An Information-Aware Framework for Exploring Multivariate Data Sets

Ayan Biswas, Soumya Dutta, Han-Wei Shen, and Jonathan Woodring

Abstract—Information theory provides a theoretical framework for measuring information content for an observed variable, and has attracted much attention from visualization researchers for its ability to quantify saliency and similarity among variables. In this paper, we present a new approach towards building an exploration framework based on information theory to guide the users through the multivariate data exploration process. In our framework, we compute the total entropy of the multivariate data set and identify the contribution of individual variables to the total entropy. The variables are classified into groups based on a novel graph model where a node represents a variable and the links encode the mutual information shared between the variables. The variables inside the groups are analyzed for their representativeness and an information based importance is assigned. We exploit specific information metrics to analyze the relationship between the variables and use the metrics to choose isocontours of selected variables. For a chosen group of points, parallel coordinates plots (PCP) are used to show the states of the variables and provide an interface for the user to select values of interest. Experiments with different data sets reveal the effectiveness of our proposed framework in depicting the interesting regions of the data sets taking into account the interaction among the variables.

Index Terms—Information theory, framework, isosurface, multivariate uncertainty

1 Introduction

Exploration of multivariate data sets is an integral part of scientific visualization as in most real world phenomena, there exist multiple factors associated with the complex interactions of different variables. To gain an in-depth understanding of a scientific process, the relationship among the variables needs to be thoroughly investigated. However, exploration of several variables simultaneously can be both tedious and confusing. In a univariate system, displaying isocontours is a popular tool for data exploration for its ability to reveal the regions of the same scalar value. The identification of salient isocontours for a single scalar field has been a well researched topic. For single variable data sets, isocontour selection has been done previously based on the shape [3, 14, 34], topology [7, 46], and geometry [27] of the contours. For multivariate data sets, this non-trivial problem of isocontour selection becomes more challenging. In the multivariate scenario, as there can be interdependences among multiple variables, isocontours of selected variables can reveal information about the associated variables and how they interact. Identification of such informative isocontours is an important aspect of multivariate data exploration. To date, a guideline for exploring specific values of different variables, and studying the relationship among them is mostly missing.

In this paper, we introduce an information-aware framework that guides the users in multivariate data exploration. In our framework, we use an information-theoretic approach to guide the user at each step of the exploration process and help towards in-depth analysis of the data sets when multiple variables are involved. In this work, the mutual information among the variables is decomposed into specific information metrics which are used to facilitate the identification of informative isocontours. Since this metric takes advantage of information overlap between variables, we make use of the mutual information to create subgroups of the variables according to their information overlap. Inside the subgroup, since not all the variables involved in that subsystem contribute equally towards the joint entropy of the subgroup, conditional entropy measure is used to calculate the relative importance of the variables. This step is non-trivial as selecting variables based solely on their entropy may not suffice as there can be variables with high information but also sharing a large amount of mutual information with other variables. In this case, selecting one variable can reveal a significant amount of information about a subgroup. After the individual variables are selected, to explore the relationship among the variables, the specific information is used to calculate the informativeness of each scalar value based on how it is related to the values of other variables. Isocontours of the selected variables can be identified which provide the uncertainty information about the other variables at the same locations. We provide an intuitive interface which allows user interaction through the whole exploration process. Using our framework, the users receive interactive step-by-step exploration guidance to examine multivariate data sets.

Our contributions in this work are threefold:

1. We use specific information to classify the isocontours of variables based on the relationship of one variable with the others, and effectively provide simultaneous visualization of the related variables.

2. We use a novel graph-based approach to analyze and cluster a system of variables and analyze the interaction among the clusters and within the clusters.

3. We introduce a new framework for multivariate data exploration which guides the users at each step of the exploration process by providing an intuitive and interactive interface.

This paper is organized as follows: in Section 2, we review research works that are related to the topic of this paper. In Section 3, we provide a brief overview of our system. Section 4 discusses our information-aware exploration system in detail, and Section 5 presents the results of applying our framework to a few multivariate data sets. In Section 6, we present domain experts' feedback on our system components. Parameter choice, performance of our proposed system, and comparison with existing multivariate analysis techniques are discussed in Section 7. We provide the conclusion and future work in Section 8.

- *Ayan Biswas is with GRAVITY group, The Ohio State University. E-mail: biswas.36@osu.edu.*
- *Soumya Dutta is with GRAVITY group, The Ohio State University. E-mail: duttas@cse.ohio-state.edu.*
- *Han-Wei Shen is with GRAVITY group, The Ohio State University. E-mail: hwshen@cse.ohio-state.edu.*
- *Jonathan Woodring is with Los Alamos National Laboratory. E-mail: woodring@lanl.gov.*

Manuscript received 31 March 2013; accepted 1 August 2013; posted online 13 October 2013; mailed on 4 October 2013.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

2 Related Works

In this section, we provide a brief review about the areas of research which are directly related to the topic of this paper: information theory and its applications, multivariate data analysis and Parallel Coordinate Plots, and salient isocontour selection.

In visualization and computer graphics, information theory [10] has been widely used to solve a variety of problems. An entropy-based solution was presented by Gumhold [17] for placing light sources in a scene for different camera parameters. Information theory has been a popular choice for view point selection [36, 4, 37]. Feixas et al. [15] analyzed the scene visibility and radiosity complexity using an information-theoretic approach. For volumetric time varying data, Wang et al. [39] proposed an importance driven approach for time-varying data visualization. In this work, they conducted a block-wise analysis to find important features from time-varying data. In another work, Chen and Jänicke [8] provided evidence that information theory can explain numerous events in visualization. Xu et al. [43] proposed an information-theoretic framework for flow visualization. Considering visualization as a visual communication channel, they evaluated the effectiveness of visualization by measuring how much information in the original data is being communicated to the users. Rigau et al. [32] presented entropy-based aesthetics measurement for paintings.

Multivariate data analysis and visualization is an active research area and it has numerous applications in diverse fields [12, 41]. Yang et al. [44] proposed analysis guided multivariate exploration by introducing a Nugget Management System (NMS). NMS first extracts valuable information hidden in the data based on the interest of users and from that nugget, other similar nuggets can be discovered. To facilitate interaction, Martin and Ward [30] added high dimensional brushing. In a later work, Jänicke et al. [23] transformed high dimensional data into a 2D attribute space where attributes are represented as a point cloud, which allowed them to analyze the multivariate data in two dimensions. Rubel et al. [33] constructed a visualization system for extremely large multivariate data sets. Proposing a statistics-based framework on existing query driven visualization (QDV), Gosink et al. [16] improved the utility of QDV for large multivariate analysis. In another work, Claessen and Van Wijk [9] used flexible linked axes for multivariate data visualization. Combining Parallel Coordinates Plot (PCP) and MDS based projection techniques, Guo et al. created a novel transfer function design interface [18] to facilitate visualization of multivariate data. In a recent work by Wang et al. [38], information theory was used for exploring the causal relationship among the variables of a time-varying multivariate data set. With the use of transfer entropy, they formulated a complete graph to show the information transfer among the variables by modulating the size and color of the nodes. In our work, we apply mutual information metrics to generate a graph model and obtain an initial clustering of the variables. From the graph, variables are selected and processed with specific information metrics which allow us to understand the variability of one variable with respect to another variable. For visualization purposes, PCP and isosurfaces have been used.

PCP [22, 21, 20] is well known for multivariate analysis where attributes are represented as parallel vertical axes scaled within their data range. However, visual cluttering in PCP can pose a significant problem towards the exploration of relationships between the neighboring axes. Ordering of the axes in PCP plays an important role in the exploration process and researchers have looked into this problem in the past [2, 31, 28, 19]. In PCP, quantification of visual clutter and its reduction [24, 11] remain an ongoing research problem. We use mutual information to reorder the PCP axes and then show the relationship for the user selected axis by the use of specific information metrics.

Isosurfaces have been well studied in the past. The saliency of isosurfaces can be decided based on the data set's statistical properties [3, 14, 34]. The topology of isosurfaces can provide important cues towards understanding the scalar field structures and can be used to determine the saliency of the isosurfaces as well [7, 46]. In some recent works, geometric properties of isosurfaces such as fractal dimensions [27] and distance transforms [6] have also been used for saliency analysis of isosurfaces. In our work, color mapped isosurfaces have been used to show the variability of the color mapped variable on the surface where the other variables' value remains constant.

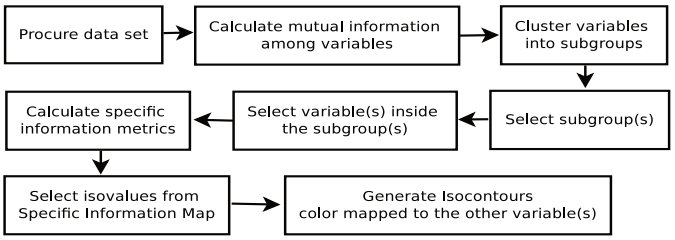


Fig. 1. A schematic representation of the workflow.

3 System Overview

In this section, a brief overview of our approach is provided. Our high level goal in this paper is to explore the multivariate data and identify regions of interest through information analysis. In this paper, we show that by using the concept of specific information measures, one can quantify the informativeness of a scalar value in one variable with respect to the uncertainty of the other variables. The specific information is based on the decomposition of the standard mutual information. To allow for more effective analysis of correlation in the data set, the variables under study should have a good information overlap. To achieve this, we subdivide the variables into smaller groups based on their mutual information. The subgroups can be selected based on their total information content which is measured by the joint entropy of the subsystem. Inside the subgroups, the variables that contain the most information about the subgroup are identified using conditional entropy. After selecting the variables, the specific information metrics are used to identify different scalar values with varying amounts of information. We present an information-aware multivariate data exploration framework with novel features such as quantitative analysis of information overlap among the variables, information-driven grouping and selection of variables for more systematic correlation study, and flexible user interface for easy selection and browsing of salient data. A schematic view of our system is provided in Figure 1, where isosurfaces have been used for variability visualization.

4 Information-aware Framework for Data Exploration

Below we discuss in detail our analysis framework, where information-theoretic approaches are used to tackle different aspects of the multivariate data exploration problem.

4.1 Information Overlap in Multivariate Data

In this section, we present our approach for multivariate data exploration by analyzing the degree of information overlap among the variables. In multivariate analysis, correlation between different variables has been a well researched topic and there are several metrics available. In information theory, mutual information between two random variables is the measure of information overlap or the correlation between the variables. For two random variables X and Y, mutual information $I(X,Y)$ is defined as

$$I(X,Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}. \quad (1)$$

As a correlation metric, mutual information has an advantage over the correlation coefficient metric as it can measure non-linear relationships as well. Mutual information between two variables measures the informativeness of one variable about the other variable.

It is also possible to measure the information associated with a specific scalar value x, $x \in X$, about another random variable Y, which is termed as *specific information*. In this case, the variable X is called the *reference variable*. In our framework, we utilize the specific information of the reference variable to identify salient isocontours, where

the saliency is calculated by how much the uncertainty of one variable is reduced after observing an isocontour from the reference variable. There exist several ways to calculate specific information, and we discuss two of them that were introduced in the works of DeWeese and Meister [13]. These two specific information metrics were named as Surprise and Predictability by Bramon et al. [5], all based on a decomposition of the standard mutual information.

Surprise: Surprise, which is also referred to as I_1, was introduced and analyzed by DeWeese and Meister [13] and it is given as:

$$I_1(x;Y) = \sum_{y \in Y} p(y|x) \log \frac{p(y|x)}{p(y)}. \tag{2}$$

$I_1(x;Y)$ is always positive as it represents the Kullback-Leibler distance between $p(Y|x)$ and $p(Y)$. A high $I_1(x;Y)$ value indicates that some infrequent occurrences $y \in Y$ have become more probable due to the observation of x which amounts to a surprising result, where x is a value of the reference variable X. The values of x, for which $I_1(x;Y)$ is high, are representative of the isovalues which are of interest to us.

Predictability: DeWeese and Meister also introduced the metric Predictability or I_2 which is given as:

$$\begin{aligned} I_2(x;Y) &= H(Y) - H(Y|x) \\ &= -\sum_{y \in Y} p(y) \log p(y) + \sum_{y \in Y} p(y|x) \log p(y|x) \end{aligned} \tag{3}$$

$I_2(x;Y)$ gives the amount of reduction in uncertainty about Y after observing the data value x. It is to be noted that, unlike I_1, I_2 can take negative values which suggests that there are certain observations x for which our uncertainty about Y may increase. The values of x, for which $I_2(x;Y)$ is high, are representative isovalues of the reference variable X that reduce the uncertainty about Y and are of interest to us. On the other hand, the values of x, for which $I_2(x;Y)$ is low or negative, represent a high uncertainty about the variable Y at the location of the isocontours, which also prompts us to perform further exploration.

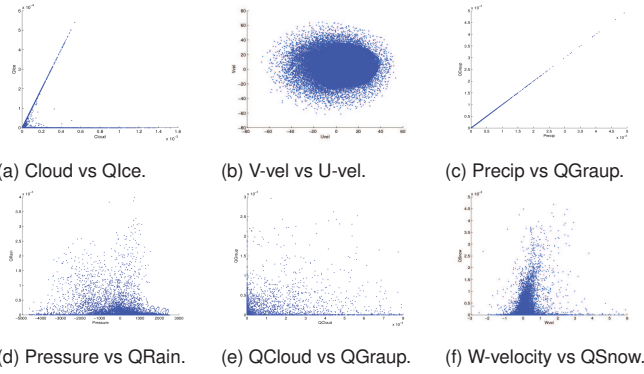
I_2 is an additive metric, which means when I_2 is measured based on two observations x and y, it amounts to the sum of I_2 calculated based on x, and I_2 based on y given that x is already known.

$$I_2(x,y;Z) = I_2(x;Z) + I_2(y;Z|x).$$

As additivity is a desirable property of a metric, Bramon et al. [5] and DeWeese and Meister [13] have considered I_2 to be more intuitive as a measure of specific information. In our work, we utilize both I_1 and I_2 for the selection of interesting scalars.

These specific information measures provide us with the tools for classifying the individual scalars of a variable. Given two variables of a data set, each of them can be considered as a random variable and for each scalar value of the reference variable chosen between the two variables, the I_1 and I_2 metrics can be computed. To guide the exploration of the data set, the scalars which have high I_1 value are identified as the surprising ones and are further classified by their I_2 values. If a surprising value has higher predictability, then the corresponding isocontour of the variable will reflect a confident state of the other variable. Conversely, if the predictability is low, then the scalar value associated with the reference variable will not be able to predict the state of the other variable with high certainty.

Since the specific information is computed between pairs of variables at specific values, for the I-metrics to work well, it is desired that the variables in question have enough information overlap. This motivates the need to explore the relationships among all the variables and group closely related variables together. In addition, to use the specific information measures, a reference variable is needed to be picked which ideally should be the variable that contains more information within a group. In the next section, we describe a graph-based approach for grouping and reference variable selection, based on the information overlap to facilitate multivariate analysis.



(a) Cloud vs QIce. (b) V-vel vs U-vel. (c) Precip vs QGraup.

(d) Pressure vs QRain. (e) QCloud vs QGraup. (f) W-velocity vs QSnow.

Fig. 2. Scatter plots between different variables showing different degrees of correlation.

4.2 Mutual Information Based Grouping of Variables

From an information-theoretic point of view, each variable in a multivariate data set carries a certain amount of information that is also shared by other variables, which can be characterized by the mutual information measure. An experiment with the Isabel Hurricane data set (discussed in Section 5.2) reveals that there exist variables in the system which show strong correlation with some variables but not so much with others. In Figure 2, we show different instances of relationship among the variables of the system. From the scatter plots shown in Figure 2a, 2b, and 2c, it is observed that a stronger relationship exists between the variables under study. A quick comparison of the scatter plots shown in Figure 2d, 2e and 2f, on the other hand, reveals that the correlations between those variables are not as strong. This study of variable relationship allows us to perform a more systematic analysis of the variables using the specific information metrics.

For each pair of variables, we measure the distance between the variables as the inverse of the mutual information between them. Considering all the variables of the system, a graph $G(V,E)$ is constructed which delineates our system of variables. Each node $v \in V$ represents a variable, and each undirected edge $e \in E$ represents the mutual information between the two variables. A hierarchical clustering is applied on this graph to decompose it into different groups. In a bottom-up clustering approach, each node represents a leaf of the cluster tree and each of them starts out as a cluster. Then the new groups are formed via a greedy algorithm which merges the two clusters with most similarity to move up the cluster tree one level. For n nodes, the general complexity of the algorithm is $O(n^3)$ and gives a locally optimal solution.

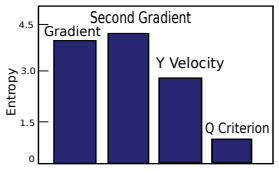
As shown in Figure 4a, the dendrogram representation of the clustered graph is presented which reveals the hierarchy of the clusters. It is evident from the figure that there exist three major subgroups which form clusters according to the information overlap. In Figure 4b, an alternative graph view of the system is provided. The layout of the graph is generated by a force-directed algorithm where the attractive force is given as the mutual information between the nodes and the repulsive force is the inverse of the mutual information. Here we see that the force-directed layout matches our hierarchical clustering results.

After subdividing the variables in the whole system into subgroups, the subgroups need to be classified according to their information content. Also, the variables contained inside the subgroups, need to be analyzed for their importance. We apply information theory to proceed with the data exploration and we discuss it in the next section.

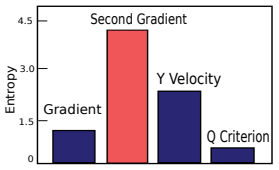
4.3 Information Based Variable Selection

Information theory provides us with a method for quantifying the information content of a random variable by using Shannon's entropy calculation. For a random variable X, Shannon's entropy $H(X)$ is defined as

$$H(X) = -\sum_{x \in X} p(x) \log p(x). \tag{4}$$



(a) Uncertainty remaining in the variables before selection.



(b) Uncertainty remaining in the variables after selection of Second Gradient.

Fig. 3. Change in uncertainty of the variables due to the variable selection.

This is a measure of the uncertainty about the given random variable. For a collection of random variables $X_1,...,X_n$, the total information content among the variables can be expressed by the calculation of joint entropy which is defined as

$$H(X_1,...,X_n) = -\sum_{x_1 \in X_1} ... \sum_{x_n \in X_n} p(x_1,...,x_n) \log p(x_1,...,x_n). \quad (5)$$

For univariate data sets, Shannon's entropy has been previously used to quantify the uncertainty of variables. The histogram of a variable is used as the probability mass function to calculate the total uncertainty of that variable using Equation 4. As discussed in the previous section, in the multivariate data sets, the variables can be grouped based on their correlation, and the joint entropy can now be applied to measure the total uncertainty within each group using Equation 5 where the joint probability distribution of the variables is used. This allows us to compute the relative importance of the groups based on their uncertainty. The groups can be selected depending on their uncertainty and the individual variables inside the selected group now need to be analyzed.

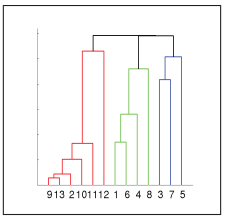
In information theory, given a group of variables, conditional entropy is used to quantify the information gain about the system when some of its variables are known. From n variables $X_1,...,X_n$, if m variables $X_{k1},...,X_{km}$ are known, then the amount of uncertainty left in the system is given by

$$H(X_1,...,X_n|X_{k1},...,X_{km}) = H(X_1,...,X_n) - H(X_{k1},...,X_{km}). \quad (6)$$

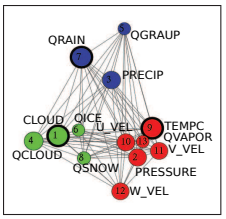
This provides a useful measure to identify variables inside subgroups that have a larger contribution towards the total uncertainty of a group. Also, this metric allows us to select some of the variables such that those variables represent the uncertainty of the whole subgroup by their information content.

An experiment with four fields of the Plume data set is used as a motivating example of how the selection of one variable can reduce the uncertainty about other variables. Plume is a simulation of the thermal downflow plumes on the surface of the sun with $126 \times 126 \times 512$ grid points. Figure 3 shows that if we want to explore the higher uncertain variables first, then just by looking at their individual entropies, the order of selection would be Second Gradient, Gradient, Y Velocity and Q Criterion as shown in Figure 3a. But, if Second Gradient field has been selected, it will now have impact on our knowledge about the other variables and as shown in Figure 3b, the second variable to be chosen will be Y Velocity as the uncertainty about Gradient field has been reduced due to the selection of Second Gradient field.

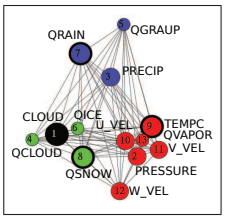
The objective of our variable selection methodology is to maximize the information gain about the subsystem after every variable selection. The selected variables also become the candidates for reference variables used to compute the I-metrics mentioned in section 4.1. It is likely that a variable of high information content, i.e., high entropy, will carry more information shared by other variables, and therefore, a good candidate to be used as a reference variable. In this sequential selection process, variables are selected and the residual information content of the remaining variables is updated to reflect the current state of the subsystem. Variables can be selected within a group and also across groups. While selecting variables across the groups, for each group, a variable can be identified which represents that subgroup and these representative variables can now be candidates for selection. For exploration within the group, the dynamic ordering of importance can be followed to guide the selection process.



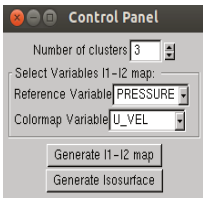
(a) The hierarchical clustering of the system of variables.



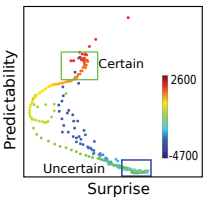
(b) The corresponding graph layout.



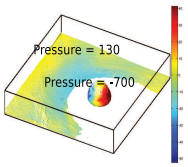
(c) The corresponding graph layout.



(d) Control panel for user's interaction.



(e) The I_1I_2 map of the user selected variables, Pressure and U-Velocity.



(f) Generated isosurface for the selected isovalue of the I_1I_2 map.

Fig. 4. A view of the system when applied on the Isabel Hurricane data set.

4.4 The Information-aware Data Exploration Framework

In this section, we combine all the previously described components together to design our framework that facilitates the exploration of multivariate data sets. In our interface, as graphs are intuitive and easier for interaction, they are chosen to represent the clustering of the variables drawn by a force-directed graph layout similar to the example shown in Figure 4b. For generating the graph layout, we have used the algorithm proposed in [26] which relates the graph layout as a dynamic spring system and the layout is generated by minimizing the total energy of the system. For variables i and j, the attractive force between the two nodes is represented as

$$F_{ij} \propto \frac{1}{d_{ij}^2}$$

where d_{ij} is the distance between the two nodes. We take d_{ij} as the inverse of the mutual information between the variables i and j for our analysis. A similar force-directed graph was also used by Zhang et al. [45] for network construction and finding an ordering in their PCP plot of the high dimensional data. In our experiments, we found that the hierarchical clustering of the variables based on all pair mutual information is reflected closely by the force-directed graph layout. The users are provided the option of selecting the number of clusters interactively and this change is automatically reflected in the graph layout by showing the individual variables colored by the colors of the clusters. The users are able to color the nodes by the joint entropy or the information content of the respective groups. Inside the groups, the information-based importance is reflected by the size of the nodes. For each group of variables, the most representative variable is highlighted for easy identification. Depending on the user's selection of variables (nodes), the relative importance of the other variables of a group is updated and shown on the display for the user to make the next selection. As shown in Figure 4b, the state of the graph shows the initial display of the system with thirteen variables of the Isabel data set. It depicts a scenario where the thirteen variables are clustered into three groups, and each group is represented by a separate color which represents its group entropy: red as high, blue as low and green in between. In each of the three groups, the representative variables are earmarked. Now if the user makes a selection of node 1, as shown

in Figure 4c, the importance of the other variables is updated and the new representative variable is highlighted as node 8.

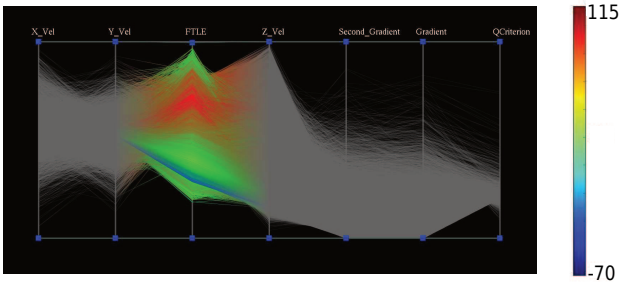


Fig. 5. Simultaneous display of multivariate system using Parallel Coordinates Plots. The user selected axis is color mapped to $I_{uncert} = I_1/I_2$.

In addition to using graphs to represent the relationships among the variables, in our framework we provide the users with two different ways to explore the data: 1) Exploration in the data domain using Parallel Coordinate Plots (PCPs), and 2) Exploration in the spatial domain using isosurfaces. For the exploration in data domain, PCPs can be effectively used to visualize the system of variables. In PCP, the ordering of the variables is an important aspect for providing a useful visualization. Since the variables are clustered into groups based on the mutual information, we use this grouping information to control the display and to find the ordering for the PCP, as described below.

Because the groups represent variables with higher correlation, the variables can be shown groupwise in the PCP. Inside the group, we find an ordering of the variables such that the mutual information is highest between the neighboring axes of the PCP, so that user can more easily understand the relationships among the multiple variables. To order the variables into a sequence of PCP axes, our goal is to maximize the total amount of information presented in the plot. With our graph model, this ordering can be solved by finding a Hamiltonian path inside the subgraph that minimizes sum of the edge weights. As the edge weights are inverse of the mutual information, the minimum weight Hamiltonian path maximizes the total information presented along the sequence of PCP axes. Although finding an optimum Hamiltonian path in a graph is an NP-complete problem, for relatively small graphs, the brute force method works well. For larger graphs, an approximated solution minimizing the inter-cluster crossings yields sufficiently good results [42]. Figure 5 shows the ordering of the seven variables of the Plume data set that maximizes the mutual information in PCP.

To supply data to the PCP interface, the users can choose to explore the volumetric multivariate data set in slices, multiple slices or the whole volume. However for large volume data, as the number of points grows larger, the PCP can get cluttered. Brushing is a popular technique in visualization to interactively select the regions of interest. We allow brushing of points according to the scalar values and also according to the specific information values. Since the specific information metric I_2 represents the uncertainty or predictability factor and I_1 describes the "surprise" of the scalar value, to show the scalar values which correspond to higher uncertainty in the other variable, a derived metric I_{uncert} is formulated such that

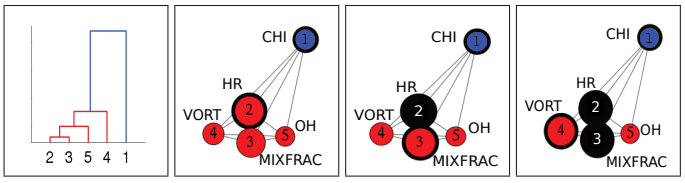
$$I_{uncert} = I_1/I_2. \tag{7}$$

I_{uncert} identifies the higher I_1 and lower I_2 values. Similarly, for the scalar values which correspond to lower uncertainty in the other variable, another derived metric I_{cert} is generated which is given as

$$I_{cert} = I_1 * I_2. \tag{8}$$

I_{cert} identifies the higher I_1 and higher I_2 values. As shown in Figure 5, the PCP plot shows the I_{uncert} metric on the user selected axis FTLE computed against Z-velocity for the Plume data set.

To facilitate exploration of volume data in the spatial domain, we incorporate the idea of specific information mentioned above. The user



(a) Hierarchical clustering of the variables. (b) Graph layout of the variables. (c) State of the graph after first variable selection. (d) State of the graph after second variable selection.

Fig. 6. Combustion data set with 5 variables: hierarchical clustering based on mutual information and a force-directed graph layout.

selects two variables X and Y from the graph. The $I_1(x;Y)$ and $I_2(x;Y)$, $x \in X$, are computed and a 2D scatter plot representing mapping of the scalars of X to the I_1I_2 space is generated, as shown in Figure 4e. In this figure, each point represents an (I_1, I_2) pair computed from a specific scalar value of X. In this map, different scalars of X are color mapped according to their value which makes it easier for users to see any relationship patterns. Users are provided with an option to select a point or a region from the I_1I_2 scatter plot with a rectangular window, and the corresponding scalar values are selected. The selected values are then used to draw the isosurfaces on the variable X which are color mapped by the scalars of Y as shown in Figure 4f. To identify the regions where the isosurface of X faithfully represents the value of Y, the higher I_1 and higher I_2 values are selected. Conversely, to identify the regions where the isosurface of X has high variation in values of Y, the higher I_1 and lower I_2 values are needed to be selected. The Figure 4e shows the I_1I_2 map of Pressure calculated with U-velocity from the Isabel data set. The x-axis represents I_1 and y-axis represents I_2. The green rectangular region selects the scalars from the Pressure field that have less variability in U-velocity values. Conversely, the blue region selects the values which have much more variability. Figure 4f shows the two examples of isosurfaces as a result of user's selection. The isosurface corresponding to Pressure value -700 comes from the blue rectangular region and it has high variability in the U-velocity. The isosurface for Pressure value 130 is selected from the green rectangular region has much less variability in U-velocity.

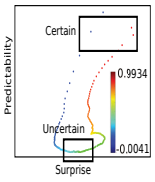
5 Results

In this section, we show the results of our framework in exploring data sets with multiple variables. The experiments were conducted on a Linux machine with an Intel core i7-2600 CPU, 16 GB of RAM and an NVIDIA Geforce GTX 560 GPU with 2GB texture memory. For the calculation of information-theoretic measures 256 histogram bins were used. The force directed graph layout was generated by the Boost Graph Library [35].

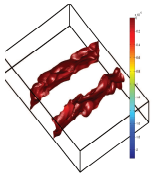
5.1 Combustion Data set

This data set is a turbulent combustion simulation data which is a time varying volume data set having five scalar variables: Mixture Fraction (MIXFRAC), Vorticity (VORT), Mass Fraction of Hydroxyl (OH) radical, Heat Release Rate (HR) and Scalar Dissipation Rate (CHI) in turbulent flames. The data set is made available by Dr. Jacqueline Chen at Sandia Laboratories through US Department of Energy's SciDAC Institute for Ultrascale Visualization. Each time step of this data set contains $480 \times 720 \times 120$ grid points. The mixture fraction denotes the proportion of fuel and oxidizer mass and this value generally provides the location of the flame where the chemical reaction rate exceeds the turbulent mixing rate. But there can be regions where the mixing rate dominates the chemical reaction rate and the flame is partially extinguished or weakly burning. To have a detailed understanding of the combustion phenomenon, only analyzing the mixture fraction may not be enough and multivariate analysis is needed for this complex process. Time step 41 from the data set was selected for our experimental purposes.

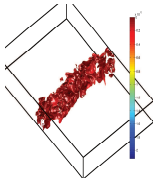
Using our framework, the all pair mutual information is calculated



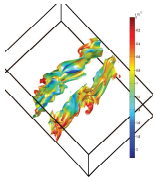
(a) I_1I_2 map of Mixture Fraction computed against Heat Release Rate.



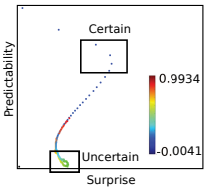
(b) Mixture Fraction isosurface for isovalue 0.01 colored by the Heat Release Rate showing low uncertainty.



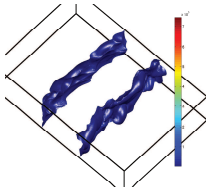
(c) Mixture Fraction isosurface for isovalue 0.95 colored by the Heat Release Rate showing low uncertainty.



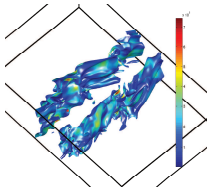
(d) Mixture Fraction isosurface for isovalue 0.40 colored by the Heat Release Rate showing higher uncertainty.



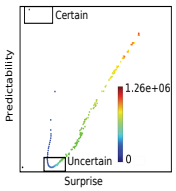
(e) I_1I_2 map of Mixture Fraction computed against Vorticity.



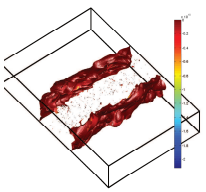
(f) Mixture Fraction isosurface colored by the Vorticity. At low Mixture Fraction value, 0.01, the uncertainty about Vorticity is low.



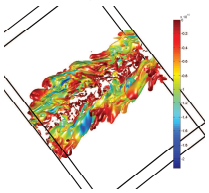
(g) Mixture Fraction isosurface colored by the Vorticity. In this selection, at Mixture Fraction value 0.5, the uncertainty about Vorticity is much higher.



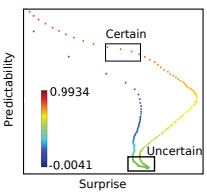
(h) I_1I_2 map of Vorticity computed against Heat Release Rate.



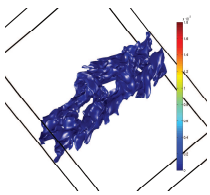
(i) Vorticity isosurface colored by the Heat Release Rate. At very low Vorticity, the uncertainty about HR is low.



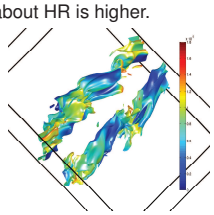
(j) Vorticity isosurface for isovalue 148750 colored by HR. At higher Vorticity value, the uncertainty about HR is higher.



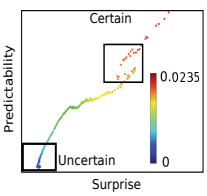
(k) I_1I_2 map of Mixture Fraction computed against Oxidizer OH mass fraction.



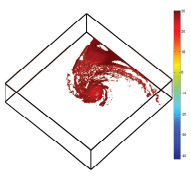
(l) Mixture Fraction isosurface colored by the Oxidizer OH mass fraction. At Mixture Fraction 0.85, the uncertainty about OH mass fraction is low.



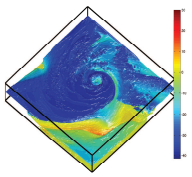
(m) Mixture Fraction isosurface colored by the Oxidizer OH mass fraction. At Mixture Fraction value 0.47, the uncertainty about OH mass fraction is high.



(n) I_1I_2 map of QVAPOR computed against Temperature.



(o) QVAPOR isosurface colored by the Temperature. At QVAPOR value 0.0188, the uncertainty about Temperature is low.



(p) QVAPOR isosurface colored by the Temperature. At QVAPOR value 0.0004, the uncertainty about Temperature is much higher.

Fig. 7. Different isosurfaces showing different amount of uncertainty for other variables.

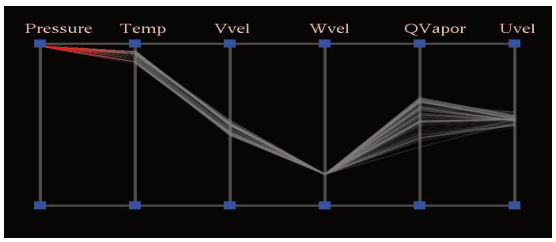
and then the graph layout is generated for user interaction as shown in Figure 6. After mutual information based grouping, Mixture Fraction, Heat Release Rate, Vorticity and OH mass fraction are placed in the same subgroup and this subgroup has higher entropy compared to the subgroup consisting of Scalar Dissipation Rate as reflected by the color of the subgroups in the graph layout. Then, we proceed to the exploration of the individual variables. In the process of variable selection within the subgroup, the first two variables suggested by our framework are Heat Release Rate and Mixture Fraction. The corresponding I_1I_2 map of the reference variable Mixture Fraction is shown in Figure 7a. Selecting the points which correspond to high I_1 and high I_2, it is observed that isosurfaces of very low and very high Mixture Fraction values all have very high Heat Release Rate, as shown with the two selected isosurfaces in Figure 7b and Figure 7c. But for the Mixture Fraction values between 0.3 to 0.55, the Heat Release Rate values are varying widely. As mentioned in [1], the stoichiometric mixture fraction for this mixture is 0.42 which corresponds to the flame. From our results, it is observed that the Heat Release Rate is not stable on or around the flame which suggests that some complex interaction around the flame is happening. To get more insight, the next variable to be analyzed is Vorticity or the flow turbulence, as suggested by our graph layout. The Figures 7e,7f and 7g present the results of isosurfaces drawn on the reference variable Mixture Fraction color mapped with Vorticity. From the results, it is to be understood that near the flame, the turbulence is higher, which is causing higher uncertainty in the region. Similar deductions can be made from the Figures 7h, 7i, and 7j where we show the results of Vorticity and Heat Release Rate. The more certain Heat Release Rate values occur where the reference variable Vorticity is much lower. As the Vorticity or the turbulence increases, the Heat Release Rate becomes more uncertain. Finally we present the results of the analysis of Mixture Fraction and Mass fraction of OH radical in the Figures 7k, 7l and 7m with Mixture Fraction as the reference variable. From the resulting isosurfaces, it is apparent that the OH mass fraction is not constant around the stoichiometric mixture fraction as noted in [1].

5.2 Hurricane Isabel Data Set

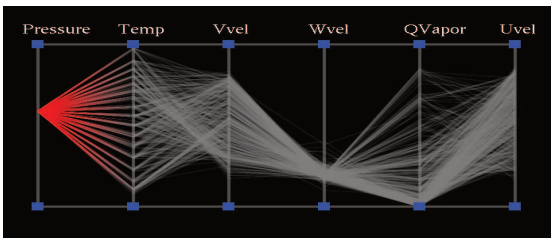
Next, we show the exploration results from the Hurricane Isabel data set. Hurricane Isabel data was produced by the Weather Research and Forecast (WRF) model, courtesy of NCAR and the U.S. National Science Foundation (NSF). This data set consists of thirteen variables and the resolution of the data set is $500 \times 500 \times 100$ for a single time step. We have selected time step 20 for our experiments.

For this data set, the initial force-directed graph layout was presented in Figure 4b. When clustering the graph into three groups, the highest entropy group consists of Pressure, TC, U Vel, V Vel, W Vel, and QVAPOR as reflected by the color of this group shown in the graph layout. From this group, choosing the variables Temperature and QVAPOR, with QVAPOR as the reference variable, we constructed the I_1I_2 map as shown in Figure 7n. From this map, the QVAPOR values are selected which correspond to less variability in Temperature as shown in 7o. On this isosurface, which is generated by QVAPOR value 0.0188, nearly the same and relatively high Temperature values are observed. Whereas, we can also identify isosurfaces as in the Figure 7p, where the variability in the Temperature value on the surface is much higher for QVAPOR value close to 0.0004. Figure 4 shows the result of selecting Pressure and U velocity for analysis, where Pressure is the reference variable. The I_1I_2 map is presented in 4e which is used to select certain and uncertain isosurfaces on Pressure corresponding to U velocity. Figure 4f shows examples of two such isosurfaces. In this figure, the Pressure isosurface for isovalue 130 has much less variation in U Velocity whereas, as we move closer to the Hurricane eye region where the Pressure goes down, there is much more variation in U Velocity as represented by the isosurface of Pressure value around -700.

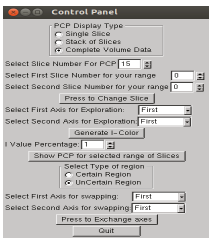
For a downsampled version of Isabel data, we use PCP to highlight informative scalars for Pressure and Temperature in Figure 8 where Pressure is used as the reference variable for specific information calculation. In this figure, we only show the variables of the selected



(a) Selection of certain scalars from PCP using I_{cert}.



(b) Selection of uncertain scalars from PCP using I_{uncert}.



(c) Sample interaction buttons.

Fig. 8. Exploration for Pressure and Temperature variables inside the subgroup of Isabel data.

group for demonstration purposes. We use I_{cert} and I_{uncert} metrics to choose the scalars from PCP. Figure 8a shows the selection of scalar values using I_{cert} where a small range of scalars of Pressure maps to a small range of scalars in Temperature. Conversely, Figure 8b shows the selection of scalars using I_{uncert} where a small range of scalar values of Pressure maps to a much larger range of scalar values on the Temperature axis. Figure 8c shows a part of the interface which facilitates the selection process. Similarly, other variables of the system can be used for exploration with effective results.

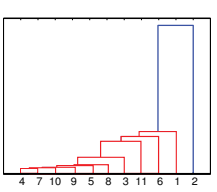
5.3 Ionization Front Instability Simulation Data Set

In our next case study, we use the data set presented in the IEEE 2008 Visualization Design Contest [40]. In this data set, researchers have intended to explore the relationships of the ionization front instabilities with the formation of the first stars of the universe. This data set contains the relative abundancies of eight chemical species, temperature, density and the velocity field. The data set size is $600 \times 248 \times 248$ and we have selected time step 99 for exploration purposes. One of the tasks of the contest was to investigate the reasons of turbulence. We used the curl magnitude computed from the velocity field as the turbulence measure and calculated the individual species number densities of the chemicals to show our exploration results.

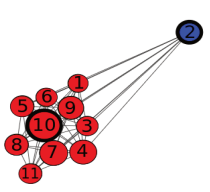
The initial clustering result and layout of the force-directed graph is presented in Figure 9a and Figure 9b. To see the relationship between Curl and H, the corresponding I_1I_2 map was generated as shown in Figure 9d using H as the reference variable. In this case, we find that, the low and high species number densities of H correspond to more certain regions in Curl. Figure 9e shows an isosurface at a low value of H. The uncertain regions occur in the intermediate values of H species density where there is more variability in Curl values, shown in Figure 9f. From our framework, if the node representing Curl is selected, then Temperature becomes the next candidate for selection. The corresponding results are shown in Figures 9g, 9h, and 9i. In this case, lower temperature values correspond to more certain Curl values and with the increase in Temperature, the Curl values become more uncertain as shown by the two isosurfaces. If Curl and Temperature nodes are selected, then the next candidate node suggested by our system is H+. The Figures 9j, 9k, and 9l represent results where lower values of H+ have more uncertain Curl values, whereas the higher H+ values correspond to more certain Curl values.

6 User Evaluation

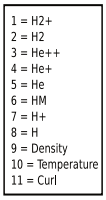
We demonstrated our system to the scientists of the Los Alamos National Laboratory to assess the effectiveness of the system for real world applications. At the beginning of the demonstration, high level explanation about the system workflow was presented to the scientists. Then we went over the details of each component of the system, providing some pre-generated results and images for an understanding of how the system helps in the exploration of the multivariate data sets. As the domain experts gained familiarity with the system, we presented the Eastern Seaboard and Gulf of Mexico region from the ocean component (LANL's POP model) of a fully coupled climate simulation using the DOE/NSF Community Climate System Model (CCSM4) with a grid resolution of approximately 1° and resolution of $320 \times 384 \times 60$. Embedded within the ocean is a model of marine



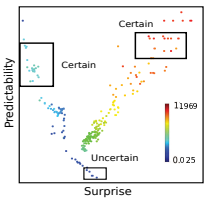
(a) Dendrogram representation of the hierarchical clustering of the variables.



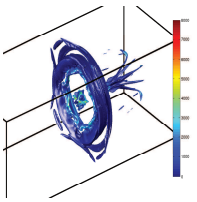
(b) Force-directed layout for graph view of the variables.



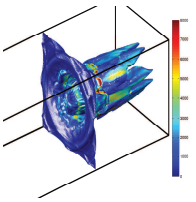
(c) Variable names.



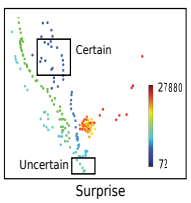
(d) I_1I_2 map of H computed against Curl.



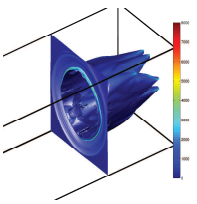
(e) H isosurface colored by Curl magnitude. At low species number density of H, e.g. 235, the uncertainty about Curl is low.



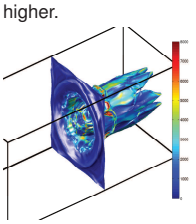
(f) H isosurface colored by Curl magnitude. At intermediate species number density of H, e.g. 3285, the uncertainty about Curl is much higher.



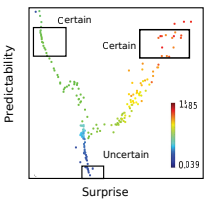
(g) I_1I_2 map of Temperature computed against Curl magnitude.



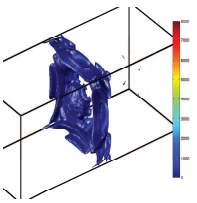
(h) Temperature isosurface colored by Curl magnitude. At low Temperature, e.g. 1150, the uncertainty about Curl is low.



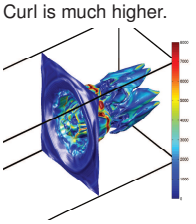
(i) Temperature isosurface colored by Curl magnitude. At higher Temperature, e.g. 8800, the uncertainty about Curl is much higher.



(j) I_1I_2 map of H+ computed against Curl magnitude.

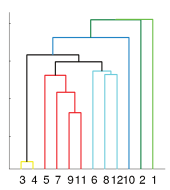


(k) H+ isosurface colored by Curl magnitude. At higher species number density of H+, e.g. 620, the uncertainty about Curl is low.

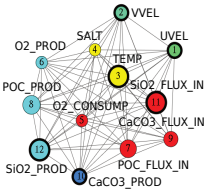


(l) H+ isosurface colored by Curl magnitude. At low species number density of H+, e.g. 62, the uncertainty about Curl is much higher.

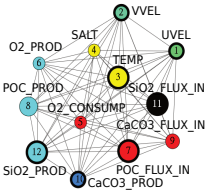
Fig. 9. Results of Ion Front data set.



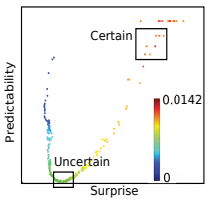
(a) Hierarchical cluster tree clustered into 6 groups.



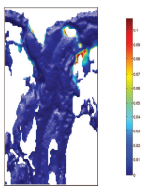
(b) The graph layout.



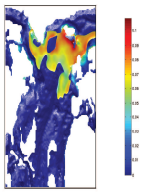
(c) State of the graph after a variable selection.



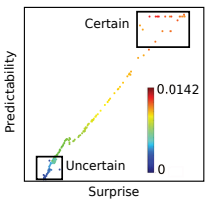
(d) I_1I_2 map of SiO2-FLUX-IN against POC-FLUX-IN.



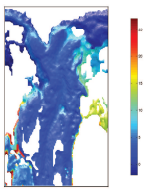
(e) Isosurface of SiO2-FLUX-IN for isovalue 0.0124 color mapped by POC-FLUX-IN.



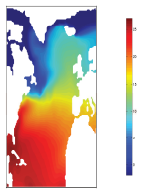
(f) Isosurface of SiO2-FLUX-IN for isovalue 0.0075 color mapped by POC-FLUX-IN.



(g) I_1I_2 map of SiO2-FLUX-IN against TEMP.



(h) Isosurface of SiO2-FLUX-IN for isovalue 0.0128 color mapped by TEMP.



(i) Isosurface of SiO2-FLUX-IN near 0 color mapped by TEMP.

Fig. 10. Application of our system on POP data set for domain expert's feedback.

ecodynamics that includes representations of plankton, nutrients, and other forms of organic and inorganic matter. This demonstration included several variables that described the production ("PROD") and sinking ("FLUX-IN") of Particulate Organic Carbon (POC), Silicate (SiO2), and Calcium Carbonate (CaCO3), as well as Oxygen (O2) production and consumption (CONSUMP), Temperature (TEMP), Salinity (SALT), and the horizontal velocity components (UVEL, VVEL).

In the exploration process, the users initially followed the suggestions made by the system and later on, they explored some of the variables by themselves. At the beginning, the users interactively selected the number of clusters of the hierarchical cluster tree to begin the exploration. Figure 10a shows the hierarchical cluster of the given set of variables where the number of clusters was 6. This grouping based on information overlap provides insight into the data. For example, grouping of SiO2-FLUX-IN, CaCO3-FLUX-IN, POC-FLUX-IN and O2-CONSUMP points to the fact that in the ocean, vertical fluxes of phytoplanktonic hard parts are often correlated and the particulate organics exert demand on dissolved oxygen. Then the users could learn about the information content of the subgroups by observing the node colors of the graph as shown in Figure 10b. For exploration within a subgroup, the users initially chose the variables SiO2-FLUX-IN and POC-FLUX-IN as suggested by the graph in Figure 10b and Figure 10c. The corresponding I_1I_2 map was generated as Figure 10d with SiO2-FLUX-IN as the reference variable. The map shows that the low and the high values of the reference variable have higher predictability but the higher values have more surprise. Then from this map, isovalues were selected according to their Predictability and Surprise for visualization of the isosurfaces as shown in Figure 10e and Figure 10f to show the variability of the POC-FLUX-IN. For exploration across the subgroups, the two representative variables SiO2-FLUX-IN and TEMP were selected and the corresponding I_1I_2 map was generated with SiO2-FLUX-IN as the reference variable which is shown in Figure 10g. From this map, it is evident that larger values of SiO2-FLUX-IN have higher predictability and surprise compared to its smaller values. This points to the fact that for this data set, the silica concentration is higher at the bottom level of the ocean where the cooler and heavier water is present which gives lower variability in temperature. Conversely, at the surface level, the silica concentration is lower and temperature variability is higher. This can be visually analyzed by generating isosurfaces of SiO2-FLUX-IN and color-mapping it with TEMP. Figure 10h shows much higher predictability where the selected isovalue is 0.0128 and Figure 10i shows much more uncertainty about TEMP when a SiO2-FLUX-IN value close to 0 is selected which gives the surface of the ocean. Apart from the variables suggested by the system, the scientists also explored some other variables that they were interested in and they verified the outcomes of our system with their knowledge base.

After the entire process was completed, we asked for general and specific feedback about our system. From the feedback provided by the domain scientists, it is apparent that they are looking for new tools and are eager to try out systems similar to what we have proposed in this paper. Given a multivariate data set, exploration of the relationships among the variables is of prime interest to them. The scientists noted that the initial dendrogram representation tied to the graph layout provided them with new ways to identify classic expected relationships among the marine systems variables but additionally the visualizations raised new kinds of research issues which they felt was very useful. If the data sets do not show the relationship pattern as expected by the scientists, they can flag this timestep for more analysis as to whether this anomaly is due to some error in the simulation or indeed something unexpected has happened. The I_1I_2 map provides information about the predictability and surprise of the reference variable with respect to the other variable. The final visualization of the color-mapped isosurfaces can also effectively convey the variability information of the other variables when compared with the reference variable. In the scientists' opinion, this system provides information about the variables from a new perspective with an interactive interface which they think is useful for multivariate exploration. The scientists concluded that besides existing tools like Matlab and Ferret, they would like to use our system for multivariate analysis in the future.

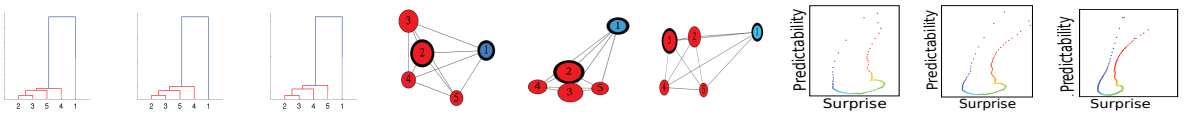
The scientists also provided suggestions to improve the system. They want to see our system integrated with the existing tools like scatter plot matrices so that it conveys more information. They also would like to see this exploration system to be extended in temporal domain for increased effectiveness. Specific to their domain needs, they would also like to explore in isopycnals (the constant density slices) that can reveal the mixing in the flow. In the graph layout, they suggested to add an option to remove the edges as they are liable to produce visual clutter when the number of variables is high. It was also apparent from the feedback that sometimes the domain scientists had some pre-selected variables in mind, like Salt, Temperature etc., which they were interested in exploring regardless of their information content. While sometimes the variables suggested by our initial graph structure did not match their primary interest, they agreed that for a large number of unfamiliar variables in the system, the knowledge of the relative information content is useful in making a choice.

7 Discussion

7.1 Parameter Choice

In this section, we discuss the choice of two parameters in our system. In the initial hierarchical clustering, the number of clusters k is a non-trivial choice. Finding the optimum k is a difficult problem [25] and it varies according to the data set. As a rule of thumb [29], we initially use $k \approx \sqrt{\frac{n}{2}}$, where n is the total number of variables in the system. The users can interactively change the value of k in their interaction phase to tune this parameter according to the property of the data sets.

The other important factor is the choice of bin size as it affects the calculation of the information-theoretic metrics. Generally, a higher number of bins will generate more precise information metrics. In Figure 11, we show the results when we vary the number of bins using



(a) Hierarchical clustering for 128 bins. (b) Hierarchical clustering for 256 bins. (c) Hierarchical clustering for 512 bins. (d) Graph layout at 128 bins. (e) Graph layout at 256 bins. (f) Graph layout at 512 bins. (g) I_1I_2 map of MIXFRAC against HR at 128 bins. (h) I_1I_2 map of MIXFRAC against HR at 256 bins. (i) I_1I_2 map of MIXFRAC against HR at 512 bins.

Fig. 11. Results with varying bin sizes for the Combustion data set.

128, 256, and 512 and it can be observed that the general pattern of the results remains the same, although there are some variations.

7.2 Performance

In this section, we discuss the performance of different components in our proposed framework as shown in Table 1. To create the joint distributions among multiple variables, instead of creating a large multidimensional array which will be very expensive in terms of the required storage space, we took advantage of the sparseness property in the array and used the Map data structure provided by C++ library to perform the Joint Entropy and all pair Mutual Information (M.I.) calculations. Without including the disk I/O time, the map creation time shown in the first column is dependent on the total number of data points for all the variables of the system. In the all pair M.I. calculation from the generated data map, it is observed that the run time varies depending on the distribution of the data values in the map. This all pair M.I. calculation is used to generate the force-directed layout of the graph. The run time of the hierarchical clustering and force-directed layout calculation is not listed here as it is in the order of milliseconds. Next, we presented the Joint Entropy calculation time within a group of variables which is required to figure out the relative importance of the variables. This is dependent on the group size and the distribution of the values in the data map. In the third column of the table, we have listed the longest running time which was measured by the time taken to update the importance of the variables belonging to the largest group. In our experimental settings, the largest group sizes are 4, 6, and 10 for Combustion, Isabel and Ion Front data set, respectively. Finally, the I_1I_2 map is generated from two selected variables and as shown in the fourth column, according to our current implementation, the calculation time is linear to the number of points in the data sets for a fixed number of bins. Generation of the maps, all pair mutual information and the joint entropy calculations can be done in a preprocessing stage, so that our framework can run interactively.

Table 1. Running Time for Different Components of the Framework

	Map Creation (Sec.)	All Pair M.I. (Sec.)	Joint Entropy (Sec.)	I_1I_2 Map Generation (Sec.)
Combustion	25.9	86.5	12.3	10.6
Isabel	38.8	1071.6	53.9	6.3
Ion Front	43.7	92.5	11.4	8.5

7.3 Comparison

In this section we compare our framework with some of the existing approaches of multivariate data analysis like scatter plots, scatter plot matrices, PCPs etc. In our framework, PCP, scatter plots and isosurfaces have been used to leverage the strengths of these visualization techniques alongside the information theory based analysis.

The traditional scatter plot is widely used for its ability to show the trend between two variables. But it is generally difficult to analyze a large amount of data in a scatter plot because of overlapping of the data points. Although this plot provides the idea of the spread of one variable for a given value of the other, the Predictability and Surprise metrics, which depend on the concept of probability distribution, may not be obvious from the plot. To remedy this, we initially calculate the specific information metrics and then use the scatter plot where the two axes are the Predictability and Surprise. Also, the number of points on this map only depends on the number of bins, whereas for the original scatter plot, the total number of points is the size of the data set, which may require efficient data structures for large data sets to provide the users with interactive brushing and selecting capabilities.

Scatter plot matrices (SPLOM) extend the idea of traditional scatter plot by plotting all pairs of scatter plots for all the variables at the same time. However, as the number of variables increases, SPLOMs tend to become cluttered. Also, it depends on the user's interpretation and memory if the user needs a global view of the system. In our framework, we present the relationship of the variables in the form of a hierarchical clustering tree which helps users get the overall structure of the variables in the tree and get the clustering information based on the information overlap of the variables. Additionally, the associated graph layout provides information about the group entropy and relative importance of these variables.

PCPs are generally regarded as an effective tool to visualize the correlations among multiple variables. But its effectiveness is governed by the order of the variable axes, the number of variables to be visualized and the total number of data points. While using PCPs, we used the mutual information based ordering and a downsampled version of the data sets to overcome the issues. Also, PCPs and SPLOMs provide information in the data domain disregarding the spatial domain. Taking this into consideration, alongside PCPs, our framework uses isosurfaces to show the spatial distribution of the scalars of the selected variable. Compared to the existing isosurface selection works, which are mostly concerned with univariate data, in the multivariate scenario we assign the importance to the isosurfaces when it depicts the variability of the other variables. As presented in this work, specific information metrics have been used to explore these relationships.

8 Conclusion and Future Work

In this paper, we presented an information-theoretic approach for the exploration of multivariate data sets. In our framework, we applied mutual information for generating an initial grouping of the variables such that the variables with high information overlap can be placed in the same group. The importance of the variables within the subgroups are identified by the calculation of conditional entropy. The variables are presented in the form of a force-directed graph layout to the users for interactive selection of variables. The selected variables are used to compute the specific information to identify the scalar values of one variable which are informative about the other variables. For the exploration in the data domain, the visualization is performed using Parallel Coordinate Plots. For exploration in the spatial domain, isosurfaces are generated which are color mapped to other variables to show the degree of uncertainty about that variable.

In future, we plan to expand our framework for time-varying data sets for selection of salient time steps and exploring additional types of data including ensemble data sets. We also plan to incorporate our framework with other existing multivariate analysis metrics for more effective exploration.

Acknowledgments

The authors wish to thank the domain experts of Los Alamos National Laboratory for their feedback on our proposed framework. The authors would also like to thank the anonymous reviewers for their comments. This work was supported in part by NSF grant IIS-1017635, IIS-1065025, US Department of Energy DOESC0005036, Battelle Contract No. 137365, and Department of Energy SciDAC grant DE-FC02-06ER25779, program manager Lucy Nowell.

REFERENCES

[1] H. Akiba, K.-L. Ma, J. H. Chen, and E. R. Hawkes. Visualizing multivariate volume data from turbulent combustion simulations. *IEEE Computing in Science and Engineering*, pages 86–93, March/April 2007.

[2] A. Artero, M. de Oliveira, and H. Levkowitz. Enhanced high dimensional data visualization through dimension reduction and attribute arrangement. In *Information Visualization, 2006. IV 2006. Tenth International Conference on*, pages 707–712, 2006.

[3] C. L. Bajaj, V. Pascucci, and D. R. Schikore. The contour spectrum. In *Vis '97: Proceedings of the IEEE Visualization*, VIS '97, pages 167–173. IEEE Computer Society Press, 1997.

[4] U. Bordoloi and H.-W. Shen. View selection for volume rendering. In *Visualization, 2005. VIS 05. IEEE*, pages 487–494, 2005.

[5] R. Bramon, I. Boada, A. Bardera, J. Rodriguez, M. Feixas, J. Puig, and M. Sbert. Multimodal data fusion based on mutual information. *Visualization and Computer Graphics, IEEE Transactions on*, 18(9):1574 – 1587, sept. 2012.

[6] S. Bruckner and T. Möller. Isosurface similarity maps. *Computer Graphics Forum*, 29:773–782, 2010.

[7] H. Carr, J. Snoeyink, and U. Axen. Computing contour trees in all dimensions. *Computational Geometry*, 24(2):75 – 94, 2003.

[8] M. Chen and H. Jänicke. An information-theoretic framework for visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1206–1215, 2010.

[9] J. Claessen and J. van Wijk. Flexible linked axes for multivariate data visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2310–2316, 2011.

[10] T. M. Cover and J. A. Thomas. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[11] A. Dasgupta and R. Kosara. Pargnostics: Screen-space metrics for parallel coordinates. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1017–1026, 2010.

[12] M. de Oliveira and H. Levkowitz. From visual data exploration to visual data mining: a survey. *Visualization and Computer Graphics, IEEE Transactions on*, 9(3):378–394, 2003.

[13] M. R. Deweese and M. Meister. How to measure the information gained from one symbol. *Network: Computation in Neural Systems*, (4):325–340, nov 1999.

[14] B. Duffy, H. Carr, and T. Möller. Integrating isosurface statistics and histograms. *IEEE Trans. Vis. Comput. Graph.*, 19(2):263–277, 2013.

[15] M. Feixas, E. D. Acebo, P. Bekaert, and M. Sbert. An information theory framework for the analysis of scene complexity, 1999.

[16] L. Gosink, C. Garth, J. Anderson, E. Bethel, and K. Joy. An application of multivariate statistical analysis for query-driven visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 17(3):264–275, 2011.

[17] S. Gumhold. Maximum entropy light source placement. In *Visualization, 2002. VIS 2002. IEEE*, pages 275–282, 2002.

[18] H. Guo, H. Xiao, and X. Yuan. Scalable multivariate volume visualization and analysis based on dimension projection and parallel coordinates. *Visualization and Computer Graphics, IEEE Transactions on*, 18(9):1397–1410, 2012.

[19] C. B. Hurley and R. W. Oldford. Pairwise display of high-dimensional information via eulerian tours and hamiltonian decompositions. *Journal of Computational and Graphical Statistics*, 19(4):861–886, 2010.

[20] A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(2):69–91, Aug. 1985.

[21] A. Inselberg. Multidimensional detective. In *Information Visualization, 1997. Proceedings., IEEE Symposium on*, pages 100–107, 1997.

[22] A. Inselberg and B. Dimsdale. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *Visualization, 1990. Visualization '90., Proceedings of the First IEEE Conference on*, pages 361–378, 1990.

[23] H. Jänicke, M. Bottinger, and G. Scheuermann. Brushing of attribute clouds for the visualization of multivariate data. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6):1459–1466, 2008.

[24] J. Johansson and M. Cooper. A screen space quality method for data abstraction. *Computer Graphics Forum*, 27(3):1039–1046, 2008.

[25] Y. Jung, H. Park, D.-Z. Du, and B. L. Drake. A decision criterion for the optimal number of clusters in hierarchical clustering. *J. of Global Optimization*, 25(1):91–111, 2003.

[26] T. Kamada and S. Kawai. An algorithm for drawing general undirected graphs. *Inf. Process. Lett.*, pages 7–15, apr 1989.

[27] M. Khoury and R. Wenger. On the fractal dimension of isosurfaces. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1198 –1205, 2010.

[28] L. F. Lu, M. L. Huang, and T.-H. Huang. A new axes re-ordering method in parallel coordinates visualization. In *Machine Learning and Applications (ICMLA), 2012 11thInternational Conference on*, volume 2, pages 252–257, 2012.

[29] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis (Probability and Mathematical Statistics)*. Academic Press, first edition, second impression edition, 1980.

[30] A. Martin and M. Ward. High dimensional brushing for interactive exploration of multivariate data. In *Visualization, 1995. Visualization '95. Proceedings., IEEE Conference on*, pages 271–, 1995.

[31] W. Peng, M. Ward, and E. Rundensteiner. Clutter reduction in multidimensional data visualization using dimension reordering. In *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*, pages 89–96, 2004.

[32] J. Rigau, M. Feixas, and M. Sbert. Informational aesthetics measures. *Computer Graphics and Applications, IEEE*, 28(2):24–34, 2008.

[33] O. Rubel, Prabhat, K. Wu, H. Childs, J. Meredith, C. G. R. Geddes, E. Cormier-Michel, S. Ahern, G. Weber, P. Messmer, H. Hagen, B. Hamann, and E. Bethel. High performance multivariate visual data exploration for extremely large data. In *High Performance Computing, Networking, Storage and Analysis, 2008. SC 2008. International Conference for*, pages 1–12, 2008.

[34] C. Scheidegger, J. Schreiner, B. Duffy, H. Carr, and C. Silva. Revisiting histograms and isosurface statistics. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1659 –1666, 2008.

[35] J. Siek, L.-Q. Lee, and A. Lumsdaine. Boost graph library. http://www.boost.org/libs/graph/, June 2000.

[36] P.-P. Vázquez, M. Feixas, M. Sbert, and W. Heidrich. Automatic View Selection Using Viewpoint Entropy and its Application to Image-Based Modelling. *Computer Graphics Forum*, 22(4):689–700, 2003.

[37] I. Viola, M. Feixas, M. Sbert, and M. Groller. Importance-driven focus of attention. *Visualization and Computer Graphics, IEEE Transactions on*, 12(5):933–940, 2006.

[38] C. Wang, H. Yu, R. Grout, K.-L. Ma, and J. Chen. Analyzing information transfer in time-varying multivariate data. In *Pacific Visualization Symposium (PacificVis), 2011 IEEE*, pages 99–106, 2011.

[39] C. Wang, H. Yu, and K.-L. Ma. Importance-driven time-varying data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1547–1554, nov 2008.

[40] D. Whalen and M. L. Norman. Competition data set and description. 2008 IEEE Visualization Design Contest, http://vis.computer.org/VisWeek2008/vis/contests.html, 2008.

[41] P. C. Wong and R. D. Bergeron. 30 years of multidimensional multivariate visualization. In *Scientific Visualization, Overviews, Methodologies, and Techniques*, pages 3–33, Washington, DC, USA, 1997. IEEE Computer Society.

[42] Y. Xiang, D. Fuhry, R. Jin, Y. Zhao, and K. Huang. Visualizing clusters in parallel coordinates for visual knowledge discovery. In *Proceedings of the 16th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining - Volume Part I*, PAKDD'12, pages 505–516, 2012.

[43] L. Xu, T.-Y. Lee, and H.-W. Shen. An information-theoretic framework for flow visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1216–1224, 2010.

[44] D. Yang, E. Rundensteiner, and M. Ward. Analysis guided visual exploration of multivariate data. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*, pages 83–90, 2007.

[45] Z. Zhang, K. T. McDonnell, and K. Mueller. A network-based interface for the exploration of high-dimensional data spaces. *Visualization Symposium, IEEE Pacific*, pages 17–24, 2012.

[46] J. Zhou and M. Takatsuka. Automatic transfer function generation using contour tree controlled residue flow model and color harmonics. *IEEE Transactions on Visualization and Computer Graphics*, 15:1481–1488, 2009.