Computers and Chemical Engineering xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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

Mixture formulation through multivariate statistical analysis of process data in property cluster space

Subin Hada, Robert H. Herring III, Mario R. Eden*

Department of Chemical Engineering, Auburn University, Auburn, AL 36849, USA

ARTICLE INFO

Article history:
Received 6 September 2016
Received in revised form 7 June 2017
Accepted 12 June 2017
Available online xxx

Keywords:
Mixture formulation
Model reduction
Optimization
Chemical product design
Visualization
Systems engineering

ABSTRACT

Data-driven modeling approaches are suitable for representing complex processes and phenomena in cases where cause-and-effect cannot be easily described from first-principles. Chemical product formulation in industrial research and development is an area where the analysis of mixture data could be utilized more effectively. Correlation, either partial or complete, is inherent in such mixture data and requires the use of multivariate statistical tools for visualization and identification of important relationships in the data. In this paper, a systematic methodology is developed by integrating data-driven chemometric techniques and property based visualization and optimization tools to solve mixture formulation problems involving multi-block data structures. Effort has been focused on: the development of mathematical models by utilizing multivariate understanding of process and product data, visually identifying design targets a priori, and decomposition of the design problem by incorporating the concept of reverse problem formulation and property clustering techniques. A case study in industrial thermo-plastic development is presented to illustrate the methodology developed in this paper.

1. Introduction

In computer-aided approaches for solving chemical product/process problems, mathematical model-based optimization strategies are often used. Mathematical programming techniques require mathematical models that describe the underlying phenomena with a system of differential and algebraic equations (DAEs). However, in real-life industrial process systems, good theoretical process models are often not available and/or cannot adequately describe these complex processes (Kettaneh-Wold, 1992; MacGregor and Cinar, 2012; Venkatasubramanian, 2009).

Here, data-driven models offer an alternative solution, where historical process data that encompasses a wide spectrum of operating conditions and existing product grades can be utilized to build mathematical models describing the relationships between final product properties and process variables (Jaeckle and MacGregor, 1998; Kettaneh-Wold, 1992; MacGregor et al., 2005; Muteki and MacGregor, 2007; Venkatasubramanian, 2009). For instance, the design of mixture formulations involving emulsions and polymers necessitate the use of data-based solution approaches. In recent decades, data-driven techniques are gaining momentum as massive amounts of data describing process and product variables can be collected relatively cheap and quickly through extensive monitoring of equipment, processes, and products at all scales (Eriksson et al., 2006; Jaeckle and MacGregor, 1998; MacGregor and Cinar, 2012; Venkatasubramanian, 2009).

Mixing/blending processes play an important role in today's manufacturing of value-added specialty chemical products such as food, cosmetics, fuel and pharmaceuticals. Mixture formulation in industrial research and development is an area where the analysis of mixture data could be utilized more effectively. However, managing such an abundance of large, complex, and information-rich multivariate data sets to build appropriate mathematical models, to control/optimize a set of variables, and to explore their effects on the final product properties remain major challenges.

In this work, chemometric and property clustering techniques are integrated to solve complex mixture design problems in a reverse problem formulation. Chemometric techniques are applied to exploit the interrelationships (latent variables) in the multivariate, noisy, and highly collinear data and to build the necessary property models. These models are further used in the property cluster space to solve formulation problems in a lower-dimensional space. Visualization and solution of the problems in this reduced dimensional space allows identification of optimum strategies for the combination of different raw materials without detail calculations. Such a practice results in fewer trials and experiments to run,

* Corresponding author.
E-mail address: edenmar@auburn.edu (M.R. Eden).

http://dx.doi.org/10.1016/j.compchemeng.2017.06.017
0098-1354/

thereby saving resources, money and most importantly reduces the product development time. The approach differs from conventional optimization techniques because it is non-iterative and avoids the combinatorial explosion when multiple components are involved.

The use of reverse problem formulation enables a two-step approach, where the property targets are identified in the first step and then the mixtures that match the targets are identified in the second step. This enables identification of optimal solutions to mixture formulation problems more easily than solving the conventional forward problems, which are iterative in nature.

Use of the property clustering technique facilitates visualization and provides a rapid targeting tool that can aide in the evaluation, analysis, and screening of feasible blend alternatives involved in mixture design problems without performing extensive enumerations. The unique feature of linear mixing rules allows for the use of simple lever arm analysis to solve the problem in the reduced cluster space. Irrespective of the number of components in the search space, the design can be achieved within a single cluster diagram, thereby avoiding the issue of combinatorial explosion. In addition, the visualization technique facilitates the solution of the design problem in a single domain, thereby minimizing the loss of information in transforming variables from property domain to cluster domain.

While the visual based approach towards the inverse solution of latent variable models has it's benefits, there are many situations in which non-linearities arise in product design and mixture design problems. Techniques have been developed to solve these problems with an optimization approach as well (Tomba et al., 2012). Specifically, sequential quadratic programming has been proven successful towards solving product design problems involving raw material and processing considerations with nonlinear latent variable models built using PLS (Yacoub and MacGregor, 2004). Additionally, mixture design problems often result in the input data matrix taking on a T-shape due to the given knowledge about material properties, mixture ratios, processing conditions and qualities. Techniques have been developed to handle this type of input data (Garcia-Munoz, 2014) that overcome limitations inherent to previous mixture design approaches.

In the following section (Section 2), details on the different techniques used in the development of a methodology for solving formulation problems that require the use of data-based solution approaches are presented. Next (in Section 3), a general solution methodology is outlined detailing the variable transformation and information passing between the steps. Finally (in Section 4), a case study involving the design of thermoplastic is presented to illustrate the developed methodology.

2. Techniques

2.1. Chemometric techniques

Chemometrics is a multivariate data analysis statistical technique that provides a multivariate understanding of process and product data by elucidating hidden relationships in data sets. In chemometrics, principal component analysis (PCA), principal component regression (PCR), and partial least-squares (PLS) provide most effective tools for developing reduced dimensional models from high dimensional processes. PCA and PLS transform a large number of correlated variables into a smaller number of uncorrelated variables called latent variables or principal components (PCs) with minimum loss of information (Eriksson et al., 2006, 1998; Wold et al., 1996). Such transformations of multivariate data ensure the proper structure required to develop latent variable models. They also provide visualization of the multivariate data and describe the correlation structure in the data. The latent variable methods of PCR and PLS are suitable compared to standard regression approaches such as multiple linear regression (MLR) or neural networks (NN), because they extract systematic variables and eliminate multi-collinearities in the multivariate mixture data sets (Eriksson et al., 2006, 1998; Jaeckle and MacGregor, 1998; MacGregor et al., 2005; Muteki and MacGregor, 2007).

Kettaneh-Wold (Kettaneh-Wold, 1992) proposed the use of a PLS model to simultaneously incorporate a process condition matrix (**Z**) together with the blend ratio matrix (**R**). Recently, Muteki et al. (Muteki and MacGregor, 2007; Muteki et al., 2007) proposed a linear mixture-property PLS model by incorporating the raw material properties (**X**), which allowed the investigation of the effect of raw material properties on the final mixture product properties. He combined the raw material properties matrix (**X**) and the blend ratios matrix (**R**) using ideal mixing rules in order to relate them to the product property matrix (**Y**):

$$\boldsymbol{Y} = f\left[(\boldsymbol{R}\cdot\boldsymbol{X}), \boldsymbol{Z}\right] + \varepsilon \tag{1}$$

where $\mathbf{X}_{mix} (= \mathbf{R}\cdot\mathbf{X})$ is the raw material property matrix of the mixtures and ε is the residual. Fig. 1 illustrates the data structures involving all three degrees of freedom available in blending processes (Muteki and MacGregor, 2007). The multi-block data structure in Fig. 1 is referred to as T-shaped when all the data blocks (**X**, **R**, and **Z**) are included in the analysis. When the process condition variable block (**Z**) is excluded, the data block is referred to as L-shaped. In this work, we use these techniques to facilitate the analysis of process data and then, identification of product targets with properties that span the desired product property space a priori to experimentations and simulations.

2.2. Property clustering technique

The property clustering technique is a property-based visualization tool for mapping a design problem from the non-conserved property domain into the conserved (component-less) cluster domain (Eden et al., 2004; El-Halwagi et al., 2004; Eljack et al., 2008; Hada et al., 2012; Shelley and El-Halwagi, 2000; Solvason et al., 2009b). Using property clusters, the problems of combinatorial explosion and visualization difficulties when dealing with multi-component mixtures can be alleviated (Chemmangattuvalappil et al., 2010; Eden, 2003; Hada et al., 2012; Solvason et al., 2009a,b). Utilizing conserved quantities called clusters, this technique facilitates mapping of property relationships into a lower dimensional space, thus allowing for visualization and insights into the problem.

For problems that can be adequately described by just three properties, the mixture design problems are solved visually on a ternary diagram. Here, the roles of the components and the property responses are interchanged from traditional ternary diagrams, and therefore, irrespective of how many components are included in the search space, the design can be achieved within a single cluster diagram. In this work, visualization of the problem allows for easy identification of optimum strategies for the combination of different raw materials, while the unique feature of linear mixing rules allow for the use of simple lever arm analysis to solve the problem in the reduced cluster space.

2.3. Reverse problem formulation

The reverse problem formulation technique decomposes mixture design into two reverse problems linked by design targets/constraints (Eden et al., 2004; Eljack et al., 2008; Solvason et al., 2009b). This enables a two-step approach, where the property targets are identified in the first step (using chemometric techniques) and then the mixtures that match the targets are identified in the second step (using the property clustering technique). This enables

Please cite this article in press as: Hada, S., et al., Mixture formulation through multivariate statistical analysis of process data in property cluster space. Computers and Chemical Engineering (2017), http://dx.doi.org/10.1016/j.compchemeng.2017.06.017