REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
30-09-2018	Final Report	04/01/15 - 09/30/18

4. TITLE AND SUBTITLE
Separating Cognition Performance from Execution Environment (SCoPE)

5a. CONTRACT NUMBER

5b. GRANT NUMBER
N00014-15-1-2376

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Ortiz-Peña, Héctor, J, Ph.D.;
McConky, Katie, Ph.D.;
Sudit, Moises, Ph.D.;

5d. PROJECT NUMBER
07993

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
CUBRC, Inc. / 4455 Genesee Street, Suite 106, Buffalo, NY 14225
Rochester Institute of Technology / 81 Lomb Memorial Dr, Rochester, NY 14623

8. PERFORMING ORGANIZATION REPORT NUMBER
CUBRC-SCOPE-FR-001

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research

10. SPONSOR/MONITOR'S ACRONYM(S)
ONR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for Public Release; Distribution Is Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The goal of SCoPE was to predict mission plan performance, taking into consideration mission plan characteristics as well as cognitive and environmental factors influencing the plan. Given the large number of features that go into any one particular mission, the team proposed to use a conceptual spaces based model to identify features spaces that correspond to successful missions. To test proposed algorithms, the project initially aimed to use datasets related to HADR missions. An HADR simulation model was created using AnyLogic for the OtK program, and this model was used to generate datasets for the SCoPE program. Developing a validated simulation model, capable of generating large sets of simulation runs, proved challenging and largely unsuccessful. Highlights of the work with the HADR simulation model are presented first, including lessons learned from working with complex simulation models.

15. SUBJECT TERMS
Mission Plan Performance Prediction, Conceptual Spaces, Feature Engineering

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Hector J Ortiz-Pena, PhD
U	U	U	SAR	28	19b. TELEPHONE NUMBER *(Include area code)* (716) 204-5196

CUBRC
Advantage through technology

R·I·T

Final Report for FY 2018 - RIT

Separating Cognition Performance from Execution Environment (SCoPE^2)

Prepared for:
Office of Naval Research (ONR)
Mr. Martin Kruger
ONR Code 30
875 North Randolph Street
Arlington, VA 22203-1995
martin.kruger1@navy.mil

Principal Investigator:
Dr. Moises Sudit
4455 Genesee Street
Buffalo, NY 14225
716-204-5155 (Office)
716-204-5448 (Fax)
sudit@cubrc.org

Technical Contact:
Mr. Héctor J. Ortiz-Peña
4455 Genesee Street
Buffalo, NY 14225
716-204-5196 (Office)
716-204-5448 (Fax)
hector.ortiz-pena@cubrc.org

Co-PI:
Katie McConky
81 Lomb Memorial Dr
Rochester, NY 14623
585-475-6062 (Office)
585-475-2520 (Fax)
ktmeie@rit.edu

Subcontractor: Rochester Institute of Technology (Co-PI: Dr. Katie McConky)

Period of Performance: FY15 - FY18

Table of Contents

1 Introduction

This report covers the final phases of the SCoPE^2 project. The goal of SCoPE^2 was to predict mission plan performance, taking into consideration mission plan characteristics as well as cognitive and environmental factors influencing the plan. Given the large number of features that go into any one particular mission, the team proposed to use a conceptual spaces based model to identify features spaces that correspond to successful missions. To test proposed algorithms, the project initially aimed to use datasets related to HADR missions. An HADR simulation model was created using AnyLogic for the OtK program, and this model was used to generate datasets for the SCoPE^2 program. Developing a validated simulation model, capable of generating large sets of simulation runs, proved challenging and largely unsuccessful. Highlights of the work with the HADR simulation model are presented first, including lessons learned from working with complex simulation models.

To be able to try our methodology on a larger dataset we turned to using sports data as a surrogate. Section 3 discusses our initial SCoPE^2 methodology, including the sports dataset chosen for study. Lessons learned are highlighted. Section 4 discusses results of our revised SCoPE^2 methodology, and resultant performance, taking into consideration the need to capture non-linear relationships between features. With the revised methodology we were able to predict sport game outcomes with an accuracy up to 73%. Results on feature sensitivity to predicted outcome are presented in Section 4 as well. The report is concluded in Section 5 with a discussion of potential future work.

2 HADR Simulation Final Results

The number of weather related disasters has increased in the most recent decade versus the decade before. These disasters on average affected over 435,000 people per event. With such large number of affected human lives, it is imperative to have the ability to quickly and efficiently respond to the affected area in order to supply food and water, provide first aid, and return the region to self-sufficiency. Critical to these tasks is the routing of supply vehicles and the collection of intelligence information. This work used an agent based simulation of a hurricane hitting New Orleans to evaluate the effects of intelligence gathering tasks on mission performance measures. The results show that when intelligence tasks are coordinated with supply vehicle routing that improvements to food supply, travel times, and road network knowledge can be observed.

One of the challenges of operating efficient HADR missions is that communication between operational and intelligence tasks is often hampered. Tasks such as delivering and distributing food and water, repairing infrastructure, providing first aid, and keeping law and order are often executed in parallel with, but not coordinated with, intelligence gathering tasks such as identifying damaged roads, tracking thieves, and reporting distribution centers' food status (Howden, 2009). Each operational task and each intelligence gathering task is often optimized independent of the other task's needs. Using a hurricane disaster relief scenario several intelligence gathering approaches were analyzed. Results suggest there exists a benefit to integrating operational and intelligence tasking.

2.1 Simulation Details

An agent based simulation framework was used to evaluate the emergency response to a hurricane event. This section provides details on the simulation framework developed for use in this study.

The simulation scenario takes place 6 days after a simulated (fictitious) hurricane has hit the coastal region of a country. The hurricane has simultaneously destroyed buildings, facilities and numerous roads, making travel difficult and leaving hundreds of thousands of people without food, water and electricity. A humanitarian aid base, located near five population centers, has been set up and stocked with food and water by two air assets which fly constantly, uninterrupted, between the base and an offshore aircraft carrier. There are five population centers and each may experience various amounts of damage from the hurricane, and thus require different types and amount of assistance. The locations of the five population centers and HADR base can be seen in Figure 1. The goal of the simulation is to return the population centers to self-sufficiency by repairing critical infrastructure. During the repair period food and water must be delivered to the population centers to prevent hunger and dehydration in the vulnerable population.

Various HADR units are tasked with delivering food and water, identifying, tracking, and detaining criminals, finding and repairing broken roads, and repairing critical infrastructure. These units include maintenance teams, supply vehicles, aerial surveillance assets, security teams and ground intelligence units. Maintenance teams are assigned to repair critical facilities such as hospitals, power plants, water treatment plants and roads. Supply vehicles transport food and water from the main HADR base to the five population centers. An unmanned aircraft systems (UAS) (e.g., a small tactical unmanned aircraft systems (STUAS)) fly over the area of operations to identify damaged roads and track criminals. Security teams actively track and capture criminals. Once captured, criminals are neutralized for a period of time. Finally, ground surveillance teams provide intelligence from the ground and report on the state of roads and facilities as well as on the supply levels of distribution centers. Reports from the ground surveillance teams are considered 100% accurate.

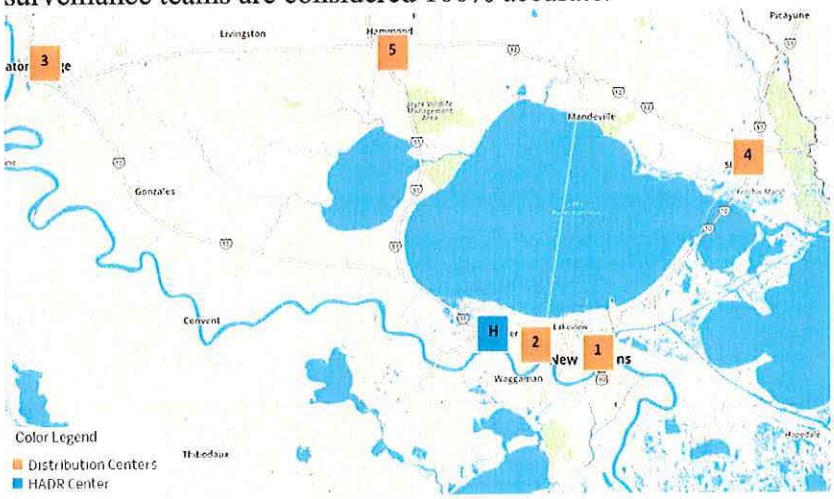

Figure 1-Distribution Center and HADR base locations.

While the main mission of the HADR units is to provide food and water to the local population and return the region to self-sufficiency by repairing critical infrastructure, there are several challenges that make these tasks more difficult. Firstly, a significant number of roads are damaged by the hurricane, and the locations of these damaged roads are unknown at the start of the simulation. If a vehicle encounters a damaged road the vehicle will travel 10 times slower than normal, therefor avoiding damaged roads is significantly advantageous for HADR units. Damaged roads can be discovered by aerial surveillance assets, and repaired by maintenance crews. In addition to a highly uncertain road network, HADR units face criminal elements that traverse the network and steal from supply vehicles and distribution centers. These criminal elements can be spotted and tracked by the aerial surveillance assets, ground surveillance teams, and security agents, and can be detained for a period of time by security forces. Finally, HADR units face the impact of corruption: reports received from NGOs on the state of supply levels at distribution centers are plagued with inaccuracy.

The main simulation engine is built using the AnyLogic agent based modelling tool (AnyLogic, 2017). The simulation component models asset dynamics and their interaction with the virtual environment. For example, making sure vehicles travel along roads. Logic to control the tasking of individual HADR assets is external to the simulation, and comprises optimization algorithms to assign tasks and routes to assets throughout the duration of the simulation. Once tasking is determined, the assignments are communicated to the simulation and distributed to the various HADR units that then execute the plans. The optimization algorithms used were developed outside of this work.

The architecture allows for the testing of two capabilities central to this current study: the ability of unmanned aircraft systems (UAS) to optimally select routes that maximize information gain, and the ability of supply vehicles to efficiently determine routes in the presence of a constantly changing road network and distribution center supply levels. The UAS route optimization algorithm can either route UAS to survey locations that will maximize information gain or route UAS in a static race track pattern that will ensure the entire area is surveyed in a periodic manner. Figures 2(a) and 2(b) illustrate the difference in UAS routing patterns over the course of a seven day period when information gain routing versus race track routing is employed. The information gain route optimization algorithm is discussed in detail in Ortiz-Peña et al. (2015).

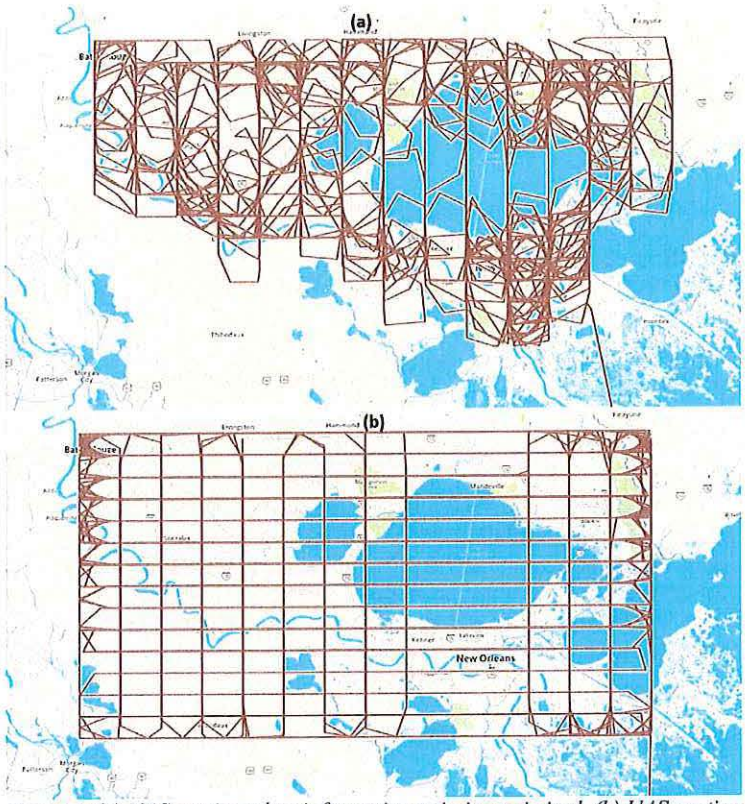

Figure 2-(a) UAS routing when information gain is maximized. (b) UAS routing when a race track pattern is used.

This study adds to previous research by statistically evaluating the benefit of information gain based routing using a simulation framework. The information gain based routing algorithm is expected to prioritize surveillance areas such that the value of information gained is maximized for the other HADR mission assets. One benefit of such routing is that damaged roads may be discovered before a supply vehicle is forced to traverse one. If a damaged road is discovered before a supply vehicle arrives to it, the supply vehicles can be dynamically rerouted, potentially reducing significantly travel times. Supply vehicles that encounter an undiscovered damaged road are forced to remain on their path and move slowly through the damaged area.

Arguably, one of the most important tasks of a humanitarian assistance mission is the supply of both food and water to the impacted regions. In order to control the assignment of supply trucks, a mathematical model was developed for optimizing supply truck assignments under variable mission objectives, road network uncertainty, and distribution center supply level uncertainty. Since the algorithm developed was to be used in a simulation context, the speed of algorithm execution was of utmost importance, but this speed would also be beneficial in a live HADR scenario. A two-phase approach was taken to route supply trucks throughout the regions. The first phase assigned supply truck routes at the start of the simulation, or iteration 0, or whenever a full supply truck requested a new route. The second phase was used after a supply truck made a delivery and was empty. The second phase ensured vehicles stop by the HADR base to resupply before making a subsequent delivery. A trip-wire system was used to trigger the use of the models. The two trip wire conditions beyond initial assignments were: i) a truck successfully made a delivery and needed a new assignment, or ii) a truck encountered a damaged road along

its predefined route. The ability to reroute supply truck assets mid-route allowed trucks to avoid damaged road segments and speed up their deliveries. Trucks could only request to be rerouted if the road they encountered had already been identified as damaged by the intelligence assets. To facilitate the fast rerouting of vehicles, a k-shortest path cache was maintained for the road network, and was updated as road network information came in from the intelligence assets.

2.2 Experiment Details

Using the HADR simulation described in the previous sections, a full factorial experiment design was conducted to understand the impact of information gain based intelligence collection and tasking operational assets with intelligence collection. The experiment manipulated three factors, each at 2 levels, for a total of 8 simulation runs per replication. The factors manipulated and their experiment levels can be seen in Table 2. The simulation was run over a period of 7 days. Three total replications were completed, where the damaged road network was changed for each simulation replication. The road damage was set at thirty percent. Each run contained one UAS (aerial surveillance asset), two ground surveillance teams, two supply vehicles for food, two supply vehicles for water, two security teams, two road maintenance teams, and four criminals.

Table 1-Factors and Their Experiment Levels

Factor	Level(s)
UAS Routing	Race Track / Info Gain
Ground Surveillance Teams Report Roads	Yes / No
Ground Surveillance Teams Report Distribution Centers	Yes / No

For each run several performance measures were captured relating to the state of the road network, the ability of HADR assets to travel the road network efficiently, criminal activity, and population well-being. While thousands of statistics were captured for each run, only a few will be reviewed here. Table 3 contains a description of each of the four categories of performance measures. An analysis of variance (ANOVA) was completed against each of these performance measures to determine which factors had significant impacts on results. All statistical tests are reported at a confidence interval of 95%. When performance measures varied over time, statistical tests were computed at daily increments and at the end of the simulation period. Statistical significance found at any of the daily increments is reported in the results.

Table 2-Experiment Performance Measures

	Performance Measure	Description
Road Network Statistics	Total Roads Discovered	• The total number of roads detected in the network over time.
	Partially Passable Roads Discovered	• The total number of partially passable roads discovered over time.
	Roads Repaired	• The total number of partially passable roads repaired over time.
Network Travel Statistics	# Ground surveillance teams Vehicles Slowed	• The cumulative number of times Ground surveillance team vehicles have been delayed by partially passable roads over time.
	# Maintenance Vehicles Slowed	• The cumulative number of times maintenance vehicles have been delayed by partially passable roads over time.
	# of Supply Vehicles Slowed	• The cumulative number of times supply vehicles have been delayed by partially passable roads over time.
	Max Travel Time to Distribution Center 3	• The maximum amount of time a delivery took to reach distribution center 3 (the farthest distribution center from the HADR base).
Criminal Statistics	# of Times Criminals Captured	• The number of criminals captured over time.
	Food Stolen	• The number of food units stolen by the criminals
	Hostile Incidents	• The number of times criminals stole food.
	The percent of time criminals tracked.	• The amount of time intelligence assets are actively tracking the location of criminals.
Population Well Being	Percent of Time Population is Well Fed	• The percent of time the distribution centers have 3 days of food supplies on hand for 90 % of their population.
	Percent of Time Population is Hungry	• The percent of time the distribution centers can feed greater than 50% of their population for three days, but less than 90%.
	Percent of Time Population is Starving	• The percent of time the distribution centers can feed less than 50% of their population for three days.
	Min Distribution Center Food Level	• The minimum food level observed across all distribution centers over the course of the simulation.
	Max Difference Across Distribution Centers	• The maximum difference in food levels observed at any point in time across distribution centers.

2.3 Results

The UAS routing algorithm had significant impacts on road network, travel and criminal statistics. With respect to improving road network knowledge and state using the Information Gain routing algorithm outperformed using a race track routing algorithm. Figure 5(a) illustrates the impact on road network discovery when the information gain algorithm is used. Initially both algorithms discover a similar amount of roads, but beyond the first 24 hours the information gain algorithm begins to see and maintain an advantage over the race track algorithm. A similar pattern can be seen in Figure 5(b) with regards to the number of damaged roads discovered although this effect was not significant in the runs completed. Finally, the number of roads repaired when using the information gain algorithm was significantly higher than using the race track algorithm as illustrated in Figure 5(c).

The improved road network as a result of the information gain algorithm resulted in improvements to vehicle travel throughout the simulation run as well. The information gain algorithm had a significant effect on reducing the number of times ground surveillance teams and maintenance vehicles were slowed down by damaged road. Supply vehicles also showed less slow downs although the UAS routing factor was not statistically significant in this case. Figure 6 shows the number of slowed vehicles over time using both the information gain algorithm and the race track routing algorithm. Finally, when the information gain algorithm was used, the maximum time to deliver supplies to distribution center 3, the farthest distribution center, was significantly reduced over the race track routing algorithm.

Finally, the UAS routing algorithm had significant effects associated with some criminal related statistics. When the information gain algorithm was used the number of hostile attacks by criminals decreased while the percent of time spent tracking criminals increased. Figure 7(a) shows that on average when using the information gain algorithm slightly more criminals were captured, however this result was not statistically significant.

Table 3-ANOVA Results for significant single factor effects. An H represents a strong significance with p-value < 0.05, while an L represents a weak effect, with p-value <=0.1.

	Performance Measure	UAS Routing	Ground Surveillance Teams Roads	Ground Surveillance Teams Distribution
Road Network Statistics	Total Roads Discovered	H		
	Partially Passable Roads Discovered			
	Roads Repaired	H		
Network Travel Statistics	# Ground Surveillance Teams Vehicles Slowed	L	H	
	# Maintenance Vehicles Slowed	H		
	# of Supply Vehicles Slowed		H	
	Max Travel Time to Distribution Center 3	H		H
Criminal Statistics	Number of Time Criminals Captured			
	Food Stolen	H	H	
	Hostile Incidents	H	H	
	The percent of time criminals tracked.	L		H
Population Well Being	Percent of Time Population is Well Fed		H	H
	Percent of Time Population is Hungry			H
	Percent of Time Population is Starving		L	H
	Min Distribution Center Food Level			H
	Max Difference Across Distribution Centers			L

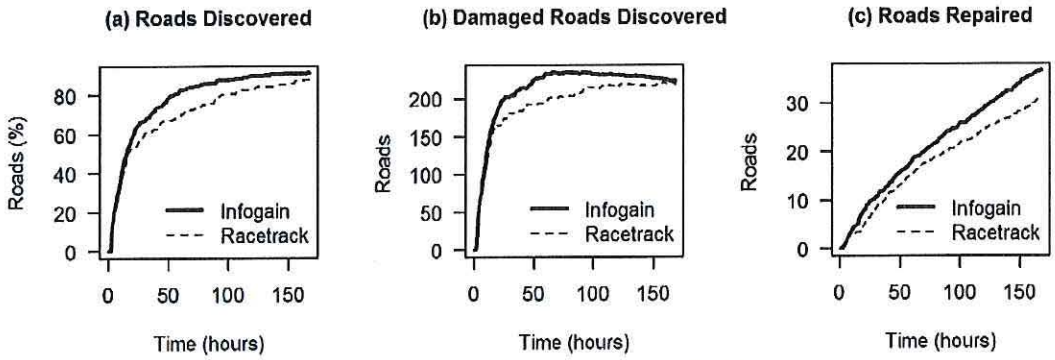

Figure 3-Average road network statistics over time versus UAS routing algorithm for: (a) the total number of roads discovered in the network, (b) the number of partially passable roads discovered, and (c) the total number of roads repaired.

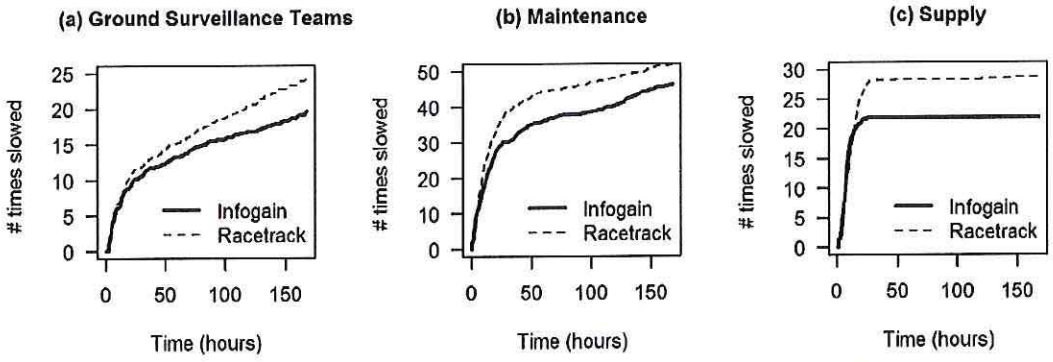

Figure 4-Number of times HADR mission vehicles are slowed by partially passable roads versus UAS routing algorithm for: (a) Ground Surveillance Teams , (b) Maintenance Vehicles, (c) Supply Vehicles.

Figure 5-Criminal statistics for (a) average number of criminals captured over time versus the UAS routing algorithm, (b) the average number of food units stolen over time versus Ground Surveillance Teams reporting on distribution centers, and (c) the amount of

When ground surveillance teams report partially passable roads they discover while traversing the network effects on vehicle movements, criminal activities, and population food levels can be observed. As the ground surveillance teams report on the road network, the number of supply vehicles and ground surveillance teams that are slowed by damaged roads decreases. Similar to the information gain algorithm for UAS routing, the number of times that criminals steal food decreases, but the amount stolen actually increases as ground surveillance report roads. Finally, as ground surveillance teams report on the road network, the average amount of time that the population is in the well fed state increases.

The ground surveillance teams reporting on distribution center food levels had the most significant impacts to the population well-being of all factors evaluated. Interestingly, the ground surveillance teams' reports on distribution centers also had significant impacts to criminal tracking and travel times as well. Figure 8 highlights the impact of ground surveillance teams reporting on distribution centers. As distribution center statuses are known with more certainty, the supply vehicles are better able to prioritize their deliveries, resulting in a larger percent of the population being well fed, and a smaller portion of the population starving. In addition to improving the number of people well fed, the ground surveillance teams reporting on distribution

centers also reduced the variability seen between distribution centers. As ground surveillance reported on distribution centers there was a significantly lower maximum food level difference across the distribution centers, indicating distribution centers were more evenly supplied with food. As ground surveillance reported on distribution centers the minimum amount of food available at any distribution center also increased, indicating that not only were distribution centers more evenly supplied, they were more evenly supplied consistently over time such that their food supply levels remained at consistently higher levels.

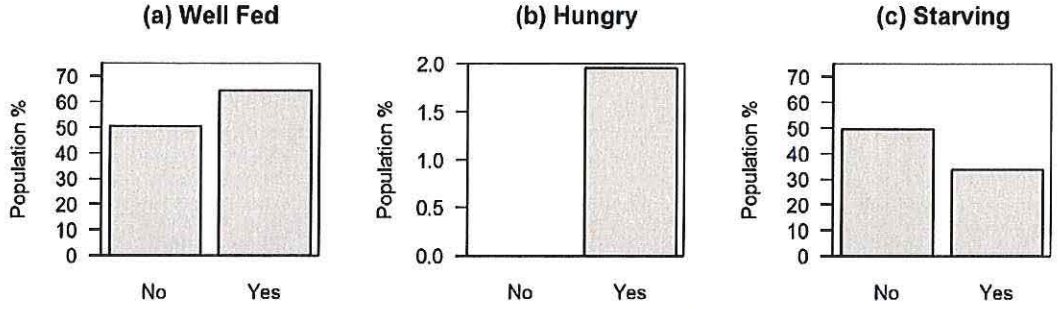

Figure 6-Effect of population hunger versus ground surveillance teams reports on distribution centers.(No is no reports on distribution centers, while Yes is when surveillance teams reported on distribution centers.)

When ground surveillance teams reported on distribution centers a few unexpected results occurred. Firstly, when ground surveillance teams reported on distribution centers, the percent of time criminals were tracked increased significantly. While tracking criminals may seem completely unrelated to distribution center reporting, the results make sense when looked into farther. When ground surveillance teams report on distribution centers, they spend several hours at the distribution centers performing inventory tasks. During this long window as they remain in one place, the likelihood that they encounter a criminal increases because criminals target the distribution centers. Secondly, when ground surveillance teams reported on distribution centers, the maximum time to deliver supplies to distribution center 3 went down. On the surface this too seems unrelated to ground surveillance teams reporting on distribution centers, but when ground surveillance teams report on distribution centers, they often travel similar routes to get to the distribution centers as supply vehicles. While traveling these roads, they can identify road damage that the supply vehicles can then avoid.

2.4 Lessons Learned and Publications

Insights gained from the simulation model developed were valuable in that they 1)validated the use of information gain based routing algorithms, and 2) provided previously undiscovered relationships between mission features and performance measures. For example, having ground teams reporting on distribution supply levels increases the number of criminals captured while at the same time decreasing the travel time to distribution centers. While the simulation model was able to provide valuable insights, gaining those insights by successfully developing and running a valid simulation model was extremely difficult. The validation process was further hampered by the long simulation times, so improvement cycles took roughly 1-2 weeks each to accomplish. Figure 7 shows the simulation modeling, analysis and validation cycle used. Of note in this cycle is that for a given set of runs, all runs needed to be completed, representing days of processing time, before results could be analyzed, only to find out that the simulation model was invalid. Additionally, the process of actually figuring out what was wrong was also challenging. For

example, the data may show that as more delivery trucks were added, less food was being delivered. Based on this odd finding, the analysts at RIT would need to dig into the raw second by second simulation data in order to identify that in fact trucks were getting stuck at a certain spot in the road network and were unable to move for long durations of the simulation. This detective work process could itself take hours, and then actually fixing the problem again took more time.

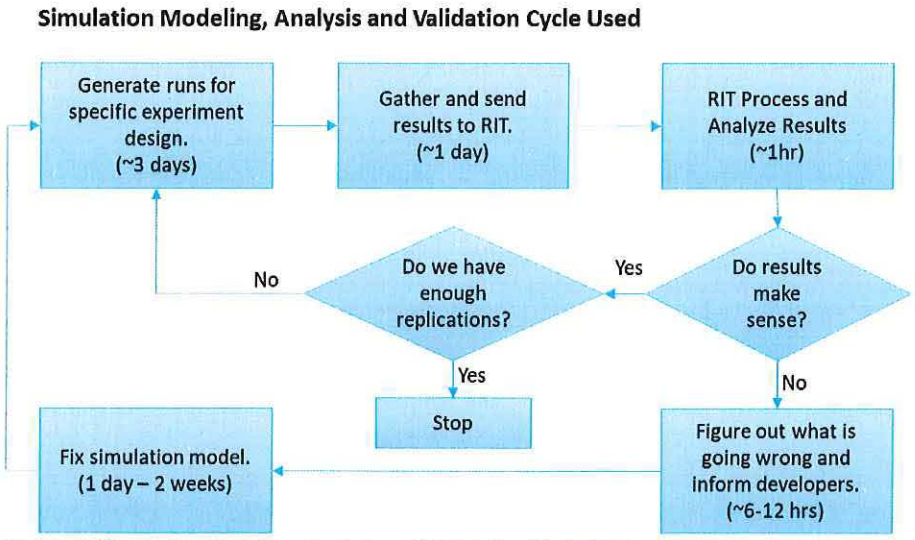

Figure 7- Simulation Modeling, Analysis and Validation Cycle Used

Figure 8 below is our proposed simulation modeling, analysis and validation cycle for future endeavors. Prior to the completion of any simulation runs, a validation plan needs to be created and maintained. This step could take a significant amount of time initially, but will pay off in the end because less detective work will be needed to identify issues, and less time will be spent generating useless simulation runs. The validation plan needs to be turned into validation software that will run in conjunction with the simulation model, validating the performance of the model as it runs in real time. If the simulation results do not pass validation checks, the runs can be immediately aborted, and the specific validation requirement can be immediately addressed.

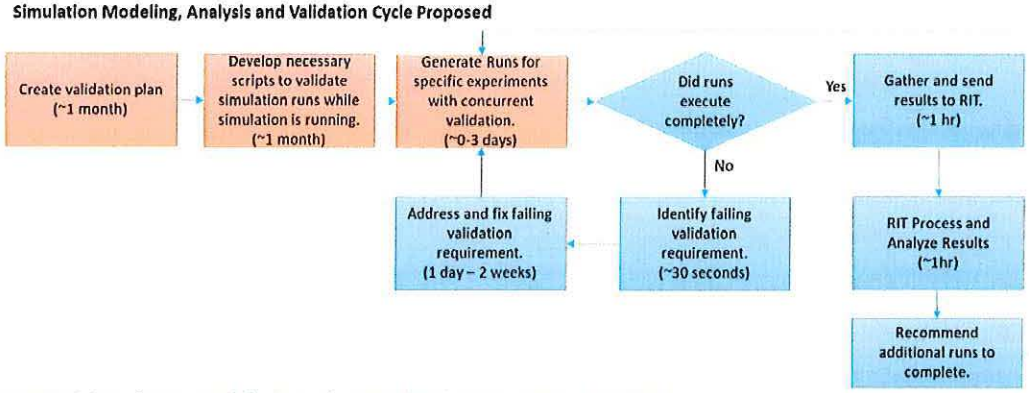

Figure 8-Simulation Modeling, Analysis and Validation Cycle Proposed

Table 4 contains a summary of the main lessons learned regarding generating large scale datasets via simulation.

Table 4 - Lessons learned regarding simulation modeling

Lesson Learned	Recommended Future Action
Simulation models of HADR or other missions can provide valuable insights into the mission planning process. Simulation can highlight unanticipated relationships between mission plan characteristics and performance measures.	• Continue to support simulation modeling, especially ways to develop and validate models faster.
Simulation Models requiring the use of Optimization Models are difficult to run in short time frames. Example – simulating 1 week of HADR effort required 5 hours of simulation time.	• When planning to use simulation to develop large scale datasets, limit the use of optimization models. If optimization models must be used, try using approximation heuristics instead. • Alternatively, explore using Generative Adversarial Networks (GANs), to produce surrogate simulation data when provided small sample datasets of actual simulation results.
Simulating road network damage needs to be done thoughtfully, as it is easy to damage a road network in such a way as to make a mission infeasible.	• Design a road damage algorithm that will ensure feasible paths. • Design the routing algorithms to more intelligently deal with infeasible situations, so vehicles do not remain motionless or stuck for extended periods of simulation time.
Validation is time consuming and challenging.	• A specific simulation validation strategy needs to be developed. This strategy should encompass every aspect of the simulation model, and should design specific tests for each simulation component. Physical systems (do cars stay on roads) and algorithmic logic (is my optimization algorithm producing reasonable results) must both be accounted for in test procedures. • A validation strategy should be similar to unit testing in software products. It should be automated, and done in real time as simulation results are being generated. Significant time was wasted running simulations that ended up needing to be thrown out due to errors that would have been obvious at the beginning of the runs had a system been in place to look for them.

The following publications and conference proceedings were generated from this work:

Conference Proceedings:
McConky, K., Ortiz-Pena, H., Poe, C., and Sudit, M. Evaluating the integration of operations tasks while optimizing ISR activities, Proceedings of SPIE: Defense and Commercial Sensing, Anaheim, Ca, April 2017.

Ortiz-Pena, H., Sudit, M., Coles, J., Poe, C., McConky, K. Automated Tracking and Assessment of Measures of Performances and Effectiveness for HADR Efforts, INFORMS International Meeting, Waikoloa Village, Hawaii, June 2016

Katie McConky, Hector Ortiz-Pena, and Moises Sudit. A Framework Supporting the Separation of Cognition Performance from Execution Environment. 2015 INFORMS Annual Meeting. INFORMS. Philadelphia, PA, 1 November

Pending Journal Publications:
McConky, K., Ortiz-Pena, H., Poe, C., Saxena, H., Sudit, M. Integration of Distribution and Intelligence Tasking for Efficient Supply Routing

3 SCOPE Initial Methodology

The SCoPE^2 project involved several hypotheses that built on one another. The goal was to systematically test these hypotheses using large scale datasets generated via simulation. The hypotheses were as follows:

Hypothesis 1: Mission plans can be characterized using conceptual spaces, and that concepts (groups of feature values) corresponding to different performance levels will be separable from one another.

Hypothesis 2: The separable concepts can be used to predict mission performance when given a plan.

Hypothesis 3: The separable concepts can be used to enable simple distance measure based approaches to mission optimization.

As discussed in section 2, the large scale simulation data set never materialized, so the team looked to existing datasets to test the above hypothesis. In lieu of military simulation data, sports data was used as a surrogate and will be discussed in section 3.1.

Figure 9 highlights the high level methodology that was initially proposed for the SCoPE^2 project. Figure 9 demonstrates the methodology, where the following three steps would be completed:
1) Cluster the data based on performance measures;
2) If separable performance clusters exist, generate feature spaces corresponding to concepts (mission spaces) that sre correlated with those performance clusters.

3) Analyze and optimize the performance measures of a given specific mission plan (see Figure 10).

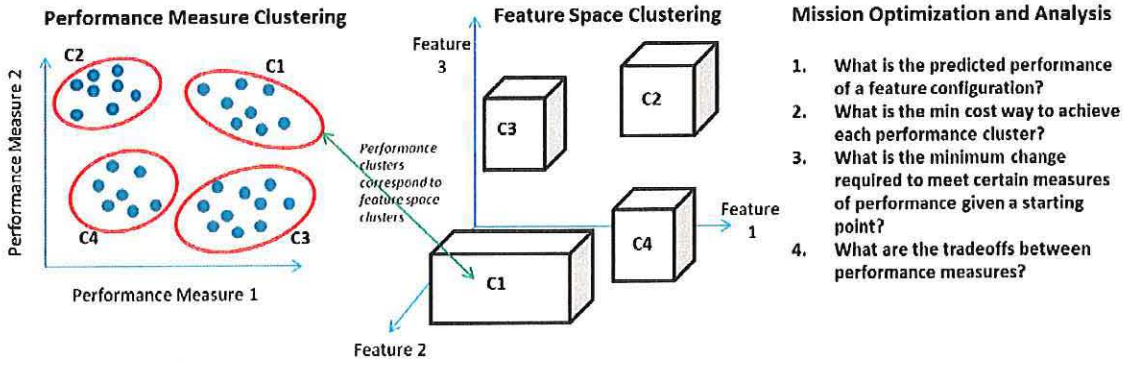

Figure 9- Initial SCOPE Methodology

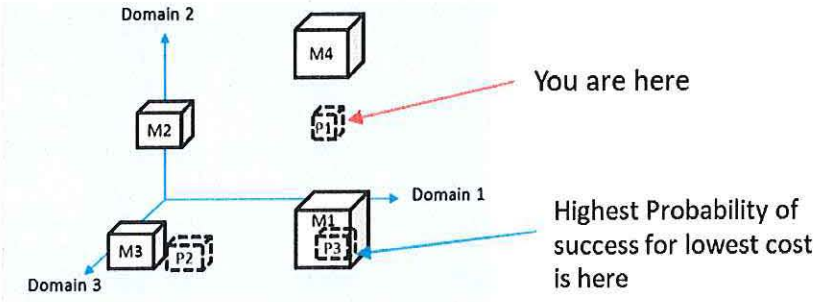

Figure 10- Can distance measures be used to recommend plan changes?

Due to a lack of a significantly large dataset to learn from, it was impossible to create performance clusters using all performance measures at once. For this reason, the initial methodology was revised to look at all combinations of 1, 2 or 3 performance measures in order to cluster, predict, and optimize individual subsets of performance measures. The following optimization model was proposed to minimize the cost of moving from a provided plan to a plan providing acceptable performance levels. The model allows for features spaces to be modelled using a probabilistic approach, such that cluster outliers could be ignored. The two constraints of the model require the model to choose just one cluster to move towards:

Sets:
$CLUSTERS = set\ of\ clusters\ 1\ to\ j\ (only\ acceptable\ clusters,\ not\ poor\ clusters)$
INPUTS= $set\ of\ dimensions\ for\ each\ cluster\ 1\ to\ i$

Parameters:
$avg_{ij} = average\ of\ input\ i\ for\ cluster\ j$
$stdDev_{ij} = standard\ deviation\ of\ input\ i\ for\ cluster\ j$
$startingInput_i = starting\ value\ of\ input\ i$
$cost_i = cost\ to\ change\ input\ i\ by\ 1\ unit$
$M_i = appropriately\ large\ M\ for\ input\ i$

$leeway_i = how\ much\ variablity\ we\ accept\ in\ an\ input\ (\#\ standard\ deviations)$

Variables:

$changetoInput_{ij} = change\ needed\ to\ input\ i\ to\ get\ into\ cluster\ j$

$$clusterChosen_j = \begin{cases} i\ if\ we\ choose\ cluster\ j,\ where\ j \in clusters \\ 0\ otherwise \end{cases}$$

$$Minimize\ Cost = \sum_{j \in CLUSTERS} \sum_{i \in INPUTS} cost_i * changetoInput_{ij}$$

Subject To:

$$\sum_{j \in CLUSTERS} clusterChosen_j = 1$$

$$startingInput_i + changetoInput_{ij} - avg_{ij} - leeway_i * stdDev_{ij} \leq M_i(1 - clusterChosen_j)$$
$$\forall\ i \in INPUTS,\ j \in CLUSTERS$$

$$-startingInput_i - changetoInput_{ij} + avg_{ij} - leeway_i * stdDev_{ij} \leq M_i(1 - clusterChosen_j)$$
$$\forall\ i \in INPUTS,\ j \in CLUSTERS$$

3.1 Surrogate Datasets

Sports data was chosen as a surrogate data set to military missions. Sports data was chosen due to its parallels to military missions and relative abundance of data available. Specifically NFL data for the 2009-2017 seasons was curated, representing 2304 individual games. Similar to military missions that have a large variety of measures of performance, so do NFL games. Table 5 highlights some of the performance measures captured for each NFL game. The features listed in Table 6 provide a sample of the types of features that were used to predict the performance measures in Table 5. Of note is that many of the features could be observed over different time frames, such as at the end of the current season, at the end of the previous season, the average of the current season leading up to the game in question, or the average of the last N games. Also of note, is that each of the features in Table 6 could be collected for both teams involved in the game.

Table 5- Performance Measures

Performance Measure	Description
1stDown	Number of first downs obtained in a game
TotYd	Number of total yards obtained in a game
PassY	Number of passing yards obtained in a game
RushY	Number of rushing yards obtained in a game
ToP	Time of possession during game
TO	Number of turnovers in a game
Skd	Number of times quarterback was sacked in a game
Pen	Number of penalties against the team in a game
3DC	Percentage of 3rd Down Conversions
Win	Whether the team won the game

Table 6 - Features

Type	Field	Description	End of Season	At Time Performance	Avg Last N Games
Rankings	RushRk	Ranking based on total number of rushing yards per season	x		
	PassRk	Ranking based on total number of passing yards per season	x		
	YrdRk	Ranking based on total number of yards gained per season	x		
	DRushRk	Ranking based on total number of rushing yards allowed per season	x		
	DPassRk	Ranking based on total number of passing yards allowed per season	x		
	DYrdRk	Ranking based on total number of allowed yards per season	x		
	Ovr	Team overall rating out of 100	x		
	Off	Team overall offensive rating out of 100	x		
	Def	Team overall defensive rating out of 100	x		
	PwrRk	ESPN's end of season power ranking	x		
	QBRk	ESPN's end of season quarterback rating	x		
Performance Stats	FirstDown	Number of first downs	x	x	x
	TotYd	Total yards gained offensively	x	x	x
	PassYd	Total yards gained via passing	x	x	x
	RushYd	Total yards gained via rushing	x	x	x
	TO	Total turn overs	x	x	
	ToP	Time of possession per game	x	x	x
	Pen	Number of penalties	x	x	x
	3DC	Percentage of third down conversions	x	x	x
	Skd	Number of times quarterback was sacked	x	x	x
Weather	Temp	Average temperature at the time and location of the game		x	
	Wtype	Cloud coverage (overcast, clear, partly cloudy, etc…)		x	
	Wind	Average wind speeds in miles per hour		x	
	Prcp	Accumulated rain in inches		x	
	Snow	Accumulated snow in inches		x	
Moral	Streak+	Number of consecutive wins		x	
	Streak-	Number of consecutive losses		x	
	Rest	How many days since the last game		x	

A survey of 128,980 NFL game predictions listed by professional handicapper services in the National Sports Monitor found that on average they were correct at predicting a game's winner only 50.39% of the time, or right around pure chance (Fox and Mayer, 2007). This suggests that predicting NFL game outcomes is difficult, even for the so called experts.

3.2 Initial Methodology Results

A number of experiments were set up to test the proposed framework and hypothesis using the surrogate sports data. However, we were never able to validate Hypothesis 1, such that separable feature spaces were created. Figure 11 shows what we expected to find versus what we largely found within our sports data.

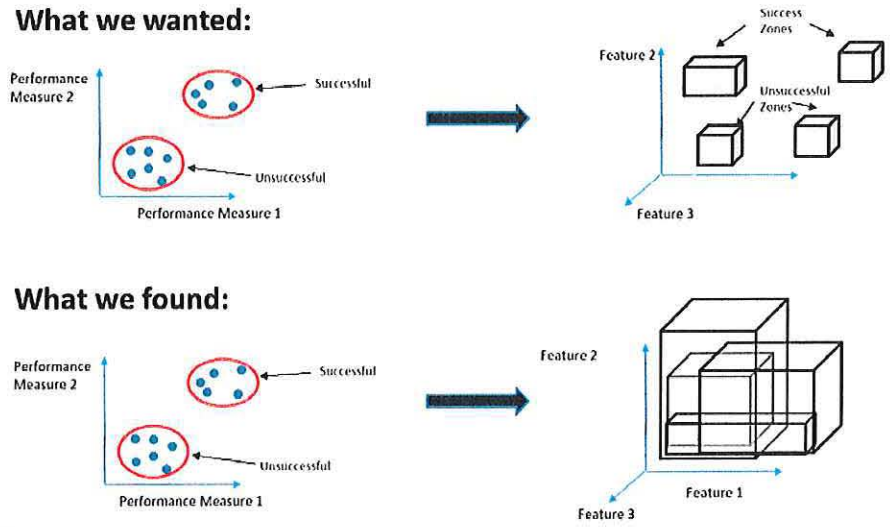

Figure 11-Unseparable feature spaces

Efforts to make reasonable performance predictions using the non-separable feature spaces proved largely futile as well. Figure 12 demonstrates typical prediction performance where the red line was our performance prediction for a certain performance measure, and the blue line was the actual performance. You can see that the predictions fail to capture the varied performance level of the actual data. This was typical across all experiments performed with performance prediction accuracies typically no better than 50% MAPE.

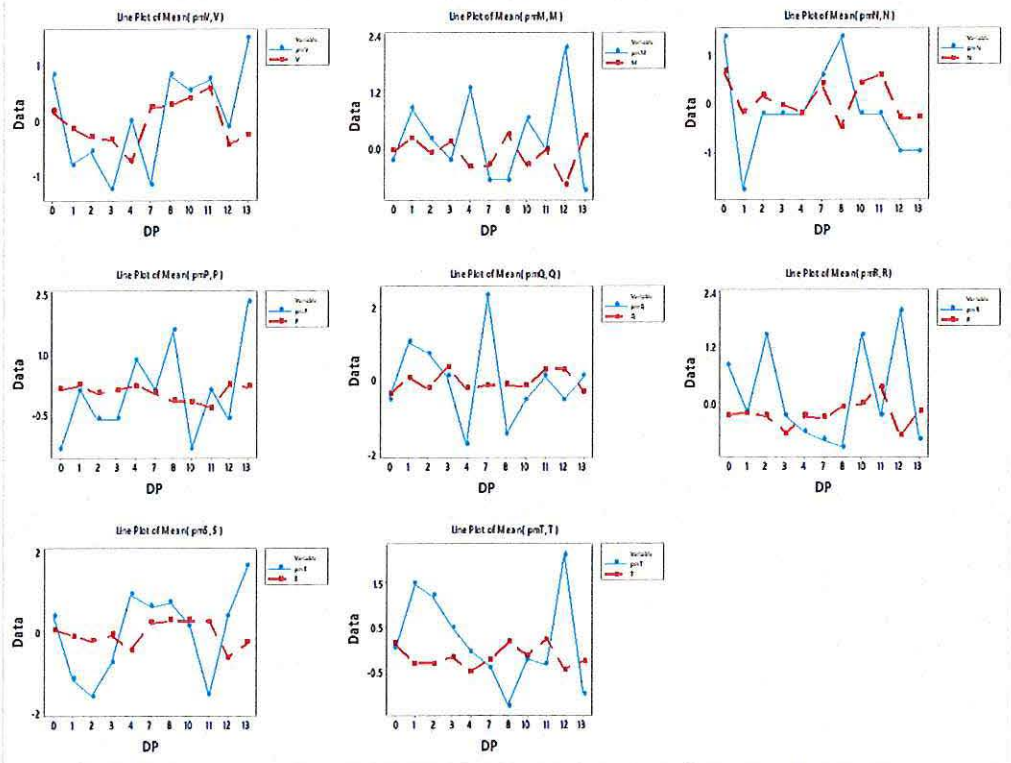

Figure 12- Performance predictions based on feature spaces (red line = predictions, blue line = actuals)

We confirmed our findings by using simple linear regression models to predict performance measures. Using linear regression models we looked at all NFL teams individually, and tried to predict certain performance measures given the complete set of features. Our hypothesis was that if linear regression models were similar across teams, we could expect the feature spaces developed to be linearly separable. However, when predicting a performance measure each team generated a regression model that used a different set of features. This suggested that a single model for all NFL teams may be difficult to create using our proposed methodology as we were assuming a single model for all teams could be created. Figure 13 visually displays the disparity between the developed models. Figure 13 shows for 4 teams which features were relevant towards predicting 8 performance measures of interest when using linear regression. The columns highlighted in red are indicating the models used to predict the number of first downs scored in a game for each of the four teams. One can see that the relevant features to this performance measure vary widely across the different teams.

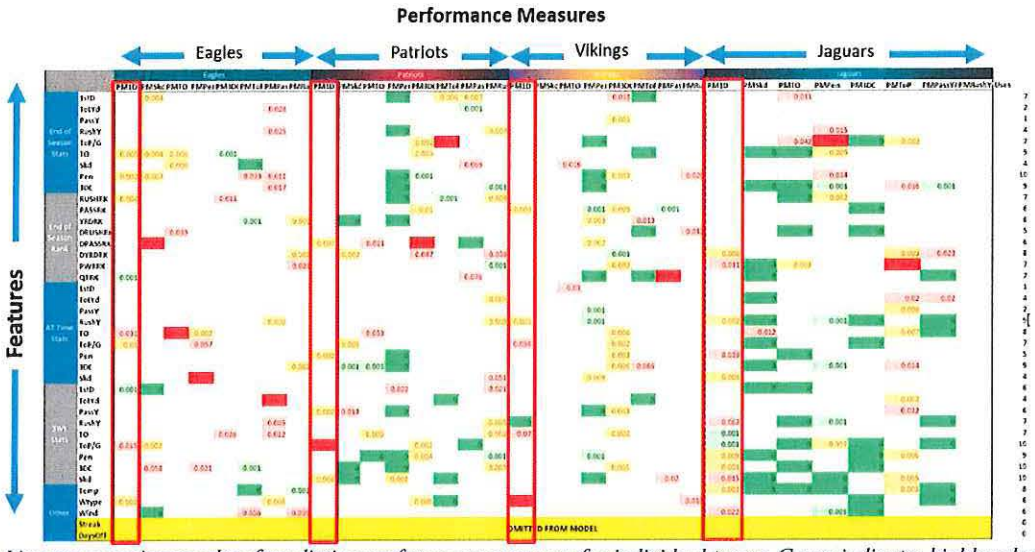

Figure 13- Linear regression results of predicting performance measures for individual teams. Green indicates highly relevant features, yellow mildly relevant features, and red are features with low relevance.

In summary we have the following results from our 3 initial hypothesis:

Hypothesis 1: Mission plans can be characterized using conceptual spaces, and that concepts (groups of feature values) corresponding to different performance levels will be separable from one another.

- Largely unsupported. Separable concepts could not be found for performance clusters for the NFL dataset.

Hypothesis 2: The separable concepts can be used to predict mission performance when given a plan.

- Largely unsupported. When the inseparable clusters were used to predict mission performance, predictions were not accurate.

Hypothesis 3: The separable concepts can be used to enable simple distance measure based approaches to mission optimization.

- Due to inseparability of the mission spaces, these algorithms went untested.

3.3 Lessons Learned

A few lessons could be learned from our initial experiments. Using these observations a new methodology was developed and tested. Firstly, the hypothesis that linearly separable clusters could be obtained by first clustering data by performance was largely unsupported. Despite extensive efforts, no linearly separable clusters were ever identified. Secondly, in the process of trying to improve our models, significant amounts of time was spent in the feature engineering process. Approximately 70-80% of our time was spent curating features that we thought may improve performance. This time could be reduced by automating the feature engineering process and efficiently searching the feature engineering space for useful feature transformations and

combinations. A process such as Kaizen Programming or Genetic Programming could be useful in reducing the time taken for feature engineering. The focus of the revised methodology shifted from predicting specific game performance measures, such as number of offensive yards, to predicting the success or failure of a game (win vs loss).

4 SCOPE Revised Methodology

To continue to make progress, once the initial methodology proved unsuccessful on the NFL data, we shifted to predicting game success (win vs loss) and using non-linear models to extract feature relationships. While many machine learning algorithms were examined, the most consistent performance was seen from decision trees, and those results are presented here.

4.1 Non-Linear Feature Space Model Performance

Starting with over 100 potential features, feature selection methods were used to identify the top 18 features that could be used to predict game success. The features and Gini-importance are outlined in the table below. Using 10-fold cross validation the average performance of these models for predicting game success was between 65 and 73%_accuracy. _____.

Table 7 - Decision Forest Model Features and Relative Importance

Feature	Description	Gini-Index
BRushDef	On average how many rushing yards did Team B allow the other team to get in one game? Averages week 1 up to that game.	0.11
AWins	Team A win percentage	0.1
BWins	Team B win percentage	0.1
ARushDef	On average how many rushing yards did Team A allow the other team to get in one game? Averages week 1 up to that game.	0.09
ARushOff	On average how many rushing yards did Team A get in one game? Averages week 1 up to that game.	0.08
APassDef	On average how many passing yards did Team A allow the other team to get in one game? Averages week 1 up to that game.	0.08
BRushOff	On average how many rushing yards did Team B get in one game? Averages week 1 up to that game.	0.07
B-QB	Team B quarter back rating out of 100	0.07
A-QB	Team A quarter back rating out of 100	0.06
APassOff	On average how many passing yards did Team A get in one game? Averages week 1 up to that game.	0.05
BPassOff	On average how many passing yards did Team B get in one game? Averages week 1 up to that game.	0.05
BPassDef	On average how many passing yards did Team B allow the other team to get in one game? Averages week 1 up to that game.	0.05
BTeam	Team B overall rating out of 100	0.04
BDef	Team B defensive rating out of 100	0.04
ATeam	Team A overall rating out of 100	0.03
AOff	Team A offensive rating out of 100	0.03
ADef	Team A defensive rating out of 100	0.03
BOff	Team B offensive rating out of 100	0.02

4.2 Feature Sensitivity

One question that may be of concern to mission planners is: "How certain does my data need to be, in order to be able to make a good prediction?" Using the decision forest models described in section 4.1 we set out to answer the following questions:

1. *Given a training dataset of 100% accuracy, which features are most sensitive to uncertainties when making predictions?*

2. *At what level of uncertainty does performance begin to diminish for each feature?*

To answer these questions the following procedure was followed. Given a dataset representing 249 NFL games (1 season with ties removed), the data set was randomly split into a training set and test set with 75% of observations belonging to the training set. A decision forest model was then trained on the test data. Fifteen sets of predictions were then made. One set of predictions was made with the 100% certain test data. For the remaining 14 sets of predictions, given an uncertainty level of α, one feature was first made more uncertain by augmenting the original observation by a percentage factor of a random number chosen between (100- α) to (100+ α). One hundred replications were completed for each feature for each level of uncertainty. Fifteen uncertainty levels were evaluated including:5,10,15,20,25,30,35,40,45,50,100,150,200,250,300.

For these initial uncertainty tests, when the value of a feature was changed, it had equal probability of being increased or decreased in value. When the feature values in the test set had equal probability of being increased or decreased by a certain uncertainty factor, prediction performance remained relatively unchanged. Across all features as uncertainty for a single feature was changed from 5 to 300%, the impact to performance was only +/-3%. Across all features there were only a few features that had marked detriments to performance when their values became more uncertain. These features were ATeam, ADef, BDef, and AOff at low levels of uncertainty and ATeam, ADef, BDef, and BRushOff at more extreme levels of uncertainty. Figure 14 and Figure 15 highlight these results.

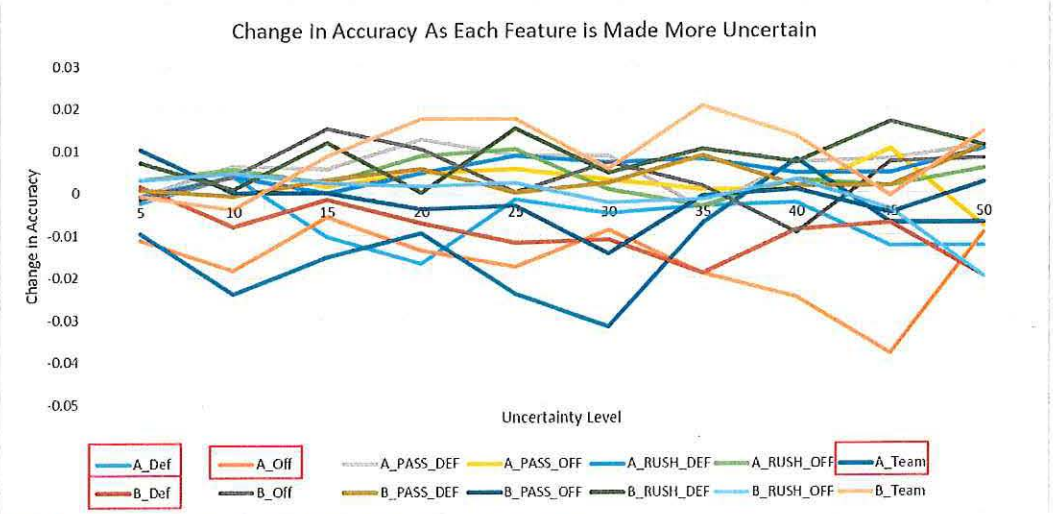

Figure 14-Low level uncertainty impact to prediction accuracy

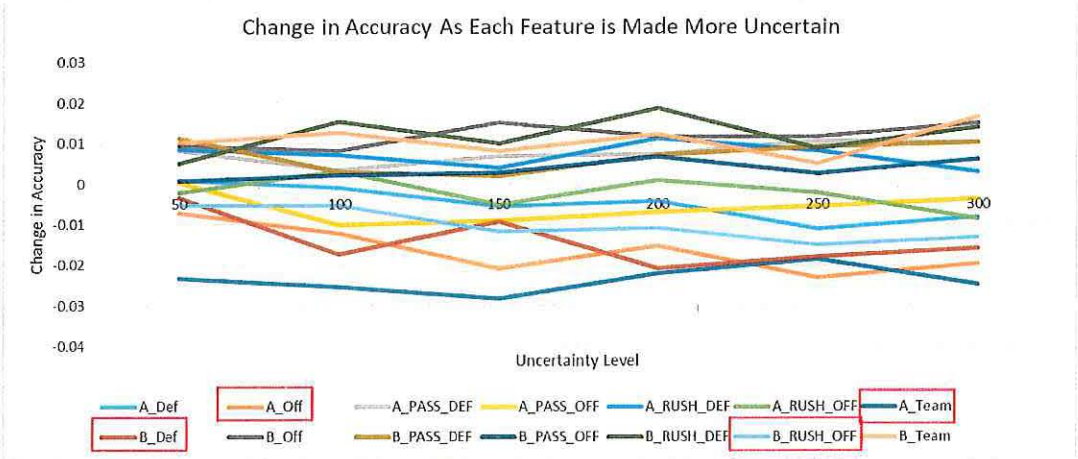

Figure 15-High level uncertainty impact to prediction accuracy

Initial results indicated if any one feature was uncertain impact to the overall model prediction performance was negligible for most features, however a few features saw a decrease of up to 3% in prediction accuracy. Since initial experimentation provided an equal probability of over or underestimating a feature, it makes sense that on average the prediction results would not change significantly as one direction of change might increase prediction accuracy while the other direction of change might decrease accuracy. The experiments were then repeated first by overestimating features by a specific uncertainty level, and then by underestimating features by a certain uncertainty level. These experiments help identify specific features that may be particularly sensitive to movement in a particular direction.

Figure 16 highlights the effects of overestimating individual feature values on prediction accuracy at high levels of uncertainty. Two features stand out: BDef and BRushDef. While most features are relatively insensitive to overestimation, when BDef is overestimated, prediction accuracy on average decreases by 4%. The opposite is true for BRushDef. When BRushDef is overestimated, prediction accuracy actually improves at high levels of uncertainty. Results such as these suggest that for these experiments, if the probability of overestimating BDef is high, spending more time refining the estimate will serve to improve prediction performance. Analogously, if the probability of overestimating any other features is high, spending more time refining those estimates would not result in improved predictive performance.

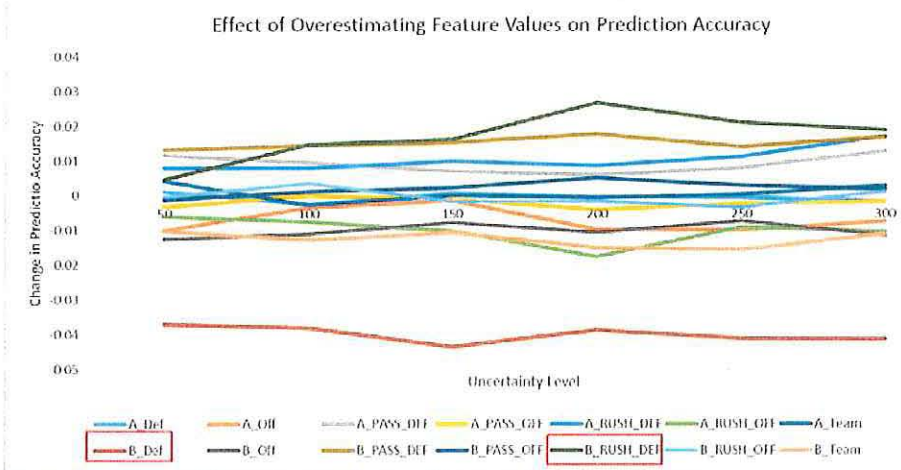

Figure 16- Effects of Overestimating Feature Values on Prediction Accuracy

Figure 17 illustrates the impacts of underestimating feature values on prediction performance. Of particular note is that underestimating the ATeam feature is detrimental to prediction accuracy, while underestimating BTeam and BOff may actually improve performance. Results such as these could be interpreted for real world application by considering being more conservative with certain feature estimates (BTeam and BOff), and potentially spending more time performing data collection on the ATeam attribute if it is likely to be underestimated.

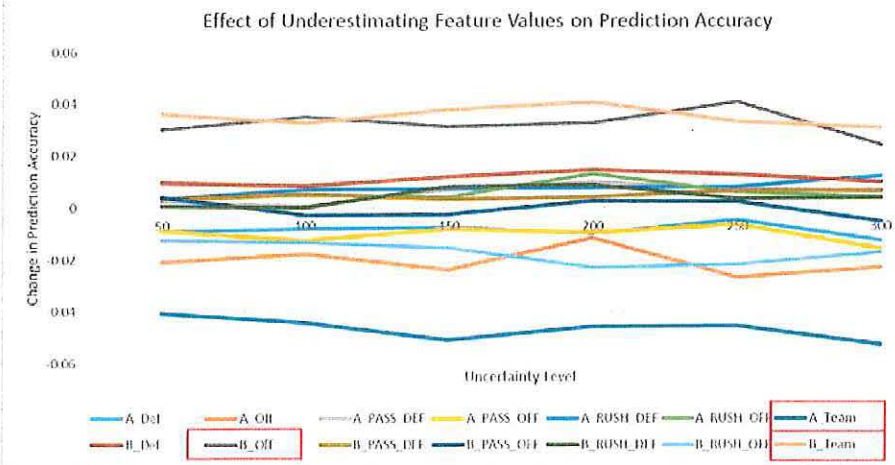

Figure 17- Effects of Underestimating Feature Values on Prediction Accuracy

These sensitivity results lead to additional questions. One investigation completed was to determine the impact of changing feature values on the predicted outcomes. Essentially we wanted to determine the range in which you could change a feature's value and be able to impact predictions. To do this, a single decision forest was trained with a training set containing 75% of available data. The remaining 25% was used as a test set. Predictions were made for the original data, and the accuracy was recorded. Then for intervals of 5% from -200 to +200 a single feature was changed by that percentage, and predictions were made. The intervals over which the prediction accuracy changed was then recorded. Results highlighted the differences between features with regards to the interval over which changes occurred, the directionality of change, and the magnitude of the change.

Figure 18 illustrates how the performance accuracy changes as each feature was adjusted by a certain percentage. Interestingly, decreasing some features results in improvements to accuracy while the opposite is true for other features. Also of interest in Figure 18 is the range of changes over which features are sensitive. These ranges are further highlighted in Figure 19 where BRushDef has one of the smallest ranges of impact and ARushOff has one of the largest ranges of impact. Features with larger ranges of impact indicate that the feature can take on a larger range of values without changing predicted outcomes than features with smaller ranges. In addition to ranges of impacts Figure 19 also lists the impact to accuracy over the range of the change. Features with large accuracy impacts indicate that more observations switch predictions over the impact range than those with smaller accuracy impacts. So for example, ARushOff has one of the largest impact ranges, meaning the feature can take on a large range of values from its starting value and have impacts to predictions. Simultaneously, ARushOff has the largest accuracy impact indicating that over the course of the impact range more predictions are changed than for other features. Conversely ARushDef has one of the smallest impact ranges and one of the smallest accuracy impacts. This would suggest that predictions change when ARushDef is changed only by a small amount, but that not many predictions actually change.

The warfighter could use similar analysis on mission data to understand the range over which features could be changed before impacting predictions, and how likely changes will actually result in a changed prediction.

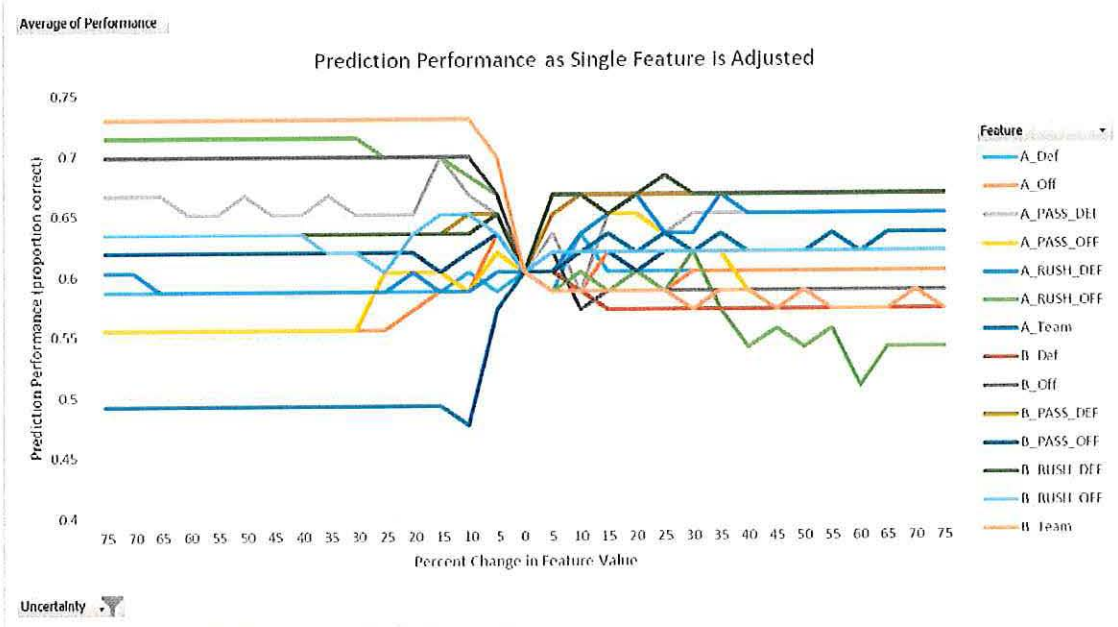

Figure 18-Prediction Performance as Single Feature Values are Adjusted

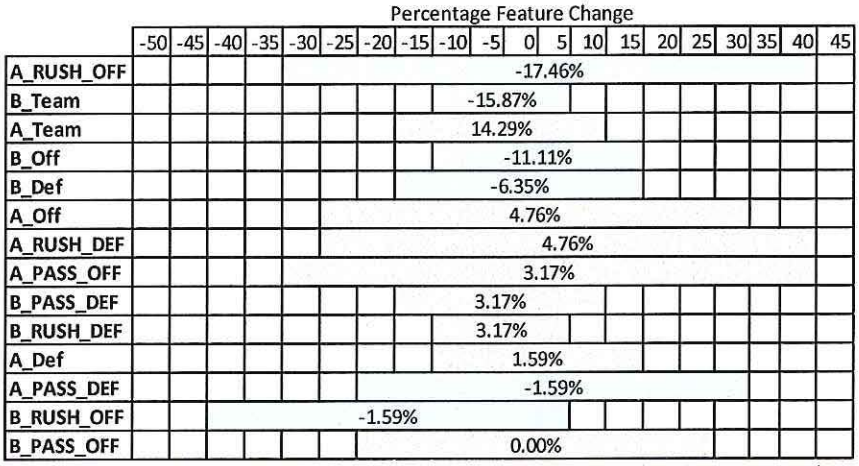

Figure 19-Feature Sensitivity to Change (Shaded regions correspond to changes in prediction, percentages listed represent the change in prediction accuracy, larger values indicate features that have more impact on predictions when changed)

5 Conclusions and Future Work

This work explored the use of a large scale detailed HADR simulation model for the generation of machine learning datasets. Small experiments conducted using the HADR simulation model were able to validate the functionality of information gain based routing algorithms, as well as identify non-intuitive relationships between mission characteristics and performance measures. However, the development of large scale datasets was hampered by the lengthy validation cycle and the long simulation run times. Future efforts utilizing simulation should have a validation functionality running in real time with the simulation. Efforts to reduce simulation runtime should be pursued, or efforts such as the use of Generative Adversarial Networks could be explored as a way to artificially produce large datasets at a faster rate than simulation alone.

This work further evaluated algorithms to efficiently predict mission performance measures using a conceptual spaces based approach. The study found that despite being able to form performance clusters, that feature space clusters were still largely non-linearly separable, preventing solid predictions of performance measures. This study found that by using non-linear models, such as decision trees the same data could be used to make reasonable performance predictions. Future work should look to augment the conceptual spaces based modeling approach with non-linear feature space development.

Finally this work began to investigate the impacts of feature uncertainty to prediction performance. Experiments were able to show that prediction performance was affected differently depending on which feature was uncertain, and that some features had larger impacts to performance than others. Additionally, the range for each feature's values was identified that had an impact on performance, and the magnitude of this performance impact was also captured. This type of analysis could be used by the warfighter in the future in order to inform intelligence collection plans or prioritize efforts to obtain better estimates of features. Future work should endeavor to use these feature sensitivities when suggesting revised mission plans.

6 References

AnyLogic. (2017) https://www.anylogic.com Accessed August 2017.

Fox, J., and Mayer, K. (2007). Assessing sports advisory services: Do they provide value for football bettors? UNLV Gaming Research & Review Journal, 11(2).

Howden, M. (2009) How humanitarian logistics information systems can improve humanitarian supply chains: a view from the field. In Proceedings of the 6th International ISCRAM Conference, Gothenburg, Sweden.

Ortiz-Pena, H., Hirsch, M., Karwan, M., and Sudit, M. (2015) The effect of decentralization and communication networks on a set of ISR-gathering assets. Proceedings of SPIE:Defense, Security, and Sensing, Baltimore, MD, April.