Building a Robust Explanatory Model from a Large Data Set

Dave Kim
Department of Mathematical Sciences
U.S. Military Academy, West Point, New York

Barry Bodt

Computational and Information Sciences Directorate

US Army Research Laboratory

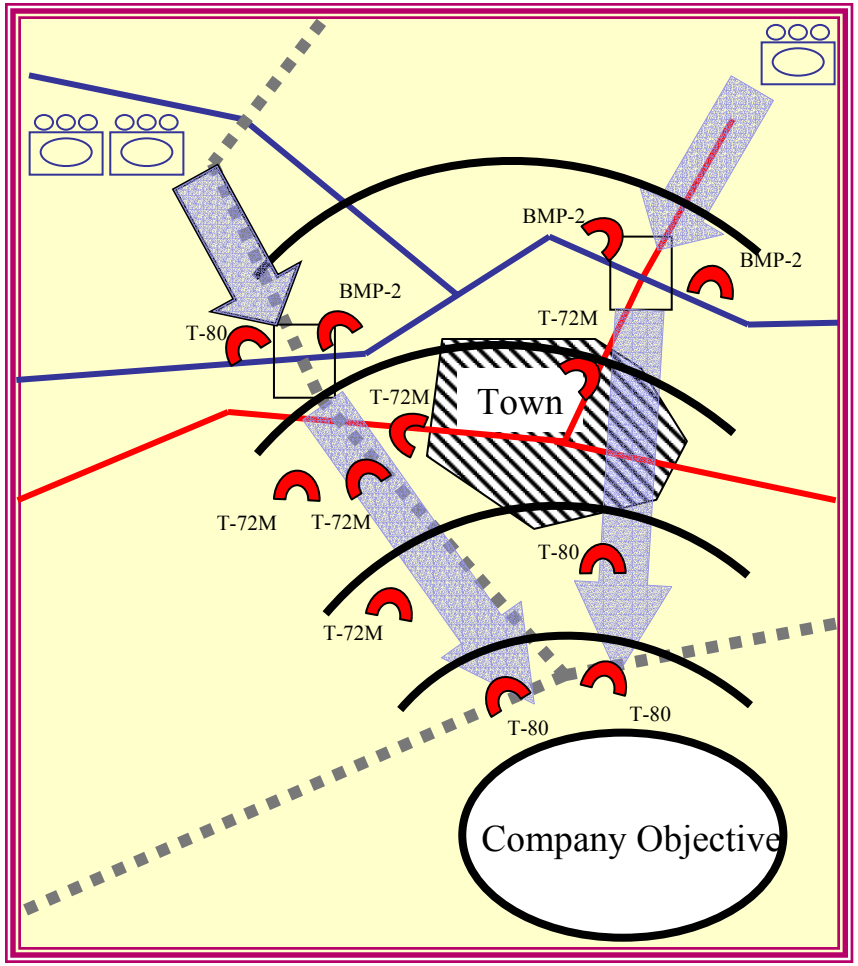
Aberdeen Proving Ground, MD

Outline

- An example of application.
- A method of robust variable selection using L_2E applied to the example.
- “BLUF”

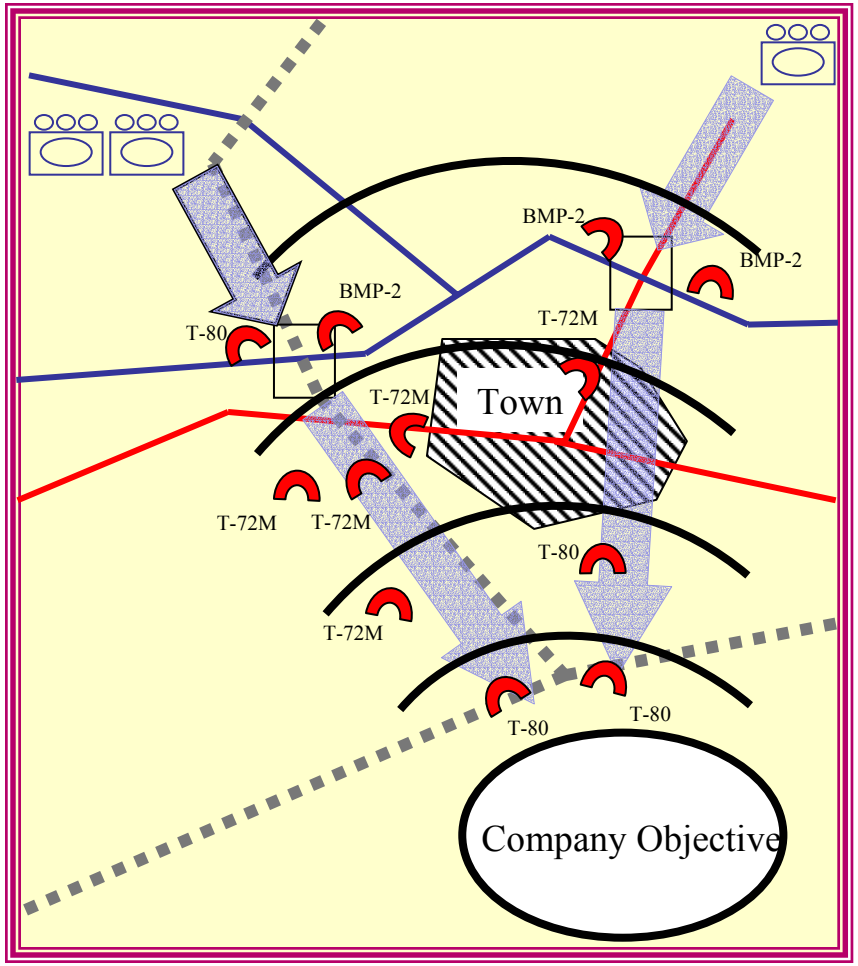
An Analysis of OneSAF Runs

- **228 OneSAF ("Semi Automated Force") runs**
- **3 situational time slices per run**
- **429 data points per run (143 per time slice)**
  - **Number of K, M/F, F, and M kills**
  - **Ammunition levels**
  - **Number of hits delivered**
  - **Range of hits**
  - **Number of side hits delivered**
  - **Distance to objective**
  - **Number of Blue on objective**
- **Response Variable: MBT Score**
- **228x430 data matrix.**



An Analysis of OneSAF Runs

- **Battle results are infinite in scope, yet the outcome of any one battle is defined by a finite set of battlefield interactions.**
- **The goal is to identify battle plan evaluation metrics based on statistical analyses of parameterized combat influences.**
- **Not about "predicting" the battle outcome; rather about explaining potential scenario developments.**



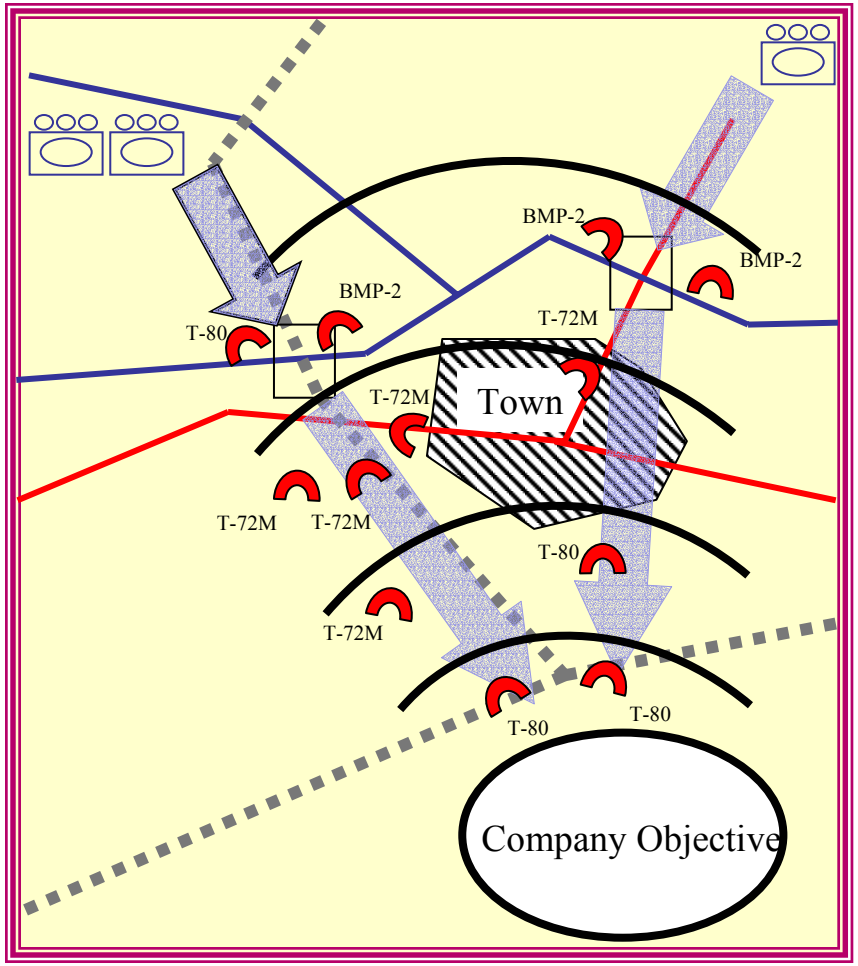
“L2E” Variable Selection
Applied to143 Variables at the Time Slice 1

- MF3S1: # tanks in P3 that are MF’d.
- F3S1:# tanks in P3 that are F’d.
- NKB1S1: # K kills on BMP by P1.
- R05HB2S1: Avg range 105 HEAT on BMP by P2
- MF1S1: # tanks in P1 that are MF’d.
- N05HB3S1: # hits 105 HEAT on BMP by P3.
- N05ST2S1: # hits 105 SABOT on T80 by P2.

Var	L_2E Coeff
Int	584.01
MF3S1	-127.12
F3S1	-108.10
NKB1S1	111.64
R05HB2S1	0.047
MF1S1	-152.55
N05HB3S1	26.36
N05ST2S1	-51.03

An Analysis of OneSAF Runs

- **What does it mean for N05ST2S1 (# hits 105 SABOT on T80 by P2) to have a negative coefficient?**
- **P2 follows behind P1 in the layout.**
- **P2 hitting T80 at the time slice 1 means P1 has been hit hard.**



Variable Selection

- Is an exploratory technique.
- Recall John Tukey's distinction between exploratory data analysis (EDA) and confirmatory data analysis.
- Tests and confidence/prediction intervals may not be valid if the same set of data is used to select variables in the model.

L_2E Regression

- Robust regression method based on L_2E criterion of Scott (2001).
- For a linear model

$$y = b_0 + b_1 x_1 + \cdots + b_k x_k + \varepsilon$$

where the error term is a normal random variable with mean 0 and variance σ^2, estimate the parameters b_i's and σ^2 by minimizing

$$L_2E = \frac{1}{2\sqrt{\pi}\sigma} - \frac{2}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\varepsilon_i^{\,2}}{2\sigma^2}\right)$$

Why L_2E?

- L_2E regression has been shown to have robustness properties similar to least median of squares (LMS) regression and other robust estimators.
- The functionals to be minimized in other robust methods need sorting of *n* residuals, whereas L_2E does not.

Comparison of multiple regression parameter estimates for L_2E vs LS

- No contamination (n=100)

	b_0	b_1	b_2	b_3	b_4	σ
True Values	0	-1	2	-1	2	0.2
Least Squares Estimates	0.0006	-1.0050	1.9982	-1.0015	2.0020	0.2055
L_2E Estimates	0.0104	-1.0068	2.0013	-1.0100	2.0022	0.2182

- 20% contamination (n=100)

	b_0	b_1	b_2	b_3	b_4	σ
True Values	0	-1	2	-1	2	0.2
Least Squares Estimates	0.0139	0.1592	1.6500	0.1679	2.0401	8.65
L_2E Estimates	-0.0357	-1.0286	2.0137	-1.0005	2.0110	0.2118

Toward Robust Variable Selection

- Most of classical variable selection methods available today are based on LS regression.
- Can we come up with a variable selection method that is not easily influenced by outliers and other data contamination?
- One of the criticisms of many variable selection schemes: the selected models often depend on the influential observations.

Example

- Simulated data set of 100 points using the model $y=3-x_2+4x_3-7x_4$ (with $\sigma=.2$).
- We want to look at how σ_{est} and L_2E_{est} vary across different subsets of explanatory variables.

Exp var	L_2E_{est}	σ_{est}
x_1	-.0159	18.19
x_2	-.0168	16.86
x_3	-.0164	18.52
x_4	-.0297	9.982

Example (cont'd)

- Simulated data set of 100 points using the model $y=3-x_2+4x_3-7x_4$ (with $\sigma=.2$).
- We want to look at how σ_{est} and L_2E_{est} vary across different subsets of explanatory variables.

Exp var	L_2E_{est}	σ_{est}
x_1 ,x_4	-.0299	9.86
x_2 ,x_4	-.0304	9.87
x_3 ,x_4	-.1114	2.69

Exp var	L_2E_{est}	σ_{est}
x_1 , x_3 ,x_4	-.0305	9.80
x_2 , x_3 ,x_4	-1.419	0.19

Heuristic Variable Selection (suggested by the prev. example)

- Choose the variable that gives the smallest L_2E and/or σ at each forward selection step, and stop when adding any of the remaining variable gives no improvement.
- The correct models were arrived at in many different models and at reasonable contamination levels.
- Interesting result: When the contamination level goes up, adding a redundant variable usually results in increase of L_2E and σ!
- The following table shows that what happens to the previous example when we have 20% contamination.

Exp var	L_2E	σ
x_2 , x_3 ,x_4	-1.000	0.19
x_1 , x_2, x_3 ,x_4	-.0557	6.73

Ongoing and Future Work

- Developing the criteria analogous to AIC and others.
- Robust F ratio?
- Continuing applications to combat simulation data.