Integrating Mixture Design within the Property Clustering Framework
Proceedings of European Congress of Chemical Engineering (ECCE-6)
Copenhagen, 16-20 September 2007

Integrating Mixture Design within the Property Clustering Framework

Charles C. Solvason[a], Fadwa T. Eljack[a,b], Ninshanth G. Chemmangattuvalappil[a], Mario R. Eden[a]

[a]*Department of Chemical Engineering, Auburn University, Auburn, AL, USA*
[b]*Department of Chemical Engineering, Qatar University, Doha, Qatar*

Abstract

Mixture Design is a Design of Experiments (DOE) tool used to determine the optimum combination of chemical constituents that deliver a desired response (or property) using a minimum number of experimental runs. While the approach is sufficient for most experimental designs, it suffers from combinatorial explosion when dealing with multi-component mixtures. To circumvent this problem, a recently developed design technique called Property Clustering is applied [1]. In this type of design the properties are transformed to conserved surrogate property clusters described by property operators, which have linear mixing rules even if the operators themselves are nonlinear. Product and process property targets are then used to describe a feasibility target region. To solve the mixture design, components are mixed according to property operator models in a reverse problem format until the mixture falls within the feasibility target region. Once candidate solutions are found, they can be screened with additional criteria per the experimenter's preference. The degree of accuracy of this modeling technique depends heavily on the ability of the property operator models to adequately describe the property within the studied design space. This work focuses on the utilization of linear Scheffe and Cox models as the property operators and discusses the implications of negative regressors on property clustering design space.

Keywords: design of experiments, property clusters, mixture design, principal component regression, pharmaceutical synthesis

1. Introduction

The terms product synthesis and design designate problems involving identification and selection of compounds or mixtures that are capable of performing certain tasks or possess certain physical properties. Since the properties of the component or mixture of components dictate whether or not the design is useful, the basis for

solution approaches in this area should be based on the properties themselves. However, the performance requirements for the desired component are usually dictated by the process and thus the identification of the desired component properties should be driven by the desired process performance. Where as numerous contributions have been made in the areas of molecular synthesis and Computer Aided Molecular Design (CAMD) e.g. Harper and Gani [2], Marcoulaki and Kokossis [3], and Eljack et al. [4], little focus has been on utilizing experimental design techniques in situations where property prediction tools are insufficient in describing the mixture's properties. Early in experimental mixture design, Scheffe [5,6] and Cox [7] developed techniques to obtain property operator models while minimizing experimental runs or design points utilizing simplex diagrams of the chemical constituent design space. However, visualization of the solution in the chemical constituent design space leads to combinatorial explosion. Viewing the problem in the property space avoids this problem while also offering insights into the effectiveness of the design. While other methods such as Principal Component Regression (PCR) and Partial Least Squares on to Latent Surfaces (PLS) are typically utilized under these conditions, it will be shown in an additional paper that these works also benefit from utilizing property clustering.

2. Objective

The overall objective of this contribution is to integrate the property clustering framework with existing mixture design techniques. Specifically, using property clusters to visualize the response in the property domain rather than the component domain to aid in the physical understanding of the experiment in cases of combinatorial explosion. The two most common mixture designs, Scheffe canonical models and Cox polynomial models, were evaluated. The results of the exercise will be used to develop additional techniques for visualization of PCR and PLS score plots under combinatorial explosion that will be published in a later document.

3. Experimental Design

Design of Experiments (DOE) is a form of experimental design that utilizes statistical methods to plan and execute informative experiments [8]. A model is postulated to represent the response surface. Experimental design points are placed in areas where observations can be collected to which the model can be fitted. In the final step, the adequacy of the model is tested. The procedure may require much iteration until the fitted equation is determined by the experimenter to be sufficient [9]. The most effective choice of model and location of design points is the focus of the experimenter. The best set of points is chosen under the following constraints: (1) the size and shape of the experimental region, (2) the number of desired experimental runs, and (3) the type of model used for constructing the "map" of the response [10]. Most often, the polynomial model is selected to represent the response surface since it can be expanded through a Taylor series to the desired level of accuracy [9]. Either a first or second degree polynomial is usually chosen to represent the surface since it requires fewer observations. Third degree or higher ordered models are seldom utilized. The point estimate forms of the models are listed in equations 1 and 2.

$$y = \beta_o + \sum_{i=1}^{u} \beta_i x_i \quad (1)$$

$$y = \beta_o + \sum_{i=1}^{u} \beta_i x_i + \sum_{i \le j}^{u} \sum_{j \ge i}^{u} \beta_{ij} x_i x_j \quad (2)$$

The response y is the point estimate of unbiased estimator of the response. β_o, β_i, and β_{ij}, are the point estimates of the regression coefficients. The total number of i or j chemical constituents is u. Least squares regression is the technique of choosing the regression coefficients that maximize the model sum of squares and minimize the residual sum of squares. Forms of least square regression can be found in many techniques such as Classical Least Squares (CLS), Inverse Least Squares (ILS), Multiple Linear Regression (MLR), Principal Component Regression (PCR), and Partial Least Squares applied to Latent Surfaces (PLS).

To derive the regressor solution for CLS, it is convenient to switch to matrix notation. Without it, the formulas become unmanageable when the number of explanatory variables increases [11]. Rewriting equation 1 in terms of matrix notation,

$$Y = BX \quad (3)$$

Where Y, B, and X are matrices of the predicted responses, estimated regression coefficients, and components fractions, respectively.

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_q \end{pmatrix} \quad (4)$$

$$B = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_u \end{pmatrix} \quad (5)$$

$$X = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,u} \\ 1 & x_{2,1} & \cdots & x_{2,u} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{q,1} & \cdots & x_{q,u} \end{pmatrix} \quad (6)$$

Where q represents the total number of experiments. In terms of matrices, equation 3 can be rewritten as equation 7 [11].

$$X'Y = X'XB \tag{7}$$
$$B = X'Y(X'X)^{-1} \tag{8}$$

Solving for the best fit regressors gives equation 8 which is the central result of Classical Least Squares (CLS) analysis. The same procedure can be extended to second order and third order models.

3.1. Multiple Linear Regression (MLR)

Instances where more than one property or response needs to be optimized simultaneously require the use of Multiple Linear Regression (MLR). MLR is an extension of classical least squares regression to more than one response. Rewriting equations 4-6 in terms of multiple responses gives equations 9-11.

$$Y = \begin{pmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,p} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{q,1} & y_{q,2} & \cdots & y_{q,p} \end{pmatrix} \tag{9}$$

$$B = \begin{pmatrix} \beta_{0,1} & \beta_{0,2} & \cdots & \beta_{0,p} \\ \beta_{1,1} & \beta_{1,2} & \cdots & \beta_{0,p} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{u,1} & \beta_{u,2} & \cdots & \beta_{u,p} \end{pmatrix} \tag{10}$$

$$X = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,u} \\ 1 & x_{2,1} & \cdots & x_{2,u} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{q,1} & \cdots & x_{q,u} \end{pmatrix} \tag{11}$$

Equation 10 demonstrates that many possible solutions exist for the regressors. However, due to the construction of the matrices, the least squares solution derived in equation 8 is still a viable method for determining which combination of regressors gives the best estimate of the responses. For the inverse of the identity matrix *X'X* to exist, *X* must have as many rows as columns. Since *X* has one row for each sample and one column for each component, then it follows that there must be at least as many samples *q* as components *u* to be able to compute equation 8. The rules governing MLR relating to samples and components are summarized by Geladi and Kowalski as the following [12]:

- For $q > u$, there is no unique solution unless one deletes the independent variables.
- For $q = u$, there is one unique solution.

- For $q < u$, a least-squares solution is possible. For $q = u$ and $q > u$, the matrix inversion can cause problems (e.g. collinearity).

Likewise, any linear dependence among the rows or columns of X will create a singular $X'X$ matrix and its inverse will not exist [13]. This is a key observation that will be discussed later.

3.2. Factorial Design

In factorial design each of the studied factors may be either quantitative or qualitative in nature. The factors are set to vary over a design range, the exact location of the design points depending on the base of the design. The most common base for designs is base 2 where the factors are observed at both the higher and lower bounds of the range. In base 3, center points are added, equidistant from the higher and lower bounds. For all of the designs, the design points are located around a nexus that is as close to the expected target as possible. The number of experiments used in the design is set according to the type of model chosen to reflect the response. It has been shown that polynomial models are the most efficient models for describing experimental designs [9]. The ability to tailor the model by expansion to higher orders and/or the inclusion of interaction terms depending on the objective of the experiment provides the necessary flexibility for successful experimental design. For instance, if a screening design is chosen, then linear models without interaction effects are usually adequate. If indicators suggest non-linearity, then the original design can be augmented with more design points and a second order model can be applied. If optimization of the design space is the objective, still more design points are required. There are many established designs to choose from, but the trend has been away from set designs and toward designs that maximize optimality. One type of optimality design is the D-optimal design which utilizes an algorithm to maximize the design space within a specified number of points [14]. The designs are particularly useful when a constrained region is studied and no classical design is available [15]. Montgomery gives a more detailed discussion of D-optimal designs and other alphabet designs [16].

The number of experiments grows rapidly with an increasing number of factors, thus when dealing with $i > 5$ variables or components, a full factorial design is not the best option for screening. Rather the fractional factorial becomes the choice in design and allows the experimenter to increase the number of factors for the same number of experiments. The trade off in using fractional factorial design is the confounding of factors with one another. With some forethought, this situation is handled by maximizing the resolution of the design. Typically, a resolution four (RIV) design is chosen that ensures that all main factors are protected from two-way interactions. Sometimes when experimental runs are constrained resolution three (RIII) designs may be chosen which confounds some 2-way interactions with the main factors. In those cases, the experimenter may use previous knowledge to dismiss the infeasible two way interactions prior to executing the experiment.

3.3. Mixture Design

Mixture design is an extension of DOE that utilizes chemical constituents as the factors in the design. By definition the constituent fractions must sum to one and each constituent fraction must lie between zero and one.

$$\sum_{i=1}^{u} x_i = 1 \tag{12}$$

$$0 \le x_i \le 1 \tag{13}$$

This equation imposes a collinearity effect by removing the independence of the factors. While it does not affect the utilization of the model, it does impact the interpretation of the regression coefficients. The reason for using the constraint is to provide a means for visually representing more data by using a ternary diagram or simplex to describe the design space as shown in figure 1. As was the case for factorial design, a model is first postulated to represent the response surface. The number and location of the design points are selected based on the objective of the experiment. Scheffe developed the first simplex-lattice designs which many researchers considered to be the foundation of mixture design [9]. To develop these designs, Scheffe noted that the location of the response of a mixture made up of exactly zero constituents must be identically zero meaning that the coefficient β_o is zero [5, 6]. Furthermore, the use of equation 12 means that the quadratic terms can be rewritten as equation 14.

$$x_i^2 = x_i\left(1 - \sum_{j=1}^{u} x_j\right) \tag{14}$$

Combining these observations, Scheffe derived the canonical models in equations 15 and 16.

$$y = \sum_{i=1}^{u} \beta_i^* x_i \tag{15}$$

$$\hat{y} = \sum_{i=1}^{u} \beta_i^* x_i + \sum_{i<j}^{u} \sum_{j>i}^{u} \beta_{ij}^* x_i x_j \tag{16}$$

The effect of these canonical models is to remove the quadratic and higher terms from analysis and leave behind only the modified pure component properties and interaction effects. However, it must be noted that the modified regressors β_i^* and β_{ij}^* do not represent only pure component properties or only interaction effects. On the contrary, because of the collinearity introduced in the derivation of the canonical

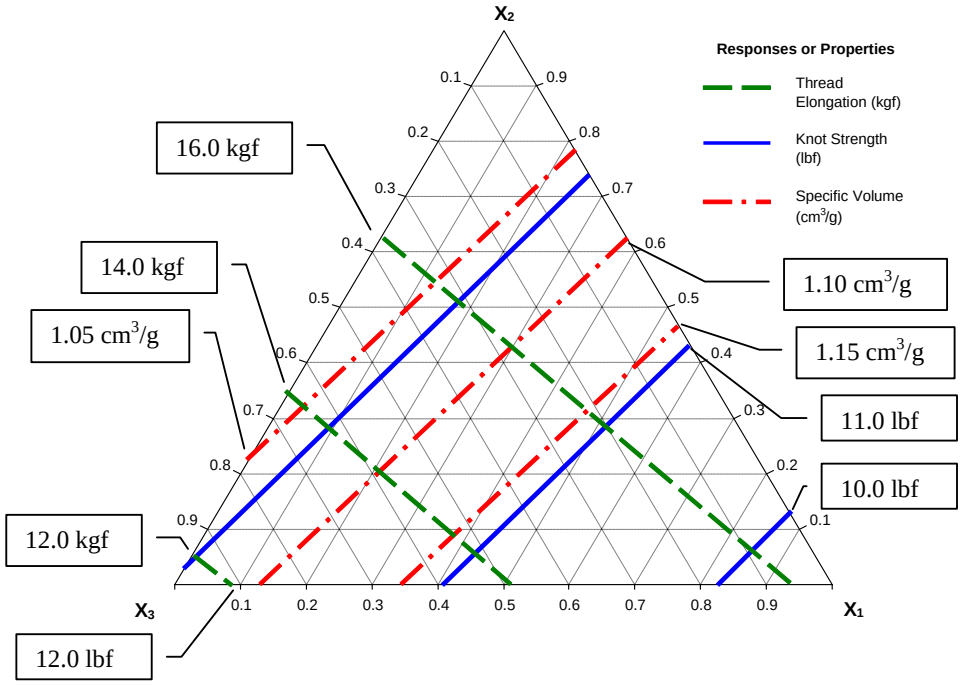


Figure 1: Mixture simplex diagram of polyethylene (x_1), polystyrene (x_2), and polypropylene (x_3) using experimental data from Cornell [9].

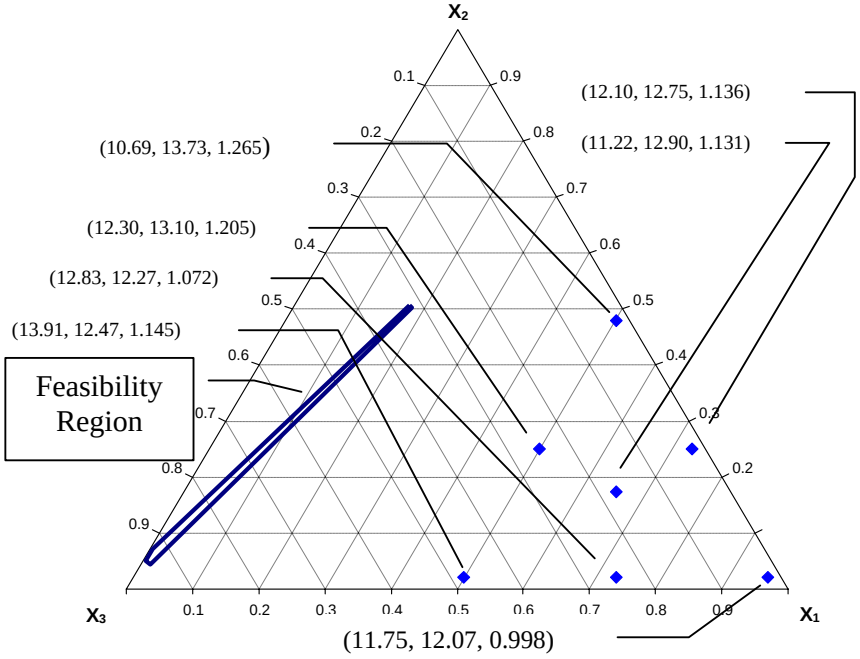


Figure 2: The design of a three-component experiment with the observed response values of thread elongation (kgf), knot strength (lbf), and specific volume (cm^3/g) at the design points [9].

models, the modified regressors are amalgams of the pure component, interaction, and quadratic effects, as shown in equations 17 and 18.

$$\beta_i^* = \beta_0 + \beta_i + \beta_{ij} \tag{17}$$

$$\beta_{ij}^* = \beta_{ij} - \beta_{ii} - \beta_{jj} \tag{18}$$

The first and second order canonical models are postulated to represent over 66% of all mixture designs encountered. In situations where a higher order cubic design is desired, the same procedure may be applied to the cubic polynomial equation to make the cubic canonical model.

$$\hat{y} = \sum_{i=1}^{u} \beta_i^* x_i + \sum_{i<j}^{u} \sum_{j>i}^{u} \beta_{ij}^* x_i x_j + \sum_{i<j}^{u} \sum_{j<k}^{u} \sum_{k>i,j}^{u} \beta_{ijk}^* x_i x_j x_k \tag{19}$$

To execute a mixture design using canonical models, Scheffe proposed using lattice points on a simplex diagram as shown in figure 2 [5, 6]. This technique simplifies the least squares calculation of regression coefficients. Since the design is a boundary design with as many points as regression terms, then the estimated terms may be taken directly from the responses at each design point [17]. It should be noted that for canonical models, the regression coefficients are no longer represented by half the response; rather, they represent the pure properties of the constituents and are no longer orthogonal. The true property value, or orthogonal effect, is estimated by taking the difference in the value of the response at pure and infinitely dilute solutions while holding all other constituents constant. Cox noted that Scheffe's simplex lattice and simplex centroid models require pure components and a range of constituent fractions from zero to one. In his paper he points out that Scheffe's models have the following drawbacks [7]:

1. If two replicate experiments on the same system have the same expected responses except for a constant difference between replicates, it is obvious that on fitting separate replicates all of the regression coefficients will be different in the two replicates.
2. The absence of squared terms makes it meaningless to consider the direction and magnitude of curvature of the response to a particular component.
3. The interpretation of the regression coefficients is in terms of the responses for simple mixtures.

In addition Kettaneh-Wold points out that the removal of the constant term to generate the canonical model makes it impossible to center these models which often leads to an ill-conditioned $X'X$ matrix in equation 23 [10]. An ill-conditioned matrix is not symmetrical about the main diagonal, which means that the order of differentiation is important which can lead to errors during inversion. It is of the utmost importance that the constituents be independent when performing CLS and MLR regressions. Collinearity may also result from additional constraints such as the upper and lower bounds on components [17].

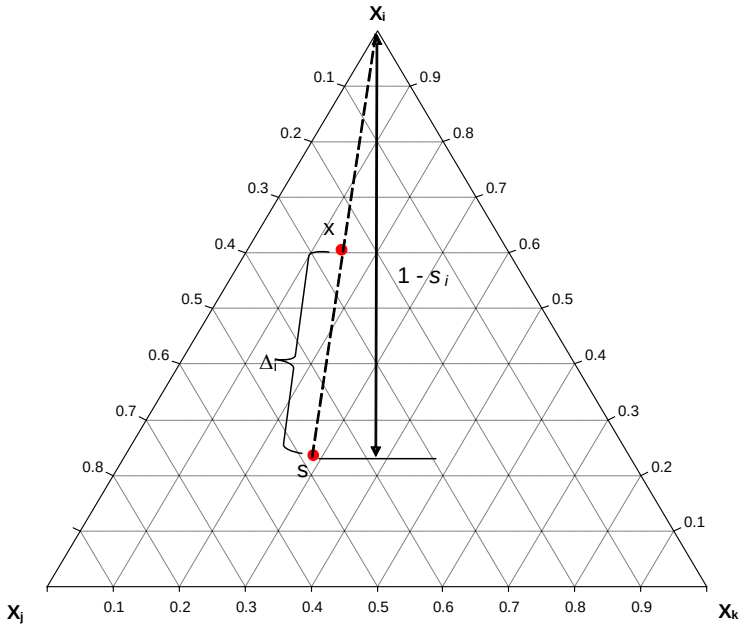


Figure 3: The parameterization of the component space in terms of a standard reference mixture (*s*) such that the incremental change Δ_i in the proportion of the component *i* is indicative of its effect on the response [9].

To address these issues, Cox proposed a variable transformation with an arbitrary reference mixture that would allow for the use of existing polynomial models with additional constraints involving a reference mixture called the standard mixture [7]. Shown in figure 3 is a simplex diagram with a standard mixture *s* and a mixture *x* with a larger proportion of constituent x_i. Noting that a *x* lies on a line from *s* to the x_i vertex, then the ratio of the other *u*-1 constituents are in the same relative proportions as the standard mixture. The proportions are written as equations 20 and 21.

$$x_i = s_i + \Delta_i \tag{20}$$

$$x_j = s_j - \frac{\Delta_j s_j}{(1 - s_i)} \tag{21}$$

Rewriting the first and second degree models, equation 1 and 2, in terms of the change in constituent *i*, Δ_i, gives equations 22 and 23 which provide a direct link between the regressors and the position of the design points to the reference mixture [9].

$$y(x) = y(s) + \sum_i^u \left(\frac{\beta_i}{1 - s_i} \right) \Delta_i \tag{22}$$

$$y(x) = y(s) + \sum_{i}^{u}\left(\frac{\beta_i}{1-s_i}\right)\Delta_i + \sum_{i}^{u}\left(\frac{\beta_{ii}}{(1-s_i)^2}\right)\Delta_i^2 \qquad (23)$$

Where *y(x)* is the expected response at design point *x* and *y(s)* is the expected response at the standard reference mixture. The estimated response surface contours generated by the two model forms are identical [9]. Hence, as Smith and Beverly point out, the gradient, or change in response per unit change in x_i, at **s** along the Cox-effect direction is called the *effect of* x_i, provided x_i is free to range from 0 to 1 [18]. While this transformation of the original Scheffe polynomials removes the primary collinearity introduced by equation 24, it leaves the secondary collinearities such as those introduced by constraints on the constituent ranges. Kettaneh-Wold suggests that the best solution maybe to refrain from interpreting the coefficients and rely on the predictions only but notes this solution is not acceptable in practice since the interpretation of regression coefficients is a necessity when the objective is to find component effects in screening situations [17]. It is in this arena that Kettaneh-Wold suggests the use of Principal Component Regression (PCR) and Partial Least Squares on to Latent Surfaces (PLS) [17]. However, Property Clustering is another method that may help in the interpretation of the regressors.

4. Property Clustering

Property Clustering was developed as a tool for tracking properties in a conserved manner by Shelley and El-Halwagi [19]. It was later applied to process and product design by Eden et al. [1]. As a design tool, the technique challenges conventional design by reversing traditional mixture designs to solve for chemical constituents that meet property constraints prior to experimentation. Essentially, the technique converts properties into conserved surrogate property clusters that are described by property operators, which have linear mixing rules, even if the operators themselves are nonlinear. In equation 53, the property, *y*, is described by a linear property operator equation.

$$\psi_k(y_k)_M = \sum_{i=1}^{u}\psi_k(y_k)_i \cdot x_i \qquad (24)$$

Although the property operator equation must be linear, the property operator itself may be nonlinear. For example, if the property operator describes density, then to meet the linear criteria imposed by equation 53 we would use specific volume as the property operator, ignoring any interaction effects from mixing, as shown in equations 25 and 26.

$$\psi_k(y_k)_i = \frac{1}{\rho_i} \qquad (25)$$

$$\frac{1}{\rho_M} = \sum_{i=1}^{u} \frac{1}{\rho_i} \cdot x_i \tag{26}$$

This step is often difficult as many properties can not be described adequately by linear models. In this situation the experimenter must weigh the benefits of the method versus the loss of solution certainty. In most cases, this caveat leads the experimenter to limit the application of this method to high volume screening designs.

After the property operator equations are defined, the method is straightforward and universal. As shown in equation 27, the property operators are non-dimensionalized by dividing by a reference property operator [1].

$$\Omega_{ki} = \frac{\psi_k(y_k)_i}{\psi_k(y_k)_{ref}} \tag{27}$$

This step serves the purpose of scaling the property design space to facilitate an easier graphical design space. As shown later in this paper, it is also used to ensure a solution to mixtures whose property operator equations contain negative regression coefficients. The non-dimensionalized properties are then summed into the Augmented Property Index (*AUP*) [19].

$$AUP_i = \sum_{k=1}^{p} \Omega_{ki} \tag{28}$$

A cluster is then defined by dividing the non-dimensionalized property by the property cluster, as shown in equation 29.

$$C_{ki} = \frac{\Omega_{ki}}{AUP_i} \tag{29}$$

Shelley and El-Halwagi have shown that this property cluster is conserved through mixing and can be used to track the property through mixing functions [19]. To graphically represent, the cluster, a ternary diagram, or simplex, is used in much the same way as a mixture design (Figure 4). Each of the vertices represent a pure property cluster. In situations where more than 3 properties are needed to describe the system, an algebraic approach may be used [20]. The properties of each pure component, as predicted from the property operator equations, are plotted on the simplex. Due to collinearities in the property operator equations, the prediction may be outside of the cluster space. However, this situation can be handled by using pseudo clusters to augment the property cluster design space as discussed later.

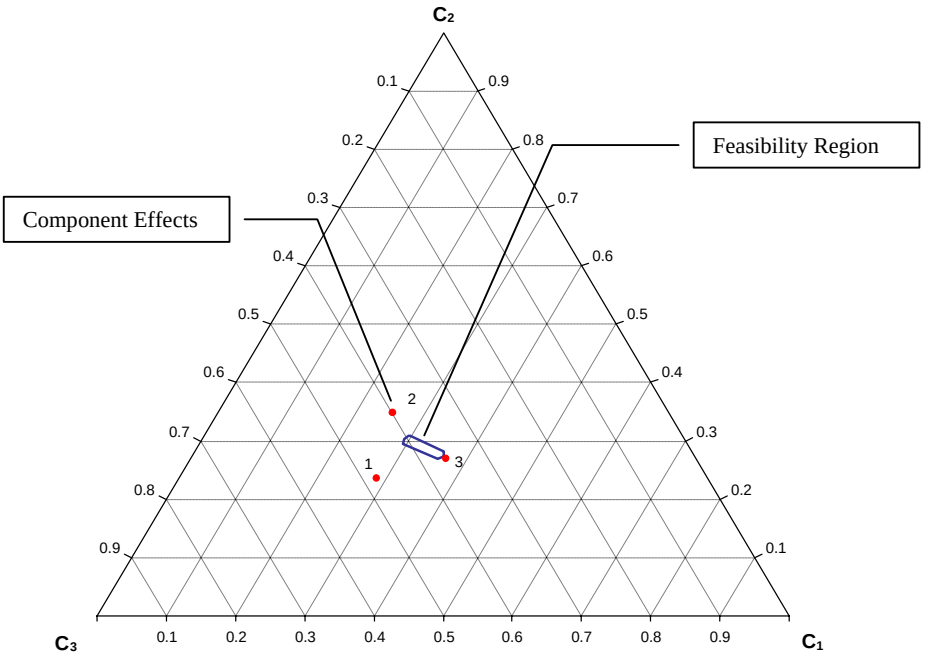


Figure 4: A property cluster simplex diagram using Scheffe property operators. Property clusters 1, 2, and 3 represent the properties thread elongation, knot strength, and specific volume. Components 1, 2, and 3 represent polyethylene, polystyrene, and polypropylene, respectively.

Also shown in figure 4 is a feasibility region, which is mapped to the design space using targets specified by the product designer. To solve the design problem, the components are mixed according to linear mixing rules [1].

$$C_{kM} = \sum_{i=1}^{u} \delta_i \cdot C_{ki} \tag{30}$$

Where the relative cluster arm δ_i is defined using the proof for inter-stream conservation of clusters given by Eden and Shelley and El-Halwagi, resulting in equation 31 [1, 21].

$$\delta_i = \frac{x_i \cdot AUP_i}{AUP_M} \tag{31}$$

To successfully meet mixing criteria, mixtures must meet three conditions [1]:

Rule 1. Cluster value of the source (or mixture of sources) must be contained within the feasibility region of the sink on the cluster ternary diagram.

Rule 2. The values of the Augmented Property Index (*AUP*) for the source or mixture of sources and the sink must match.

Rule 3. The flow rate of the source (or mixture of sources) must lie within the acceptable feed flow rate for the sink.

The original rule 3 must be modified slightly to meet the criteria of mixture design. Since each source must be considered as a pure component, then the constituents are subject to the additional constraint of summing to unity as imposed by equation 28. Therefore, a more appropriate statement of Rule 3 is

Rule 3. The constituent fraction of the candidate mixtures must match the constituent fractions of the sink

In addition to these three rules, the experimenter may wish to apply additional constraints, such as limiting the mixture size, type, or cost. These constraints assume an optimization role in the design to further reduce the generated list of potential candidate mixtures.

In summary, property clustering can represent the multitude of response plots normally associated with experimental design on a single simplex for systems with up to 3 properties. Additional properties are either represented with additional diagrams or solved algebraically. Property clustering also consolidates the numerous effects of components on the response into distinguishable values which may be used to surmise which of the components has the largest effect and is thus most important. Both of these property clustering attributes can be used to increase the efficiency of the existing experimental designs.

5. Integration of Property Clustering with Mixture Design

As discussed earlier, property clustering is a transformation technique which facilitates a reverse problem solution. Instead of sampling to determine property values, or estimating property values based on constituent values, the property value is specified and the appropriate constituent mixture is determined empirically to match the solution. Utilizing property clustering in a reverse problem solving role not only avoids combinatorial explosion, but offers the potential for solving process, mixture, and molecular design problems simultaneously [4]. The key to successful modeling with property clustering is to utilize appropriate property operator models. Traditionally these have been based on established property models tailored to a linear format. However, in situations where models with the necessary degree of accuracy do not yet exist, models drafted from experimental research are used. These models can be incorporated into the existing property clustering framework while providing additional benefits not found in the traditional experimental design. First, the experimental design points on which the model is based can be mapped into the property cluster space to determine if the feasibility region is completely explored as

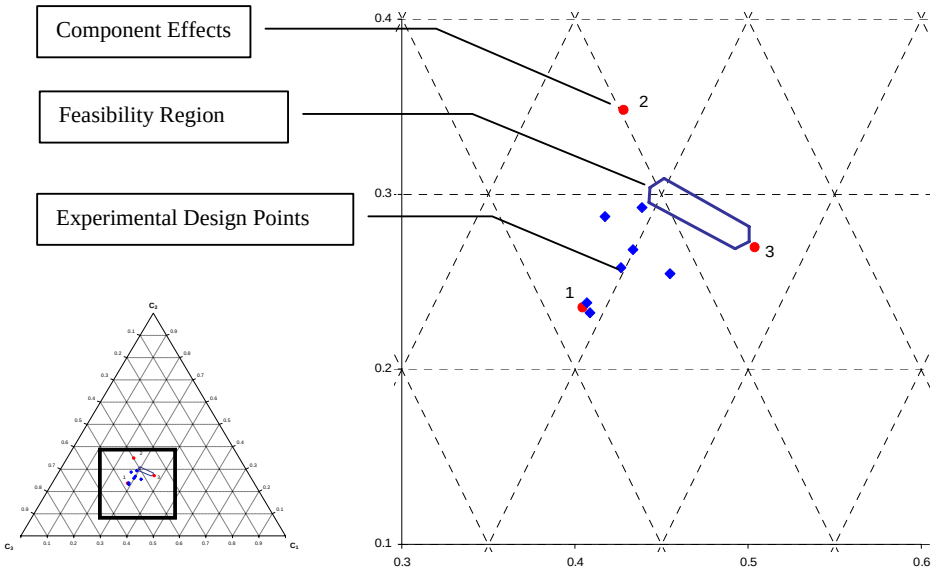


Figure 5: The experimental design points of a three-component mixture of polyethylene, polystyrene, and polypropylene mapped to property cluster space.

shown in figure 5. This visualization technique is appropriate regardless of the number of components investigated, thereby providing a means for ensuring the space has been properly explored.

Second, most experimental based models utilize regression coefficients as estimates of the effects of each property on the response. Depending on the type of regression utilized, the interpretation of these regression coefficients can be quite difficult. Visualizing the problem in property space consolidates each components impact on the mixture, aiding in the ability to screen components.

5.1. Visual Solution using Scheffe Models

Avoiding the interpretation of the regression coefficients, the Scheffe model can be used for modeling the design space. In fact the solution achieved by using the Scheffe and Cox models is the same even though the regression coefficients are different [9]. The reason for this result is that the regression coefficients represent different entities in each model. In the Scheffe model, the values of each constituent in the cluster space represents a combined effect comprised of contributions from the pure component property, collinear effects, and nonlinear effects. The Cox models remove some of the collinearities in mixture design, but depend on good experimental design controls to limit the effect of other collinearities and nonlinearities. This observation also explains why some regression coefficients can be negative. If the coefficients only represented pure component values, then the property would always be positive. It follows then that a negative regression coefficient is indicative of a strong collinearity and/or nonlinearity effect overwhelming the weak linear effect.

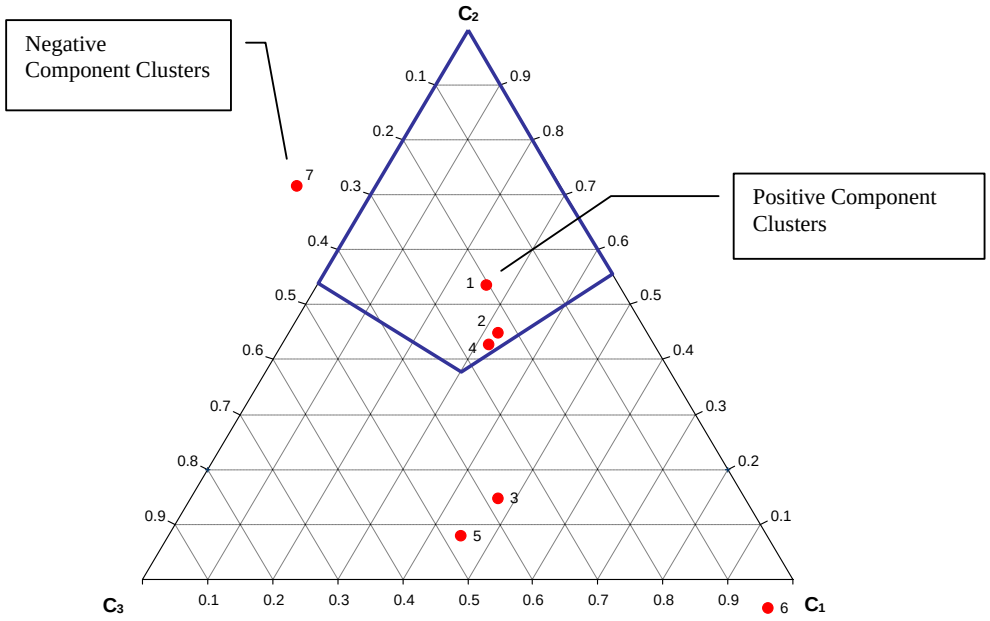


Figure 6: A property cluster simplex diagram of a seven-component system using property operators derived from an Acetaminophen excipient design [21]. Clusters 1, 2, and 3 represent the properties repose angle, water content, and compressibility.

This situation can cause problems within the traditional property clustering framework, but is not unsolvable. In figure 6 a ternary cluster diagram containing property models with negative regression coefficients is shown. In the situations with negative regression coefficients, the combined effect of the component resides outside the ternary diagram. Hence, it can be theorized that when using models comprised of regression coefficients, some solutions may require mixing with chemical constituents outside of the cluster space defined by the ternary diagram. The region outside of the cluster space is defined as negative cluster space and the region inside the simplex shall be referred to as positive cluster space. The total number of visual cluster regions, N_R, is a function of the number of properties, P.

$$N_R = 2P \tag{32}$$

The negative cluster region is comprised of six distinct regions for a three property solution. These regions are of two types: those with property clusters greater than one are called Type I regions and those with negative property clusters are called Type II regions.

$$C_{ki} > 1, \text{ Type I Regions} \tag{33}$$

$$C_{ki} < 0, \text{ Type II Regions} \tag{34}$$

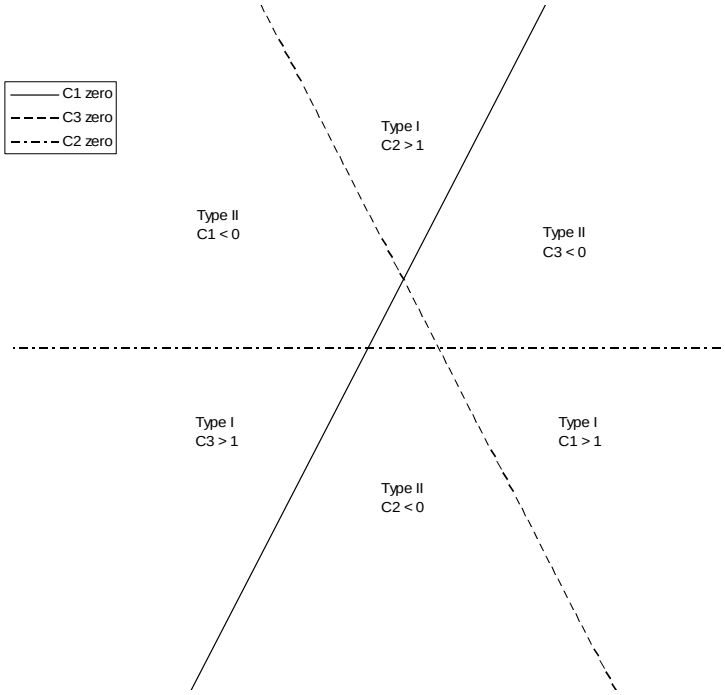


Figure 7: Details of negative property cluster space for a three-property system.

These regions are related to one another since in the derivation of property clusters, Eden notes that the clusters must sum to one [1].

$$\sum_{k=1}^{p} C_{ki} = 1 \tag{35}$$

For the three property example equation 35 implies that two negative property clusters equate to a third cluster with a value greater than one. Likewise, two clusters with a value greater than one equate to a third cluster with a value less than one. The number of different types of regions N_T are related to the number of properties evaluated according to equation 36.

$$N_T = P - 1 \tag{36}$$

For example, in the case of four properties, three types of negative cluster space regions would occur. The first two types plus a third type called Type III.

$$TypeIII = \left\{ \begin{matrix} C_{ki} > 1 \\ C_{li} > 1 \\ k \neq l \end{matrix} \right\} \tag{37}$$

Like the three property case before, the Type III region can also be determined from a combination of two clusters greater than one or less than one.

$$TypeIII = \left\{ \begin{matrix} C_{ki} < 1 \\ C_{li} < 1 \\ k \neq l \end{matrix} \right\} \tag{38}$$

Property clusters in the Type I and Type II regions may also be estimated from three negative clusters and three clusters greater than one, respectively.

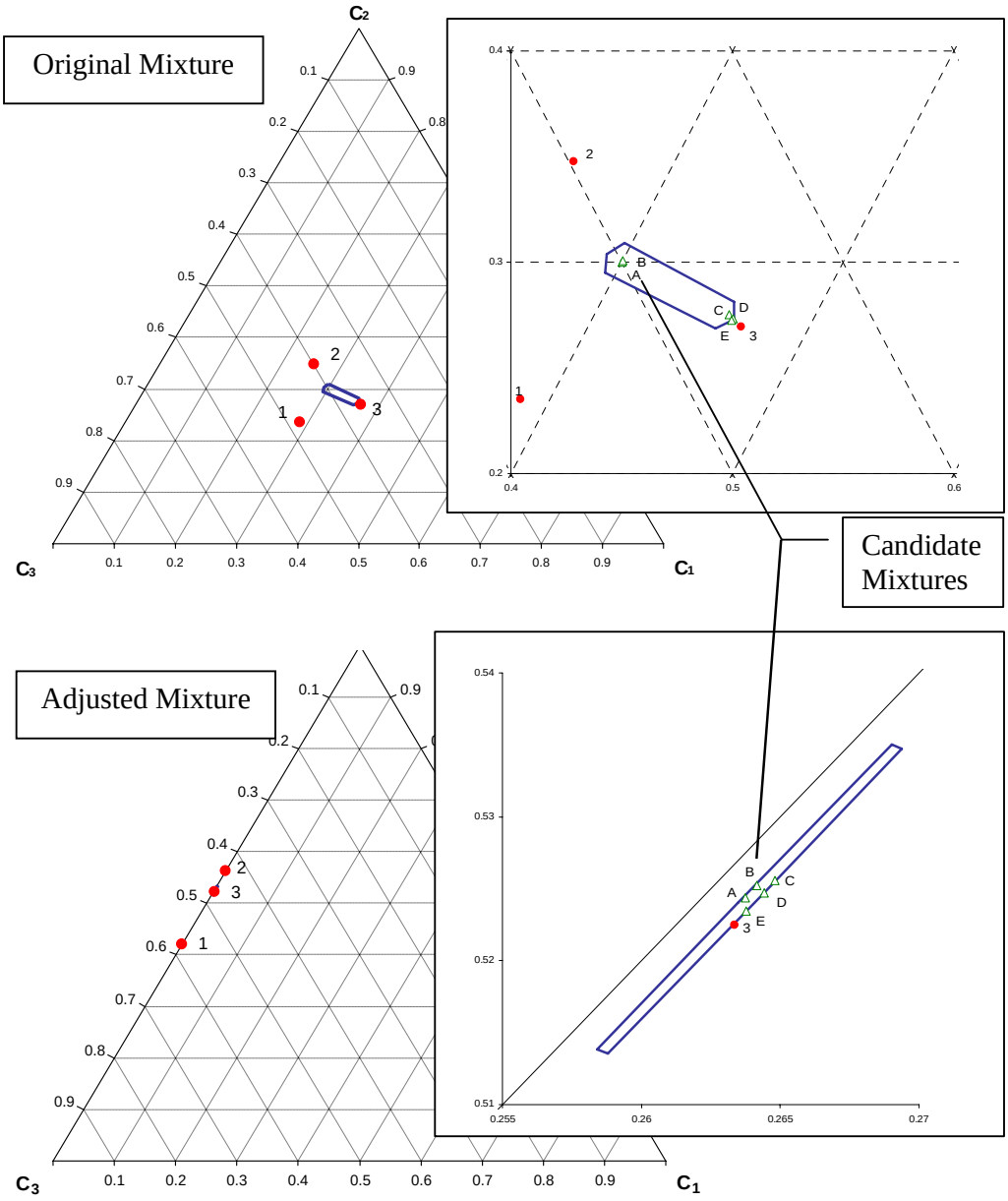
Also noteworthy when dealing with negative property clusters are the constraints on Augmented Property Index (*AUP*). In equation 28, the non-dimensionalized properties are summed to create the *AUP* [1, 21]. While a negative regression coefficient may create a negative property operator, the advent of such constructs must be constrained by the following rule:

Rule 4. All *AUP* values must be positive.

In equation form, the rule is written as equation 39.

$$AUP > 0 \tag{39}$$

Without the constraint of equation 39, a negative *AUP* is mathematically possible. When applied to the linear mixing rules, it gives an infeasible solution to the lever arm and violates inter-stream conservation as shown in equation 31. This situation is therefore avoided by the applying the constraint of equation 39. To ensure a positive *AUP* value, the property operator references are adjusted. Although the adjusted references give different values for the clusters, the underlying property values are unchanged. For example, the cluster diagram of the three component and three property mixture shown in figure 4 is altered by changing the reference values for the property operators, resulting in figure 8. Although the clusters of the three components reside in a different location, their relative proximity to each other and the feasibility region remains the same. Visual differences might change, but because the construct preserves the monotonically increasing rule, the comparative order of difference is the same. For instance, for the three component mixture of figure 8, component 3 (polypropylene) will always be closest to the feasibility region while component 1 (polyethylene) will always farthest from the feasibility region regardless of the property operator reference chosen. Furthermore, when candidate mixtures from the adjusted cluster diagram are transformed back to component space, the



Candidate Mixtures	Property Clusters						Component Fractions					
	Original Reference			Adjusted Reference			Original Reference			Adjusted Reference		
	C_1	C_2	C_3	C_1	C_2	C_3	x_1	x_2	x_3	x_4	x_5	x_6
A	0.3003	0.3007	0.3990	0.0015	0.5252	0.4732	0.178	0.508	0.314	0.177	0.510	0.313
B	0.3616	0.2748	0.3636	0.0020	0.5256	0.4724	0.171	0.513	0.316	0.171	0.514	0.315
C	0.3636	0.2734	0.3630	0.0021	0.5247	0.4732	0.000	0.083	0.917	0.000	0.078	0.922
D	0.3635	0.2726	0.3639	0.0021	0.5234	0.4745	0.000	0.057	0.943	0.000	0.057	0.943
E	0.3002	0.3002	0.3996	0.0015	0.5244	0.4741	0.010	0.051	0.940	0.010	0.050	0.940

Figure 8: A three component system of polyethylene, polystyrene, and polypropylene mapped to cluster space using different property operator reference values.

resulting solutions are identical to the candidate solutions derived using the original references. This is an important aspect of property clustering as the flexibility to adjust the property operator reference values maintains the positive *AUP* criteria of equation 39.

5.2. Visual Solution using Cox Models

The above discussion is a valid solution for mixing and developing list of candidate solutions. However, if the objective is to screen constituents, then some knowledge of the pure component effects will be needed in order to avoid additional experimentation each time a new constituent is added to the list of candidates. The goal of the screening design is to determine what characteristics in the constituents need to be avoided so that appropriate candidates can be selected in the future. Since the objective of the screening design is to understand the effects of the constituents on the mixture, then the component clusters need to represent the pure property values void of collinearities and nonlinearities. One method for removing the collinearities is to utilize the Cox modifications on the Scheffe canonical models. In this method the major collinearity introduced by equation 12 is removed. It has been shown by Cornell that the response surfaces for the Scheffe and Cox models are identical [9]. This result is also true of the property clustering solutions utilizing the Scheffe and Cox models. First it must be noted that the property operator equation now assumes the form of equation of equation 38 where

$$\sum_{i}^{p}\left(\frac{\beta_i}{1-s_i}\right)\Delta_i = \sum_{i}^{u}\beta_i x_i = y^z \tag{40}$$

Inserting equation 40 in equation 22, the property operator expression assumes the form of equation 41.

$$y(x) = y(s) + y^z \tag{41}$$

Where *y(s)* is the response at the standard mixture and y^z is the pseudo property value that represents the property, *k*, contribution to the mixture. It has been shown that the response at the standard mixture is equivalent to β_o regressor [9]. The resulting property operator expression is then normalized by dividing by the reference property operator as shown in equation 42.

$$\Omega_k = \Omega_k^s + \Omega_k^z \tag{42}$$

Likewise the augmented property index is rewritten as equation 43.

$$AUP = AUP^s + AUP^z \tag{43}$$

Redefining the property cluster of equation 29 in terms of the pseudo property cluster C_k^z and the standard property cluster C_k^s gives equations 44 and 45.

$$C_k^z = \frac{\Omega_k^z}{AUP^z} \tag{44}$$

$$C_k^s = \frac{\Omega_k^s}{AUP^s} \tag{45}$$

The sum of these clusters will not give the true cluster without correcting for the different *AUP* values. This is done using a set of correction factors as shown in equations 46 and 47.

$$F^z = \frac{AUP^z}{AUP} \tag{46}$$

$$F^s = \frac{AUP^s}{AUP} \tag{47}$$

F^z is the pseudo correction factor and F^s is the standard correction factor. These are combined in equation 48 to give the relationship between the true property cluster, the pseudo property cluster, and the standard property cluster.

$$C_k = F^s C_k^s + F^z C_k^z \tag{48}$$

Rewriting equation 48 in terms of the mixture with pseudo properties and standard properties gives equation 49.

$$C_{kM} = \frac{\sum_i^u x_i \Omega_{ki}^z}{AUP_M} + \frac{\Omega_k^s}{AUP_M} \tag{49}$$

Rewriting the AUP of the mixture in terms of the correction factor gives equation 50.

$$C_{kM} = \frac{\sum_i^u x_i \Omega_{ki}^z}{AUP_M} + \frac{\Omega_k^s}{AUP_M^s} \cdot F_M^s \tag{50}$$

Inserting the cluster definition of equation 45 in equation 50 and rearranging gives equation 51.

$$C_{kM} - F_M^s C_{kM}^s = \frac{\sum_i^u x_i \Omega_{ki}^z}{AUP_M} \tag{51}$$

Noting that the left hand side of equation 51 is the same as the product of the pseudo correction factor and the pseudo property cluster of the mixture, the equation is rewritten as equation 52.

$$F_M^z C_{kM}^z = \frac{\sum_i^u x_i \Omega_{ki}^z}{AUP_M} \tag{52}$$

Inserting the pseudo cluster definition to remove property operator in favor of the cluster and rewriting the *AUP* of the mixture in pseudo terms gives equation 53.

$$C_{kM}^z = \frac{\sum_i^u x_i AUP_i^z C_{ki}^z}{AUP_M^z} \tag{53}$$

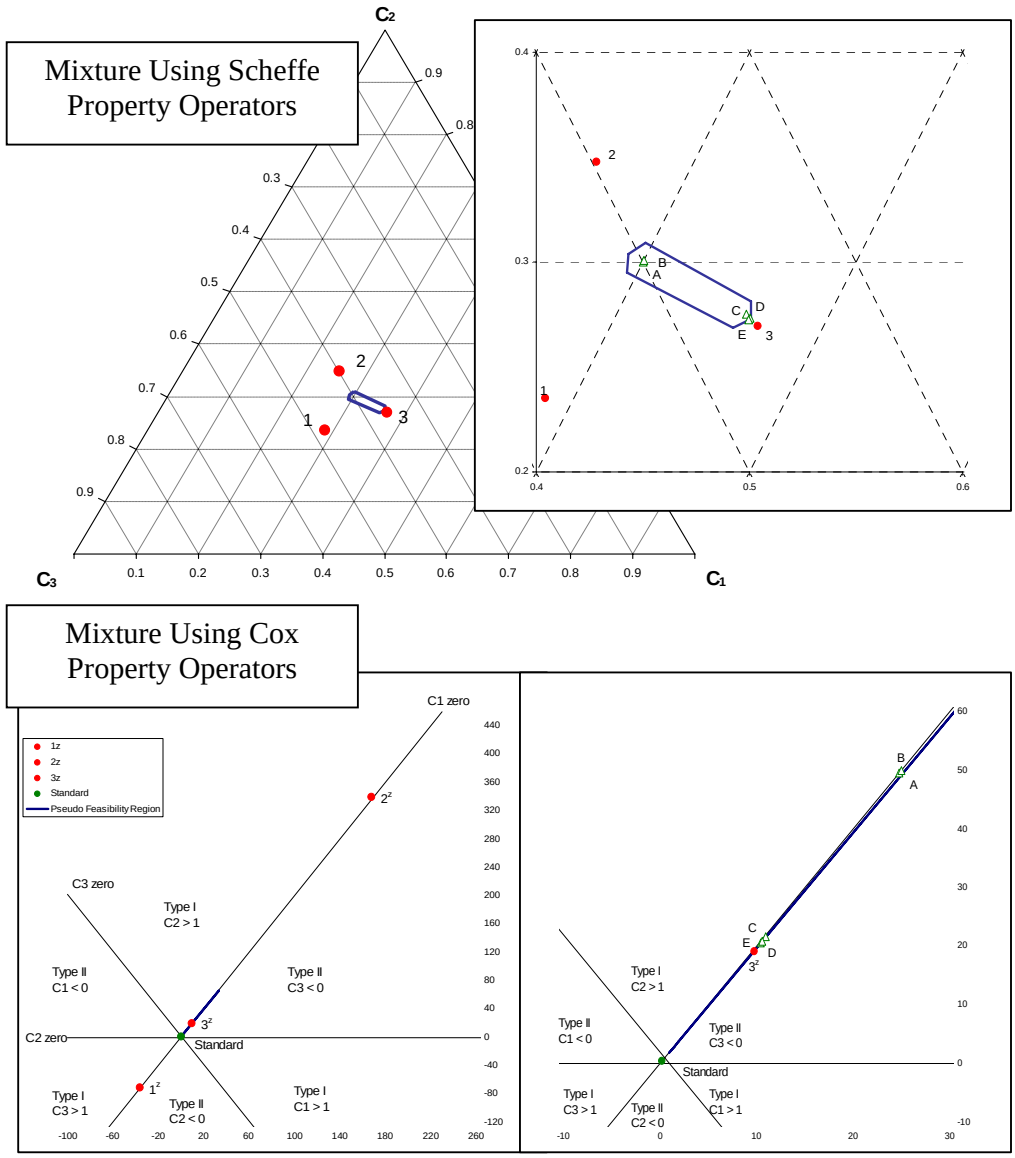
Equation 53 assumes a familiar form of the relative cluster arm using the pseudo relative cluster arm δ^z as defined by equation 54.

$$\delta^z = \frac{x_i AUP_i^z}{AUP_M^z} \tag{54}$$

The pseudo relative cluster arm maintains the monotonically increasing criteria imposed by Eden et al. provided that the all of the augmented property indices used in the solution of the problem are positive; a constraint placed on the solution by equation 39 [1]. A negative *AUP* would violate this relationship and prevent the proper cluster solution form being obtained. Combining equations 53 and 54 into the final form for mixing pseudo clusters gives equation 55.

$$C_{kM}^z = \sum_i^u \delta^z C_{ki}^z \tag{55}$$

Visually inspecting the pseudo mixing rule in figure 9 shows that the relative cluster arms are indicative of a mix involving a pseudo feasibility region. The pseudo feasibility region is defined in the same manner as the true feasibility region but corrects for the standard mixture property values using equation 48. The resulting pseudo relative cluster arms represent the addition of the pseudo components to move the mixture into this region, only now each pseudo component better represents its contribution to the mixtures' properties, void of most collinearities. It can be seen in



Candidate Mixtures	Property Clusters						Component Fractions					
	Scheffe Property Operator			Cox Property Operator			Scheffe Property Operator			Cox Property Operator		
	C_1	C_2	C_3	C_1	C_2	C_3	x_1	x_2	x_3	x_1	x_2	x_3
A	0.300	0.301	0.399	-0.010	49.548	-48.538	0.177	0.510	0.313	0.177	0.510	0.313
B	0.362	0.275	0.364	-0.010	50.043	-49.032	0.171	0.514	0.315	0.171	0.514	0.315
C	0.364	0.273	0.363	0.160	21.595	-20.755	0.000	0.078	0.922	0.000	0.078	0.922
D	0.364	0.273	0.364	0.163	20.908	-20.071	0.000	0.057	0.943	0.000	0.057	0.943
E	0.300	0.300	0.400	0.163	20.534	-19.697	0.010	0.050	0.940	0.010	0.051	0.939

Figure 9: A three component system of polyethylene (x_1), polystyrene (x_2), and polypropylene (x_3) mapped to cluster space using both Scheffe and Cox property operator models.

This is expected since the normal Scheffe and Cox mixing designs achieve the same property response plots [9]. Hence, for ease of use, the traditional property clustering method should be used with the Scheffe property operators, while the pseudo property clustering method using Cox property operators should be reserved for interpretation of component effects on the mixture properties. An interesting benefit of using pseudo property clustering with Cox models over the traditional method is the ability to visualize the components impact on the mixtures properties simultaneously. In the traditional techniques, each response plot is mapped onto the component space. Ignoring the combinatorial explosion issue for a moment, it can be seen that when two or more iso-properties are parallel or conflict in direction, it can be difficult to know where to move the mixture visually. This problem is compounded exponentially when multiple components are evaluated, leading researchers to either limit the number of components in the experiment or use powerful statistical techniques such as PLS. Pseudo property clustering offers a medium ground between the two methods and in some cases can be used in conjunction with PCR and PLS to further clarify solutions, especially when performing screening designs.

Rules governing the interpretation of the cluster points in the property cluster space are listed as follows:

Rule 5. The visual distance from the standard reference mixture to a component cluster point is indicative of the magnitude of the components effect on the response.

Rule 6. If the constituents lie on opposite sides of a line which passes through the standard reference mixture, then the constituents are said to be inversely related.

6. Case Study – Polymer Blend for Spun Yarn

This case study is a combination of selected illustrative examples presented by Cornell [9]. Two of the three properties are taken directly from Cornell's work. The first of the two properties is thread elongation of spun yarn. The second is knot strength. Both of these properties are important attributes of fibers used for high tech rope selection in modern racing sailboats. A third attribute added to the mix is density since flotation of rope in marine applications can prevent technical maneuvers from becoming catastrophic. Other properties such as abrasion resistance, stiffness, or breaking strength may have been used, although the breaking strength would not have been completely independent of knot strength and would have needed to be handled accordingly.

With the three properties chosen, the important step of choosing the property operator expressions comes next. Using the mixture data for a three component system in figures 1 and 2, linear Scheffe and Cox models were developed for thread elongation and knot strength. For the third property, density, no experimental data was gathered. However, a pure component property operator model was previously developed and will be applied here [1].

$$y_1 = 11.70x_1 + 9.40x_2 + 16.40x_3 \quad (56)$$

$$y_2 = 9.59x_1 + 12.85x_2 + 11.98x_3 \quad (57)$$

$$y_3 = 1.30x_1 + 0.98x_2 + 1.07x_3 \quad (58)$$

Where the chemical constituents x_1, x_2, and x_3 represent polyethylene, polystyrene, and polypropylene, respectively. Using the traditional mixture design simplex, the chemical constituents are placed at the vertices and each of the three properties are mapped to the design space along with the feasibility region which is set by product targets. The interpretation of the design is difficult since knot strength and density are nearly mutually exclusive. This means that the influence of one chemical constituent may have competing effects on the mixture response; adversely effecting one property to the benefit of another. Likewise, should an additive be chosen to supplement the design, an additional figure would be required to determine its impact and relationship to the previous design. This is a less than ideal situation. To prevent this combinatorial explosion and to provide an easier method of which to examine the impact of components on all the properties simultaneously, the property operator models are analyzed using property clustering. First the property operator models are non-dimensionalized using a set of references chosen to assure a positive *AUP*. The resulting property clustering diagram is shown in figure 4 with the vertices representing each of the three properties in their cluster forms. As was shown in the traditional simplex, the third chemical constituent, polypropylene, is closest to the feasibility region, followed by polystyrene and then polyethylene. It is now also clear that the addition of polyethylene and polystyrene have a much greater impact on the properties of the mixture than polypropylene, suggesting that polypropylene should be used as a filler. Of the two remaining polymers, polyethylene appears to have a larger impact on the mixture properties. However, since the properties were derived using Scheffe models, inherent collinearities exist in the property operator models. To circumvent this problem, the Scheffe models are reparameterized as Cox models using the methods outlined by Cox [7] where the standard reference mixture is at location (0.653, 0.173, 0.173) as provided by Cornell [9].

$$y_1 = 12.10 - 0.40x_1 - 2.70x_2 + 4.30x_3 \quad (59)$$

$$y_2 = 10.56 - 0.97x_1 + 2.29x_2 + 1.42x_3 \quad (60)$$

$$y_3 = 1.204 - 0.096x_1 - 0.223x_2 - 0.133x_3 \quad (61)$$

Separating the models into standard and pseudo clusters results in figure 9. The pseudo mixtures represent the mixtures void of most collinearities and better represent each chemical constituents impact on the mixture properties. Here it is confirmed that polypropylene has the smallest effect on the combined mixture properties. However, by removing most of the collinearity in the model, the result now suggests that polystyrene has the strongest effect on the combined mixture properties and that polystyrene and polyethylene have inverse effects. From the figure it can be seen that

polystyrene has the strongest effect on density (C_3) and thread elongation (C_1), both of which are negative effects. It also has the strongest effect on knot strength (C_2), which is a positive effect. Conversely, polyethylene has the weakest effect on thread elongation, a smaller negative effect on knot strength, and a smaller positive effect on density.

To evaluate the candidate solutions for the design it is necessary to create a pseudo feasibility region as shown in figure 9. The feasibility region matches the true feasibility region when corrected with the standard reference mixture. Using linear mixing rules a list of candidate solutions was found to match the candidate solutions found using Scheffe derived property operator models. Since the resulting candidate solutions are the same using either the pseudo cluster method or the true cluster method, then the true cluster method with Scheffe models should be used to determine candidate solutions because it is easier to visualize. If, however, insights regarding the impact of each of the chemical constituents are sought, then the pseudo cluster method with Cox models should be used.

By evaluating the placement of the experimental design points in the property cluster space, insights into the effectiveness of the design are also gained. In figure 5 the experimental design points are translated to the property cluster space. Unfortunately, the design points are all outside the feasibility region and none of the candidate mixtures fall between the design points. This is the same inference made when investigating figure 2. This means that to obtain the candidate solutions the property operator models must be extrapolated, which introduces unneeded error into the solution and suggests an insufficient design. To prevent this situation, the design points should be repositioned so that they cover the feasibility region. The procedure for executing the repositioning must take into consideration the increase in accuracy of the property space at the expense of the optimality of the component space. The added benefit by viewing the design in property space is that the points may always be viewed on a single diagram, regardless of the number of components studied as long as the number of properties measured are three or less. In cases of three or more properties are studies, additional diagrams may be used or algebraic methods applied [20].

7. Conclusions

In this work, a systematic property based framework for solution of mixture design problems using property clustering has been presented. The recently introduced property integration framework has been extended to include experimentally derived property operator models: specifically first order Scheffe canonical and Cox polynomial models. When interpretation of the chemical constituent's impact on the mixture property is warranted, Cox derived property operator models are utilized such that the location of the pseudo chemical constituent relative to the standard reference mixture is indicative of its impact on the mixture's properties. The accuracy of the design is visually observed by placing the design points in the property design space.

A significant result of the developed methodology is that for problems that can be satisfactorily described by just three properties, the experimental mixture design problems are analyzed visually on a simplex diagram, irrespective of how many chemical constituents are included in the search space. However, algebraic and optimization based approaches can easily extend the application range to include more properties.

8. References

1. Eden M.R., Jørgensen S.B., Gani R., and El-Halwagi M.M. (2003) Reverse Problem Formulation based Techniques for Process and Product Design. *Computer Aided Chemical Engineering*, 15A, 451-456.
2. Harper P.M. and Gani R. (2000) A multi-step and multi-level approach for computer aided molecular design. *Computers & Chemical Engineering*, 24(2-7), 677-683.
3. Marcoulaki E.C. and Kokossis A.C. (1998) Molecular design synthesis using stochastic optimisation as a tool for scoping and screening. *Computers & Chemical Engineering*, 22(Supplement 1), S11-S18.
4. Eljack F.T., Eden M.R., Vasiliki K., Kazantzi Q., and El-Hawagi M.M. (2007) Simultaneous process and molecular design - A property based approach. *AIChE Journal*, 53(5), 1232-1239.
5. Scheffe H. (1958) Experiments with mixtures. *J. R. Stat. Soc. B.*, 20(2), 344-360.
6. Scheffe H. (1963) Simplex-centroid design for experiments with mixtures. *J. R. Stat. Soc. B.*, 25(2), 235-263.
7. Cox D.R. (1971) A note on polynomial response functions for mixtures. *Biometrica*. 58(1), 155-159
8. Box G., Hunter W., and Hunter J. *Statistics for Experimenters*. John Wiley & Sons, New York, NY (1978).
9. Cornell J.A. *Experiments with Mixtures*. John Wiley & Sons, New York, NY (2002).
10. Kettanah-Wold N. (1991) Use of experimental design in the pharmaceutical industry. *Journal of Pharmaceutical and Biomedical Analysis*, 9(8), 605-610.
11. Larsen, P.V. (2003) *ST111: Regression analysis and analysis of variance.* (**http://statmaster.sdu.dk/courses/st111/**)
12. Geladi P. and Kowalski B.R. (1986) Partial Least Squares Regression: A Tutorial. *Analytica Chimica Acta*, 185, 1-17.
13. Kramer R. *Chemometric Techniques for Quantitative Analysis*. Marcel Dekker, New York, NY (1998).

14. Meyer R.K. and Nachtsheim C.J. (1995) The Coordinate-Exchange Algorithm for Constrtucting Exact Optimal Experimental Designs. *Technometrics*, 37, 60-69.

15. Eriksson L., Johansson E., Wikstrom C. (1998) Mixture design – desgn generation, PLS analysis, and model usage. *Chemometrics and Intelligent Laboratory Systems*, 43, 1-24.

16. Montgomery D.C. and Meyers R.H. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments.* John Wiley & Sons, USA (1995).

17. Kettanah-Wold N. Analysis of mixture data with partial least squares. (1992) *Chemometrics and Intelligent Laboratory Systems*, 14, 605-610.

18. Smith W.F. and Beverly T.A. (1997) Generating linear and quadratic Cox mixture models. *J. Qual. Technol*., 29, 211-224.

19. Shelley M.D. and El-Halwagi M.M. (2000) Component-less design of recovery and allocation systems: a functionality-based clustering approach. *Computers & Chemical Engineering*, 24(9-10), 2081-2091.

20. Qin X., Gabriel F.B., Harell D.A., and El-Halwagi M.M. (2004) Algebraic Techniques for Property Integration via Componentless Design. *Industrial and Engineering Chemistry*, 43, 3792-3798.

21. Martinello T., Kaneko T.M., Velasco M.V.R., Taquedo M.E.S., and Consiglieri V.O. (2006) Optimization of poorly compactable drug tablets manufactured by direct compression using the mixture experimental design. *Internation Journal of Pharmaceutics*, 322(1), 87-95.