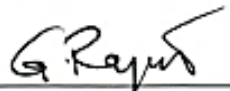
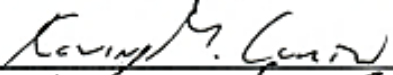

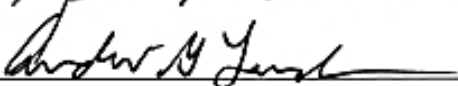
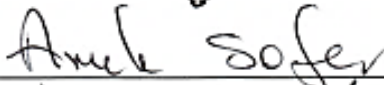



USING OPERATIONAL PATTERNS TO INFLUENCE ATTACKER DECISIONS ON
A CONTESTED TRANSPORTATION NETWORK

by

Daniel E. Stimpson
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Systems Engineering and Operations Research

Committee:

	Dr. Rajesh Ganesan, Dissertation Director
	Dr. Kevin Curtin, Committee Member
	Dr. Karla Hoffman, Committee Member
	Dr. Andrew Loerch, Committee Member
	Dr. Ariela Sofer, Department Chair
	Dr. Kenneth S. Ball, Dean Volgenau School of Engineering

Date: 5/3/17 Spring Semester 2017
George Mason University
Fairfax, VA

Using Operational Patterns to Influence Attacker Decisions on a Contested
Transportation Network

A Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy at George Mason University

by

Daniel E. Stimpson
Masters of Science
Naval Post Graduate School, 2005
Bachelor of Science
United States Naval Academy, 1992

Director: Rajesh Ganesan, Professor
Department of Systems Engineering and Operations Research

Spring Semester 2017
George Mason University
Fairfax, VA



This work is licensed under a [creative commons attribution-noncommercial 3.0 unported license](https://creativecommons.org/licenses/by-nc/3.0/).

DEDICATION

This effort is dedicated to the brave man and women who have, and currently are, sacrificing both life and limb in service to our great republic, the 50 United States of America.

ACKNOWLEDGEMENTS

Since it is only by God's grace that I live and breathe, I am foremost thankful to Him for the opportunity and ability to pursue this work. Additionally, I am grateful to my wife and children (Susan, Caroline, and William) for their patience and understanding during this endeavor. Also, I thank Professor Rajesh Ganesan for his technical expertise and willingness to explore new concepts.

Thank you to Professor Mark Pullen, whose support provided the opportunity for me to transition from active duty in the Marine Corps to these studies at George Mason University. Finally, thanks to Professor David Schum, whose course based on his book *The Evidential Foundations of Probabilistic Reasoning* provided a seed for much of my thinking on this subject.

TABLE OF CONTENTS

	Page
List of Tables	viii
List of Figures	ix
List of Equations	xi
List of Abbreviations and Symbols.....	xii
Abstract	xiii
Chapter One - Introduction	1
Overview	1
The Improvised Explosive Device	3
The Logistician’s Challenge	6
The OODA Loop.....	7
Learning by Induction	11
The Operational Problem	13
Making Predictions	15
A System View	18
Reinforcement Learning.....	20
Research Gap.....	21
Research Contribution.....	24
Dissertation Structure.....	25
Chapter two – Literature Review	27
Vehicle Routing.....	27
Network Interdiction.....	28
Route Clearance Operations.....	32
Attack Pattern Recognition	36
Clustering.....	40
Accounting for Defender Activity	41
Addressing Current Methodological Shortcomings	47

Chapter Three – Model Design and Formulation	51
Model Description.....	52
Model Formulation.....	54
Decision Variable	58
System State	60
Resource State	61
Decision Constraint	63
Information State	63
Knowledge State.....	65
Attack Probability.....	66
Exogenous Information	68
Cost and Objective Functions.....	69
Convoy Operating Cost	69
Delivery Reward.....	70
Customer Inventory Holding Cost.....	71
Unmet Demand Cost	71
Attack Costs.....	72
One step cost function	73
State Value Function Transition	73
Policy	75
Chapter Four – Experimental Results and Analysis	77
Computational Results	77
Base Case Experiment.....	77
Examining the Base Case	83
Policy Response to Different Environments	88
Base Case Policy Summary	93
Simulation Results.....	93
Alternative Threat Profile.....	96
Hybrid Network Simulation Example	102
Summary of Experimental Findings	104
Chapter Five – Conclusion.....	106
Conclusion.....	106

Improved Modeling.....	109
Future Research.....	110
Addressing Dimensionality	110
Time Series Pattern Recognition	111
Symbolic Aggregate Approximation.....	113
T-Patterns.....	115
Close Motifs	117
Appendix A: The Nine Principles of IED Combat	120
Appendix B: Excerpt from Convoy Operations Handbook.....	121
Movement Control (U.S. Marine Corps, 2001)	121
Appendix C: Base Case Policy Adaptations	122
References	123

LIST OF TABLES

Table	Page
Table 1: Base case scenario constraints	79
Table 2: Base case cost parameters.....	81

LIST OF FIGURES

Figure	Page
Figure 1: IED incident counts in IRAQ, 2003-2009 (Cordesman, 2015)	2
Figure 2: IED attack counts in Afghanistan, 2009-2012 (International Security Assistance Force (ISAF), 2013)	3
Figure 3: Categories of effective enemy-initiated attacks (EIAs) in Afghanistan, 2014-2015 (U.S. Department of Defense, 2015)	5
Figure 4: The simplified conception of John Boyd's OODA loop (Osinga, 2001)	8
Figure 5: John Boyd's complete OODA loop (Boyd, 1995)	10
Figure 6: System Block Diagram	18
Figure 7: Agent-environment interaction in RL	20
Figure 8: Risk associated with Decision Analysis Gaps (Connable, Perry, Doll, Lander, & Madden, 2014)	23
Figure 9: Dantzig logistics game illustration	29
Figure 10: Diagram of railway network of Western Russia and Eastern Europe (Harris & Rose, 1955)	30
Figure 11: Route clearance in Afghanistan	35
Figure 12: Example display of IED activity in Iraq during September, 2006 (Ardohain, 2016)	38
Figure 13: Visualization and description of EIAs in Afghanistan (Center for Army Analysis, 2016)	42
Figure 14: Example EIA pattern analysis and description for Afghanistan (Center for Army Analysis, 2016)	43
Figure 15: Representation of unequal discovery probabilities. Red and blue delineate the direction of travel. The numbers are counts of traversals associated with the two discoveries shown (Koyak, 2009b)	45
Figure 16: Illustration of 436 randomly-generated emplacement events, illustrating the complexity of the operational environment (Koyak, 2010)	46
Figure 17: Problem illustration	54
Figure 18: Supply routes in Iraq during spring of 2009 (Center for Army Analysis, 2012)	55
Figure 19: Supply routes in Afghanistan during spring of 2009 (Center for Army Analysis, 2012)	55
Figure 20: Example network and adjacency matrix	59
Figure 21: Example of defender action choices for some state (\mathbf{St}) in a two path case ..	60
Figure 22: Example attack probability functions	67
Figure 23: Time relationships between information flow and state transitions	74

Figure 24: Example simulation output.....	78
Figure 25: Four base case attacker threat profiles	80
Figure 26: Ten base case customer demand profiles (Poisson)	80
Figure 27: The distribution of accrued objective function value for simulations applying RL and myopic decision policies in four different operating environments'	82
Figure 28: Consolidated base case simulation results for RL agent.	84
Figure 29: Median performance of RL and myopic agents in the base case	85
Figure 30: Impact of added RCPs	86
Figure 31: Detailed base case comparison.....	87
Figure 32: Count of system states with each action choice for myopic and RL decision policies (labeled by Demand level Threat level).....	89
Figure 33: Comparison of policy dynamics, myopic and RL agent in base case (lower threat case with 1 RCP).....	91
Figure 34: RL agent policy response to changes in expected customer demand and attacker threat level in base case (1 RCP).....	92
Figure 35: Example RL and myopic simulation output (lower demand and threat profiles)	94
Figure 36: Operational dynamics during example simulation runs. Top: cost. Bottom: activity proportions	95
Figure 37: Inverted attack probability profile and agent performance comparison.....	97
Figure 38: Increasing attack probability profile and agent performance comparison	97
Figure 39: RL agent policy response to alternative attacker threat profiles	99
Figure 40: Policy comparison showing the percentage of system states with given agent decisions for the three threat profiles.....	101
Figure 41: RL agent path utilization in three route case, lower threat profile	103
Figure 42: RL agent path utilization in three route case, higher threat profile ²⁸	103
Figure 43: Representation of an activity sequence on a road segment ending in an attack.	112
Figure 44: Example of a subset of possible activity sequences	113
Figure 45: A hierarchy of various time series representations in the literature (Lin, Williamson, Borne, & DeBarr, 2012)	114
Figure 46: Example time series that appears to be random	115
Figure 47: Example time series with hidden, but detectable, patterns.....	116
Figure 48: The formation of a T-Pattern from simple to complex (Jonsson, 2011).	117
Figure 49: Constructing activity tree for closed motif detection (Nguyen, Ng, & Yew- Kwong, 2014)	118

LIST OF EQUATIONS

Equation	Page
Equation 1: Transition probability rule	56
Equation 2: Reward bounds in any given time step.....	57
Equation 3: Long run total value of rewards (Stewart, 1999).....	57
Equation 4: Agent action vector	59
Equation 5: State variable definition.....	61
Equation 6: Customer inventory update	62
Equation 7: RCP inventory update	62
Equation 8: Delivery decision constraint.....	63
Equation 9: Traffic density update.....	64
Equation 10: Attack density update	64
Equation 11: Path probability of attack	66
Equation 12: Fixed probability of attack for each path.....	68
Equation 13: Variable probability of attack for each path.....	68
Equation 14: Deterministic convoy operating cost	70
Equation 15: Deterministic delivery reward	71
Equation 16: Expected inventory holding cost	71
Equation 17: Expected penalty for unmet demand	72
Equation 18: Expected attack penalty	73
Equation 19: One-step cost function.....	73
Equation 20: One-step state value approximation	74
Equation 21: Value function approximation (learning equation)	74
Equation 22: Myopic one-step state value approximation.....	77

LIST OF ABBREVIATIONS AND SYMBOLS

The Actionable Hot Spot	AHS
Approximate Dynamic Programming	ADP
Blue Force Tracker	BFT
Critical Interval	CI
Combine Information Data Network Exchange	CIDNE
Counterinsurgency	COIN
Counter Radio Electronic Warfare system	CREW
Combat Service Support Element	CSSE
Discrete Fourier Transforms	DFT
Dynamic Programming	DP
Enemy-Initiated Attacks	EIA
Force Protection Resource	FPR
Improvised Explosive Device	IED
International Security Assistance Force.....	ISAF
Irregular Warfare	IW
Kernel Density Estimator.....	KDE
Joint IED Defeat Organization	JIEDDO
Line of Communication	LOC
Marine Air Ground Task Force.....	MAGTF
Mobility Control Center.....	MCC
Markov Decision Process	MDP
Maximum Flow Network Interdiction Problems.....	MFNIP
Mine Resistant Ambush Protected (military vehicle).....	MRAP
Maximize Shortest Path problems	MXSP
Non-homogeneous Poisson Process	NHPP
Non-deterministic polynomial-time.....	NP
Observe, Orient, Decide, and Act	OODA
Operations Research	OR
Route Clearance Patrol	RCP
Reinforcement Learning	RL
Symbolic Aggregate Approximation	SAX
Significant Activities	SIGACTS
Transition Probability Matrix	TPM
Traveling Salesman Problem	TSP
Vehicle Routing Problem.....	VRP

ABSTRACT

USING OPERATIONAL PATTERNS TO INFLUENCE ATTACKER DECISIONS ON A CONTESTED TRANSPORTATION NETWORK

Daniel E. Stimpson, Ph.D.

George Mason University, 2017

Dissertation Director: Dr. Rajesh Ganesan

The recent counterinsurgency (COIN) campaigns conducted in Iraq and Afghanistan motivate this dissertation which examines how reinforcement learning can be applied to the conduct of continuous military operations on a contested road network under the constant threat of attack by ambush. Specifically, this research studied the application of reinforcement learning (RL) methods to learn operational dynamics important to route selection and timing that improve network defender performance.

Recent warfare has been characterized by ambushes with concealed bombs – known as improvised explosive devices (IEDs) – used to disrupt and harass military operations on the roadways of Iraq and Afghanistan. If history provides any indication of the future, it is reasonable to expect that U.S. forces will likely be engaged in these kinds of operations again.

The IED ambush is the attacker's prediction of the future. The choices of time, place, and technique of attack are made based on the attacker's expectation of an attack

opportunity. Accordingly, it is natural to ask how he came to his conclusion. That is, what did he observe that led him to his particular conclusion? Further, we should back up and ask if he had observed something different, how would his attack decision have changed? Here we are asking the counterfactual question, what would have happened if the target (hereafter referred to as the defender) had operated differently? In this way, the problem being addressed goes beyond the tactical problem of maximizing IED detection and avoidance, or minimizing damage and delay. Rather the problem is one of using the defender's operational choices (that are being observed by the attacker) as a direct means to shape the attacker's expectations and therefore his subsequent actions.

The foundation for this approach is the late Colonel John Boyd's well known OODA loop conceptual model of warfare. Boyd's model, which is widely accepted in modern military theory, contains four recurring functions from which competitor behaviors emerge: Observation, Orientation, Decision, and Action. The unique nature of the network counter-IED problem is that it is ongoing and repetitive with nearly constant interaction between the opponents who are engaged in their own individual OODA loop decision processes.

Recurring military operations on roadways are such that the network defender cannot escape the observation of his attacker. This gives the attacker a distinct advantage. Therefore, any complete model of military logistic processes under contested conditions must not only provide solutions to the supply distribution problem, but also must address the IED ambush problem. This dissertation seeks to incorporate the well-

established OODA loop principles into a reinforcement learning scheme in order to craft improved operational plans.

The general approach taken is to greatly relax the normal Vehicle Routing Problem (VRP) constraints in order to provide maximum flexibility in routing and timing choices. The goal is to effectively address the attack problem by crafting vehicle movement patterns that satisfy the military distribution problem, but are principally oriented on shaping the attacker's expectations and choices. Thus, in contrast to most previous work, there is an explicit assumption of dependence between the defender's actions and the attacker's choices. To date, this approach has not been pursued in the operations research (OR) literature in the context of vehicle routing and scheduling.

RL is an algorithmic method for solving sequential decision problems where an agent learns through trial and error, interacting with its environment. As such, the agent is connected to the environment via perception and action such that the agent seeks to discover a mapping of system states to optimal actions. The goal of RL is to find a decision policy that maximizes the long-run measure of reinforcement which describes the goal to be achieved.

This dissertation introduces and demonstrates a fundamental RL model for determining convoy schedules and route clearance assignments, in light of attack costs on a contested transportation network, subject to deliberate ambushes. Our computational results show meaningful performance improvements over one-step, myopic decision rules. Further, the decision policies that are discovered by the RL agent would be difficult for unaided human planners to duplicate. Thus, our principle contribution to the

field of Operations Research is the development of an underpinning argument and model demonstration of a fundamentally different approach to address the attack prediction problem when conducting repetitive operations on a contested road network. In this we have produced a learning algorithm that doesn't just identifying statistically significant attack patterns and adjust to them, rather it seeks to learn from opponent interaction to influence and exploit attacker behavior.

CHAPTER ONE - INTRODUCTION

Overview

The recent counterinsurgency (COIN) campaigns conducted in Iraq and Afghanistan motivated this dissertation which examines how reinforcement learning (RL) can be applied to provide an improved model of military logistic operations on a contested road network. Specifically, this research focused on the use of RL via approximate dynamic programming (ADP) to represent competitive interaction and improve defender outcomes by learning attacker behavior.

Compared to conventional (force-on-force) conflicts, the protection of military logistic activities takes on greater significance during COIN operations. Historically, in COIN, insurgent attackers have intentionally sought to engage logistic units, seeing them as poorly defended targets that offer a high-payoff and even a potential source of supplies. For example, in the 1930's Mao Zedong expressed his belief that the enemy's rear was the guerrillas' front (U.S. Department of the Army, 2009). Juxtaposed are the counterinsurgent defenders who generally operate from fixed bases and conduct regular transportation activities to maintain logistical support and conduct other military operations.

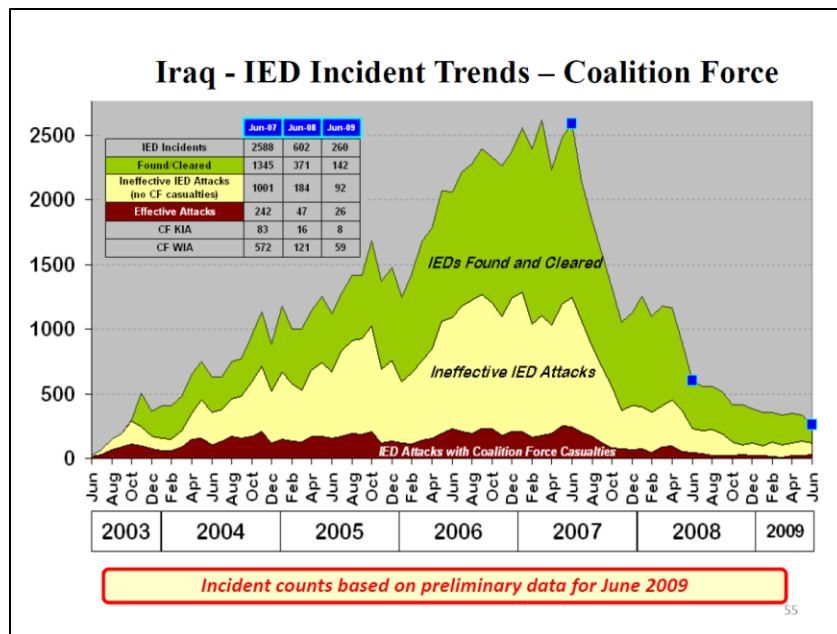


Figure 1: IED incident counts in IRAQ, 2003-2009 (Cordesman, 2015)

Despite efforts to the contrary, modern, large-scale distribution practices can provide attackers with a nearly constant stream of targets, causing the domestic roadways to become the main battle area (U.S. Department of the Army, 2009). Accordingly, the protection of military logistic activities can take on greater significance during COIN warfare compared to conventional conflicts. This dynamic was clearly displayed in Iraq and Afghanistan, where insurgents commonly utilized concealed bombs, known as improvised explosive devices (IEDs)¹ to routinely ambush coalition force maneuver with significant damaging effect (see Figure 1 and Figure 2). In fact, from 2001 through 2013 more than 60% of U.S. combat casualties were caused by IED ambushes (Barbaro, 2013).

¹ The term Improvised Explosive Device (IED) was introduced by the British in the 1970's during their conflict with the Irish Republican Army (IRA)

If history is any indication of the future, it is reasonable to expect that U.S. forces will likely conduct similar operations and face similar challenges again in the future, making this an important and enduring military problem.

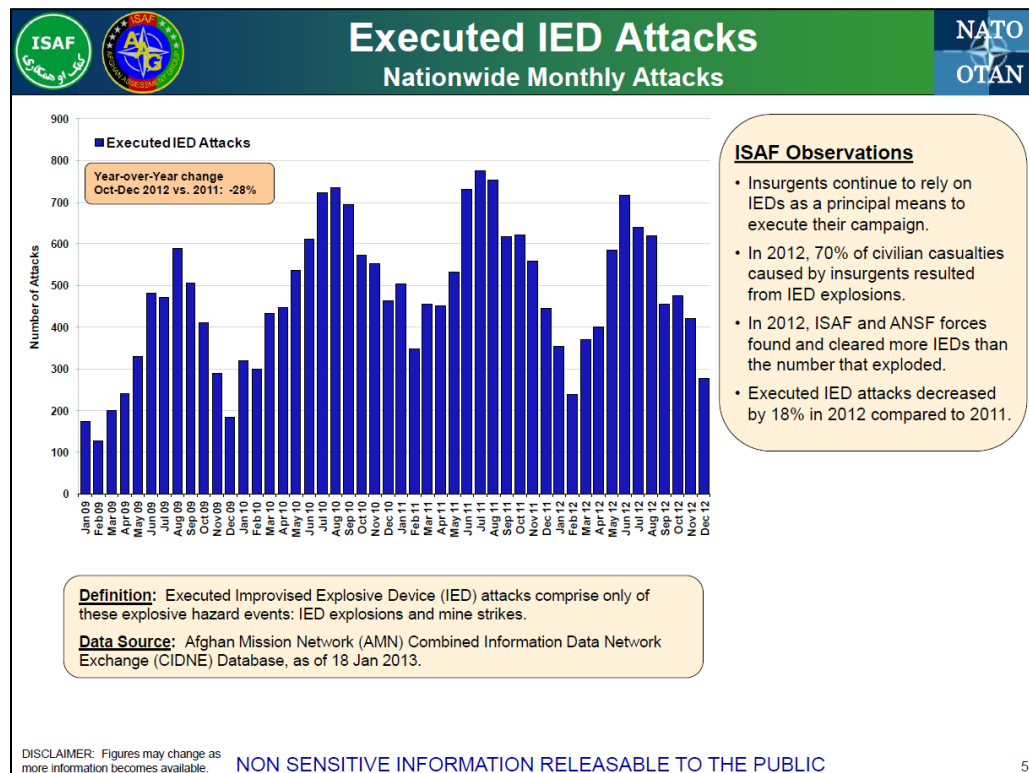


Figure 2: IED attack counts in Afghanistan, 2009-2012 (International Security Assistance Force (ISAF), 2013)

The Improvised Explosive Device

Because of their low cost, ease of manufacture, and significant psychological effects, IEDs provide aggressors with a lethal means to maintain a facade of strength as they appear to freely choose the time and place of engagement against an established

power which they seek to portray as helpless and vulnerable. Further, IEDs can take many forms because there are ample raw materials with which they are made. From household chemicals to modified military ordnance, their construction can take virtually unlimited forms and their sheer simplicity often makes them hard to defeat, even with the most sophisticated equipment.

In the recent Iraq and Afghanistan experiences, insurgent attackers continually and deliberately adapted countermeasures to overcome many of the latest and best military defensive strategies. In fact, for more than a decade, this occurred almost naturally as IED technologies and tactics were developed and proliferated to counter coalition technologies and interrupt their freedom of movement. Thus, despite billions of dollars invested to defeat IEDs, visual observation remained the most common and effective method of detecting their presence² (Joint IED Defeat Organization, 2010).

During this period, the proliferation of IEDs was such that they clearly became one of the attacker's weapons of choice (see Figure 3). But, IEDs are not new. An early example of a coordinated, large-scale IED campaign was the Belarussian Rail War, launched by Belarussian guerrillas against the Germans during World War II. Both command-detonated and delayed-fuse IEDs were used to derail thousands of German trains in 1943-44 (Stockfish & Yariv, 1970). Also, during the Vietnam conflict, mines and booby traps killed more U.S. servicemen than IEDs did in Iraq and Afghanistan

² In response to the escalating use of IEDs in Iraq, in 2003 the Army Chief of Staff established the Army IED Task Force. Then in February 2006, the Deputy Secretary of Defense established the Joint IED Defeat Organization (JIEDDO) under DoD Directive 2000.19E. From 2006 through 2013 JIEDDO's annual budget greatly exceeded \$1 billion per year (Barbaro, 2012).

combined (Associated Press, 1970; Barbaro, 2013).^{3,4} Historically, IEDs have been used by the Irish Republican Army, the Medellin cartel, Hamas, Boko Haram, and others such that since 2001, across the globe, there have been tens of thousands of IED detonations per year with seemingly no end in sight (Barbaro, 2013).

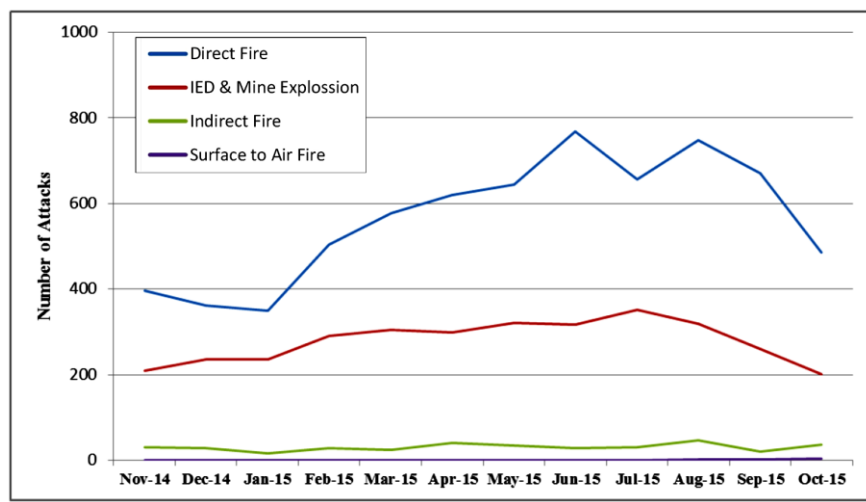


Figure 3: Categories of effective enemy-initiated attacks⁵ (EIAs) in Afghanistan, 2014-2015 (U.S. Department of Defense, 2015)

³ According to the September 9, 1970 article cited (published before the end of the Vietnam conflict), Pentagon sources claimed half of the 43,000 killed to date died from booby-traps. Separately, the official Pentagon estimate attributed 6,500 total deaths to mines, booby traps, and grenades. In either case, this exceeds those reported as killed in action (KIA) by IEDs in both Iraq and Afghanistan to date.

⁴ As of May, 2014 at least 60 percent of U.S. combat casualties reported in Iraq and Afghanistan were due to IEDs, meaning approximately 3,200 were Killed in Action and 33,100 Wounded in Action by IEDs.

⁵ EIAs are a subset of security incidents that do not include any friendly-initiated actions. Effective attacks result in combat-related non-insurgent casualties (killed-in-action or wounded-in-action) which are a subset of all reported EIAs (U.S. Department of Defense, 2015). Thus, the charted data do not include ineffective and unexploded IEDs (IEDs that exploded but caused no casualties and those found and cleared without detonating). Historically these have far outnumbered the number of effective IEDs.

The Logistician's Challenge

"Good generals study tactics; great ones study logistics."
General Omar Bradley, U.S. Army

The logisticians' fundamental responsibility is to provide the right items, in the right quantity, to the right place, at the right time to support often widely dispersed military operations. Thus, despite a threat of attack, they will always strive for the most efficient and timely deliveries possible.

It is not hard to appreciate the significant challenges of maintaining a large distribution and mobility network under austere conditions across a foreign road network. Where infrastructure exists, there may be several alternative modes of transportation (i.e., maritime, rail, air, and surface) to move personnel and material throughout the distribution network. But, even when all of these alternatives exist, the primary means of distribution has historically remained surface transportation over roadways, with trucks and armored vehicles. In many cases, the distribution challenge is intensified due to poor infrastructure or even a total lack of improved roads where routes can be cross-country and ad hoc across open terrain.

One way U.S. forces defended against the threat of IEDs was to form dedicated clearance units (Route Clearance Patrols or RCPs) to conduct periodic route clearing operations in a defensive effort to minimize their exposure and the impact of IED ambushes. Since these route clearing resources were limited, their efforts typically focused on the roads with the highest perceived risk to find and neutralize as many IEDs as possible.

The OODA Loop

One approach to designing attack prediction algorithms is to account for the critical activities leading to the attacker's ambush decision as a means to understand his choices of time and place. The late U.S. Air Force Colonel, John Boyd provided a framework to do this when he developed the well-known Observe-Orient-Decide-Act (OODA) Loop model through his studies of air-to-air combat, human knowledge acquisition, and competitive decision making. He argued that, for people to make timely decisions, they "must be able to form mental concepts of observed reality, as [they] perceive it, and be able to change these concepts as reality itself appears to change" (Boyd, 1976). Further, he noted that as people strive to build an accurate understanding of reality, they iteratively improve their understanding in a cycle of informational creation and destruction that is repeated until they arrive at an internally consistent perception of reality. Finally, Boyd believed that the extent to which people are able to match their mental image to physical reality, determines the extent to which they can make informed decisions (Osinga, 2001).

These insights are incorporated into Boyd's OODA loop conceptual model of warfare, which contains the following four primary functions (Boyd, 1986):

- Observation: The utilization of surveillance, reconnaissance, and other means to fill in knowledge gaps.
- Orientation: The act of making sense of what is learned and building perceptions of reality.
- Decision: Determining what course of action (COA) to pursue.
- Action: The action taken to disrupt or destroy the opposing side's functions.

Popularized interpretations of the OODA loop conceive of gaining advantage in competitive situations by out-pacing and out-thinking one's opponent by means of repeatedly cycling through the OODA loop more rapidly than the opposing side (Osinga, 2001). In popular lexicon, it is often assumed that whichever side can operate "inside" of the other's OODA loop (meaning operate faster) will gain the advantage and dictate the terms of the conflict (Weber, 2007). While this concept is in keeping with Boyd's ideas, to simply think of the OODA loop concept as a speed contest is a gross oversimplification of his strategic thinking and only a partial application of his theory (Osinga, 2001). This common view of the OODA loop is illustrated in Figure 4.

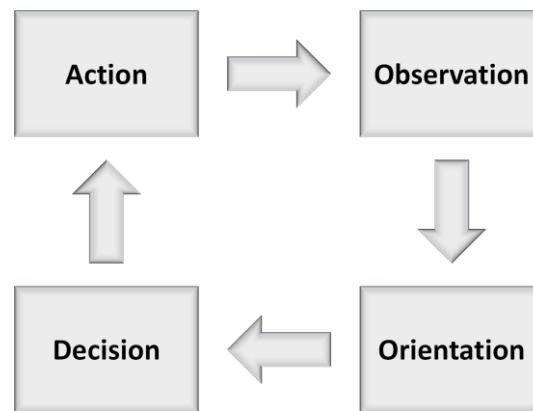


Figure 4: The simplified conception of John Boyd's OODA loop (Osinga, 2001)

The OODA loop is easy to comprehend, but it is based on a concept that runs much deeper than the popular conception. Full understanding of it reveals that it provides a comprehensive framework in which to think about competitive human

undertakings, including modern warfare. Further, it has been stated that “Boyd’s loop can apply to operational, strategic, and political levels of war” and that one of its great strengths is its “elegant simplicity” which makes it useful in many domains (Grey, 1999). This is why the OODA loop has become a well-accepted part of the conceptual mainstream of all western militaries (Osinga, 2001).

There are several shortcomings of condensing Boyd’s theories to a simplified four-step loop. First, such simplifications tend to emphasize speed in decision making. This obscures the complexity and richness of its governing themes. Second, overemphasis on the basic OODA loop conceals Boyd’s broader ideas about developing organizations that are agile and adaptive so they can survive and prosper in the face of fierce competition. Finally, oversimplifications can lead to thinking that the OODA loop is just a recipe to be followed when it represents a way of strategic thinking that is based on insights from history, science, philosophy, and other military theory (Osinga, 2001).

As shown in Figure 5, Boyd’s OODA construct is an intricate network of interconnections and feedback. Further, because it is conceived as a loop, it is continuous, repeating and cascading.

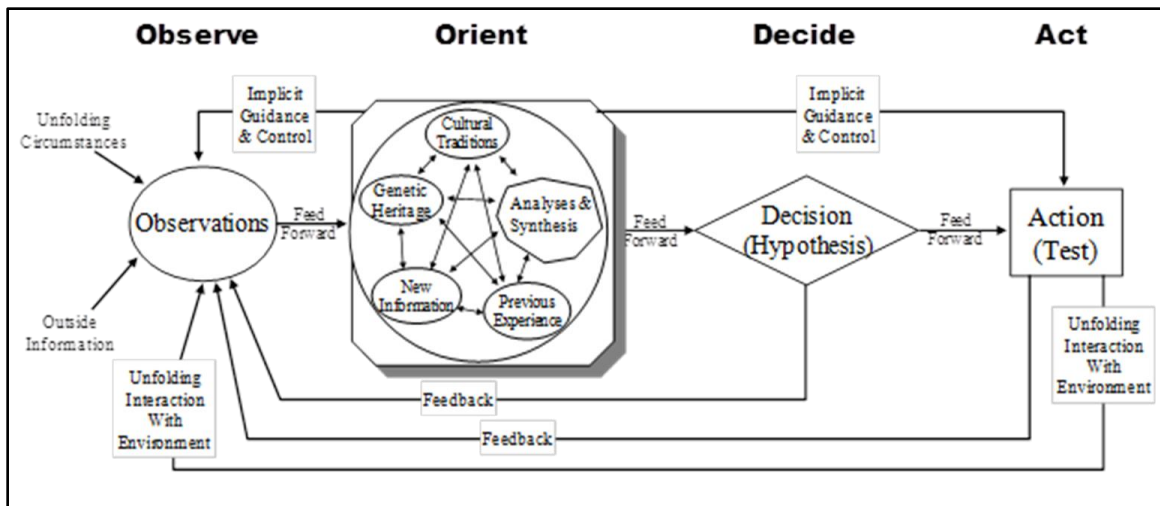


Figure 5: John Boyd's complete OODA loop (Boyd, 1995)

Boyd said not all stages of the OODA loop are created equal, but that Orientation is the *schwerpunkt* (or the decisive point) of the model and of all human decision making in general (Ford, 2010). According to Boyd, “Orientation shapes the character of present observation-orientation-decision-action loops - while these loops shape the character of future orientation” (Boyd, 1987). If Boyd is correct, orientation is the most critical function not only to the current decision, but to future decisions and ultimately to the long-term success of a competitive undertaking. This would mean that a key to prevailing in any competitive situation is the ability to remain well-oriented, from start to finish. Then the converse is also true; systemic disorientation will be a significant cause of difficulties.

The attacker’s choice to use IEDs is not arbitrary, rather it is one part of a broad effort designed to gain advantage (Koyak, 2009b). The essential nature of the counter-IED problem on road networks is that it is ongoing and repetitive, with nearly constant

interaction between opponents, each engaged in their own OODA loop decision processes. As such, the defenders' every observable action is shaping his opponents' mental image and his expectations about what will likely occur in the future.

In the problem at hand the network defender is focused on delivering personnel and material throughout the battle space, but they cannot escape the fact they are operating under the direct observation of their attacker, who has his own planning cycle directed at interdicting them. Thus, any complete model of logistic processes under such contested conditions must address both the supply distribution problem and the opponent interaction problem. The approach taken in this dissertation is to incorporate the military's well understood OODA loop principles into an RL scheme in order to improve current operational planning approaches.

Learning by Induction

"The insurgency in Iraq took merely weeks to adapt to the MRAP armor upgraded vehicles⁶" (Garaux, 2010).

Learning from interaction is the fundamental idea underlying nearly all theories of learning and intelligence (Sutton & Barto, 1998). Further, learning from observation involves the use of inductive mental processes, but no amount of observation can guarantee knowledge of the future. This presents the decision maker with what is known as the problem of induction which has been summarized by saying "in the future, the future will resemble the past because, in the past, the future has resembled the past" (Schum, 1994). John Stuart Mill explained the problem of induction by noting that no amount of observations of white swans can allow the inference that all swans are white,

⁶ MRAP refers to Mine-Resistant Ambush Protected vehicles, specially designed to protect from IED attacks

but discovering a single black swan is sufficient to refute a conclusion based on all past observations (Mill, 1843; Taleb, 2004). In a competitive environment, the problem is amplified because we can never expect an opponent to remain passive and allow all of his actions to be openly observed. This not only means no amount of past observation will ever perfectly or continuously predict the future, but many actions of an opponent will remain permanently hidden from observation. Further, an adversary can be expected to not only vary his observable patterns, but to also engage in active deception.

Thus, IED warfare on a road network confronts both sides with a situation where they are engaged in probabilistic reasoning about their opponent's future actions, but each side's observations provide them with asymmetric information. That is, both sides look backward to predict forward in time, but there is an important difference. The attacker has nearly perfect observation of the current defender transportation movements (which are carried out in open view) while the defender has very little direct observation of the attacker who generally remains clandestine. This means that the attacker has current information about where the defender's vehicles are, as well as historical information about where they have been and at what frequency. Meanwhile, the defender is limited to historical information based on IED discoveries which only indicate that an act of emplacement occurred; he often cannot determine exactly when or how the IED ambush was established. This means the information distribution is strongly skewed to the attacker's advantage, and the defender's problem of induction is significantly more difficult than the attacker's.

Where asymmetry exists, or where it would be advantageous to create, can only be determined through a robust observe function. In the OODA loop lexicon, failing to recognize significant asymmetries is to fail to maintain correct orientation. Thus, observations must focus on one's own processes and assumptions as much as on those of the opponent. Additionally, for the decision process to be anticipatory it must incorporate an understanding of how every action taken will be perceived by the opponent and how he is likely to orient, decide and act in response.

One of the more obvious asymmetries is the Western military's long held advantage in the use of complex material technologies which can influence decision makers' to favor technological solutions to virtually every problem (Dunlap, 1998). This held recently in Iraq and Afghanistan where technological solutions to the IED problem were continually emphasized. While most of these provided improved survival and detection, insurgents continually found effective asymmetric counterstrategies to maintain their attack effectiveness. This provides a strong reminder of the importance of continuously pursuing non-material solutions that emphasize learning, decision making, and operational art.

The Operational Problem

If we view an IED ambush from the attacker's perspective, we see that the IED emplacement is the attacker's prediction of the future. The choices of time, place and technique are made by the attacker based on his expectation of a future attack opportunity. Accordingly, it is natural to ask why the attacker came to their particular conclusion. For any rational actor, this will be based on what was observed. But, it is

important to back up one step further and ask the counterfactual question; if the attacker had observed a different sequence of defender activity, would his attack decision have been different? If such differences are manifested and can be learned, then it may be possible to make better attack predictions and improve outcomes for the defender.

This is the central problem motivating this research. It goes beyond the tactical problem of maximizing IED detection and avoidance, or minimizing damage and delay. It is to pursue a method for using the defender's operations as a direct means of influencing and anticipating an attacker's decisions - or in the common military vernacular, "to get inside the attacker's OODA loop" (Boyd, 1986).

In military parlance, this perspective means that the defender's individual actions must be viewed operationally, as opposed to tactically.⁷ In fact, failure to coordinate operationally, across time and space, may unwittingly simplify the attacker's prediction problem.⁸ For example, take a hypothetical unit arbitrarily choosing to move on some route (A) one morning and changing to some other route (B) the following afternoon; meanwhile, some another unit operating independently might unknowingly choose route (B) in the afternoon of the first day, then route (A) in the morning of the second day. In this example, both units change their routes and schedules in an apparently unpredictable manner, but taken together, from the attacker's perspective, days one and two are

⁷ In the current context, what is meant by the tactical action relates to the conduct of direct counter-IED tasks, such as the choices in route, vehicle formation, detection equipment employed, and IED neutralization methods. In contrast, the operational level encompasses decisions such as synchronizing individual movements and determining priorities of effort (such as the allocation of IED clearance efforts on various routes). This is consistent with military doctrine which describes tactics as actions related to winning individual engagements, where operations focus on sequencing tactical engagements to achieve broader objectives (US Department of Defense, 2013).

⁸ In practice, higher headquarters may specify the route or the determination may be left to the convoy commander (U.S. Marine Corps, 2001).

identical. In such a case, coordination between the units could have prevented the oversight. Further, a remarkable irony emerges as the attacker actually has a better understanding of the defender's operational maneuver pattern than the defender himself. If we then extrapolate this dynamic to a larger number of participants, it's not hard to imagine how 100's or even 1,000's of individual choices might result in a regular, even uniform overall pattern being presented to the attacker. Accordingly, two conclusions follow. First, we see that seemingly unpredictable individual unit commander choices do not equate to force-level operational unpredictability. Second, the probability of attack on each individual unit action cannot be assumed to be independent of the actions taken by others on the road network.

Making Predictions

“For the Marine patrol, the first step is to understand the enemy and adjust accordingly. Age-old patrolling axioms like ‘vary your routes’ and ‘avoid establishing patterns’ are imperative” (Powledge, 2005).

U.S. military doctrine highlights the importance of not establishing patterns and predictable forms of behavior as a means to improve survivability (U.S. Department of Defense, 2014; Joint IED Defeat Organization, 2010). While this doctrinal statement is sound, its application does not always produce the intended force level result. This is partly because there is often a misunderstanding at the individual unit level of how to produce unpredictability under the prevailing operational conditions.

Predictability, by this definition, depends upon a prevailing system's constraints. For example, imagine an unordered list of the heights of ten individuals. If we attempt to

randomly match one of the individuals to one of the heights on the list, there is a 0.1 probability of being correct. Now consider lining the ten individuals up so that the person to everyone's left is shorter than themselves (ordering them by height); the matching can now be made with certainty. Note, the introduction of just one constraint completely changed the nature of the problem. In fact, the constraint is strong enough to remove all uncertainty. Accordingly, the intensity of a constraint determines the number of possible arrangements a system can take (Ashby, 2011).

The power of constraints is often so ubiquitous that it can be easily overlooked, but constraints are not only a powerful ally to making predictions, but without them prediction is impossible. That is, unconstrained systems are chaotic, and therefore totally unpredictable. For example, in search and rescue situations constraints are what allow a search area to be defined. There is always a maximum range that the missing object can travel from its last known position based on the existing conditions. Thus, it is impossible for the object to be outside the area defined by the prevailing physical constraints.

U.S. military doctrine recognizes that organizations operate in certain modes rather than randomly. It states that there are unique types and levels of insurgent groups which have different strategies and capabilities; understanding these can help reveal operational patterns and help predict tactics, techniques, and procedures (U.S. Department of the Army, 2014). This is to say that insurgent organizations have significant constraints on their activities - some obvious, some subtle. But, this is clearly true of all human organizations, not just insurgents', i.e., the defender's logistic

operations. While this is generally acknowledged, it can become a peripheral consideration in logistical planning. Significant improvements may be possible by thoughtful, systematic examination of existing constraints on both the attacker and defender in seeking improved operational planning concepts.

As discussed, organizational behavior stems from the collective behavior of the individual members, where every individual has physical, psychological, and other limitations that bound what they can do, while their perceptions and preferences bound what they are willing to do. This means that while human decisions may be unpredictable, no human activity is ever truly random. Therefore, what is really intended by operators seeking to be unpredictable is to maximize their operational variation. Then from the concept of variation follows the capacity for generating surprise (Ashby, 2011).

The fourth of the nine Principles of IED Combat is “avoid setting patterns” (Joint IED Defeat Organization, 2010). But even as so-called random individual actions don’t necessarily lead to operational unpredictability, neither does maximizing operational variety automatically lead to surprise. Here again we must carefully consider the observer’s perspective and whether or not the observer will be able to distinguish one sequence of actions from another. If so, as Boyd instructed we must also consider what conclusion we think the observer is likely to reach. In other words, will making some adjustment to defender operations be enough of a difference for the observing opponent to notice and adjust his own behavior? Then, will the defender be able to detect and measure the attacker’s behavior change? This leads to the conclusion that generating surprise requires the ability to create distinguishable elements in the set of possible

outcomes (Ashby, 2011). This lies at the heart of Boyd's OODA loop. Both sides are orienting on the activity patterns they can observe, yet they can only recognize patterns according to their ability to discriminate among the different individual actions and sequences of actions.

A System View

If we consider the attacker and defender systems interacting, we might conceive of the defender's distribution system (A) and the attacker's ambush system (B) interacting to bring about outcomes (C). This yields a many-to-many transformation of states such that $S_A \times S_B \rightarrow S_C$. Then as Figure 6 shows, the attack process (B) occurs in a black box and the central problem is to understand what the attack system is doing. Because the black box is unobservable, it may be possible to experiment with the sequence on inputs in time $s_{A_t}, s_{A_{t+1}}, s_{A_{t+2}}, \dots$, and observe the sequence of outcomes, $s_{C_t}, s_{C_{t+1}}, s_{C_{t+2}}, \dots$, to gain knowledge of the dynamics in (B) (Heylighen & Joslyn, 2001).

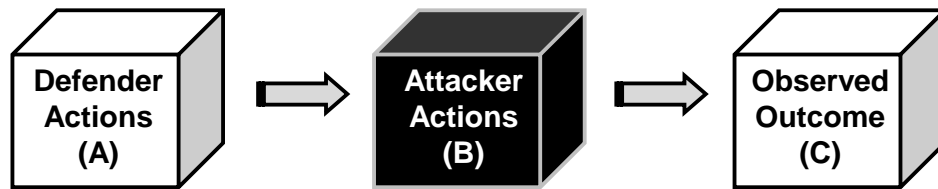


Figure 6: System Block Diagram

Clearly, if we only have random unstructured inputs at (A), we are unlikely to learn much about the attacker processes at (C). In fact, the outcome at (C) may appear to be random also. In contrast, well-crafted defender action plans (for convoys, route clearance, and other observable activities) provide a means to make associations and learn some of the action-reaction dynamics of the overall system. This requires careful record keeping and inference methods to detect whatever patterns may be present or emerging over time. Ultimately, learning and understanding how the activities in boxes (A) and (C) may be related is only possible by taking a network-wide, historically informed, operational perspective of the logistics operations being undertaken.

In a complex system, the dilemma is to discover which part of a measured pattern should be ascribed to “randomness” and which part to “order.” That is, can we find and understand usable information and determine what information to ignore? It is this interplay, between order and randomness, which makes the problem at hand complex as opposed to merely complicated.⁹

Here enters the need for innovative analytics and modern learning algorithms to discover better ways to detect and exploit the structure within a seemingly chaotic environment. When the analytical process begins, the patterns, parameters, and constraints that govern the system are unknown, but to the extent they can be learned, they must be discovered by observing the system’s behavior (Crutchfield, 1994).

⁹ A complicated process (a large system consisting of many components, subsystems, degrees of freedom, etc.) is not necessarily complex (involving randomness and unpredictability) (Crutchfield, 1994).

Taking the principles herein together suggests an improved modeling approach to better characterize the attacker's black box compared to standard probabilistic risk modeling approaches.

Reinforcement Learning

RL is an algorithmic method for solving sequential decision problems where an agent learns through trial and error interacting with its environment. As such, the agent is connected to the environment via perception and action such that the agent seeks to discover a mapping of system states to optimal agent actions (see Figure 7). The goal of RL is to find a decision policy that maximizes a long-run measure of reinforcement that describes the goal to be achieved (Kailbling, Littman, & Moore, 1996). A decision policy is defined as a rule or function that determines the agent's choice given the environmental information available from the observed system state (Powell, 2007).

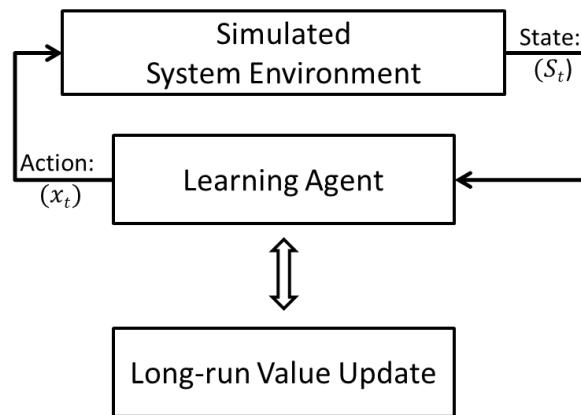


Figure 7: Agent-environment interaction in RL

In RL, the agent is not told which actions to take, but instead must discover them. Since action choices are made to bring about the highest long run value, rather than the highest immediate payoff, a well-structured RL algorithm can naturally discover multi-move sequences to arrive at high payoff states (Sutton & Barto, 1998). This makes it a clear choice for the competitive problem domain under study in this dissertation. Further, by properly adjusting the step-size parameter within an RL algorithm¹⁰ the agent can adjust its decision policy as the environment (i.e., the opponent's behavior) changes (Sutton & Barto, 1998). Since the network counter-IED problem is a dynamic, sequential decision making problem, RL is a well-suited solution method for discovering effective operational schemes.

Research Gap

Commanders appreciated analysis that predicted the location of IEDs, but felt there was still a need for additional capability to predict where and when IED emplacements would be active (Connable, Perry, Doll, Lander, & Madden, 2014).

General Michael Barbero, then Director of Joint IED Defeat Organization (JIEDDO), identified what he saw as the two most significant capabilities produced by the U.S.'s multi-billion dollar counter-IED effort; he called them "game changers." The first is forensic data collection and exploitation which allowed massive improvements in the ability to investigate and establish the link between attacks and the attackers. The second is wide-area surveillance which drastically improved the military's situational awareness (Barbaro, 2013). These innovations significantly reduced one of the attacker's

¹⁰ This is accomplished through Temporal Difference (TD) learning where the step-size parameter is not reduced all the way to zero (Sutton & Barto, 1998).

major advantages - his anonymity. But, after ten years and billions of dollars spent, we continue to see that almost as fast as a technological solution is implemented, the adversary adapts. For example, the Iraq insurgency adjusted to the introduction of improved armored vehicles (MRAPs) in just a few weeks (Garaux, 2010). These experiences demonstrate that even “game changing” technologies are ultimately employed as process improvements and generally do not immediately neutralize an opponent.

To date, even with vastly improved technologies, little has been done to develop and establish any truly new operational concepts in logistic distribution and battlefield circulation. This research effort has sought to engage in this needed area and apply operations research (OR) techniques to explore, develop, and demonstrate a fundamentally new approach to the conduct of recurring transportation operations in a contested environment.

Generally, the force protection¹¹ issue of greatest concern to commanders during Iraq and Afghanistan operations was the IED threat. This drove unit commanders to emphasize the need for understanding the factors driving IED trends and patterns they observed (i.e., location, time, device types, and frequencies), their origin (caches, logistics, and financial networks), and to predict future attacks (Connable, Perry, Doll, Lander, & Madden, 2014). Typically, trend analysis took the form of summary statistics of historical events which were used to inform operational and tactical leaders as to the

¹¹ Force Protection refers to preventive measures taken to mitigate hostile actions against Department of Defense personnel (to include family members), resources, facilities, and critical information (U.S. Department of Defense, 2010).

most likely times and types of attacks, target types, discovery rates (i.e., either finding IEDs or being attacked by them), and the like. But generally, these analytical products did not provide the needed predictions of specific times and locations of future events in a way that could drive surveillance or clearance asset allocations (Ardohain, 2016).

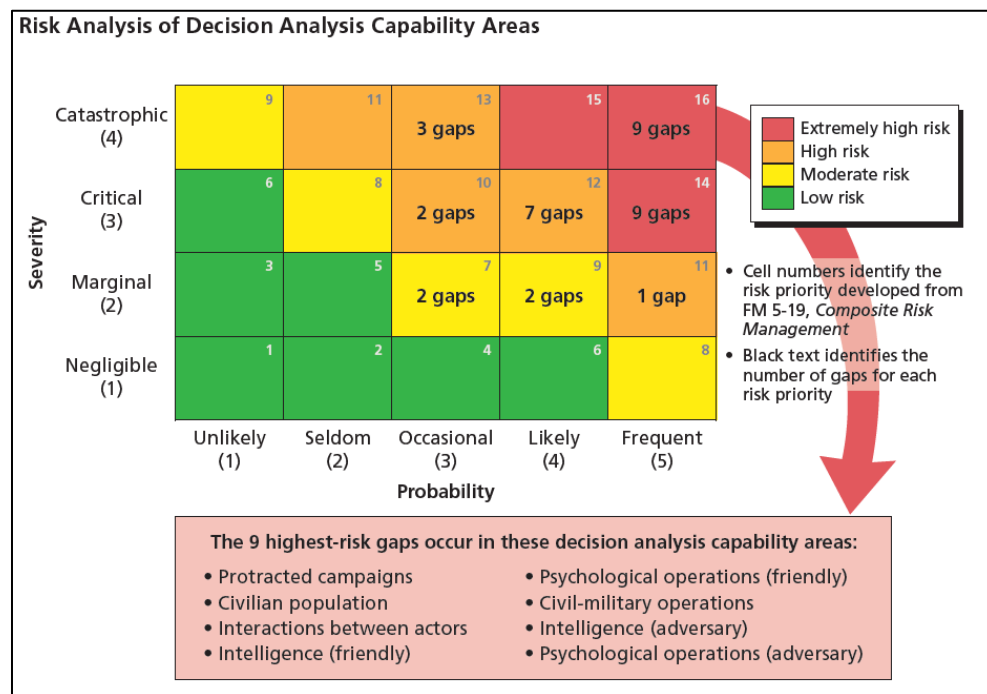


Figure 8: Risk associated with Decision Analysis Gaps (Connable, Perry, Doll, Lander, & Madden, 2014)

Moreover, a recent study by (Connable, Perry, Doll, Lander, & Madden, 2014) stated that most decision support derived from simple analyses, not complex modeling. This remained true even while DoD and the supporting community strived to develop models and simulations in support of Irregular Warfare (IW). Figure 8 shows an

assessment of gaps in DoD analytic capabilities. The horizontal axis displays assessed probability that each gap will effect IW decision making. Then the vertical axis displays the expected damage a lack of that analytic capability might cause to a commander's decision making ability. The highest risk gaps (those expected to be both most frequent and most severe) are colored red. The top nine "extremely high risk" gaps are listed under the chart; note that *interaction between actors* is in this category. (Larimer, Checco, & Persons, August 18, 2008; Connable, Perry, Doll, Lander, & Madden, 2014)

Research Contribution

"Whenever the present state of knowledge concerning something of interest is deemed inadequate, methods of investigation may be considered to improve understanding" (Bhattacharyya & Johnson, 1977)

This dissertation makes a contribution to OR practice by introducing a new approach to military distribution planning consistent with the principles previously expressed. It is based in military and human decision theory as articulated by the military theorist John Boyd. In this aspect, the techniques developed demonstrate a meaningful step toward closing the gap between theory and practice for current decision support algorithms related to military logistic distribution under contested conditions.

Methodologically, this dissertation demonstrates a means to achieve significant improvement over optimizing immediate operational choices based on concurrent risk assessments. This is accomplished through the application of RL via ADP which inherently coordinates individual action choices across the planning horizon, forecasting the downstream effects of its current decision.

The fundamental contributions of this dissertation to the field of OR are:

1. We provide an underpinning argument and a model to demonstrate the potential value of a fundamentally different approach to operational movement control of military transportation assets on a contested road network. This proposes operational coordination of defender activities across the network to shape attacker expectations and improve defender outcomes.
2. Unlike most previous work, this research is not focused on improvement of, and direct application to, existing military practices. Rather we develop and demonstrate a new approach via a learning algorithm that doesn't just identifying attack patterns and adjust activities to accommodate or avoid them, but rather it seeks to influence such patterns in order to exploit them. This concept has not been explored in any published research and is an important extension of current game theoretic and statistical approaches in the field.
3. We integrate understandings from three broad analytical fields related. These are vehicle routing, route clearance and attack pattern recognition, and ADP methodologies for solving RL problems. Each is of these is critical to the implementation of our modeling approach, and when applied in combination, expand current OR practice and can provide meaningful responses to differing environmental conditions which would be difficult for unaided human planners to duplicate.

Dissertation Structure

The remainder of this dissertation is organized as follows: Chapter 2 provides a literature review. Chapter 3 presents the model design and discusses the essential

elements of the ADP formulation. Chapter 4 describes the experimental design, computational results, and analysis of the simulation results. Chapter 5 summarizes the research with conclusions, observations, and recommendations for further work.

CHAPTER TWO – LITERATURE REVIEW

The literature applicable to this dissertation falls into three broad categories. The first is previous work done in the fields of vehicle routing and network interdiction. These form a foundation upon which this work builds. The second is military route clearance problems and attack pattern recognition applied to the network context. The third is the application of ADP to RL which is the core methodology applied. Insights from each of these research areas are critical to the implementation of the modeling approach, and when applied in combination, expand current OR modeling approaches.

Vehicle Routing

The vehicle routing problem (VRP) is a generalization of the Traveling Salesman Problem (TSP) and is non-deterministic polynomial-time hard (NP-hard).¹² It was introduced to the OR community by Dantzig and Ramser in 1959, since then a large body of literature has been associated with it (Dantzig & Ramser, 1959). The classic formulation uses a fleet of capacitated vehicles located at a common depot to deliver goods to a set of customers. Modifications to the VRP introduce various constraints related to vehicle types, travel time, and delivery time windows. Recent surveys have

¹² The polynomial time reduction from the Satisfying Assignment Problem (which is NP-complete) to the Hamiltonian Cycle (finding a cycle that ends and begins at the same node, visiting each node only once) is given in (Karp, 1972). Then the Hamilton Cycle problem (NP-complete) reduces to the Euclidean TSP (finding the minimum cost tour of a set of points in a plane, (Papadimitriou, 1977)). All variants of the VRP are at least as hard as the TSP while none of these problems admit an efficient solution technique unless it is proven that $P=NP$ (Hinton, 2010).

been offered by (Pillac, Gendreau, Gueret, & Medaglia, 2013; Hinton, 2010; Kumar & Panneerselvam, 2012).

Typically the VRP focuses on determining how to distribute goods, in a given time period, by a set of vehicles, to a set of customers. Generally the vehicles are sourced from one or more depots, move throughout a defined network, and are operated by a set of crews (drivers). VRP solutions determine a set of routes for each vehicle, starting and ending at their own depot with all customer demands and operational constraints satisfied. The VRP objective is most often to minimize the overall transportation cost (Toth & Vigo, 2001). A comprehensive treatment of VRPs modeled as Markov decision processes with fully developed ADP formulations is provided by (Goodson, 2010).

Network Interdiction

Network interdiction models have been thoroughly studied by the OR community. Figure 9 shows a representation of an early logistic game, based on a 1953 paper for the U.S. Navy's Operations Evaluation Group. It describes a non-linear, two-person, zero-sum game that allows for separate defender allocation of ships and escort vessels to various routes while an attacker allocates submarines (Danskin, 1962). Later, Wollmer provides an extension to this model to determine the placement of intercepting units in order to maximize the probability of preventing an opposing force from proceeding from a particular node to another in an undirected network. In both treatments, standard gaming assumptions are utilized so that the attacker knows the defender's strategy for placing interceptors (Wollmer R. D., 1970).

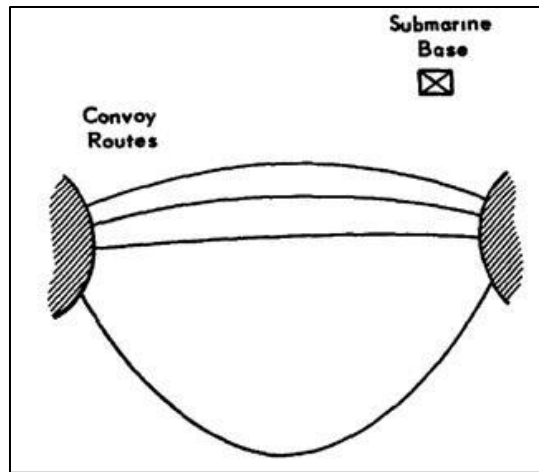


Figure 9: Dantzig logistics game illustration

(Harris & Rose, 1955) provided another early network interdiction example in a classified SECRET paper that was released to the public in 1991. In it they described how to determine a rail network's "bottleneck" (or minimum cut) in order to interdict its flow capacity with air power. The illustrative case was the Western Russia rail network, see Figure 10.

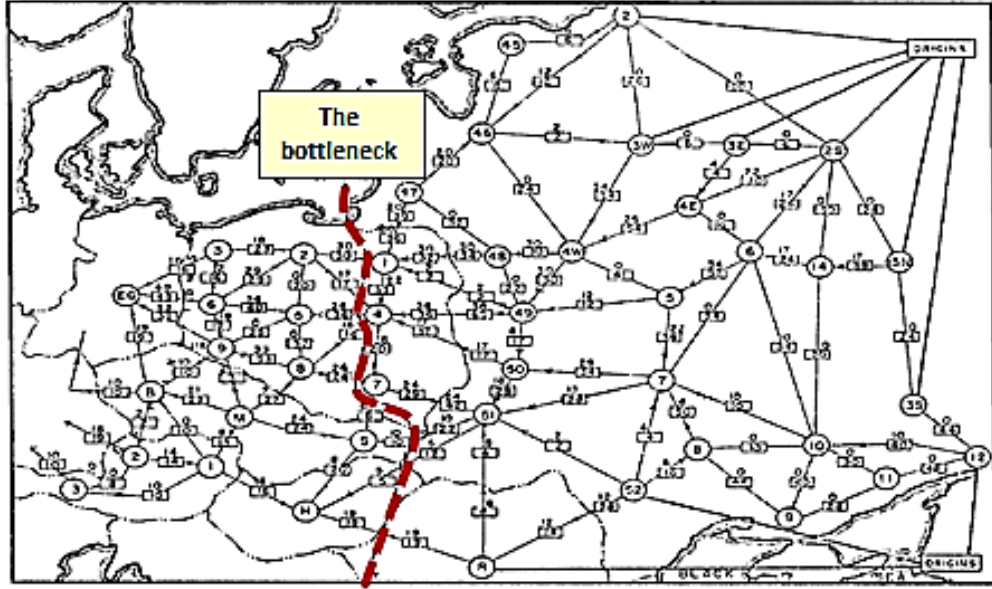


Figure 10: Diagram of railway network of Western Russia and Eastern Europe (Harris & Rose, 1955)

The first explicit simultaneous-play network interdiction game formulation¹³ is (Washburn & Wood, 1995). Unlike previous sequential-play games (Stackelberg Games), the interdicator and evader act simultaneously, or at least without knowing each other's strategy (known as a Cournot Game). (Washburn & Wood, 1995) present a two-person, zero-sum game for determining an optimal inspection strategy for single evader and a single (detector) inspection point. The evader determines a probabilistic path with minimal probability of detection while the interdicator maximizes his probability of detection with an arc-inspection strategy. This formulation is extended by (Unsal, 2010) to allow multiple inspector types (e.g., aircraft, ground-based inspection teams).

¹³ According to (Unsal, 2010)

Network infrastructure games have also been widely developed in two problem classes known as the Maximize Shortest Path problems (MXSP) and the Maximum Flow Network Interdiction Problems (MFNIP). These models generally assume predetermined attack effects and mitigations which are known by both players. Neither provides any mechanisms for explicit player learning, rather they provide worst case performance estimates for one-time assaults on the network infrastructure.

In MXSP, a network user wishes to traverse a shortest path from a specified starting node to a specified ending node in a directed or undirected network whose arc lengths (costs) are known. An attacker interdicts (destroys or lengthens) arcs to maximize the shortest-path. Contributors in this area include (Fulkerson & Harding, 1977; Golden, 1978; Isreali, 1999; Wevley, 1999; Israeli & Wood, 2002; Akgun, 2000).

MFNIP is another classic network interdiction model applied when the attacker's goal is to minimize the defender's throughput, isolate nodes, or otherwise diminish system function. It was originally motivated by efforts to destroy enemy supply lines during the Vietnam War. Ford-Fulkerson formulated the simplest of all interdiction problems where the interdictor breaks the source node from the sink node by eliminating all possible paths¹⁴ (Ford & Fulkerson, 1962). Wollmer extended this approach by limiting the available attack resources (Wollmer R. D., 1964; Wollmer R. D., 1970). There are several other alternative approaches to solving Wollmer's original problem, including dynamic programming approaches. (Steinrauf, 1999) provides a good review

¹⁴ This resulted in development of the max flow-min cut theorem, a basis for most network interdiction models. By interdicting the arcs in the minimum cut set, the maximum reduction to the network's maximum flow capacity is achieved.

of the early literature covering these developments. The MFNIP has been thoroughly studied to include significant contributions by (Wood, 1993; Cormican, 1995; Isreali, 1999) and recently by (Dhami, Pande, & Tamata, 2013).

Fundamentally, MXSP and MFNIP are foundational, but different from the problem at hand in that both these formulations model attacks made against the network's architecture to limit network functionality. In contrast, the problem under study in this dissertation is characterized by attacks against the vehicles and resources moving across the network where damage to the network infrastructure is not the primary concern.

Route Clearance Operations

Much of the specific analytical work on the IED problem remains classified by the Department of Defense, but to date there is no standard comprehensive modeling approach currently accepted to address the routing and scheduling of ground transportation under contested conditions. In the academic literature, several related models have been proposed.

(Washburn A. , 2006) presents a model in which he assumes IEDs are a low level concern to the defender in terms of the fraction of shipped material that is lost; therefore, defender movement patterns are determined by considerations other than the IED threat. Accordingly, the objective is to minimize the rate at which vulnerable classes of traffic (convoys) take lethal hits by intentionally using more resistant vehicle classes (RCPs) to either find the IED or suffer the attack. Thus, he states that the primary question is whether an IED will be removed by a dedicated clearance operation or by some other defender traffic. He uses a game theoretic to allocate route clearance missions according

to attack frequency. The model assumes a continuous process where traffic levels are given and known to both sides, and new IEDs are implanted at a known rate as old ones are removed. He also provides an alternative formulation for cases when defender losses become excessive (Washburn A. , 2006).

In a later paper, (Washburn & Ewing, 2011) explain the model rationale by observing that the defender's object is not to maximize the rate of IED removal by clearance units because the removal rate is the attacker's emplacement rate and the defender eventually removes every IED, one way or another. This perspective is continued in a third paper in which (Lin & Washburn, 2010) address the use of decoy IEDs. There are four essential assumptions made in these three route clearance papers:

1. Indefiniteness. The battle is assumed to proceed indefinitely.
2. Logistic ineffectiveness. The attacker's efforts are assumed to have a negligible effect on defender's logistic operations.
3. Independence. The various types of defender traffic and the attacker process of placing mines on roads are all assumed to be independent time-homogeneous Poisson processes.
4. Scalar Damage. Vehicles lost or damaged, cargo lost, people killed or wounded, and any other effects can be put on a single damage scale.

(DeGregory, 2007) provided a model that employs a two stage optimization approach for allocating a suite of force protection resources¹⁵ (FPRs) to guard scheduled logistics movements in an asymmetric environment. The algorithm does not specifically

¹⁵ FPRs include a fixed-wing aerial platform, armed helicopter platforms, motorized infantry platoons capable of performing route security and route reconnaissance, and convoy escorts.

provide route clearance mission planning, but considers the comprehensive use of shared resources (like aerial electromagnetic jamming assets) and dedicated resources (such as armed ground and airborne escorts). In the model's first stage, a convoy plan is generated by satisfying supply and movement requirements. The second stage is a binary integer program that determines the optimal employment of FPRs to the convoy plan. The resulting output is an overall convoy plan with integrated FPRs that produces the lowest expected number of casualties for an individual convoy. DeGregory represents attacker risk by integrating a probabilistic threat model with trend analysis and intelligence considerations, but acknowledges that adding a dynamic threat modeling technique is important. Finally, he offers a program development methodology for integrating his methodology into existing U.S. Army systems and processes.

Like DeGregory, (Marks, 2009) employs a multi-stage technique, but instead focuses on the route clearance scheduling problem. First he estimates IED activity on the road network as a two-state Markov process (similar to a queuing model). Then he performs column generation with an ADP algorithm to generate a set of feasible route clearance missions which are input to a mixed integer program. The result is a route clearance schedule and an associated risk-reduction measure. In practice, the model produces a solution characterized by a tendency to concentrate clearance efforts on a limited number of important roads which Marks suggests might be candidates for static counter-IED efforts, such as permanent observation posts.



Figure 11: Route clearance in Afghanistan¹⁶

(Kolesar, Leister, Stimpson, & Woodaman, 2012) address the importance of the attacker's IED emplacement tempo on the appropriate timing of the defender's clearance operations. They present a simple interaction model and analysis which asserts that the rate and timing of attacker IED emplacement substantially dictate the optimal route clearance schedule. They demonstrate that the more rapidly IEDs are being emplaced, the more sensitive is the timing between the clearance operation and follow-on traffic. Attack risk levels are determined from historical patterns which are reduced by the passage of dedicated clearance patrols. Then after some time, according to the attacker's emplacement tempo, the risk level returns to its original intensity according to a user defined "reseeding" function. This approach was employed by (Leister & Hudson, 2009) in a mixed integer programming algorithm which calculates optimal RCP's schedules (in terms of routing and timing) for a specified number of RCPs, given scheduled traffic

¹⁶ Source: counteriedreport.com

movements. They developed a mixed integer program and user interface that calculates a minimum IED risk movement schedule for a road network over a 24-hour horizon.

Attack Pattern Recognition

Bottom line—the only way to get ahead of the enemy’s decision cycle is to constantly and thoroughly analyze every scrap of information you can get your hands on and try to “see” patterns (U.S. Army Counterinsurgency Center, 2011).

In Iraq and Afghanistan, military commanders wanted to understand where IEDs would be encountered by their troops; however, consistent attack prediction was rarely achieved. Aside from the fact that such prediction is enormously difficult, much of the trouble is that data quality has remained generally poor and inconsistent. Since comprehensive data collection is difficult and time consuming, and because operations analyses often have not produce relevant, timely, and actionable enough products to support critical decision processes, military operators focused their energy on collecting information to support their immediate operational needs rather than to feed specific analytical processes (Shankar, 2014; Connable, Perry, Doll, Lander, & Madden, 2014). Then since analysts generally could not enter combat zones to collect their own data, a reinforcing cycle developed, discouraging the development of sophisticated analytical models.

Therefore straightforward pattern and trend analyses remain the predominant approaches for determining IED risk in the recent military campaigns. These types of analyses endeavored to use past IED discovery trends to predict future attacks. Frequently, past observations were the sole basis for IED activity depictions and threat

predications, occasionally bolstered by statistical methods and other computational tools (Connable, Perry, Doll, Lander, & Madden, 2014).

Beyond basic trend analysis, several models have been proposed to include stochastic Markov chains and game theory. Additionally, during the recent conflicts, the Joint IED Defeat Organization (JIEDDO) employed a Crime Pattern Analysis Team (CPAT), made up of mathematicians and law enforcement experts who developed predictive IED models based on crime analysis techniques (Shankar, 2014). Even so, the most common statistical representation in the literature has been point processes, most often the non-homogeneous Poisson process (NHPP).

A basic statistical understanding of temporal patterns and probabilistic structure of IED events is provided by (Kolesar, Leister, & Woodaman, 2008). This work showed that when aggregated across two broad regions and over two different twelve month periods, IED incidents followed a NHPP. This finding held for both time-of-day and day-of-week effects. But, when applied further it could not be generalized to predict more specific trends, cycles, or seasonality of IED incidents in other contexts (Kolesar, 2009). Further, (Shankar, 2014) found that the NHPP did not adequately characterize local (non-aggregated) IED discoveries during foot patrolling, primarily because the events exhibit spatial and temporal clustering rather than uniformity.

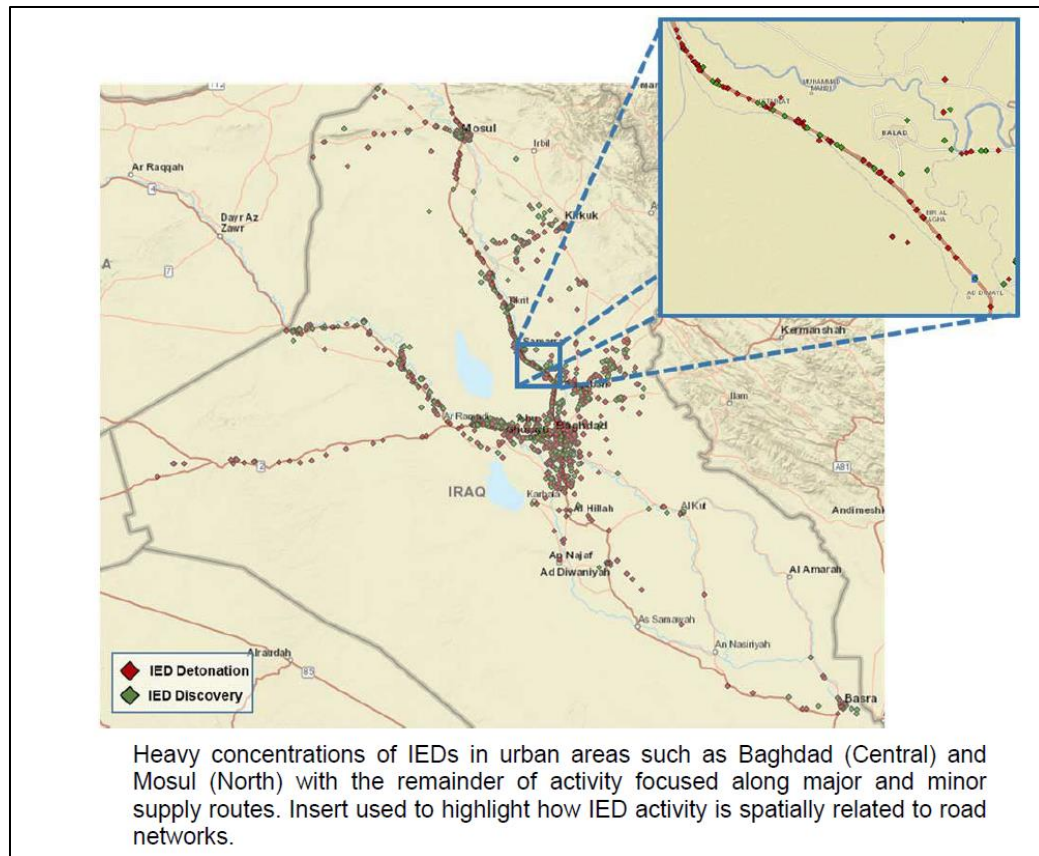


Figure 12: Example display of IED activity in Iraq during September, 2006 (Ardohain, 2016)

Recently, (Ardohain, 2016) also found IED patterns did not conform to a Poisson process when focusing on specific areas and short time periods. Specifically, he found that predictable IED patterns were most often local. This conclusion drove him to formulate three different IED prediction models based on the Hawkes point process, non-linear sine function optimization, and discrete Fourier transforms (DFT). All of these models used the inter-arrival time between IED events as the only model input and were tested against data from several well-defined geographic areas in Iraq and Afghanistan. Of the IED patterns analyzed, he found that the non-linear optimization and DFT models

both outperformed a mean inter-arrival time model. These techniques were used to distinguish IED attack patterns from randomness through a “test-two” methodology where, given N observations, the first N-2 inter-arrivals were used to predict the N-1 and Nth inter-arrivals. The results from this work were mixed, but suggested some ability to make attack predictions from identifiable discovery patterns. (Ardohain, 2016).

(Ardohain, 2016) continued by associating various observed IED discovery patterns with two broad conditions¹⁷ based on the supplies of attacker funding, materials, and labor for IED production. Without direct observation of insurgent logistics processed, this explanation for observed patterns appears speculative. Further, the categorizations fails to consider the influence defender activity patterns may have had on the attack patterns observed.

(Stafford, 2009) explored sequence pattern detection and time series analysis to develop predictive models for the timing and frequency of IED attacks. These models used historical attack patterns to identify trends and relationships for forecasting the number of monthly IED attacks based on aggregate force levels and holiday observances of Ramadan. He concluded that neither of these were major factors in predicting the number of monthly attacks (Stafford, 2009). But, these broad findings, based on aggregated environmental conditions, likely say little about localized attacker decisions.

¹⁷Condition one is unconstrained IED supply and condition two is limited supply broken into large supply, short supply, and steady supply categories.

Clustering

"I invoke the first law of geography: everything is related to everything else, but near things are more related than distant things." (Tobler, 1970)

Research addressing the use of spatial analysis has adopted the term “hot spot” to indicate areas with higher concentrations of disorder events. (Keefe & Sullivan, 2011) provide a formal definition of an IED hot spot:

- An area that contains a cluster of observations whose spatial dependence has been established using statistical testing; with a reasonable amount of confidence, it can be determined that the clustering pattern could not have occurred randomly, and
- The concentration of problem events in the cluster is greater than the average concentration of events in other parts of the study area.

Further, an actionable hot spot (AHS) is a hot spot of adequate size, shape, and sequence to justify the application of counter-IED resources. Essentially, an AHS uses recent temporal and spatial IED-related activities to detect clustered patterns that are expected to indicate continued threat in the immediate area (Keefe & Sullivan, 2011).

The AHS prediction approach was tested on historical data in support of six Army brigades and one logistics unit in Baghdad, Iraq. It produced variable results with better prediction performance in areas where clustering occurred, but results were poor

elsewhere. Overall the tool provided a five percent improvement compared to results without the tool¹⁸ (Connable, Perry, Doll, Lander, & Madden, 2014).

Accounting for Defender Activity

A weakness common to every approach reviewed thus far has been they have relied nearly exclusively on defender IED discovery data without endeavoring to account for influences of the defender's activity on the IED emplacement and discovery patterns. Figure 13 and Figure 14 are such examples. These depictions show EIAs temporally and spatially without any reference to defender movement patterns. While this is useful for understanding past trends, it is of limited value in determining how future operations might be adjusted to make improvements. Thus, analyses based solely on such data must implicitly assume either the defender's activity will remain essentially constant across the analysis period or the attacker's activities are independent of the defender's. Clearly, the latter can only hold if the attacker is not a learning adversary engaged in a deliberate planning process. Further these approaches make several lesser unlikely assumptions such as consistent attacker effort and defender IED detection probabilities across time, space, and environmental conditions.

The issue at hand is how much difference will including additional explanatory factors improve understanding and performance in attack detection, prevention, and/or prediction. (Ardohain, 2016) concludes that such transient factors cannot be ignored based on his finding that discernable IED discovery patterns generally do not endure for sequences of more than twenty-five events as the underlying process dynamics rarely

¹⁸ Over the period of the test, the tool proved to be accurate 30 percent of the time which is 5 percent better than results obtained without the tool.

endure for extended periods. He adds that any series of less than six events is not enough to make reasonable predictions. So a rapid learning approach is needed. This concern is also expressed by Shankar who stated that although the study of the attacker's IED emplacement behavior is a critical, it has had little focus to date (Shankar, 2014).

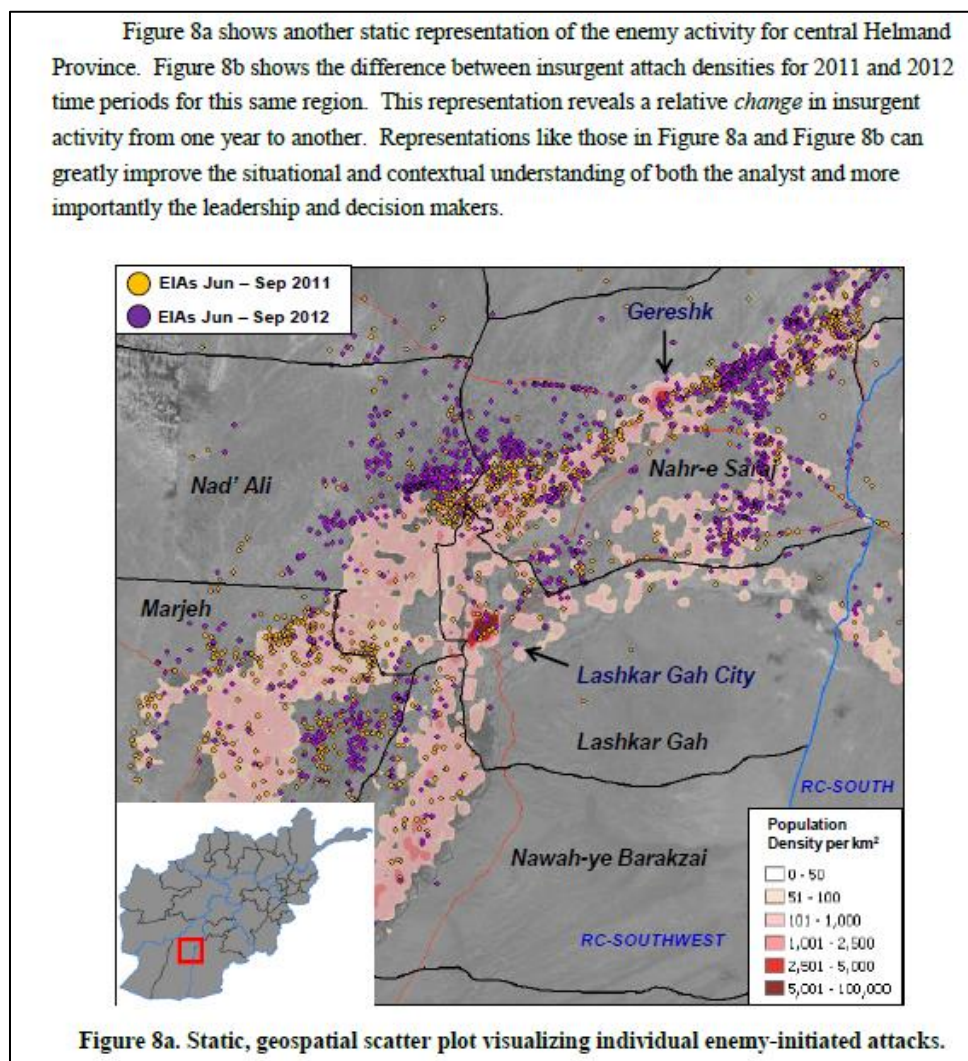


Figure 13: Visualization and description of EIAs in Afghanistan (Center for Army Analysis, 2016)

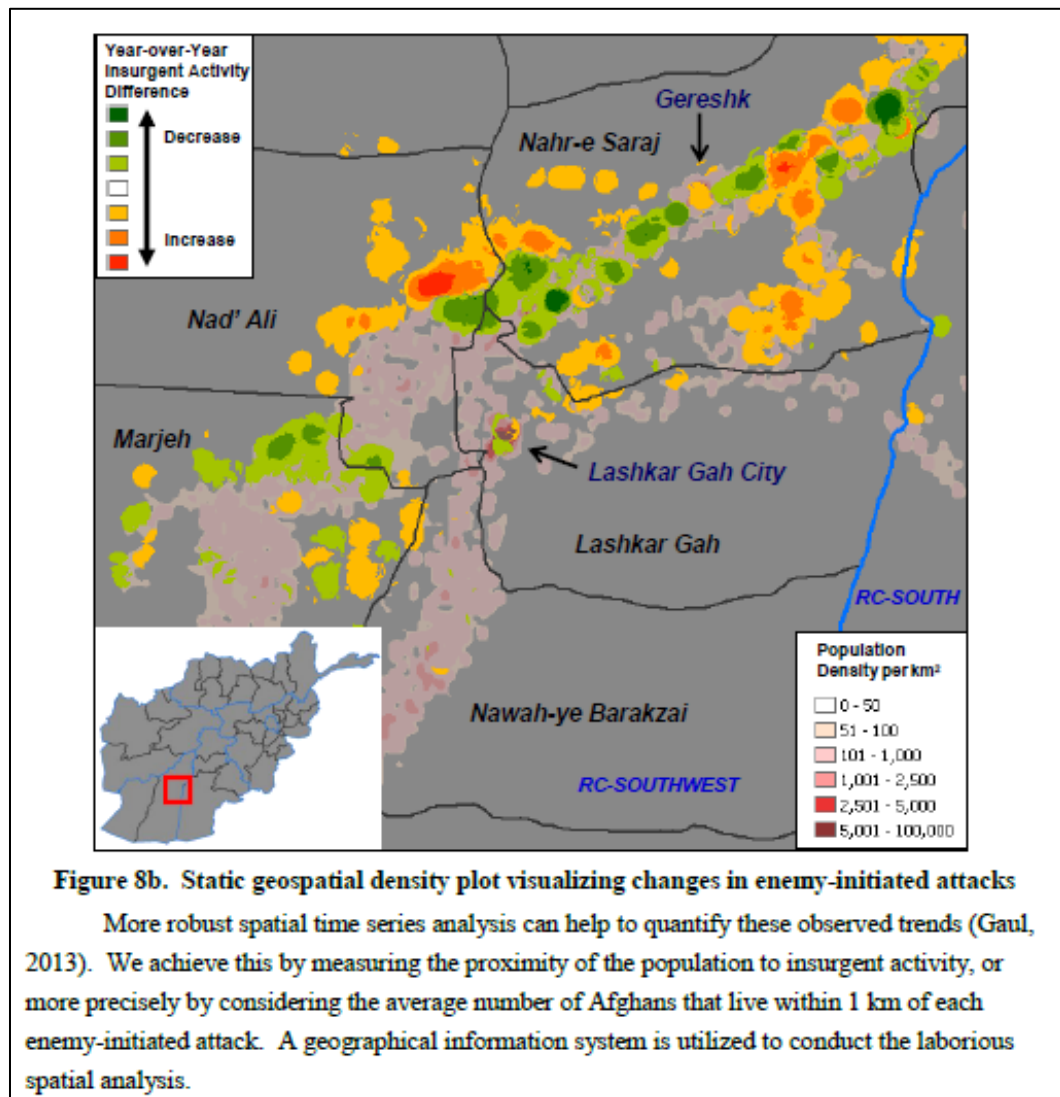


Figure 14: Example EIA pattern analysis and description for Afghanistan (Center for Army Analysis, 2016)

The common reliance on basic, one sided analytical approaches has occurred primarily because the highest quality data available to analysts has historically been

reports of the attacker activities (referred to as enemy significant actions, or SIGACTS¹⁹). While some analyses have incorporated defender activities (in the form of Blue Force Tracker (BFT) data²⁰) with more sophisticated modeling, these efforts have been limited because such information has generally been incomplete and inconsistently available (Connable, Perry, Doll, Lander, & Madden, 2014; Shankar, 2014; Ahner & Spainhour, 2015; Ardohain, 2016)

Koyak incorporates BFT data in a group of models he developed in a series of three papers. In these he discusses how the observable IED discovery process can provide insights and predictions related to the unobservable emplacement process. He describes several statistical estimation models which explicitly recognize the separate roles of the attacker observation and action processes. He represents IED emplacement activity as either a NHPP or extensions of historical patterns (Figure 15 and Figure 16) and calculates the probability of encountering an IED on a particular stretch of road based on both previous IED discoveries and defender traffic patterns (Koyak, 2009a). These models are of varying complexity and can be tailored to the user's situation. They share the following basic attributes (Koyak, 2009a; Koyak, 2009b; Koyak, 2010):

- The event of interest is the time and location of IED emplacement, not the time of the defender's IED discovery

¹⁹ CIDNE (Combine Information Data Network Exchange) is the USCENTCOM (U.S. Central Command) database of record for retaining SIGACTS. CIDNE was originally adopted during by the Multi-National Force – Iraq in 2006 (Center for Army Analysis, 2012).

²⁰ BFT is a GPS-enabled system that provides real-time friendly force location information.

- Discovery times are right-censored emplacement times (because discovery times are observable and necessarily greater than the corresponding, non-observable, emplacement time)
- Discoveries occur mainly during defender road traversals
- Discovery is a random event, with probability that can vary according to factors such as the IED type, vehicle type, and vehicle speed

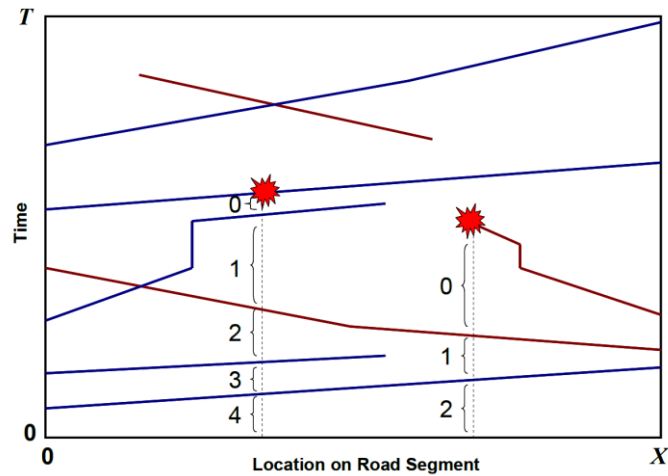


Figure 15: Representation of unequal discovery probabilities. Red and blue delineate the direction of travel. The numbers are counts of traversals associated with the two discoveries shown (Koyak, 2009b).

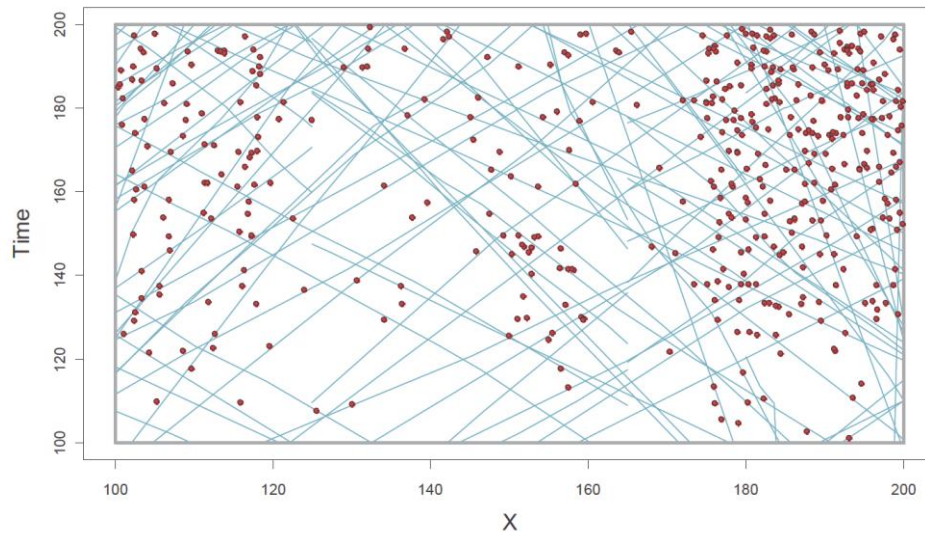


Figure 16: Illustration of 436 randomly-generated emplacement events, illustrating the complexity of the operational environment (Koyak, 2010)

Recently (Ahner & Spainhour, 2015) incorporated BFT data into a logistic regression methodology for analyzing time-dependent factors relating defender presence to IED discoveries. They propose an approach for providing improved planning and attack prediction by understanding the effect of spatial and temporal BFT density on IED discoveries. They also analyze time of day and cache discovery effects. They don't explicitly examine activity patterns, but nonetheless develop a useful method of incorporating aggregated BFT data into pattern analyses that improves the explanatory power of their regression models.

Highlighting the data availability issues, Shankar conducted his own unique, onsite data collection effort to model IED activity against dismounted (or foot-mobile) patrols in Afghanistan. The dismounted patrol problem is distinctively different from

mounted (and convoy) problem because foot-patrol routes are not confined to distinct network paths as vehicles are to roadways. The increased variability of foot paths means specific routes are not repeated regularly and there is an increased probability that any particular foot-patrol will not encounter some particular IED.

In this work, Shankar gained the cooperation of active combat units to collect individual patrol reports which he used to develop a spatial cluster model, an IED emplacement model, and a simulation to describe and predict the IED activity being observed. Like Koyak, he focused on modeling emplacement time rather than IED discovery, using an overlapping travel zone methodology and probabilities related to individual patrols encountering individual IEDs (Shankar, 2014). While this level of detail is indispensable, the aspect of structuring defender activities was not addressed.

Addressing Current Methodological Shortcomings

The current logistic interdiction literature addresses several important aspects of the problem, but it does not directly consider the critical role of opponent observation, reactivity, and learning from experience. Here, it is critical to recall John Boyd's insight, connecting observation to action via orientation and human decision making. The unique nature of the network counter-IED problem is that it is ongoing and repetitive with opponents who have asymmetric information, means and goals. The common assumption of independence between attacker and defender as in (Washburn A. , 2006;

Lin & Washburn, 2010; Washburn & Ewing, 2011) is an important shortcoming that ignores this critical problem feature²¹.

The attack pattern recognition works cited do not provide a method for understanding how variations in activity sequences and movement patterns might change the attack patterns observed. In contrast, the modeling approach we are pursuing focuses on the critical role of attacker observation on attack decisions. It seeks to learn attacker preferences, to the extent they are exhibited, by making the explicit assumption of dependence between the defender's actions and the attacker's choices. To assume independence is to assume the attacker has no specific preferences of target type. Further, it assumes the attacker does not react to variations in target activity which is to assume attacker indifference and the absence of specific attacker goals.

Dynamic Programming (DP) is a method for solving sequential decision problems that can be expressed as Markov Decision Problems (MDPs). While its discovery provided a revolutionary conceptual framework, the classic DP approach suffers from two principle draw backs that ADP helps to mitigate. The first is known as the "*curse of modeling*." This so-called curse refers to the difficulty of knowing state transition probability distributions related to outcomes from agent decisions. This curse relates firstly the knowing the probabilities related to expected outcomes given an agent choice such that they can be enumerated in a transition probability matrix (TPM). And secondly, the expected reward (or penalty) received after any given system state

²¹ There is a significant branch of research focused on projecting attacker cognition (in the form of preferences and objectives) to determine likely attack scenarios. This psychology approach is distinctly different from the one taken here which is strictly concerned with observable actions without regard for motivation.

transition has occurred. In most real world applications, these two types of probability are not known and must be learned through trial and error. Additionally, since the underlying model is an MDP, the systems states are “memoryless;” meaning these transition and reward probabilities depend solely on the observed state and the decision being made. Thus, they are independent of any prior states or agent decisions (Denardo, 2003).

The second principle drawback of classic DP is the well-known “*curse of dimensionality*.” This stems from exponential growth in problem dimensionality when attempting to expressly enumerate multiple dimensions and levels in a problem’s state space, outcome space, and action space. In practice, DP generally requires each of these three spaces be defined by a vector of attributes that can quickly grow and become intractable. For example a customer’s inventory might consist of N different products that can be held in M different quantities. This means there are M^N different inventory states.

Even though the DP method provides guaranteed optimality in polynomial time, which is exponentially faster than exhaustively searching each possible decision policy to provide the same guarantee (Sutton & Barto, 1998), and while not all problems suffer from all three sources of dimensionality growth, the curse of dimensionality is the most commonly cited reason for why DP is not usable in many real world applications (Powell, 2007).

Therefore, ADP has emerged as a powerful technique for solving complex RL problems as a means to mitigate the curses of modeling and dimensionality. This is accomplished principally through use of a transition function (rather than a TPM) to

model agent behavior and interaction with the environment. Classic ADP techniques are described by (Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998) and a comprehensive survey with more than 100 references is provided in (Gosavi, 2009). Further, (Powell, 2007) provides comprehensive strategies for further reducing dimensionality through use of post-decision state values and (Balakrishna, 2009) presents a diffusion wavelet treatment to further address the curse of dimensionality. These improve the means for solving an increased number of operationally relevant problems.

ADP has several appealing features that make it an appropriate choice for the competitive problem under study; chiefly that it easily incorporates the stochastic nature of both the environment and the outcome of the agent decisions without having pre-defined transition probabilities. They are learned through the algorithm by trial and error. This approach is not yet common in the transportation literature, but (Powell, Simao, & Bouzaïene-Ayari, 2012) offer a unifying framework for applying ADP to transportation problems which we employ in our model.

Where DP provides a provably optimal decision policy in a limited number of problems, ADP is effective for discovering good overall policies (not necessarily optimal) in a much wider array of problems, where optimality cannot be readily achieved through any other OR techniques.

CHAPTER THREE – MODEL DESIGN AND FORMULATION

To date, despite vastly improved counter-IED technologies, the literature reveals that little has been done to develop, publish, and establish improved modeling concepts for sustained logistics under contested conditions. Many solution methods to dynamic routing and scheduling problems have ignored current knowledge to forecast the future because solutions to the VRP are already very difficult (Spivey & Powell, 2004). As such, a central goal of this formulation is to keep the vehicle assignment and routing problem as simple as possible to focus the computational effort on the challenge of learning from the opponent interaction.

In order to determine the best way to schedule route clearance and other vehicle movements under contested conditions, there are two objectives any complete logistical model must be meet. The first is to satisfy the operational requirement of moving troops and cargo throughout the operational area. This involves solving some version of the VRP. The second is the minimization of loss and damage from ambushes which requires an operational scheme to counter the attacker. But, it is not enough to simply optimize the immediate (myopic) choices of what, when, where and how to move across the road network based on static risk assessments, scripted attacker behavior, or long run averages. Rather, we claim the goal should be to synchronize individual movement

decisions across the fleet of vehicles and entire planning horizon, accounting for downstream attacker-defender interactions.

Our model represents opponent interaction by assuming dependence between attack probabilities and targeted traffic patterns. There are currently few analytical approaches explicitly make this assumption, but RL algorithms offer opportunities for meaningful improvements in this area. To our knowledge this approach has not been pursued anywhere in the OR literature related to this problem. Our goal is to effectively address the attack problem by crafting vehicle movement schedules that not only satisfy the military distribution problem, but also maximize the defender’s performance, given the attacker’s reaction to changes in defender activity patterns. The model herein is an initial step in this direction.

Model Description

Since the VRP is already difficult to solve in its own right (Hinton, 2010)²², the general approach taken will be to greatly relax common VRP constraints (such as those related to distance traveled, crew endurance, fuel cost, etc.) in order to provide maximum flexibility and free computational effort to crafting vehicle movement plans that minimize successful ambushes. This is accomplished by choice of convoy size, convoy and RCP route assignments, and activity timing. Thus, we employ a full truckload, single stop, route selection formulation of the VRP, as these are relatively easy to solve (Ropke, 2005).

We make four principle assumptions in this model formulation:

²² VRP and NP-completeness are discussed under Vehicle Routing, on page 26.

1. Infinite horizon: A clear characteristic of the network ambush problem we are addressing is its repetitive nature with ongoing, nearly constant interaction between the attacker and the defender. We are not endeavoring to predict rare events, but rather systematic and sustained attack situations. Thus, the infinite horizon assumption is quite natural.
2. Dependence: Assuming dependence between the attacker's and defender's actions is a key distinguishing feature of our model. It is taken to mean that the network defender cannot escape the direct observation of the attacker while conducting activities across the network. As such, we assume that the locations, types, and frequency of defender traffic bears directly on the attacker's ambush decisions. In practice this interaction must be continuously learned over time and this model is a foundational step toward potentially developing a family of RL algorithms to more fully address learning opponents in similar circumstances. Thus, this formulation, explicitly assumes dependence between the defender's actions and the attacker's ambush choices. The military theory underlying this assumption is discussed in (Stimpson, 2011).
3. Scalar function cost and reward: All values are measured on a single scale (e.g., delivery rewards, operating costs and damage penalties). Variations in the value scheme can easily be tailored to a different modeling objective. For example, since the objective function minimizes total operating costs, the penalty for unmet customer demand can be increased or decreased relative to

the other penalties and benefits to vary the policies' responsiveness to customer demand.

4. Latent IEDs: We assume every IED ambush established by the attacker is discovered by the defender with some probability of damage. Thus, there are no latent (undiscovered and unexploded) IEDs remaining on the network roadways following defender traffic passage and IEDs do not accumulate.

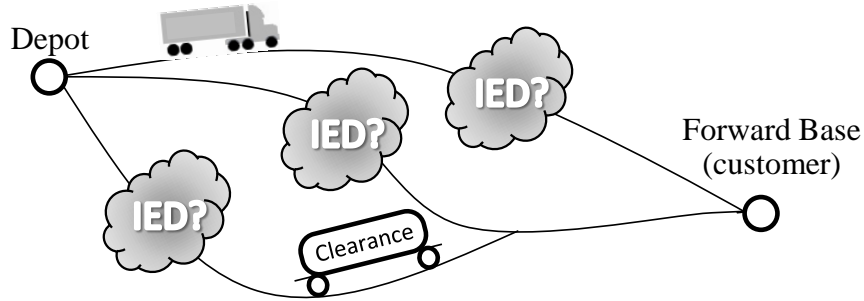


Figure 17: Problem illustration

Model Formulation

Our goal is to present a simple, but generalizable, model in keeping with current military operational concepts. Thus, as Figure 17 shows, in our basic formulation we model a single depot supplying an outlying operational base (customer) across a network of preplanned network paths (routes). As such, we define a graph of multiple undirected paths between a single depot and customers where the customer locations and routes are fixed. This is consistent with an established system of bases and approved routes which are typical for military distribution (see Figure 18 and Figure 19).



Figure 18: Supply routes in Iraq during spring of 2009 (Center for Army Analysis, 2012)

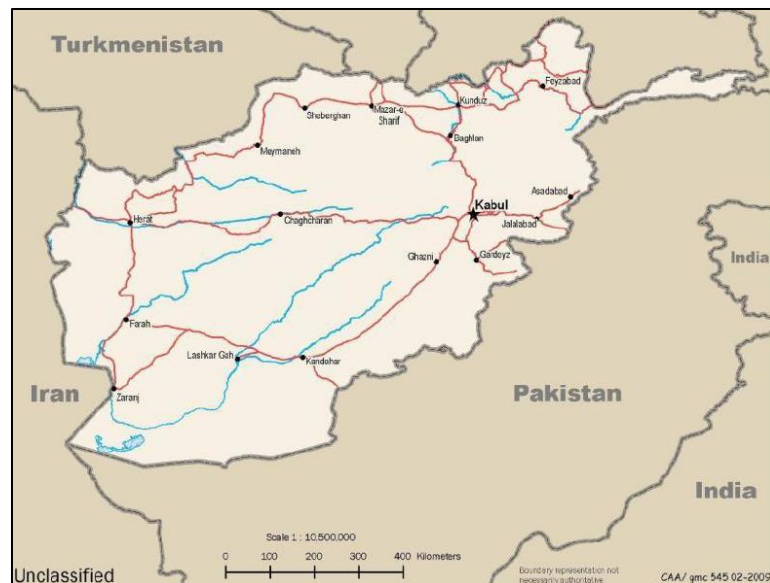


Figure 19: Supply routes in Afghanistan during spring of 2009 (Center for Army Analysis, 2012)

In our formulation each time step is an interval beginning at time t . The agent is a decision maker that observes the system at equally spaced time intervals $0,1,2, \dots$ without limit. At each time interval, the agent observes the system which can only be in one of (N) states, defined by the system state variables. The transition from state to state is governed by chance based on the agent's decision, stochastic system state changes, and stochastic exogenous information that may or may not arrive. (P_{ij}^k) is the probability that if the agent makes decision $k \in K$ in the current system state (i) , the agent will observe system state (j) in the next time period. Each state transition generates some immediate, one-step reward (calculated according to the one-step cost function) (C_i^k) , which can be negative (i.e., a penalty). State transitions occur with probability 1.0 as shown in Equation 1 (Denardo, 2003). In reality the agent does not know the transitions probabilities a priori, rather they must be learned through observation by trial and error.

Equation 1: Transition probability rule

$$\sum_{j=1}^N P_{ij}^k = 1 \quad \forall i, k$$

At each time step, the agent seeks to maximize the total reward (or minimize the total penalty) which is the sum of the immediate reward for the current decision (k) and long run discounted value of all future (downstream) rewards it expects to realize over the entire planning horizon. Thus, at each time step, the agent faces the tradeoff decision between maximizing its current reward against the long range payoff for future rewards that can be realized as a consequence of the current and future decisions. As such, the agent learns the best decision for each system state in which it may find itself. This

provides us with a state–dependent decision rule (or function) that we refer to as a *policy* (Powell, 2007; Denardo, 2003).

Equation 2: Reward bounds in any given time step

$$m = \min_{i,k} \{C_i^k\} < C_i^k < \max_{i,k} \{C_i^k\} = M$$

Per Equation 2, when the problem can be solved explicitly, the immediate reward available to the agent in every state is between the minimum and maximum possible reward. Then, since we apply a discount factor, $0 < \gamma < 1$, to all future rewards, the present value of the entire future decision stream is between $m/(1 - \gamma)$ and $M/(1 - \gamma)$. As such, the reward available to the agent in any given state is according to Equation 3 which is no more than the sum $M/(1 - \gamma)$, of the geometric series $M + M\gamma + M\gamma^2 + \dots$ (Denardo, 2003) .

The value of γ determines how “greedy” the agent will act, in balancing its choice of current verse future rewards. When $\gamma = 0$, the agent only values the current payoff without regard for the future consequences of the current choice. Increasing the value of γ causes the agent to increasingly weigh the value of future system states, in which it may find itself, as a significant component of the current choice. If $\gamma = 1$, the agent will value all future states equally.

Equation 3: Long run total value of rewards (Stewart, 1999)

$$Total\ Reward = \sum_{t=0}^{\infty} \gamma^t C_t \quad 0 < \gamma < 1$$

We say the model is in *discrete-time* since agent decisions are made at each time step and that it is *finite* since the state and decision spaces are defined as finite sets.

These characterizations, together with our definition of the transition probabilities, make this model *Markov* (Denardo, 2003).

In order for the system state to be effective and informative, the state variables must provide an adequate means for calculating the subsequent state transition and the value of future rewards. As with the all MDPs, transition probabilities and rewards are functions of the current state variables and the action choice without regard for how the current state was reached. Thus, except for the historical information held in the state variables themselves, the system state is said to have the Markov property, i.e., it is *memoryless*, or independent of the path by which the system arrived in its current state (Sutton & Barto, 1998). Since historical context is necessary in this setting, we define an information state that gives the agent limited historical data within the system state definition.

Decision Variable

(P) is the set of assignable network paths for convoys and RCPs. We define the network such that each path from the depot only services one customer $h \in H$. While alternative architectures are available, this definition simplifies the formulations and is adequate for the intended application (see Figure 20). Additionally, as discussed previously, historical geospatial analysis shows that IED attack regions have been localized, with activity in confined clusters or hot spots (Keefe & Sullivan, 2011), which are generally limited to an attacker's sphere of influence.

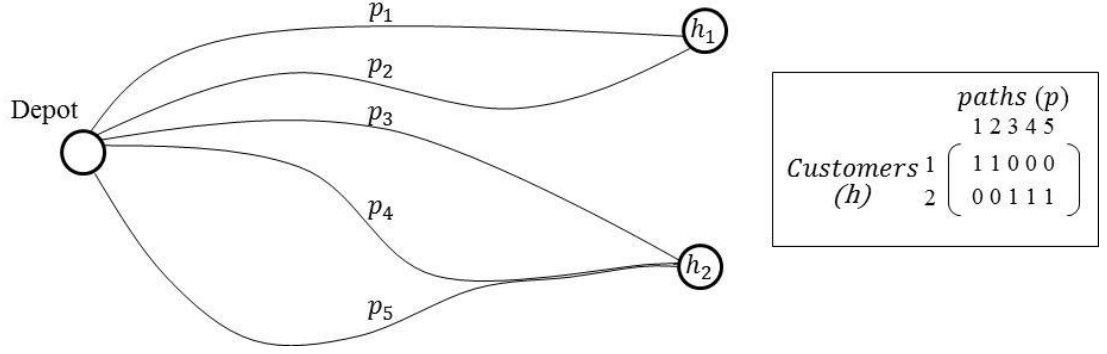


Figure 20: Example network and adjacency matrix

In general we expect P to be the same for convoys and RCPs, but (P) can be indexed by traffic type according to constraints such as the accessibility of certain routes to certain vehicle types or convoy size limitations that may be path specific. In our base case, we model only two classes of vehicles, Convoys and RCPs, but additional classes can be defined as required.

From any feasible system state, the agent chooses from the set of (X) feasible actions at time (t) , where $x_t \in X$ as determined by the system state such that $q_p \in Q$, and $l_p \in (0,1)$. $Q = \{0,1,2, \dots q^{max}\}$ is a set of feasible convoy sizes in homogeneous “trucking units” that can be shipped on any path $p \in P$. At each time step, feasible agent actions are to ship some feasible number of supply units on each path, and/or assign route clearance to some feasible number of paths as described by the vector (x_t) . This yields Equation 4.

Equation 4: Agent action vector

$$x_t = (q_{t,1}, q_{t,2}, \dots, q_{t,P}, l_{t,1}, l_{t,2}, \dots, l_{t,P})$$

$(q_{t,p})$ is the choice of convoy size (or quantity) to be shipped from the depot to the customer on path (p) at time (t) . When $q_{t,p} = 0$, the action is to ship zero units, that is, do not ship anything, on path (p) (we also refer to this choice as “wait”). Then, $(l_{t,p})$ is the RCP path assignment, where $l_{t,p} = 0$ is the action of choosing not to clear path p . A simultaneous assignment occurs when $q_{t,p} \geq 1$ and $l_{t,p} = 1$ for any (p) at time (t) (see Figure 21). In practice, this is understood to be a direct RCP escort of the convoy on the assigned path. Defining the action vector in this way, only allows zero or one convoy and/or RCP assignment per network path in each time step. But this is not an actual limitation in practice as the agent can vary the convoy size as desired.

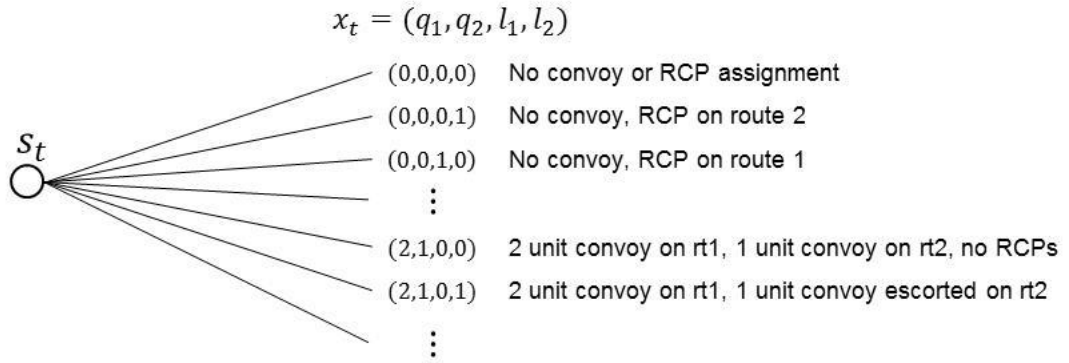


Figure 21: Example of defender action choices for some state (S_t) in a two path case

System State

The state variable is a minimally dimensioned function of history that is necessary and sufficient to compute the decision function, the transition function, and the

contribution (or cost) function (Powell, Simao, & Bouzaiene-Ayari, 2012). It defines every particular system state. In order to discover good decision policies, the state definition must well represent conditions under which each decision is to be made. This requires representing historical detail from recent activity patterns and interactions such that useful projections, conditioned on the current decision, are possible. We compose the system state variable (S_t) with three components, the resource state (R_t), information state (I_t), and knowledge state (K_t) as shown in Equation 5:

Equation 5: State variable definition

$$S_t = (R_t, I_t, K_t)$$

The resource state (R_t) includes customer inventory levels and RCP readiness status. The information state (I_t) maintains attacker and defender activity information in the form of density measures. While, the knowledge state (K_t) is the current belief about the expected outcome for every feasible decision (x_t) with respect to expected customer demands and attack probabilities.

Resource State

The resource state is defined as $R_t = (u_{t,H}, j_{t,n})$, where $u_{t,h} = [0, u_h^{max}]$ defines feasible inventory levels for each customer $h \in H$ at time (t) and $j_{t,n} = [0, j_n^{max}]$ is the readiness state of RCP (n) at time (t).

Following every RCP assignment, there is a recovery period required before the RCP can be reassigned. Therefore, j_n^{max} is the RCP state immediately following an RCP usage and its value decays as function of (t), until $j_{t,n} = 0$. Then the RCP is available for reassignment. All RCPs are held in a central inventory and can be individually assigned

to any network path so long as they are in the recovered state. Customer inventory and RCP state transitions are defined as follows:

Equation 6: Customer inventory update

$$u_{t+1,h} = \max\{0, u_{t,h} + q_{t,h} - \hat{d}_{t,h}\}, \quad \forall h$$

Equation 7: RCP inventory update

$$j_{t+1,n} = \begin{cases} j_n^{\max}, & \text{if } l_{t,p} = 1 \text{ (RCP is assigned to path } p\text{)} \\ \max\{0, j_{t,n} - f(t)\}, & \text{otherwise} \end{cases}$$

Here, $(\hat{d}_{t,h})^{23}$ is the stochastic demand from customer (h) in the current time step which only becomes known after the agent decision (x_t). Thus, the agent schedules convoys in anticipation of customer demand rather in reaction to immediate demands. If the observed demand in the current time step is lower than expected, the system state will reflect increased inventory in the next time step. Over time, in the absence of some additional knowledge, the observed historical demand will be the basis by which the expected demand distribution is updated. But if other information is available, like an anticipated change in force level or operational requirement, these can be used to update the agent's expectation.

We treat supply as homogeneous to minimize the state space dimensionality, but the addition of different supply classes can be accommodated. In practice this is likely unnecessary as shipping capacity is generally fungible within supply classes and vehicle types (like fuel verses bulk cargo); thus, this treatment provides adequate resolution for most applications. Also, since the customer has storage capacity, immediate customer

²³ The “hat” (e.g., $\hat{d}_{t,h}$) denotes exogenous information arriving from outside the system and not within the agent's control.

demand can be filled by on hand customer inventories. This allows the convoy delivery schedule to be flexible, providing opportunities to use the most advantageous schedule and path choices to satisfy demand with minimum operational costs and attack penalties.

Decision Constraint

Now that $(u_{t,h})$ and $(q_{t,h})$ have been defined, we note an agent decision constraint we apply to improve solution times by minimizing the agent decision space; namely we do not allow the agent to schedule a delivery quantity $(q_{t,h})$ that will exceed any customer's storage capacity (u_h^{max}) . In practice this is unrestrictive, so long as the customer capacity exceeds the max shipping capacity. Alternatively, the model could be constructed to allow excess shipments with a penalty when the sum of current inventory and deliveries exceeds customer storage capacity.

Equation 8: Delivery decision constraint

$$x_t = (q_{t,1}, q_{t,2}, \dots, q_{t,P}, l_{t,1}, l_{t,2}, \dots, l_{t,P}) \quad S.T. \quad u_{t,h} + q_{t,h} \leq u_h^{max} \quad \forall h$$

Information State

The information state (I_t) must maintain enough historical, environmental, and pattern data to make associations between defender actions and attacker preferences. From a dimensionality perspective, this is the most challenging aspect of this problem since there is virtually no limit to the number of environmental factors that might be recommended for inclusion. Additional techniques and strategies to address this challenge are outside the scope of this dissertation, but are discussed further in under Future Research (page 110).

From the information state comes the agent's current attack expectation (defined in the knowledge state below). We define a simple information state, $I_t = (b_{t,P}, e_{t,P})$ where $(b_{t,p})$ is a measure of defender traffic density on network path ($p \in P$) at time (t) and $(e_{t,p})$ is the measure of the concurrent attack density on network path ($p \in P$). Both density states are calculated as measures of the recent activity on each path across the network, updated after the current agent (defender) decision (x_t) and the observed attacker response:

Equation 9: Traffic density update

$$b_{t,p} = b_{t-1,p} + f(x_t) \quad \forall p$$

Equation 10: Attack density update

$$e_{t,p} = f(e_{t-1,p}, a_{t-1,p}) \quad \forall p$$

As defined, the current agent decision updates each path's traffic density measure ($b_{t,p}$) prior to decision (x_t), while the previously observed attack behavior is used to update the attack densities ($a_{t-1,p}$) for each path. Also, note we make no distinction between convoy and RCP traffic to reflect an assumption that the attacker is equally likely to attack all defender traffic, but this assumption is easily relaxed.

Depending on what constraints are applied; the decision space is relatively small compared to the information state space because the action choices available to the defender in any time step are limited to the timing and volume of shipments and the assignment of route clearance (previously shown in Figure 21). For example, if there are five network paths, ten convoy size options, and at least five RCPs, then the unconstrained agent has 3.2 M possible actions ($10^5 \cdot 2^5$). But, if we don't allow

unrealistic choices like using all paths simultaneously (i.e., sending separate convoys on all five paths in the same time step), but instead constrain the agent to just one convoy per time step, then there are only 320 action choices ($10 \cdot 2^5$). This strategy significantly reduces dimensionality, allowing computational resources to be focused on the most likely actions the agent will take.

Meanwhile even the minimal information state space calculation we employ is significantly larger. For example in the same road network, if we have two customers with ten inventory levels, three RCPs with a four time period recovery function, and five traffic and attack density states, there are 12,500 resource states ($10^2 \cdot 5^3$) and 9,765,625 information states ($5^5 \cdot 5^5$), yielding over 122 billion total system states. Thus, the information state cannot be enumerated, but must be calculated as a function in any realistic application.

Knowledge State

While we are implementing a simple attack preference model in this formulation, obviously an attacker behavior model can become quite complex, incorporating several dimensions, such as observed attacker preferences related to resource availability, environmental, spatial, temporal, seasonal, historical, and other features. Further, these models can range from a probability of attack based simply on event counts and trends analysis to numerous other prediction measures such as those discussed under Attack Pattern Recognition (page 36) and Future Research (page 110).

Recall we assume the IED ambush is the attacker's prediction of the future. Thus it is our primary task to understand the basis of this prediction, which is what the attacker

observes about the defender's actions. The extent this can be accomplished is the extent to which improved predictions can be crafted. Therefore, the knowledge state necessarily includes both the customer demand and attack expectations, which are informed by the information state, according to the action choices at hand.

The knowledge (or belief) state, $K_t = f(x_t, I_t)$, defines the agent's expectation of customer demand and being attacked given each feasible action it can take, including the discounted downstream expectation. Thus, we calculate the expected probability of attack component according to the current traffic and attack densities, independent of customer demand.

Attack Probability

We conceive of a fixed and variable attack probability in acknowledgement that there is always some probability of attack when traveling through a high risk area, but we posit this risk increases with repeated path usage. As such, we incorporate this idea with a simple, but reasonable, path specific definition for probability of attack:

Equation 11: Path probability of attack

$$P[\hat{a}_{t,p}] = P[\hat{a}_{t+1,p}^{fixed}] + P[\hat{a}_{t+1,p}^{variable} | x_t, b_{t,p}, e_{t,p}] \quad \forall p$$

Here we assume without regular defender traffic (i.e., $b_{t,p} = 0, \forall t$), the attacker does not anticipate the defender; therefore, does not deliberately prepare IED ambushes so $P[\hat{a}_{t+1,p}^{variable}] = 0$. But, as traffic density increases, so will the attacker's expectation of ambush opportunities which can be emplaced in the lengthy temporal gaps between the occasional defender transits. Then, we postulate, as defender traffic volume continues to increase, the shrinking temporal gaps between defender transits provide less

emplacement opportunity and begin to increase the attacker's risk of being interdicted (caught in the act of emplacing); therefore, decreasing his attack opportunities and the probability of attack.

In practice $P[\hat{a}_{t+1,p}^{variable}]$ would not be defined a priori, but must be learned over time on every network path. Figure 22 shows two conceptual examples. We note here that there are other forms of defender activity we do not model (to include various forms of surveillance) that might affect the attacker's perceived emplacement opportunities to the extent he is sensitive to them. While these could be included in the formulation determining the expected attack probability, they are outside our immediate treatment.

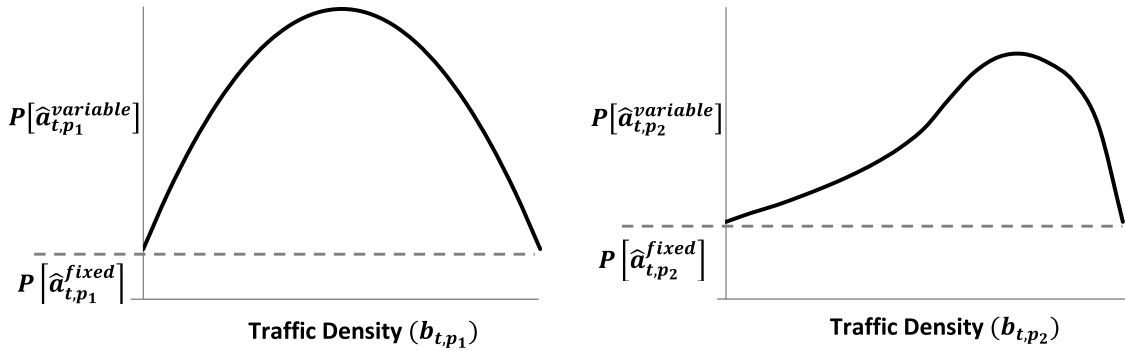


Figure 22: Example attack probability functions

This means a decision to move along a certain path may increase or decrease $P[\hat{a}_{t,p}]$ depending on (I_t) . Further, the agent's current travel decision not only bears on

the current knowledge state (K_t), but it directly affects the downstream knowledge states (K_{t+1}, K_{t+2}, \dots) as well.

It should be noted that a consequence of assuming dependent attack probabilities is that the defender has a level of control over the attacker's emplacement actions. This is an important distinguishing feature of our model that relies on a robust information state and learning from correlations between defender traffic frequency and the observed attacks. The attack probability transitions are calculated as follows:

Equation 12: Fixed probability of attack for each path

$$P[\hat{a}_{t+1,p}^{fixed}] = f(e_{T,p}) \quad \forall p$$

Equation 13: Variable probability of attack for each path

$$P[\hat{a}_{t+1,p}^{variable}] = f(x_t, b_{t,p}, e_{t,p}) \quad \forall p$$

This path-based treatment assumes the attacker does not discriminate between defender traffic types. To the extent they can be learned, a simple extension would provide individual attack probabilities for each traffic type on each network path.

Exogenous Information

Exogenous information, $W_{t+1} = (\hat{D}_t, \hat{A}_t)$ defines information from outside the system that arrives after decision (x_t) is made. Thus, the expected demand state is defined historically as $\hat{D}_t = f(\dots, d_{t-2,h}, d_{t-1,h}, \hat{d}_{t,h})$, $\forall h$ where current demand for customer (h) is an unobserved random variable ($\hat{d}_{t,h}$) before the shipping decision is made. It becomes a known value after the decision (x_t), so that (\hat{D}_t) is the expected demand at time (t).

Likewise, $\hat{A}_t = f(\dots, a_{t-2,p}, a_{t-1,p}, \hat{a}_{t,p})$, $\forall p$, is the attack expectation where $\hat{a}_{t,p}$ is a stochastic, binary post-decision variable describing the agent's attack expectation on each network path. It only becomes known after decision (x_t). Information of an attack ($a_{t,p}$) can be observed anytime the defender vehicles are operating on the network or by any means of monitoring the defender may employ. This information is used to update \hat{A}_{t+1} .

Cost and Objective Functions

The system progresses according to the system model (S^M) where the state variable (S_t) is updated according to the current decision (x_t) and the arrival of exogenous information (W_{t+1}). As such, we can generically describe how the state variable changes over time as $S_{t+1} = S^M(S_t, x_t, W_{t+1})$ where (S_t) captures the changes in the resource, information and knowledge states as the system progresses through time. According to the objective function and value measure, decisions are then made to avoid low value states and arrive at high value states across the entire planning horizon (Powell, Simao, & Bouzaïene-Ayari, 2012). The equations that follow provide the cost components used to evaluate every feasible decision and calculate the value of each state transition.

Convoy Operating Cost

The military has long understood that operating in convoys provides many benefits, chiefly mutual support and protection. Because convoys are often afforded significant security protection with armed vehicle escorts and even aircraft at times, they

entail significant fixed cost. (DeGregory, 2007) provides a useful discussion of these FPR considerations.

Here, we differentiate between convoy security escorts and RCPs which can also “escort” a convoy by clearing IEDs immediately in front of it. In contrast, security escorts are organic combat forces within the convoy. This leads to a path specific convoy cost equation that deterministically sums the fixed cost of security, planning, and other requirements associated with every convoy with the variable costs associated with convoy size. It should be noted that there are no RCP costs charged in our formulation, as these are considered to be outside the immediate system.

Equation 14: Deterministic convoy operating cost

$$c_t^1(q_{t.P.}) = \sum_{p=1}^P c_{t,p}^{convoy}$$

Where,

$$c_{t,p}^{convoy} = \begin{cases} c^{fixed} + q_{t,p}c^{variable} & \forall p \text{ and } q_t > 0 \\ 0 & otherwise \end{cases}$$

Delivery Reward

Let ($c_h^{deliver}$) be the reward for every unit of supply a convoy delivers to each customer (h). Obviously the long range objective of the model is to provide sustaining supplies to the customer. Hence, the successful delivery of each supply unit is rewarded deterministically based on the shipping decision. As formulated, this is the only positive contribution the agent can earn through its operations which reflects the logisticians’

primary responsibility and ultimate measure of success. While we only model one class of supply here, indexing for multiple supply classes is easily accommodated.

Equation 15: Deterministic delivery reward

$$c_t^2(q_{t,H}) = \sum_{h=1}^H q_{t,h} c_h^{deliver}$$

Customer Inventory Holding Cost

Moving excessive inventory to a remote operating base for storage entails security and storage costs on the customer. For this reason we exact a cost penalty for inventory levels that exceed customer storage capacities. Recall customer demand ($\hat{d}_{t,h}$) is an unknown stochastic variable at the time of decision (x_t), so the inventory cost are determined for feasible agent decisions according to the expected inventory level after customer demand is observed.

Equation 16: Expected inventory holding cost

$$\mathbb{E}[c_t^3(u_{t,H}, q_{t,H}, \hat{d}_{t,H})] = \begin{cases} \sum_{i=0}^{u_{t,h}+q_{t,h}} c^{holding} \max(0, u_{t,h} + q_{t,h} - i) P[\hat{d}_{t,h} = i] & \text{if } u_{t,h} + q_{t,h} \leq d_h^{max} \\ \sum_{i=u_{t,h}+q_{t,h}-d_h^{max}}^{u_{t,h}+q_{t,h}} c^{holding} (u_{t,h} + q_{t,h} - i) P[\hat{d}_{t,h} = i] & \text{if } u_{t,h} + q_{t,h} > d_h^{max} \end{cases} \quad \forall h$$

Unmet Demand Cost

The logistician has failed in his basic responsibility if customer demand ever exceeds the sum of current inventory and shipments. Therefore, we assess a large penalty

for unmet (exogenous) customer demand. The agent evaluates this possibility at each time step by calculating the expectation according to Equation 17.

Equation 17: Expected penalty for unmet demand

$$\mathbb{E}[c_t^4(u_{t,H}, q_{t,H}, \hat{d}_{t,H})] = \begin{cases} \sum_{i=u_{t,h}+q_{t,h}}^{d_h^{max}} c^{unmet}(i - u_{t,h} + q_{t,h})P[\hat{d}_{t,h} = i] & \text{if } u_{t,h} + q_{t,h} < d_h^{max} \\ 0 & \text{otherwise} \end{cases} \quad \forall h$$

Attack Costs

Anecdotal historical evidence has suggested that the presence of route clearance activities (RCTs) had no significant impact on the number of IEDs emplaced by insurgents; however, we have observed that RCPs have typically maintained higher discovery rates and lower casualty rates than standard maneuver forces and logistic units. This performance difference is due in part to the specialized equipment RCPs employed which allowed them to identify and interrogate potential IEDs from increased distances. Additionally, because RCPs typically search the same areas repeatedly, and for extended periods, they maintained more intimate knowledge of the roadways and surroundings under their purview compared to passing units. This is why they generally demonstrated greater ability to detect subtle environmental changes that might indicate the presence of an emplaced IED (Ardohain, 2016).

In our model, the costs of attacks are charged when the defender's convoy and RCP routing and scheduling decision is coincident with the attacker's ambush. In this formulation we have chosen to use the same attack probability calculation for all traffic

types on a given network path, but indexing for traffic type is a trivial extension. We do index attack penalties by traffic type (assigned according to decision x_t) to reflect the robustness of different vehicle and equipment, but attacks do not consume supply.

Equation 18: Expected attack penalty

$$\mathbb{E}[c_t^5(x_t, \hat{a}_{t,p})] = \begin{cases} \sum_{p \in x_t} P[\hat{a}_{t,p}](c_{x_t}^{attack}) & \forall p \\ 0 & otherwise \end{cases}$$

One step cost function

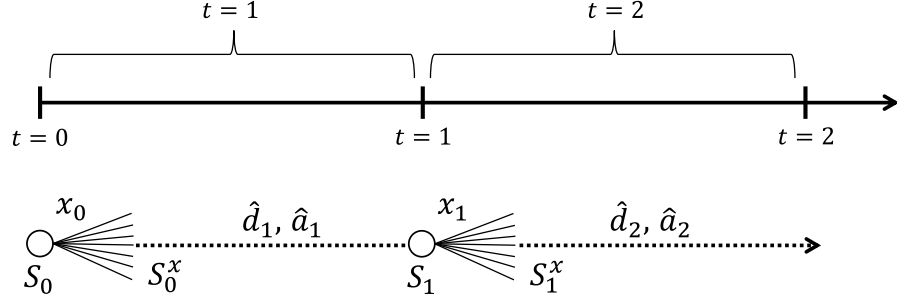
In any system state, (S_t) , the above costs yield a distinct one-step contribution function for every feasible action (x_t) available to the agent according to the system state at time (t) . Since the objective function maximizes total expected value over time, the rewards for deliveries are added while operating and attack costs are subtracted as follows:

Equation 19: One-step cost function

$$C_t(S_t, x_t, W_{t+1}) = -c_t^1(q_{t,P}) + c_t^2(q_{t,H}) - \mathbb{E}[c_t^3(u_{t,H}, q_{t,H}, \hat{d}_{t,H})] - \mathbb{E}[c_t^4(u_{t,H}, q_{t,H}, \hat{d}_{t,H})] - \mathbb{E}[c_t^5(x_t, \hat{a}_{t,p})]$$

State Value Function Transition

By letting $C_t(S_t, x_t, W_{t+1})$ be the contribution for being in state (S_t) , and taking action (x_t) , we can calculate the value of being in any post-decision state at some time (t) by recursively calculating the expected value of each feasible next state (S_{t+1}) . As shown in Figure 23, we find this value with the system model $S_{t+1} = S^M(S_t, x_t, W_{t+1})$ by splitting the model into two components – the deterministic $S_t^x = g(S_t, x_t)$ and the stochastic $S_{t+1} = f(W_{t+1}, S_t^x)$ where $x_t \in X_t$ is the set of all feasible actions in state (S_t) .



$$S_{t+1} = f(W_{t+1}, S_t^x)$$

$$\text{where } S_t^x = g(S_t, x_t) \text{ and } W_{t+1} = (\hat{D}_t, \hat{A}_t)$$

Figure 23: Time relationships between information flow and state transitions

With the deterministic Equation 20, we select the action (x_t) that maximizes the sum of the current one-step contribution and the long run (discounted) value of the logistic effort given the agent's current expectation. With Equation 21 we update the value of the post-decision state using the learning version of Bellman's recursive equation. Then RL is achieved via several simulations of the system, observing the system state (S_t), making decisions using Equation 20, and learning the value of the post-decision states using Equation 21.

Equation 20: One-step state value approximation

$$\hat{v}_t^n = \max_{x_t \in X_t^n} \left(C_t(S_t^n, x_t, W_{t+1}) + \gamma \bar{V}_{t+1}^{n-1}(S_{t+1}^x) \right)$$

Equation 21: Value function approximation (learning equation)

$$\hat{V}_t^n(S_t^x) = (1 - \alpha_{n-1}) \hat{V}_t^{n-1}(S_t^x) + \alpha_{n-1} \hat{v}_t^n$$

We start by initializing $\hat{V}_t^0(S_t) = 0, \forall S_t$, where (α) is the learning rate parameter and (γ) is the discount rate parameter. Then in each iteration, solving Equation 20 provides an observation (or sample realization) to update our belief in what is the “real” value of the current state. Accordingly, we want this to be a correct step along the stochastic gradient according to (α) which defines the step size. However, since the gradient is stochastic, we have no guarantee that this step is in the direction of the true post-decision state value. Despite this uncertainty in both the contribution function value and the state transition $(S_t \text{ to } S_{t+1})$, the algorithm can be proven to converge by selecting appropriate learning parameters and revisiting the state a sufficient number of times. Thus, we continue the learning process by increasing counter n until the convergence criteria is met (Powell, 2007).

Finally, when learning is complete, the algorithm has determined the values of several feasible post-decision states of the system. We test the system in the learned phase of the implementation, where in a given pre-decision state (S_t) , the optimal decision is derived by executing Equation 20. Several value function approximation techniques are available for cases of higher dimensionality to mitigate computational storage issues arising from the storage of the post-decision states values.

Policy

For dynamic, stochastic problems, such as the one at hand, the decision (x_t) for $t \geq 1$ is a random variable because we don’t know at $t = 0$ what the system state will be at any future time. In ADP, we don’t focus on finding the best decision in real time, but rather on learning the best decision rule given the information at hand. In any large

problem, a good (approximately optimal) policy will take significant time to learn. But, once the decision policy is learned, it can be applied in real-time, as soon as the current real-world system state is determined.

CHAPTER FOUR – EXPERIMENTAL RESULTS AND ANALYSIS

Computational Results

To demonstrate the formulation provides meaningful policy responses in differing environmental conditions, we tested the algorithm results against several representative scenarios of varying customer demand, subject to various attacker profiles. This gives us several different regimes, with consistent costs, rewards, and environmental models to evaluate the RL model’s look-ahead performance against a myopic decision strategy benchmark.

Our myopic decision algorithm applies a pure, one-step, greedy decision rule that does no learning. Rather it chooses the maximum expected reward for the current decision without consideration for downstream system states or values. Mathematically, this is performed by removing the state value for (t+1) in the current decision objective function (V_{t+1}) which is easily accomplished by setting the discount parameter (γ) equal to zero. This renders Equation 20 as:

Equation 22: Myopic one-step state value approximation

$$\hat{v}^n = \max_{x^n \in X^n} (C^n(S^n, x^n, W^n))$$

Base Case Experiment

We learn the decision policy, both myopically and with ADP, in various combinations of attacker and customer behavior. Then we test the learned policies in a stochastic simulation where we track several fundamental performance measures, such as

convoy sizes, convoy and RCP frequency, and attacks against convoys and RCPs (see Figure 24). The principle comparison measure we apply is the total objective function value achieved by the agent at the end of each simulation. This summarizes each policy's overall performance given the rewards and penalties in each simulated environment.

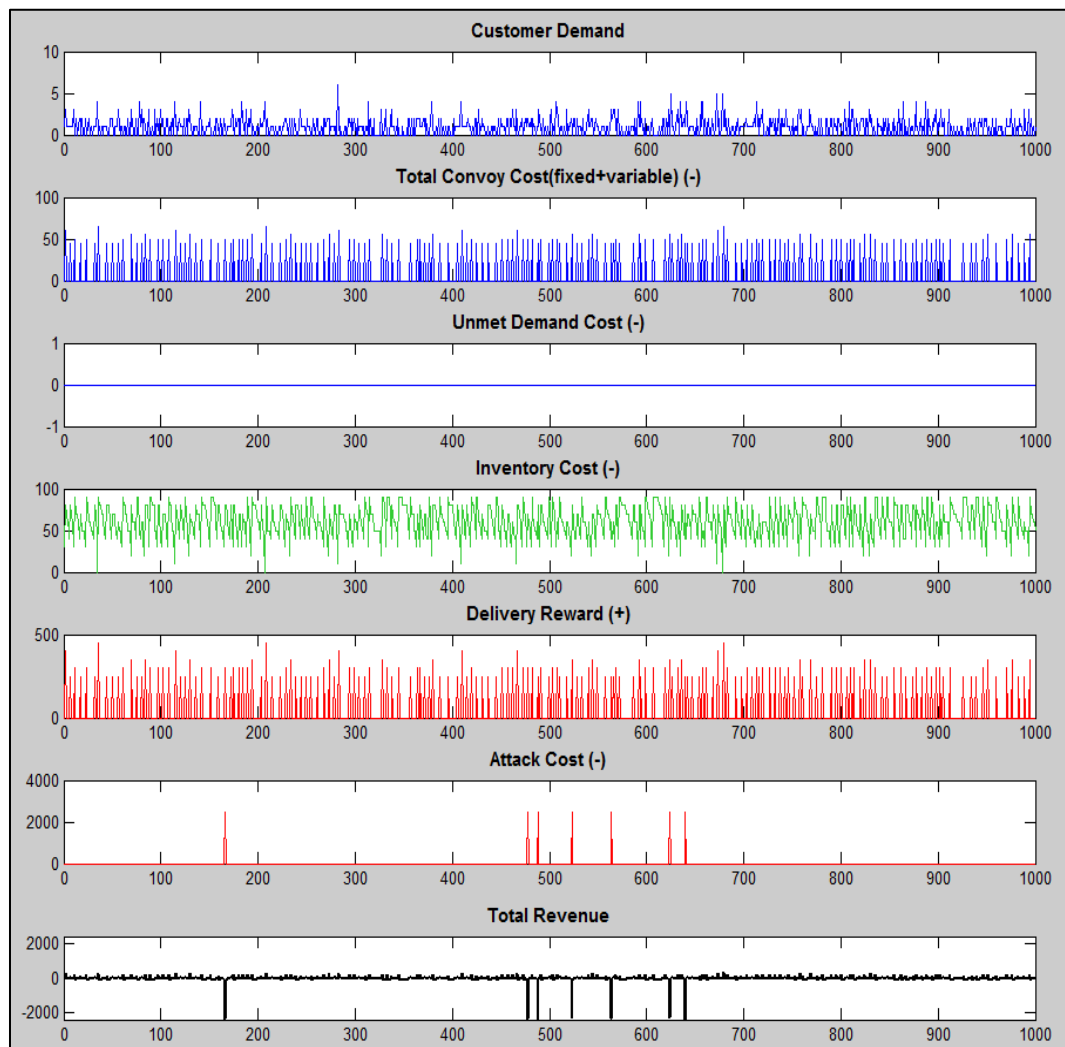


Figure 24: Example simulation output

The base case scenario uses the constraints listed in Table 1 for a problem consisting of a single depot and customer, connected by a single network path, and one assignable RCP.

Customer capacities (u)	21
Convoy sizes (q)	11
Traffic density levels (b)	8
Attack intensity levels (e)	2
RCP recovery period (j^{max})	5
RCPs	1-3

Table 1: Base case scenario constraints

These agent choices are applied across a set of scenarios that span four attack probability distributions (shown in Figure 25) and ten customer demand distributions (shown in Figure 26). This gives us a set of forty environmental models (one for each combination of attack and demand distributions) to test and compare the learning and myopic agents' behavior.

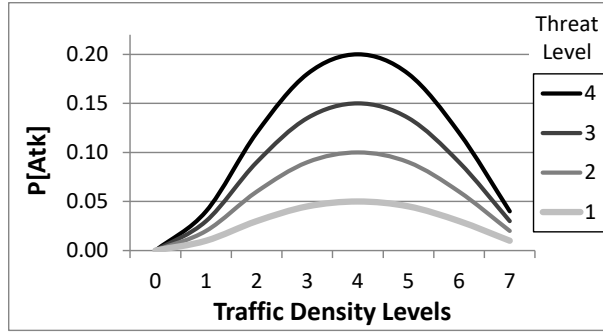


Figure 25: Four base case attacker threat profiles

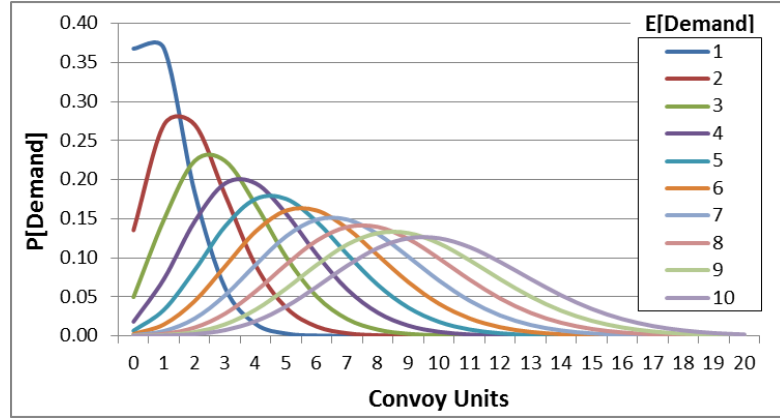


Figure 26: Ten base case customer demand profiles (Poisson)

Thus, we have a two-step process. In step one the agent learns the decision policy for every system state it visits under the applicable environmental parameters. Then in step two, the policy is implemented in a series of 10 independent, unscripted, stochastic simulations of 1,000 time steps each, maintaining consistent customer demand and attacker threat behavior. Throughout the simulations, the learned policy is applied at each simulated time step and new system state, with rewards and penalties calculated

according to the parameters in Table 2. Finally the performance of the agent in each simulation is evaluated by the total reward earned under the prevailing cost structure.

Convoy operating cost, fixed (c^{fixed})	20
Convoy operating cost, variable ($c^{variable}$)	5
Delivery reward ($c^{deliver}$)	50
Customer inventory holding Cost ($c^{holding}$)	10
Unmet demand cost (c^{unmet})	500
Unescorted attack cost ($c^{attack,convoy}$)	5000
Escorted attack cost ($c^{attack,escort}$)	2500
RCP attack cost ($c^{attack,RCP}$)	2000

Table 2: Base case cost parameters

The performance comparison of each policy within four simulated scenarios are shown in Figure 27 as the total accrued value each agent achieved (i.e., the sums of the rewards and penalties). As shown, the RL approach significantly reduces the defender’s median operational cost in all four regimes. Since the RL and myopic decision policies are applied in identical simulated environments, the performance gaps are entirely due to the decision strategies applied. Therefore they show the value of the RL approach of accounting for the multi-step effects of current decisions over the myopic benchmark. The RL algorithm significantly improves performance by satisfying customer demand in a way that accounts for the dependence of future attack probabilities on the current choices available to the agent. This is accomplished primarily through use of the

knowledge state. Thus, through learning a superior decision policy, the RL agent is able to better negotiate the dynamic operating environment, arriving at high value states and avoiding low value states more often than the myopic agent.

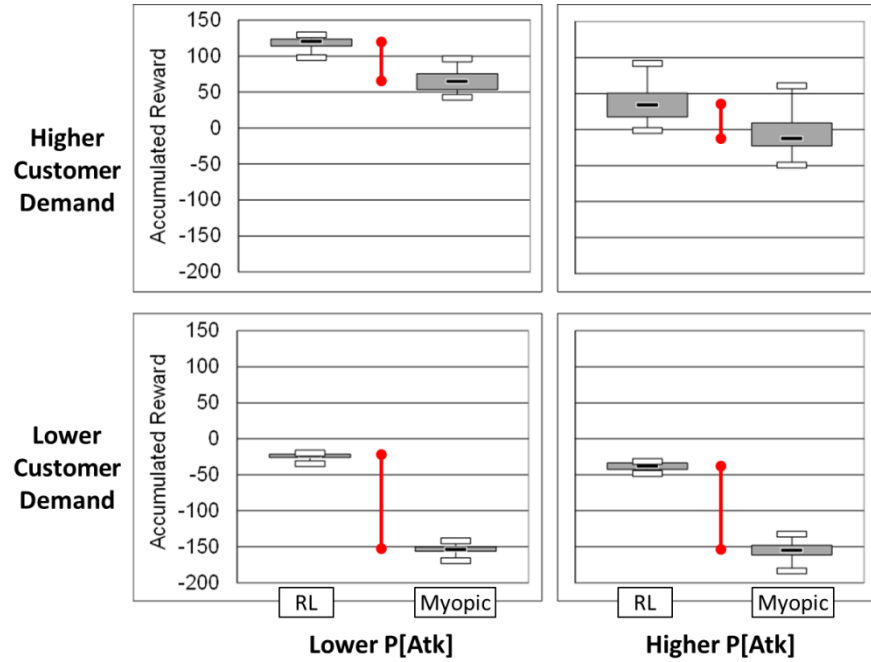


Figure 27: The distribution of accrued objective function value for simulations applying RL and myopic decision policies in four different operating environments^{24,25}

²⁴ The figure shows the max, min, median, and middle two quartile results for the objective function value in four simulated environments - demand levels 1 and 6 (lower and higher) in combination with threat levels 1 and 4 (lower and higher) of the base case.

²⁵ All four charts have the same vertical axis with the levels determined by the cost structure shown in Table 2. Under the base case cost structure, satisfying customer demand (earning the reward for delivery) is the biggest factor in the objective function values achieved.

Examining the Base Case

There are four broad observations we make about Figure 27. The first is that the lower left box plot comparison is where the biggest RL over myopic improvement occurs (this is when threat and demand are lowest). Second, the least cost variance (the narrowest box plots) occurs in the same lowest demand and threat regime. Then moving either up or right from the lower left comparison (either up to higher customer demand or right to higher attack probability) RL improvements decrease. Third, by the height of the box plots, we observe that in every case the RL results are more consistent (that is they have less variation) than the myopic results. Finally, the higher demand and higher risk case (the upper right comparison) is the only place where the myopic decision rule is competitive with RL policy (the box plots have significant (44%) overlap).

These four observations show that the learning agent is able to reliably achieve better and more consistent performance in the uncertain, simulated environment than the myopic agent. Further, that these improvements are greatest when there is the most slack in the system (where *slack* is defined as excess agent delivery capacity over expected customer demand); in other words, when agent has the most ability to choose its courses of action. Conversely, when system slack decreases (that is when customer demand is near the agent's delivery capacity as in the higher demand environment) there is little opportunity for the agents to make meaningful scheduling decisions, as they must utilize nearly constant full capacity convoys to satisfy customer demand. Thus, in both cases the agents have less variety in their activity patterns and less distinguishable performance outcomes.

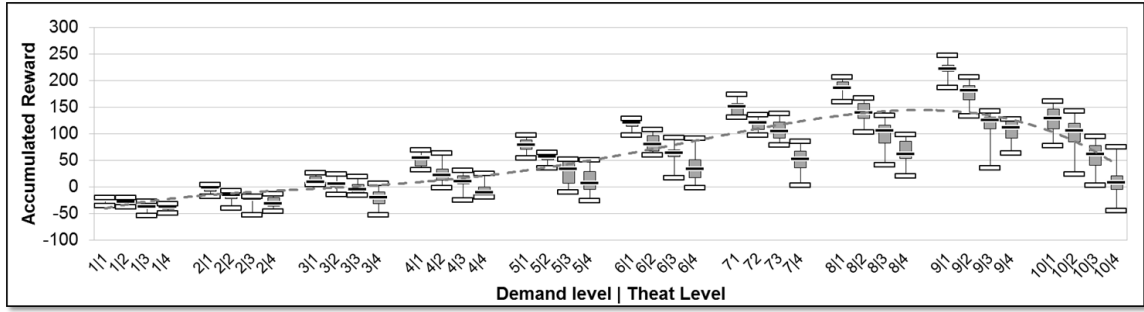


Figure 28: Consolidated base case simulation results for RL agent.

Figure 28 shows the results for the RL agent across the full set of 40 base case environments. The results appear in groups of four box plots where each group of four display results for each attack threat level (1-4) for each customer demand level (1-10). From these, we can see two additional system dynamics. First, the increasing negative slope within each successive customer demand level (i.e., for each group of four box plots) shows the increasing effect the attacker is able to exert as customer demand increases. By this, we see more clearly the importance of system slack and variety for the RL agent to learn and counter the observed attacker behavior. Second, the increasing payoff level for increasing customer demand levels (shown by the dashed trend line) results from the agent earning increased reward for accomplishing its primary objective of making deliveries until customer demand nears the agent's delivery capacity (at nine expected convoy units in this scenario). At this point, increases in customer demand regularly exceed the agent's delivery capacity, so the penalties for unmet customer demand begin to drive overall performance lower.

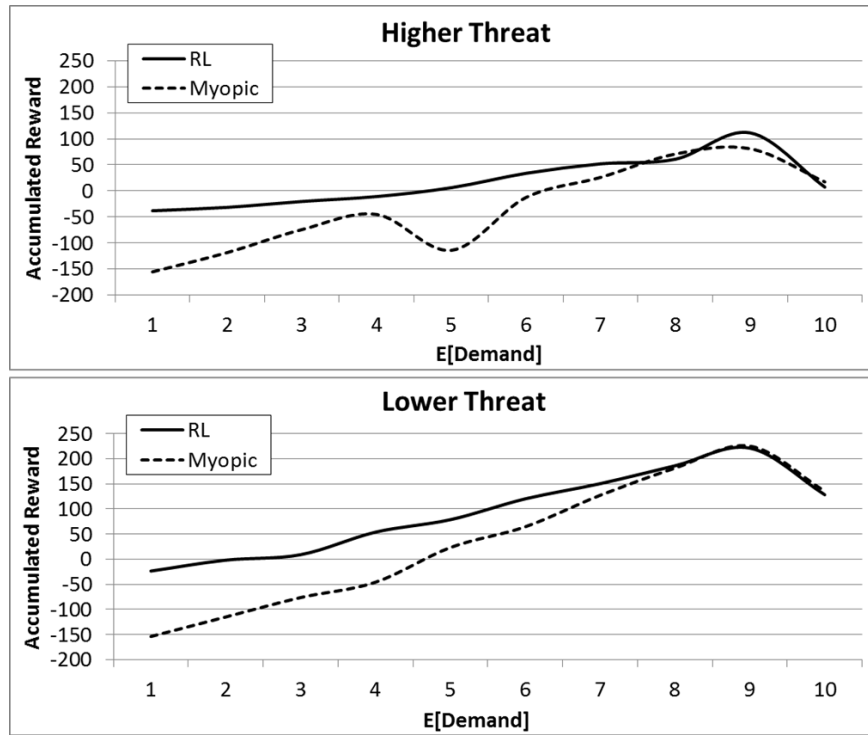


Figure 29: Median performance of RL and myopic agents in the base case

Figure 29 shows the results of extracting the overall trend lines (as shown by the dashed line in Figure 28) for both the myopic and RL simulation runs. This provides a means to compare general results across 160 different simulated environments. Here we observe the mean performance difference between the RL and myopic agent's decreases as customer demand increases and the logistic system reaches capacity. This again reflects the degree to which slack in delivery capacity is needed for the learning agent to make significant improvements, at any threat level.

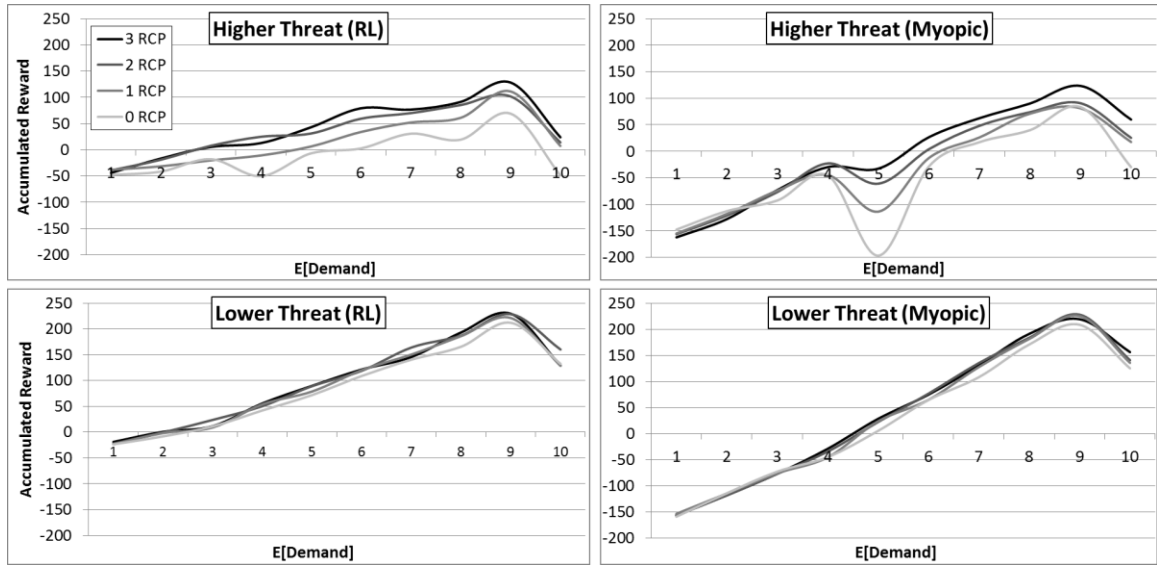


Figure 30: Impact of added RCPs

Within the base case, we are also interested in the marginal improvement increasing RCP protection can provide. Figure 30 shows that under the base case cost parameters, RCPs provide the most meaningful improvements at the highest attacker threat levels (shown in the top two charts). Moreover, these improvements are greatest when customer demand is greatest. Conversely, RCPs made little difference in the low attack probability environment. While this outcome is a fairly obvious, it serves to validate the model formulation.

In the upper right hand chart of Figure 30 there is a severe drop in the myopic agent's performance when the expected customer demand is five units (note the local minima in the upper right chart at all RCP capacities). This seemingly inconsistent result is further illustrated in Figure 31 which shows three side-by-side comparisons of the base case with zero, one, and two RCPs (i.e., each chart shows the results for 40

environmental scenarios). In these plots, since attack probability is on horizontal axis, the slope of each line depicts the effect of threat level in each modeling regime. This helps to illustrate two important dynamics.

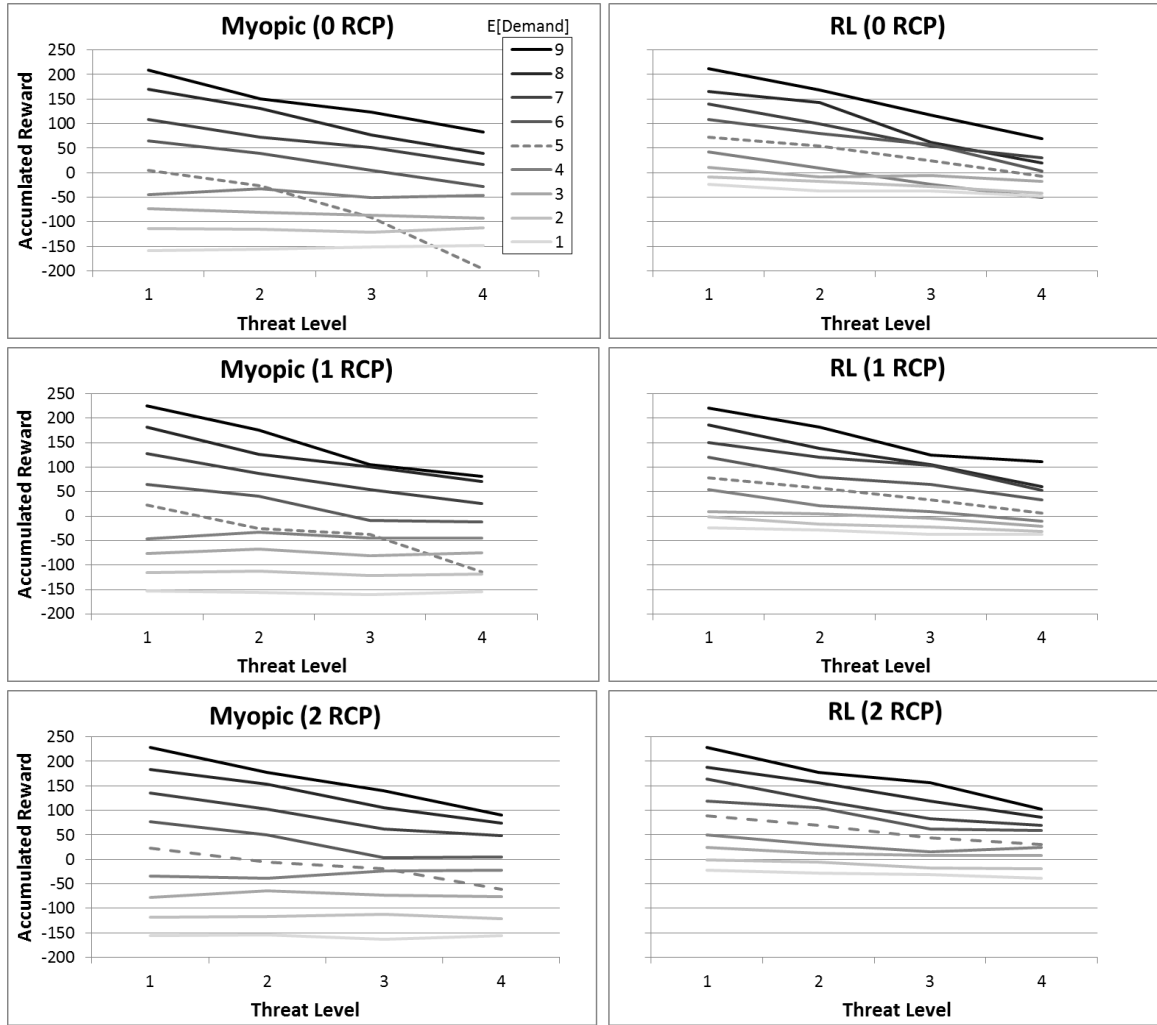


Figure 31: Detailed base case comparison

First we observe that the negative slope of the lines within each chart (Figure 31) increases as demand increases (toward delivery capacity). At the lowest demand levels

(shown by the lighter shaded lines) the slope of the lines is shallow, confirming the earlier observation that the attacker has only minimal effect on the overall agent performance when demand is low. Then as demand increases (depicted by the darker shaded lines) the slope of the lines increase, indicating the greater attacker effect with increasing customer demand. Second, Figure 31 shows that when the expected demand is five units (depicted with a dashed line) there is a significant break in the smooth progression of the lines in all three myopic cases. This anomaly occurs where myopically satisfying customer demand results in the optimal operational tempo (traffic volume) for the attacker, who is able to maximize his effect on the defender. By comparing all three sets of charts right to left, it is clear that unlike the myopic agent, the RL agent is able to detect this disadvantageous operational regime and adjust its policy to avoid the significant performance degradation.

Policy Response to Different Environments

In each pair of simulation runs, the RL and myopic agents have identical resources and modeled environments; thus, the performance differences being observed result from each agent's learned decision policies. To better understand how the RL agent achieves superior results we want to closely examine the decision policies themselves.

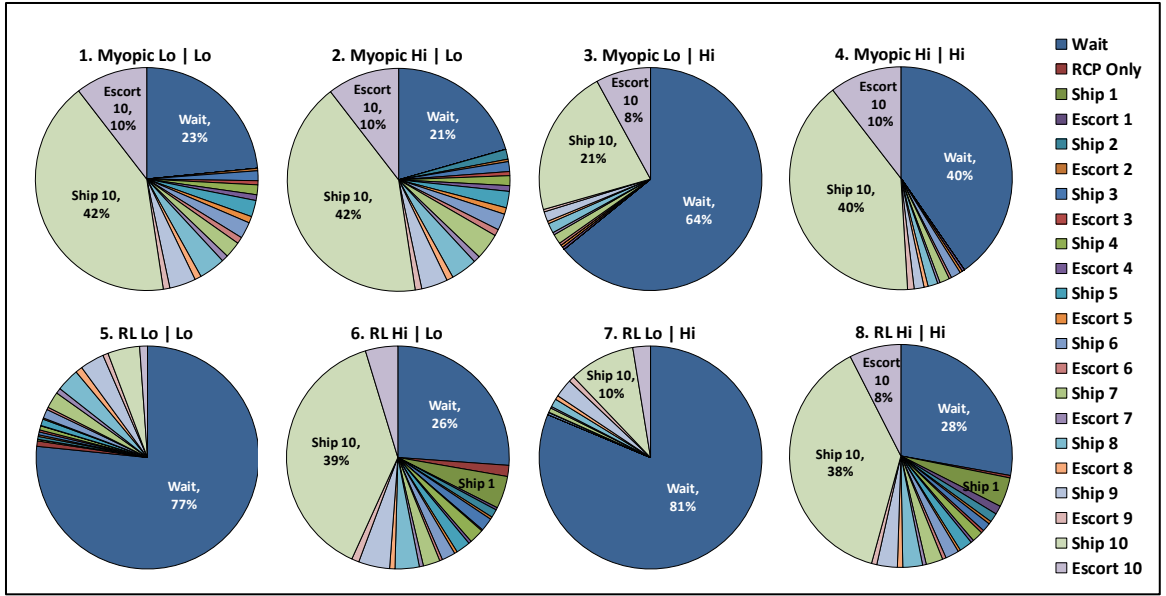


Figure 32: Count of system states with each action choice for myopic and RL decision policies (labeled by Demand level | Threat level)

Figure 32 is a set of pie charts depicting the proportion of states in which the learned policies assign each available agent action (for both the RL and myopic agents). The look-ahead effect on the learned policy can be seen by making vertical comparisons of the (upper) myopic policy pie charts to the (lower) RL policy pie charts. For example, comparing charts #1 and #5 shows that the myopic agent chooses to “wait” in 23% of the system states (chart #1) compared to 77% for the RL agent (chart #5).²⁶ Conversely, the RL agent only uses the maximum capacity convoy (labeled “Ship 10” and “Escort 10”) 6% of the time compared to 52% in the myopic case. This reveals that the RL policy chooses to ship in significantly fewer system states than the myopic policy, but when the RL agent does choose to ship, it uses a more evenly distributed (or more

²⁶ The pie charts are based on the numbers in Appendix C: Base Case Policy Adaptations.

variable) allocation of the available actions (i.e., it includes a higher proportion of small convoys). This more dispersed assignment policy is a consistent feature of all the RL policies shown.

The effects of changing demand or threat levels can be seen by comparing the pie charts in Figure 32 horizontally. For example, comparing charts #1 to #2 shows that the myopic policy is essentially unresponsive to increased customer demand. Then comparing #5 to #6 shows the RL policy changes significantly under the same changes in expected customer demand (from choosing "wait" in 77% under low demand and low threat, to just 26% under high customer demand). Thus, we see that the RL agent policy chooses to ship in far fewer states when demand is low, but greatly expands the number of states when demand increases. Finally, Figure 32 also reveals that both the myopic and RL agents react to increased threat by decreasing the number of states in which they choose to ship (comparing chart #1 to #3 and #5 to #7).

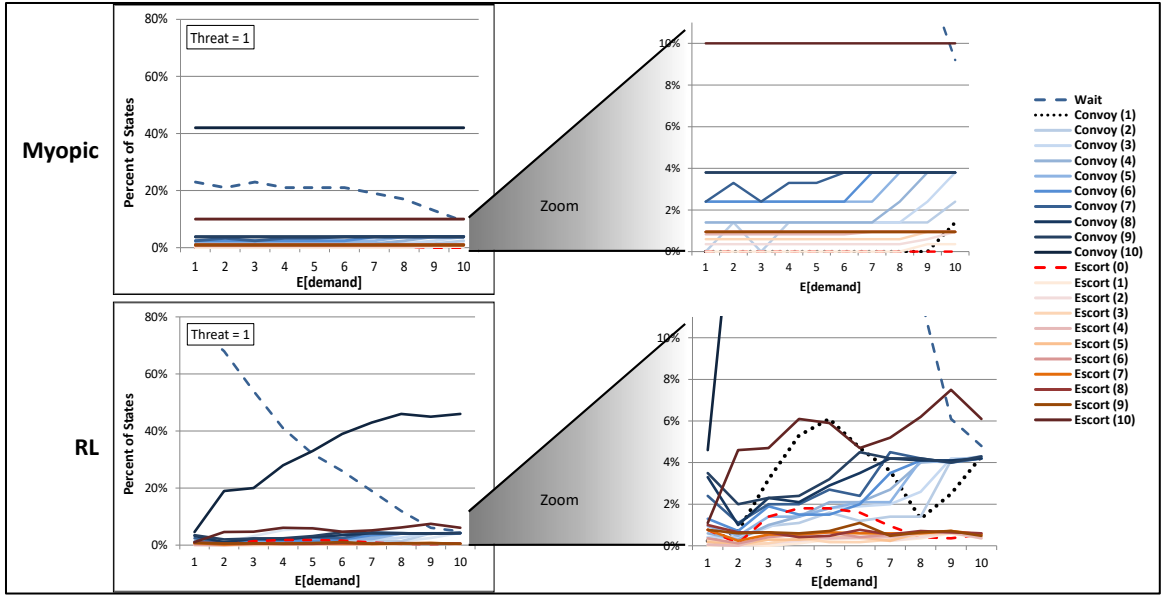


Figure 33: Comparison of policy dynamics, myopic and RL agent in base case (lower threat case with 1 RCP)

Figure 33 provides a more detailed decision policy comparison which shows the RL agent's policy response is appreciably more dynamic than the myopic agent's when observed customer demand changes. Observe the top half of Figure 33 which shows that the myopic agent policy is reminiscent of a typical $(s; S)$ policy as is expected in general inventory models (Feinberg & Lewis, 2016; Scarf, 1960)²⁷. This stasis in the myopic case occurs because the system moves from state-to-state, isolated from downstream estimates of exogenous customer and attacker behavior. While both the myopic and RL agents account for the immediate attacker threat, only the RL agent has an extended view for expected downstream payoffs and penalties in its objective function. Thus, we see in the lower two charts of Figure 33 that no such general policy emerges for the RL agent.

²⁷ In 1959 Herbert Scarf proved optimality of $(s; S)$ policies for finite-horizon problems with continuous demand.

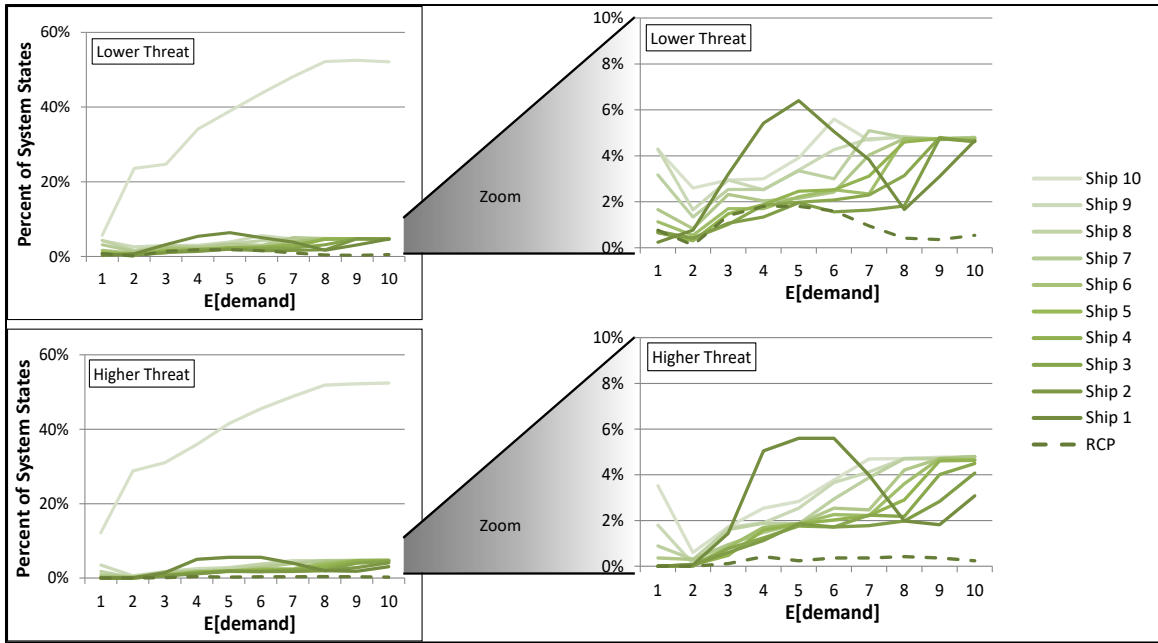


Figure 34: RL agent policy response to changes in expected customer demand and attacker threat level in base case (1 RCP)

Figure 34 compares the RL agent's dynamic responses to two different expected attacker threat levels for the same ten customer demand levels. While there are several similarities and differences in the pattern of agent actions within the policies displayed, one of the most noticeable features is the dashed line showing the increased use of independent RCPs for dedicated route clearance in the low threat case (when expected demand is between three and six units, see upper right chart). This shows that a common operational policy which maximizes RCP usage may be sub-optimal under some environmental conditions. A second noteworthy element is that the minimum number of states where convoys of less than ten units are assigned occurs when expected customer demand is two units (this is where all the lines reach their minimums). At this point, the

RL agent chooses ten unit convoys over every other convoy size (at a 2.7:1 ratio in the lower threat case and 18.5:1 for the higher threat). This inflection point for every line (except the ten unit convoy line) is unexpected as the general trend above three units of demand, for all convoy assignments is generally correlated to demand as most logistics planners would likely expect.

These two nuanced variations in the RL agent's policy provide improved performance and would be difficult for unaided human planners to determine; demonstrating the often unintuitive results an RL algorithm can provide.

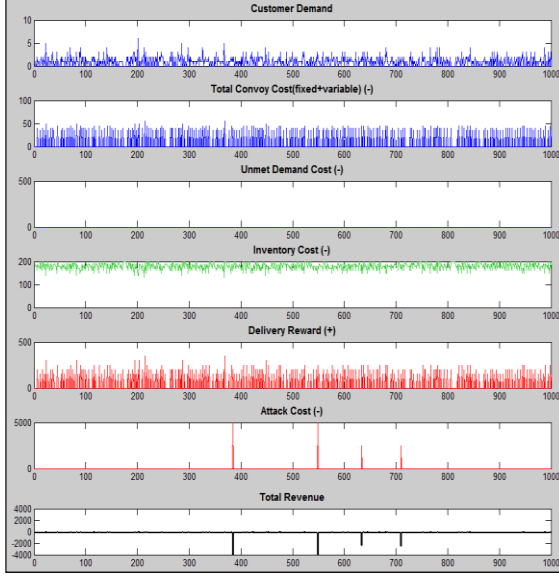
Base Case Policy Summary

In summary, examining the base case policies shows that the RL agent is far more responsive to environmental changes than the myopic agent. Additionally, compared to the myopic agent, the RL agent generally maintains greater variety of action in the states it chooses to act. Finally, the RL agent's policy response to varying environments is often unexpected when viewed across system state space.

Simulation Results

It is important to observe how a policy is implemented in the actual environment because during real-world execution the system will only visit a subset of the total possible system states, often never visiting the majority of the possible state states. Therefore, the distribution of actual agent decisions during implementation may be significantly different from what the policy mapping suggests. For example, large convoys may be the policy choice in a majority of system states, but during execution they are conducted rarely because the agent chooses to avoid these states.

Myopic



RL

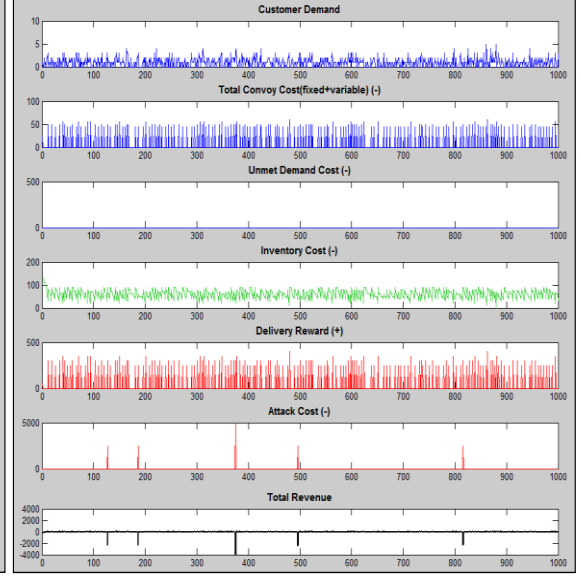


Figure 35: Example RL and myopic simulation output (lower demand and threat profiles)

During the simulations we tested the agent policies for their effectiveness. Figure 35 provides a visual comparison of the RL and myopic agent actions subject to identical environmental models. By comparing these simulation outcomes we see the RL agent chooses action less often in the form of fewer, larger convoys while consistently maintaining a lower customer inventory than the myopic agent in the same circumstances (see the delivery reward and inventory cost lines).

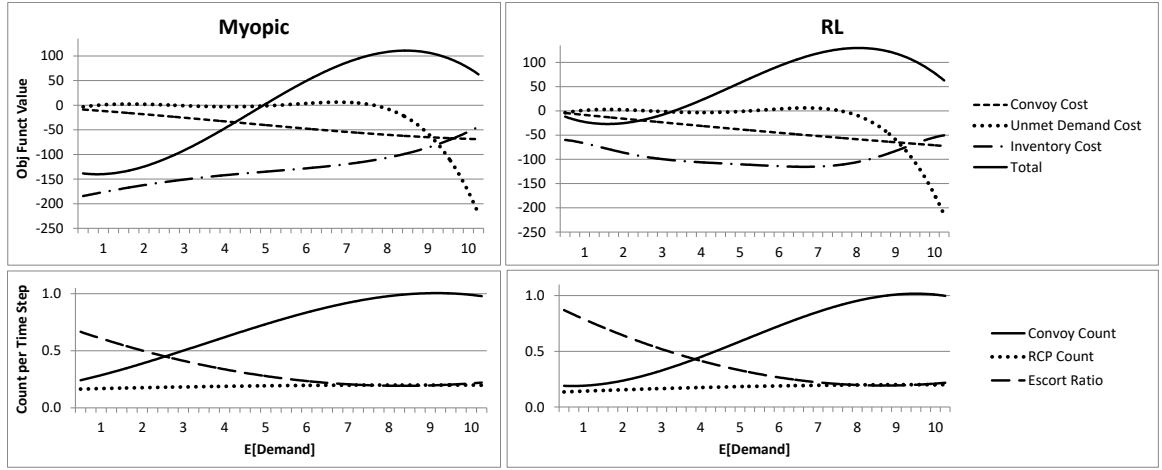


Figure 36: Operational dynamics during example simulation runs. Top: cost. Bottom: activity proportions

The top two charts in Figure 36 show agent activities levels observed in the base case simulation. The cost profiles shown for convoy and unmet inventory costs (solid and dotted line) have similar contours in response to changes in expected demand. But, there is a significant difference in inventory levels maintained (dot-dash line). The inventory cost curve in the RL chart (dashed-dotted line in upper right hand chart) forms a distinct bathtub shape that is absent from the myopic agent chart. This indicates that the RL agent consistently chooses to maintain lower customer inventory levels in the lower demand regimes compared to the myopic agent.

The bottom two charts in Figure 36 show the proportions of asset utilization. These reveal that during execution, the RL agent’s convoy count (solid lines) exhibits “s-shaped” response to increasing customer demand, where myopic agent response is nearly linear until nearing delivery capacity. Thus, in response to the attacker threat, the RL agent consistently seeks to utilize fewer convoys with a higher ratio escorted (dashed

line) than the myopic agent; thus, conserving its logistics assets so long as it has sufficient slack capacity. This is unlike the typical results for threat free (uncontested) environments where the most efficient response to increasing demand is increased convoy sizes, paying only variable shipping costs.

Finally, as is expected, the dotted line in upper two charts in Figure 36 shows that unmet customer demand penalty increases significantly when customer demands begin to exceed delivery capacity (dotted line, top two charts).

Alternative Threat Profile

Extensions of the Base Case are useful for exploring the learning and myopic agents' responses to diverse environmental conditions. In Figure 37 and Figure 38 we change the attacker threat profiles to demonstrate that just as in the base case, the RL approach can significantly reduce the defender's median operational cost in differing threat environments (see Figure 37 and Figure 38).

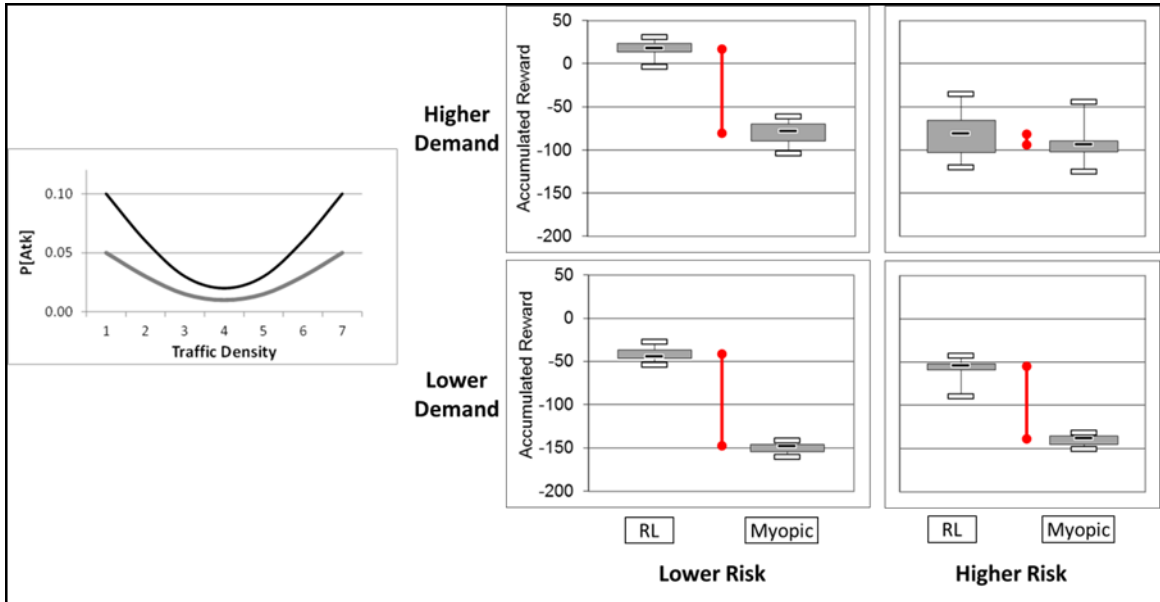


Figure 37: Inverted attack probability profile and agent performance comparison.

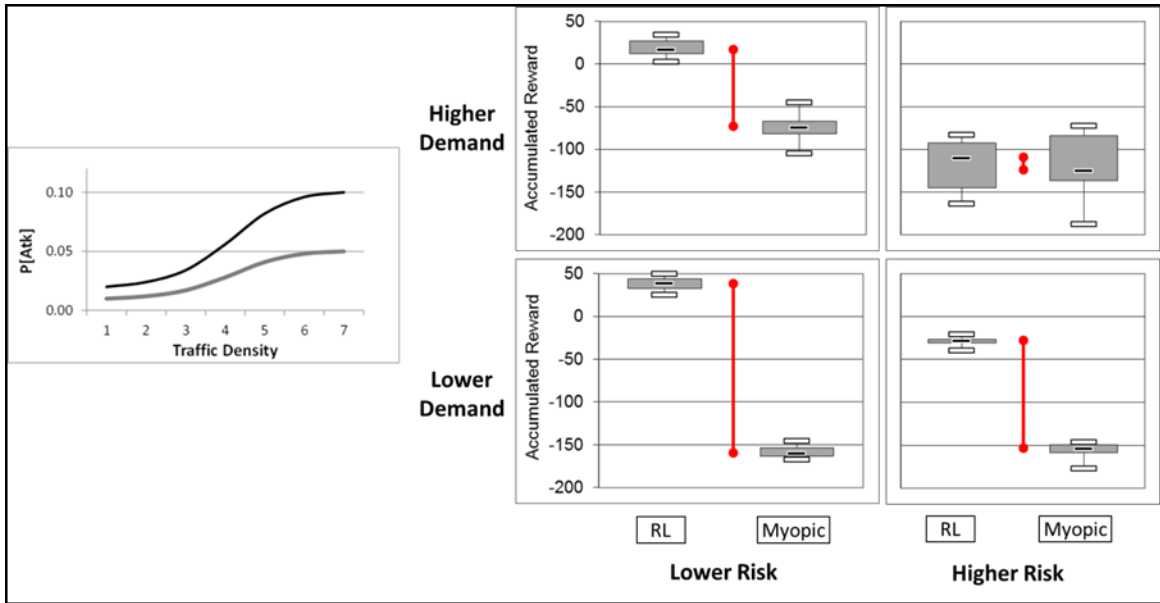


Figure 38: Increasing attack probability profile and agent performance comparison

Again we note that the learning agent consistently achieves better performance than the myopic agent in these alternative threat environments where just as we observed earlier these improvements are greatest when system slack is greatest. Additionally, the four observations we made about the base case (Figure 27) continue to hold under these alternative threat profiles, namely:

1. The biggest performance improvement from RL over the myopic occurs when both the threat and demand are lowest.
2. The least cost variance (the narrowest box plots) occurs in the lowest demand and threat case.
3. RL results are generally more consistent (has less performance variation) than the myopic results.
4. The highest demand and risk case (upper right comparisons where system slack is minimum) is the only place where the myopic decision rule is competitive with RL policy.

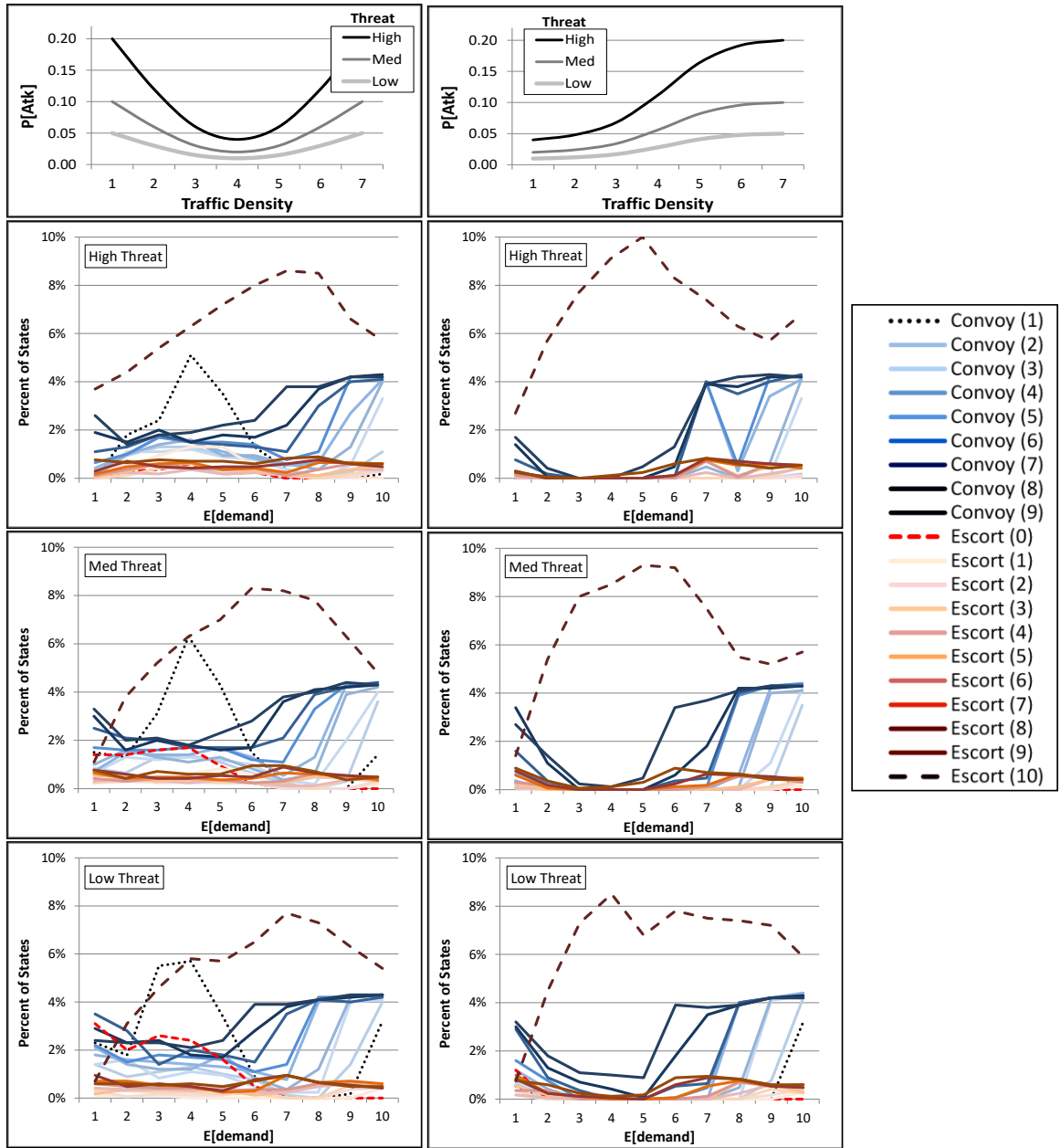


Figure 39: RL agent policy response to alternative attacker threat profiles

Figure 39 shows the RL agent's policy response to the two alternative threat profiles (displayed in the top two charts). A comparison of the policy profile charts on

the left (inverted threat profile) to those on the right (increasing threat profile) shows the RL agent has distinctly different responses to the two attacker behavior profiles. Most notably, under the increasing threat behavior model (right side charts) the curves take on a distinct, deep “bath-tub” shape which is clearly missing from the inverted threat profile charts (left side). In fact, the RL agent’s response to the two threat profiles are nearly opposite in the lower customer demand environments. Additionally, the inverted threat profile charts (left side) contain a noteworthy feature, namely the spike of single unit convoys (when demand is four units, dotted black line) and increased use of RCPs (dashed red line) when expected demand is less than six units. Again we note that these dynamics occur in the lower customer demand regimes, when the RL agent has the most slack delivery capacity. Further, these distinct shifts in the RL agent’s strategy provide improved performance and would be difficult for human planners to determine without the aid of a learning algorithm.

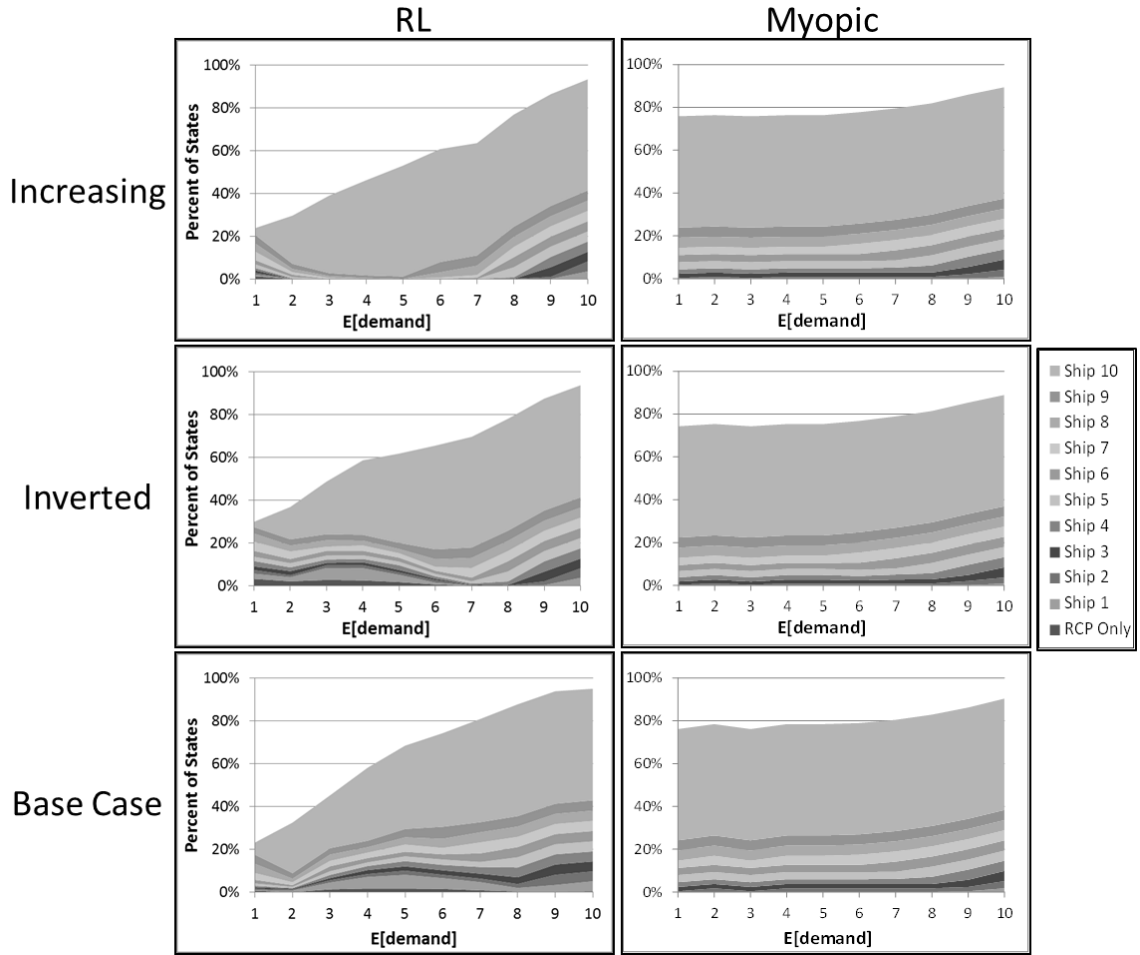


Figure 40: Policy comparison showing the percentage of system states with given agent decisions for the three threat profiles.

In Figure 40 we show a summary of the RL and myopic agent policy responses under each of the three attacker threat profiles presented. From this we see the dynamic response of the RL agent juxtaposed to the nearly uniform response to the myopic agent. Not only are the RL agent policies different from the myopic agents', they clearly vary from each other greatly according to the threat profile being faced. This shows the

sensitivity of the learning agent to detect changes in its operating environment and the potential value of adopting this modeling approach to vehicle scheduling.

Hybrid Network Simulation Example

To demonstrate the network performance of the learning algorithm, we provide a final three route network example with each route exhibiting a different threat profile - one of the three previously analyzed²⁸. Comparing the pie charts in Figure 41 (lower threat) to those in Figure 42 (higher threat) shows that during simulation, the RL agent makes significantly different route utilization choices in response to each observed variation in both threat and customer demand behavior. The pie charts show that even when the attacker threat profiles do not change within the network, but only the threat level, the RL agent still adjusts its decision policy in a way that produces widely varying route utilization choices. Further, the choices of route vary significantly with changes in observed customer demand (comparing the pie charts horizontally). Again we note, these policy refinements provide measureable performance improvement and are not likely to be realized by unaided human planners.

²⁸ This problem has 430,080 systems states and 124 (constrained) agent action choices

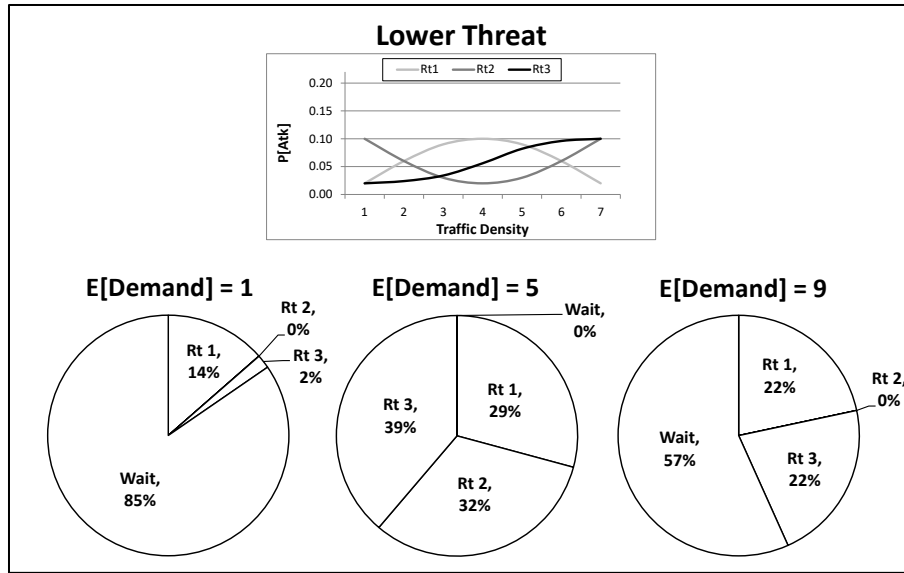


Figure 41: RL agent path utilization in three route case, lower threat profile²⁹

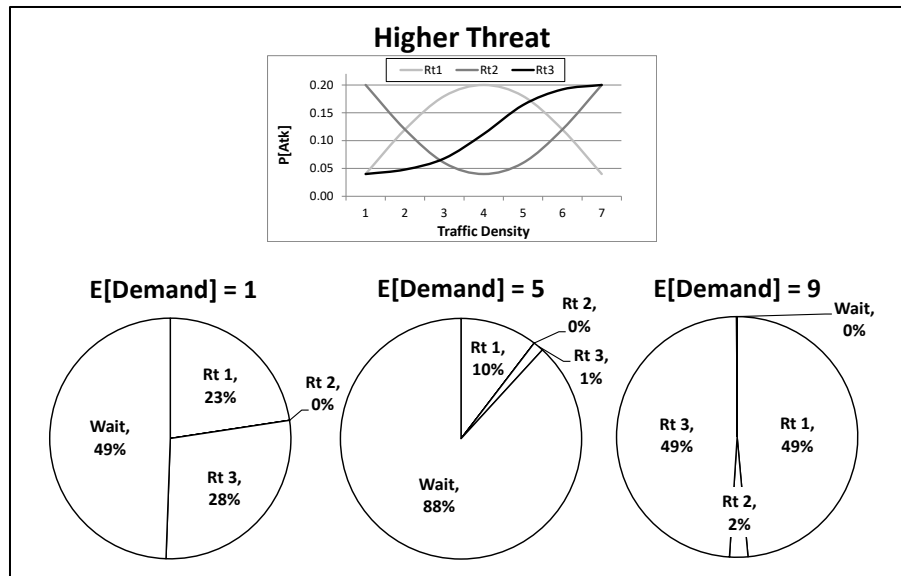


Figure 42: RL agent path utilization in three route case, higher threat profile²⁸

²⁹ Values rounded to nearest percent

Summary of Experimental Findings

1. During simulation, not only does the RL agent achieve consistently better overall performance under all environmental models, but it also achieves more consistent results. Only when there is little or no logistic slack does the myopic agent achieve results that are competitive with the RL agent's.
2. Our RL algorithm shows marked improvement over the myopic benchmark in all simulations when there was slack in the agent's delivery capacity while decreasing slack resulted in increasing attacker effects. Thus, delivery capacity slack is necessary for RL performance improvements because slack allows the learning agent maximum policy latitude to avoid low value states and arrive at high value states. Conversely, when the logistic system is operating near its capacity, the learning agent has fewer decision choices available and its behavior becomes less variable (i.e., more predictable).
3. In achieving improved results over the myopic benchmark, the RL agent responded to increased customer demand by increasing the variety of actions across the state space. Further, it responded to increased threat by decreasing number of states in which it chose action. This is seen in the RL agent's convoy count metric which exhibited "s-shaped" response to increasing customer demand, where myopic agent response was nearly linear. Thus the frequency of convoys utilized by the RL agent to satisfy the same customer demand is generally lower than the myopic agent's.
4. The consistent RL performance improvements are generated by often subtle and unexpected shifts in the agent's decision policy relative to the myopic agent's decision rule. These generally unforeseeable policy refinements provide measureable

performance improvement, demonstrating the value of the RL approach to provide insights which unaided human planners would likely be unable to discover.

CHAPTER FIVE – CONCLUSION

Conclusion

This dissertation is motivated by the stark reality of two recent military campaigns in Iraq and Afghanistan - both characterized by intense combat on public roadways where more than 60% of U.S. combat casualties were caused by IED ambushes (Barbaro, 2013). After more than ten years of massive investment by the United States to address the IED threat, today the threat remains as ominous as ever. This work is further motivated by the belief that protecting one's own supply lines and attacking those of an enemy will remain a fundamental military strategy for the foreseeable future.

The approach taken in this dissertation is to incorporate the military's well understood OODA loop principles into an RL scheme in order to provide a means to improve current operational planning approaches. The OODA loop is a means to structure our thinking about the problem and points to a methodology emphasizing the importance of observation in the attacker decision cycle (Boyd, 1986).

If we view an IED ambush from the attacker's perspective, we see that the IED emplacement is the attacker's prediction of the future. The attacker's choices of time, place and technique are based on his expectation of a future attack opportunity. Accordingly, it is natural to ask why the attacker came to any particular conclusion. For a rational actor, this will be based on what was observed and understood. This reasoning leads us to seek a modeling approach that will account for the dependence between the

defender's observable actions and a learning attacker's observations, orientations, decisions and actions in response. Hence, in this dissertation we have argued for two conclusions. First, seemingly unpredictable individual unit commander choices do not equate to force-level operational unpredictability. Second, the probability of attack on each individual unit action cannot be assumed to be independent of the past and future actions taken by others on the road network.

The principle approach of most previous OR models addressing contested network scenarios fit into two broad categories. First are those that focus on the network itself, seeking to maintain the maximum network flow or minimum cost paths in the face of changes to network architecture due to attack. While these models provide insight and context to our work, they do not directly apply to the sustained repetitive logistics problem faced in this research. Second are statistical and game models that use traditional aggregation methods, point processes, or presumed opponent strategies. These generally rely on regularity and long-run averages; thus, they do not capture the nuanced nature of the day-by-day, even minute-by-minute ebb and flow of adversarial interaction and competitive learning. While this research effort is not divorced from any of these approaches, it did take a deliberate step in a new direction, toward assuming dependence between the attacker's and defender's action choices.

Another important distinction between this effort and much of the previous work is that this research is not focused on improvement of, and direct application to, existing military practices. Rather, we are proposing a new operational paradigm in contrast to current OR models in the literature which envision either a defender or an attacker who

exhibits some pattern of activity while the opposing party adjusts (Washburn A. , 2006; DeGregory, 2007; Marks, 2009; Lin & Washburn, 2010; Washburn & Ewing, 2011; Kolesar, Leister, Stimpson, & Woodaman, 2012). Recent experience, in two theaters of war, indicates that such assumptions are only justified when operations are viewed with a high level of aggregation.

In contrast, we propose that the defender's counter-IED problem during repetitive transportation movements is more than finding and avoiding IED ambushes, but includes the critical element of shaping the attacker's expectations by learning how to best influence and anticipate the attacker's behavior through carefully organized activities. Thus our model assumes dependence between attack probabilities and targeted traffic patterns. While there are currently few analytical approaches that explicitly make this assumption, this is a distinguishing feature of our approach. It relies on a robust information state to enable a learning agent to uncover the action-reaction dynamics between the attacker and defender.

To our knowledge our approach has not been pursued anywhere in the OR literature related to this problem. Our goal is to effectively address the attack problem by crafting vehicle movement schedules that not only satisfy the military distribution problem, but also to significantly improve the defender's performance.

We employ this approach in an RL algorithm which offers unique opportunities for meaningful improvements in this complex problem area because it provides a means to evaluation immediate choices under uncertainty in the context of long run objectives. The application of these techniques in our basic formulation produces insights that can

not only inform logistic planning methodologies and improve understanding of the action-reaction dynamics, but the non-obvious, dynamic responses also demonstrate the need for continued research on the use operational patterns to influence attacker decisions and improve defender outcomes.

Improved Modeling

Developing and validating our modeling technique is hindered by data availability and the practical reality that there is no opportunity for live trials. Archived BFT data is of limited use for learning attacker-defender interactions, not only because it is generally incomplete, but more so, this effort rests on the counterfactual claim that if the defender acted differently, so too would have the attacker. Testing this assertion entails more than access to historical data, but requires a robust trial and error testing through carefully designed operational experiments.

The ADP technique we have utilized also has several practical limitations with respect to scalability, parameter selection, and computing capacity. Here enters the need for continued innovation and additional modeling techniques to discover better ways to detect, characterize, and exploit the structure within the seemingly chaotic environment. In this context, the chief technological hurdle we face is the increased dimensionality of the already difficult, NP-hard VRP.

Thus, we face a squeeze between the curse of dimensionality and the requirement for a solution method with robust enough content to improve prediction and shape the ongoing competitive interaction. Nonetheless, this dissertation is an initial step toward

developing a fundamentally different way of addressing the problem and there are several directions future efforts can take this work.

Future Research

During the course of this research several complementary fields of study came to light that, while outside the immediate scope, may provide important enabling capabilities to developing a fully operational learning algorithm for military logistic movement control.

Addressing Dimensionality

The model presented herein is very simple; nonetheless, the curse of dimensionality was significant. Several aggregation approaches were applied, but more work is needed in this area to support the level of detail required. Potential state space factors include increased time and environmental resolution, additional network activities types, attack methods, and terrain attributes. Other factors have been recommended, such as mosque, police station, and checkpoint locations, significant religious and political events, economic factors, demographics, terrain, weather, and lunar cycles (Ahner & Spainhour, 2015). Without creative new approaches, the addition of these attributes is intractable with the current formulation. But, their inclusion is likely important to any fully functional operational model. Thus, innovative approaches for representing increased temporal and spatial resolution across a network are needed. Several researchers have provided good foundational work in this area (Okabe, Yomono, & Kitamura, 1995; Okabe & Yamanda, 2001; Yamanda & Thill, 2007; Lu & Chen, 2007; Xie & Yan, 2008; Eckley & Curtin, 2013).

The requirement for additional resolution need not discourage efforts to move forward because experience has shown that most attack networks are local, characterized by spatially limited activity hot spots (Keefe & Sullivan, 2011; Connable, Perry, Doll, Lander, & Madden, 2014; Ahner & Spainhour, 2015). Thus, there is no need to model a complete road network for a county or major city. This fact means that models with a small number of network paths can be useful, making the dimensionality challenge more tractable. Additionally, since an agent will only visit a subset of the total system states during real-world execution, there is no need to exhaustively model every possible state. These considerations provide some obvious relief from the curse of dimensionality.

Time Series Pattern Recognition

In a complex system, the dilemma is to discover which part of a measured pattern should be ascribed to “randomness” and which part to “order” (Crutchfield, 1994). That is, can we find and understand usable information and determine what information to ignore? The sequences of attacker and defend activities across the network are of particular interest for understanding if the attacker is exhibiting preferences or aversions to certain defender activity patterns. Further, real-time (or online) awareness of when attack conditions are developing is critical to the network defender.

In any real-world environment, with multiple actors, there can be many meaningful patterns of activity and interactions that go unnoticed. Detecting such patterns is the first step toward predicting them. If we assume that there is a temporal structure in a set of behaviors, such that certain actions are typically part of some repeated sequence of events, then the occurrence of the first event is a precursor

indicating increased likelihood of the successor action. When such patterns occur repeatedly, with temporal consistency, they may be detectable and therefore be predictable (Magnusson, 2000).

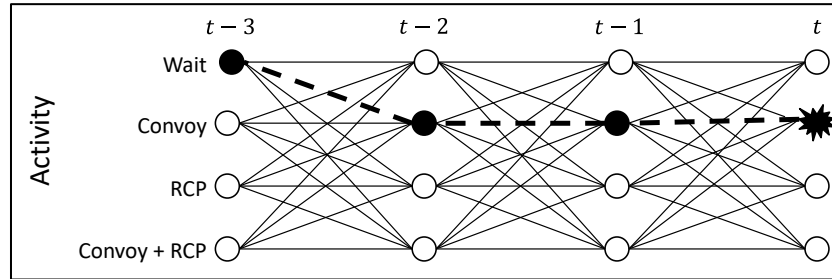


Figure 43: Representation of an activity sequence on a road segment ending in an attack.

Traffic activity on a roadway can be viewed as time series, a chronological sequence of events indexed by time and categorized at any level of resolution desired. For example patrols, civilian traffic, convoys, RCPs with attack events can be tracked and indexed. Figure 43 shows a simple sequence of four possible activities that might be tracked on a road segment. The four activities across the four time steps allows $4^4 = 256$ possible activity patterns. The simple sequence highlighted is (wait – convoy – convoy – convoy:attack) can be represented by unique alphabetical characters such as: A, B, B, B_a.

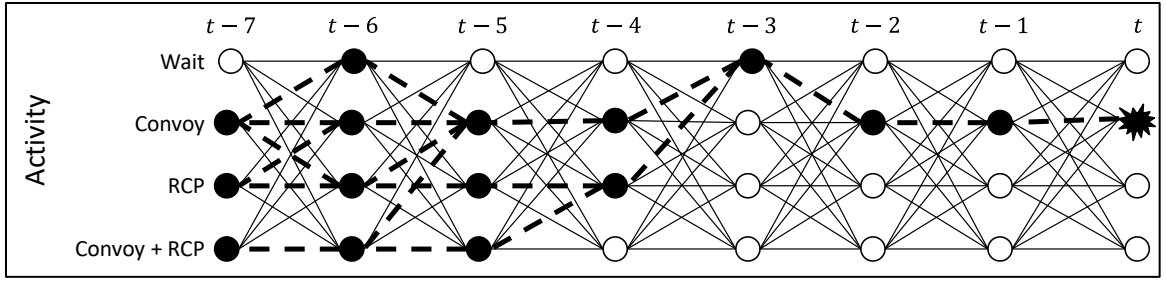


Figure 44: Example of a subset of possible activity sequences

Figure 44 expands the time series to eight time steps such that there are $4^7 = 16,384$ possible activity patterns (paths) leading from $(t - 7)$ to the attacked state at time (t) , but there are only 256 possible paths (1.56% of the total) that match from $t - 3$ to t . If it can be learned that the attacker favored attacking some subset of the total possible activity patterns (such as A, B, B, B) over others, this information could be used to augment a real-time decision algorithm.

Given the large number of locations and activities that might require monitoring, an efficient means of recording and processing multiple time series data streams is needed. This places a premium on developing an efficient representation of streaming network activities and a processing algorithm for effectively identifying high risk conditions. When the analytical process begins, the patterns, parameters, and constraints that govern the system are unknown, but to the extent they can be learned, they must be discovered by observing the system's behavior (Crutchfield, 1994).

Symbolic Aggregate Approximation

Many high level representations of continuous time series have been proposed for data mining, including Fourier transforms, wavelets, Eigenwaves, piecewise polynomial

models, and others (see Figure 45). But, the dimensionality of their representation using these approaches is the same as in the original data; therefore, they tend to scale poorly.

(Lin, Keogh, Lonardi, & Wei, 2007)

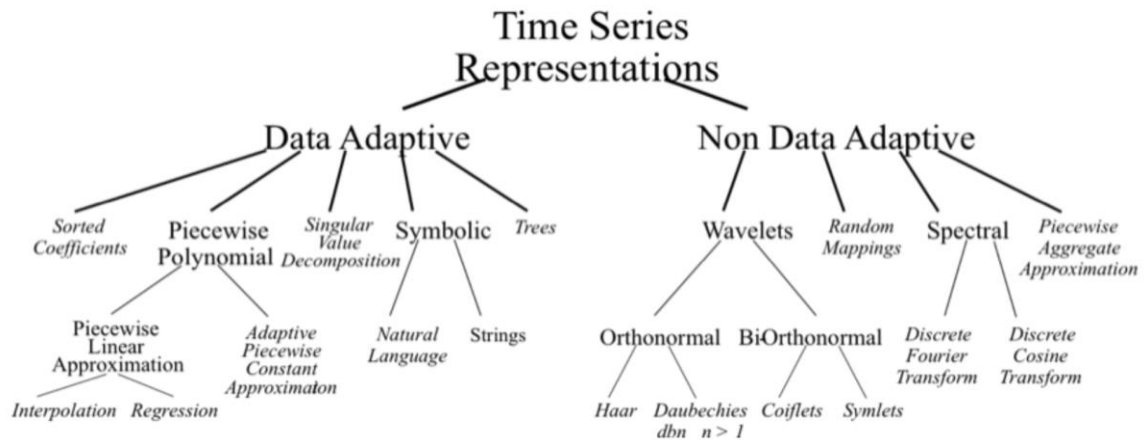


Figure 45: A hierarchy of various time series representations in the literature (Lin, Williamson, Borne, & DeBarr, 2012)

Symbolic Aggregate Approximation (SAX) is a time and space efficient method for recording continuous time series data while providing a means to process and compare lengthy activity sequences with minimal computational overhead. Further, with such recorded chronological data it is important to identify frequently repeated subsequences (referred to as motifs) which may have significance in the attacker's decision process (Lin, Keogh, Lonardi, & Wei, 2007; Nguyen, Ng, & Yew-Kwong, 2014). The next two sections summarize two potentially useful methods for accomplishing this using symbolic data streams.

T-Patterns

The T-Patterns algorithm uses symbols to describe discrete events in the time series events. T-patterns occur when both hidden and manifest behavior patterns involve similar relationships in their structure that can be identified. To distinguish these patterns, it is possible to test for statistical significance against the null hypothesis that each of the pattern's components is independent and randomly distributed over time at the observed frequency (Magnusson, 2000). For example, in Figure 46 we see that the letter "b" occurs 5 times in the string of 41 characters.

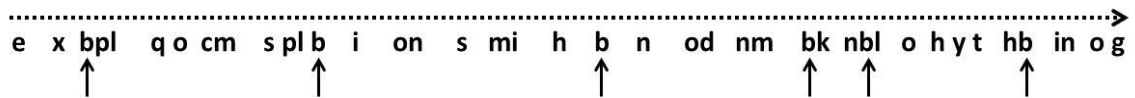


Figure 46: Example time series that appears to be random

Figure 47 shows that even the highly regular and repeated patterns of b-o-m-b can be difficult to detect when interspersed with other events. Especially in more complex cases, where patterns might occur over periods ranging from seconds to years, such patterns are easily missed by human observers - even with the use statistical detection software (Jonsson, 2011). This becomes more apparent when considering many event patterns are not strictly behavioral; that is they can be environmental and even psychological.

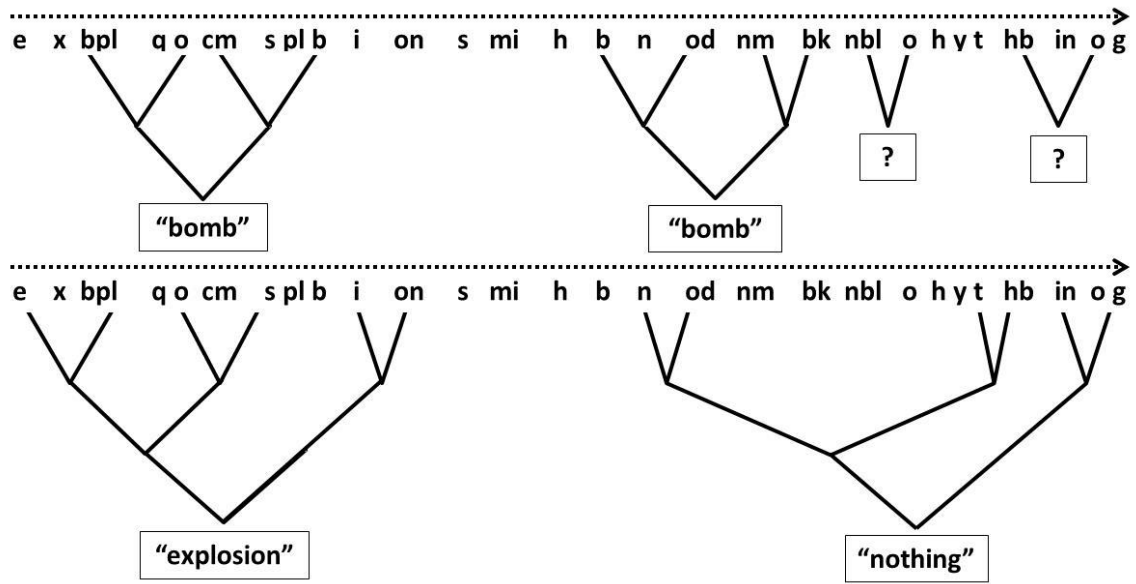


Figure 47: Example time series with hidden, but detectable, patterns

T-Patterns are a particular type of temporal pattern that relies on the hypothesis that patterns can be identified by their parts. Often between the components of a T-Pattern there may be a mixture of various other activities that vary greatly across multiple instances of the same T-Pattern. This makes them difficult to identify with detection routines that rely on a consistent event sequences. To overcome this, Magnusson developed a search algorithm that begins with a breadth-first search and groups pairs of events by the simplest patterns first, illustrated in Figure 48 (Magnusson, 2000). The algorithm is distinguished from traditional statistical approaches in that it only keeps the most complete patterns. Then, by building on them in a completion and selection process, it avoids the common problems associated with erroneously detecting partial and redundant portions of the same pattern (Jonsson, 2011).

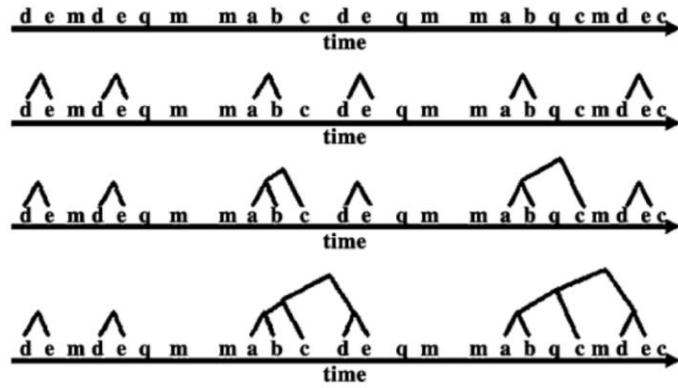


Figure 48: The formation of a T-Pattern from simple to complex (Jonsson, 2011).

Close Motifs

Another promising approach that might prove useful in attack pattern detection was more recently proposed by (Nguyen, Ng, & Yew-Kwong, 2014). In their paper they define a motif as a frequent pattern or subsequence in streaming data and introduce the concept of a *closed motif*. In a data stream, they state that a motif is *closed* if it is not a subsequence of any longer sequence having the same number of occurrences.

(Nguyen, Ng, & Yew-Kwong, 2014) provide a method for discovering closed motifs of variable length in a single scan (or online) through use of a flexible suffix tree structure that allows fast detection and classification of all repeated sequences. Notably, this technique works efficiently through their depth-first search and discovery algorithm without being confined to a predefined motif length (see Figure 49). This methodology also provides a means to avoid the common problems of being overwhelmed by redundant, uninteresting patterns.

To find important patterns, the event tree can be traversed from the tree root according to the necessary condition to exhaustively determine every repeated subsequence in the data stream. Then those that may be important can be identified according to a probabilistic model.

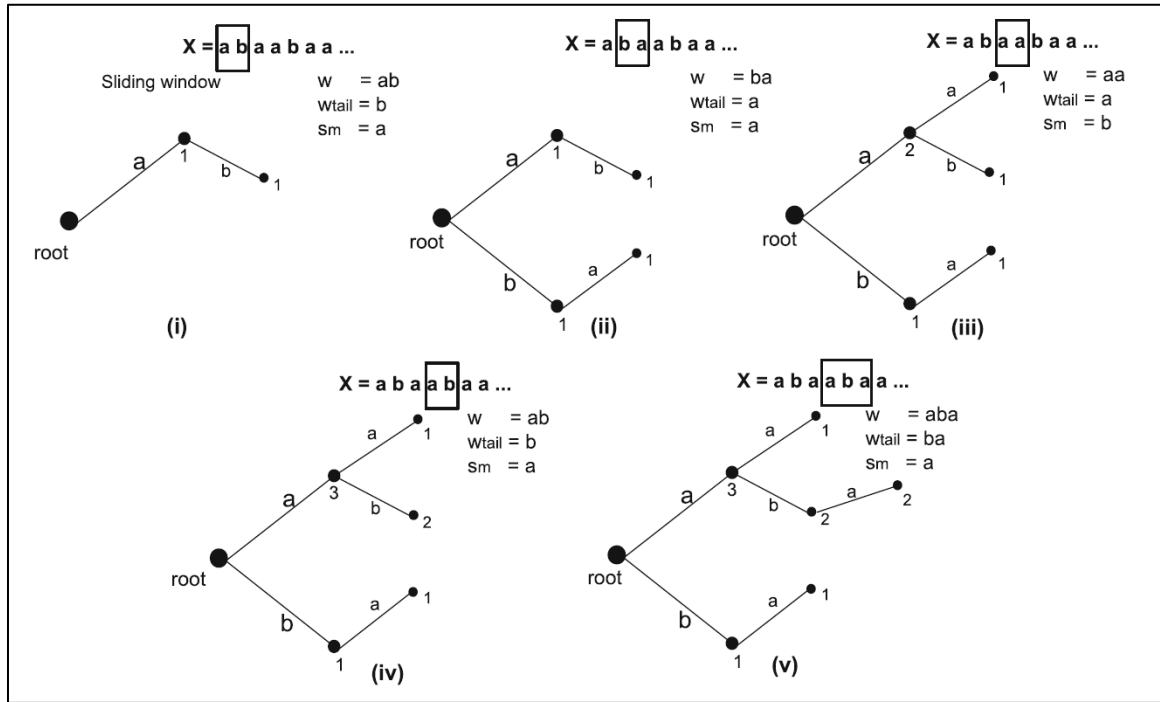


Figure 49: Constructing activity tree for closed motif detection³⁰ (Nguyen, Ng, & Yew-Kwong, 2014)

Future research could explore how a modified SAX procedure and one of these pattern detection algorithms might be combined to efficiently represent and analyze road network activity to improve RL policy applications in real-time. Then with improved

³⁰ Each tree node represents a word constructed by following a path from the tree's root to this node. The number below a node denotes its weight (or count).

data collection and availability, these approaches could provide significant improvement of the currently state of the art.

APPENDIX A: THE NINE PRINCIPLES OF IED COMBAT

1. Maintain an offensive mindset
2. Develop and maintain situational awareness
3. Stay observant
4. Avoid setting patterns
5. Maintain standoff
6. 360-degree security
7. Maintain tactical dispersion
8. Utilize blast/fragmentation protection
9. Know and use CREW (Counter Radio Electronic Warfare system)

Source: (Joint IED Defeat Organization, 2010)

APPENDIX B: EXCERPT FROM CONVOY OPERATIONS HANDBOOK

Movement Control (U.S. Marine Corps, 2001)

Movement control is the planning, routing, scheduling, and control of personnel and cargo movements over lines of communication (LOCs).

The MAGTF (Marine Air Ground Task Force) commander may be required to establish a highway traffic regulation system or regulate the movements of units in accordance with a traffic regulation system of a senior headquarters. The military police of the Combat Service Support Element (CSSE), in coordination with the motor transport officer, develop highway regulation plans. If necessary, a traffic circulation plan, normally prepared as an overlay, is prepared and distributed.

It may include—

- Route restrictions, route designations, and direction of movements.
- Locations of unit boundaries, highway regulating points, traffic control points, and principal supply points.
- Major geographic features and light line.

To coordinate movements, the CSSE may also be required to establish a Mobility Control Center (MCC). The MCC plans, schedules, routes, and controls movement. When established, that organization would—

- Issue operating procedures for the highway/road net.
- Receive and process convoy clearance requests.
- Plan traffic routing.
- Coordinate traffic scheduling.
- Coordinate and approve movement credit for controlled routes.
- Establish movement priorities in accordance with the commander's guidance.
- Prepare and maintain road movement table and critical time and point graphs that monitor and control traffic movement.

APPENDIX C: BASE CASE POLICY ADAPTATIONS

Scenario	Myopic Lo Lo	Myopic Hi Lo		Myopic Lo Hi		Myopic Hi Hi	
Myopic Agent Action	Count of States	Count of States	Change from Lo Lo	Count of States	Change from Lo Lo	Count of States	Change from Lo Lo
Do Nothing	394	346	-2.86%	1078	40.71%	678	16.90%
RCP Only	0	0	0.00%	0	0.00%	0	0.00%
Ship 1	0	0	0.00%	0	0.00%	0	0.00%
Escort 1	0	0	0.00%	0	0.00%	0	0.00%
Ship 2	0	24	1.43%	0	0.00%	0	0.00%
Escort 2	6	6	0.00%	0	-0.36%	0	-0.36%
Ship 3	24	24	0.00%	0	-1.43%	0	-1.43%
Escort 3	10	10	0.00%	0	-0.60%	0	-0.60%
Ship 4	24	24	0.00%	0	-1.43%	0	-1.43%
Escort 4	14	14	0.00%	6	-0.48%	6	-0.48%
Ship 5	40	40	0.00%	0	-2.38%	0	-2.38%
Escort 5	16	16	0.00%	6	-0.60%	6	-0.60%
Ship 6	40	40	0.00%	0	-2.38%	24	-0.95%
Escort 6	16	16	0.00%	6	-0.60%	6	-0.60%
Ship 7	40	64	1.43%	24	-0.95%	24	-0.95%
Escort 7	16	16	0.00%	6	-0.60%	6	-0.60%
Ship 8	64	64	0.00%	24	-2.38%	24	-2.38%
Escort 8	16	16	0.00%	6	-0.60%	10	-0.36%
Ship 9	64	64	0.00%	24	-2.38%	24	-2.38%
Escort 9	16	16	0.00%	6	-0.60%	16	0.00%
Ship 10	704	704	0.00%	360	-20.48%	680	-1.43%
Escort 10	176	176	0.00%	134	-2.50%	176	0.00%

Scenario	RL Lo Lo	RL Hi Lo		RL Lo Hi		RL Hi Hi	
RL Agent Action	Count of States	Count of States	Change from Lo Lo	Count of States	Change from Lo Lo	Count of States	Change from Lo Lo
Do Nothing	1287	439	-50.48%	1366	4.70%	467	-50.48%
RCP Only	13	27	0.83%	0	-0.77%	6	0.83%
Ship 1	4	79	4.46%	0	-0.24%	72	4.46%
Escort 1	0	6	0.36%	0	0.00%	21	0.36%
Ship 2	7	20	0.77%	0	-0.42%	22	0.77%
Escort 2	4	6	0.12%	0	-0.24%	7	0.12%
Ship 3	7	32	1.49%	0	-0.42%	21	1.49%
Escort 3	5	3	-0.12%	0	-0.30%	7	-0.12%
Ship 4	10	36	1.55%	0	-0.60%	27	1.55%
Escort 4	1	7	0.36%	0	-0.06%	7	0.36%
Ship 5	16	35	1.13%	0	-0.95%	32	1.13%
Escort 5	3	7	0.24%	0	-0.18%	6	0.24%
Ship 6	22	33	0.65%	5	-1.01%	33	0.65%
Escort 6	6	7	0.06%	1	-0.30%	9	0.06%
Ship 7	40	41	0.06%	11	-1.73%	41	0.06%
Escort 7	13	10	-0.18%	4	-0.54%	9	-0.18%
Ship 8	56	59	0.18%	20	-2.14%	48	0.18%
Escort 8	17	13	-0.24%	10	-0.42%	13	-0.24%
Ship 9	59	76	1.01%	45	-0.83%	50	1.01%
Escort 9	13	18	0.30%	14	0.06%	13	0.30%
Ship 10	78	647	33.87%	161	4.94%	642	33.87%
Escort 10	19	79	3.57%	43	1.43%	126	3.57%

REFERENCES

- Ahner, D. K., & Spainhour, R. (2015). Development of Analytical Models of Blue Force Interaction with Improvised Explosive Device Incidents. *Military Operations Research*, 20(N2), 5-17.
- Akgun, I. (2000). *The K-group Maximum-Flow Network-Interdiction Problem*. Naval Postgraduate School. Monterey, CA.
- Ardohain, C. M. (2016). *IED Pattern Recognition Using Sinusoidal Models*. Monterey CA: Naval Post Graduate School, Operations Research Master's Thesis.
- Ashby, R. W. (2011). Variety, Constraint, and the Law of Requisite Variety. *Emergence: Complexity & Organization*, 12(1-2), 190-207.
- Associated Press. (1970, September 9). Booby Traps A Major Killer In Vietnam. *Saint Petersburg Independent*, pp. 12-A. Retrieved from <http://news.google.com/newspapers?nid=950&dat=19700909&id=Dd4LAAAAI BAJ&sjid=hFcDAAAAI BAJ&pg=3855,1850946>
- Balakrishna, P. (2009). *Scalable Approximate Dynamic Programming Models with Applications in Air Transport*. Systems Engineering and Operations Research. PhD. Dissertation. Fairfax, VA: George Mason University.
- Barbaro, M. L. (2012). *Statement before the US House of Representatives*.
- Barbaro, M. L. (2013, May 17). Improvised Explosive Devices are Here to Stay. *Washington Post*.
- Bellman, R. E. (1957). *Dynamic Programming*. Princeton, NJ: Princeton University Press.
- Bertsekas, D. P., & Tsitsiklis, J. (1996). *Neuro-Dynamic Programming*. Belmont, MA: Athena Scientific.
- Bertsekas, D., Tsitsiklis, J., & Wu, C. (1997). Rollout Algorithms for Combinatorial Optimization. *Journal of Heuristics*, 3(3), 245–262.
- Bhattacharyya, G. K., & Johnson, R. A. (1977). *Statistical Concepts and Methods*. New York, New York: John Wiley & Sons Inc.

- Boyd, J. R. (1976). *Destruction and Creation (unpublished brief)*.
- Boyd, J. R. (1986). *Patterns of Conflict, (unpublished brief)*.
- Boyd, J. R. (1987). *Organic Design for Command and Control, (unpublished brief)*.
- Boyd, J. R. (1995). *The Essence of Winning and Losing (unpublished brief)*. Retrieved August 7th, 2016, from www.dnipogo.org/richards/boyds_ooda_loop.ppt
- Center for Army Analysis. (2012). *Deployed Analyst History Report - Volume 1*. Fort Belvoir, VA: Center for Army Analysis.
- Center for Army Analysis. (2016). *Deployed Analyst Handbook*. Fort Belvoir, VA: Center for Army Analysis.
- Connable, B., Perry, W. L., Doll, A., Lander, N., & Madden, D. (2014). *Modeling, Simulation, and Operations Analysis in Afghanistan and Iraq*. Sant Monica, CA: RAND National Defense Research Institute.
- Cordesman, A. H. (2015). *Trends in Iraqi Violence, Casualties and Impact of War: 2003-2015*. Washington, DC: Center for Strategic Studies.
- Cormican, K. J. (1995). *Computational Methods for Deterministic and Stochastic Network Interdiction Problems, Masters Thesis*. Naval Post Graduate School, Monterey, CA.
- Crutchfield, J. P. (1994). *Is Anything Ever New? Considering Emergence in Complexity: Metaphors, Models, and Reality, SFI Series in the Sciences of Complexity XIX*. (G. Cowan, D. Pines, & D. Melzner, Eds.) Redwood City: Addison-Wesley.
- Danskin, J. (1962, November-December). Game Theory Model of Convoy Routing. *Operations Research*, 10(6), 774-785.
- Dantzig, G. B., & Ramser, J. H. (1959). The Truck Dispatching Problem. *Management Science*, 6(1), 80-91.
- DeGregory, K. (2007). *Optimization-based allocation of force protection resources in an Asymmetric Environment, Operations Research Master's Thesis*. Cambridge, MA: Massachusetts Institute of Technology.
- Denardo, E. V. (2003). *Dynamic Programming Models and Applications*. Mineola, NY: Dover Publications.
- Dhami, H. S., Pande, B. P., & Tamata, P. (2013, February). Reduction of Maximum Flow Network Interdiction Problem: Step towards the Polynomial Time Solutions. *International Journal of Applied Information Systems*, 5(3).

- Dunlap, C. J. (1998). *Challenging the United States Symmetrically and Asymmetrically: Can America Be Defeated? (Ed.), Preliminary Observations: Asymmetrical Warfare and the Western Mindset*. Carlisle, Pennsylvania: Strategic Studies Institute.
- Eckley, D. C., & Curtin, K. M. (2013). Evaluating the Spatiotemporal Clustering of Traffic Incidents. *Computers, Environment and Urban Systems*, 37, 70-81.
- Feinberg, E. A., & Lewis, M. E. (2016). On the Convergence of Optimal Actions for Markov Decision Processes and the Optimality of (s,S) Policies for Inventory Control (version 2). *Cornell University*, 1-27.
- Ford, L. R., & Fulkerson, D. R. (1962). *Flows in Networks, R-375-PR*. Santa Monica, CA: Rand Corporation.
- Fulkerson, D. R., & Harding, G. C. (1977). Maximizing the Minimum Source-Sink Path Subject to a Budget Constraint. *Mathematical Programming*, 13, 116-118.
- Garaux, J. (2010, January). The IED Fight. *The Marine Corps Gazette*, 94(1), pp. 8-12.
- Golden, B. L. (1978). A Problem in Network Interdiction. *Naval Research Logistics Quarterly*, 25, 711-713.
- Goodson, J. C. (2010). *Solution Methodologies for Vehicle Routing Problems with Stochastic Demands, PhD Thesis*. University of Iowa.
- Gosavi, A. (2009). Reinforcement Learning: A Tutorial Survey and Recent Advances. *Journal on Computing*, 21, 178-192.
- Grey, C. S. (1999). *Modern Strategy*. Oxford, England: Oxford University Press.
- Harris, T. E., & Rose, F. S. (1955). *Fundamentals of a Method for Evaluating Rail Net Capacities*. Santa Monica, CA: The RAND Corporation.
- Heylighen, F., & Joslyn, C. (2001). *Cybernetics and Second-Order Cybernetics - in Encyclopedia of Physical Science & Technology (3rd ed.)*. (R. A. Meyers, Ed.) New York: Academic Press.
- Hinton, T. G. (2010). *A Thesis Regarding The Vehicle Routing Problem Including a Range of Novel Techniques for its Solution*. University of Bristol, Engineering Department of Computer Science.
- International Security Assistance Force (ISAF). (2013). *ISAF Monthly Data: Trends through December 2012*. International Security Assistance Force (ISAF). Retrieved from <http://augengeradeaus.net/2013/02/falsche-zahlen-aus-afghanistan/>

- Israeli, E., & Wood, K. (2002). Shortest-Path Network Interdiction. *Networks*, 40(2), 97-111.
- Israeli, E. (1999). *System Interdiction and Defense, Doctoral Dissertation*. Monterey, CA: Naval Post Graduate School.
- Joint IED Defeat Organization. (2010). *Counter-IED Smart Book, For Pre-Deployment and Field Use*. Kwikpoint.
- Joint IED Defeat Organization. (2010). *Joint Improvised Explosive Defeat Organization Annual Report*. US Government.
- Jonsson, G. K. (2011). *Hidden Temporal Pattern in Interaction*. University of Aberdeen.
- Jonsson, G. K. (2011). *Hidden Temporal Pattern in Interaction*. Aberdeen: PhD Thesis, University of Aberdeen.
- Kailbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, 4, 237-285.
- Karp, R. M. (1972). Reducibility among combinatorial problems. In R. E. Miller, & J. W. Thatcher (Eds.), *In Complexity of Computer Computations* (pp. 85-103). New York: Plenum Press.
- Keefe, R., & Sullivan, T. (2011). *Resource-Constrained Spatial Hot Spot Identification*. NATIONAL SECURITY RESEARCH DIV. ARLINGTON VA: RAND CORP.
- Kolesar, P. (2009). *Poisson Trending of IED Event Frequencies*. Fairfax, VA: George Mason University, JIEDDO Annual Technical Report.
- Kolesar, P., Leister, K., & Woodaman, R. (2008). *A Time Series Analysis of Improvised Explosive Device Incidence In Red Blue Interaction Modeling for the Joint Improvised Explosive Device Defeat Organization*. Fairfax, VA: George Mason University C4I Center.
- Kolesar, P., Leister, K., Stimpson, D., & Woodaman, R. (2012, April). A Simple Model of Improvised Explosive Device Clearance. *Annals of Operations Research*.
- Koyak, R. (2009a). *Risk on Roads: A Modeling Approach (part 1)*. Monterey, CA: Naval Post Graduate School.
- Koyak, R. (2009b). *Risk on Roads: A Modeling Approach (part 2)*. Monterey, CA: Naval Post Graduate School.
- Koyak, R. (2010). *Risk on Roads: A Modeling Approach (part 3)*. Monterey, CA: Naval Post Graduate School.

- Kumar, S. N., & Panneerselvam, R. (2012). A survey on the Vehicle Routing Problem and its Variants. *Intelligent Information Management*, 4, 66-74.
doi:10.4236/iim.2012.43010
- Leister, K., & Hudson, T. (2009). *Route Clearance Team Scheduling, Final Report, Report Prepared Masters Degree Project Course*. Fairfax, VA: George Mason University.
- Lin, J., Keogh, E., Lonardi, S., & Wei, L. (2007). Experiencing SAX: a novel symbolic representation. *Data Mining and Knowledge Discovery*, 15(2), 107-144.
- Lin, J., Williamson, S., Borne, K., & DeBarr, D. (2012). *Advances in Machine Learning and Data Mining for Astronomy, Chapter One: Pattern recognition in time series*. Boca Raton, FL: CRC Press, Taylor & Francis Group.
- Lin, K., & Washburn, A. (2010). *The Effect of Decoys in IED Warfare. Report prepared for Joint IED Defeat Organization*. 5000 Army Pentagon, Washington D.C.
- Lu, Y., & Chen, X. (2007). False Alarm of Planar K-Function when analyzing Urban Crime Distributed along Streets. *Social Science Research*, 36(2), 611-632.
- Magnusson, M. S. (2000). Discovering Hidden Time Patterns in Behavior: T-Patterns and Their Detection. *Behavior Research Methods, Instruments, & Computers*, 32, 93-110.
- Marks, C. E. (2009). *Optimization-Based Routing and Scheduling of IED-Detection Assets in Operations Research Master's Thesis*. Cambridge, MA: Massachusetts Institute of Technology.
- Mill, J. S. (1843). *Being a Connected View of the Principles of Evidence, and the Methods of Science*. Harper & Brothers. Retrieved 01 10, 2016, from https://ebooks.adelaide.edu.au/m/mill/john_stuart/system_of_logic/index.html
- Nguyen, H.-L., Ng, W.-K., & Yew-Kwong, W. (2014). Closed Motifs for Streaming Time Series Classification. *Knowledge and Information Systems*, 41(1), 101-125.
- Okabe, A., & Yamanda, I. (2001). The K-Function Method on a Network and its Computational Implementation. 33, 271-290.
- Okabe, A., Satoh, T., Furuta, T., Suzuki, A., & Okano, K. (2008). Generalized Network Voronoi Diagrams: Concepts, Computational Methods, and Applications. *International Journal of Geographical Information Science*, 22(9), 965-994.
doi:10.1080/13658810701587891

- Okabe, A., Yomono, H., & Kitamura, M. (1995). Statistical Analysis of the Distribution of Points on a Network. *Geographical Analysis*, 27(2), 152-175.
- Osinga, F. (2001). *Science, Strategy and War: The Strategic Theory of John Boyd*. London, England: Routledge.
- Papadimitriou, C. H. (1977). The Euclidean travelling salesman problem is NP-Complete. *Theoretical Computer Science*, 4(3), 237-244.
- Pillac, V., Gendreau, M., Gueret, C., & Medaglia, A. L. (2013). A Review of Dynamic Vehicle Routing. *European Journal of Operational Research*, 225(1), 1-11. doi:10.1016/j.ejor.2012.08.015
- Powell, W. B. (2007). *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. New York: John Wiley & Sons.
- Powell, W. B., Simao, H. P., & Bouzaiene-Ayari, B. (2012). Approximate Dynamic Programming in Transportation and Logistics. *European Journal of Transportation and Logistics*, 1(3), 237-284.
- Powledge, T. R. (2005, May). Beating the IED Threat - A Company Commander's Perspective. *Marine Corps Gazette*, 89(5), pp. 64-66.
- Puterman, M. L. (1994). *Markov Decision Processes*. New York: John Wiley and Sons, Inc.
- Ropke, S. (2005). *Heuristic and Exact Algorithms for Vehicle Routing Problems*, PhD Thesis. Copenhagen, Denmark: University of Copenhagen.
- Salah, A. A., Pauwels, E., Tavenard, R., & Gevers, T. (2010). T-Patterns Revisited: Mining for Temporal Patterns in Sensor Data. *Sensors*, 10, 7496-7513.
- Scarf, H. (1960). *The optimality of (s,S) Policies in the Dynamic Inventory Problem*. Stanford, CA: Stanford University Press.
- Schum, D. (1994). *The Evidential Foundations of Probabilistic Reasoning*. Evanston, Illinois: Northwestern University Press.
- Shankar, A. (2014). *Spatial and Temporal Modeling of IED Emplacements against Dismounted Patrols*, PhD Dissertation. Fairfax, VA: George Mason University.
- Spivey, M. Z., & Powell, W. B. (2004). The Dynamic Assignment Problem. *Transportation Science*, 38(4), 399-419.

- Stafford, W. B. (2009). *Sequential Pattern Detection and Time Series Models for Predicting IED Attacks, Masters Thesis*. Monterey, CA: Naval Post Graduate School.
- Steinrauf, R. L. (1999). *Network Interdiction Models, Masters Thesis*. Monterey, CA.: Naval Post Graduate School.
- Stewart, J. (1999). *Calculus, Early Transcendentals (fourth edition)*. Pacific Grove, CA: Brooks/Cole Publishing Company.
- Stimpson, D. (2011, September). Thinking About IED Warfare. *Marine Gazette*, 95(9), pp. 35-42.
- Stimpson, D., & Ganesan, R. (2015, December). A Reinforcement Learning Approach to Convoy Scheduling on a Contested Transportation Network. *Optimization Letters*, 9(8), 1641-1657.
- Stockfish, D., & Yariv, E. (1970). *Dokshyc-Parafianow Memorial Book — Belarus (Sefer Dokshitz-Parafianov)*. (D. H. Mechnikov, Trans.) Tel Aviv: Association of Former Residents of Dokshyce-Parafianow in Israel. Retrieved November 2014, from <http://www.jewishgen.org/yizkor/dokshitsy/dok274.html>
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning, an Introduction*. Cambridge, MA: The MIT Press.
- Taleb, N. (2004). *Fooled by Randomness*. New York: Random House.
- Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, 46(2), 234-240.
- Toth, P., & Vigo, D. (2001). *The Vehicle Routing Problem*. Philadelphia, PA: Society for Industrial and Applied Mathematics.
- U.S. Army Counterinsurgency Center. (2011). *Counterinsurgency Lessons Learned*. Fort Leavenworth, KA: U.S. Army.
- U.S. Department of Defense. (2010). *Joint Publication 1-02 Department of Defense Dictionary of Military and Associated Terms*. Washington DC.
- U.S. Department of Defense. (2014). *Joint Publication 3-10, Joint Security Operations in Theater*. Washington DC: US Department of Defense.
- U.S. Department of Defense. (2015). *Enhancing Security and Stability in Afghanistan*. Report to Congress, December, 2015.

- U.S. Department of the Army. (2009). *FM 3-24.2 Tactics in Counterinsurgency*. US Department of The Army.
- U.S. Department of the Army. (2014). *FM 3-24/MCWP 3-33.5 Insurgencies and Countering Insurgencies*. Washington DC: US Department of the Army.
- U.S. Marine Corps. (2001). *Marine Corps Reference Publication, Convoy Operations Handbook*. Washington DC: United States Marine Corps.
- Unsal, O. (2010). *Two-Person Zero-Sum Network-Interdiction Game with Multiple Inspector Types, Masters Thesis*. Monterey, CA: Naval Post Graduate School.
- Washburn, A. (2006). *Continuous Network Interdiction, Report NPSOR-06-007*. Monterey, CA: Naval Postgraduate School.
- Washburn, A., & Ewing, P. L. (2011). Allocation of Clearance Assets in IED Warfare. *Naval Research Logistics*, 58, 180–187.
- Washburn, A., & Wood, K. (1995). Two-person zero-sum games for network interdiction. *Operations Research*, 43, 243–251.
- Weber, R. H. (2007). *Methods for conducting Military Operational Analysis, Analyzing, Intelligence, Surveillance & Reconnaissance. Chapter 15*. Alexandria, Virginia: Military Operations Research Society.
- Wevley, C. M. (1999). *The Quickest Path Network Interdiction Problem, Master's Thesis*. Monterey, CA: Naval Post Graduate School.
- Wollmer, R. D. (1964, November-December). Removing Arcs from a Network. *Operations Research*, 12, 807–1076.
- Wollmer, R. D. (1970, June). Interception in a Network. *Naval Research Logistics*, 17(2), 207-216.
- Wood, K. R. (1993). Deterministic Network Interdiction. *Mathematical and Computer Modeling*, 17(2), 1-18.
- Xie, Z., & Yan, J. (2008). Kernel Density Estimation of Traffic Accidents in a Network Space. *Computers, Environment, and Urban Systems*, 35(5), 396-406.
- Yamanda, I., & Thill, J. (2007). Local Indicators of Network-Constrained Clusters in Spatial Point Patterns. *Geographical Analysis*, 39(3), 268-292.

BIOGRAPHY

Daniel E. Stimpson graduated from Helix High School, San Diego, California, in 1984. He enlisted in the U.S. Marine Corps before earning a Bachelor of Science degree (with honors) from the United States Naval Academy in 1992. He served as a Marine Officer until 2008 which included earning a Master of Science degree in Operations Research from the Naval Postgraduate School in Monterey, California in 2005. Upon retirement from the Marine Corps, he began research and analytical work on the IED problem in Iraq and Afghanistan which included a deployment to Regional Command Southwest in Helmond, Afghanistan. He is married to Susan and the father of Caroline and William. He currently serves as a civilian Marine with Headquarters, Marine Corps.