



INSTITUTE FOR DEFENSE ANALYSES

Reasoning about Uncertainty in the Stan Modeling Language

Heather Wojton, Project Leader

John Haman
Thomas Johnson
Heather Wojton

April 2019

Approved for public release. Distribution
is unlimited.

IDA Document NS D-10578-NS

Log: H 2019-000153



The Institute for Defense Analyses is a non-profit corporation that operates three federally funded research and development centers to provide objective analyses of national security issues, particularly those requiring scientific and technical expertise, and conduct related research on other national challenges.

About This Publication

This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-14-D-0001, Task BD-9-229990, "Test Science Applications," for the Office of the Director, Operational Test and Evaluation. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments

Review was conducted by Mr. Robert R. Soule, Director, and Heather M. Wojton from the Operational Evaluation Division.

For more information:

Heather M. Wojton, Project Leader
hwojton@ida.org • 703-854-6811

Robert R. Soule, Director, Operational Evaluation Division
rsoule@ida.org • (703) 845-2482

Copyright Notice

© 2019 Institute for Defense Analyses
4850 Mark Center Drive, Alexandria, Virginia 22311-1882 • (703) 845-2000.

This material may be reproduced by or for the U.S. Government pursuant to the copyright license under the clause at DFARS 252.227-7013 (a)(16) [Jun 2013].

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-10578

Reasoning about Uncertainty in the Stan Modeling Language

Heather Wojton, Project Leader

John Haman
Thomas Johnson
Heather Wojton

Executive Summary

This briefing discusses the practical advantages of using the probabilistic programming language (PPL) called Stan to answer statistical questions, especially those related to the quantification of uncertainty. Stan is a relatively new statistical tool that allows users to specify probability models and reason about the processes that generate the data they encounter.

Stan has quickly become a popular language for writing statistical models because it allows one to specify rich (or sparse) Bayesian models using high-level language. Further, Stan is fast, memory efficient, and robust.

Stan requires users to be explicit about the model they wish to evaluate, which makes the process of statistical modeling more transparent to users and decision makers. This is valuable because it forces practitioners to consider assumptions at the beginning of the model-building procedure rather than at the end (or not at all). In this sense, Stan is the opposite of a “black box” modeling approach. Stan may be tedious and labor intensive at first, but the payoffs are large. For example, once a model is set up, inferential tasks are all essentially automatic, as changing the

model does not change how one analyzes the data. This is a generic approach to inference.

To illustrate these points, we use Stan to study a ballistic miss distance problem. In ballistic missile testing, the p -content circular error probable (CEP) is the circle that contains p percent of future shots fired, on average. Statistically, CEP is a bivariate prediction region, constrained by the model to be circular.

In Frequentist statistics, the determination of CEP is highly dependent on the model fit, and a different calculation of CEP must be produced for each plausible model. However, with Stan, we can solve the CEP problem independent of the model.

Using simple summary statistics, we show how to use Stan to calculate CEP and uncertainty intervals for the parameters.

Statistical practitioners can access Stan from several programming languages, including R and Python.

Reasoning about Uncertainty in the Stan Modeling Language

John Haman
Thomas Johnson

10 April 2019

IDA What is Stan?

- Stan is a Bayesian model specification (and compilation) language
 - Similar to BUGS and other model languages, but more flexible and easier to use
- Started at Columbia University in 2012
- Open source
- Under active development



IDA Why do we use Stan at IDA?

- Model building in Stan is transparent and structured
 - Model assumptions are specified at the beginning of the model process
- Inferential tasks are generic
 - Uncertainty quantification is done by calculating summary statistics from posterior samples
- Stan is fast, memory efficient
- Available as an R package (`rstan`)

IDA What Stan buys us

- Access to Bayesian statistical models
 - High level modeling language
 - Easy to create statistical models
- Uncertainty intervals for all model parameters
- Flexibility to fit a large array of models

IDA Ballistic Data (notional)

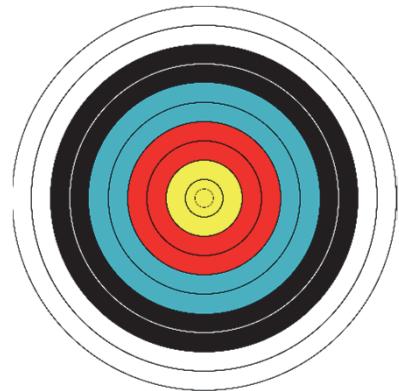


IDA The Notional Experiment



Old

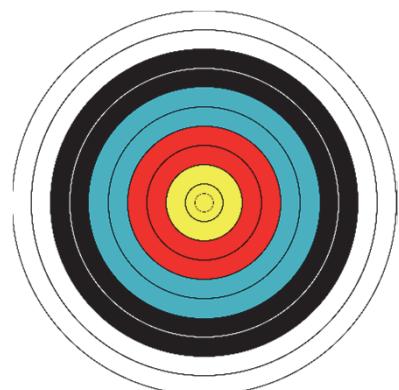
Far



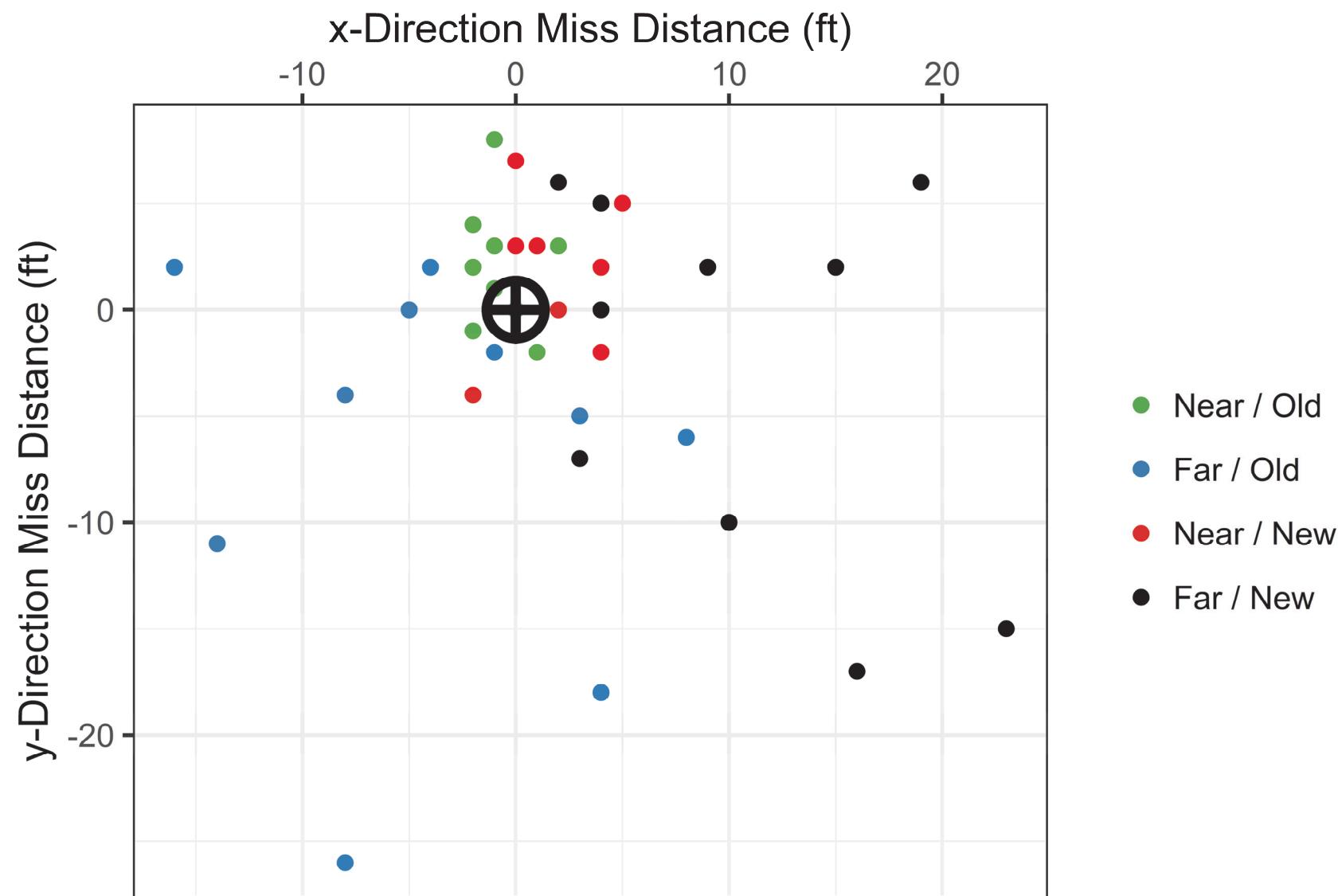
New



Near



IDA The Notional Experiment

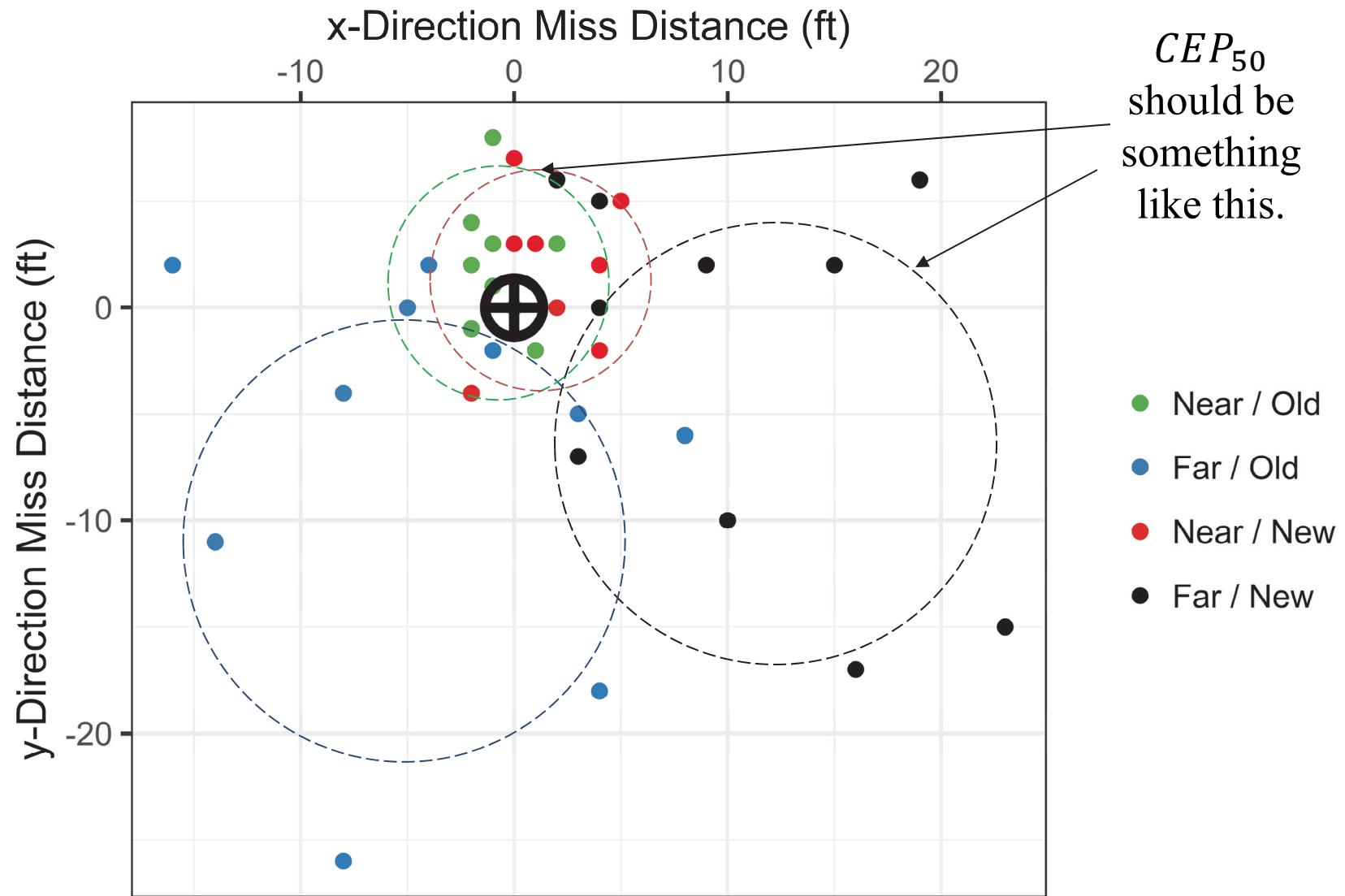


IDA Circular Error Probable (CEP)

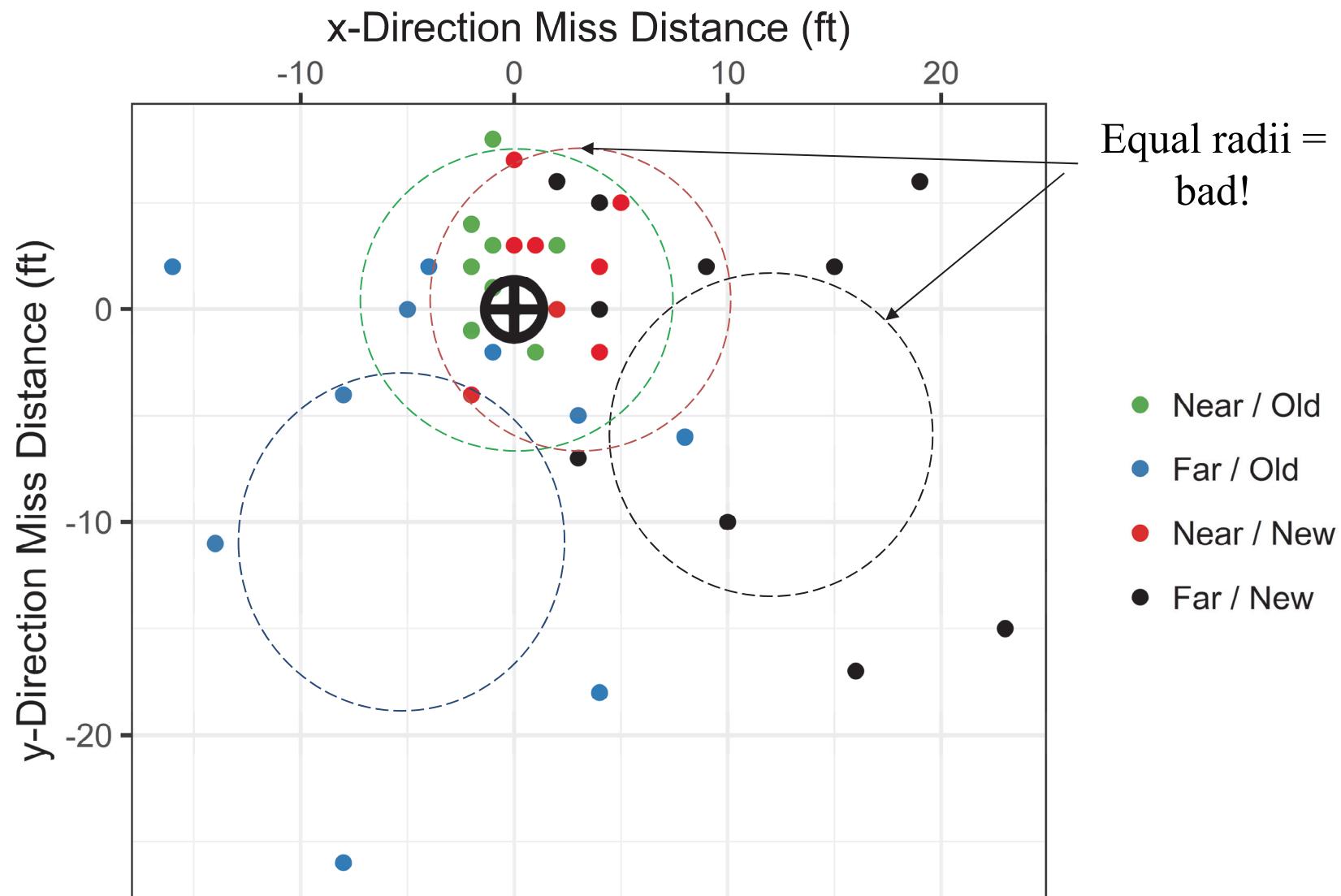
Want to define a circle in which we predict the next shot will land with probability 50%.

Challenge: Want CEP radius to change with experimental factors

IDA Circular Error Probable (CEP)



IDA Constant variance model is inadequate



IDA Proposed Model*

$$Y = \begin{pmatrix} x \\ y \end{pmatrix} \sim Normal \left(\begin{pmatrix} M & 0 \\ 0 & M \end{pmatrix} \begin{pmatrix} \rho \\ \zeta \end{pmatrix}, \Sigma \right)$$

Observed x and y direction miss distances

x direction parameters

Design Matrix

y direction parameters

Σ is a matrix that supplies different variances depending on factor settings in M .

This is just a complicated linear model,

$$Y = X\beta + \epsilon$$

IDA The Traditional Way

- Fit model with `nlme`, or `statmod` in R
- Code custom approximate F-statistic
- Doable, but the math is hard
- Not easy to extend

IDA The Stan way – more transparent

All Stan programs start with three “blocks”

Sampling statement

```
// CEPmodel.stan (abridged)
data {
    // import data from R...
    matrix[n, k] X;
}
parameters {
    vector[p] rho; // mean parameters
    vector[p] zeta;
    matrix[n, n] Sigma; // variance parameters
}
model {
    // Sample from our model
    y ~ normal(X * [rho, zeta]', Sigma);
}
```

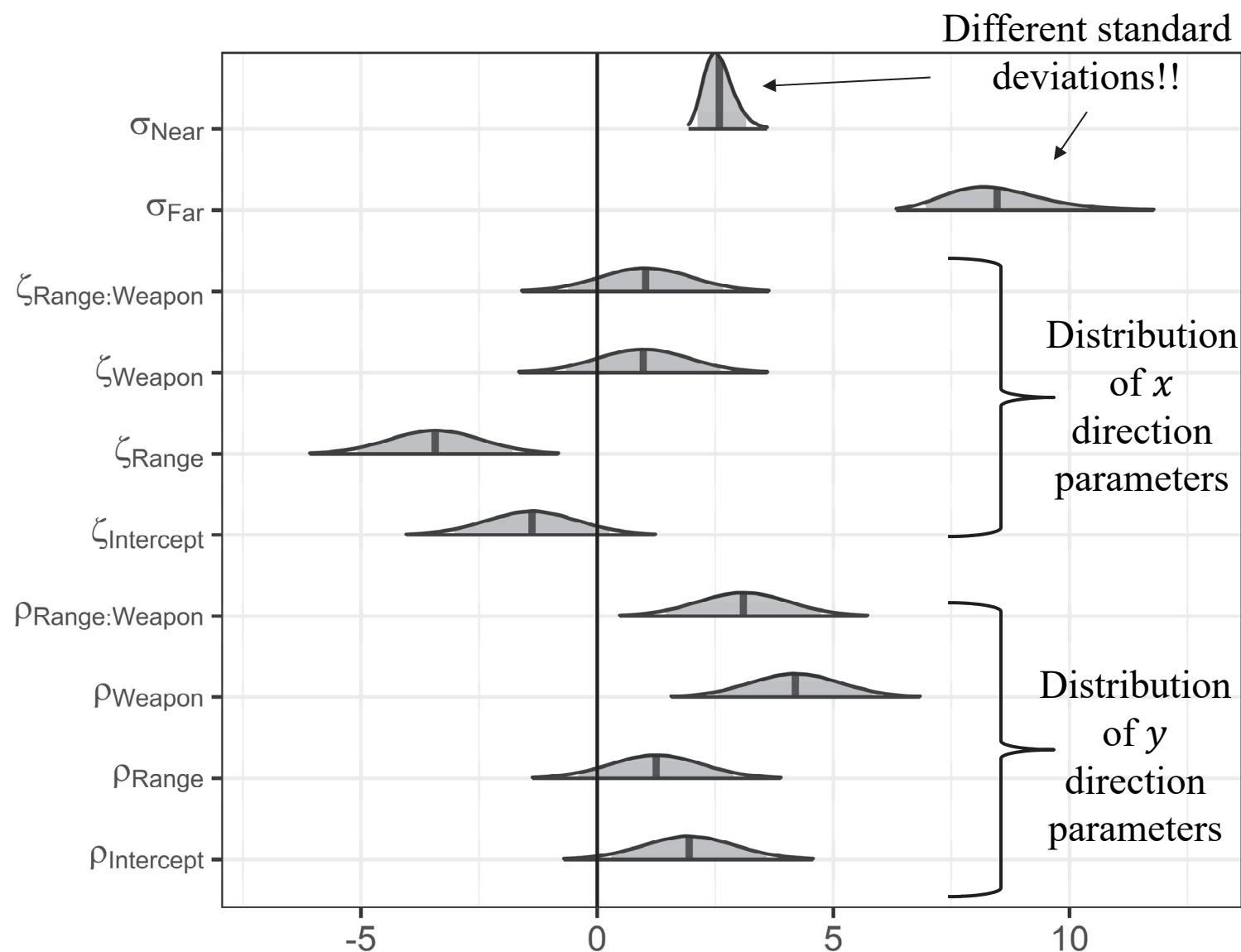
IDA Stan Output: Posterior Samples

Each column is a vector of samples

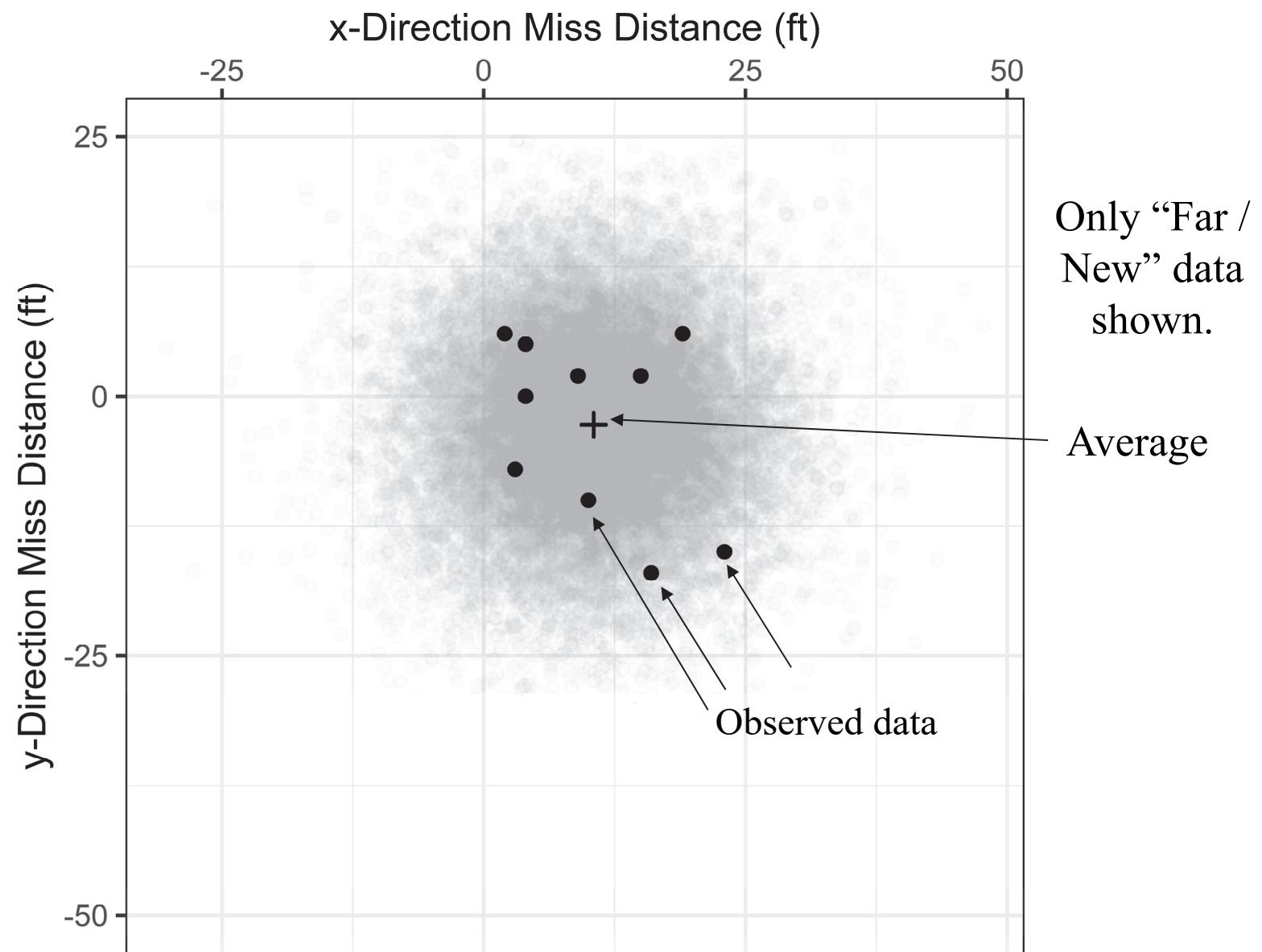
Sample # /Term	σ_{Far}	σ_{Near}	ρ_{Weapon}	ρ_{Range}	...
1	7.75	2.31	3.01	5.39	...
2	9.29	2.49	2.45	4.31	...
3	9.73	2.45	2.71	4.39	...
4	7.86	2.48	2.39	5.66	...
5	8.24	2.26	2.67	4.32	...
:	:	:	:	:	:

All inference done by computing column standard deviations or quantiles.

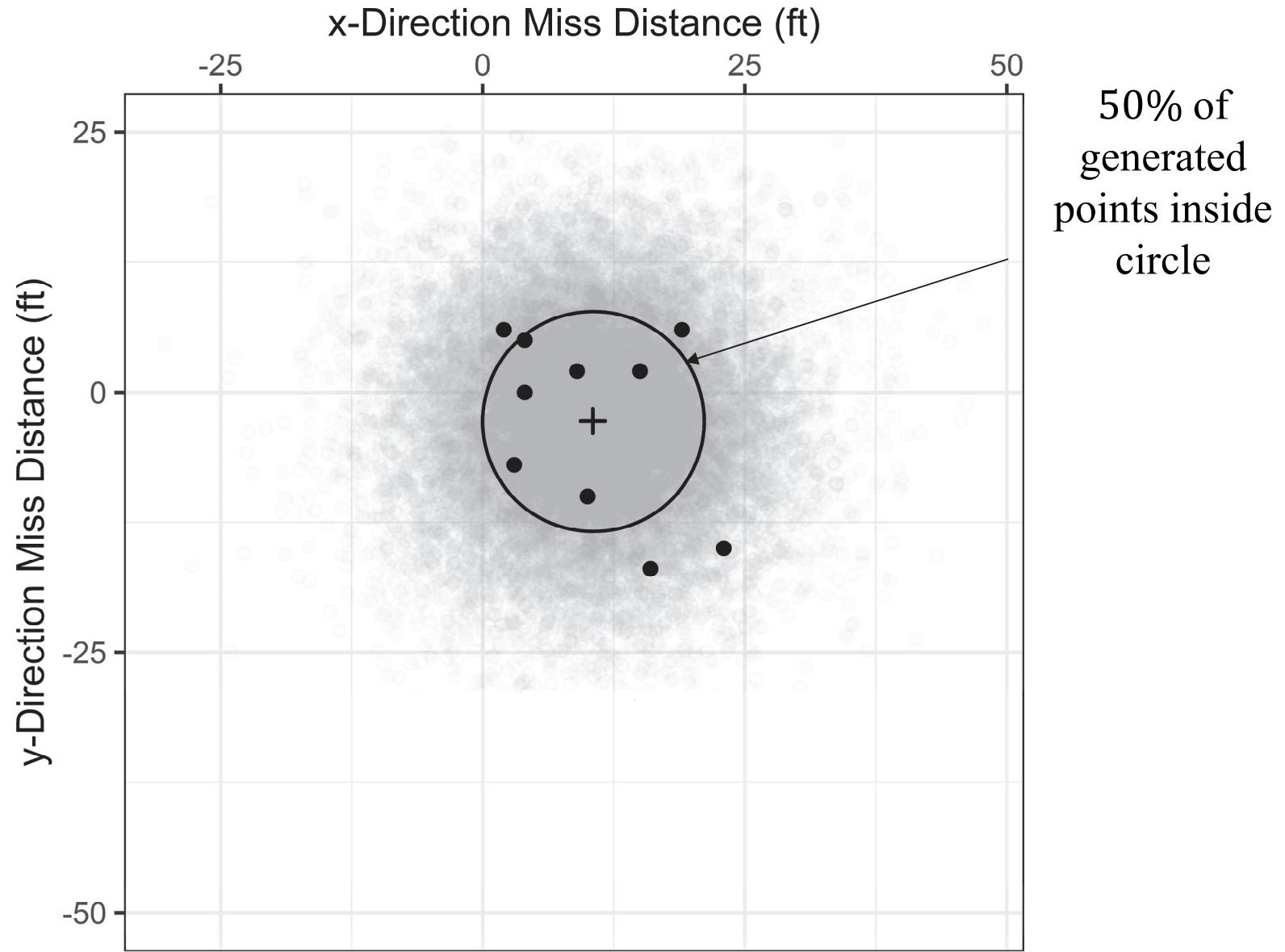
IDA Posterior Distributions of All Model Parameters



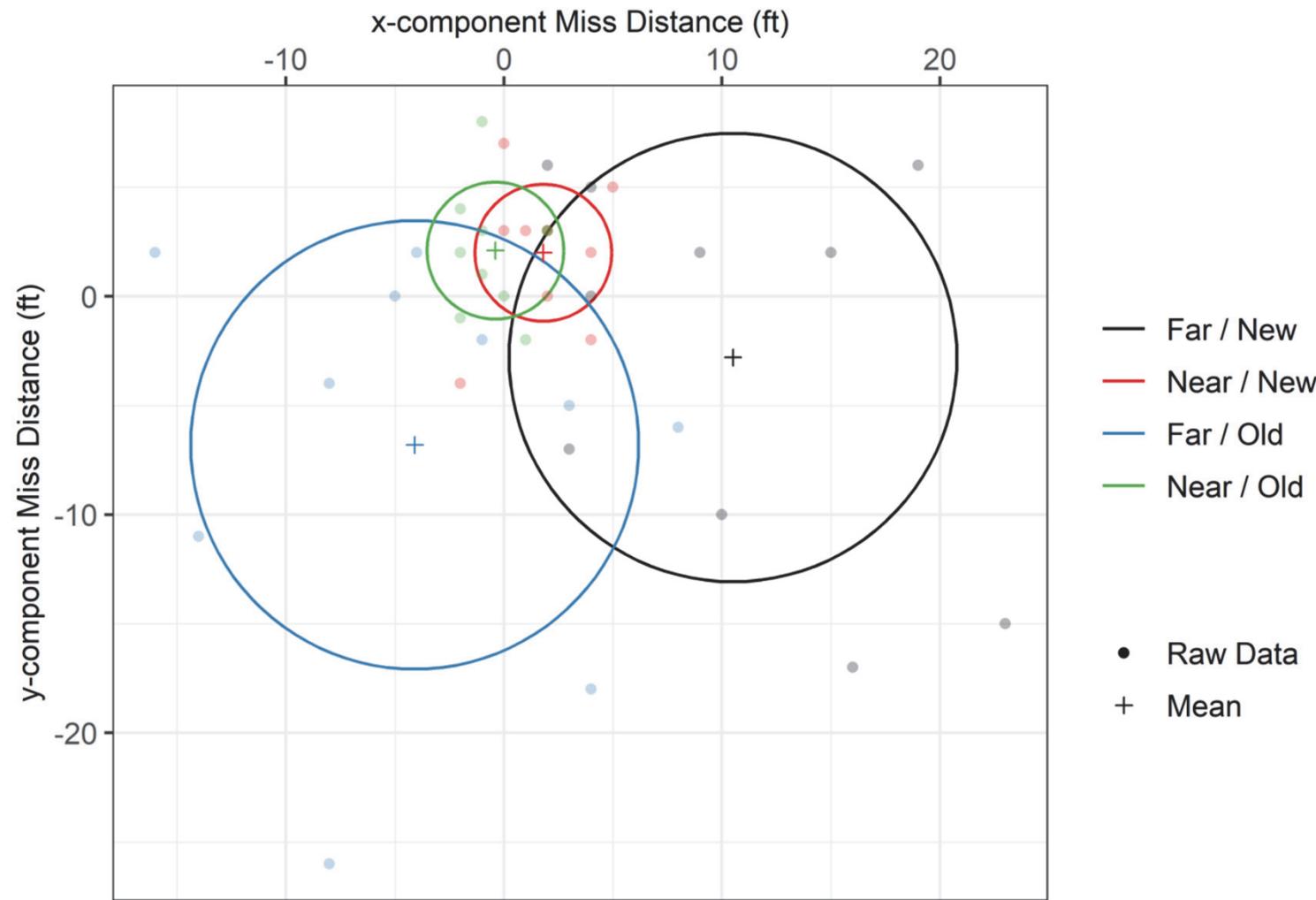
IDA Generate New Data from the Stan Model



IDA Size the Circle (CEP)



IDA Calculate the Desired CEP for each Factor



IDA Trade-offs

- Pros:
 - Reliable uncertainty quantification
 - Uncertainty intervals for all parameters, no bootstrapping required
 - Extensible
 - Model checking available
- Cons:
 - More work upfront
 - Have to learn something new

IDA Resources

mc-stan.org

- Tutorials
- Videos
- ***Reference Manual***
- Case Studies

Stan Modeling Language

User's Guide and Reference Manual

Stan Development Team

Stan Version 2.17.1

Monday 11th December, 2017



mc-stan.org

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE (DD-MM-YYYY)			2. REPORT TYPE		3. DATES COVERED (From - To)	
4. TITLE AND SUBTITLE			5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER			
6. AUTHOR(S)			5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)					8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)					10. SPONSOR/MONITOR'S ACRONYM(S) 11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT						
13. SUPPLEMENTARY NOTES						
14. ABSTRACT						
15. SUBJECT TERMS						
16. SECURITY CLASSIFICATION OF: a. REPORT b. ABSTRACT c. THIS PAGE			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON	
					19b. TELEPHONE NUMBER (Include area code)	

