Interactive Exploration of Ingredient Mixtures Using Multiple Coordinated Views

Sara Johansson*, Kristina Knaving*, Amanda Lane†, Mikael Jern* and Jimmy Johansson*

* Linköping University, Dept of Science and Technology, Sweden
† Unilever R&D, Port Sunlight, United Kingdom
sara.johansson@itn.liu.se, krikn195@student.itn.liu.se, amanda.lane@unilever.com
mikael.jern@itn.liu.se, jimmy.johansson@itn.liu.se

Abstract
The complex nature of multivariate data sets calls for high interactive performance and intuitive metaphors. A specific type of multivariate data is where the variables sum up to a constant, here defined as multicomponent data. This application paper presents an interactive application for analysis of modelled multicomponent data. The aim is to find high performance variable combinations that fulfill some requested properties. The application is based on coordinated views that include parallel coordinates, a ternary diagram, a 2D scatter plot and a line plot. It supports numerous interaction techniques enabling fast analysis of complex patterns in multicomponent data sets. The application is developed in collaboration with researchers within the fields of statistics and chemistry. An informal usability evaluation indicates that the interactive nature of the application clearly facilitates the analysis process.

Keywords—Information visualization, multicomponent data, ternary diagram, interactivity, modelled data.

1 Introduction
Over the past decades multivariate data sets have become increasingly common in a variety of domains. Major challenges are to extract meaning, discover structure and find patterns. Improvements of existing visualization and interaction methods hold great potential to provide valuable and previously unknown information that can advance our understanding of complex phenomena and systems.

This application paper presents a tool for analysis of multicomponent data, which is here defined as a specific type of multivariate data where the variables sum up to a constant. This type of data is common in areas such as chemistry, statistics, mineralogy and biology. The focus of this paper is on ingredient mixture data which is multicomponent data that includes ingredient variables and performance variables. For this data the goal of the analysis is to find mixing proportions that yield a desirable performance and at the same time generate minimum variability due to the influence of uncontrollable variables, referred to here as process variables.

To be able to investigate the trade-offs needed to find suitable ingredient mixtures, taking all aspects into consideration, specialists are needed who are aware of the restrictions in specific situations. With large amounts of numerical data this is not easily done, however. Interactive visualization tools can facilitate the analysis considerably by:

- Providing an overview of large amounts of data;
- Taking advantage of the specialists’ knowledge to investigate trade-offs and determine optimum solutions;
- Speeding up investigation, decision making and product development by supplying interaction possibilities.

This paper describes an interactive visualization tool for multicomponent data called VIME (Visual Ingredient Mixture Exploration). VIME incorporates traditional information visualization techniques, ternary diagrams and a number of interaction techniques to discover optimal mixtures of ingredients. Ternary diagrams are extensively used in disciplines such as chemistry [16] and mineralogy [14] to analyse mixtures of ingredients or mineral compositions, but are less common in information visualization. The ternary diagram in VIME is extended with several interaction techniques to further enhance the ability to analyse mixtures.

The application development has taken place in close collaboration with domain experts in the areas of statistics and chemistry. VIME has been developed using the Geo-Analytics Visualization (GAV) framework [10].

The outline of the paper is as follows. In section 2 the background of this paper and previous research related to
2 Background and Related Work

Exploration of ingredient mixture data differs from exploration of common multivariate data sets due to the internal relationship between component variables and the specific goal of the analysis. The data visualized in VIME contains ingredient variables as well as performance attributes. The performance of different mixtures can vary due to ingredient proportions and process variables, such as consumer habits and environmental conditions. The goal of the analysis is to find compositions that yield a desirable result and, at the same time, generate minimum variability due to the influence of process variables.

Traditionally, multicomponent data has been visually analysed by using static diagrams such as ternary diagrams. A ternary diagram is a plot of three variables in the shape of an equilateral triangle, where each corner of the triangle represents 100 percent of one of the variables. Every point in the triangle represents a unique composition of the three variables which sums up to 100 percent. Figure 1 shows an example of how to interpret the variable composition of a point in a ternary diagram. To read the proportion of variable \( A \) at the point in figure 1 a line, \( L_A \), which is parallel to the triangle side opposite to \( A \)'s corner, is drawn through the point. The proportion of variable \( A \) at that point is given by the shortest distance, \( P_A \), between \( L_A \) and the triangle side opposite to the variable corner [3]. In the same way the proportion of variable \( C \) at the point in figure 1 is read using the line \( L_C \) and distance \( P_C \). In this example the point contains approximately 50 percent of \( A \) and 15 percent of \( C \), hence the amount of variable \( B \) is approximately 35 percent.

The use of equilateral triangles to represent mixtures can be traced back to post-Newtonian work on mixing colours [14], and in 1897 Bancroft [3] suggested several ways to use the characteristic of an equilateral triangle to calculate the proportions of a point inside it.

Ternary diagrams are commonly used in several research areas, for example in mineralogy, which includes studies of economic minerals, petrology, classification and phase relationships [14], in chemistry [16, 6] and in computer science [5]. An example of ternary diagrams being used for visualization of mixtures is given in [13] where properties of glass produced from waste streams, which depend upon waste stream composition, are visualized.

Using static representations, such as ternary diagrams, makes the analysis process both difficult and time-consuming since the number of mixtures which need to be analysed is large and often contains many ingredient variables. Having the ability to interactively construct visual representations of ingredient variables will significantly speed up the analysis process since the user can freely manipulate the visual representations to suit their needs. Using multiple linked views, which display different aspects of the data simultaneously, is a well established concept [12] that would further enhance the analysis process.

To the best of our knowledge no application exists that provides this interactivity and, at the same time, can be easily adapted to the specific structure of ingredient mixture data in a way that makes use of the knowledge and previous experiences of domain experts. What does exist, however, are numerous applications and toolkits that can be used for interactive analysis of multivariate data sets, some examples are Spotfire [1], InfoVis Toolkit [7], GAV [10] and XmdvTool [19]. These are all built on the principle of coordinated and linked views [2, 4]. Coordinated and linked views have been successfully used to overcome the difficulty of presenting large amounts of data in one display and concurrently make it possible to find detailed structures. Some statistical tools, such as [8], provide traditionally used visual representations for multicomponent data, but lack the possibility of coordination of multiple views.

The work presented in this paper uses interactive ternary

---

**Figure 1:** A ternary diagram plotting ingredients \( A \), \( B \) and \( C \). The ingredient proportions of a point are read using straight lines that are parallel to the triangle sides. As an example, looking at variables \( A \) and \( C \) the red point shown contains 50 percent of ingredient \( A \) (as shown by the line \( L_A \) and distance \( P_A \)) and 15 percent of ingredient \( C \) (as shown by the line \( L_C \) and distance \( P_C \)).
diagrams together with more common information visualization techniques such as parallel coordinates [9, 20], scatter plots and line plots in a coordinated and customised environment to enable interactive exploration of multicomponent data.

3 Data Modelling

The multicomponent data used in this work is calculated using a regression model; a full second-order model (Scheffé [15]). This type of regression model is commonly used in robust formulation optimisation [17, 11]. The data used in this paper, having five ingredients and two process variables, is structured as follows:

\[ \eta(x, z) = \sum_{i=1}^{5} \gamma_i^0 x_i + \sum_{i=1}^{5} \sum_{j>i}^{5} \gamma_{i,j}^0 x_i x_j + 2 \sum_{k=1}^{2} \left[ \sum_{i=1}^{5} \gamma_i^k x_i + \sum_{i=1}^{5} \sum_{i<j}^{5} \gamma_{i,j}^k x_i x_j \right] z_k + \sum_{i=1}^{5} \gamma_i^{1,2} x_i + \sum_{i=1}^{5} \sum_{i>j}^{5} \gamma_{i,j}^{1,2} x_i x_j \]

where \( x_i \) and \( x_j \) are the proportions of ingredients \( i \) and \( j \), respectively. \( z_1 \) and \( z_2 \) are process variables and \( \gamma \) is a parameter estimate. Varying the sampling level of the ingredients results in data sets of different resolutions.

In this paper data sampled at either 10 or 5 percent is used. Based on equation 1 two models were extracted providing values for the mean and variance. The mean value is a measure of how effective a specific combination of ingredients are, and the variance a measure of the variation due to the uncontrollable variables. Sampling with a resolution of 10 percent provides 1001 valid mixtures that need to be analysed. Changing the sampling resolution to 5 results in 10626 valid mixtures, more than ten times as many.

4 Application

VIME is a tool for exploration of the performance of ingredient mixtures. The graphical user interface of VIME is displayed in figure 2 and the following sections describe its features in further detail.

4.1 Pre-processing

The ternary diagram is used to display variables summing up to a specific constant, typically 100 percent, and, hence, is well suited for representation of multicomponent data. However, it is only able to display three ingredients (components) at a time if the proportions of all are to be read individually. For representation of data containing more than three ingredients in the ternary diagram in VIME, the data is reduced to subsets containing two ingredient variables and a variable representing the sum of the remaining ingredients. For every combination of two ingredients, there is a number of items with varying values for the remaining ones, as shown in table 1 where all mixtures containing 80 percent of \( \text{Ing}_1 \) and 10 percent of \( \text{Ing}_2 \) are shown.

The first step in creating subsets is to average the performance variables (\( \text{Mean} \) and \( \text{Variance} \)) in table 1 into a new variable \( \text{Ing}_{3+4+5} \) by taking the sum of the individual values, see table 2. The result is a single mixture containing ingredient components \( \text{Ing}_1, \text{Ing}_2 \) and \( \text{Ing}_{3+4+5} \) together with their average performance values.

<table>
<thead>
<tr>
<th>( \text{Ing}_1 )</th>
<th>( \text{Ing}_2 )</th>
<th>( \text{Ing}_{3+4+5} )</th>
<th>\text{Mean}</th>
<th>\text{Variance}</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>0.1</td>
<td>0.1</td>
<td>13.26</td>
<td>3.02</td>
</tr>
<tr>
<td>0.8</td>
<td>0.1</td>
<td>0.1</td>
<td>11.87</td>
<td>1.04</td>
</tr>
<tr>
<td>0.8</td>
<td>0.1</td>
<td>0.1</td>
<td>17.75</td>
<td>1.86</td>
</tr>
</tbody>
</table>

Table 1: All ingredient mixtures with 80 percent of \( \text{Ing}_1 \) and 10 percent of \( \text{Ing}_2 \) in a data set sampled at every ten percent. Each row corresponds to a specific ingredient mixture.

Table 3: The last step in creating subsets is to average the performance variables (\( \text{Mean} \) and \( \text{Variance} \)) in table 2. The result is a single mixture containing ingredient components \( \text{Ing}_1, \text{Ing}_2 \) and \( \text{Ing}_{3+4+5} \) together with their average performance values.

<table>
<thead>
<tr>
<th>( \text{Ing}_1 )</th>
<th>( \text{Ing}_2 )</th>
<th>( \text{Ing}_{3+4+5} )</th>
<th>\text{Mean}</th>
<th>\text{Variance}</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>0.1</td>
<td>0.1</td>
<td>14.29</td>
<td>1.97</td>
</tr>
</tbody>
</table>

Table 2: The result of combining \( \text{Ing}_3, \text{Ing}_4 \) and \( \text{Ing}_5 \) in table 1 into a new variable \( \text{Ing}_{3+4+5} \). The result is a single mixture containing ingredient components \( \text{Ing}_1, \text{Ing}_2 \) and \( \text{Ing}_{3+4+5} \) together with their average performance values.

4.2 Visual Representations

VIME is made up of four views; a robustness plot, a ternary diagram, a performance plot and a parallel coordinates representation, as shown in figure 2. The parallel coordinates representation (bottom view in figure 2) provides an overview of the whole modelled data set, containing all ingredient variables as well as performance variables. A
The layout and visual representations of VIME. The top left view contains a robustness plot, displaying Mean against Variance. In the top centre view is a ternary diagram representing Ing1 and Ing5. This diagram is coloured using bi-variate colouring as described in detail in section 4.3. To the top right is a performance plot representing the average performance of different proportions of Ing1 and Ing5 together. In the bottom half of the application the whole modelled data set is displayed using parallel coordinates.

Ten percent sampling resolution (1001 items) is used to prevent visual clutter. The parallel coordinates representation also serves as a visual interface to control what is displayed in the other views.

The ternary diagram (top, centre view in figure 2) effectively displays compositions of ingredients, and since every point in the diagram represents a possible ingredient mixture, high resolution data can be visualized without generating clutter. Although a ternary diagram can display data with very high resolution, the data used in this application is sampled using a 5 percent resolution. To still make use of the full triangular area and to facilitate discovery of trends, all pixels are coloured according to the Mean and Variance of the sample points closest to them, calculated using Euclidean distances between pixels and sample points. The colouring of the ternary diagram is shown in figure 3 and described in detail in section 4.3.

The robustness plot (top, left view in figure 2) is a scatter plot displaying Mean against Variance for all ingredient mixtures that are selected in the parallel coordinates representation or in the ternary diagram. If a point is selected in the ternary diagram then each ingredient mixture that this point is made up of is represented as a glyph in the robustness plot. In figure 2, three points close to the right corner in the ternary diagram, representing Others, have been selected. As described in section 4.1 these points are made up from a number of unique ingredient mixtures. All ingredient mixtures from which the three selected points are made, are displayed as individual glyphs in the robustness plot in figure 2. Through this the robustness plot provides a focused view of the performance distribution of selected ingredient combinations.

In the performance plot (top, right view in figure 2) the analyst can focus on the performance of certain ingredients or combinations of ingredients. In the figure the combination of Ing1 and Ing5 is displayed. The performance plot
Figure 3: Colour is used to display both *Mean* and *Variance* in the ternary diagram. The left view displays *Mean* using a colour scheme where the hue ranges from blue to red with green and yellow in between. In the centre view *Mean* and *Variance* is displayed simultaneously, where opacity ranging from fully opaque to completely transparent is used to represent *Variance*. In the right view *Variance* is displayed using grey scale. The left colour legend, representing *Variance*, is always mapped to the grey scale colour scheme. To make use of the full triangle area of the ternary diagram all pixels are coloured according to the colour of the sample point closest to the pixel.

is a line plot that displays the average performance (*Mean*) at different summed proportions of the selected ingredients. The proportions are represented along the x-axis. An average performance, represented by the dark blue line, is computed from all ingredient mixtures where the selected ingredients together make up the given proportion. The individual mixtures that the average is computed from are represented by small red dots in the plot. As an example, the points positioned along the vertical grid line which correspond to the proportion 0.5, represent the performance of ingredient mixtures where *Ing*$_1$ and *Ing*$_5$ together make up 50 percent of the mixture. The position of the blue line at the 0.5 proportion corresponds to the average performance value of the ingredient mixtures.

4.3 Colouring

The aim of the colouring in VIME is to be able to represent both performance variables (*Mean* and *Variance*) simultaneously. This is achieved using bi-variate colouring, based on the principles presented in [18]. According to these the selected colours should be perceived to maintain order and to be different when pointing out data differences. Furthermore the two colours should never obscure each other.

In VIME, hue is used to describe the *Mean* variable and opacity to describe the *Variance* variable. Both hue and opacity preserve the order of the variables, and clearly point out differences between high and low values. Figure 3 displays the different colouring available for the ternary diagram in VIME. The left-hand diagram is coloured according to the *Mean* variable, with hue ranging from blue to red with green and yellow in between. The centre diagram is coloured according to the *Mean* and *Variance* simultaneously using opacity ranging from fully opaque to completely transparent to represent the *Variance*. As can be seen from the centre diagram, the top corner of the diagram (representing mixtures with high amounts of *Ing*$_3$) is more transparent than the other parts of the diagram, and so indicates a high variance for ingredient mixtures containing mostly *Ing*$_3$. Since there might be some difficulty assessing two colour spans at once, opacity can be displayed separately in the ternary diagram using a grey scale, as shown in the right-hand diagram in figure 3.

4.4 Interactivity and Coordination of Views

The views of VIME are coordinated by colour, selection and thresholding. Any action performed in one view is immediately reflected in all of the others. All views display different aspects and samplings of the modelled data. The parallel coordinates display all ingredients and performance variables, the ternary diagram displays subsets of two ingredients, the robustness plot displays the mean and variance of selected ingredient mixtures and the performance plot shows average performance against proportions of selected ingredients. Due to this diversity the linking between the different subsets and views is complex.

Thresholding, for example based on unwanted performance properties or restrictions on certain ingredients, can be performed by using the threshold sliders of the parallel coordinates axes. The thresholding is reflected in the robustness plot and, if performed on performance variables or ingredients of the current subset, the thresholding is also reflected in the ternary diagram. In the performance plot the thresholding is reflected by a re-computation of average performance for selected variables, displaying only the performance of data items that are within the thresholds in the parallel coordinates representation.

The data subsets displayed in the ternary diagram have
Figure 4: An illustration of the one-to-many linking between the ternary diagram and parallel coordinates. The selected point in the ternary diagram corresponds to a mixture having 10 percent of $\text{Ing}_1$, 70 percent of $\text{Ing}_2$ and 20 percent of Others. The data set contains 5 ingredients meaning that Others consists of different combinations of $\text{Ing}_3$, $\text{Ing}_4$ and $\text{Ing}_5$, as seen in the parallel coordinates representation. In this particular example there are 6 different mixtures.

a one-to-many connection to the original data set since one mixture in a subset can represent multiple mixtures in the original data (as previously discussed in section 4.1). Hence, if a single mixture is selected in the ternary diagram showing a subset of the data, all mixtures in the original data are highlighted in the parallel coordinates representation. This is illustrated in figure 4 where the selected point in the ternary diagram corresponds to a mixture containing 10 percent of $\text{Ing}_1$, 70 percent of $\text{Ing}_2$ and 20 percent of Others. Since the data contains 5 ingredients, Others consists of more than one combination of $\text{Ing}_3$, $\text{Ing}_4$ and $\text{Ing}_5$, as seen in the parallel coordinates representation.

As described in section 4.2, the robustness plot displays mixtures selected in either the ternary diagram or in the parallel coordinates. Due to the relationship between the data sets, selection of one point in the ternary diagram results in several glyphs being displayed in the robustness plot, whereas the selection of a line in the parallel coordinates only results in one glyph being displayed in the robustness plot.

As mentioned in section 4.2, the parallel coordinates are used as a visual interface to control the other views. By selecting ingredient headers in the parallel coordinates the user controls which data subset to display in the ternary diagram, as well as which ingredients to present in the performance plot.

The ternary diagram view contains two legends where thresholding based on performance variables can be accomplished, as well as interaction with the colours. By interacting with these sliders the analyst can expand or contract the span where a specific colour sequence is used, which is then reflected in the other views. Expanding the colour span for areas of interest may enhance the understanding of variable variation and can be used to quickly find all areas above or below a certain value.

In addition to thresholding and selection, the ternary diagram offers some additional information retrieval aids.

Figure 5: Parallel coordinates when only ingredient mixtures with high Mean values remain. As visible, $\text{Ing}_1$ and $\text{Ing}_3$ have the greatest impact on performance.

When hovering over the triangle, guidelines linking from the position of the mouse pointer to the specific ingredient proportions at the triangle sides simplifies the interpretation of the diagram. Simultaneously, text information on proportions and performance is displayed beside the diagram. The legends and some of the interactive aids are shown in figure 2.

5 Example Scenario

This section provides a description of how an analyst can work with VIME to find an optimum multicomponent mixture, based on the analyst’s knowledge on performance efficiency and data constraints. For the data set that is to be analysed in this example, the desired performance is a high mean value and a stable result with low variance. In addition to demands on performance there are restrictions on two of the ingredients. Due to high cost, ingredient one should not make up more than 50 percent of the final mixture, and due to environmental issues, the maximum amount allowed of ingredient five is 20 percent.

When the data has been loaded into the application, the
analyst gets a quick understanding of which ingredients have the most influence on the result by removing all ingredient mixtures with low mean value, using the axis sliders in the parallel coordinates. As shown in figure 5, it is easily perceived that large amounts of ingredient one or five would give the desired performance result, however there are restrictions on both of these ingredients.

The next step the analyst takes is to remove all combinations not fulfilling the restrictions of ingredient one and five, again using the axis sliders of the parallel coordinates to perform the thresholding. To gain an understanding of the performance of each ingredient based on given restrictions, the ingredients are examined individually in the performance plot (see figure 6). It is found that high amounts of ingredient two and low amounts of ingredients three and four yield the best performance (highest Mean).

Based on this initial view the analyst decides to continue by examining the combinations of ingredients two and four more closely in the ternary diagram. Using the variance legend in the ternary view the ingredient mixtures mostly affected by uncontrollable variables are removed. By modifying the colour regions using the mean legend, the regions representing ingredient mixtures resulting in best performance are identified as yellow regions. The result of the interaction is displayed in figure 7.

The high performance ingredient mixtures are then selected in the ternary diagram, highlighted by black points in figure 7, and the corresponding combinations that these are made up of in the original data set are displayed in the robustness plot. As seen in figure 8 it is clearly visible that the performance varies between the selected combinations, and by examining them in the robustness plot the analyst can identify the combinations with the highest performance that also have the lowest variance and hence are least affected by uncontrollable variables. These combinations are selected within the robustness plot and their individual profiles can be examined in the parallel coordinates plot, as shown in figure 9.

The selected ingredient mixtures are suggestions of combinations that both fulfill the given constraints and have a high and stable performance. From here on the analysis can continue, for instance by examining other combinations of ingredients more closely in the ternary diagram and robustness plot.

6 Usability

The usability of VIME has been established and examined through an informal qualitative study halfway through the application development, as well as through discussions and feedback from potential users throughout the de-
development process. The application has hence been developed and adapted based on preferences, needs and requests of potential future users and domain experts.

The usability evaluation was carried out with seven potential future users, using an early layout of VIME. The participants became familiar with the application by being presented with three simple tasks to perform using VIME. Afterwards they were asked to answer a short questionnaire.

All except two, who graded themselves as novices regarding mixture data sets, thought this kind of application could be useful to them in their work. The participants were asked to specify what they considered the application could be used for, and suggestions for use were hypothesis forming, result demonstration and as a tool for understanding the decision making process. The usability of the application was graded high by all participants and some comments on the general usability were:

“It’s very useful for getting a feel for your data and understanding the relationships between the variables”

“Very good, intuitive controls, easy to pick up”

“Good, use of colour and hue make it easy to distinguish lines on parallel plot and spot patterns”

“Good. I had a problem highlighting some of the points/lines - it didn’t do what I thought it would do”

The results and comments of the usability evaluation were used to guide further development and to improve and adapt VIME to the needs and preferences of potential future users. For instance, the original layout contained a 3D scatter plot and a second parallel coordinates plot displaying the same data subset as the ternary diagram, these were both removed since several users found them redundant, and since the 3D scatter plot “looks pretty but is difficult to interpret and doesn’t add much”.

7 Conclusions and Future Work

This paper presents VIME (Visual Ingredient Mixture Exploration), an interactive tool for visualization and exploration of ingredient mixture data. The data that is analysed in VIME is calculated using a mathematical model where performance results and variance, due to process variables, are extracted. VIME provides an interactive environment for exploration of data extracted from these models and efficiently incorporates specialist knowledge to investigate and determine optimum solutions of ingredient mixtures.

To test the usability of the application, an informal qualitative evaluation with domain experts within the areas of statistics and chemistry has been performed. The results of the preliminary study were promising and indicated that the visual representations and interactivity of VIME clearly facilitate the analysis process compared to existing methods used.

Future work includes a thorough evaluation of the application and further generalizations, such as making it possible to use data models resulting in more than two performance variables. Furthermore the usability and generality could be further established by adapting the application to experimental data and multicomponent data from other research areas such as, for instance, environmental or energy data.

Acknowledgements

We would like to thank Cristina Vozian for her extensive work on upgrading and implementing parts of the application. We would also like to thank Johan Franzén for his indispensable support during the early implementation phase. Finally, we would like to thank the scientists at Unilever R&D, Port Sunlight, for their comments, ideas and feedback during the project.
References


