Todo list

Evaluating Statistical Methods for Nuclear Forensics Analysis

by

Arrielle C. Opotowsky

A preliminary report submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Nuclear Engineering & Engineering Physics)

at the

UNIVERSITY OF WISCONSIN–MADISON

November 2017

Preliminary Examination Committee:

Rebecca M. Willet, Professor, Electrical & Computer Engineering

Charles F. Weber, Distinguished Scientist, Oak Ridge National Laboratory

Jake P. Blanchard, Professor, Nuclear Engineering & Engineering Physics

Douglass L. Henderson, Professor, Nuclear Engineering & Engineering Physics

Paul P.H. Wilson, Professor, Nuclear Engineering & Engineering Physics

For Steve

ACKNOWLEDGMENTS

This proposal would not be possible with the wisdom and patience of my advisor, Paul Wilson. I honestly don't have words for how grateful I am to you. I'm also appreciative of the CNERG community for technical and non-technical assistance; may quiche recipes be forever shared during important phone calls. Kelly Burton and Max Lagally have invested much effort into my success and convinced me that graduate school was the right path for me—more than once. My GERS friends have given me so much in and out of school, especially José Roberto, Richard, and Chandler. I have also received generous funding from the National Science Foundation and the Department of Homeland Security.

Mountains of personal support motivated me here and kept me here, which I do not take for granted. Steven W. Harrell, my chosen family, inspired me to get all my KSAs, no matter where I wanted to find them. The "If you're gonna be dumb, you gotta be tough" mentality hilariously applies to a PhD program. Robin, you have been such a light in my life for over a decade and always remind me why I came back to grad school. And to my friends for 15 years, thanks for keeping in touch despite the gaps. Denise and April, it's been amazing to watch your wonderful transformations. Ruthie, you push me to be fierce, spittin' truths and slaying your way through life. Maurice, thanks for the help when I was struggling. Lou, thanks for being one of my best friends and sources of laugh lines. Liz, I

love finding your inspirational notes around my home and thanks for all the meals. For endlessly encouraging my academic pursuits, I'm appreciative of my California family, Mel, Bonnie, Joelle, and Jamie. Finally, my Madison family has blessed me in countless ways: Shan, Ninja, Peter, Drax, Heather, Burnie, Marit, Sarah, James, BLou, Fetal, Matt,

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The purpose of this work is to evaluate the utility of statistical methods as an approach to determine forensics-relevant quantities for commercial spent nuclear fuel as less information is available. Machine learning algorithms will be used to train models to provide these values (e.g., reactor type, time since irradiation, burnup) from the available information. The training data will be simulated using ORIGEN, which will provide an array of nuclide concentrations as the features (X) and the parameters of interest (y) are provided from the simulation inputs. Information reduction will be carried out using computationally generated gamma spectra; the radionuclide concentrations from the simulations can be converted into gamma energies, which then undergo a detector response calculation to represent real gamma spectra as closely as possible. Machine learning best practices will be used to evaluate the performance of the chosen algorithms, and inverse problem theory will be used to provide an interval of confidence in the model predictions.

copied from goals, fix up

1 INTRODUCTION

The realm of nuclear security involves many parallel efforts in nonproliferation (verification of treaty compliance, monitoring for smuggling, proper storage and transportation of nuclear materials), cyber security, minimizing stocks of weaponizable materials, disaster response training, and nuclear forensics. All of these efforts have been continually improving, but there was a gap regarding the ability of the United States (US) to coordinate and respond to a nuclear incident, especially with the technical portion of nuclear forensics: characterization and analysis. After all, the first textbook on the topic was published in 2005 [10]. In 2006, the US Department of Homeland Security (DHS) founded the National Technical Nuclear Forensics Center (NTNFC) within the Domestic Nuclear Detection Office (DNDO). The mission of the NTNFC is to establish a robust nuclear forensics capability to attribute radioactive materials with demonstrable proof.

There are many fields that contribute to the nuclear forensics capability, such as radiochemical separations, material collection techniques, improving detector technology, material library development, and identifying forensic signatures. These needs vary based on whether the material being collected is post-detonation (e.g., bomb debris) or pre-detonation (e.g., spent nuclear fuel (SNF)). In the pre-detonation realm, this project focuses on statistical methods to identify correlated material characteristics, which can lead to new forensic signatures.

1.1 Motivation

Nuclear forensics is an important aspect of deterring nuclear terrorism even though it is not, at first glance, thought to be preventative nuclear security. The most common defense of the field is that nuclear forensics capability deters state actors, not terrorist organizations. While it is true that a strong capability encourages governments to be more active in prevention of nuclear terrorism, it can also deter the terrorist organizations as well by increasing their chances of failure. Less destructive success tends to be more valued than high-risk mass destruction. In addition to influencing governments and making nuclear terrorism higher risk for organizations, nuclear forensics can assist in cutting off certain suppliers of nuclear materials or technologies (e.g., nuclear specialists that are only involved for financial reasons, access to state suppliers). Shutting off the sources builds a concrete barrier to nuclear terrorism. Therefore, nuclear forensics is considered impede this form of terrorism in both tangible and abstract ways [8].

Following the prevention value of nuclear forensics, it is important to understand the process of the technical portion of the investigation and how that can be improved. In the event of a nuclear incident, such as the retrieval of stolen special nuclear material (SNM) or the detonation of a dirty bomb, it is necessary to learn as much as possible about the source of the materials in a timely manner. In the case of non-detonated SNM, knowing the reactor parameters that produced it can point investigators in the right direction

in order to determine the chain of custody of the interdicted material. Section 1.1.1 covers the specific needs of the nuclear forensics community for SNF provenance, and Section 1.1.2 discusses how alternative computational approaches are useful, with a focus on why statistical methods in particular are being pursued.

1.1.1 Needs in Nuclear Forensics

The process of technical nuclear forensics includes the analysis and interpretation of nuclear material to determine its history, whether that be intercepted SNF, uranium ore concentrate (UOC), or the debris from an exploded nuclear device. After the technical portion is complete, intelligence data can be used to aid in material attribution; this is the overall goal of nuclear forensics.

After a nuclear incident, the material or debris is sampled and evaluated through many techniques that provide the following information: material structure, chemical and elemental compositions, and radioisotopic compositions and/or ratios. These measurements or ratios comprise the forensic signatures of the sample in question. These signatures can be analyzed with specific domain knowledge; for example, UOC will have trace elements depending on where it was mined from. They can also be analyzed against a forensics database in the case of SNF.

Measurement needs and techniques vary vastly depending on the material, as does the type of signature. This study focuses on non-detonated materials,

specifically, SNF. It is important to determine if some intercepted SNF is from an undisclosed reactor or a commercial fuel cycle to attribute it to an entity or state. This is typically done by obtaining select chemical and elemental signatures and isotopic ratios, and comparing these measurements to those in an existing forensics database of reference SNF. The signatures for SNF correlate to characteristics that can, in a best case scenario, point to the exact reactor from which the fuel was intercepted. The reactor parameters of interest are the reactor type, fuel type and enrichment at beginning of irradiation, cooling time, and burnup [1, 19, 20].

The current and future work of this study are designed based on two primary needs to bolster the US nuclear forensics capability: post-incident rapid characterization, and forensics database challenges and imperfection.

First, our best measurement techniques may not be available in an emergency scenario, and fast measurements typically yield inaccurate results. Currently, both radiological measurements and mass spectrometry are used in nuclear forensics exercises. Because these techniques have a multitude of variants within each category, there are differing levels of certainty of the results. Easily deduced is that the faster and cheaper methods also provide the most uncertain values. Thus, the main tradeoff is between time/cost and amount of information gained. A lofty goal would be to develop methods that provide instantaneous information, reliable enough to guide an investigation (e.g., within 24 hours). In the case of SNF, it takes weeks in a lab to measure isotopes via advanced (cooled detector) gamma

spectroscopy and mass spectrometry equipment. A handheld detector that measures gamma spectra could provide the fast measurements to calculate isotopic ratios for the above-mentioned fuel parameters of interest. However, while this nondestructive analysis is rapid, it is also difficult to evaluate because of the presence of overlapping peaks and the fact that uncertainties differ significantly because of the detector reponse, environment, storage, electronics, etc. Broadly speaking, gamma spectra give less information at a higher uncertainty than the near-perfect results of some destructive mass spectrometry techniques, like TIMS.

Second, forensics databases are imperfect; this is three-fold. Because of the values needed for material provenance and the number of measurement types, the forensics databases are 1. highly multidimensional and have 2. inconsistent uncertainties or missing data entries, respectively. Thus, direct comparison between measurement results and a database therefore may not yield accurate parameter predictions. Furthermore, 3. forensics databases are kept by individual countries, and the reactor operation history information is somewhat well guarded, so it may be difficult to study SNM from a country that has a different fuel cycle. It is proposed that using a machine-learned model may be able to combat these issues; this is introduced next in Section 1.1.2.

find MS technique and citation

find best citation

1.1.2 Contribution of Statistical Methods

As previously mentioned, there are two main issues that are being addressed for forensics of SNF: database issues and speed of characterization. Many have begun considering computational techniques developed by nuclear engineers to calculate the parameters relevant to nuclear forensics analysis. One example is the INverse DEPletion THeory (INDEPTH) tool [1, 19, 20]. INDEPTH uses an iterative optimization method involving many forward simulations to obtain reactor parameters of interest given some initial values.

Another approach utilizes artificial intelligence to solve nuclear forensics problems, such as implementing searching algorithms for the database comparison step [3] and machine learning for determining reactor parameters from SNF characteristics [2, 6, 7, 11, 12, 13, 17]. A variety of statistical and machine learning tools have been used to characterize spent fuel by predicting categories or labels (e.g., reactor type, fuel type) as well as predicting values (e.g., burnup, initial enrichment, cooling time). The former uses classification algorithms and the latter uses regression algorithms, many of which can be altered to perform both classification and regression. There is some promising work discussed in Section 2.4 that shows certain applications of machine learning can provide an additional tool for solving the forensics problem, both qualitatively (for visualization) and quantitatively (for prediction).

Statistical methods have the uniqueness of requiring minimal domain knowledge via machine learning algorithms that predict the characteristics or values of interest [2, 6, 7, 11, 12, 13, 17]. They first create a black-box

statistical model using the database entries, and can predict the reactor parameters of an unknown sample based on that model. Having a machine-learned model based on a large number of simulations may also overcome the challenges of missing data, irregular uncertainty, or lack of information on other fuel cycles. This logic also follows for other computational methods using a large number of simulations. Although not encompassed in a resuable model, they also could overcome missing data, irregular uncertainties, or ignorance of different or non-commercial fuel cycles. Also, it is generally known that statistical methods will be able to either use or reduce the dimensions in the forensics databases, which is another unique characteristic.

Figure 1.1 compares the INDEPTH and statistical methodologies, both of which use simulated SNF. While not all steps are required to be equivalent, the only difference here is the method one chooses to obtain reactor parameters. Both workflows address speed of characterization, as it is intended to have gamma spectra as the inputs. Both workflows also address many of the database issues, described above.

Because INDEPTH is better studied and validated than statistical methods, this work focuses on a statistical approach but with the intention to compare methodologies. Since focusing on rapid characterization is also a main goal, the data input to the tool will be manipulated to reflect the information reduction of a gamma detector measuring the SNM. Thus, this work evaluates to what degree statistical methods will be able to predict reactor parameters with respect to the type of training data used.

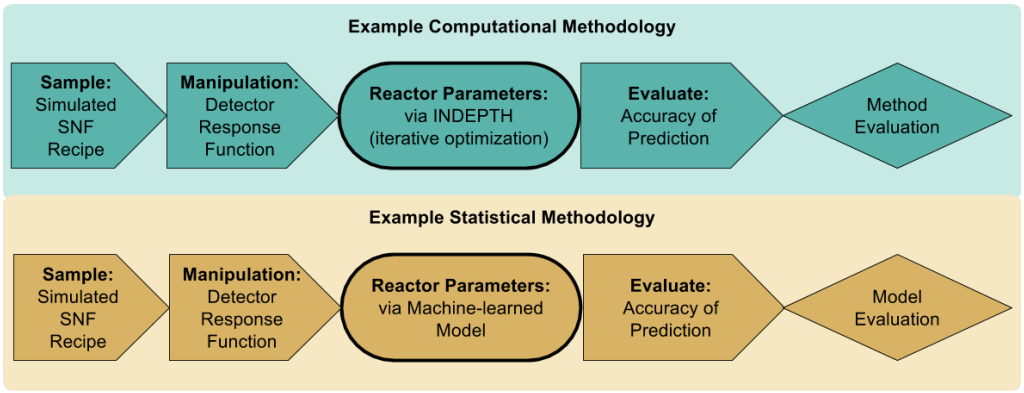


Figure 1.1: Diagram showing examples of forensics research using computational methodologies.

1.2 Methodology

As previously mentioned, the typical workflow of the technical portion of a forensics investigation is to take measurements of an unknown material and compare those measurements to databases filled with previously measured standard materials. As this work focuses on SNF, these measurements are elemental, chemical, and radiological in nature. Because creating databases from real measurements to represent SNF from reactor technologies from around the world is impossible, the database in this study will be created from high-fidelity simulations via Oak Ridge Isotope GENeration (ORIGEN) [16] within the SCALE code system [14] for modeling and simulation.

In the simulation and statistical learning paradigm, we need to determine how much information to what quality is needed to train a machine-learned model; the model must give appropriate predictions of reactor parameters given a set of measurements from a test sample of interdicted SNF. Thus,

the space that the training set encompasses must be chosen carefully so as to represent the typical scenario for stolen SNF.

The next step is to choose an algorithm that performs statistical learning. Statistical learners have varied strengths and weaknesses based on what is being predicted and how they implement optimization. Chosen for this study are simple regression algorithms for burnup prediction: nearest neighbor and ridge regression. For comparison, support vector regression is used because it is known to handle highly dimensional data sets well. These algorithms are introduced in Section 2.2.1.

After the training is complete, the results of each models' predictions must be evaluated. Typically, a test set is used to compare against the model created from the training set. The testing error can therefore be tabulated with respect to various specifications such as the training set size, number of features, or algorithm parameters (regularization terms, etc). These results are broadly known as diagnostic plots and show if the algorithms' predictions are due to good performance or bad fitting.

After the models are evaluated using machine learning best practices, it will be important to compare them both against each other and against other computational forensics parameter methods. Thus, a Bayesian approach from the field of inverse problem theory will be used to give the probability density of the predictions so that the statistically generated predictions can be evaluated directly against other solutions, such as optimization-based methods or direct computations.

Next, information reduction (within the training and/or testing data sets) must be investigated to extend this workflow to mimic that of the real world. The primary example investigated here is the reduction of information quality via gamma ray detectors, as they can provide fast results. If an algorithm could overcome the limitations of gamma detection and still provide useful results, this would warrant further studies and perhaps be field-applicable.

Thus, ultimately, the goal is to answer the question *How does the ability to determine forensic-relevant spent nuclear fuel attributes degrade as less information is available?*.

1.3 Goals

The main purpose of this work is to evaluate the utility of statistical methods as an approach to determine nuclear forensics-relevant quantities as less information is available. Machine learning algorithms are used to train models to provide these values (e.g., reactor type, time since irradiation, burnup) from the available information. The training data is simulated using the SCALE 6.2 code suite [14], which provides an array of nuclide concentrations as the features (X) and the parameters of interest (y) are provided from the simulation inputs. Information reduction is carried out using computationally generated gamma spectra; the radionuclide concentrations from the simulations can be converted into gamma energies, which then

undergo a detector response calculation to represent real gamma spectra as closely as possible. Machine learning best practices are used to evaluate the performance of the chosen algorithms, and inverse problem theory is used to provide an interval of confidence in the model predictions.

The necessary background is covered in Chapter 2. First, an introduction to the broader field of nuclear forensics is in Section 2.1 to place this work in the context of the technical mission areas. After that, a short discussion of the field of machine learning, the algorithms used, and validation methods are in Section 2.2. Section 2.3 includes information about the codes used to generate the training data, via fuel cycle simulation, detector response function, and isotope identification of gamma spectra. Lastly, a review of statistical methods being used in studies of forensics analysis is covered next in Section 2.4.

After the existing work is discussed and the gap that this work will fill is identified, the methodology and a demonstration of the experimental components is introduced next in Chapter 3. This will cover the simulated training data in Section 3.1, the parameters behind the learned models in Section 3.2, and the process of model evaluation in Section 3.3.

Finally, Chapter 4 summarizes the official thesis research proposal. After the preparatory tasks are covered in Section 4.1, there are three experiments outlined in Sections 4.2, 4.3, and 4.4. Qualitiative hypotheses as well as alternativev directions for risk mitgation are discussed throughout this chapter as well.

2 BACKGROUND AND LITERATURE REVIEW

This chapter provides a background and literature review of the necessary components for this project. Section 2.1 outlines the broader field of technical nuclear forensics, with a focus on the area that motivates this project. Section 2.2 introduces the field of machine learning for an uninitiated audience, covers the relevant algorithms, and presents the methods field practitioners use for validation. Next, Section 2.3 covers the computational methods used to generate the training data for the machine learning input. Finally, the marriage of Sections 2.1, 2.2, and 2.3, is presented in Section 2.4, which is a review of previous work applying statistical methods to the nuclear forensics analysis of pre-detonated nuclear materials.

2.1 Nuclear Forensics

Nuclear forensics comprises a large part of an investigation into a nuclear incident, such as interdicted nuclear material or the detonation of a weapon containing radioactive components. The forensics portion of the investigation encompasses both the analysis of nuclear material and/or related paraphrenalia as well as the interpretation of these results to establish nuclear material provenance. The former has many technical aspects, relying on a range of nuclear science and chemistry. The latter involves intelligence and political considerations of the material analyses for attribution. This review

will only consider the technical portion of the nuclear forensics workflow.

First discussed are the types of forensic investigations in Section 2.1.1, followed by an introduction to inverse problem theory in Section 2.1.2 as a way to evaluate the results of forensic methods.

2.1.1 Types of Nuclear Forensics Investigations

The technical programs researching improvements to the US's nuclear forensics capabilities are split between the type of material being investigated. The analysis of irradiated debris from a weapon has different collection and measurement requirements than recovered SNF from a commercial reactor. This separates the field into post-detonation and pre-detonation nuclear forensics. While both are discussed below in Sections 2.1.1.1 and 2.1.1.2, respectively, there is more focus on pre-detonation topics since this work is based on SNF.

2.1.1.1 Post-Detonation

Post-detonation nuclear forensics requires a diverse set of measurements to obtain the following information: identification of nuclear material, reconstruction of the weapon device design, and reactor parameters for nuclear material provenance. This could apply to an improvised nuclear device or a nuclear bomb. In conjunction with the measurements and characterization are a large array of logistical concerns, including recovery efforts, personnel safety, and material collection cataloging and transportation.

In the case of a full explosion using fissile material, the collection of materials and debris occurs as quickly as possible. It can be in the crater created by the explosion, further away from the center in the fallout, and in the atmosphere above or downwind from the detonation. These are collected by finding glass-like material near the epicenter, debris swipes in the fallout region, and advanced particle collection in the atmosphere via an airplane, respectively. While the epicenter cannot be reached for some time, the debris and atmosphere measurements of radioactive material can provide the yield of the weapon and whether it was made using uranium or plutonium. This along with other physical and chemical measurement allow device reconstuction to begin. Attribution begins to narrow to specific countries or organizations based on this information. [8]

The research needs for post-detonation focus on material collection and analysis as well as nuclear device modeling for rescontruction purposes. Ideally, most material sample collection would be done using automatic instrumentation. Additionally, bolstering the existing device modeling code for reverse engineering is needed. And, as with pre-detonation, a database of standard materials must be both strengthened and centralized. [8]

2.1.1.2 Pre-Detonation

Pre-detonation nuclear forensics investigations occur for every scenario in which non-detonated nuclear material has been found or intercepted. Although this could be an intact bomb, it is more likely that SNM intended

for a weapon would be the target of an investigation. Thus, the range of intact materials for measurement could be as small as one fuel rod. The goal is to determine the provenance of the SNM, which is generally done by reconstructing the irradiation process that created the material.

For SNF, where the material was obtained is the first step of the investigation. This would be gleaned from the reactor parameters and storage history (e.g., reactor type, cooling time, burnup), which requires first measuring and calculating certain values: isotopic ratios, concentration of chemical compounds, or existence of trace elements. Both radiological methods (e.g., gamma spectroscopy) and ionization methods (e.g., mass spectrometry) measure these quantities.

Although this is less of a humanitarian emergency than a post-detonation investigation, it is still important to have rapid characterization capabilities via on-site non-destructive analyses. As previously discussed in more detail in Section 1.1, however, the faster measurements result in poor measurement quality. Also, there is a need for research to combat the database issues, as an insufficient forensics database can reduce the accuracy and/or certainty of a reconstructed set of reactor parameters. Another area of research is deeper study of known forensics signatures or discovering new signatures with modeling, simulation, or statistical methods.

The Real World Methodology section in Figure 2.1 shows a typical workflow for the technical portion of a pre-detonation forensics investigation. After a sample is obtained, characterization begins. Next, the results of these

techniques are then compared against existing standard materials databases to obtain the desired reactor parameters. These steps would be performed iteratively in a real investigation, first using non-destructive measurements, and destructive measurements last. The following steps in Figure 2.1 are seeking out reactor history information, if available, and reporting all results to the investigators.

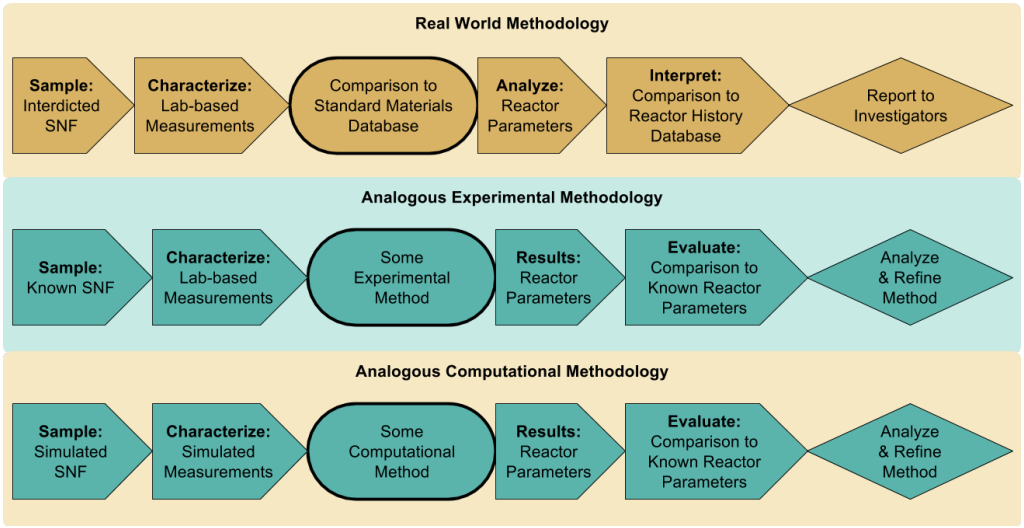


Figure 2.1: Diagram showing how research on experimental and computational methodologies parallel the typical real-world scenario.

Below the example of a real-world workflow in Figure 2.1 are analagous experimental workflows, both physical and computational. For researchers studying alternative measurement techniques or a slight difference in the overall approach, it is necessary to iterate through multiple studies using known materials to probe sensitivities or other weaknesses in the procedure.

Fix up diagram to be more aligned with discussion

2.1.2 Nuclear Forensics as an Inverse Problem

Nuclear forensics is a traditional inverse problem, which has been well documented in mathematics and many scientific disciplines. Understanding inverse problem theory can help systematically define both the solution methods and their limitations. This section provides an introduction to the topic as well as its application to nuclear forensics.

As outlined in a textbook on the formal approach to inverse problem theory [18], the study of a typical physical system encompasses three areas:

1. *Model parameterization*
2. *Forward problem:* predict measurement values given model parameters
3. *Inverse problem:* predict model parameters given measurement values

First, this shows that it is important to consider the parameters that comprise a model; this is denoted as the *model space*. This is not every measurable quantity; domain knowledge is necessary to determine the model space. In the nuclear forensics context for spent nuclear fuel, this would consist of, e.g., several isotopic ratios because they are known to have a relationship with the reactor parameters that created the fuel of interest.

Second, understanding the physical system also requires an understanding of the forward problem: predicting how a certain set of model parameters will affect the resulting measurements. The breadth of these end measurements provides the *data space*, which are all the conceivable results of a given

forward problem. So for spent nuclear fuel this would be, perhaps, the range of isotopic ratios typical of a commercial reactor.

Lastly, the inverse problem is statistical in nature: given some solution, there is a probability that the data measured is caused by some value(s) of a model parameter. Including measurement uncertainties broadens the linear model to a probability density of the parameters. The opposite is also true in the forward case: including parameter uncertainties broadens the forward problem results to a probability density of the potential measurement values.

In this way, we can define some probability that the answer is correct, given a set of measurements and their uncertainty. Inverse problem theory states that this follows the general form of Bayes' theorem, which is commonly expressed as follows:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.1}$$

where A and B are events, $P(A)$ and $P(B)$ are the probabilities that events A and B will occur, respectively, $P(B|A)$ is the prior probability that event B will occur given a known result for A, and $P(A|B)$ is the posterior probability that event A will occur given a known result for B.

This is can be mapped easily to the inverse physical system problem scenario. A would represent an occurence of a parameter in the model space, and B would represent the measurement of some value. Thus, $P(A)$ is the probability of a parameter existing without any knowledge of B. This is known as the likelihood, usually given by some theory about the

system. $P(B)$ is the probability of some measurement existing without any knowledge of A. This is known as the marginal likelihood, which is some homogeneous concept for the potential measurements that could be made (this only serves to scale to absolute probabilities and does not affect the relative probabilities). The prior probability $P(B|A)$ is the chance that a measurement is observed from a given parameter, representing the forward problem. Lastly, the posterior probability is the chance of some parameter existing given some measurement, representing the inverse problem solution [18]. It may be more intuitive to consider the conceptual version of Bayes' theorem below. A discussion of how these values are obtained takes place in Section 2.2.3.3.

$$Posterior = \frac{Prior * Likelihood}{Marginal\ Likelihood} \tag{2.2}$$

This framework is helpful for an experiment that intends to compare different methods for calculating the posterior probability of a system given some measurements. In the nuclear forensics context of pre-detonated materials, this would be a a set of probabilities for different parameters of interest, e.g., reactor type, burnup, cooling time, and enrichment of some interdicted spent fuel. The prior probabilities are obtained by a large set of forward problems, e.g., a database of spent fuel recipes and parameters. The likelihoods are obtained in differing ways. One method is expert-elicited values. Another is a predicted model from some theory or previously known relationship, e.g., empirical relations between isotopic ratios and certain

reactor parameters.

2.2 Machine Learning

Machine learning is a sub-field of artificial intelligence (AI) within the broad category of computer science. The goal of AI is to create computer systems that respond to their environment according to some set of criteria or goal. For example, self-driving vehicles have computers on board that learn to avoid curbs and humans. It is common knowledge that the use of AI has been expanding at a rapid rate in recent years. News stories of major tech companies' AI advancements are frequent and news articles abound with data on which jobs will be replaced with AI in the near future.

While its use has been increasing in the commercial sector, there is also much anecdotal evidence to support the existence of a rapid increase of AI use in academic research across many disciplines beyond robotics. AI systems have been used in detection (e.g., fraud or spam), medical diagnostics, user analysis (e.g., Netflix ratings), and a host of scientific disciplines that have increasing amounts of multivariate data.

Much of the recent advances to the field of AI have occured in the statistical realm, which forgoes domain knowledge in favor of large data sets. Thus, machine learning and statistical learning has become somewhat of a separate field. Machine learning research focuses on the underlying algorithms using mathematical optimization, methods for pattern recogni-

could explain more, cite changing science of

tion, and computational statistics. As an application, however, this study is not concerned with computational time, but rather the ability to correctly predict values and categories relevant to the nuclear forensics mission. This restricts the relevancy of the algorithms to the underlying theory and its impact on the resulting model's accuracy.

Machine learning algorithms can be separated into two main categories: unsupervised and supervised learning. The former groups or interprets a set of input data, predicting patterns or structures. The latter includes both the input and output data, enabling the trained model to predict future outputs. Broadly speaking, the unsupervised learning algorithms are designed for clustering data sets or dimensionality reduction (i.e., determining some subset or linear combination of features most relevant to the input data) of data sets. Supervised learning algorithms predict both discrete and continuous values via classification and regression, respectively. Some algorithms can perform both classification and regression, and neural networks can even be modified to perform either supervised or unsupervised learning.

As shown in Figure 2.2, a typical (supervised) machine learning workflow begins with a training data set, which has a number of *instances*, or rows of observations. Each instance has some *attributes*, also referred to as *features*, and a label, which can be a categorical label or discrete/continuous values.

The training data are then inserted into a statistical learner; this calculates some objective, minimizes or maximizes that objective, and provides some model. This model is typically evaluated using a testing set that has

the same set of attributes and labels (but different instances). The comparison of what the model predicts and the actual label gives the *generalization error*. Depending on the performance and application, the model may need improvement from more training and/or some changes in the algorithm parameters. Once the model is performing well enough and validated, it is finalized; then a user can provide a single instance and a value can be predicted from that.

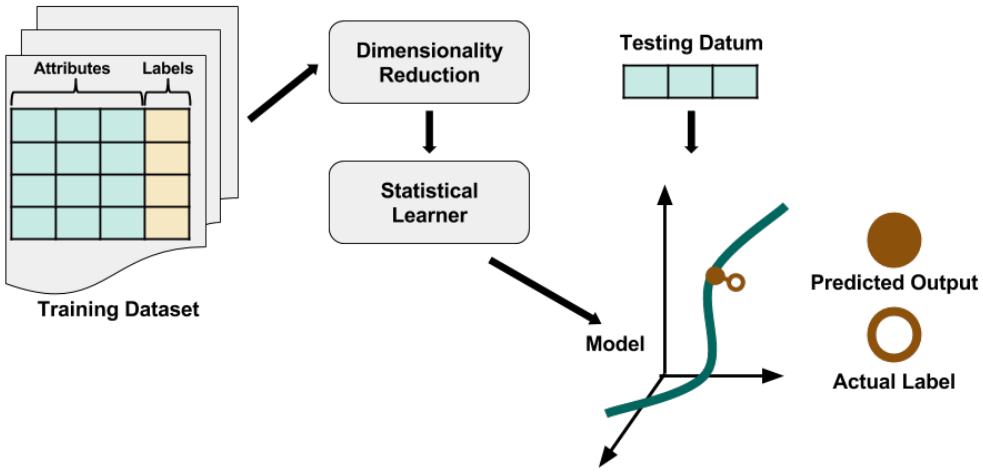


Figure 2.2: Schematic representing the workflow of a statistical learning regression algorithm

This study performs both classification and regression tasks using supervised learning algorithms. Differences among the underlying mathematics of the algorithms impact the trained models. Therefore the algorithms used in this study will be discussed in Section 2.2.1. Next, algorithm selection and assessment is covered in Section 2.2.2. Evaluating and optimizing algorithm performance is discussed in Section 2.2.3, as well as robustly comparing

different algorithms for validation.

2.2.1 Classification and Regression Algorithms

For relevant nuclear forensics predictions, both classification and regression algorithms must be used. For example, one may want to predict the reactor type based on some measurements (referred to as features) of spent fuel of an unknown source, and this would require a classification algorithm. Or perhaps the input fuel composition is relevant to an investigation on weapons intent, so a regression algorithm would be used to train a model based on some set of features. Since algorithm formulation impacts the resulting performance, they are discussed in detail below.

should add useful vocab: training data X and y, instances, features, etc

2.2.1.1 Linear Models

One of the simplest and most obvious methods of prediction is a linear model using a least-squares fit.

not sure about this organization

2.2.1.2 Nearest Neighbor Methods

Nearest neighrbor is

intro, alg math, explain regular-

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i \tag{2.3}$$

Nearest neighbor regression calculates a value based on the instance that is closest to it. The metrics for distance differ, but in this study, Euclidian distance was used. There is no learning in this regression, per se; the training set populates a space and the testing set is compared directly to that. [4]

2.2.1.3 Support Vector Machines

Support vector regression (SVR) is an extension of the popular classification algorithm, support vector machine (SVM). This algorithm was chosen because of its ability to handle highly dimensional data well, which in this study is approximately 300 features.

SVM classifies two classes by determining an optimal hyperplane, given by wx+b, between them. As seen in Figure ?, the algorithm evaluates the quality of the line that separates two classes by maximizing the width of the margin given the contraints surrounding the line. Some problems are not linearly separable, and thus a penalty term is introduced to allow for misclassifications. As shown in Figure ?, the algorithm then simultaneously minimizes the misclassifications while maximizing the margin.

This can be extended easily to multidimensional analysis via what is called the *kernel trick*. First, using a nonlinear kernel function maps the data into higher dimensional feature space. Then the algorithm can find

a linear separation in this space, as shown in Figure ?. Further, this can be upgraded from classification to SVR by doing similar math but instead minimizing the margin, as shown in Figure ?.

The kernel chosen for this study is the Gaussian radial basis function, shown below. This has two tuneable parameters, gamma and C. Gamma influences the width of influence of individual training instances, and strongly affects the fitting of the model. Low values correspond to underfitting because the instances have too large of a radius (low influence) and high values correspond to overfitting because the instances have a small radius (high influence).

The C parameter also affects the fitting of the model by allowing more or less support vectors, corresponding to more or less misclassification. A lower C smooths the surface of the model by allowing more misclassifications, whereas a higher C classifies more training examples by allowing fewer misclassifications. Too low and too high of a C can cause under- and overfitting, respoectively.

Since there is a tradeoff of fitting strength provided by both parameters, it is common to run the algorithm on a logarithmic grid from 10'-3 to 10'3 for each parameter. If plotted on a heatmap of accuracies given gamma and C, there will be a diagonal of ideal combinations that emerges. The lowest of these is usually chosen.

2.2.2 Model Selection and Assessment

After a model is trained, the first step is model selection and assessment. Selection is estimating model performance among a set of trained models using a single validation set. After one model is chosen, assessment takes place by determining the prediction capability on new data via a previously unseen testing set. Both selection and assessment can be done in a single step using k-fold cross-validation, which is described below.

2.2.2.1 Sources of Error

In statistical learning, there are two sources of error that need to be simultaneously minimized: bias and variance. Bias is caused by simplifications in the model, so the error is caused by missed relationships in the data; an underfit model is due to high bias. Variance is caused by including random noise in the model, so the error is caused by oversensitivity to that noise; an overfit model is due to high variance.

Include math or just reference it? Talk about irreducible error too?

As shown in Figure 2.3, there is a minimum in the total error, showing that there is a tradeoff between the bias and variance. Some bias is desired in order to generalize to future unknown data. But some variance is also positive for the model because it captures the relationships in the data that the bias counteracts.

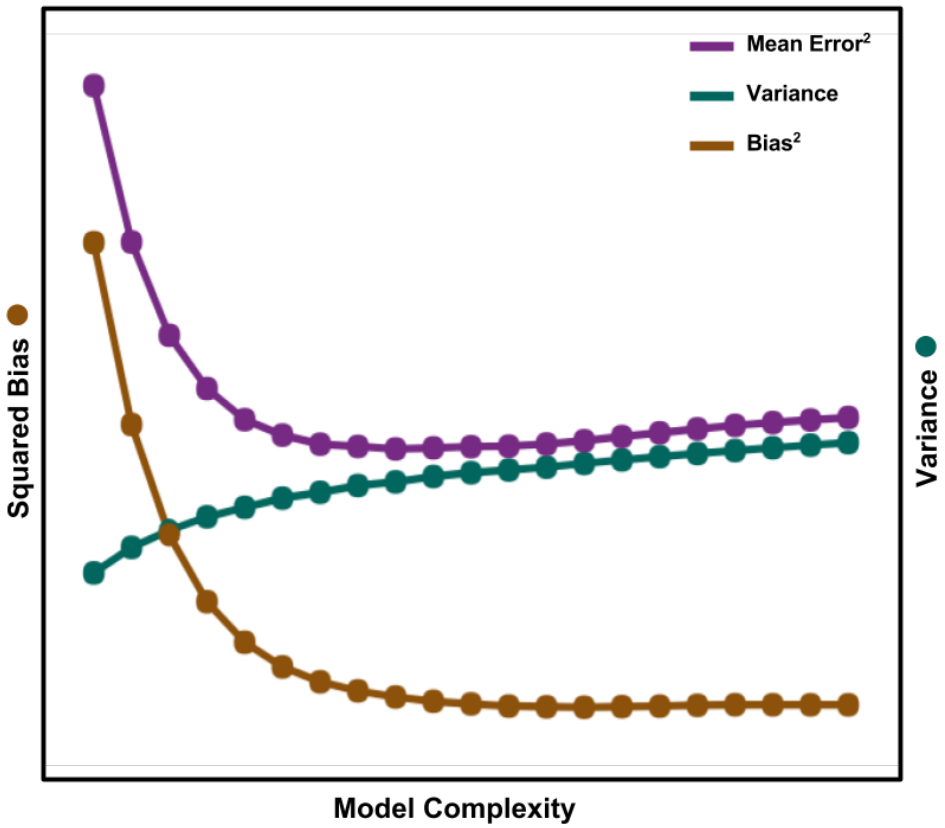


Figure 2.3: Total prediction error comprised of bias and variance

2.2.2.2 Types of Error

While the sources of the model prediction error are well known, the creation of a statistically learned model is a hidden process. Although the model emerges from a black box, there are ways to evaluate the generalization (i.e., prediction) capability of it. This is done by removing a small portion of the data for use as a testing set. The rest of the data set is known as the training set and is used to train a model. After training, the test set is used

to test the model.

The generalization error is typically referred to as the *testing error*, as it is measuring the ability of the model to predict future cases that were not introduced in the training phase (i.e., the testing set entries). Next, the *training error* is provided by comparing the model predictions to the training set, as the model would likely be smoother than the potential noise the training set would include. This is useful to determine the fitness of the model, the application of which is discussed below in Section 2.2.3.

Although one could just train and test their model, there is a way to test the model while still in the training phase. A testing set that would be used during training to give feedback, a *cross-validation* set, can provide a faster convergence to a satisfactory model. As shown in Figure 2.4, this can be done by splitting the data set into three groups: a large training set, a small cross-validation set, and a small testing set.

However, in practice, multiple rounds of cross-validation steps are used, referred to as *k-fold cross-validation*. This allows a user to use all data entries as a testing entry once. As illustrated in Figure 2.4, this splits the dataset into k subsets. One set is designated as the testing set, and a model is trained with the rest. Following the first training phase, another begins, this time with a different subset as the testing set. This process is performed k times to give k models, and the models are then averaged, providing an additional level of model validation than can be achieved with a single testing set.

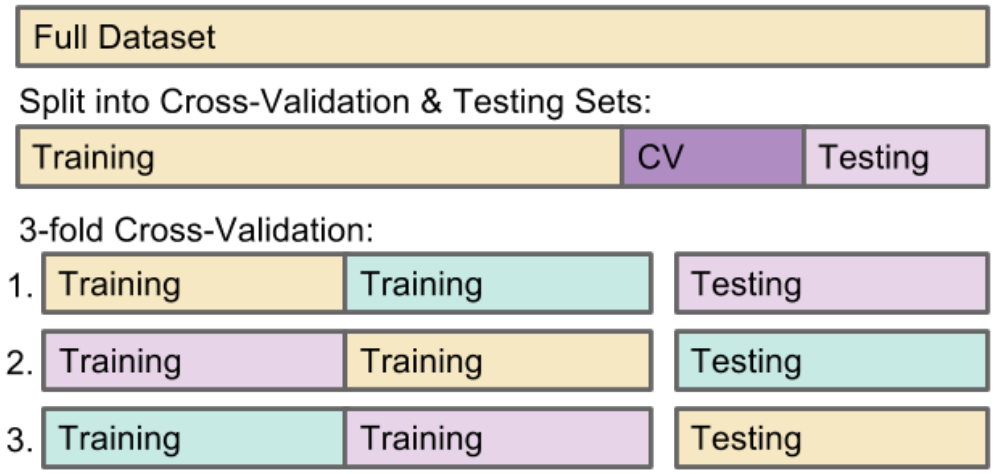


Figure 2.4: Illustration of how a dataset can be split up for model evaluation

2.2.3 Model Optimization and Validation

It is unlikely to have a model perform as one expects the first time. There are therefore a few techniques for optimizing the performance. It should be noted that much of the discussion here and in Section 3.3 focuses on the diagnostics aspect rather than the validation aspect of these techniques. In practice, these are used for both purposes, but in this work the formal comparison of model performance will be used, introduced and demonstrated in Sections 2.2.3.3 and 3.3.2, respectively.

However, the increase in performance from over-optimization could be linked to the training set performance and might not generalize outside of the specific type of input data used. A workaround for this scenario is to obtain more data for the set or to obtain a completely different data set

altogether.

2.2.3.1 Training Set Size

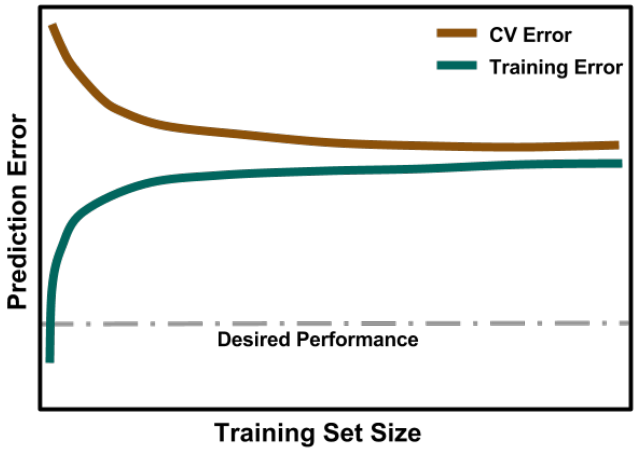
The first diagnostic plot for optimizing the model performance is called a *learning curve*, which provides information about the bias-variance tradeoff with respect to the data set size. More specifically, learning curves compare the training and cross-validation errors to the size of the training set (i.e., number of instances in the training set). This is done by randomly selecting a percentage of the the training set, inputting that into a statistical learner, and tabulating the error of the learned model.

Typically, a learning curve will look somewhat like one of the three examples in Figure 2.5. A learning curve tests the model for high bias or high variance, which can correspond to an undefit or overfit model, respectively.

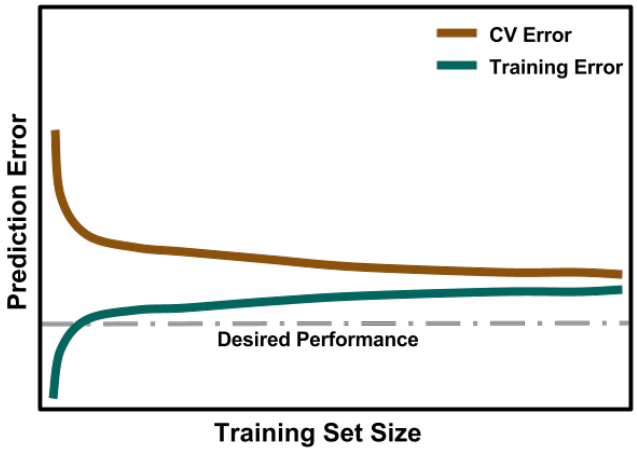
add reference

Figure 2.5a suggests underfitting because the model is missing important features in the data. It is characterized by a small gap between the curves but high overall errors. The cross-validation error remains consistently high and the training error increases drastically with increasing data, since it is not generalizing well.

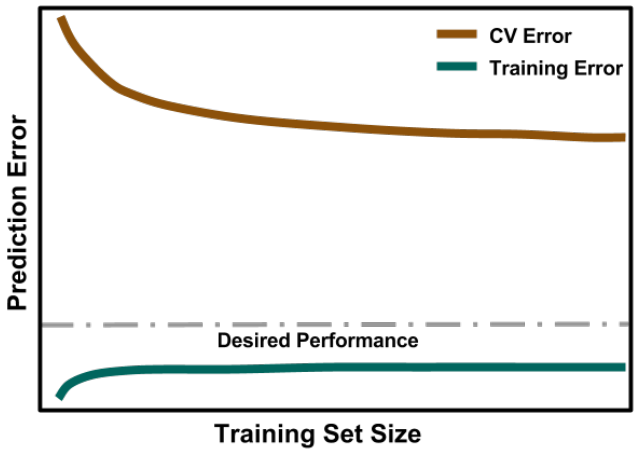
Figure 2.5c suggests overfitting because the model has too much sensitivity to variations in the data. It is characterized by a very large gap between the curves. It has an extremely low training error, as it has taken



(a) High bias



(b) Ideal



(c) High variance

Figure 2.5: Learning curves for three training scenarios

into account every detail of the training set, but a high cross-validation error because it cannot generalize beyond the testing set.

Figure 2.5b is an example of a more ideal model fit. It is characterized by a small gap between the two errors, and they are at a reasonable level with respect to the desired performance. The training error should increase with respect to the training set size due to a larger amount of bias (preventing overfitting). But the cross-validation error should decrease quickly with respect to the training set size due to being close to the minimum of the bias-variance tradeoff.

2.2.3.2 Model Complexity

After ensuring the appropriate training set size is selected, the models must be further optimized using *validation curves*. These provide information on the bias-variance tradeoff with respect to model complexity. Two main factors affecting model complexity can cause the model to be under- or overfit to the data: number of features in the data set and algorithm parameters that vary the regularization.

Regularization is a component of many machine learning aglorithms a

describe figure

In practice, plotting learning and validation curves can be iterative. But as previously mentioned, too many optimizations will result in a poorly performing model when exposed to data outside of the training set.

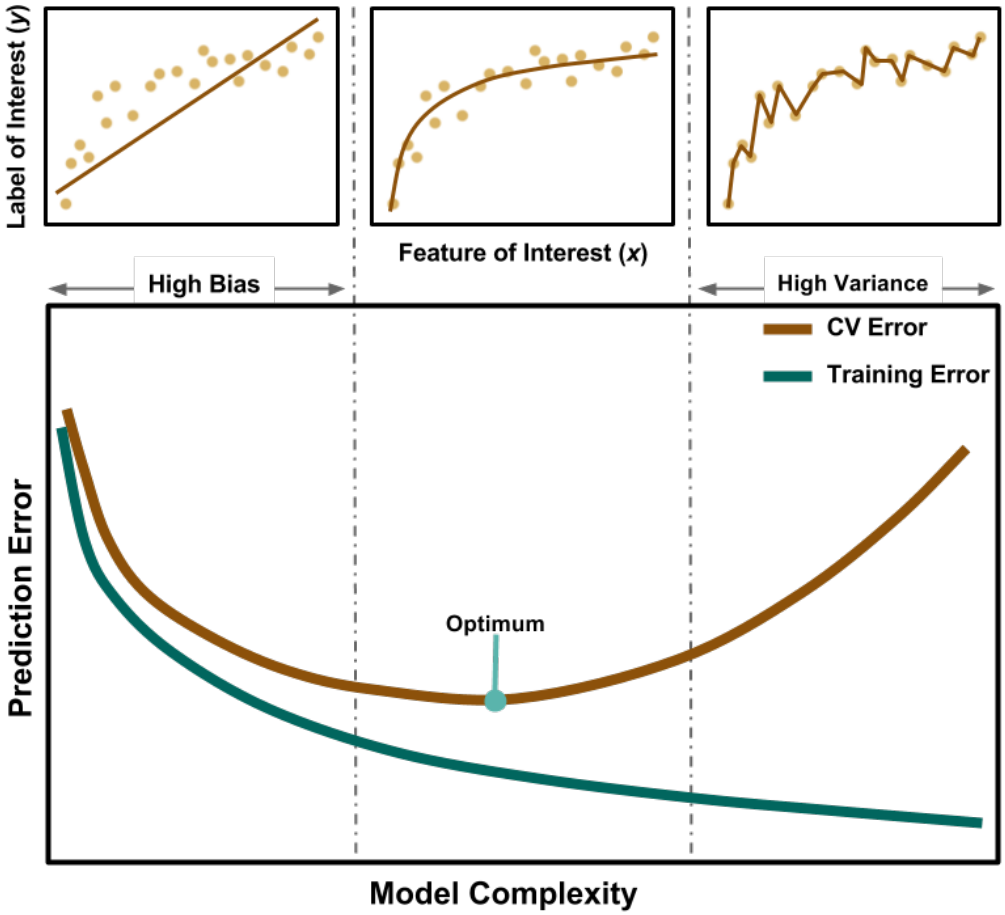


Figure 2.6: Validation curve showing examples of different fittings

2.2.3.3 Comparison of Methods

Inverse stuff.

Priors given by set of forward problems (input data space - ORIGEN sims)

Likelihood given by model space, e.g.:

Model params as determined from ML alg

(related) MLE from bayesian inference

Direct inversion of matrix A

Expert-elicited params (i.e. initial guesses for INDEPTH? Or optimized ones?)

Marginal likelihood only needed for absolute measures - doesnâŁ™t affect relative probabilities

2.3 Computational Methods

I'm a section on the computational methods!

2.3.1 Fuel Cycle Simulation

Short blurb on FC sim (capabilities enable other materials to be studied beyond SNF)

2.3.2 Data Modification

Perfect data -> real world data

2.3.2.1 Detector Response Functions

GADRAS?

Discuss methods, and potential options. Am I implementing for prelim?

2.3.2.2 Isotope Identification from Gamma Spectra

Same as above, but prob won't implement for prelim.

2.4 Applications of Statistical Methods to Nuclear Forensics Analysis

I'm the lit review section on previous applied ML work in the NF field.

It is first important to determine if statistical methods can overcome the inherent database deficiencies. After that, the statistical methods must be considered in such a way as to represent a real-world scenario. Although mass spectrometry techniques provide extremely accurate isotopic information for analytical methods, they are time-consuming and more expensive. And although gamma spectroscopy can give extremely fast results cheaply, it only measures certain radiological signals and is influenced by many environmental factors, storage, and self-attenuation. As different machine learning algorithms and parameters are investigated, this work focuses on probing the amount of information required to obtain realistic results.

2.4.1 Special Nuclear Materials Studied

The review on nf for the whole fuel cycle is useful here, perhaps. This is also important when I discuss my risk management section later.

2.4.2 Statistical Methods Employed

Very short details from lit review outline and success rates should be discussed here

3 METHODOLOGY AND DEMONSTRATION

This chapter first covers the methodology of the proposed work by introducing each experimental component and a demonstration of each component. This has been split into three sections, summarized below.

Section 3.1 discusses how the training data is obtained. After the initial training data is simulated in Section 3.1.1, with a possible information reduction step in Section 3.1.2, it will be input to a statistical learner.

Section 3.2 is about algorithms that use the features and labels of the training data to statistically formulate a model. Algorithm choice and parameters are discussed in Section 3.2.1. Next, the main goal for these machine-learned models is to supply reactor parameters associated with some unknown SNF. Section 3.2.2 shows the results of testing this goal: the prediction of a new instance that has only features and no label.

Finally, the algorithms are evaluated for accuracy and validated, as shown in Section 3.3. To both understand the performance and validate the models, the results are then evaluated for over- or under-fitting, which is in Section 3.3.1. But validation is more than just making sure the models are properly fit to the data. Perhaps the training set was not representative of the actual data space, whereas other methods do not rely on the data space for results. So, lastly, comparison against other algorithms as well as other methods is described in Section 3.3.2.

This work incorporates some methods and suggestions from previous

work on the subject [2] regarding machine learning model performance with respect to information reduction. This is to establish some baseline expectations of reactor parameter prediction and how the different algorithms perform.

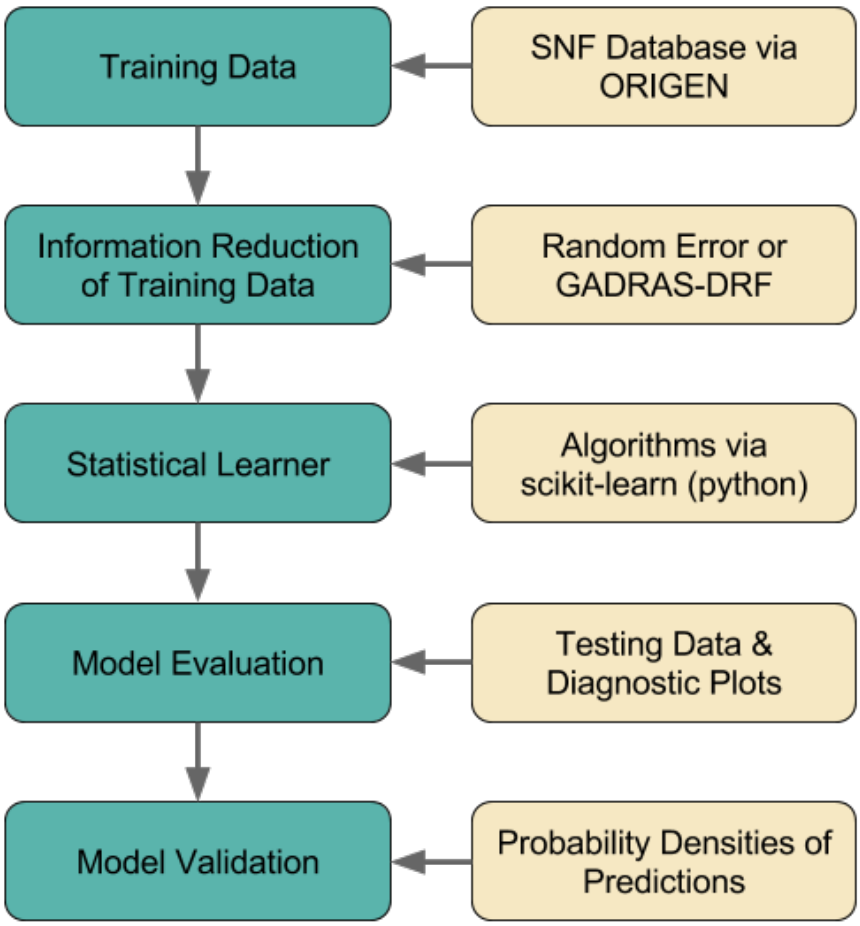


Figure 3.1: Methodology of the proposed experiment.

Next, this work will expand upon the previous work in two ways. The

first is adding a different information reduction technique via applying a gamma spectroscopy detector response function (DRF) to the SNF nuclide recipes, which can calculate various spectra based on the types of gamma detectors available to the forensics community. Secondly, a more advanced machine learning algorithm, support vector regression, is included so as to compare more complex models against simpler models. A schematic of the workflow involving the experimental components is shown in Figure 3.1.

3.1 Training Data

3.1.1 Spent Nuclear Fuel Simulations

Because creating databases from real measurements to represent reactor technologies from around the world is impossible, the database in this study will be created from high-fidelity simulations via ORIGEN [16], an activation and depletion code within the SCALE 6.2 modeling and simulation suite [14]. Specifically, the ARP module of the activation and depletion code ORIGEN was used: ORIGEN-Automatic Rapid Processing (ORIGEN-ARP).

A set of simulations of SNF at different burnups and cooling times will comprise the database. Of interest to an entity trying to create a weapon is partially irradiated fuel if they have plutonium separations capabilities or any radioactive substance in the case of a dirty bomb. Addressing the former, a smaller burnup than is typical for spent fuel from a commercial reactor is used in the previous work.

Fuel	Reactor	Enrichment
CE14x14	PWR	2.8
CE16x16	PWR	2.8
W14x14	PWR	2.8
W15x15	PWR	2.8
W17x17	PWR	2.8
S14x14	PWR	2.8
VVER440	PWR	3.60
VVER440_3.82	PWR	3.82
VVER440_4.25	PWR	4.25
VVER440_4.38	PWR	4.38
VVER1000	PWR	2.8
GE7x7-0	BWR	2.9
GE8x8-1	BWR	2.9
GE9x9-2	BWR	2.9
GE10x10-8	BWR	2.9
Abb8x8-1	BWR	2.9
Atrium9x9-9	BWR	2.9
SVEA64-1	BWR	2.9
SVEA100	BWR	2.9
CANDU28	PHWR	0.711
CANDU37	PHWR	0.711

(a) Reactor types and uranium-235 enrichment [weight%].

	PWR	BWR	PHWR
Power Density [MW/MTU]	32	23	22
Burnup [MWd/MTU]	600–17700	600–12300	600–12300
Cooling Time		{1m, 7d, 30d, 1y}	

(b) Simulation space defining reactor parameters and cooling time.

Table 3.1: Design of the training set space.

It should be noted that many algorithms are developed on an assumption that the training set will be independent and identically distributed (i.i.d.). This is important so that the model does not overvalue or overfit a certain area in the training space. A truly i.i.d. training set would go beyond the lower burnups, but this is purely for demonstration with a single use case in mind. The training database is thus constructed by simulating the same training set space as described in Ref. [2], shown in Table 3.1. For each entry shown here the simulations included

While in most machine learning studies the testing set is chosen randomly from the training set, the previous work used an external one, shown in Table 3.2. Although the test set was designed to have values in between the trained values of burnup, it was chosen systematically. Therefore it was implemented in this study for comparison, but cross-validation will be used moving forward. More specifically, using k-fold cross-validation is expected to better indicate the model performance.

Fuel	Reactor	Enrichment	Cooling Time	Burnup
CANDU28	PHWR	0.711	{1m, 7d, 30d, 1y}	{1400, 5000, 11000}
CANDU28	PHWR	0.711	{3m, 9d, 2y}	{5000, 6120}
CE16x16	PWR	2.8	{1m, 7d, 30d, 1y}	{1700, 8700, 17000}
CE16x16	PWR	2.8	{3m, 9d, 2y}	{8700, 9150}
CE16x16	PWR	3.1	{7d, 9d}	{8700, 9150}
GE7x7-0	BWR	2.9	{1m, 7d, 30d, 1y}	{2000, 7200, 10800}
GE7x7-0	BWR	2.9	{3m, 9d, 2y}	{7200, 8800}
GE7x7-0	BWR	3.2	{7d, 9d}	{7200, 8800}

Table 3.2: Design of the testing set space.

3.1.2 Information Reduction

Since the overall goal of this project is to determine how much information to what quality is needed to train a machine-learned model, there will be an information reduction manipulation applied to the training data set. This study evaluates the impact of randomly introduced error of varying amounts on the ability of the algorithms to correctly predict the burnup.

The three algorithms will be evaluated with error applied to each nuclide vectors in the training set. A maximum error is ranging from $0 - 10\%$ is chosen for each round of training, and a random error within the range of $[1 - E_{max}, 1 + E_{max}]$ is applied to each component of the nuclide vector.

However, since error in a nuclide vector is not random, in fact it is systematic and dependent on a number of known sources of uncertainty, the next study will introduce error by limiting the nuclides to only those that can be measured with a gamma spectrometer. This will use the code GADRAS-Detector Response Function (GADRAS-DRF) [5] to computationally generate gamma spectra from the nuclide vectors, and is the next step in the future work. This is discussed in more detail in Section 4.1.

3.2 Statistical Learning for Models

3.2.1 Algorithms Chosen

Choosing which algorithms to test is usually based on what is being predicted and intuition regarding strengths and weaknesses of different optimization methods.

Algorithm	Parameter	Value
Nearest Neighbor Regression	*n*-neighbors	1
	Weights	uniform
	Distance Metric	L2: Euclidian Distance
Ridge Regression	Regularization, α	1.0
	Normalization	False
	Stopping Tolerance	0.001
Support Vector Regression	Kernel	Radial Basis Function
	Gamma, γ	0.1
	C	1.0
	Epsilon, ε	0.1
	Stopping Tolerance	0.001

Table 3.3: Algorithm paramters used for initial model evaluation

For a benchmarking excercise, some machine learning approaches here were chosen based on previous work [2]: nearest neighbor and ridge regression. These are useful because they are simple, providing a dissimilarity-based model and a linear regression-based model, respectively. If more complex

algorithms are not required to obtain useful results, then there is no need to use more computationally expensive options. However, hedging on the fact that more complex models will be needed, this work also employs an algorithm that is known to handle highly dimensional data sets well: support vector regression. These algorithms were introduced in Section 2.2.1.

A python-based machine learning toolkit, scikit-learn [15], is used to train the models. The default parameters used for the algorithms are in Table 3.3.

3.2.2 Reactor Parameter Prediction

The prediction of reactor parameters here is done with the burnup of the SNF to provide algorithm and corresponding model generalizability. Following the training phase of the models, next it is important to estimate the reactor parameter predication capabilities of those models. This is done with a set of measurements from a test data set with samples that mimic interdicted SNF. The testing set has the same features as the training set, with labels that are compared to the predicted labels. The results of each model's prediction errors are shown in Table 3.4.

First shown in Table 3.4 are testing set errors and cross-validation errors, because although previous work uses the former, it is expected that the latter will provide better estimates. The models evaluated by the testing set do not have a validation set for pre-evaluation. The models that are evaluated via cross-validation do not use the testing set.

Algorithm	Error Origin	MAPE	RMSE
Nearest Neighbor Regression	Testing Set	78.24	3479.0
	5-fold Cross-Validation	127.84	4401.4
Ridge Regression	Testing Set	0.44	44.96
	5-fold Cross-Validation	0.05	3.24
Support Vector Regression	Testing Set	4.32	428.52
	5-fold Cross-Validation	0.49	163.61

Table 3.4: Model burnup prediction errors for three algorithms

Next there are two error types. For the sake of comparison to previous work and convenient interpretation, mean absolute percentage error (MAPE) is tracked. However, MAPE requires that no true values are 0. The preferred method in the community is to use root-mean-squared error (RMSE) for model error estimation, so both are tabulated. The MAPE shows that there are some extremely high and extremely low errors depending on the algorithm, both of which are quite concerning. Thus, next introduced are some diagnostic and optimization procedures that can shed light on the errors shown here.

3.3 Validation

To obtain reliable models, one must both choose or create a training set carefully and study the impact of various algorithm parameters on the error. Although the title of this section suggests final steps of confirming a model's

usefulness for predictions, what follows is more of a troubleshooting exercise. In practice, these analyses are used for both purposes.

3.3.1 Model Diagnostics

Machine learning algorithms are heavily dependent on the inputs and parameters given to them, such as training set sizes, regularization, learning rates, etc. From the results shown in Section 3.2, it is clear there is room for improvement. Diagnostic plots show the errors of the predicted burnup values to the actual burnup values with respect to some variable on the x-axis. As previously introduced in Section 2.2.3, the errors are compared to the training error to understand the generalization strength with respect to training set size (learning curves) and the algorithm parameters governing model complexity (validation curves).

In addition to machine learning best practices, another layer of comparison is added here. Because it is difficult to ensure consistently representative testing data, the accuracy of a learned model should not depend on only one testing set. The learned model's accuracy is better estimated by using a validation set, or even better, k-fold cross-validation, introduced in Section 2.2.2. This work includes both the testing error (using the testing set described in Section 3.1) and cross-validation error. The predetermined testing set will allow for comparison against the previous work it was obtained from [2], but it is assumed that cross-validation will provide a better indication of model performance.

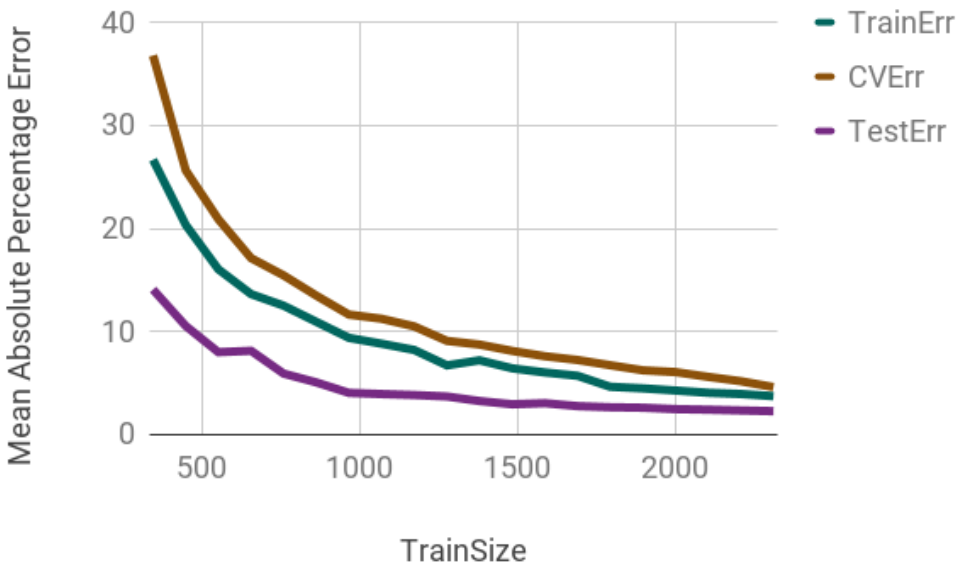


Figure 3.2: Learning curve for burnup prediction, $\gamma = 0.001$

The learning curves are obtained as follows, shown in Figure 3.2. For a given (randomly chosen) training set size between 15 and 100% of the total data set, training and prediction rounds were performed for each. The testing error scenario performs this k times and averages those results. This is equivalent to the k in k-fold cross-validation to provide some semblance of equivalent statistics. The cross-validation error scenario has no need for averaging because it is performed automatically. In both cases, the learning curves do not provide a clear picture of over- or undertraining upon first glance.

The validation curves are obtained as follows, shown in Figure 3.3. The γ parameter in support vector regression (SVR), which influences model

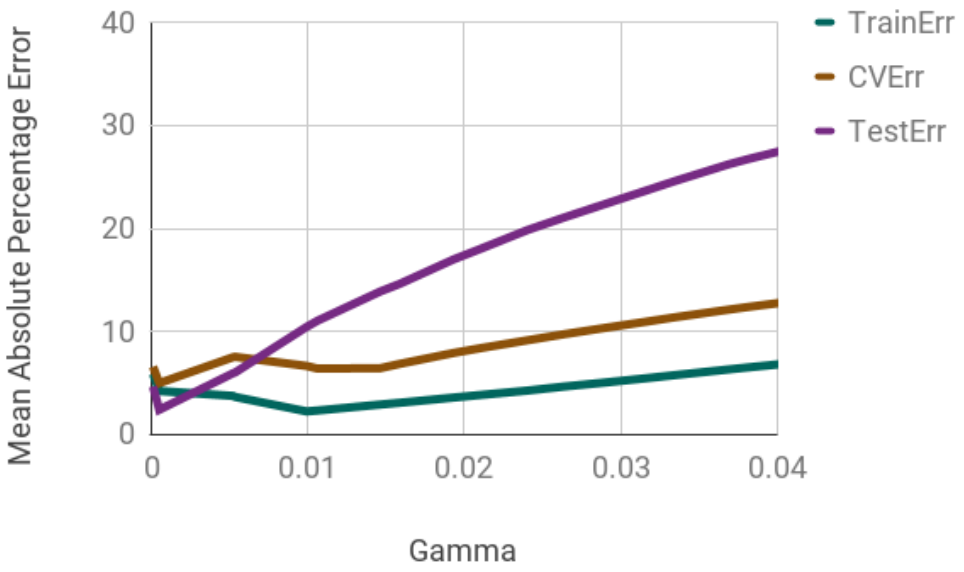


Figure 3.3: Validation curve for burnup prediction, $TrainSize = 2313$

complexity, was varied from 10^{-4} to 10^{-1}. Training and prediction rounds were performed for different γ values in this range. Again, the testing and cross-validation errors are both used as described above. As with Figure 3.2, determining the robustness to over- or undertraining is difficult here although there is possibly a minimum at $\gamma = 0.001$.

Although there is no example behavior of Figure 3.2's peculiar learning curve in Figure 2.5, the curve mimics the squared bias curve from Figure 2.3. This indicates that the bias in the model is much higher than the variance. Next, the testing error is lower than the training error; this should never be the case, and indicates an issue with the systematically chosen testing set. While the cross-validation error is correctly higher than the training

error, it follows along in parallel, producing no information on model fitness other than confirming a very high bias. It is presumed this is not the fault of the algorithms, but the training set itself. It is likely covering too small of a range of the simulation space.

Additionally, Figure 3.3's validation curve shows the testing error dropping below the training error for extremely small γ, around where a minimum might be. Since the model suffers from high bias, no amount of model complexity can be optimized. The resolution to the underfitting is discussed in Section 4.1.

3.3.2 Model Comparison

In addition to evaluating a single learned model, it may be beneficial to compare models. Options for comparison of algorithms: inverse bayesian stuff, Scatter plots, Pairwise t-tests. Confidence intervals on predictions to understand true error versus sample error Test set must be > 30 instances, Can easily calculate $N\%$ confidence interval.

4 RESEARCH PROPOSAL

This document previously demonstrated the performance of machine learning on a set of nuclear material isotopics to calculate a reactor parameter of interest: burnup. Additionally, there have been various methods discussed for understanding the learned model's behavior and thus the quality of the results. Moving forward to a set of experiments is now possible.

Before describing the experiments, some topics and issues are addressed in Section 4.1. Finally, the proposed research experimental design is presented Sections 4.2, 4.3, and 4.4.

4.1 Experiment Preparations

Expanding Training Set

As identified in Section 3.3, the testing set used for the demonstration was not suitable for further study without being expanded. Many algorithms are developed on an assumption that the training set will be i.i.d.. This is important so that the model does not overvalue or overfit a certain area in the training space. The next step is to provide a larger, more diverse training set to the algorithms so they can better predict when faced with new instances. This diversity will be suggested from the spent fuel isotopic composition (SFCOMPO) database [9], as it includes many common domestic and international reactors.

The SFCOMPO-2.0 relational database [9] has approximately 750 SNF measurements from 44 reactors. While this is not sufficient as a training set, it provides a better framework for simulating a larger training set using ORIGEN. After cross-validation, diagnostics, and optimization, the trained models can be tested against the entries in this database to provide a clear estimate of of the model performance.

Finalizing Set of Algorithms

The three algorithms in the demonstration (nearest neighbor, ridge, and support vector regression) are not necessarily the set that will be evaluated for the experiments. After these are used to train new models on a larger training set for comparison, other algorithms can be speedily assessed as well. Since support vector-, distance-, and linear-based models are already represented, other obvious choices include Bayesian methods, decision trees, neural networks, or ensemble methods.

mention in litrev or exclude....or cite book chapter

Computational Framework and Resources

Thus far, all simulations and training have not required more processing than available on a personal laptop. However, some algorithms do require larger amounts of computational time (e.g., artifical neural networks). If necessary, the training stage can be done using the Center for High Throughput Computing (CHTC), which is available to University of Wisconsin

researchers.

4.2 Experiment 1

Viability of Statistical Learning on Direct Isotopics

The first experiment will be a purposefully constructed version of the demonstration: evaluating the model performance with known isotopics. This sheds light on how this methodology will perform on the simplest scenario, providing an estimate on the maximum level of performance. *The main purpose of this experiment is to iteratively probe the usefulness of statistical methods for determining reactor parameters.*

Figure 4.1 shows what can be used for training and testing (or prediction). The two horizontal boxes show the physical and computational forms of what these experiments are simulating, respectively. The lab-measured mass spectra correspond to the perfect information being referred to here. In the computational context, these measurements instead come from simulations. Mass spectra results are thus approximated as the direct isotopics given from the simulations, since mass spectrometry provides highly reliable and accurate information.

The variables for this experiment will include the complexity of the machine learning algorithm used, feature reduction (e.g., different subsets of isotopes: top n, fission products), and different subsets of the desicion space (e.g., simplifying the regression task by fixing the reactor type). It is

expected that a more complex algorithm (e.g., SVR) will be needed, and that preprocessing and/or manual feature reduction will assist in creating higher quality models. Simplifying the decision space should always improve prediction, but it is not obvious how much it will be needed for burnup prediction specifically.

It is possible that statistical models trained on direct isotopic information do not perform well enough. Other than attempting different types of algorithms, it is possible to preprocess the data, statistically performing feature reduction via principal components analysis (PCA). If this is not sufficient, it is possible SNF is has too many or too few correlated features to provide reliable models across the space of current reactor technologies. Since separated plutonium and UOC have been also studied using these techniques, it is possible these materials can provide useable learned models. Additionally, this methodology would also work if applied to post-detonation materials. There is work on creating standard materials to represent the "urban canyon", so this is another subject that could benefit from statistical correlations. cite

4.3 Experiment 2

Viability of Statistical Learning on Gamma Spectra

The second experiment will be the previously discussed extension of the demonstration by applying detector response functions to the SNF isotopics:

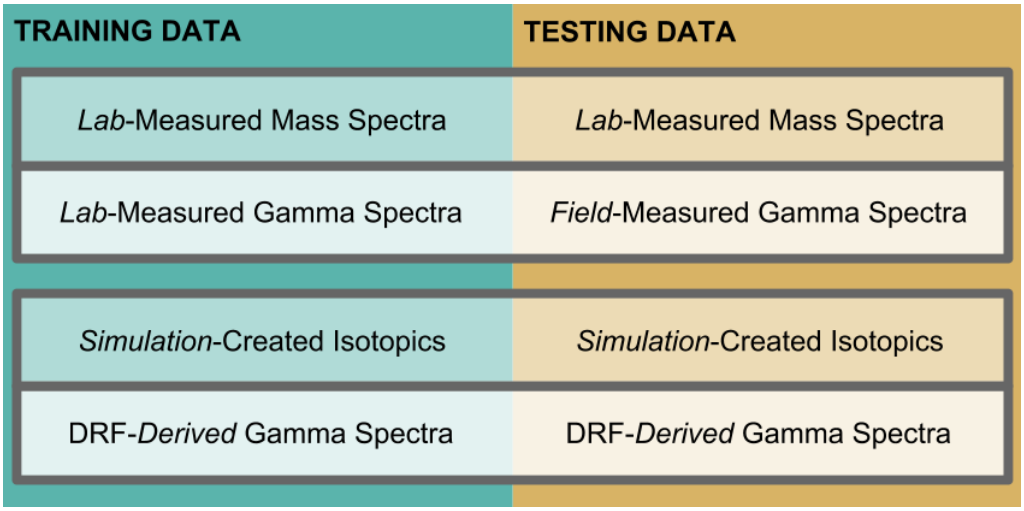


Figure 4.1: Physical and Computational Comparisons for Experiments 1 and 2

evaluating the model performance with reduced isotopic information. This demonstrates the usefulness of this methodology in a real-world scenario where exact isotopics are not always known. *The main objective of this experiment is to measure the reduction in statistical model parameter prediction reliability as the quality of the training information is reduced.*

The two bottom portions of the boxes in Figure 4.1 represent a more realistic measurement scheme, involving a model trained from gamma spectrometers rather than the lengthy process of performing mass spectrometry on the samples. In the physical context, the measurements for training would be done using a semiconductor gamma detector, but the testing or prediction step may be done outside of the lab on a different detector. This will be captured by applying different detector response functions to the radionuclide inventories from the simulations.

The variables for this experiment will include the complexity of the machine learning algorithm used and quality of the training data set. Feature reduction is implicit here, since gamma detection only includes radionuclides within the SNF isotopics. The indirect isotopic training data are likely going to reduce the prediction capability of the models, but it is not yet clear if a response function simulating a hand-held NaI gamma detector can provide any useful predictions. And while it is still expected that the complex algorithms will perform better, it is not obvious if different algorithms than the ones used in Experiment 1 will be needed .

It is possible that statistical models trained on indirect isotopic information do not perform well enough. Again, here, different algorithms may perform better than others due to the underlying optimization processes. Further feature reduction could also prove useful, focusing on particular energy regions or particular peaks throughout the spectrum. The quality of the isotopic information can be improved slightly by using an isotope identification algorithm; this may improve the performance, as they are developed to automatically report isotopics from gamma spectra. If this still is not sufficient, it may be that only direct isotopic information (i.e., that obtained from mass spectrometry) is required for reliable statistical models of SNF. Although preprocessing could also be investigated here, the materials discussed above may also be more disposed to defined statistical correlations.

4.4 Experiment 3

Viability of Statistical Learning on Other Fuel Cycle Flows

Pending success

REFERENCES

[1] Broadhead, Bryan L, and Charles F Weber. 2010. Validation of inverse methods applied to forensic analysis of spent fuel. In *Proceedings of the Institute of Nuclear Materials Management 51st Annual Meeting*. Baltimore, MD, USA. https://www.osti.gov/scitech/biblio/1001291.

[2] Dayman, Kenneth, and Steven Biegalski. 2013. Feasibility of fuel cycle characterization using multiple nuclide signatures. *Journal of Radioanalytical and Nuclear Chemistry* 296:195–201. http://link.springer.com/article/10.1007%2Fs10967-012-1987-4.

[3] Gey, Frederic, Chloe Reynolds, Ray Larson, and Electra Sutton. 2012. Nuclear forensics: A scientific search problem. In *Proceedings of the Lernen, Wissen, Adaption (Learning, Knowledge, Adaptation) Conference*. Dortmund, Germany. http://metadata.berkeley.edu/nuclear-forensics/Paper_9-12-12_lwa-2012-nuclear-forensics-scientific-search-problem_v7.pdf.

[4] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2001. *The Elements of Statistical Learning*. Springer Series in Statistics, New York, NY, USA: Springer New York Inc.

[5] Horne, Steven M., Gregory G Thoreson, Lisa A. Theisen, Dean J. Mitchell, Lee Harding, and Wendy A. Amai. 2014. Gamma De-

tector Response and Analysis Software - Detector Response Function (GADRAS-DRF). User's Manual, Sandia National Laboratories, Albuquerque, New Mexico, USA. Version 18.5; SAND2014-19465, http://www.osti.gov/scitech/servlets/purl/1166695.

[6] Jones, Andrew, Phillip Turner, Colin Zimmerman, and J.Y. Goulermas. 2014. Machine learning for classification and visualisation of radioactive substances for nuclear forensics. In *Techniques and Methods for Safeguards, Nonproliferation and Arms Control Verification Workshop*. Portland, Oregon. https://www.researchgate.net/publication/264352908_Machine_Learning_for_Classification_and_Visualisation_of_Radioactive_Substances_for_Nuclear_Forensics.

[7] Jones, Andrew E., Phillip Turner, Colin Zimmerman, and John Y. Goulermas. 2014. Classification of spent reactor fuel for nuclear forensics. *Analytical Chemistry* 86:5399–5405. http://pubs.acs.org/doi/ipdf/10.1021/ac5004757.

[8] May, Michael, Reza Abedin-Zadeh, Donald Barr, Albert Carnesale, Philip E. Coyle, Jay Davis, William Dorland, William Dunlop, Steve Fetter, Alexander Glaser, Ian D. Hutcheon, Francis Slakey, and Benn Tannenbaum. 2007. Nuclear Forensics: Role, State of the Art, and Program Needs. Tech. Rep., Joint Working Group of the American Physical Society and the American Association

for the Advancement of Science. https://www.aaas.org/report/nuclear-forensics-role-state-art-program-needs.

[9] Michel-Sendis, Franco, Jesus Martinez-GonzÃ¡lez, and Ian Gauld. 2017. Sfcompo 2.0 âŁ“ a relational database of spent fuel isotopic measurements, reactor operational histories, and design data 146:06015. www.oecd-nea.org/sfcompo/.

[10] Moody, K.J., P.M. Grant, and I.D. Hutcheon. 2005. *Nuclear Forensic Analysis*. 1st ed. Boca Raton, Florida, USA: CRC Press. https://books.google.com/books?id=Q9mgDnWoPLYC.

[11] Nicolaou, G. 2006. Determination of the origin of unknown irradiated nuclear fuel. *Journal of Environmental Radioactivity* 86:313–318. http://nuclear.ee.duth.gr/upload/A13%20%20%20identification.pdf.

[12] ———. 2009. Identification of unknown irradiated nuclear fuel through its fission product content. *Journal of Radioanalytical and Nuclear Chemistry* 279(2):503–508. http://link.springer.com./article/10.1007%2Fs10967-007-7300-x.

[13] ———. 2014. Discrimination of spent nuclear fuels in nuclear forensics through isotopic fingerprinting. *Annals of Nuclear Energy* 72:130–133. Technical Note, http://www.sciencedirect.com.ezproxy.library.wisc.edu/science/article/pii/S0306454914002308.

[14] Oak Ridge National Laboratory. 2016. SCALE: A Comprehensive Modeling and Simulation Suite for Nuclear Safety Analysis and Design. Code Suite, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA. Version 6.2.1, ORNL/TM-2005/39, Available from Radiation Safety Information Computational Center as CCC-834, `http://scale.ornl.gov`.

[15] Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830. `http://scikit-learn.org/stable/index.html`.

[16] Rearden, B.T., and M.A. Jessee. 2016. Ch. 5 Depletion, Activation, and Spent Fuel Source Terms. In *SCALE Code System: User Documentation*, 5–1–5–263. Oak Ridge, Tennessee, USA: Oak Ridge National Laboratory. Version 6.2.1; ORNL/TM-2005/39, `https://www.ornl.gov/sites/default/files/SCALE%20Code%20System.pdf`.

[17] Robel, Martin, Michael J. Kristo, and Martin A. Heller. 2009. Nuclear forensic inferences using iterative multidimensional statistics. In *Proceedings of the Institute of Nuclear Materials Management 50th Annual Meeting*. Tuscon, AZ, USA: Institute of Nuclear Materials Man-

agement. LLNL-CONF-414001, https://e-reports-ext.llnl.gov/pdf/374432.pdf.

[18] Tarantola, Albert. 2005. *Inverse Problem Theory and Methods for Model Parameter Estimation*, chap. 1. The General Discrete Inverse Problem, 1–40. Philadelphia, Pennsylvania, USA: Society for Industrial and Applied Mathematics. http://epubs.siam.org/doi/pdf/10.1137/1.9780898717921.ch1.

[19] Weber, Charles F, Vladimir A Protopopescu, Michael H Ehinger, Alexander A Solodov, and Catherine E Romano. 2011. Inverse solutions in spectroscopic analysis with applications to problems in global safeguards. In *Proceedings of the Institute of Nuclear Materials Management 52nd Annual Meeting*. Palm Desert, CA, USA. https://www.osti.gov/scitech/biblio/1031530.

[20] Weber, Chuck F, and Bryan L Broadhead. 2006. Inverse depletion/decay analysis using the scale code system. In *Transactions of the American Nuclear Society Winter Meeting*, vol. 95, 248–249. Albuquerque, NM, USA. Track 4: Nuclear and Criticality Safety Technologies.