a probability-of-occurrence-related dependent variable. The choice of variable is also important in that it has to be one that is well explained by the regression (in terms of adjusted R^2).

-Use a calibrated TRL scale (preferably calibrated to a relevant dimension of maturity).

Chapter 10

Chapter 10. Conclusions and future research

10.1 Summary of results

In this thesis, we proposed a 4-level taxonomy of TRL-based models. When we were performing our research on TRL cost and schedule models, we discovered that different models departed from different assumptions, and that the only way of consistently evaluating and comparing those models was by grouping them according to the increasing assumptions that they make. This natural grouping of TRL models not only helped us group and compare existing TRL models, but the fact that the assumptions were now clearly stated allowed a better understanding of the models. It also allowed us to make recommendations and propose alternative methodologies that make better use of the assumptions. Each level was analyzed theoretically (mainly by using properties of the TRL scale), and empirically (mainly by using the NASA dataset). The NASA dataset was also used in introducing and evaluating the new methodologies at levels 2 and 3, as a well as in making recommendations at levels 1 and 4. Data scarcity was a major concern throughout the thesis, and we often had to use specific methodologies and propose special recommendations to deal with this fact.

The framework, although theoretical, allowed us to answer four practical research questions, one at each assumption level:

At level 1, we found that the model strongly supported GAO maturation risk recommendations, both theoretically and empirically.

At level 2, the data confirmed that TRL transition times were statistically differentiated enough for the study of transition time distributions to make sense. Furthermore, we proposed the bootstrap method as the most suitable technique to make estimations on those variables with small datasets.

At level 3, we found that there is significant positive correlation between early TRL transition times. Furthermore, we were able to use this correlation to propose a method (the autoregression method) that improved forecast accuracy by 15% over level-2 fixed estimate models.

At level 4, we found that although the assumption was only partially supported, regressions on TRL still explained a significant percentage of variability in maturity variable risk. However, we proposed some modifications to the methodologies found in the literature so that they avoid performing mathematical operations on TRLs, that they correct for TRL's ordinality, and that they limit the model to probability-of-occurrence terms only.

All those results allow us to say that TRL can indeed be used beyond simply the exit criteria of a Science and Technology (S&T) program.

Table 10.1 summarizes the results of the 4-level model:

Assumption level	Theoretical support	Empirical support	Practical usefulness	Potential for improvement and better use of assumptions
Level 1	Very high	Very high	Medium	Low
Level 2	High	High	High	Medium
Level 3	Medium	High (for TRL 1 to 5)	High	High
Level 4	Low	Medium	Very High	High

Table 10.1 Summary of the results for the support of the assumptions, and usefulness of the models

Ideally a method needs to have both high theoretical and empirical supporting evidence. As expected, we note that there is an inverse relationship between the strength of the supporting evidence, and the usefulness/potential usefulness of the models

All the models still have potential for improvement (although with varying degrees):

- Level 1: By improving the measurement of variance reduction.
- Level 2: By using better bias reduction bootstrapping algorithms, and by using regression analysis to identify relevant factors in predicting maturity variables.
- Level 3: By developing more complex, robust methodologies for larger datasets, and by better isolating factors other than TRL that affect technology transition.
- Level 4: By developing more models, integrating measures of advancement degree of difficulty, calibrating the scales, and reducing the models to probability-of-occurrence variables.

10.2 Future and related research

The presented models are in no case a silver bullet in solving the problems in DoD's acquisition process. A lot of those problems can be mainly attributed to managerial issues, poor risk management practices, external changes in funding or requirements, or to wrong incentive structures (for example, the contracting structure). This research main contribution is to help future researchers understand and develop more realistic cost and schedule models to improve the risk analysis part of risk management.

There are many directions of research that directly follow the work presented so far:

First, although special care was taken in choosing the methods and in making conclusions and recommendations, the empirical part of this thesis is still based on a very small dataset. The results need to be confirmed as soon as more data becomes available.

Second, although there are reasons to believe that the four assumptions can be extended to other types of technologies and other agencies, let us note that our empirical evidence only supports the assumptions for NASA technologies for now. The results in Crepin, El-Khoury, and Kenley (2012) are in support of a generalization of results to the Department of Energy's TRL, but they are also based on a small dataset and only on a few of the 8 TRL transitions. More statistical comparison work can be done as other agencies adopting TRL start releasing data.

Third, a very important extension to this research is to augment the models by looking at other complementary measures of maturity. As we mentioned in section 9.4, we can try to capture a larger proportion of variability in maturity variables by performing multiple regression analysis. After all, we know that TRL alone only partially captures the risk of transition variables.

Fourth, there is still work to be done to improve level-4 regression models. For a start, we can find what exact versions of cost and schedule variables best correlate with TRL. Also, we can develop and test calibrations of the TRL scale that are relevant on certain dimensions of maturity.

Fifth, we proposed one way of decomposing maturity using the Enterprise Architecting framework in table 3.1. This might not be an extensive decomposition, or the most relevant way of looking at maturity dimensions. There is still conceptual work to be done on defining the relevant aspects of maturity, then empirically testing their representativeness (for example, through multiple regression analysis). Furthermore, table 3.2 shows that some of those dimensions of maturity are still not measured by any specific scale. Hence more work needs to be done to make sure all the relevant maturity dimensions can be measured with scales tailored to those dimensions.

Sixth, there are other completely different modeling approaches that can be used. For example El-Khoury and Kenley (2012) propose a dynamic programming methodology to integrate decision-making into Cost-Schedule bivariate modeling. One paper proposes a system dynamics approach to model the acquisition system (Dixon, 2007), another paper proposes an epoch–based framework to model program development time and program risk (Szajnfarber, 2011). Yet another paper applies real options to technology portfolios (Shishko et al, 2004). While TRL-based models are the most commonly used in risk management, those other models can be useful for managerial applications, even if they are not as quantitative.

Seventh, there are other approaches that address DoD's acquisition problems more directly as opposed to our approach, which was through the improvement of risk management. Dorey (2011) proposes a risk-driven contracting structure that tries to address the incentive structure in government contracting especially when the government is the party assuming the high cost risks. Other recommendations address the internal political and organizational incentives inside DoD that lead to an early start of production for immature projects. Nolte (2008) notes that Acquisition programs have much more money than S&T programs, which forces much of the technology maturation process to take place within an acquisition program rather than within the S&T area. Furthermore, DoD often has to proceed in early immature product development just to get needed funds and management support, simply to keep the program going.

Finally, a new line of recommendations is to allow performance to be variable (Gansler, 2012), in technically acceptable terms. This would add new feasible combinations in the CPS trade space, which could lead to dramatic improvements in cost and schedule.

References

References

[1]- Albright, S., Winston, L., & Zappe, C. (2006). *Data Analysis and Decision Making with Microsoft Excel- 3rd Edition.* South-Western Publication Co.

[2]- Alexander, A. (1981). The linkage between technology, doctrine, and weapons innovation: experimentation for use. *Santa Monica, CA: RAND Corporation.*

[3]- Azizian, N. (Oct 2009). A Review and Analysis of Maturity Assessment Approaches for Improved Defense Acquisition. *NDIA 12th Annual.* San Diego, CA: Systems Engineering Conference.

[4]- Bilbro, J. (Feb 2007). Mitigating the Adverse Impact of Technology Maturity. *Project Management Challenge 2007.* Fourth Annual NASA Project Management.

[5]- Bilbro, J. (Sep 2008). Using the Advancement Degree of Difficulty (AD2) as an Input to Risk Management. *Multi-Dimensional Assessment of Technology Maturity- Technology.* Virginia Beach, VA.

[6]- Broadus, E. (25 Oct 2006). *Update on the Process for Evaluating Logistics Readiness Levels (LRLs).* Presentation to Booz Allen Hamilton Inc.

[7]- Browning, T., & Eppinger, S. D. (Nov 2002). Modeling Impacts of Process Architecture on Cost and Schedule Risk in Product Development. IEEE Transactions On Engineering Management, Vol. 49, No. 4.

[8]- Clay, R., Marburger, S., Shneider, M., & Trucano, T. (2007). *Modeling and Simulation Technology Readiness Levels.* Sandia National Laboratories.

[9]- Conover, W. (1999). *Practical Nonparametric Statistics.* Wiley.

[10]- Conrow, E. (1995). Some Long-Term Issues and Impediments Affecting Systems Acquisition Reform. *Acquisition Research Symposium.* 1995 Acquisition Research Symposium Proceedings and Defense Systems Mana.

[11]- Conrow, E. (2003). Effective Risk Management: Some Keys to Success- 2nd Edition. *American Institute of Aeronautics and Astronautics.*

[12]- Conrow, E. (2009). Estimating Technology Readiness Level Coefficients. *AIAA Paper No. 6727.*

[13]- Conrow, E. (Jun 2007). *Have Performance Requirements Historically Been Met in Systems Developed for the U. S. Military?* Invited white paper response from the 1997 SCEA Acquisition Reform Model Sharing Workshop.

[14]- Conrow, E. (Nov-Dec 2011). Technology Readiness Levels and Space Program Schedule Change. *Jornal of Spacecraft and Rockets Vol. 48 No. 6.*

[15]- Cornford, S., & L. , S. (2004). Quantitative Methods For Maturing and Infusing Advanced Spacecraft Technology. (pp. pp. 663-681). IEEE Aerospace Conference Proceedings.

[16]- Crepin, M., El-Khoury, B., & Kenley, C. (Jul 2012). It's All Rocket Science: On the Equivalence of Development Timelines for Aerospace and Nuclear Technologies. *Proceedings of the 22nd INCOSE Symposium.* Rome, Italy.

[17]- De Angelis, D., & Young, A. (Apr 1992). Smoothing the Bootstrap. *International Statistical Review/ Revue Internationale de Statistique*, pp. Vol.60, No. 1, pp 45-56.

[18]- Department of Energy. (Oct 2009). *Technology Readiness Assessment Guide.* DOE G 413.3-4.

[19]- Dews, E., Smith , G., Barbour, A., Harris, E., & Hesse, M. (Oct 1979). Acquisition Policy Effectiveness: Department of Defense Experience in the 1970s. *The Rand Corporation, R2516-DR&E.*

[20]- DiCiccio, T., & Efron, B. (1991). *Comparison of procedures for constructing confidence limits in exponential families. Technical Report.* Stanford University: Department of Statistics.

[21]- Dion-Schwarz, C. (Feb 2008). How the Department of Defense Uses Technology Readiness Levels. Office of the Director, Defense Research and Engineering.

[22]- Dixon, R. (Sep 2007). The Case for aKnowledge Based DoD Software Enterprise: An Exploratory Study Using System Dynamics. Naval Postgraduate School.

[23]- DoD. (2009). Technology Readiness Assessment (TRA) Deskbook. *Prepared By The Director, Research Directorate (DRD).* Office Of The Director, Defense Research And Engineering (DDR&E).

[24]- DoD. (2011). *F-35 Joint Strike Fighter Concurrency Quick Look Review.* DoD.

[25]- DoD. (Jan 2012). *Defense Acquisition Guidebook.* Retrieved from http://at.dod.mil/docs/DefenseAcquisitionGuidebook.pdf

[26]- DoD. (Nov 2007). *The Defense Acquisition System.* DoD Directive 5000.

[27]- Dorey, P. (Mar 2011). Enhancing Cost Realism Through Risk-Driven Contracting: Designing Incentive Fees Based On Probabilistic Cost Estimates. Air Forces Fellows.

[28]- Dubos, G., & Saleh, J. (2010). Spacecraft technology portfolio: Probabilistic modeling and implications for responsiveness and schedule slippage. *Acta Astronautica Volume 68(7–8),* 1126–1146.

[29]- Dubos, G., & Saleh, J. (Apr-May 2011). Spacecraft technology portfolio: Probabilistic modeling and implications for responsiveness and schedule slippage. *Acta Astronautica,* Vol. 68, Iss. 7–8, Pg. 1126–1146.

[30]- Dubos, G., Saleh, J., & Braun, R. (July–Aug 2008). Technology Readiness Level, Schedule Risk, and Slippage in Spacecraft Design. *Journal of Spacecraft and Rockets,,* Vol. 45, No. 4, , pp. 836–842. doi:10.2514/1.34947.

[31]- Efron, B., & Tibshirani, R. (1993). *An Introduction to the Bootstrap.* Boca Raton, FL: Chapman & Hall/CRC.

[32]- El-Khoury, B., & Kenley, C. (May 2012). An Analysis of TRL-Based Cost and Schedule Models. Monterey, California: 9th Annual Acquisition Research Symposium.

[33]- Fernandez, J. (2010). *Contextual Role of TRLs and MRLs in Technology Management-SAND2010-7595.* ST&E Integration Department: National Security Studies and Integration Center.

[34]- Fiorino D., T. (Jun 2003). Engineering Manufacturing Readiness Levels (EMRLs): Fundamentals, Criteria and Metrics. *Presentation to OSD/ATL Program Manager's Workshop.*

[35]- Foden, J., & Berends, H. (2010). Technology Management at Rolls-Royce. *Industrial Research Institute.*

[36]- Gansler, J. (May 2012). *Defense Affordability.* Monterey, CA: NPS 9th Annual Acquisition Research Symposium.

[37]- GAO. (1999). *Best Practices: Better Management of Technology Development Can Improve Weapon Outcomes.* NSIAD-99-162: GAO.

[38]- GAO. (2006). *Best Practices: Stronger Practices Needed to Improve DoD Technology Transition Process.* -06-883: GAO.

[39]- GAO. (2008). *Defense Acquisition: Assessments of Selected Weapon Programs.* -80-476SP: GAO.

[40]- GAO. (2009, April 30). *Defense Acquisitions: Charting a Course for Lasting Reform.* Retrieved from GAO-09-663T: http://www.gao.gov/new.items/d09663t.pdf.

[41]- GAO. (Apr 2010). *Best Practices: DoD can achieve better outcomes by standardizing the way manufacturing risks are managed.* GAO-10-439.

[42]- Garvey, R. (2000). *Probability Methods for Cost Uncertainty Analysis : A Systems Engineering Perspective.* Bedford, Massachusetts, USA : CRC Press.

[43]- Graben, J. (2009, June 15). *USM/BIAC.* Retrieved from Business & Innovation Assistance Center : http://www.usm.edu/biac/

[44]- Graettinger, C., Garcia, S., & Ferguson, J. (2003). *TRL Corollaries for Practice-Based Technologies.* Carnegie Mellon University, Software Engineering Institute.

[45]- Hobson, B. (2006). A Technology Maturity Measurement System for the Department of National Defence The TML System. *DRDC Atlantic CR 2005-279.*

[46]- Holt, L. (2007). *A Tool For Technology Transfer Evaluation: Technology Transfer Readiness Levels (TTRLS).* The Netherlands:), 58th International Astronautical Congress 2007, 18th Symposium on Space Activity and Society (E5.),Innovating Through Technology Spin-in and Spin-off (1.).

[47]- Hoy, K., & Hudak, D. (1994). Advances in Quantifying Schedule/Technical Risk. *The 28th DoD Cost Analysis Symposium.* Leesburg, VA: The Analytical Sciences Corporation,Xerox Document University.

[48]- Hubbard, D. (2007). *How to Measure Anything (2007).* Hoboken, New Jersey: Wiley.

[49]- Kenley, R., & Creque, T. (1999). *Predicting Technology Operational Availability Using Technical Maturity Assessment.* INCOSE Annual Meeting.

[50]- Lee, T., & Thomas, L. (2003). Cost Growth Models for NASA's Programs. *Journal of Probability and Statistical Science*, Vol. 1, No. 2, pp. 265–279.

[51]- LeGresley, P., Bathke, T., Carrion, A., Cornejo, J., Owens, J., Vartanian, R., et al. (2000). *1998/1999 AIAA Foundation Graduate Team Aircraft Design Competition: Super STOL Carrier On-board Delivery Aircraft.* the Society of Automotive Engineers, Inc and the American Institute of Aeronautics and .

[52]- Mahafza, S., Componation, P., & Tippett, D. (Dec 2004-Mar 2005). A performance-based technology assessment methodology to support DoD acquisition. *Defense Acquisition Review Journal*, Volume 11, Number 3.

[53]- Mankins, J. (1995). *Technology Readiness Levels.* NASA Office Of Space Access And Technology White Paper.

[54]- Mankins, J. (2002). *Approaches to Strategic Research and Technology (R&T) Analysis and Road Mapping.* ACTA Astraunotica, Vol. 51, No. 1-9, Pp.3-21 , 2002.

[55]- Mankins, J. (Mar 1998). *Research & Development degree of difficulty (R&D3).* Advanced Projects Office, Office of Space Flight, NASA Headquarters.

[56]- Mathews, H., Datar, T., Feely, K., & Gauss, J. (2009). *Patent No. 7,627,494 B2,.* Washington DC, U.S.A.

[57]- Ministry of Defence (MoD). (Mar 2006). Acquisition Management System, ., (p. release (v10.1)). United Kingdom.

[58]- Minning, C., Moynihan, P., & Stocky, J. (2003). Technology Readiness Levels for the New Millennium Program. *Aerospace Conference 03.* Proceedings IEEE, Vol. 1, pp. 417-426.

[59]- Mooney, C., & Duval, D. (1993). *Bootstrapping: A nonparametric Approach to statistical inference, quantitative applications in the social sciences.* Newbury Park, CA: Sage Inc.

[60]- Moore, G. A. (1991). *Crossing the Chasm: Marketing and Selling High-Tech Products to Mainstream Customers.* New York: Harper Business Essentials.

[61]- Morgan, J. (2008). *Manufacturing Readiness Levels (MRLs) and Manufacturing Readiness Assesments (MRAs).* (AFRL/RXMT, Performer) Wright PAtterson, AFB. OH.

[62]- Nightingale, D. (Jun 2009). Principles of Enterprise Systems. *Second International Symposium on Engineering Systems.* Cambridge, MA.

[63]- Nightingale, D., & Rhodes, D. (Mar 2004). Enterprise Systems Architecting: Emerging Art and Science within Engineering Systems. *MIT Engineering Systems Symposium.* Cambridge, MA.

[64]- Nolte, L. (2008). *Did I ever tell you about the whale? Or measuring technology maturity.* Information Age Publishing, Inc.

[65]- Nolte, W. (2005, April 28). Technology Readiness Level Calculator. *Technology Readiness and Development Seminar.* The Aerospace Corporation: Space System Engineering and Acquisition Excellence Forum.

[66]- North Atlantic Treaty Organisation. (Sep 2010). *Development of an Assessment Methodology for Demonstrating Usability, Technical Maturity, and Operational Benefits of Advanced Medical Technology.* RTO TECHNICAL REPORT, TR-HFM-13.

[67]- Olive, D. (2005). *A simple confidence interval for the Median", , 2005.* Southern Illinois university.

[68]- OSD Manufacturing Technology Program. (Jul 2011). Manufacturing Readiness Level (MRL) Deskbook., (p. Version 2.01).

[69]- Peisen, D., & Schulz, L. (1999). *Time Required To Matureaeronautic Technologies To Operational Readiness.* Task Order 221 Case Studies.

[70]- Perry, R., Smith, G., Harman, A., & Henrichsen, S. (Jun 1971). System Acquisition Strategies. *The Rand Corporation, R-733-PR/ARPA.*

[71]- Phillips L, E. (2010). The Development and Initial Evaluation of the Human Readiness Level Framework. *Naval Postgraduate School.*

[72]- Sadin, S., Povinelli, F., & Rosen, R. (1989). The NASA Technology push towards Future Space Mission Systems. *Acta Astronautica*, 20:73-77.

[73]- Sadin, S., Povinelli, Frederick, P., & Rosen, R. (1989). NASA technology push towards future space mission systems. (pp. pp 73-77, V20). Bangalore, India: Selected Proceedings of the 39th International Astronautical Federation Congress, Acta Astronautica.

[74]- Salah-Esa, A., & Dobbyn, T. (2011). *Price of F35 fighter soars.* Reuters.

[75]- Salter, J. (2006). Having Confidence in Non-Parametric Data. *Oxford Pharmaceutical Sciences Ltd.* Oxford, UK.

[76]- Sauser, B., Emmanuel Ramirez-Marquez, J., Magnaye, R., & Tan, W. (2008). A Systems Approach to Expanding the Technology Readiness Level within Defense Acquisition , . *nternational Journal of Defense Acquisition Management*, Vol. 1,pp. 39.

[77]- Sauser, B., Gove, R., Forbes, E., & Emmanuel Ramires-Marquez, J. (Jan 1, 2010). Integration maturity metrics: Development of an integration readiness level, –. *IOS Press* (p. Volume 9 (1)). Information Knowledge Systems Management.

[78]- Shachter, R., & Kenley, C. (1989). Gaussian Influence Diagrams. *Management Science 35(5)*, 527-550.

[79]- Sharif, A., Junfang, Y., & Stracener , J. (2012). *The U.S. Department of Defense Technology Transition: A Critical Assessment.* INCOSE.

[80]- Shishko, R., Ebbeler , D., & Fox, G. (2004). NASA technology assessment using real options evaluation. (pp. pp 1-13). Systems Engineering Volume 7, Issue 1.

[81]- Smith, J. (2005). An Alternative to Technology Readiness Levels for Non-Developmental Item (NDI) Software. *Proceedings of the 38th Hawaii International Conference on System Sciences.* Hawaii.

[82]- Smith, P., & Merritt , G. (Jun 2002). Proactive Risk Management: Controlling Uncertainty in Product Development. Productivity Press.

[83]- Smoker , E., & Smith, S. (2007). System Cost Growth Associated with Technology-Readiness Level. *Journal of Parametrics*, Volume 26, Issue 1.

[84]- Systems and Software Engineering/Enterprise Development Directorate. (Aug 2006). *Risk Management Guide for DoD Acquisition, 6th ed., version 1.0, OUSD(AT&L)*. Retrieved from http://www.dau.mil/pubs/gdbks/docs/ RMG%206Ed%20Aug06.pdf.

[85]- Szajnfarber, Z. (n.d.). ff.

[86]- Tan, W., Emmanuel Ramirez-Marquez, J., & Sauser, B. (Aug 2009). A Probabilistic Approach to System Maturity Assessment,. *Stevens Institute of Technology*.

[87]- TAO, L., Probert, D., & Phaal, R. (April 2008). Developing the concept of 'Innovation Readiness Levels'. *The 17th IAMOT.* Dubai, UAE: International Association for Management of Technology.

[88]- TEC-SHS. (Sept 2008). *Technology Readiness Levels Handbook for Space Applications.* issue 1 revision 6 – March 2009TEC-SHS/5551/MG/ap.

[89]- The undersecretary of defense. (2011). *Memorandum for component acquisition executives.* Washington. DC.

[90]- U.S. Army Medical Research and Materiel Command. (June 2003). Biomedical Technology Readiness Levels (TRLs). *Prepared for the Commander, Science Applications International Corporation*, under Contract number DAMD17-98-D-0022.

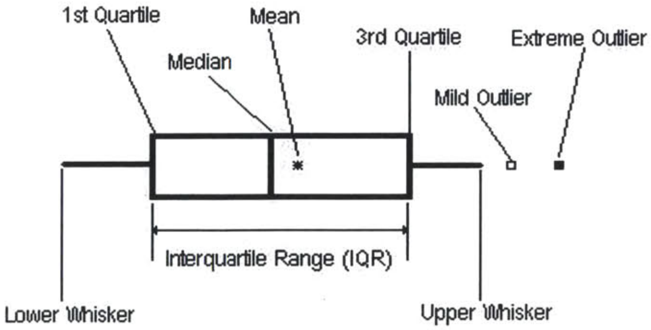
[91]- Valerdi, R., & Kohl, R. (Mar 2004). An Approach to Technology Risk Management. *Engineering Systems Division Symposium.* Cambridge, MA.

[92]- Weapon Systems Acquisition Reform Act. (2009). *Puplic Law 111–23.*

Annex 1

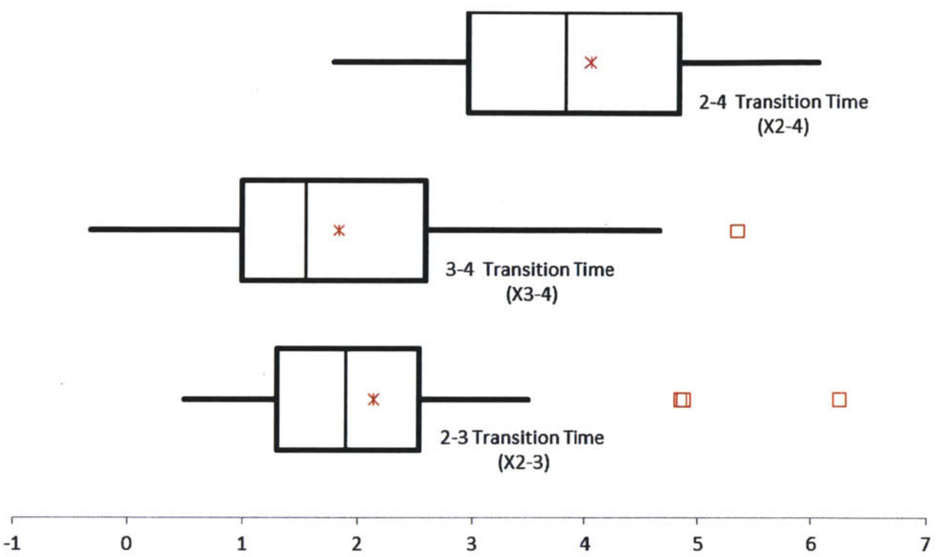
Annex 1: Box-Whisker plots of Army TRL transition times

This Annex presents Box-Whisker plots in support of the ANOVA analysis for the level-2 assumption. For each triplet of transition times $X_{i/i+1}$, $X_{i+1/-i+2}$ and $X_{i/-i+2}$, the plots show that $X_{i/i+1}$, $X_{i+1/-i+2}$ are differentiated from $X_{i/-i+2}$ (the top plot) in a statistically significant manner . In other terms, the means are far from each other, while the variance is small enough so that the two samples can be considered taken from different populations at a 95% significance level.

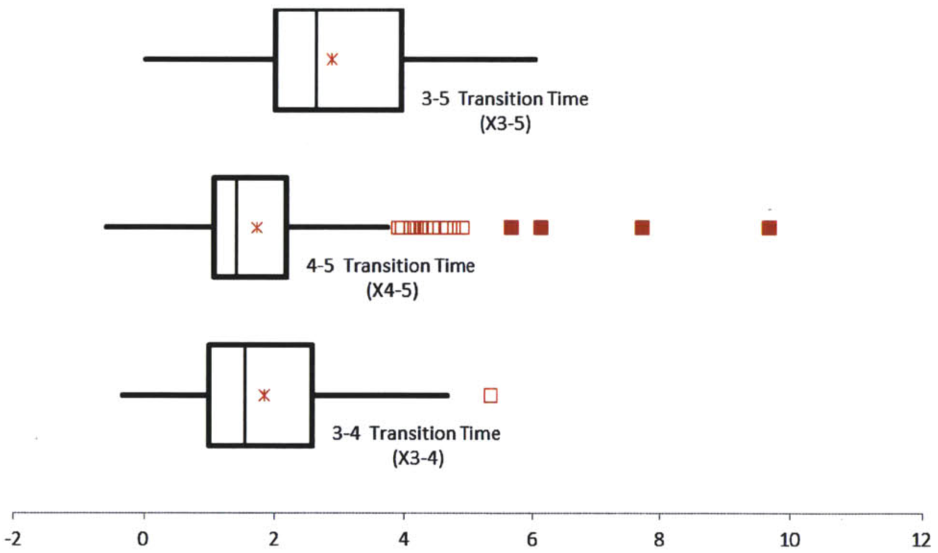


Whiskers extend to the furthest observations that are no more than 1.5 IQR from the edges of the box. Mild outliers are observations between 1.5 IQR and 3 IQR from the edges of the box. Extreme outliers are greater than 3 IQR from the edges of the box.

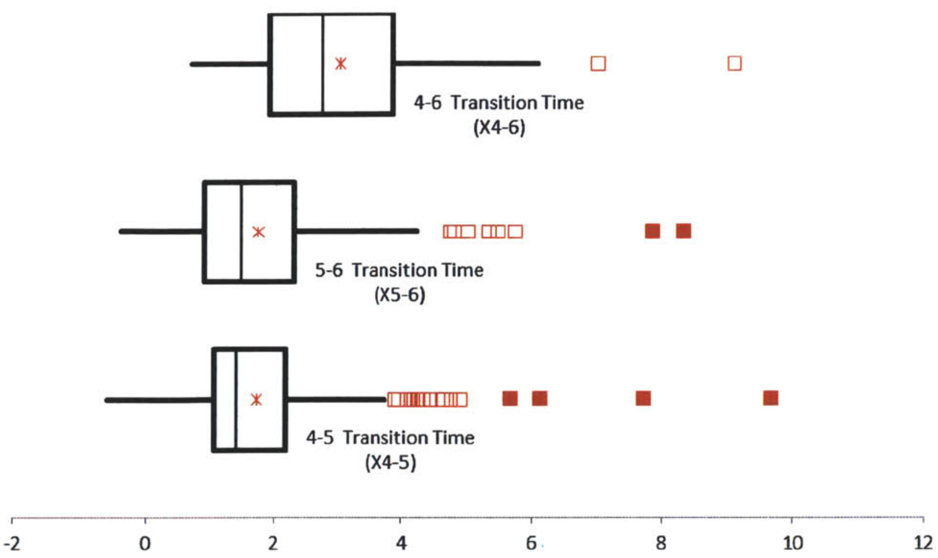
Box-Whisker Plot Comparison of Transition Times 2 to 4



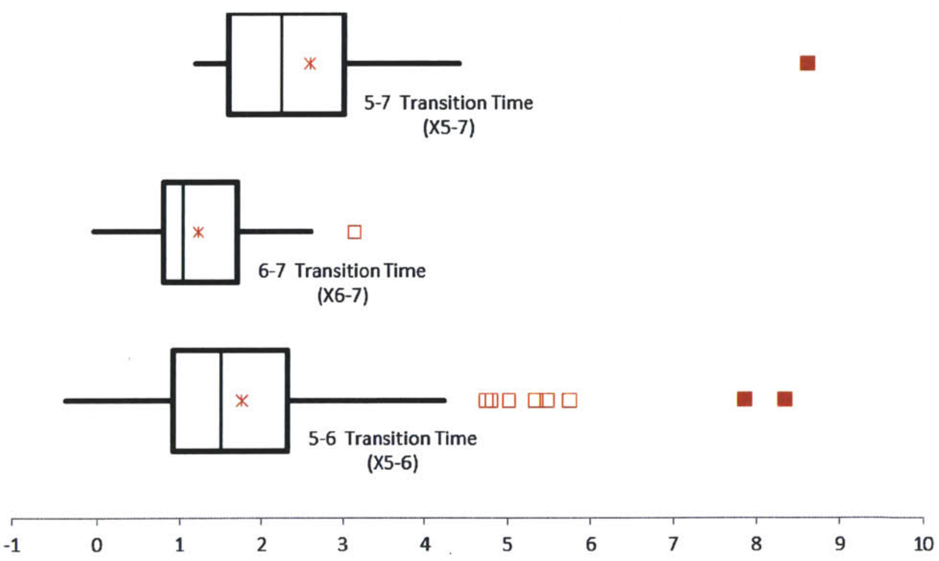
Box-Whisker Plot Comparison of Transition Times 3 to 5



Box-Whisker Plot Comparison of Transition Times 4 to 6



Box-Whisker Plot Comparison of Transition Times 5 to 7



Annex 2

Annex 2: VBA Excel functions

Regular bootstrap function VBA code

```
Function BootstrapInv(Arr As Range, intIteration, p)

   n = Arr.Rows.Count
   ReDim Hold(n) As Single          'Bootstrapped array
   ReDim Hold2(intIteration) As Single 'Array for bootstrapped medians

   Randomize

   For j = 1 To intIteration
      'Read values into array
      For i = 1 To n
         Hold(i) = Arr(Int(Rnd * n) + 1)
      Next i
      'Store computed medians into array
      Hold2(j) = Application.WorksheetFunction.Median(Hold)
   Next j
   Call Sort(Hold2)

  'Compute the value relative to the percentile
  BootstrapInv = Application.WorksheetFunction.Percentile(Hold2, p)


End Function

'Sort Function
Sub Sort(Arr() As Single)
```

```
   Dim Temp As Single
   Dim i As Long
   Dim j As Long
   For j = 2 To UBound(Arr)
      Temp = Arr(j)
      For i = j - 1 To 1 Step -1
         If (Arr(i) <= Temp) Then GoTo 10
         Arr(i + 1) = Arr(i)
      Next i
      i = 0
10    Arr(i + 1) = Temp
   Next j
End Sub
```

Inverse bootstrap function VBA code

```
Function BootstrapInv(Arr As Range, intIteration, k)

   n = Arr.Rows.Count
   ReDim Hold(n) As Single               'Bootstrapped array
   ReDim Hold2(intIteration) As Single  'Array for bootstrapped medians

   Randomize

   For j = 1 To intIteration
      'Read values into array
      For i = 1 To n
         Hold(i) = Arr(Int(Rnd * n) + 1)
      Next i
      'Store computed medians into array
      Hold2(j) = Application.WorksheetFunction.Median(Hold)
   Next j
   Call Sort(Hold2)
  BootstrapInv = Application.WorksheetFunction.PercentRank_Inc(Hold2, k)

End Function


'Sort Function
Sub Sort(Arr() As Single)
   Dim Temp As Single
   Dim i As Long
   Dim j As Long
   For j = 2 To UBound(Arr)
      Temp = Arr(j)
      For i = j - 1 To 1 Step -1
         If (Arr(i) <= Temp) Then GoTo 10
```

```
        Arr(i + 1) = Arr(i)
      Next i
      i = 0
10    Arr(i + 1) = Temp
   Next j
End Sub
```

TransTime function VBA code

```
Function TransTime(TRL1 As Byte, TRL2 As Byte, p As Double)
 Dim Tbl1(1, 8) As Single 'Table of average transition times
 Tbl1(1, 1) = 1.46
 Tbl1(1, 2) = 1.17
 Tbl1(1, 3) = 1.41
 Tbl1(1, 4) = 1.5
 Tbl1(1, 5) = 2.46
 Tbl1(1, 6) = 2.06
 Tbl1(1, 7) = 2.74
 Tbl1(1, 8) = 2.32

 Dim Tbl2(8, 9) As Single 'Table of standard deviations
 Tbl2(1, 2) = 1.271574058
 Tbl2(1, 3) = 2.111095722
 Tbl2(1, 4) = 3.669137923
 Tbl2(1, 5) = 4.379652914
 Tbl2(1, 6) = 6.789409538
 Tbl2(1, 7) = 9.561772848
 Tbl2(1, 8) = 9.055569189
 Tbl2(1, 9) = 11.43291368

 Tbl2(2, 3) = 1.270630916
 Tbl2(2, 4) = 2.867553922
 Tbl2(2, 5) = 3.669665839
 Tbl2(2, 6) = 6.0745659
 Tbl2(2, 7) = 8.903365894
 Tbl2(2, 8) = 8.457953044
 Tbl2(2, 9) = 10.93730065

 Tbl2(3, 4) = 1.6999656
 Tbl2(3, 5) = 2.492089825
 Tbl2(3, 6) = 5.433726172
 Tbl2(3, 7) = 8.396301495
 Tbl2(3, 8) = 8.044436248
 Tbl2(3, 9) = 10.69209209

 Tbl2(4, 5) = 1.029626913
 Tbl2(4, 6) = 4.990792692
```

```
Tbl2(4, 7) = 8.131867727
Tbl2(4, 8) = 7.944518789
Tbl2(4, 9) = 10.71078117

Tbl2(5, 6) = 4.948585066
Tbl2(5, 7) = 8.175391521
Tbl2(5, 8) = 8.019676323
Tbl2(5, 9) = 10.78195304

Tbl2(6, 7) = 2.457070811
Tbl2(6, 8) = 3.66010432
Tbl2(6, 9) = 5.627664352

Tbl2(7, 8) = 3.47313281
Tbl2(7, 9) = 3.959444023

Tbl2(8, 9) = 3.19170326

If TRL1 = 0 Then
MsgBox ("Starting TRL is empty")
TransTime = "error"
ElseIf TRL2 = 0 Then
MsgBox ("Ending TRL is empty")
TransTime = "error"
Else
  TransTime = 0
  TransStdev = 0
  For n = TRL1 To TRL2 - 1
  TransTime = TransTime + Tbl1(1, n)
  Next n
End If
TransStdev = Tbl2(TRL1, TRL2 + 1)
X = Application.WorksheetFunction.T_Inv(p, 18)
TransTime = TransTime + X * TransStdev / Sqr(18)
End Function
```

ransTimeInv function VBA code

```
Function TransTimeInv(TRL1 As Byte, TRL2 As Byte, p As Double)
 Dim Tbl1(1, 8) As Single 'Table of average transition times
 Tbl1(1, 1) = 1.46
 Tbl1(1, 2) = 1.17
 Tbl1(1, 3) = 1.41
 Tbl1(1, 4) = 1.5
 Tbl1(1, 5) = 2.46
 Tbl1(1, 6) = 2.06
 Tbl1(1, 7) = 2.74
```

```
Tbl1(1, 8) = 2.32

Dim Tbl2(8, 9) As Single 'Table of standard deviations
Tbl2(1, 2) = 1.271574058
Tbl2(1, 3) = 2.111095722
Tbl2(1, 4) = 3.669137923
Tbl2(1, 5) = 4.379652914
Tbl2(1, 6) = 6.789409538
Tbl2(1, 7) = 9.561772848
Tbl2(1, 8) = 9.055569189
Tbl2(1, 9) = 11.43291368

Tbl2(2, 3) = 1.270630916
Tbl2(2, 4) = 2.867553922
Tbl2(2, 5) = 3.669665839
Tbl2(2, 6) = 6.0745659
Tbl2(2, 7) = 8.903365894
Tbl2(2, 8) = 8.457953044
Tbl2(2, 9) = 10.93730065

Tbl2(3, 4) = 1.6999656
Tbl2(3, 5) = 2.492089825
Tbl2(3, 6) = 5.433726172
Tbl2(3, 7) = 8.396301495
Tbl2(3, 8) = 8.044436248
Tbl2(3, 9) = 10.69209209

Tbl2(4, 5) = 1.029626913
Tbl2(4, 6) = 4.990792692
Tbl2(4, 7) = 8.131867727
Tbl2(4, 8) = 7.944518789
Tbl2(4, 9) = 10.71078117

Tbl2(5, 6) = 4.948585066
Tbl2(5, 7) = 8.175391521
Tbl2(5, 8) = 8.019676323
Tbl2(5, 9) = 10.78195304

Tbl2(6, 7) = 2.457070811
Tbl2(6, 8) = 3.66010432
Tbl2(6, 9) = 5.627664352

Tbl2(7, 8) = 3.47313281
Tbl2(7, 9) = 3.959444023

Tbl2(8, 9) = 3.19170326

If TRL1 = 0 Then
```

```
 MsgBox ("Starting TRL is empty")
 TransTimeInv = "error"
 ElseIf TRL2 = 0 Then
 MsgBox ("Ending TRL is empty")
 TransTimeInv = "error"
 Else
   TransTimeInv = 0
   TransStdev = 0
   For n = TRL1 To TRL2 - 1
   TransTimeInv = TransTimeInv + Tbl1(1, n)
   Next n
 End If
 TransStdev = Tbl2(TRL1, TRL2 + 1)
X = (p - TransTimeInv) / (TransStdev / Sqr(18))
TransTimeInv = Application.WorksheetFunction.T_Dist(X, 18, 1)
End Function
```

Annex 3

Annex 3: Matlab codes for the forecasting algorithms

1-Code generating the influence diagram

```
%generates an influence diagram B from a covariance matrix C.

function B=IDgenerate(C)

P=inv(C);
n=length(C);
B=zeros(n,n);
v=zeros(1,n);
S=zeros(1,n);
t=1;

for i=1:n
    B(i,i)=0;

    for j=1:i-1
        for l=1:i-1
            B(j,i)=B(j,i)+P(j,l)*C(l,i);
        end
        B(i,j)=0;
    end
    if C(i,i)>1000000
        v(i,i)=1000001;
    else
        for k=1:i-1
        S(i)=S(i)+C(i,k)*B(k,i);
        end
        v(i)=max(C(i,i)-S(i),0);
    end

    if v(i)==0||v(i)>1000000
        P(i,i)=0;
        for j=1:i-1
            P(i,j)=0;
            P(j,i)=0;
        end
    else
        P(i,i)=1/v(i);

        for j=1:i-1
            t=P(i,i)*B(j,i);
            for k=1:j-1
                P(j,k)=P(j,k)+t*B(k,i);
                P(k,j)=P(j,k);
            end
            P(j,j)=P(j,j)+t*B(j,i);
        end
        for j=1:i-1
            P(i,j)=-P(i,i)*B(j,i);
            P(j,i)=P(i,j);
        end
    end
end
```

2-Code updating the influence diagram

```
%updates the means in an ID by knowing the result of a the exact times of

%the first n transitions indicated in the entry vector v.

%v must be smaller in length than m and B, who should have the same length.


function M=updater(B,m,u)

M=m;

M(1:length(u))=u;

for i=length(u)+1:length(m)

   M(i)=m(i)+B(1:i-1,i)'*(M(1:i-1)-m(1:i-1))';

end
```

3-Code generating the full bounded Influence Diagram forecast

```
% creates forecasting matrices using ID method of variables vector V, with 18 zeros in

% between. Bounds the forecasts by a 2.5 upper limit, and -1.7 lower limit.


function megaforecaster(B,m,V)


A=zeros(2);


    for i=1:length(V)

    A((27*(i-1)+1):(27*(i-1)+8),1:8)=forecaster(B,m,V(:,i)');

    end


    %this second part is  a verification for "explosion correction"

    for i=1:length(A(:,1))

        for j=1:length(A(1,:))-1

            if A(i,j)>2.5

                A(i,j)=2.5;

            elseif A(i,j)<-1.7

                A(i,j)=-1.7;

            end

        end
    end
A=circshift(A,[0,1]);
 xlswrite('forecast',A);
```

4- Code generating the autoregression forecast

```
%this function's aim is to perform all of the forecast in the forecasting
%table for a technology. For every coefficient of the table (matrix A, the output), the algorithm
%regresses the current variable against all the previous (available) ones
%in the selected subset(S), then performs the forecast.



function A=regression_forecaster(S,u)


A=zeros(length(u));
A(:,end)=u';


for i=1:length(u)-1
    A(1:i,i)=u(1:i)';


    for j=i+1:length(u)
        A(j,i)=[1,u(1:i)]*regress(S(:,j+1),S(:,1:i+1));
    end


end
```

4.5 Code generating the closest neighbor forecast

```
%this algorithm picks the closest neighbor at each step, and then assumes
%that the growth factors will be similar between this technology and the
%"so-far-closest-one" (unless the last value was zero, then it will be the same additive growth).



function A=closestneighbour(V)
```

```
A=zeros(2);
n=length(V);
h=length(V(:,1));


for i=2:h-1
    VV=V(1:i,:);
    C=Cov(VV)-2*eye(n);
    u=max(C);
    u=ones(n,1)*u;
    U=(u==C);

    %U is a matrix with zeros, and only one "1" on every column, the position
    %of the "1" in the column indicating the variable that this column variable
    %correlates with the best. This matrix has great chances of being
    %symmetric.

    a=zeros(1,12);
    for  j=1:n
        for k=1:n
            if U(k,j)==1
                a(j)=k;
                break
            end
        end
    end


    %the generated "a" is a (1,n) vector with u(i) being the variable that correlates the most
    %with variable i (u(i) corresponds to the first maximum correlation this double loop finds
when it runs down every column).



    %i's are the forecast tables' columns, j's are rows, and k's are technologies
    for j=i+1:h
        for k=1:n

            A(27*(k-1)+1:27*(k-1)+i,i)=V(1:i,k);

            if V(j-1,a(k))==0
                 A(27*(k-1)+j,i)=A(27*(k-1)+j-1,i)+(V(j,a(k))-V(j-1,a(k)));
            else
                A(27*(k-1)+j,i)=A(27*(k-1)+j-1,i)*V(j,a(k))/V(j-1,a(k));
            end
            A(27*(k-1)+1:27*(k-1)+8,1)=V(1,k)*ones(8,1);
            %Filling the first column with a constant forecast.
            A(27*(k-1)+1:27*(k-1)+8,8)=V(:,k);
            %adding a last column of the actual values, before rotating
            %(just to make the format suitable for the excel)
        end
    end

end

 A=circshift(A,[0,1]);
 xlswrite('forecast',A);
```