Categorical Variables in NNSA Cost Estimating Relationships

Zachary Matheson, Greg Stamp, & Charles Loelius

National Nuclear Security Administration

DICEROLLER – A case study

- **D**&D **I**ntegrated **CER** for **O**ne-for-one and **L**ifecyc**L**e **E**stimate **R**anges (DICEROLLER)

- NNSA Office of Planning, Analysis and Evaluation (PA&E) requires the ability to estimate costs associated with Deactivation and Demolition (D&D) of NNSA facilities.

- This capability will support:
  - Lifecycle cost estimates for capital acquisition projects, and
  - “One-for-one” replacement cost estimates, meaning that new construction at DOE sites “is offset by the sale, declaration of excess, or demolition of building area of an equivalent or greater size.”

Model objectives

Background: PA&E leads Analyses of Alternatives and develops early-stage planning estimates like those in the SSMP.

Model requirements:

- High-level for early-stage estimates
- Easy to use
  - Small number of parameters, which should be easy to identify at early stages
- Covers a wide range of project scope, size, costs, etc.
- Based on historic data
- Capable of producing AACE Class 5 quality estimates

Preparing the data

- Total of 41 data points used to construct cost estimating relationship (CER)
    - NNSA Office of Safety, Infrastructure & Operations (NA-50)
    - DOE Office of Environmental Management (DOE EM)
    - G2 Planning Database
    - Sandia National Laboratory
  - Range of facility size:
    - 240 ft^2 – 319,742 ft^2
  - Range of total project costs:
    - $3,764 - $343,000,000
  - Range of hazard categories:
    - Nuclear Category 2, 3; Radiological; Chemical; Biological; No Hazard
  - Range of contamination types:
    - Radiological, Lead/asbestos, No contamination
  - Range of building types:
    - Permanent technical; Permanent non-technical; Temporary
- Data adjusted to account for escalation, location

Creating the CER

- Data cross-referenced with DOE's facility management database to identify cost drivers:
  - Facility gross square footage *(GSF)*
  - Contamination type *(Contam)*
    - Radiological, Lead/asbestos, No contamination
  - Building construction type *(Type)*
    - Permanent technical; Permanent non-technical; Temporary

- Tested several model forms to generate a cost estimating relationship to predict future D&D project costs.

- The dataset and cost estimating relationship were made into a user-friendly tool for use by PA&E

What is a categorical variable?

- A categorical variable is used in a model to describe characteristics that can't be directly quantified.
  - The DICEROLLER CER uses two categorical variables: contamination and building type.

- We'll cover two ways of incorporating categorical variables: Label Encoding and One-Hot Encoding.

Label encoding

- Uses integers to represent lists of categories.
  - For example, contamination type*:*
    - 1 – Radiological
    - 2 – Lead/asbestos
    - 3 – No contamination
- By its nature, label encoding imposes a hierarchy or an ordering upon your data.
- Using label encoding, the model form of the CER is:
  - $\log(TPC) = \alpha + \beta \cdot \log(GSF) + \gamma \cdot \boldsymbol{Contam} + \delta \cdot \boldsymbol{BldgType}$
    where $Contam \in \{1, 2, 3\}, BldgType \in \{1, 2, 3\}$.

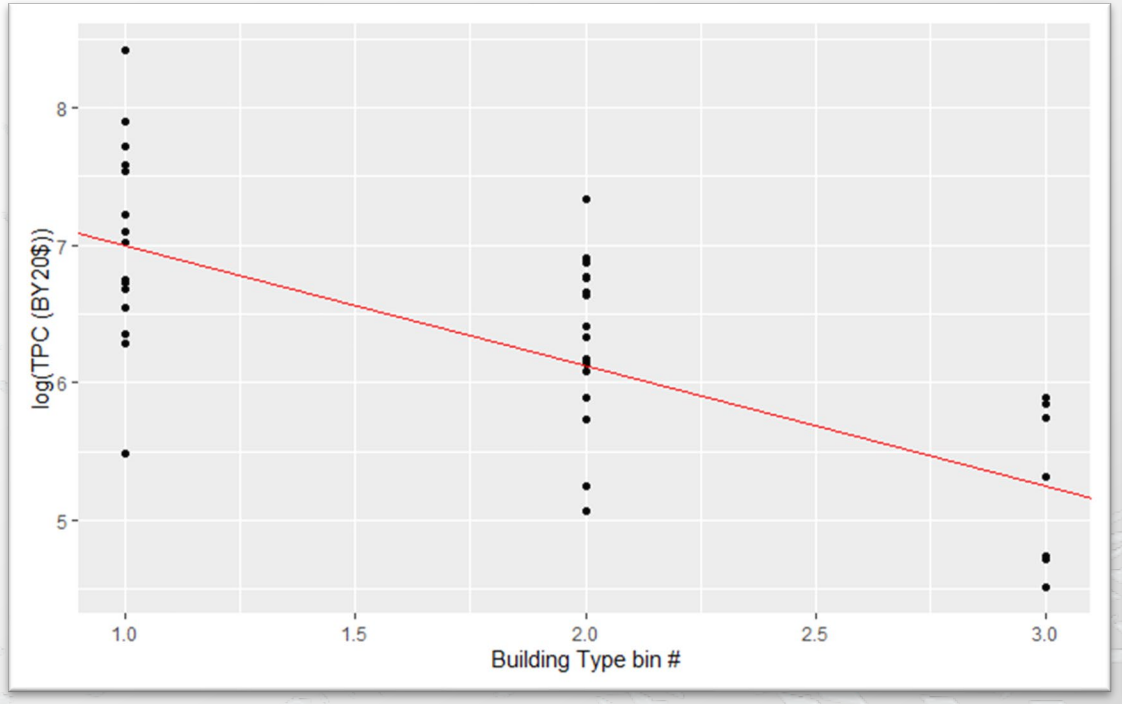
Contamination	Building Type	Contam	Bldg
Radiological	Technical	1	1
Lead/asbestos	Technical	2	1
None	Technical	3	1
Radiological	Non-technical	1	2
Lead/asbestos	Non-technical	2	2
None	Non-technical	3	2
Radiological	Temporary	1	3
Lead/asbestos	Temporary	2	3
None	Temporary	3	3

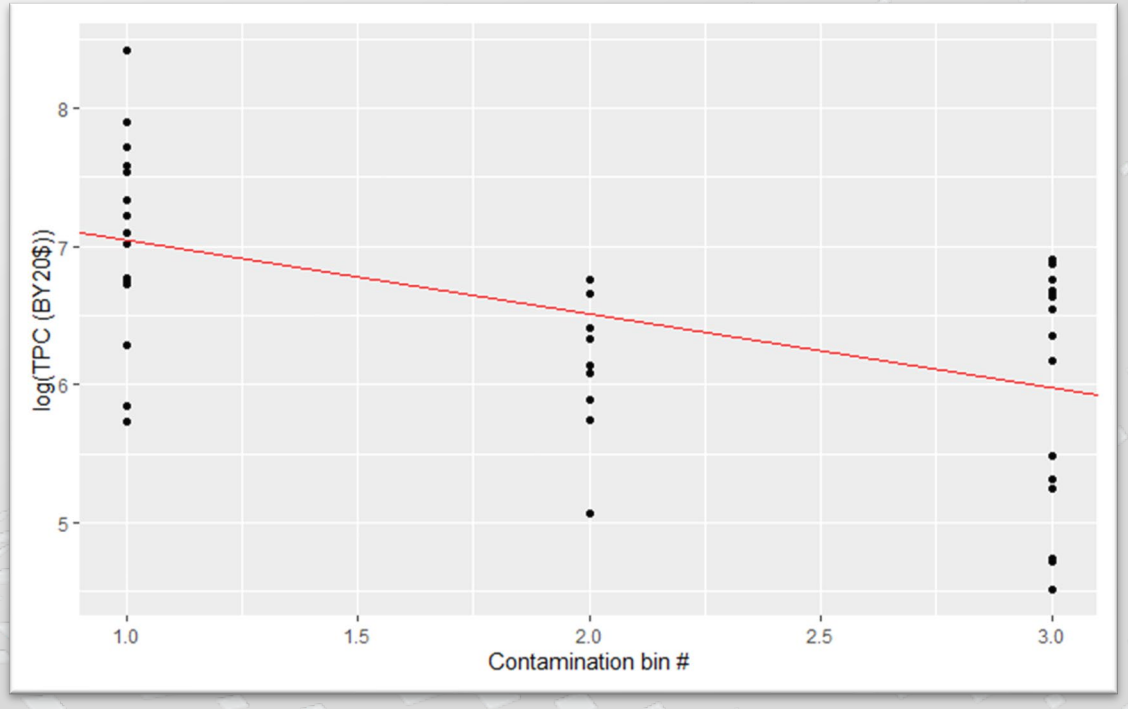
2 parameters

The trouble with label encoding

- This model form implicitly assumes that the cost difference (in log space) between contamination bins #1 and #2 is the same as the difference between bins #2 and #3.
  - You can see this if you separate out the relevant part of the regression equation:
    - $\log(TPC) = \cdots + \gamma \cdot \boldsymbol{Contam} + \cdots$
  - Does cost really increase linearly with contamination bin number?

Label encoding with DICEROLLER





When can we use label encoding?

- When the ordering/hierarchy makes sense.
  - You'll probably do this automatically when you look for cost drivers in your dataset.
- When the spacing between labels makes sense.
  - In DICEROLLER, we changed the value of the second contamination category from 2 to ~2.14 so that it better lined up with the line connecting categories 1 and 3.
  - Essentially, we've added an extra step to the regression and additional parameters to the model.

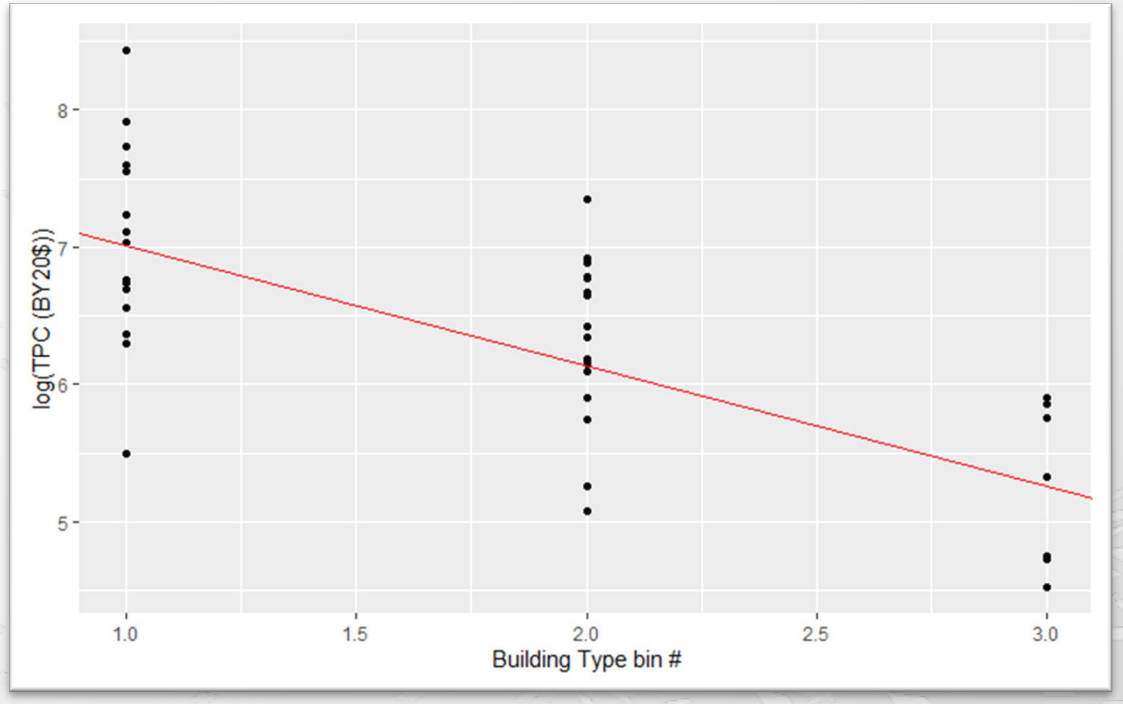
Contamination	Building Type	Contam	Bldg	Contam	Bldg
Radiological	Technical	1	1	1	1
Lead/asbestos	Technical	2	1	2.14	1
None	Technical	3	1	3	1
Radiological	Non-technical	1	2	1	1.98
Lead/asbestos	Non-technical	2	2	2.14	1.98
None	Non-technical	3	2	3	1.98
Radiological	Temporary	1	3	1	3
Lead/asbestos	Temporary	2	3	2.14	3
None	Temporary	3	3	3	3

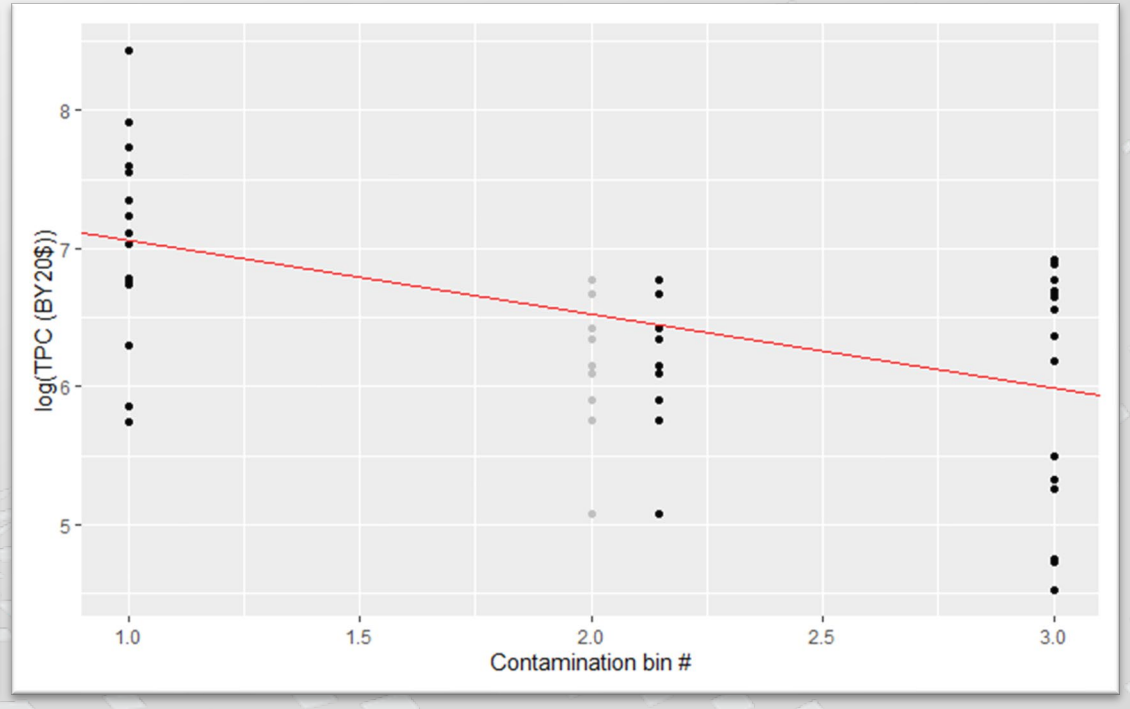
One-hot encoded

2 parameters

4 parameters (2 coefficients + the middle category labels)

Improved label encoding





One-hot encoding

- One-hot encoding is a term used by the machine learning community.
  - Also called dummy encoding.*
- Assign data a value of 1 if it belongs to a particular group within a category, and 0 if not.
- If you have k groups within a category, then use k-1 dummy variables.
- For example, contamination in DICEROLLER:
  - $\log(TPC) = \cdots + \delta_1 \cdot C_1 + \delta_2 \cdot C_2$

	EC1	EC2
Radiological contamination	1	0
Lead or asbestos contamination	0	1
No contamination	0	0

Multiple categorical variables

- Interactions between multiple categorical variables can – and should – be accounted for in regression models.
  - Not clear how to do this in Label Encoding.
  - Straightforward with One-Hot Encoding, but rapidly drives up the number of parameters.
- The most general model should include all possible interactions between variables.
  - You can then pare this model back by removing terms which are not statistically significant to the regression.
  - For DICEROLLER, the most general model form would be 18 terms (3 contamination categories times 3 building types, times 2 to account for interactions with/without GSF):

$$\log(TPC) = \beta_1 + \beta_2 D_{11} + \beta_3 D_{12} + \beta_4 D_{13} + \beta_5 D_{21} + \beta_6 D_{22} + \beta_7 D_{23} + \beta_8 D_{31} + \beta_9 D_{32} + \cancel{\beta_{10} D_{33}}$$
$$+\log(GSF) * (\beta_{11} + \beta_{12} D_{11} + \beta_{13} D_{12} + \beta_{14} D_{13} + \beta_{15} D_{21} + \beta_{16} D_{22} + \beta_{17} D_{23} + \beta_{18} D_{31} + \beta_{19} D_{32} + \cancel{\beta_{20} D_{33}})$$

Contamination	Building Type	Contam	Bldg	Contam	Bldg	Group							
Radiological	Technical	1	1	1	1	1	0	0	0	0	0	0	0
Lead/asbestos	Technical	2	1	2.14	1	0	1	0	0	0	0	0	0
None	Technical	3	1	3	1	0	0	1	0	0	0	0	0
Radiological	Non-technical	1	2	1	1.98	0	0	0	1	0	0	0	0
Lead/asbestos	Non-technical	2	2	2.14	1.98	0	0	0	0	1	0	0	0
None	Non-technical	3	2	3	1.98	0	0	0	0	0	1	0	0
Radiological	Temporary	1	3	1	3	0	0	0	0	0	0	1	0
Lead/asbestos	Temporary	2	3	2.14	3	0	0	0	0	0	0	0	1
None	Temporary	3	3	3	3	0	0	0	0	0	0	0	0

2 parameters

4 parameters (2 coefficients + the middle category labels)

8 parameters

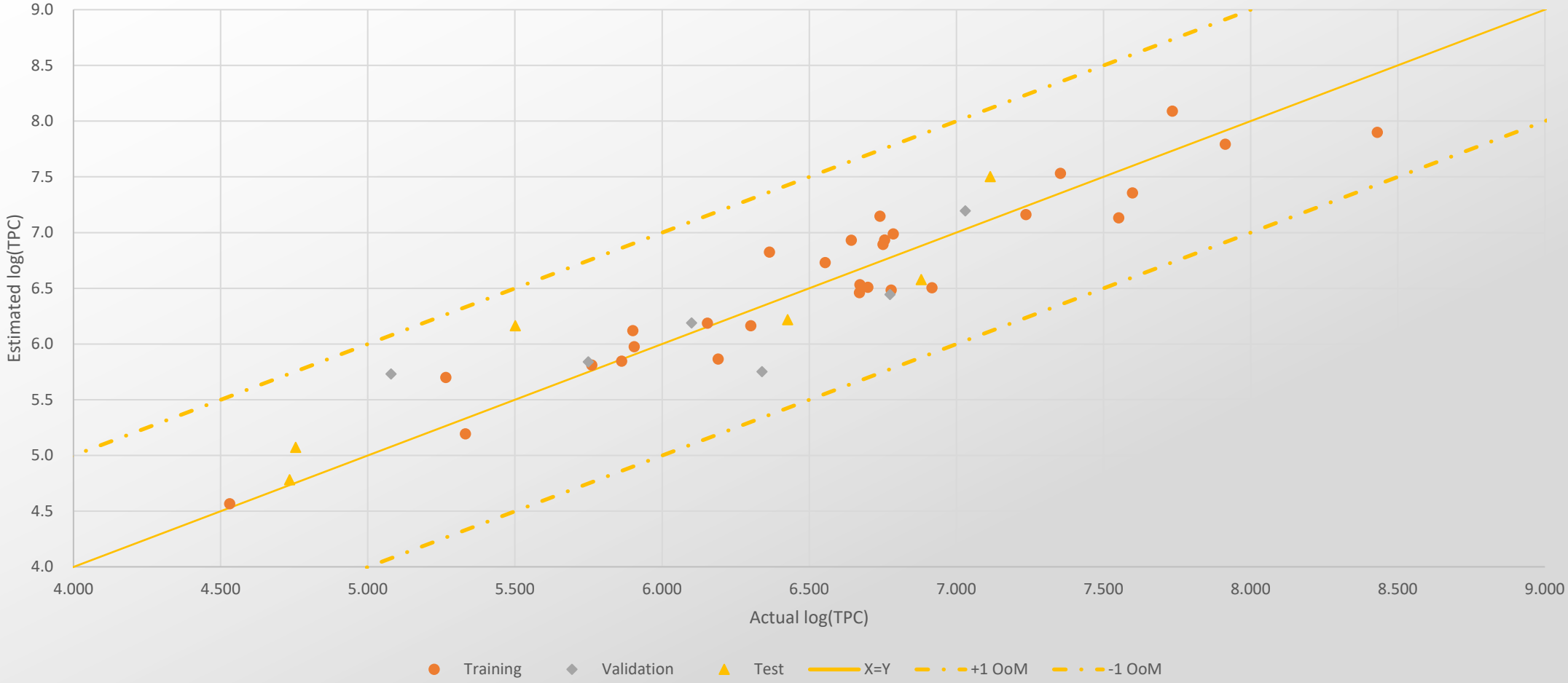
The trouble with one-hot encoding

- **Lots** of parameters
  - The number of parameters increases quickly with the number of categorical variables and the number of categories within each.
  - Fewer remaining degrees of freedom
  - Risk of overfitting
- Unreliable if you have few data points per group within a category.
  - May lead to false claims of statistical significance.

Two DICEROLLER models

- Option 1: Modified label encoding
  - $log(TPC) = \alpha + \beta \cdot log(GSF) + \gamma \cdot Contam + \delta \cdot BldgType$
  - where $Contam \in \{1, 2.14, 3\}, BldgType \in \{1, 1.98, 3\}$
  - Predicts log(TPC) with mean squared error of 0.28 for the training data set, 0.39 for the validation data set, and 0.31 for the test dataset.
- Option 2: One-hot encoding
  - $log(TPC) = \alpha + \beta \cdot D_{NoneTemp} + log(GSF)\ (\gamma + \delta \cdot D_{RadTech})$
  - where $D_{RadTech} = \begin{cases} 1\ if\ rad\ contaminated\ technical\ facility \\ 0\ otherwise \end{cases}$, etc.
  - Predicts TPC with mean squared error of 0.32 for the training data set, 0.42 for the validation data set, and 0.33 for the test dataset.

Model validation

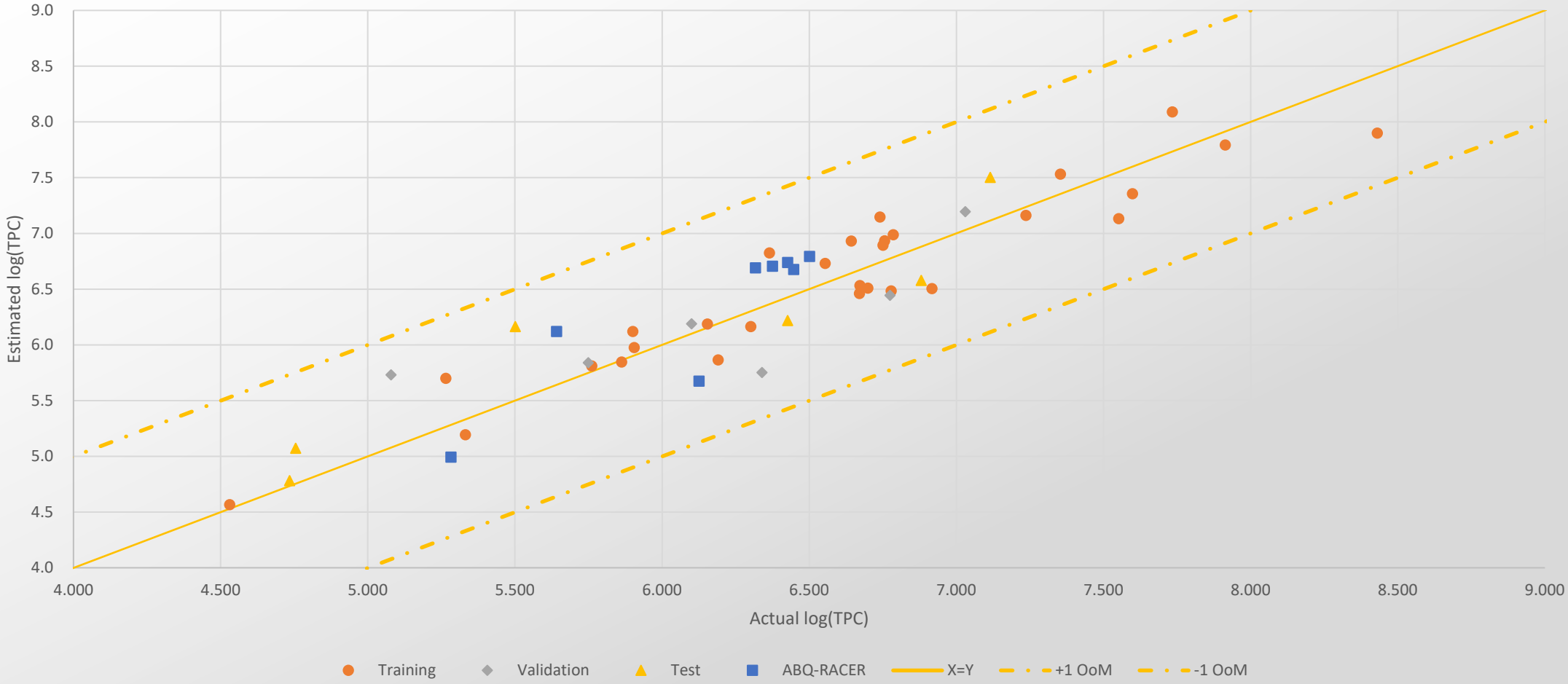


Model-to-model comparison

- "Remedial Action Cost Engineering and Requirements (RACER) is a cost estimating system that was developed under the direction of the U.S. Air Force for estimating environmental investigation and cleanup costs for the annual budgeting and appropriations process." [1]
- RACER is more detailed than DICEROLLER, which was designed for early-stage estimates.
- RACER was used to derive a cost estimate when the NNSA Albuquerque was planned for demolition.

[1] Source: Remedial Action Cost Engineering Requirements (RACER™) - https://frtr.gov/ec2/ecracersystem.htm

Model-to-model comparison



Conclusions

- We developed a model that met our requirements:
  - High-level for early-stage estimates
  - Easy to use
  - Small number of parameters, which should be easy to identify at early stages
  - Covers a wide range of project scope, size, costs, etc.
  - Based on historic data
  - Capable of producing AACE Class 5 quality estimates

Recommendations

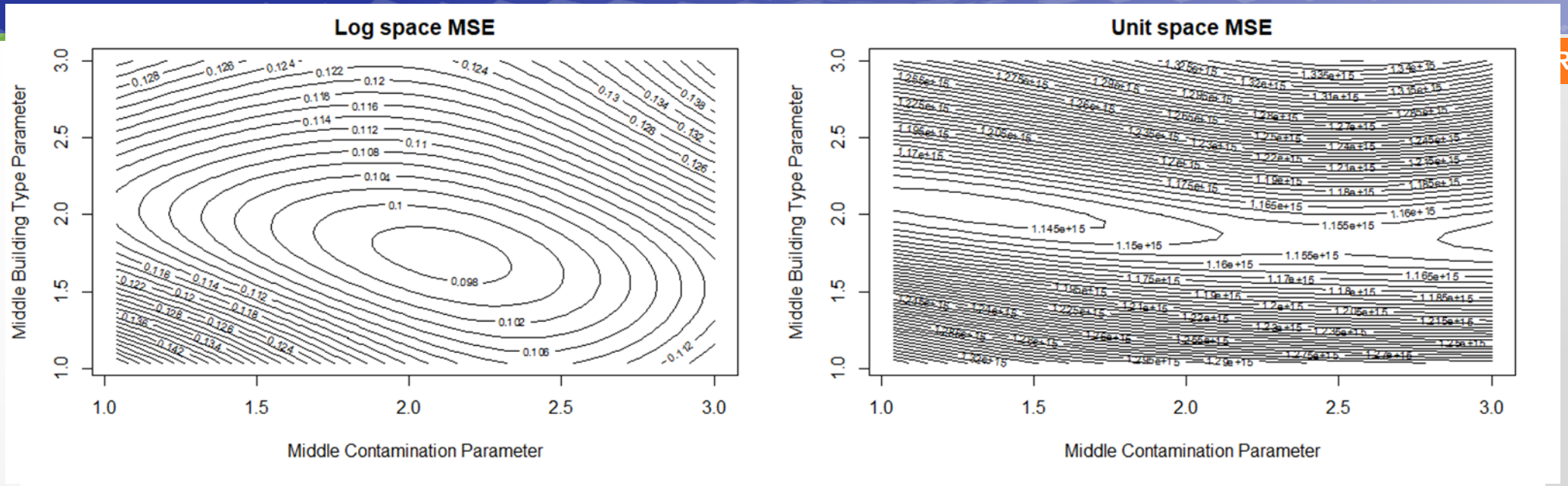
- Consider having a separate regression equation for each category, instead of a “one-size-fits-all” equation.
- Label encoding can be okay if you’re sure there is a hierarchy in your data and if you space it out properly.
- Try to have at least three data points per category.
  - A category containing a single data point means you have a parameter [over-] tuned to that individual point.
- Compare the variance between different groups.

To learn more:

- *Using Dummy Variables in CER Development* Dr Shu-Ping Hu and Alfred Smith, CCEA
  - https://www.iceaaonline.com/ready/wp-content/uploads/2021/10/JCAPv10i1Oct2021.pdf
- *Categorical encoding using Label-Encoding and One-Hot-Encoder* Dinesh Yadav, Towards Data Science
  - https://towardsdatascience.com/categorical-encoding-using-label-encoding-and-one-hot-encoder-911ef77fb5bd
- Also worth checking out the Wikipedia articles on Dummy variables (statistics) and Categorical variables

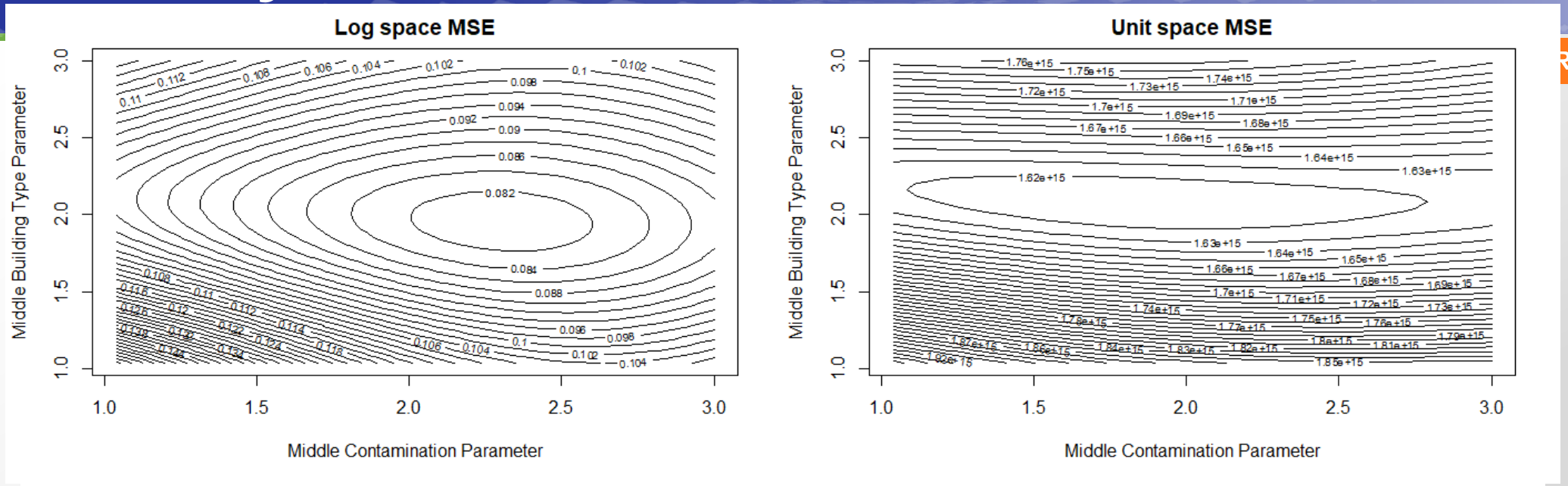
Backup

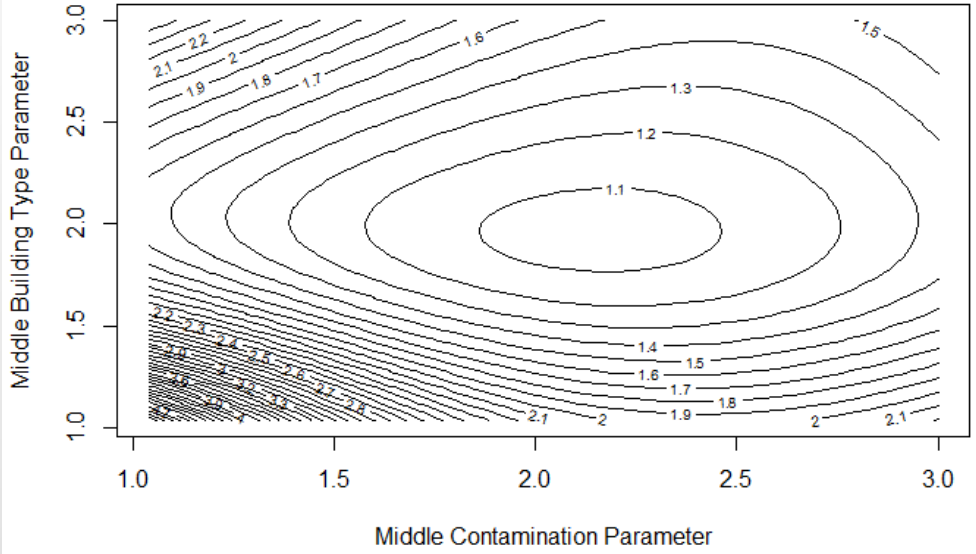
All data

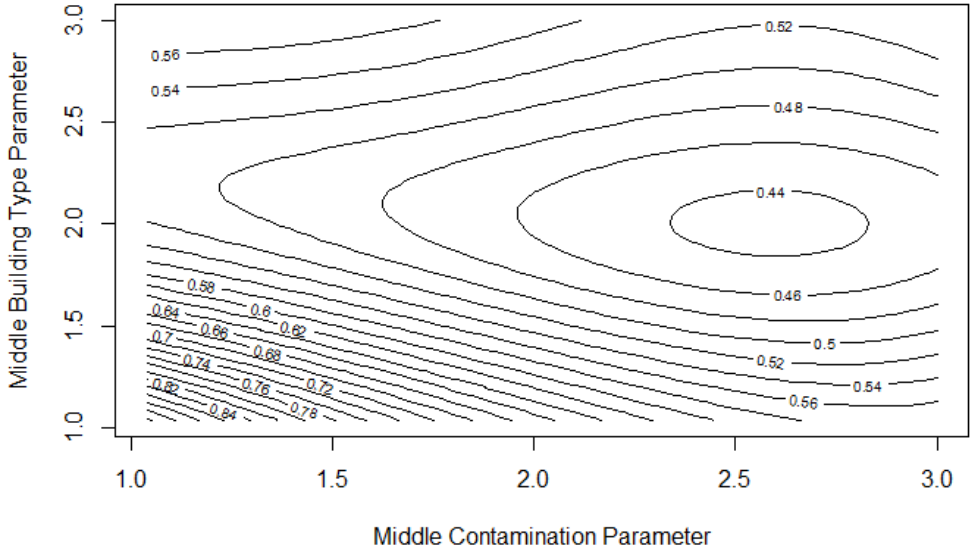


National Nuclear Security Administration

Training data only

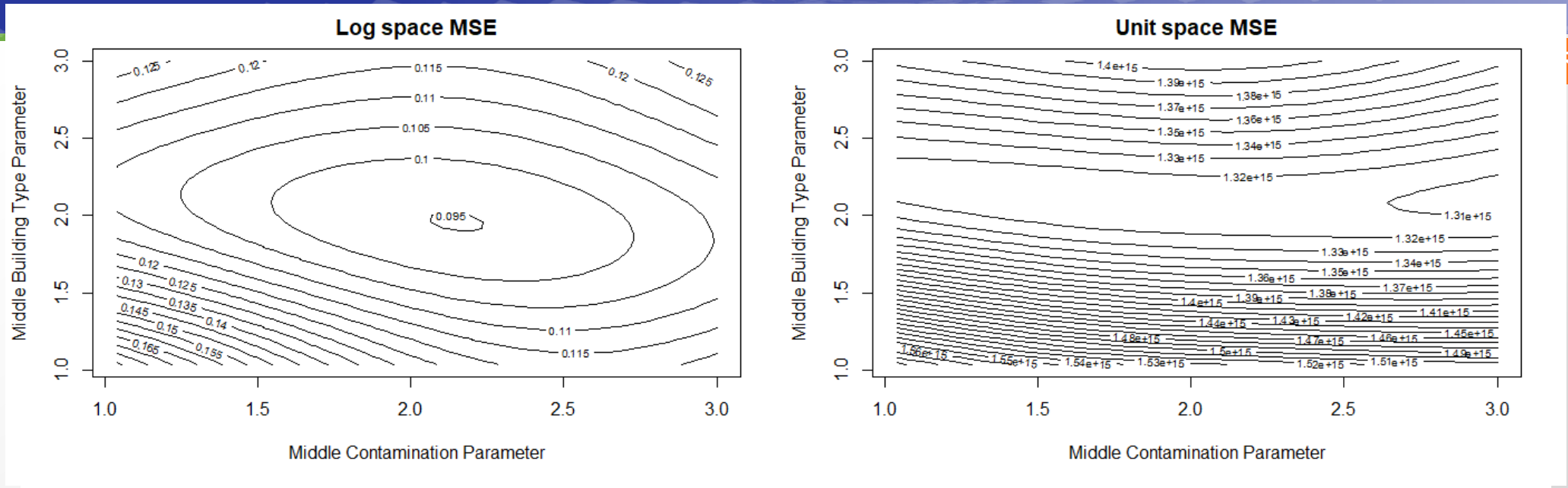


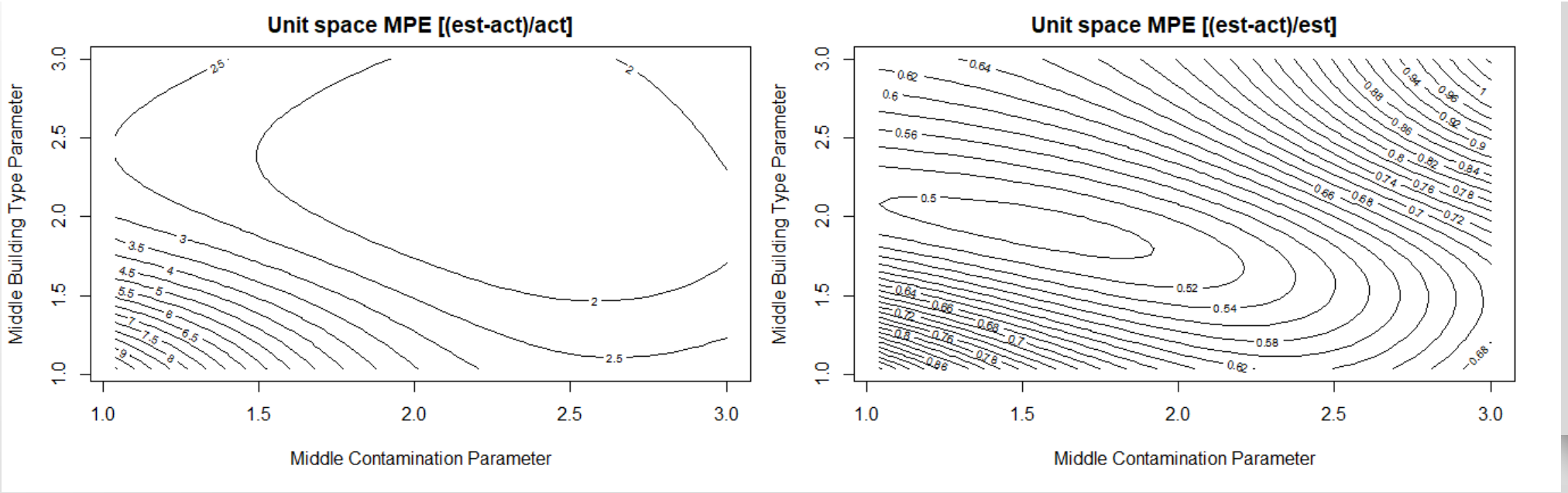




National Nuclear Security Administration

Training + validation data





National Nuclear Security Administration