Chemometrics and Intelligent Laboratory Systems 149 (2015) 107–117

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Design of experiments and data analysis challenges in calibration for forensics applications

Christine Anderson-Cook, Tom Burr *, Michael S. Hamada, Christy Ruggiero, Edward V. Thomas

30th Anniversary Issue in Chemometrics and Intelligent Laboratory Systems

ARTICLE INFO

Article history:
Received 17 April 2015
Received in revised form 7 July 2015
Accepted 8 July 2015
Available online 15 July 2015

Keywords:
Design of experiments
Measurement error
Multivariate calibration
Nuclear forensics

ABSTRACT

Forensic science aims to infer characteristics of source terms using measured observables. Our focus is on statistical design of experiments and data analysis challenges arising in nuclear forensics. More specifically, we focus on inferring aspects of experimental conditions (of a process to produce product Pu oxide powder), such as temperature, nitric acid concentration, and Pu concentration, using measured features of the product Pu oxide powder. The measured features, *Y*, include trace chemical concentrations and particle morphology such as particle size and shape of the produced Pu oxide power particles. Making inferences about the nature of inputs *X* that were used to create nuclear materials having particular characteristics, *Y*, is an inverse problem. Therefore, statistical analysis can be used to identify the best set (or sets) of *X*s for a new set of observed responses *Y*. One can fit a model (or models) such as $Y = f(X) + error$, for each of the responses, based on a calibration experiment and then "invert" to solve for the best set of *X*s for a new set of *Y*s. This perspectives paper uses archived experimental data to consider aspects of data collection and experiment design for the calibration data to maximize the quality of the predicted *Y*s in the forward models; that is, we assume that well-estimated forward models are effective in the inverse problem. In addition, we consider how to identify a best solution for the inferred *X*, and evaluate the quality of the result and its robustness to a variety of initial assumptions, and different correlation structures between the responses. We also briefly review recent advances in metrology issues related to characterizing particle morphology measurements used in the response vector, *Y*.

1. Introduction

Forensic science is a broad discipline that aims to infer characteristics of source terms using measured observables. In many cases, the inferred source terms can be traced to an attribution goal. For example, if certain ranges of input Pu concentration could be ruled out, then particular processes that otherwise might have produced the product could be ruled out. Our focus is experiment design and data analysis challenges arising in nuclear forensics. We focus on inferring aspects *X* of experimental reaction conditions using measured features *Y* of the reaction product. The word "signature" is a useful qualitative term that conveys shifts in the responses *Y* as a function of changes in the processing conditions *X*. Such a task falls within the scope of the Department of Homeland Security National Technical Nuclear Forensics Center, which is sponsoring a Plutonium processing signatures multiyear project to identify signatures of nuclear forensic value in plutonium (Pu) materials that can be related to the processing conditions used to produce them. One initial goal is to identify possible signatures derived from PuO_2 produced via a Pu(III) oxalate precipitation process shown in Fig. 1: [11,14,20,61]. The inferred processing conditions could help indicate what facility and settings were used to make the interdicted special nuclear material.

The context we focus on in this perspectives paper is to produce a variety of PuO_2 materials via a statistically designed experiment that is currently in the planning stages in which certain process factors are deliberately varied over a course of experimental trials. Other process factors not discussed here will be held constant. Each experimental trial involves specific settings for each process factor. Once produced, the materials are analyzed to characterize their morphological, chemical and physical properties (which may include trace element concentrations, surface area, crystalline phase, and porosity). Collectively, it is expected that the properties of the materials produced will span a wide range. Design of experiments enables a causal relationship between the factors and responses to be systematically explored and functional models describing the relationship between them to be estimated. Ultimately, the goal is to be able to infer the set of experimental conditions used to produce the material, based on its observed properties using an inverse modeling approach.

In anticipation of the new experimental data in the near future, this perspectives paper uses archived experimental data from a previous similar process to consider aspects of data collection and experiment

* Corresponding author.
E-mail address: tburr@lanl.gov (T. Burr).

http://dx.doi.org/10.1016/j.chemolab.2015.07.008
0169-7439/

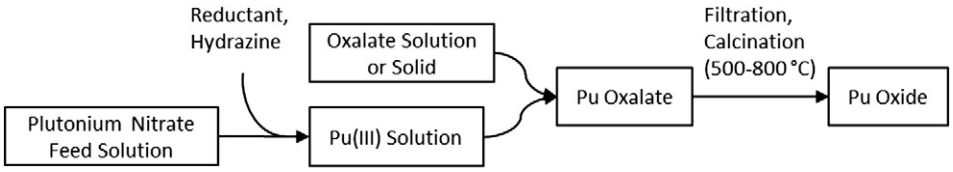

Fig. 1. PuO2 production via a Pu(III) oxalate process that is one of many possible Pu precipitation and crystallization methods for the conversion or recovery of plutonium. Pu(III) oxalate precipitation has been used since the Manhattan Project because of its good Pu recovery and ability to be filtered. Experimental reaction conditions include, for example, the form of the oxalate and the order of addition (Pu to oxalate or vice-versa, reagent/reactant concentrations and addition rates, nitric acid concentration. temperature and duration, stir rates). These *X*s create different features *Y* in the PuO_2. The *Y* features include particle morphology (size, shape, structure, porosity), particle agglomeration, and trace element containment concentrations. Literature suggests that the process removes impurities such as Al, Fe, and UO_2, but there is less "decontamination" from Na, Ca, K; and none from Americium [22]. Different precipitation processes have different decontamination factors, so these trace elements can be a process signature.

design for the calibration data to maximize the quality of the predicted *Y*s in the forward models; that is, we assume that well-estimated forward models are effective in the inverse problem. In addition, we consider how to identify the best solution for the inferred *X*, and evaluate the quality of the result and its robustness to a variety of initial assumptions, and different correlation structures between the responses. We also briefly review recent advances in metrology issues related to characterizing particle morphology measurements used in the response vector, *Y*, and describe the needed statistical tools that have been developed for chemometrics and other applications. Also, new technical challenges arise in this forensics context, including: the need to characterize measurement error, with possible censoring (data reports as "less than a particular threshold" or "more than a threshold"), making use of interlaboratory sample exchange programs to help estimate measurement biases, new assay protocols, particularly for particle morphology measurements using modern image analysis to characterize particle size and shape, the need to allow for a combination of qualitative and quantitative factors, and the fact that regardless of what functional form is assumed to relate *X* to *Y*, once the data are collected, many functional forms will be evaluated.

We have begun to generate candidate experiment designs following established optimality criteria for the forward model with an approximately known functional form. Simulation is expected to be a key analysis tool to address the metrology, experiment design assessment, calibration, and model selection challenges. This allows candidate designs to be compared across different potential outcomes incorporating scenarios described in Sections 5 and 6.

This perspectives paper reviews multivariate calibration, experiment design for multivariate calibration, and describes to what extent previous literature helps address the example nuclear forensics application. Section 2 gives more background. Section 3 describes multivariate calibration and experiment design as used in chemometrics, with a focus on the inverse problem. Section 4 describes related metrology issues, focusing on morphological measurements with particle morphology measurements included the predictors *X* and with recent ongoing efforts to improve morphology measurements. Section 5 describes current progress in the context of the motivating nuclear forensics example. Sections 6 and 7 include research directions and a summary.

2. Background

We use the following notation. There are *p* factors (predictors) and *q* responses.

Factors = Processing conditions such as temperature, nitric acid and Pu concentrations $X_1,X_2,\ldots,X_p$. Complete factor set: $X = \{X_1,X_2,\ldots,Xp\}$.

Responses = Measurements of processed material such as particle size and shape

$Y_1,Y_2,\ldots,Y_q$

Signature = Complete set of responses, $Y = \{Y_1,Y_2,\ldots,Yq\}$

Forward (causal) models are often developed using the results of a controlled experiment. Forward models are often (but not necessarily) expressed in terms of a low-order polynomial, which can be thought of as a Taylor series approximation to the true underlying relationship. Such models relate a specific response (Y_i) to the complete set of factors. Generically, this relationship can be expressed as $Y_i = f_i(X_1,X_2,\ldots,X_p)$. The main forensics goal is to use the signature acquired from material of an unknown pedigree to infer the conditions used to process the material. Inverse prediction, generically expressed by $\hat{X}_j = g_j(Y_1, Y_2, \ldots, Y_q)$, can be used to predict the value of the j^{th} process factor based on measured properties. One way to perform inverse prediction is with a collection of fitted forward models from the signature of the unknown material. Alternatively, one might use the data acquired from the experiment to form a training set from which an inverse model can be developed directly,without using forward models [29,30,37,38,46]. While physics and chemistry dictate the causal relationship between factors and responses, experimenters can influence control over the complexity of the fitted forward models via selection of factors, factor levels, and the set of trials performed. Experimenters also influence the effectiveness of the inverse modeling process by selecting a sufficiently informative/discriminating set of response variables (the "signature") that can be used to unambiguously resolve the various factors. For a given sample, the number of responses comprising the signature may be limited by difficulty, expense, and/or availability of material. We believe that there should be at least as many response variables as factors ($q \geq p$) to successfully deduce the complete set of conditions used to process an unknown material. With fewer responses ($q < p$), there is considerable potential for ambiguous results with non-unique solutions.

To illustrate, Fig. 2a is a principal coordinate plot [62] of the distances between the *X* vectors for each of the 72 samples. The *X* vector is the percentages of fat, sucrose, dry flour, and water in a cookie recipe. The response vector, *Y*, is 700 near-infrared (NIR) reflectance from 1100 to 2498 nanometers in steps of 2 nm. This is freely available data from the R [49] package ppls (which provides functions for linear and nonlinear regression based on partial least squares and penalization techniques; see Section 3.1). From Fig. 2a, we select samples 24 and 51 (in the bottom left corner of the plot) that are close in the *X*-space, and sample 19 (top right) far away. We plot the corresponding three spectra in Fig. 2b. Qualitatively, in this case, a small (large) distance in *X*-space corresponds to a small (large) distance in *Y* space between samples, which indicates that one could anticipate using *Y* to predict *X* with reasonable success.

In processes to produce Pu oxide, we estimate the relationship between the chemical engineering processing parameters employed and the physical, chemical, and morphological characteristics of the produced materials. The response *Y*s used to infer processing conditions include morphological features and trace chemical concentrations, but could also include other analytical chemistry measurements (such as crystallographic phase, surface area, chemical form, heterogeneity, oxidation state, isotopic composition, nitrate/chloride/sulfate/hydroxide/carbide or other anion or organic contaminants present). Inferred processing conditions such as temperature, Pu and nitric concentration could indicate the facility used to produce any future interdicted Pu powder. Due to the large number of chemical engineering flow sheets used historically, process variations within each flow sheet, and the potential for even more variations among international and subnational