Visual Exploration of High Dimensional Scalar Functions

Samuel Gerber, Peer-Timo Bremer, Valerio Pascucci, and Ross Whitaker

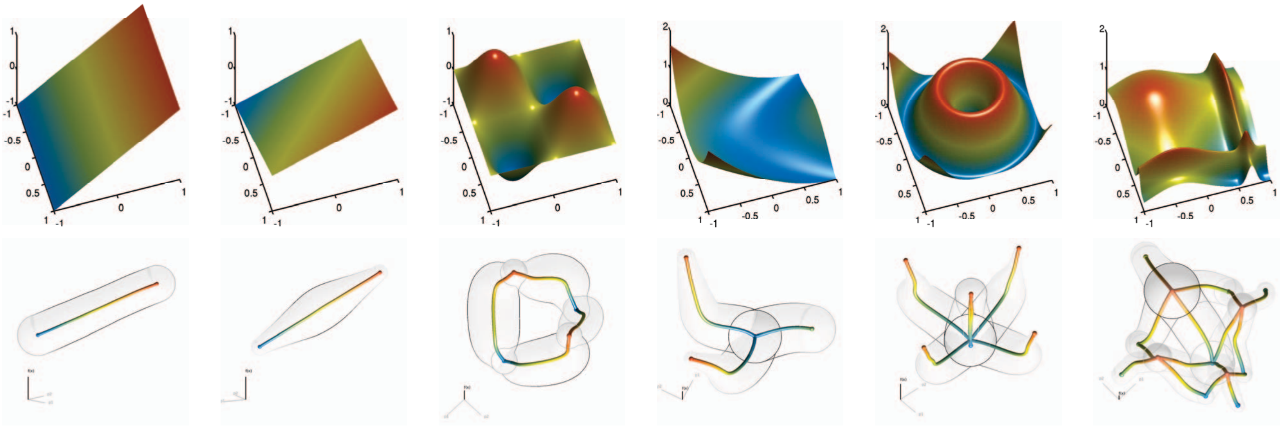


Fig. 1. The proposed visualization illustrated on several two-dimensional scalar fields. In the bottom row, each curve represents a monotonic region of the 2D domain, a geometric summary for each crystal of the Morse-Smale complex of the function above.

Abstract—An important goal of scientific data analysis is to understand the behavior of a system or process based on a sample of the system. In many instances it is possible to observe both input parameters and system outputs, and characterize the system as a high-dimensional function. Such data sets arise, for instance, in large numerical simulations, as energy landscapes in optimization problems, or in the analysis of image data relating to biological or medical parameters. This paper proposes an approach to analyze and visualizing such data sets. The proposed method combines topological and geometric techniques to provide interactive visualizations of discretely sampled high-dimensional scalar fields. The method relies on a segmentation of the parameter space using an approximate Morse-Smale complex on the cloud of point samples. For each crystal of the Morse-Smale complex, a regression of the system parameters with respect to the output yields a curve in the parameter space. The result is a simplified geometric representation of the Morse-Smale complex in the high dimensional input domain. Finally, the geometric representation is embedded in 2D, using dimension reduction, to provide a visualization platform. The geometric properties of the regression curves enable the visualization of additional information about each crystal such as local and global shape, width, length, and sampling densities. The method is illustrated on several synthetic examples of two dimensional functions. Two use cases, using data sets from the UCI machine learning repository, demonstrate the utility of the proposed approach on real data. Finally, in collaboration with domain experts the proposed method is applied to two scientific challenges. The analysis of parameters of climate simulations and their relationship to predicted global energy flux and the concentrations of chemical species in a combustion simulation and their integration with temperature.

Index Terms—Morse theory, High-dimensional visualization, Morse-Smale complex.

1 Introduction

Visual representations of high-dimensional scalar fields are becoming an increasingly important challenge in a variety of fields. To illustrate the problem, consider the manufacture of concrete. The recipe, or ingredients, for concrete consists of various mixtures of a variety of constituents, such as rock, cement, and water, as well as age. A quantitative measure of the success of such a particular recipe is *compressive strength*. Different aspects, or parameters, of the concrete recipe can interact to impact the compressive strength in complicated, nonlinear relationships. A typical regression analysis provides the mathematical relationship, but visualizing and understanding the resulting high-dimensional structure is still quite difficult and does not directly answer many of the relevant questions. In particular, a civil engineer might like to know if there are multiple distinct recipes for *strong* concrete. Additionally, one may want to understand how the recipes for weak concrete differ from these optimal mixtures, and what particular deviations from ideal should be avoided. Furthermore, an engineer might like to know how to make small modifications to a current recipe in order to realize incremental improvements, and what the risk is that these changes could make things worse. A similar set of problems arises in numerical simulations, where a great variety of free parameters can interact to affect the results. Indeed, the parameters in a simulation are the *recipe* for achieving certain quantitative outcomes, and there exists a set of questions analogous to those in the concrete example. Our proposition is that this kind of analysis demands new visualization tools that can aggregate data and effectively reduce the dimensionality while respecting the important structure introduced by the output variable. These tools need to capture not only global information, such as the overall topology of these relationships, but also local information, such as the geometry of these functions.

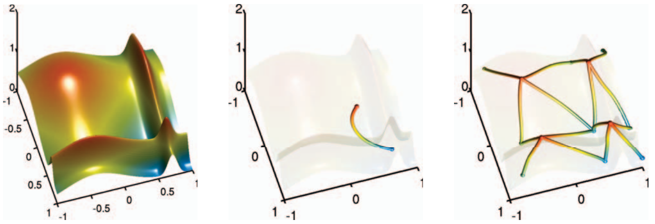
The relationship of concrete mixtures and compressive strength can be represented as a high dimensional scalar function $y = f(\mathbf{x})$, where $\mathbf{x} \in \mathbf{R}^d$ are the parameters (ingredients and recipe) and $y \in \mathbf{R}$ is the output (compressive strength). Conventional multiple regression of f assumes a set of samples $y_i = f(\mathbf{x}_i)$, and attempts to reconstruct f for the entire domain. Of course, the number of samples must be larger than the degrees of freedom in the model, and in high-dimensional spaces model selection becomes a critical problem. The resulting surrogate model of f may subsequently be used to predict the output for new inputs and for analysis in lieu of f. The goal of this paper is sub-

- *Samuler Gerber, Valerio Pascucci and Ross Whitaker are with the Scientific Computing and Imaging Institute, University of Utah.*
- *Peer-Timo Bremer is with the Center of Applied Scientific Computing (CASC), Lawrence Livermore National Laboratory.*

Manuscript received 31 March 2010; accepted 1 August 2010; posted online 24 October 2010; mailed on 16 October 2010.
For information on obtaining reprints of this article, please send email to: tvcg@computer.org.

tly, but significantly, different from typical multiple regression. We are not aiming to interpolate or extrapolate f, but to analyze and visualize its structure using the existing samples to provide insight into the relationship between the parameters and the output. In particular, our goal is to understand: i) the extreme output values (how many there are and their location); ii) their connection in parameter space (how can one continuously modify the inputs to get from some local minimum to some local maximum); and iii) the inverse relationships indicating which combinations of inputs are responsible for which output. More generally, we are interested in the *topology* of f, which describes the extreme values of the output, as well as the geometry of the regions connecting them. Due to the large number of parameters, a simultaneous visualization of all of the data is impossible. Therefore, we propose to aggregate the data into topologically-based *summaries* of distinct regions or regimes in the parameter space.

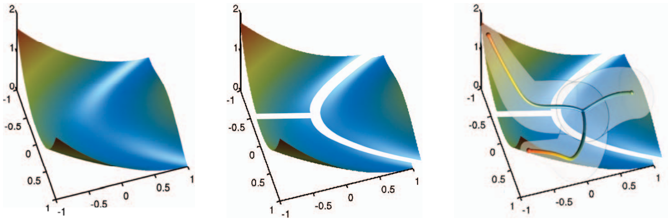
Rather than modeling f directly, we propose to model some approximate of the inverse relationship $\mathbf{g} : \mathbf{R} \mapsto \mathbf{R}^d$, using multivariate kernel regression. Here, $\mathbf{g}$ describes the conditional expectation $E[X|y]$, and each $\hat{\mathbf{x}} = \mathbf{g}(y)$ is the representative of the parameters that achieve a particular output y. This reverses the relationship of independent and dependent variables compared to multiple regression analysis of f. However, if f is not monotonic and thus not invertible, the data aggregation implied by $\mathbf{g}$ can lead to a loss of important information. For example, if different maxima are averaged their combined location will not resemble the actual locations, see Fig. 2(b). Instead, we propose to decompose f's domain into *piecewise monotonic* cells (or crystals) using an approximate Morse-Smale complex, see Fig. 2(c). Furthermore, filtering the data, based on topological properties, allows to analyze $\mathbf{g}$ at different scales, making the framework highly flexible and robust against noise.



(a) (b) (c)

Fig. 2. Sum of four Gaussian kernels of various shapes on $\mathbf{R}^2$ (a) represent by a locally linear regression curve (b) and the proposed approach that computes regression curves of piecewise monotonic regions (c). Note that, the independent variable of the regression corresponds to the vertical axis.

The framework described in this paper combines topological and geometric information to generate simplified lower dimensional representations that preserve important information about the high-dimensional scalar fields. This approach leads to a set of tools for interactive exploration of the input domain as well as statistical analysis. For each crystal we construct a single regression curve connecting the minimum to the maximum which results in a sparse yet highly informative representation of the input domain, see Fig. 3. The curves are embedded into two dimensions for visualization, using conventional dimension reduction techniques. The third dimension in the visualization is used to to represent the independent variable of g, which in this case is the output of f. Thus, the visualization can be understood as an approximation to the d-dimensional surface described by f. The visualization describes not only the topology, but also provides additional information about each cell's geometry, including local information such as tangents and curvature, as well as global information such as size and sampling density, see Fig. 4.



(a) (b) (c)

Fig. 3. Schematic illustration of the proposed method. The scalar function (a) is decomposed into piecewise monotonic regions (b) and each region is approximated by a regression curve (c).

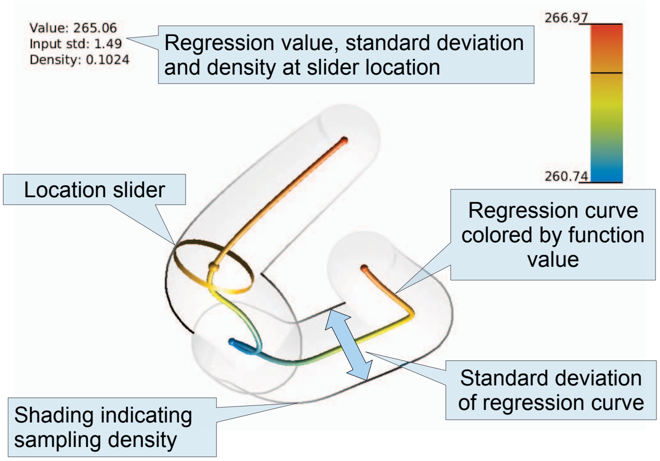

Value: 265.06
Input std: 1.49
Density: 0.1024
Regression value, standard deviation and density at slider location
266.97
260.74
Location slider
Regression curve colored by function value
Standard deviation of regression curve
Shading indicating sampling density

Fig. 4. Visualization of the approximate Morse-Smale complex. Each curve represents a segmented region of the input. Additional geometric and statistical properties of the regression curves provide additional insight into the structure of the segmented regions.

The main contributions in this paper are:

1. A new inverse regression scheme based on the Morse-Smale complex for functions defined on point clouds;
2. A sparse representation of the high dimensional Morse-Smale complex based on the inverse regression scheme;
3. A visualization approach of the sparse representation using a two layered dimension reduction approach;
4. A linked view interface allowing the user to interactively explore the geometry and topology of a high dimensional function at multiple scales; and
5. Detailed studies on the effects of parameter choices influencing the computation and the expressive power of the proposed approach.

2 Background and Related Work

Visualization in High Dimensions. By far the most common approach to explore high dimensional spaces is to use projections onto one, two, or three dimensional subspaces and showing scatter plots, smooth approximations, or labels corresponding to the projected positions of the data within this subspace. Common projection approaches, based on statistical measures of the data, are principal component analysis [36] and projection pursuit [24]. In a slightly different direction researchers have proposed to look at curves that describe sequences of projections of the data onto different directions as in Andrews plots [1], which are closely related to parallel coordinates [34]. The *grand tour* approach [2] extends this strategy to a sequence of projections onto multiple planes. More recently, the projection approach to visualize high dimensional data has been used in conjunction with nonlinear dimensionality reduction methods [58, 39]. An interesting variation with embeddings onto the hyperbolic plane is proposed by Walter et al. [54]. In the machine learning community embedding is a common strategy for quantifying the effectiveness of manifold learning approaches [5, 33, 47, 50].

Dimension Reduction and Regression. Dimension reduction and manifold learning approaches [5, 33, 47, 50] are appropriate for scattered data that has an underlying low dimensional structure which is not explicit in the formulation of the problem. When analyzing high dimensional scalar functions, however, one is less interested in the geometry of the domain but in the geometry induced by the dependent variable. In the context of scalar functions, dimension reduction approaches operate on the domain of the function. Often the set of samples in the domain is not reducible, i.e. the function is truly defined on the complete data domain and not only on a lower dimensional structure embedded in the domain. Thus, classical dimension reduction approaches are not applicable. The proposed method combines a topological decomposition with dimension reduction to arrive at a low dimensional representation that succinctly describes the scalar field.

Another related area of research is regression analysis focusing on the parametrization and/or segmentation of the *independent* variables. Multivariate adaptive regression splines (MARS) [25] or kernel regression [43, 55] can be formulated as averaging local models in the parameter domain—the independent variables. In regression and classification trees [8, 13] the independent variables are explicitly partitioned into regions and the regression estimate is calculated by a smooth averaging scheme. However, understanding the parameters from the point of view of the output variable is another matter. The inverse approach of partitioning the independent variables based on the topology of the function f is a unique contribution of this paper.

Topological Analysis. For lower dimensional scalar data, topology-based techniques have been proposed in a wide variety of applications. Topological structures such as the Reeb graph [46, 45, 7, 35, 11] or the Morse-Smale complex [23, 9, 29, 30] provide an abstract representation of scalar fields well suited for analysis. They can be used to define a wide variety of features in various applications, ranging from medical [12], to physics [38, 10] and material science [28].

Algorithms have been proposed for computing topological structures [45, 11, 30] on n-dimensional manifolds. Furthermore, there exist some extensions to point cloud data. Harvey and Wang [31] use a k-nearest neighbor graph to compute a contour tree which is subsequently displayed as an improved version of a topological landscape [56] . In a related approach Oesterling et al. [44] use the landscape of merge trees to illustrate point cloud densities.

The Morse-Smale complex is defined as the intersection of the Morse complex of f and $-f$. Computing a Morse complex is a well known concept in several areas, albeit under different names. In computational geometry the Morse complex is often described in terms of a filtration of the sub-levelsets of f. Chazal et al. [14, 15] use nested Rips complexes to compute the Morse complex of a function f given at sample points of a (low-dimensional) manifold embedded into high dimensional space. To remove noise and artifacts they construct the persistence diagram of f and prove that it is stable under small perturbations. For the final segmentation they use the algorithm of Zhu et al. [60] which is, apart from the choice of *steepest* edge and simplification procedure, very similar to the one used in Section 3.

In image processing the Morse complex is known as watershed segmentation [22, 6, 40] and has been described for n-dimensional grids as well as for abstract graphs [53]. Finally, in the context of pattern recognition and machine learning, the Morse complex can be thought of as a variant of mean shift clustering [16, 26, 19] in which the kernel density estimation is replaced with f. Recent graph based variants such as medoid shift [49] and quick shift [52] are very similar to the algorithms described in Zhu et al. [60] and Chazal et al. [14] except for the choice of neighborhood and simplification metric. In general, any gradient ascent style clustering [27] could be used to construct a Morse complex like segmentation.

Morse Theory. The techniques presented in this paper are based on Morse theory [42, 41] and in particular on the notion of Morse-Smale (MS) complexes [23]. We briefly outline the basic theory and discuss a few of the more common approximations. We refer the reader to [23] for a more formal discussion.

Let $\mathbf{M}$ be a smooth manifold without boundary and $f : \mathbf{M} \rightarrow \mathbf{R}$ a smooth function with gradient ∇f. A point $\mathbf{x} \in \mathbf{M}$ is called *critical* if $\nabla f(\mathbf{x}) = 0$ or *regular* otherwise. At any regular point $\mathbf{x}$ the gradient (vector) is well-defined and integrating it in both directions traces out an *integral line* γ, $\gamma(s) = \nabla f(\gamma(s))$ which starts at a minimum and ends at a maximum. The ascending/descending manifold of a critical point $\mathbf{c}$ is defined as all points whose integral lines start/end at $\mathbf{c}$. The descending manifolds form a complex called the Morse complex and the ascending manifolds define the Morse complex of $-f$. The set of intersections of ascending and descending manifolds creates the MS-complex of f.

The complex consists of a set of *crystals* formed by the union of integral lines that start and stop at the same extremal points. These crystals yield a decomposition into monotonic, non-overlapping regions $\mathbf{D}_i \subset \mathbf{M}$ of the domain. This observation is important when representing the function with one regression curve per crystal. The monotonicity ensures that the level sets of f within each $\mathbf{D}_i$ are topological disks of the appropriate dimension. As a result, computing $\hat{\mathbf{x}} = \mathbf{g}(y) = E[X|y]$ is guaranteed to not average topologically distinct structures, such as multiple extrema, which could distort the results.

3 Methodology

In this paper we consider functions represented as a finite set of points X in a high dimensional space $\mathbf{R}^d$ and a set of corresponding scalar values Y, i.e. a discrete set of samples of a function $f : \mathbf{R}^d \rightarrow \mathbf{R}$ with $y_k = f(\mathbf{x}_k)$.

The proposed approach consists of three steps to arrive at a 2D representation for visualization of the high dimensional scalar function:

1. **Morse-Smale Approximation:** Compute segmentation X_i and Y_i using the Morse-Smale approximation of f based on X and Y.
2. **Geometric Summaries:** Construct regression curves $\mathbf{r}_i$ as a geometric summary of each segmentation X_i and Y_i.
3. **Dimension Reduction:** Embed regression curves in 2D using a two-step dimension reduction approach.

The first step captures the topological properties of the data. The second step provides geometric information about the data while preserving the topological structure. Finally, the third step provides a representation suitable for visual exploration. In the following we describe each step in detail and illustrate how the proposed technology is used to gain insight into the structure of high dimensional scalar fields.

Morse-Smale Approximation. To compute the ascending and descending manifolds we use a variant of the quick shift algorithm [52]. At each vertex, we compute the k-nearest neighbor graph of X and among these choose the steepest (a/descending) edge to represent the gradient. All vertices that have no a- or descending gradient assigned are local extrema and we label all vertices of X according to the local extrema its a-/descending gradient will terminate. The resulting complexes contain a region for each local extrema that represent the a-/descending manifolds. Note that our complexes can be over-segmented and may need to merge regions as compared to the initial under-segmentation of [52]. Subsequently, we collect all vertices with the same label pair into *crystals* X_i and add the extrema to all crystals that share the corresponding label. This set of crystals is then used as an approximation to the MS-complex of f.

Geometric Summaries. For each crystal of the Morse-Smale complex, a geometric summary is constructed by an inverse regression. This yields a 1D curve in the d-dimensional domain of f.

Formally, the input domain for each crystal C_i with samples (X_i, Y_i) of the MS-complex is summarized by a parametric curve $\mathbf{r}_i : [\min_{\mathbf{x} \in C_i} f(x), \max_{\mathbf{x} \in C_i} f(x)] \mapsto \mathbf{R}^d$. Modeling the curve by the conditional expectation $r_i(y) = E[X \in C_i | Y = y]$ yields a representation of the crystal C_i as the average of the level sets $\{x : f(x) = y, x \in C_i\}$ within the partition. The conditional expectation $E[X \in C_i | Y]$ is estimated with locally linear regression [17] and can be written as

$$\mathbf{r}_i(y) = (\bar{Y}_i W(y) \bar{Y}_i^T)^{-1} \bar{Y} W(y) X_i^T .\mathbf{u}_1. \tag{1}$$

Let $n_i = |X_i|$, the number of points in crystal i. Then $W(y)$ is a $n_i \times n_i$ diagonal matrix with $W_{k,k} = K(y, y_k)$ and K a kernel function. $\bar{Y}_i = (\mathbf{1}, Y_i)$ a matrix with the first row all ones and the second row the function values of crystal i, and $\mathbf{u}_1 = (1,0)^T$. Thus, $\mathbf{r}_i(y)$ is estimated by a weighted linear fit to X_i, a first order kernel regression, with contribution of point x_k decreasing with increasing distance of $\|y_k - y\|^2$. In this paper, we use a Gaussian kernel $K(y, y_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{|y-y_i|^2}{2\sigma^2}}$ with the kernel bandwidth σ a free parameter — since the bandwidth choice is over the range, i.e. a scalar value, it can be readily set according to the given data.

The geometric properties of the curve provide additional information about the Morse-Smale crystal that we visualize in the low dimensional embedding. The gradient, or tangent vector, of the regression curve is directly available from the linear approximation in the regression computation,

$$\frac{d}{dy}\mathbf{r}_i(y) = (\bar{Y}_i W(y) \bar{Y}_i^T)^{-1} \bar{Y} W(y) X_i^T \mathbf{u}_2, \tag{2}$$

and gives a measure of the local *sensitivity* of the input coordinates. The average distance of the data to the curve as a function of the parameter y gives important information about the shape of the crystal. The coordinate-wise standard deviation is calculated by

$$\delta_i(y) = \sqrt{(\sum_j^{n_i} \frac{K(y, y_j)\rho_i(\mathbf{x}_j)}{\sum_k^{n_i} K(y, y_j)})}, \tag{3}$$

with $\rho_i(\mathbf{x}_j)$ the vector of squared projection residuals of $\mathbf{x}_j$ onto $\mathbf{r}_j$ computed by $\rho_i(\mathbf{x}_j)_m = (\mathbf{r}_i(y_j) - \mathbf{x}_j)_m^2$, here $m = 1 \ldots d$ indicates the component of the vector. The *average*, direction independent, standard deviation is computed by

$$s_i(y) = \sqrt{(\sum_j^{n_i} \frac{K(y, y_j) d_i(\mathbf{x}_j)}{\sum_k^{n_i} K(y, y_j)})}, \tag{4}$$

with $d_i(\mathbf{x}_j)_m = \|\mathbf{r}_i(y_j) - x_j\|^2$. The sampling density along the curve

$$p_i(y) = \frac{1}{|X|} \sum_j^{n_i} K(y, y_j), \tag{5}$$

provides information about the number of points used in the computation of the regression curve point at y and gives an indication, in combination with the standard deviation, of how densely sampled the crystal in that region is.

To ensure consistency of the endpoints of regression curves that share a maximum or minimum, we add the points that share crystals for computing the regression curve. For each neighboring crystal X_n we include its points x_k in the kernel estimation, but modify the corresponding scalar values $y_k \in Y_n$ by $y_k = 2\max(Y_i) - y_k$ for maxima and by $y_k = 2\min(Y_i) - y_k$ for minima. This ensures a more accurate estimation of extremal point locations (i.e. more data) and a smooth transition into curves associated with adjacent crystals, while guaranteeing that end points of distinct, adjacent curves coincide at extremal points. At the same time this approach does not significantly distort the regression curve with points from other crystals.

We represent the curves as polylines with a dense sampling. Thus, an equidistant sampling $S = \{s_1, \ldots, s_{\#samples}\}$ with $s_j = \min Y_i + jh$, with $h = \frac{\max(Y_i) - \min(Y_i)}{\#samples - 1}$ of the range of Y_i yields a piecewise linear approximation $L = \{\mathbf{l}_1, \ldots, \mathbf{l}_{\#samples}\}$ with $\mathbf{l}_j = \mathbf{r}_i(s_j)$ to the high dimensional regression curve. This linear approximation is in the next step embedded into 2D for visualization.

Dimension Reduction. The set of regression curves can be represented by a graph embedded in R^d with each edge corresponding to a curve and vertices corresponding to extremal points. For visualization, we embed this graph into the plane preserving the spatial relation among extrema and the geometry of the partitions that connect them as best as possible. It is important to point out that the goal of this dimension reduction is to provide an informative illustration of f rather than manifold learning of X. We compute the projection into the plane using a three step approach: first, vertices are embedded; second edges are embedded individually; and third, the resulting two-dimensional curves are attached to the projected vertices through affine transformations.

The extrema are embedded into two dimension using the principal components of the corresponding point set. Specifically, the extremal points $E = [\mathbf{e}_1, \ldots, \mathbf{e}_k]$ are projected onto their first two principal components $C_e = [\mathbf{c}_1, \mathbf{c}_2]$ with the projections $P_e = C_e^T E$. For the second step, each linearly approximated regression curve L_i is separately projected by $P_{L_i} = C_{L_i}^T L_i$, with C_{L_i} the first two principal components of L_i. In the third step, the projected curve P_{L_i} is *connected* to its corresponding minimum and maximum with an affine transformation. In this way the directions of maximal variance for each curve are retained and capture how much the curve deviates from a straight line connecting the minimum and maximum.

Alternatively, one could consider a direct embedding of the high-dimensional graph obtained from the Morse-Smale complex. However, this graph is very sparse (few edges) and this can lead to distortions that do not reflect the relative locations of the extremal points. For a graph based embedding approach, the distortions are less predictable. Depending on the structure of the graph, extrema that have no direct edge connecting them can get pushed far apart. A PCA based dimension reduction is easier to interpret, since the distortion induced can only move points closer to one another and not further apart. In any case, the structure of the graph does not depend on the embedding method. However, the user needs to be aware of the specific distortions, depending on the approach used, that are induced on extrema locations and the regression curves connecting them.

If the support of the domain is suspected to form a lower dimensional subspace one should consider reducing the dimensionality of the data set first. If the reduced dimension is small enough, simpler approaches for visualization of the scalar function can be pursued. Otherwise, the proposed approach can be applied to the dimension reduced data. Optionally, a more sophisticated graph embedding could be considered, for example, a manifold learning approach that jointly embeds the sampled regression curves S_i and the original data X to avoid the pitfalls of sparse graphs. However, the focus of the presented methodology is on visualizing scalar functions defined on a high-dimensional domain. A discussion of the many possibilities and challenges involved with manifold structured domains is beyond the scope of this paper.

3.1 Level of Detail

As shown in [14], the complexes of ascending and descending manifolds can be simplified by *persistence*, which, in this case, is directly related to the L_∞-norm. Furthermore, it has be shown that *features* with large persistence are stable under small perturbations and thus robust against noise and/or sampling variations. We extent the concepts of [14] by performing two simultaneous simplifications on the two complexes which amounts to simplifying both maxima (in the descending complex) and minima (in the ascending complex). The simplification merges neighboring extrema and we adapt the MS complex accordingly.

The crystals of the simplified MS-complexes will not be monotonic with respect to the original function, but they will be monotonic on filtered versions of f, which if necessary, can be constructed within an L_∞ error no larger than the half of the persistence of the last cancellation [23, 9]. At the coarsest level, only a single maximum and minimum remain and the method is equivalent to a multivariate regression analysis. At finer levels, more detailed information about the topology is represented. The standard deviation of the regression curve representing the crystal holds information about the crystal. For example, if during simplification two far apart extrema are merged the standard deviation will increase accordingly. The sampling density provides additional information about the reliability of a crystal. A low sampling density indicates that the crystal is potentially spurious due to an under-sampling of the corresponding region in the domain.

To understand how the MS complex changes with increasing simplification we use a modified version of a persistence diagram [18]. Instead of showing the traditional bar codes indicating the lifetime of a feature, we use a persistence graph that shows the number of features as a function of persistence scaled by the global function range, see Fig. 6. While the traditional diagram corresponds to a single filtra-

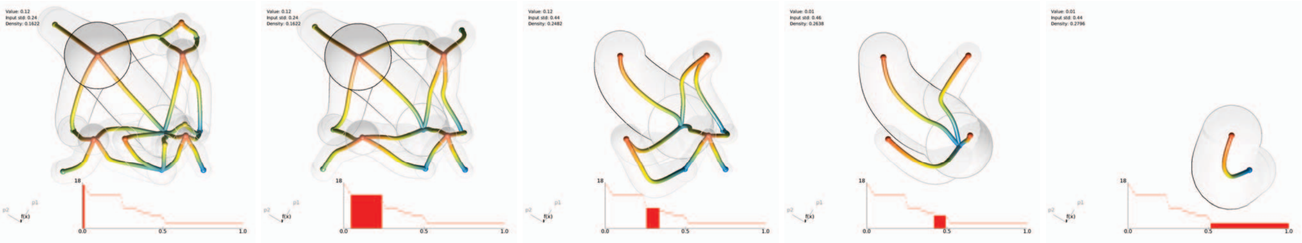


Fig. 5. The 2D Gaussian example at different levels of detail of the Morse-Smale approximation.

tion, we combine the information from the ascending and descending manifold simplification by showing the number of extrema (both maxima and minima) as a function of persistence. Therefore, the graph indicates which extremal points are due to noise and how many can be accurately represented with the amount of data available: Extremal points with a low persistence are most likely due to noise and/or under sampling while plateaus represent a stable number of extremal points that need a large amount of change to be simplified.

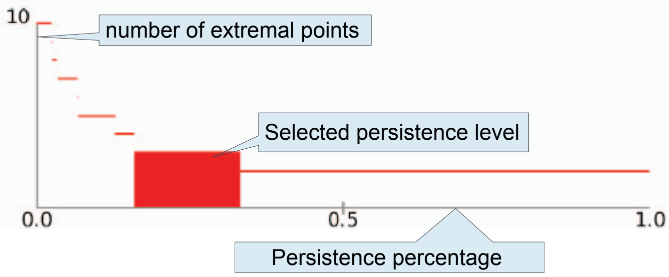


Fig. 6. The persistence graph shows the number of extremal points as a function of their persistence. The persistence graph provides information about the level of detail of the visualization and an indication of the number of reliable extremal points.

Fig. 5 shows a sequence of increasing persistence levels on the 2D Gaussian example. The topological changes can result in very different geometries, and to avoid disorienting transitions between embeddings, we anchor the embedding using the single crystal arising from to the lowest level of detail (highest persistence). Each subsequent level of detail is translated such that the maximum of the anchor is matched with the corresponding maximum in the current embedding and then rotated and scaled such that the shared extremal points of the current level of detail with the previous level match as best possible (minimum L_2 norm).

3.2 Discussion of Morse-Smale Complex Approximation

While the described method provides a powerful set of tools to analyze high dimensional functions, it is important to realize their limitations. In the lower dimensional cases, a significant effort has been made to guarantee that the resulting MS complex is structurally correct, meaning there exists some smooth function $\bar{f}$ with the given MS complex. This notion of consistency does not generalize when intersecting independently computed stable and unstable manifolds. Furthermore, by representing crystals implicitly as intersections, there exists no sense of saddles or boundaries between crystals. While this makes the computation tractable, we can no longer distinguish two crystals with the same extrema pair. Instead, such pairs are combined into a single crystal and represented by a single regression curve. Since such crystals accumulate points from spatially distinct locations one would expect their regression to exhibit large standard deviations (large widths).

Finally, the approximate MS-complexes are dependent on the size of the neighborhood, the number of samples with respect to dimensionality and number and *size* of the features present. To analyze how accurately the MS-complex recovers features, we ran various experiments on synthetic examples with two, three, and four maxima, equidistantly distributed on the diagonal of hypercubes with dimensions two to six. For each combination, we computed 50 random sets of $64, 128, \dots, 4096$ samples and 5, 10, 20, and 50 nearest neighbors. We then computed the number of maxima present in the MS-complex versus persistence for each combination and averaged the resulting curves over the 50 instantiations. These persistence curves are a good measure on how well the approximate MS-complex is capable to detect features—for a good approximation, a significant plateau is visible for the correct number of peaks.

The analysis is based on the function f defined by a diagonal sine curve of amplitude 0.5 with frequency determined by the number of maxima and multiplied with a Gaussian kernel orthogonal to the diagonal (Fig. 7). Thus, f varies between -0.5 and 0.5 and is symmetric around 0. We define the *feature size* as the percentage of the domain with $f(x) \geq 0.1$ divided by the number of features present. Table 1 shows the feature sizes for the different number of maxima as well as the lowest number of samples among our experiments that still showed a noticeable plateau, see Fig. 7.

Table 1. Feature volume and number of necessary samples to distinguish all features for the synthetic test function at various dimensions and numbers of features.

	Feature Volume in % | Necessary # of samples					
	2 Maxima		3 Maxima		4 Maxima	
2D	14.9	32	8.9	64	6.2	128
3D	8.5	64	4.7	256	3.2	512
4D	4.5	128	2.3	256	1.6	1024
5D	2.3	128	1.1	512	0.8	2048
6D	1.2	256	0.5	1024	0.4	4096

Contrary to what one might expect our experiments suggest that the ability of the MS-complex to correctly detect features shows virtually no dependence on the dimension of the domain—as long as the volume or spatial extent of the features grows proportionally to the space. Instead, the necessary number of samples is primarily defined by the relative volume of the features present and to a lesser extent by how many features exist overall. In the case of a uniform sampling of the domain, the probability of a sample falling into a feature of p percent of the domain volume is p. Given n samples, the probability of a sample falling into that feature is $(1-(1-p)^n)$. This yields estimates about the size of features that can be expected to be detected for a given number of samples. A more sophisticated analysis would consider the number of samples required for sampling a given partition, as induced by the Morse-Smale complex, of the domain. The coupon collector problem [4] yields the expected number of samples for k equal sized partitions and is approximately $k\log(k)+0.6*k$. However, this alone does not guarantee that the Morse-Smale complex computation will succeed, but provides a lower bound.

The ideal number of neighbors, however, depends directly on the dimension. Higher dimensions need a larger set of neighbors to identify features, see Fig. 8. The more neighbors are used, the more stable the segmentation becomes as the neighborhood acts as a low-pass filter. However, once the number of neighbors becomes to large relative to the overall number of samples, features are lost due to excessive smoothing. For a regular grid of dimension n there are 3^n-1 neighbors within a distance of $\sqrt{n}$. Therefore, in a simplicial mesh one expects the number of neighbors to rise exponentially with respect to n. Fortunately, our experiments suggest that the necessary numbers of neighbors behaves closer to linear as a function of dimension. One possible explanation is that the numbers of necessary neighbors is only related to the number of closest neighbors which in a regular grid would be $2n$.

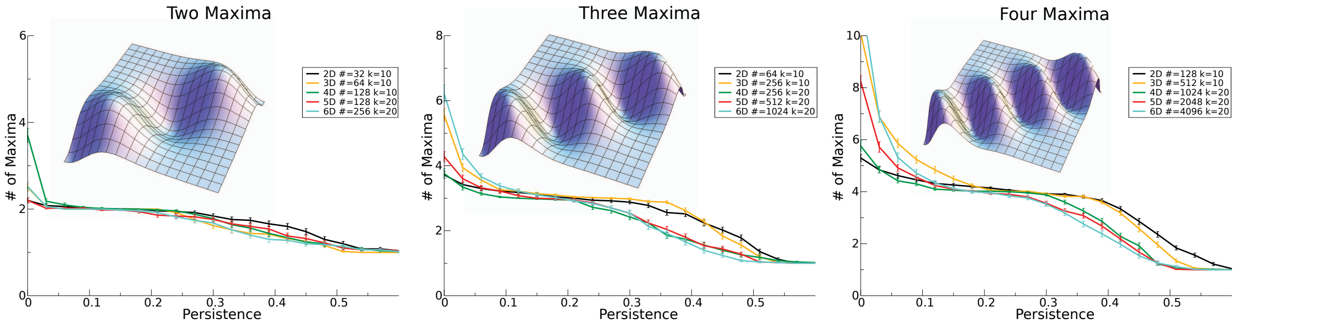


Fig. 7. Average persistence graphs showing the number of maxima vs. persistence for the sensitivity study for different number of features and different dimensions. Each curves shows the mean and standard error of 50 random instantiations. Among our experiments we manually picked the curve with lowest number of samples that showed a distinct plateau at the correct count and thus could differentiate all (positive) features.

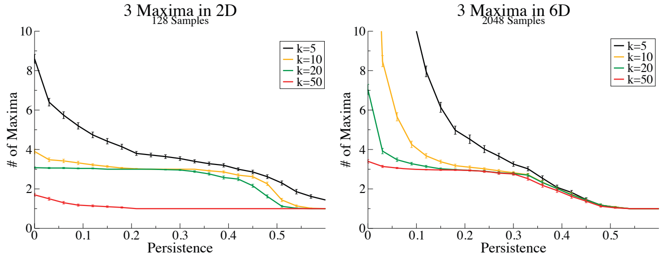


Fig. 8. Persistence curves for the three maxima test function in two and six dimensions with 128 and 2048 samples respectively for different neighborhood sizes. A larger neighborhood stabilizes the curves but in the extreme can smooth away features.

In many cases the intrinsic dimension of the data does not correspond to the dimension of the observations. For example, in physical simulations the observed data points are constrained by the physical model and thus the data has an intrinsic dimension corresponding to the degrees of freedom of the physical model. In machine learning, manifold estimation approaches are built based on this observation. The techniques employed are often based on building a nearest-neighbor graph [50, 47, 5] that approximates the manifold structure in the data. In cases with manifold structured data, i.e. a scalar function defined on a lower dimensional manifold in the data space, the approximate MS-complex implicitly takes the manifold structure into account. However, as discussed in Section 3, it might be beneficial for such data sets to use a dimension reduction approach as a preprocessing step.

4 Visualization

The visualization consists of two parts, the MS-complex window (MSW) and an application specific domain inspection window (DIW) relaying additional information about the input domain.

The MSW is primarily used to gain an overview over the topological and geometric structure of the scalar field and to drive interactive exploration of the input domain. Fig. 4 shows the components of the window; each tube corresponds to a Morse-Smale crystal. The colored tube is the embedding of the regression curve with color corresponding to the predictor value (of the scalar function). The geometry of the curve is an approximation to the actual geometry in the high dimensional input domain. The width of the outer transparent tube corresponds to the standard deviation of the regression curve (proxy for width of the crystal) and can be turned on and off. Information about the sampling density is visualized with coloring of the silhouette of the transparent tube, dark color indicates high density and light color low density areas. On the bottom right the persistence graph is shown and the user can select the visualization of a particular level of detail, as described in Section 3.1. For complicated topologies, the user can select a particular crystal of interest and the rest of the MS-complex is faded out around the crystal of interest, as illustrated in Fig. 9. The user can inspect a crystal in detail by moving the slider (ring) along the tube, more detailed geometric information is then interactively displayed in the DIW.

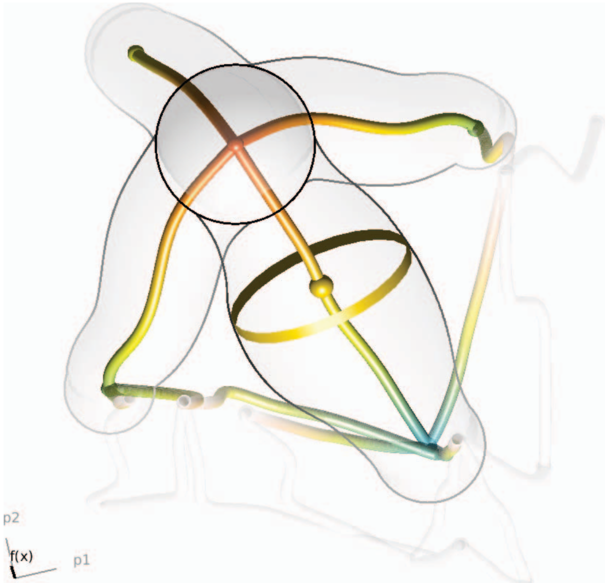


Fig. 9. Dealing with visual clutter by focusing on Morse-Smale crystals of interest. Here four selected tubes are rendered solid, the remaining tubes fade out with increasing distance from the selected tubes to keep a sense of the global layout.

Fig. 10 shows two domain inspection modes that present information about the local geometry of the selected location such as mean, standard deviation, and gradient of the regression curve. In its most general form, the DIW presents this information in a box plot type visualization for each input coordinate at the selected slider location. Alternatively, the regression curve of a crystal can be coordinate-wise graphed as a function of the output value. This provides a more global representation over the geometry of the crystal. For input domains

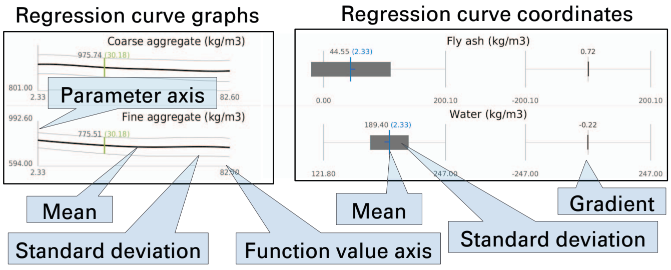


Fig. 10. Domain inspection windows, (left) coordinate wise graphs of the regression curves and (right) local information at slider location.

with additional structure, this information can be visualized on the specific form of the input domain, for example images for grids as illustrated in Fig. 14 or heat maps for gene expressions.

5 Demonstration of the Methodology

To demonstrate our approach we tested the proposed visualization on datasets for which the main trends are known. This process is valuable to validate the approach as well as to develop compelling examples that illustrate its effectiveness in practice. The two datasets [1] that we use for this purpose are: (i) analysis of manufacturing parameters of concrete production and the quality of the product and (ii) on demographic information of US counties in relation to crime rate.

5.1 UCI Concrete Compressive Strength

The concrete compressive strength data set examines the effect of different cement mixtures on the compressive strength of the resulting concrete. The dataset consists of 1030 samples of different concrete cores tested for compressive strength. The input variables are cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, and age. The output is the compressive strength in MPa. Yeh [59] describes the dataset in detail and performs an analysis with a linear regression model and neural networks. The regression model $f_c(age) = a\left[\frac{water}{cement}\right]^b c[ln(age)+d]$ with a, b, c and d regression coefficients, is tailored to the effects of the water/cement ratio and age and does not take into account any other variables. Both the regression model and the neural network are used as a predictive model and do not lend themselves readily to insight into the structure of the relation between mixtures and concrete strength. Fig. 11 shows a typical result using a regression analysis that confirms that a low water/cement ratio results in a stronger concrete. Fig. 12 shows the proposed approach at the coarsest level of detail (highest persistence) with a single maximum and minimum. The water/cement ratio relationship observed in the conventional regress is reflected in the curves for water and cement with increasing compressive strength. Additionally, an inverse relationship between fly ash and compressive strength is readily visible. At a finer level of detail, multiple interesting interactions occur, as shown in Fig. 13 on three selected crystals. One can build several hypotheses for testing with specific regression models. First, it is immediately visible that relatively different mixtures can lead to similar strength concrete mixtures. The minima differ in their settings of fly ash, blast furnace and coarse/fine aggregation ratio — this suggests that fly ash and blast furnace can lead to weak mixtures if the coarse/fine aggregation ratio and cement amount is not properly adjusted. The relationship between coarse and fine aggregates is visible in (c), but this relationship also depends on the amounts of fly ash and blast furnace. While (b) and (c) show that a relatively large amount of coarse aggregates with little blast furnace results in a strong mixture, in (a) the inverse is also visible. A similar inverse relationship is indicated by cement and fly ash—increasing the amount in fly ash requires the reduction of the cement amount to obtain strong concrete.

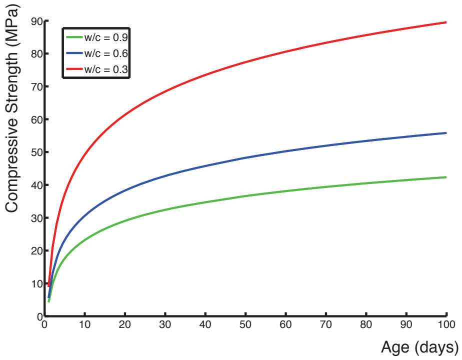
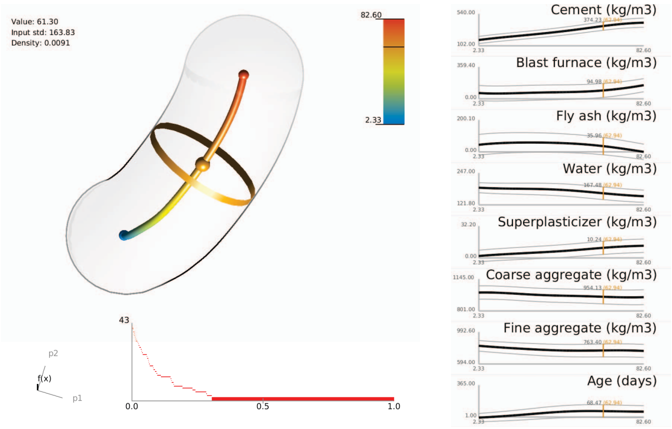


Fig. 11. Concrete compressive strength modeled with a regression model on age and water/cement ratio.

5.2 Brain Magnetic Resonance Images

This example illustrates how the proposed method can be adapted to an application with more structured input domains. This is a data set of 416 brain volumes from the OASIS brain database [2] with mini mental state examination scores (MMSE), the scalar output. Each brain represents a data point in R^d with d the number of voxels in the volume, which is $176 \times 208 \times 176$. Since only 416 data points are available, the computations are performed in the 415 dimensional subspace spanned by the data. The MMSE score is a scalar ranging from 0 to 30 measuring cognitive abilities with a score of above 25 being normal and scores below indicating mild (21-24), moderately (10-20) and severe (below 9) cognitive impairment. Fig. 14 shows the location on the regression curve as an image corresponding in the input domain, standard deviation is indicated with increasing red shade on the pixels. Above the gradient is shown (green decreasing intensity and red increasing intensity). The coarsest level of detail shows that decreasing MMSE scores correlate with increasing ventricle size. It is well known that the size of the ventricle increases with age or disease, such Alzheimer's, which in turn affects the MMSE scores significantly [21, 3]. At finer levels, multiple minima appear, related to atrophies in the left versus right ventricle horn. However, these minima are potentially artificial due the small number of subjects with low MMSE scores, as indicated by the sampling density in the visualization (light silhouette).



(a) (b)

Fig. 12. (a) Visualization of UCI concrete compressive strength data set at coarsest level of detail. The domain inspection window (b) and the corresponding composition as a function of strength (a). Note the inverse relationship between water and cement as well as increasing strength with age. Additionally, we can see that fly ash tends to zero for strong mixtures.

6 Applications

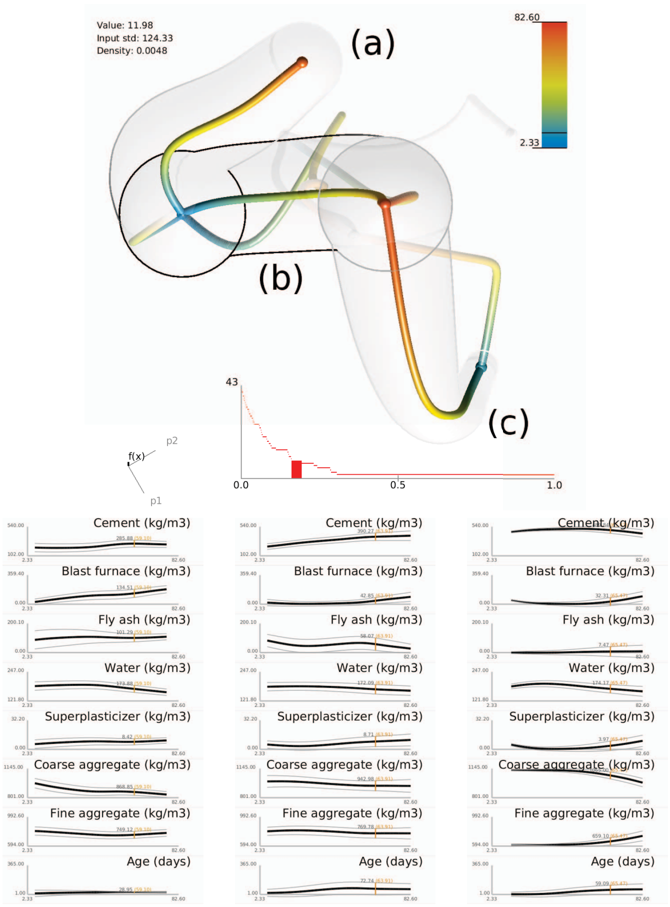
This section presents results from explorations of scientific data in collaboration with domain scientists. The first case is focused on the analysis of the relationship between the composition of chemical species in a turbulent combustion process and its efficiency in terms of fuel consumption and pollutants generated. The data is from a time dependent simulation of jet flames. The second application involves the analysis of the parameter space of a climate modeling simulation and the ability to estimate the uncertainty associated with a given prediction. In both cases the scientists have found that the use of our tools revealed new insights that existing techniques did not provide.

6.1 Combustion Simulation

This data consists of 700K samples of chemical composition and temperature extracted pointwise (samples in space and time) from temporal jet simulations of turbulent CO/H2-air flames, as described by Hawkes et. al [32], with detailed chemistry, thermodynamics, and transport. The data includes extinction and reignition phenomena. Several chemical components form and evolve during the combustion reaction and in turn effect the amount of heat released. In this analysis we explore the temperature in relation to the chemical composition. The chemical species involved are O2 (Oxygen gas / Oxidizer),

[1] From the UCI machine learning repository http://archive.ics.uci.edu/ml

[2] http:www.oasisbrains.org



(a) (b) (c)

Fig. 13. Finer level of detail of the compressive strength data set. Three crystals are selected for closer examination.

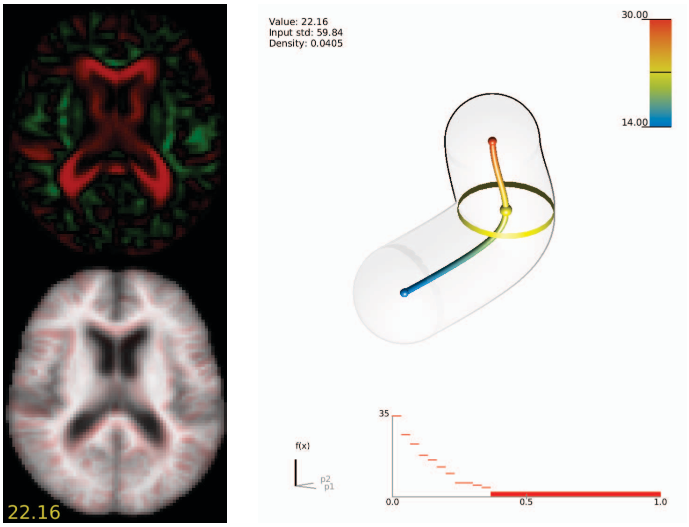
O (Oxygen), OH (Hydroxide), H2O (Water), H (Hydrogen), HO2, CO (Carbon monoxide), CO2 (Carbon dioxide) and HCO—a 10 dimensional scalar function with amount of heat released as output.

The data is incommensurate among the different chemical species. Thus, we normalized the data such that each coordinate has a standard deviation of one. This is important for the nearest-neighbor computation and places equal weight on the different coordinates for distance computations. To reduce computation time, we randomly, uniformly without replacement, sampled 10K points of the 700K data set. The subsample shows a nearly identical mean, standard deviation, and histogram as the complete data set.

Fig. 15 shows, that at a high persistence, three distinct minima with low standard deviation and a single maximum with a larger standard deviation emerge. This confirmed the expectation of the domain experts of four distinct modes of combustion in one intuitive and easily accessible illustration. In particular, as can be seen from the regression curves, the maximum in temperature corresponds to the main combustion mode where fuel and oxidizer are present in stoichiometric proportions. In this case, fuel and oxygen react to mutually annihilate each other and form products—corresponding to the large standard deviation of the peak—-releasing heat in the process. The three distinct minima correspond to:

(a) extinction, where the mixing of fuel and oxidizer is highly turbulent and blows the flame out, resulting in a large amount of HO2;
(b) pure fuel (H2 and CO), where no chemical reaction occurs due to the lack of oxidizer; and
(c) pure oxidizer (O2), again with no chemical reaction due to the lack of fuel.

For comparison, Fig. 16(a) shows a projection of all points onto the first two principal components. Two of the minima are readily visible, however the third minima is hidden in this projection.



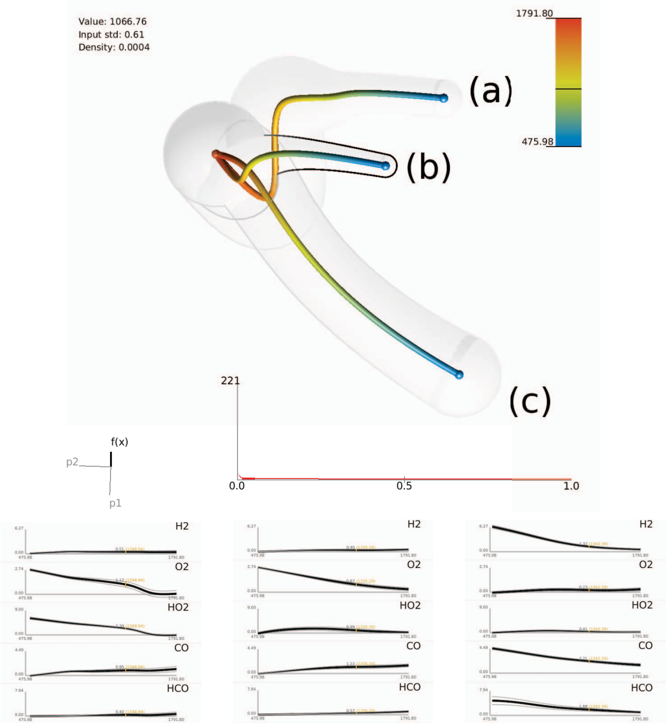
(a) (b)

Fig. 14. The proposed approach illustrated on a structured input domain. Mini mental state examination (MMSE) scores as a function on magnetic brain resonance images. The regression curve (b) and an axial slice of the constructed volume (a, bottom) from the regression curve at MMSE score 23.14 and the corresponding gradient (a, top) with red a positive (increasing voxel intensity) and green negative (decreasing voxel intensity) direction. The gradient indicates that with decreasing MMSE scores the ventricle size increases.

6.2 Climate Simulations

With the rising sophistication and accuracy of current climate simulations their predictive capabilities are increasingly called upon to inform national and international policy. In this context, the ability to estimate the likelihood of a given prediction and a quantification of potential uncertainties is crucial. To answer this need a concerted effort has been made to better understand the uncertainties involved in climate simulations [57, 37, 51]. One of the most common techniques to analyze uncertainties in a climate simulation is to create an ensemble of simulations for varying input parameters. In this case our data set consists of 593 runs of a recent version of the Community Atmosphere Model [3], a global atmosphere model developed at the National Center for Atmospheric Research (NCAR). Within the ensemble, 21 input parameters describing various aspects of the atmospheric physics are varied within ranges determined by experts. The resulting 21-dimensional domain is scaled into the unit box. For each instantiation a large number of local and global variables are recorded, such as, the global energy output or average temperature. As an example, we use the total upward long wave flux, a measure of how much long wave (i.e. thermal) radiation is leaving the planet. Analyzing, the long wave flux as a function of the inputs reveals a single global minimum and two strong local maxima, see Fig. 17. The two resulting crystals show two markedly different regimes leading to an equally high long wave flux. In particular, one maximum shows a minimal value for *tau* and a maximal value of *cmftau* while the other shows the opposite behavior. Interestingly, both parameters are related to convection, the thermal driven upwelling of warm, moist air. While *tau* defines the time scale (and thus the rate of energy conversion) for *deep convection* happening above 500 hPa; *cmftau* does so for *shallow convection* happening below 500 hPa. A possible explanation for the observed strong inverse relation of these two convection schemes is the need for both schemes to work in tandem to create clouds which prevent long wave energy from escaping. An imbalance in the two convection schemes may ultimately result in fewer clouds and thus a higher long wave flux.

[3] http://www.ccsm.ucar.edu/models/atm-cam



(a) (b) (c)

Fig. 15. Chemical composition in relation to heat released during a jet flame combustion simulation. The three distinct minima correspond to pure fuel, pure oxidizer and extinction/reignition. Graphs of chemical composition plotted against temperature for the crystals corresponding to extinction (a), pure oxidizer (b) and pure fuel (c) minima compositions.

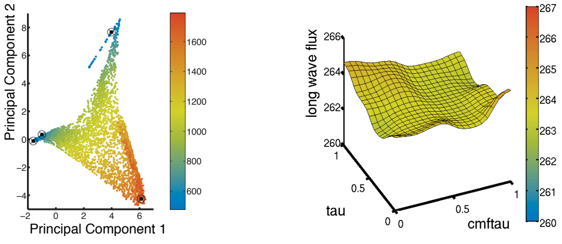
Our analysis, for the first time, demonstrated a significant influence of *tau* and *cmftau* on the longwave flux something not apparent in standard statistical approaches. For example, direct linear regression on *cmftau* and *tau* in relation to flux results in p-values of 0.005 and 0.6822, respectively. A multiple linear regression model with feature selection based on the Bayesian information criterion (BIC) [48] rejects both *cmftau* and *tau* from the best-scoring models. Even a kernel regression surface of long wave flux on *cmftau* and *tau* (shown in Fig. 16(b) for a particular kernel bandwidth (the observation holds regardless of bandwidth) fails to reveal their combined strong relation to thermal radiation.

7 Conclusion

The applications demonstrate that the approach presented here is capable of detecting complicated interactions in high dimensional data sets. Validation of the visualization by domain experts confirms this claim and provides strong evidence of the usefulness of the technique. This represents an important first step to build further applications based on the proposed framework.

For a future study, linear instead of a nonparametric regression could be used to support the automatic detection of statistical significant correlations. In very high dimensional settings, such quantitative evaluations would provide a tool to guide users towards interesting regions and parameters. The climate model simulations take significant amount of resources to run and an important open challenge is to provide guidance for interesting future parameter settings based on previous runs.

For the dimension reduction of the simplified Morse-Smale complex, several alternate approaches are possible. Extension to manifold structured domains is an interesting future direction. However, several points need to be addressed. The inverse regression could be based on manifold coordinates or the original data space changing the interpretation of the regression curves. If the manifold learning is used as a preprocessing step the effect on the Morse-Smale complex needs to be investigated. The dimension reduction could introduce spurious extrema, on the other hand the effects of noisy observations can be reduced.



(a) (b)

Fig. 16. (a) A projection of the combustion data onto the first two principal components, colored by temperature. The ⊗ show the location of the extremal points computed from the approximate Morse-Smale complex on the high dimensional data. The projection reveals only two of the three persistent minima. (b) A kernel regression of the thermal radiation in the climate data on *cmftau* and *tau* does not show a strong influence on thermal radiation, even with a small kernel bandwidth of 0.15.

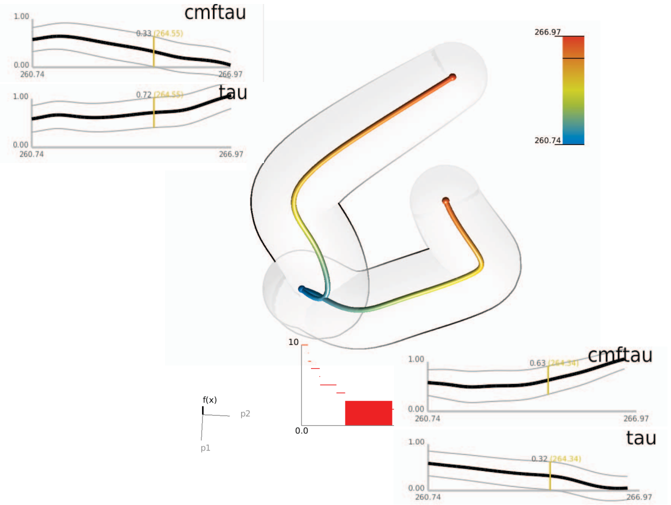


Fig. 17. Climate simulation parameter analysis in relation to upward long wave flux. A very strong inverse relationship between *tau* and *cmftau* is visible, both related to cloud formation. An imbalance in those parameters may prevent cloud formation which would lead to a large flux.

Acknowledgments

Thanks to Chems Touati for video production and Prof. James C. Sutherland for combustion simulation expertise and data.

This work was funded by the National Institute of Health grants U54-EB005149 and 2-P41-RR12553-08, and NSF grant CCF-073222.

This was work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-CONF-446831. This work was supported by the Director, Office of Advanced Scientific Computing Research, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 through the Scientific Discovery through Advanced Computing (SciDAC) program's Visualization and Analytics Center for Enabling Technologies (VACET).

References

[1] D. F. Andrews. Plots of high-dimensional data. *Biometrics*, 28(1):125–136, 1972.

[2] D. Asimov. The grand tour: a tool for viewing multidimensional data. *SIAM J. Sci. Stat. Comput.*, 6(1):128–143, 1985.

[3] S. A. Barron, L. Jacobs, and W. R. Kinkel. Changes in size of normal lateral ventricles during aging determined by computerized tomography. *Neurology*, 26(11):1011–, 1976.

[4] L. E. Baum and P. Billingsley. Asymptotic distributions for the coupon collector's problem. *Ann. Math. Stat.*, 36:1835–1839, 1965.

[5] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, 2003.

[6] S. Beucher. Watersheds of functions and picture segmentation. In *Proc. IEEE ICASSP*, pages 1928–1931, 1982.

[7] R. L. Boyell and H. Ruston. Hybrid techniques for real-time radar simulation. In *Proc. 1963 Fall Joint Comp. Conf.*, pages 445–458, 1963.

[8] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

[9] P.-T. Bremer, H. Edelsbrunner, B. Hamann, and V. Pascucci. A topological hierarchy for functions on triangulated surfaces. *IEEE Trans. on Vis. and Comp. Graphics*, 10(4):385–396, 2004.

[10] P.-T. Bremer, G. Weber, V. Pascucci, M. Day, and J. Bell. Analyzing and tracking burning structures in lean premixed hydrogen flames. *IEEE Trans. on Vis. and Comp. Graphics*, 16(2):248–260, 2010.

[11] H. Carr, J. Snoeyink, and U. Axen. Computing contour trees in all dimensions. *Comput. Geom. Theory Appl.*, 24(3):75–94, 2003.

[12] H. Carr, J. Snoeyink, and M. van de Panne. Simplifying flexible isosurfaces using local geometric measures. In *IEEE Visualization '04*, pages 497–504. IEEE Computer Society, 2004.

[13] P. Chaudhuri, M. ching Huang, W. yin Loh, and R. Yao. Piecewise-polynomial regression trees. *Statistica Sinica*, 4:143–167, 1994.

[14] F. Chazal, L. Guibas, S. Oudot, and P. Skraba. Analysis of scalar fields over point cloud data. In *SODA '09: Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1021–1030, 2009.

[15] F. Chazal, L. Guibas, S. Oudot, and P. Skraba. Persistence-based clustering in riemannian manifolds. Technical Report RR-6968, INRIA, 2009.

[16] Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(8):790–799, 1995.

[17] W. S. Cleveland. LOWESS: A Program for Smoothing Scatterplots by Robust Locally Weighted Regression. *Amer. Statistician*, 35:54, 1981.

[18] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. *Discrete Comput. Geom.*, 37(1):103–120, 2007.

[19] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE TPAMI*, 24:603–619, 2002.

[20] M. M. W. Conover and R. Beckman. comparison of three methods for selection values of input variables in the analysis of output from a computer code. *Technometrics*, 22(2):239–245, 1978.

[21] M. de Leon, A. George, B. Reisberg, S. Ferris, A. Kluger, L. Stylopoulos, J. Miller, M. La Regina, C. Chen, and J. Cohen. Alzheimer's disease: longitudinal CT studies of ventricular change. *Am. J. Roentgenol.*, 152(6):1257–1262, 1989.

[22] H. Digabel and C. Lantuejoul. Iterative algorithms. In *Proc. Symp. Quantitative Analysis of Microstructures in Material Science, Biology and Medicine*, pages 85–99, 1978.

[23] H. Edelsbrunner, J. Harer, and A. Zomorodian. Hierarchical Morse-Smale complexes for piecewise linear 2-manifolds. *Discrete Comput. Geom.*, 30:87–107, 2003.

[24] J. Friedman and J. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput*, C-23(9):881 – 890, 9 1974.

[25] J. H. Friedman. Multivariate adaptive regression splines. *Ann. Statist.*, 19(1):1–141, 1991. With discussion and a rejoinder by the author.

[26] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic Press, 2nd edition, 1990.

[27] K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Information Theory*, 21:32–40, 1975.

[28] A. Gyulassy, M. Duchaineau, V. Natarajan, V. Pascucci, E.Bringa, A. Higginbotham, and B. Hamann. Topologically clean distance fields. *IEEE TVCG*, 13(6):1432–1439, 2007.

[29] A. Gyulassy, V. Natarajan, V. Pascucci, P.-T. Bremer, and B. Hamann. Topology-based simplification for feature extraction from 3D scalar fields. *IEEE TVCG*, 12(4):474–484, 2006.

[30] A. Gyulassy, V. Natarajan, V. Pascucci, and B. Hamann. Efficient computation of Morse-Smale complexes for three-dimensional scalar functions. *IEEE TVCG*, 13(6):1440–1447, 2007.

[31] W. Harvey and Y. Wang. Generating and exploring a collection of topological landscapes for visualization of scalar-valued functions. In *Proc. Symposium on Visualization*, volume 29, page to appear, 2010.

[32] E. R. Hawkes, R. Sankaran, J. C. Sutherland, and J. H. Chen. Scalar mixing in direct numerical simulations of temporally evolving plane jet flames with skeletal co/h2 kinetics. *Proceedings of the Combustion Institute*, 31(1):1633 – 1640, 2007.

[33] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504 – 507, July 2006.

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

[35] T. Itoh and K. Koyamada. Automatic isosurface propagation using an extrema graph and sorted boundary cell lists. *IEEE Trans. Vis. and Comp. Graph.*, 1(4):319–327, 1995.

[36] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.

[37] J. Kiehl, C. Shields, J. Hack, and W. Collins. The climate sensitivity of the community climate system model version 3 (ccsm3). *Climate*, 19(11):2584–2596, 2006.

[38] D. Laney, P.-T. Bremer, A. Mascarenhas, P. Miller, and V. Pascucci. Understanding the structure of the turbulent mixing layer in hydrodynamic instabilities. *IEEE TVCG*, 12(5):1052–1060, 2006.

[39] J. X. Li. Visualization of high-dimensional data with relational perspective map. *Information Visualization*, 3(1):49–59, 2004.

[40] F. Maisonneuve. Sur le partage des eaux. Technical report, School of Mines, Paris, France, 1982.

[41] J. Milnor. *Morse Theory*. Princeton University Press, New Jersey, 1963.

[42] M. Morse. Relations between the critical points of a real functions of n independent variables. *Transactions of the American Mathematical Society*, 27:345–396, July 1925.

[43] E. A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 9(1):141–142, 1964.

[44] P. Oesterling, C. Heine, H. Jänicke, and G. Scheuermann. Visual analysis of high dimensional point clouds using topological landscape. In *Proc. IEEE Pacific Visualization*, page to appear, 2010.

[45] V. Pascucci, G. Scorzelli, P.-T. Bremer, and A. Mascarenhas. Robust on-line computation of reeb graphs: simplicity and speed. *ACM Trans. Graph.*, 26(3):58, 2007.

[46] G. Reeb. Sur les points singuliers d'une forme de pfaff completement intergrable ou d'une fonction numerique [on the singular points of a complete integral pfaff form or of a numerical function]. *Comptes Rendus Acad.Science Paris*, 222:847–849, 1946.

[47] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(550), 2000.

[48] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978.

[49] Y. Sheikh, E. Kahn, and T. Kanade. Mode-seeking by medoidshifts. In *Proc. IEEE International Conference on Computer Vision*, 2006.

[50] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(550):2319–2323, 2000.

[51] L. Tomassini, P. Reichert, R. Knutti, T. Sticker, and M. Borsuk. Robust bayesian uncertainty analysis of climate system properties using markov chain monte carlo methods. *Climate*, 20(7):1239–1254, 2006.

[52] A. Vedaldi and S. Soatto. Quick shift and kernel methods for mode seeking. In *Proc. European Conf. on Computer Vision*, pages 705–718, 2008.

[53] L. Vincent and P. Soille. Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(6):583–598, 1991.

[54] J. A. Walter and H. Ritter. On interactive visualization of high-dimensional data using the hyperbolic plane. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 123–132, 2002.

[55] G. Watson. Smooth regression analysis. *Sankhya, Series*, A(26):359–372, 1964.

[56] G. Weber, P.-T. Bremer, and V. Pascucci. Topological landscapes: A terrain metaphor for scientific data. *IEEE Transactions on Visualization and Computer Graphics*, 13:1416–1423, 2007.

[57] M. Webster, C. Forest, J. Reilly, M. Babiker, D. Kicklighter, M. Mayer, R. Prinn, M. Sarofim, A. Sokolov, P. Stone, and C. Wang. Uncertainty analysis of climate change and policy response. *Climate Change*, 61(3):295–320, 2003.

[58] C.-C. Yang, C.-C. Chiang, Y.-P. Hung, and G. C. Lee. Visualization for high-dimensional data: Vishd. In *IV*, pages 692–696, Washington, DC, USA, 2005. IEEE.

[59] I. C. Yeh. Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research*, 28(12):1797 – 1808, 1998.

[60] X. Zhu, R. Sarkar, and J. Gao. Shape segmentation and applications in sensor networks. In *Proc. INFOCOM*, pages 1838–1846, 2007.