ColorMapND: A Data-Driven Approach and Tool for Mapping Multivariate Data to Color

Shenghui Cheng, Wei Xu, *Member, IEEE*, and Klaus Mueller, *Senior Member, IEEE*

Abstract—A wide variety of color schemes have been devised for mapping scalar data to color. We address the challenge of color-mapping *multivariate* data. While a number of methods can map low-dimensional data to color, for example, using bilinear or barycentric interpolation for two or three variables, these methods do not scale to higher data dimensions. Likewise, schemes that take a more artistic approach through color mixing and the like also face limits when it comes to the number of variables they can encode. Our approach does not have these limitations. It is data driven in that it determines a proper and consistent color map from first embedding the data samples into a circular interactive multivariate color mapping display (ICD) and then fusing this display with a convex (CIE HCL) color space. The variables (data attributes) are arranged in terms of their similarity and mapped to the ICD's boundary to control the embedding. Using this layout, the color of a multivariate data sample is then obtained via modified generalized barycentric coordinate interpolation of the map. The system we devised has facilities for contrast and feature enhancement, supports both regular and irregular grids, can deal with multi-field as well as multispectral data, and can produce heat maps, choropleth maps, and diagrams such as scatterplots.

Index Terms—Multivariate data, color mapping, color space, high dimensional data, pseudo coloring

1 Introduction

Mapping data to color has a rich history and several well-tested color schemes have emerged (e.g., [1], [6], [35]). Most of these, however, are defined for *scalar data* where a scalar value indexes a one-dimensional table that returns an RGB color triple. Other schemes assign colors to different, usually disjoint materials and then use standard blending functions to handle areas where materials overlap or mix together. The latter often occurs in the graphical rendering of simulations or imaged data, while the former is frequently encountered in pseudo-coloring for heat maps or choropleth maps.

In this paper, we are interested in colorizing *multivariate data*. Here we mainly focus on numerical data (categorical data can be converted into numerical data [34]). These types of multivariate data occur frequently in many applications, such as demographic assessments, environmental monitoring, scientific simulations, imaging, business [10] and others. The domain can be a geographic map, an image, or a volume. They are a subset of multi-field data which also include multi-channel, multi-attribute, multi-modal, and multi-material data, among others. Visualizing these types of data in their native domain remains challenging, and there is so far little support to map these data vectors directly into color.

• *The authors are with the Visual Analytics and Imaging Lab at the Computer Science Department, Stony Brook University, Stony Brook, NY 11794 and with the Computational Science Initiative, Brookhaven National Lab, Upton, NY 11973-5000. E-mail: {shecheng, wxu, mueller}@cs.stonybrook.edu.*

Manuscript received 23 Jan. 2017; revised 11 Feb. 2018; accepted 19 Feb. 2018. Date of publication 0 . 0000; date of current version 0 . 0000.
(Corresponding author: Shenghui Cheng.)
Recommended for acceptance by I. Hotz.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2018.2808489

A common practice is to visualize multivariate data as multiple images where each channel is mapped to a separate plot with a simple color scale. Fig. 1d shows such an arrangement for four scalar images. However, a disjoint display of this nature makes it difficult to recognize correlations (or a lack thereof) that may exist among the different channels (variables) in the image.

For this reason, we wish to fuse the individual images into a single multi-color image. Correlations can then be easily perceived by similarity of color, while dissimilarities become apparent by color variations. At the same time we can use the color as a label to reveal which of the factors dominate or co-exist in certain areas. Essentially, we retain color as a visual representation of the relative strength of a given variable for each pixel in the image.

One way to achieve this fusion is by interpolation or blending. Let us assume we have $n \leq 3$ variables. Then each variable is assigned to one of n primary colors, and a mapped color is produced via bilinear (for $n = 2$ variables) or barycentric (for $n = 3$ variables) interpolation [36]. Alternatively, we can assign each variable to one of a monitor's three (RGB) primaries and blend the three variables directly in hardware into an RGB image.

One drawback of this concept is that it is difficult to extend to $n > 3$. Hardware blending is infeasible since monitors typically only have three primary colors. Conversely, interpolation could be realized using advanced schemes like generalized barycentric interpolation [25]. A severe drawback of interpolation and blending is that they do not yield a perceptually uniform result. Both map the data into an RGB color cube which is not a perceptual color space. It gives rise to the *rainbow color map* which renders some value differentials invisible while overly emphasizing others [4], [31]. This is not the case for the established 1D

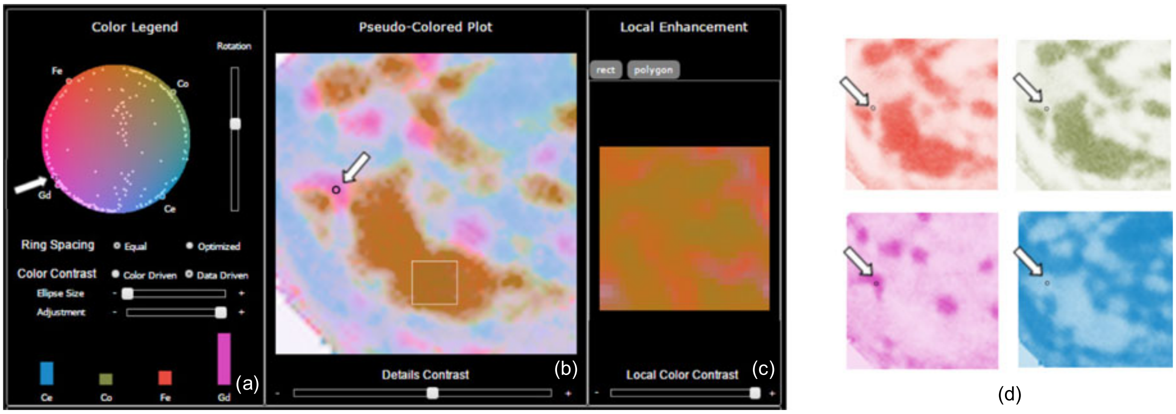

Fig. 1. System interface with all major displays and components (using the battery data, see Section 6.2 for more detail). Users can select a multivariate data point in any of these displays via mouse click. The system responds by highlighting the selected data point with a small circle both in the targeted display as well as in the other, synched displays (see arrows, added for illustration). (a) Integrated CIE HCL (Hue Chroma Luminance) interactive multivariate color mapping display (ICD, top) with control panel (middle), and the selected point's multivariate spectrum display (bottom). (b) Multi-field / hyperspectral image, pseudo-colored via the multivariate color map in (a). (c) Locally enhanced colorization of the selected rectangular region in (b). (d) Individual scalar images (usually displayed on the bottom of the interface in a *channel view* partition) colorized via the attribute-linked color primaries marked and labeled at the circle boundary of the multivariate color map in (a). The image in (b) constitutes a joint colorization of these individual channel images.

color maps which are the result of psycho-physical experiments and are perceptually uniform.

The system we have devised combines a multivariate data embedding scheme [7] inspired by generalized barycentric interpolation with a perceptually uniform colorspace, CIE HCL. The teaser image of Fig. 1 gives an overview of our approach by ways of an example. Fig. 1d shows the four channel images we wish to fuse. Stacked up, each image pixel is a 4D data point. We embed the data points into what we call *circular interactive multivariate color mapping display (ICD)*, shown in Fig. 1a. The attributes are arranged on the ICD's boundary in terms of their similarity. Using the ICD, the color of a multivariate data sample is then obtained via generalized barycentric coordinate interpolation. The generated image (see Fig. 1b) clearly shows at what locations pixels correlate and what the dominant factors are.

Our paper is structured as follows. Section 2 presents related work. Section 3 gives an overview of our tool and framework. Section 4 presents its basic features, while Section 5 describes additional functionalities we developed in response to requirements we discovered during practical use. Section 6 showcases several case studies. Section 7 presents a user study and feedback. Section 8 ends with conclusions and future work.

2 Related Work

A color map is also frequently referred to as *color palette* or *color scheme*. Color palettes are most often designed for univariate data, and they are almost always due to some path in a given color space. A very simple method to generate a color palette is to linearly interpolate between $\mathrm{RGB} = (0, 0, 255)$ and $\mathrm{RGB} = (255, 0, 0)$, which is equivalent to varying the hues in HSV color space from red to purple. This gives rise to the infamous *rainbow colormap*. While straightforward to implement, the rainbow colormap is less than ideal since it is not isoluminant. This means that it has sub-ranges that have little perceivable contrast and consequently any scalar detail mapping into these sub-ranges is difficult to distinguish [4], [28].

There has been significant work on designing more effective standard color maps for scientific data visualization. Well known here is the IBM PRAVDA system [2]. In addition, a prominent guide is also the Color Brewer [6] which presents a variety of color schemes for cartography applications, broken down into sequential, diverging, and qualitative schemes. For the former two schemes the site suggests decompositions into up to 9 elements. More could be obtained via interpolation, either piecewise linear to preserve the original elements or via higher-order functions. The Brewer schemes are highly respected and widely applied. According to the authors [16] they were designed "using both experience and trial and error". Later, in more analytical research Wijffelaars et al. [35] show that the Brewer palettes generally follow curved paths in the hue slices of the CIE LUV color space, but that the elements are not iso-distant from one another. The authors then describe an analytical tool by which lightness-ordered palettes of any hue can be created and which follow optimally lightness-sampled paths.

Choosing colors in CIE LUV color space is preferable since it is perceptually uniform. Perceptual uniformity means that any two equidistant colors elicit the same perceived color contrast in a human observer. These perceptually well-defined distance relationships enable a convenient mapping of geometric operations into color space. We take advantage of these relationships in our work. Once the mapping is done we convert to RGB for display.

The perceptual uniformity of CIE LUV space has also proven to be effective for the rendering of photographic (RGB) volume datasets. It allowed for meaningful opacity mappings as well as gradient calculations [12]. Finally, more recent research on color palettes includes that of Fang et al. [13] who presented a method for maximizing the perceptual distances among a set of colors assigned by users for categorical data. Gramazio et al. [14], on the other hand, described work that sought to optimize color palettes for user-defined discriminability and preferences.

2.1 Bivariate and Trivariate Color Palettes

We are specifically interested in color schemes that can support multivariate data. Stevens presents an online how-to guide [36] for constructing a 3×3 *bivariate* color palette from

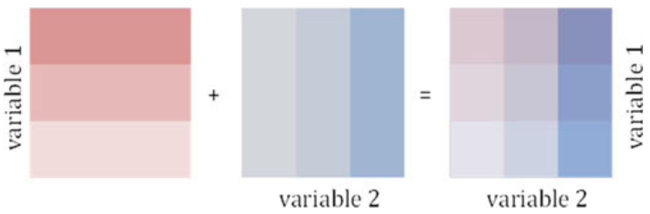

Fig. 2. Constructing a bivariate color palette from two univariate color palettes (see Stevens [36]).

two three-element 1D color palettes (see Fig. 2). It constructs a 2D palette cell by blending two 1D palette cells together. Stevens writes that this requires some manual tweaking in hue and saturation to make the mixed cells along the diagonal more distinguishable. In fact, this manual tweaking of cell colors is not unlike the more principled and algorithmic techniques that have been published in the visualization community to address the problems arising from the blending of colors in two or more semi-transparent layers [1], [9], [33]. One of these problems is the appearance of false (third) colors that can be generated when blending two colors together. Given these problems, it is unclear how Stevens' scheme would extend to color palettes of an order greater than two. It is also not a proven perceptually uniform scheme.

Another way to construct *bivariate* color palettes is via *interpolation* or *blending*. We have already discussed this approach and its shortcomings in the introduction.

2.2 Color Mapping for Multivariate Data

The colorization of data of more than three variables has received less attention so far. Work in this area includes that of Hagh-Shenas et al. [15]. They compare two techniques for the visualization of 6-dimensional data on choropleth maps: (1) blending using six separate color ramps and (2) texturing with spectral noise. Their user studies reveal that while the error rate for blending significantly rises already for three variables, the increase in the error rate for texturing is only statistically significant for the case of six joint variables (five was not tested). Our approach also performs blending but users can visually map a color back into the ICD (see Fig. 1a) to gain insight about the multivariate proportions (using intensity to determine the overall strength). Conversely, in the system by Hagh-Shenas et al. users need to mentally decode the blended value into its k constituents via the k disjoint color ramps which is arguably difficult. Their more promising noise textures, on the other hand, have limited use in our case since they cannot be used in a continuous domain without severe loss in resolution.

Others have looked at the problem from the perspective of dimension reduction. These methods have been mainly described in the context of mapping hyperspectral image data into RGB space. Ready and Witz [28] perform Principal Component Analysis (PCA) [19] and map the top three PCA vectors into color space. However, while this preserves as much of the data variance as possible, it offers little control about the colors assigned and their relations to the variables.

On the other hand, Lawrence et al. [23] use Multidimensional Scaling (MDS) [22] for dimension reduction and enforce constraints on the colors used in different areas of the image by adding a value constraint into the MDS stress equation. This requires a suitably colored input image to specify this value constraint. As such this algorithm is more of a framework for painting colored images from multispectral image data since the constraints are given in the image domain and not in the attribute domain. And so, imposing color constraints on the data attributes themselves is not easily done. In that respect, there is no color legend and no concrete color map.

2.3 Multivariate Data Visualization

Our ICD (see Fig. 1a) embeds multivariate data into a 2D display. We use a technique that is essentially an optimized version of RadViz [17], which we presented in [7]. There, we also showed that the equations of RadViz are equivalent to those of Generalized Barycentric Coordinate interpolation [7], [25] when formulated as a mapping problem and substituting the convex polygon by a ring. There are also other embedding techniques, such as ISOMAP [32], t-distributed stochastic neighbor embedding (t-SNE) [24], multidimensional scaling (MDS) [22], locally linear embedding (LLE) [29] and others but all of these only map the data samples but cannot retain the data attributes. The latter is important for us however, since we wish to enable the user to relate the blended color to the respective channels (see our discussion in Section 2.2).

RadViz [17] fulfills this goal, but similar to Star Coordinates [20] and Generalized Barycentric Coordinates [25] it may result in an ambiguous display where data points far apart in high-dimensional space can map closely in the 2D display. The three-way optimization scheme we presented in [7] absolves that, creating a display in which (1) similar (correlated) attributes map closely on the RadViz ring, (2) data points close (far) in high-dimensional space also map close (far) from one another in the 2D display (gauged by Euclidian distance), and (3) the display locations the data points are mapped to are proportional to the values they have for the corresponding attributes located on the RadViz ring. We note that we normalize all dimensions into a [0, 1] interval prior to mapping.

Finally, another paradigm we might use is the data context map [8]. While it also maps attributes and sample points into a common space, it intersperses them which makes integration with a color map difficult.

3 Overview

Multi-field data [18] often come on irregularly and possibly sparsely sampled geo-domains. This can lead to visualizations that are difficult to interpret due to a lack of continuity. Suppose we have m sample points and for each such sample point P_i, there are n attributes. For the sample point P_i, its attribute vector D_i can be recorded as

$$D_i = [d_{i1}, d_{i2}, \ldots, d_{in}],$$

where d_{ij} is the jth channel value of the ith sample point. Conversely, we can also construct a vector for each of the n attributes, comprised of the m samples. For instance, the jth attribute V_j, is then represented as:

$$V_j = [d_{1j}, d_{2j}, \ldots, d_{mj}].$$

The geolocation of P_i, can be represented as the 2-tuple

$$[P_{iX}, P_{iY}],$$

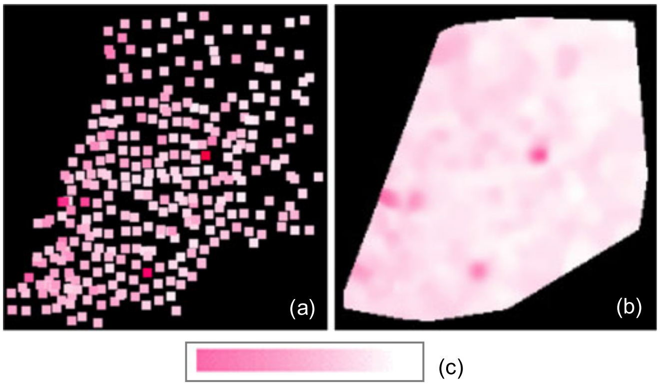

Fig. 3. Visualizing the "As" factor in the pollution data (a) Irregularly sampled observations. (b) AKDE interpolation (c) Color legend - range is [1.61, 30.13].

and it is the sample or pixel location in the original geodomain or image, respectively. Alternatively, the geolocations can also be determined by a two-dimensional space embedding, such as MDS, PCA, etc. (see Section 6.4) of the high-dimensional data. In the latter case the multivariate data vector plays a dual role—it determines the color *and* the geolocation.

As a running example, we will use a dataset of 300 multivariate pollution samples obtained at irregularly placed sensors in a large Asian metropolitan area. This dataset consists of spatial measurements of several heavy pollutant chemicals—As, Cd, Cr, Cu, Hg, Ni, Pb, and Zn. Fig. 3a shows a visualization of the As factor with concentration mapped to luminance and each sample represented by a small tile. Fig. 3b shows the same data now interpolated with adaptive kernel density estimation (AKDE) [21]. AKDE adapts the kernel used for interpolation to the local sparseness of the data, using a wider kernel over samples situated in low-density regions, and vice versa. The interpolated map makes it much easier to appreciate isolated and grouped hot spots as well as uneventful areas. For this reason, we will only use the AKDE-interpolated domain for irregularly spaced data.

Fig. 4d shows the AKDE-interpolated maps for all eight pollutants arranged into small multiples. We observe that the disjoint display makes it difficult to appreciate spatial correlations that may exist among the pollutants. In the next sections we describe our interface, ColorMapND, designed to overcome this challenge.

3.1 The ColorMapND Interface

Fig. 4 shows the interface of our ColorMapND system for the aforementioned pollution dataset. It consists of the following four components: (a) Color Legend Panel, (b) Pseudo-Colored Plot, (c) Local Enhancement Panel, and (d) Channel View.

The *Color Legend Panel* (a) contains the *circular interactive multivariate color mapping display (ICD)* with the color map doubling as a color legend. The vertical slider on the right can be used to rotate the ICD's outer ring and with it the attributes and the assembly of data points, and so alter the mapping's color assignments. The *ring spacing* check box allows users to choose the attribute layout scheme along the ring—uniformly spaced or correlation-optimized. The *color contrast* check box sets the system into the color-preserving or data-driven color enhancement mode. The *ellipse size and magnification sliders* are

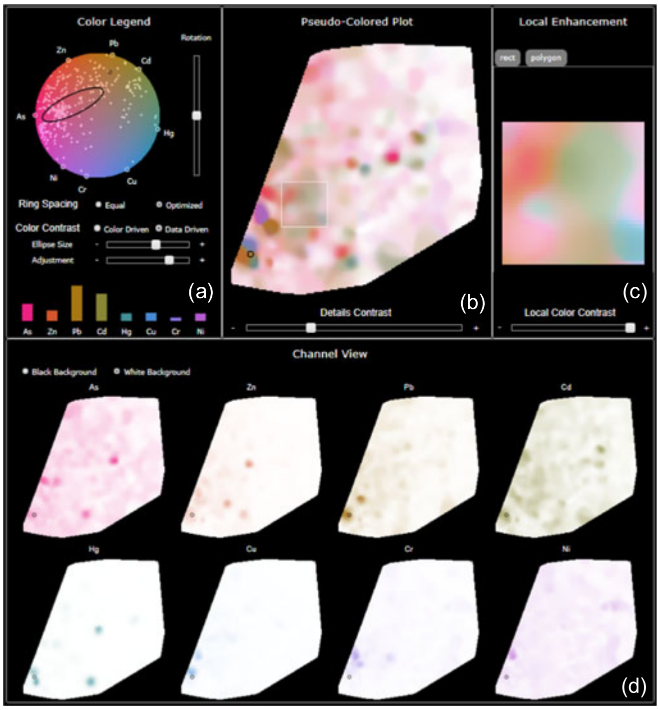

Fig. 4. The interface of our system, using the pollution dataset as a demonstration example.

used for detail enhancement (see Section 5.4 for all). Finally, the bar charts on the bottom visualize the true values for each attribute of a given point (see below).

The *Pseudo-Colored Plot* (b) in the center shows the colorized image. The *details contrast* slider can be used to control the strength of the length-to-opacity mapping (see Section 5.1). The *Local Enhancement Panel* (c) displays the locally color-enhanced area chosen by a rectangle or polygon drawn into the colorized image (shown here as a white box). The degree of color enhancement θ can be controlled by the slider below the image. Optionally, users can also color-enhance the entire colorized image.

The *Channel View* (d) on the bottom is a small multiple view of all attribute/channel images, each colorized by the color selected by their respective node points in the ICD's outer ring. This display allows users to focus on one attribute at a time.

Our system is fully interactive (after an initial 3-4s setup time for a newly loaded dataset) and lends itself well to exploratory scenarios. Moving the mouse over the colorized image or within the ICD updates the bar chart of the Color Legend Panel with the channel values of the moused-over point. This gives users quantitative information about the point and can further help them recognize the fusion of the colors.

Mouse interactions in one display are conveyed in the other displays as well, essentially linking them together for ease of visual information retrieval. Observe that in Fig. 4, each of the displays has a point circled in black (bottom left in the images, top center in the ICD). The dots move synchronously no matter in which physical display the mouse actually is. In this particular example we can easily learn that the (circled) heavy pollutant area has high "Pb" and "Gd".

In the following sections we will first describe the basic framework and then move to the more advanced algorithms and operations.

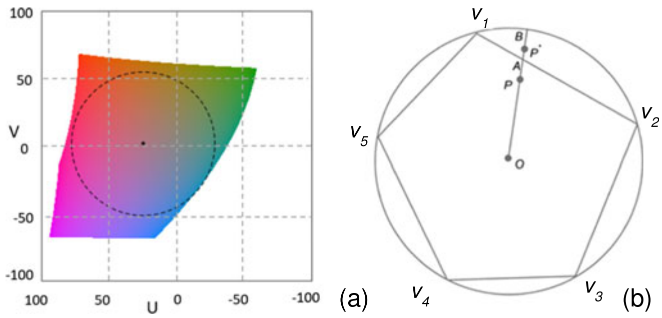

Fig. 5. Effective use of the HCL color space: (a) the optimal HC slice at $L = 55$ with the maximal circle; (b) the polygonal mapping region of RadViz and our extension to a circle to enable the full use of the CIE HCL color space.

4 The Basic Framework

The three fundamental tasks of our multivariate color mapping framework are as follows:

1. Convey dissimilarities in the multivariate data as perceivable differences in color → visually encode the data sample to data sample relationships.
2. Convey dissimilarities of the attributes as perceivable differences in color → visually encode the attribute to attribute relationships.
3. Convey associations of a data sample with the attributes as a perceivable labeling in color → visually encode the attribute to data sample relationships.

The mediating interface of our framework is the representation gained by fusing the optimized RadViz display with an equally-shaped color map, forming the ICD. The accuracy of both of these components is prerequisite to the accuracy in the three main tasks listed above.

In terms of the spatial embedding of the multivariate data into the ICD, the first and third tasks have been addressed to a large extent by the framework published in [7]. We summarize it in Section 4.3 and describe how we adapted it for the circular boundary of the ICD. The second task is addressed by a novel similarity-based attribute ordering and spacing. This is described in Section 4.2.

Having achieved a faithful spatial embedding of the multivariate data we next require a perceptually accurate color mapping framework which can convert these spatial relationships to perceivable color relationships. This is one of the main contributions of this work and is described in detail in Section 4.1.

4.1 Color Mapping in the CIE HCL Color Space

Color mapping is the process of assigning color to data. It can occur in any color space. We have three requirements for this color space: (1) it should be perceptually uniform, (2) it should be disk-shaped, and (3) the HS (Hue Saturation) slices of the color space should be iso-luminant. The former two are needed to afford the geometrical mapping operations and interactions inherent to our framework, while the last is needed so that we can use the slice-orthogonal direction for vector length encoding.

Requirements (1) and (3) rule out the HSV and HSL color spaces which have a disk-shaped cross-section but have non-linear intensity variations within the HS slices. A better choice in these respects is the CIE LUV color space which is perceptually uniform [27], [30]; its shape, however, is far from circular, violating requirement (2).

Fortunately there is a lesser known color space—the CIE HCL (Hue Chroma Luminance) color space [40]—which fits our three constraints. It is a cylindrical representation of the CIE LUV color space and removes the non-linear intensity variations within a HS slice. However, even though the CIE HCL color space seems to fulfill our three requirements, there are still some inherent adverse properties which we discovered in practical use of our system. The solutions we propose to overcome these shortcomings are described in Section 5.

When dealing with color spaces it is important to note that color monitors are only capable to display colors within the triangular sRGB space which is a sub-region of the CIE space (see Fig. 2 in the supplement material for a visual depiction). The CIE HCL space we are using has regions that fall outside the sRGB space and hence our mapping may produce some colors that are not displayable. These are mainly colors in the green range bordering to blue which are located around the three o'clock position on the ICD ring. A possible solution to this problem might be to provide visual cues, such as a shaded ring segment, that would alert users to avoid these locations for the placement of important primaries. At the same time, the sRGB space includes colors that are not contained in the CIE HCL space. These are the most vibrant shades of blue and red which, however, can be recovered by our color contrast enhancement facility described in Section 5.

4.1.1 Optimal HC Slice and ICD Size and Placement

It turns out that the diameter of the HC slice changes as a function of L, and it does so in a non-linear fashion. This can be explained by the non-regular shape of the associated CIE LUV space. What this means in practice is that the capacity of an HC slice to provide a sizable set of human-distinguishable colors is dependent on L. Maximizing this number is thus desirable.

We therefore aim to find the CIE HCL slice for which the diameter is maximized. This *optimal CIE HCL slice* is the one where the associated slice in the CIE LUV space can pack the largest circle. Further, in order to provide an unbiased spectral coverage in the color map, we require the center of this circle to coincide with the CIE LUV slice's white point.

Using iterative search, we found the optimal HC slice to be at $L = 55$. We denote this optimal slice as HCL_{55} and define a coordinate system bounded by $\pm$ 100 along each of the two axes with origin at [0, 0]. The white point on this slice is at $O = (26.147,\ 1.1344)$ and the radius of the maximal circle with the white point at its center is $R_0 = 53.2$. Fig. 5a shows the optimal slice and ICD disk.

A remaining concern is that the LUV color space outside this maximal circle is essentially wasted (see again Fig. 5a). We will return to this issue in Section 5.4.5 where we describe our detail enhancement option which utilizes the colors of the entire CIE LUV space.

4.1.2 Encoding Vector Magnitude

We note that the ICD embeds the data points in terms of their *affinity* to the attributes positioned at the circle's boundary. Data points with a *relatively higher* value in attribute A (as compared to attribute B) will map closer to the

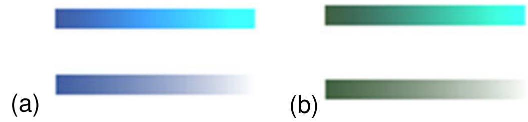

Fig. 6. Using lightness (top, range $[1 \dots 100]$) vs. opacity (bottom) for value encoding. (a) and (b) are two different colors, A and B.

boundary node of attribute A than that of attribute B. On the other hand, a data point that has the same value ratios but overall higher values than another data point will map to the same map location. Both points will then be assigned the same color and will be indistinguishable in the colorized geo-spatial display or image.

As an extra visual channel, we can use L to encode the vector length. This, however, proves problematic in CIE HCL. Consider two colors A ($\mathrm{H_A}$, $\mathrm{C_A}$, $\mathrm{L_A}$) and B ($\mathrm{H_B}$, $\mathrm{C_B}$, L_B). If we fix ($\mathrm{H_A}$, $\mathrm{C_A}$) and ($\mathrm{H_B}$, $\mathrm{C_B}$), and only change $\mathrm{L_A}$ and $\mathrm{L_B}$ from 1 to 100, we observe the upper two bars in Fig. 6. We make the following two observations: (1) the change in lightness is not linear, and (2) the color changes over the range of L. In fact, in this case, the two different colors end at the same color when $L = 100$.

Instead, we can keep the optimal HC slice at $L = 55$ and only increase transparency τ which is equivalent to decreasing opacity α, from left to right, using a white background. This is shown in the bottom two bars in Fig. 6. We observe a linear change, an L-like appearance, and a preservation of the original base colors throughout. Thus, in practice we use α to encode vector length, increasing α with increasing vector length. This will render points with greater vector magnitude in a darker color. We will denote this color space as the $\mathrm{HCL}_{55}\alpha$ color space.

4.2 Mapping the Attributes to the ICD Boundary

Placing an attribute node at the ICD boundary labels the attribute with the color at this position. We call it the attribute's *primary color*. This color is used to colorize its channel image and it allows users to quickly spot regions in the fused image which are dominated by this attribute.

The procedure we use to embed the data points into the ICD (Fig. 1a) is driven by the arrangement of attributes about the ICD's circular boundary. Each arrangement produces different data layouts and colorizations, emphasizing the criteria enforced by the arrangement.

As mentioned in Section 3.1 users have the ability to choose the attribute layout scheme along the ICD ring—uniformly spaced or correlation-optimized. In addition they also have the ability to freely position the attribute nodes on the ICD ring per their own preference, for example to highlight a certain attribute of interest in the colorization, or give it a color associated with some semantics such as blue for a variable called "Winter".

The optimized placement makes sure that the primary colors are optimally used. There are two criteria to consider for an arrangement: (1) the order of the attribute nodes, and (2) the spacing between them. Both use the pairwise (1-correlation) distance metric as the input.

4.2.1 Determining the Order of the Attributes

To determine the order of the attribute nodes on the ICD ring, we require an algorithm that can construct a closed

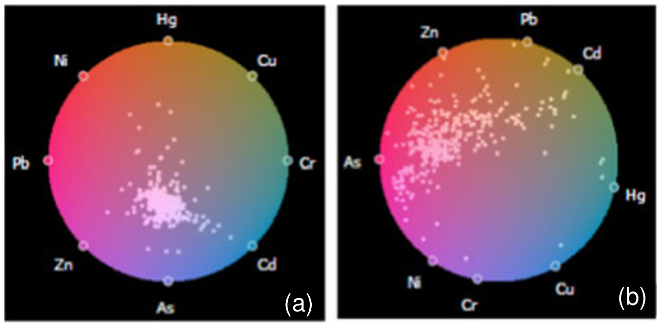

Fig. 7. Layouts as a function of attribute spacing on the color map boundary. (a) Equidistant spacing. (b) Optimized spacing.

loop since we need to place the attribute nodes along a circle. This excludes a tour generated by solving a Traveling Salesman Approximation since the ends of the salesman tour are not connected and therefore not properly spaced apart. Instead we express the task as a Hamiltonian Cycle Problem (HCP).

We solve an approximation of it (since the HCP is NP-complete) using a dynamic programming approach [3] inspired by the original scheme independently developed by Bellman, and Hell and Karp. Initially, we divide the entire set of connections into different subsets. Then we optimize for the best solution over subsets and eventually expand to the whole set. The output is an ordered set of attribute nodes which can be placed on the ICD ring, equally spaced.

4.2.2 Determining the Spacing of the Attributes

If we also wish to obtain optimal spacing between the nodes on the ICD circular boundary we can use the metric $(1 - \rho_{ij})$ where ρ_{ij} is the correlation of attribute i and j as follows:

$$s_{ij} = \frac{1 - \rho_{ij}}{\sum_{k,l \epsilon HC}(1 - \rho_{kl})} \; s_{ICD},$$

512

Here, s_{ICD} is the circumference of the ICD ring and s_{ij} is the distance between two attributes i, j on the ICD ring. The spacing we obtain groups similar attributes close together, which are then assigned similar colors. This is in some sense a dimension reduction, saving any distinct primary colors for more independent attributes.

Optimizing the arrangement of the attributes around the circle also leads to a better embedding of the data points. Fig. 7 compares the layouts obtained with (a) an equidistant ordering, and (b) an optimized ordering. We observe that in (a) the data points are lumped together and overlap in, while in (b) they are more scattered which in turn will yield more diversity in the colorization.

4.2.3 Upper Bound on the Number of Attributes

There are natural limits rooted in human color contrast perception which bound the number of attributes that can be reasonably encoded. For the CIE LUV color space, the least noticeable difference (JND) ΔE_{uv} in the UV plane is $\sqrt{\Delta u^2 + \Delta v^2} = 13$ which is equivalent to the difference in brightness $\Delta L = 1$, assuming a color cube sized ± 100 [26]. The disk of our HC color space (see Section 4.1.1) has a

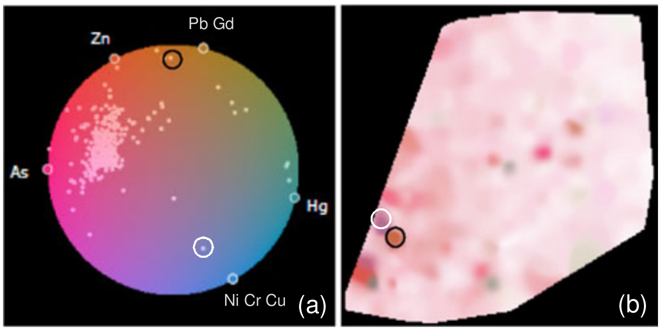

Fig. 8. Interactive color assignment for attributes using the pollution dataset (a) Color map with point display, (b) colorized geo-spatial domain. The black and white circled points are interesting outliers.

circumference of $s = 2\pi\ R_o = \ 2\pi{\cdot}53.2 = 334.26$. Thus the number of distinguishable primaries in a ring layout with uniform spacing is $s/\Delta E_{uv} = 334.26/13 = 25.6 \approx 25$[1].

This number is equivalent to an angular spacing of 14.4° of the attributes on the ICD ring. Hence any attributes spaced closer in an optimized layout will not be well distinguished. This places a certain advantage for the uniformly spaced layout, but on the other hand, it encodes highly correlated attributes in a similar color which is semantically meaningful.

4.2.4 Interactive Arrangement of the Attributes

Apart from the automated attribute ordering and spacing sometimes a targeted interactive placement can help in gaining insight into the data. Consider Fig. 8a where we interactively grouped the most correlated (> 0.75) attributes "Pb" and "Gd" as well as "Ni", "Cr", "Cu". We observe that most points form a single cluster, but we also observe some outliers. These outlier points are dominated by different attribute combinations. For example, the point circled in black is dominated by "Pb" and "Gd" and the point circled in white is dominated by "Ni", "Cr" and "Cu". After checking their spatial locations in the colorized image (Fig. 8b), we see the black circled area which is dominated by "Pb" and "Gd". Such a finding can be important for residents living in that area, or to their environmental control agency.

4.3 Embedding the Data Points Into the ICD

As mentioned, for embedding the data points into the ICD we adopt the layout scheme described in [7], which is an optimized version of RadViz [17]. In native RadViz, the location P of a data sample $D = [d_1, d_2, \ldots, d_n]$ mapped into the interior of the RadViz disk is computed as:

$$P \ = \sum_{j=1}^{n} w_j v_j \qquad w_j = d_j \ / \sum_{k=1}^{n} d_k,$$

where v_j is the location of attribute node j on the disk's boundary.

As discussed in Section 2.3, the optimized version of the scheme is designed to enforce that (1) similar data points are driven to similar plot locations, and (2) data points with an affinity for certain attributes are driven more closely to these nodes. We accomplish the latter with an iterative layout error reduction and the former with a force-directed sample adjustment. The interested reader is referred to [7] for a detailed description of these two schemes.

1. This is somewhat of an approximation since we approximated the Euclidian distance with a curve. But the error is not large.

4.3.1 Extending the RadViz Polygon to a Circle

The linear equations that underlie RadViz (and also our optimization of it) map data points into a convex polygonal region defined by the attribute vertices v_j. However, the CIE HCL color space has a circular boundary. Therefore, as shown in Fig. 5b, there are pocket regions outside the polygonal extent in the CIE HCL color space that would never be considered in the colorization.

To accommodate the full CIE HCL space we devised a method that enlarges the polygonal mapping to a disk. Suppose a point P located inside the polygon. Its new position $P*$ in the color space disk with center O can then be obtained by (see Fig. 5b):

$$\frac{OP}{OA} = \frac{OP^*}{OB}.$$

4.3.2 Looking Up the Color

The CIE HCL color space is a cylinder where each slice is indexed in polar coordinates, H and C, and the slice itself is selected by L. H is the angular and C is the radial coordinate. The color $(H,\ C,\ L)$ of $P*$ can then be calculated as:

$$H \ = \tan^{-1}\left(\frac{P_Y^*}{P_X^*}\right) \qquad \mathrm{C} = \sqrt{{P_X^*}^2 + {P_Y^*}^2} \qquad L = 55$$

where P_X^* and P_Y^* are the components of point $P*$.

To display the HCL color, converting it into RGB is necessary. This takes three steps. First, convert the HCL color into LUV space. This is a simple transform from polar coordinates to Cartesian coordinates. Second, convert the LUV color into XYZ by first obtaining the white point and then performing a transform via non-linear mapping. Finally, convert the XYZ color into RGB by a linear transform.

5 Additional Functionalities

When testing the basic framework with some real-world datasets, such as the pollution data presented so far as well as others, we came across a few shortcomings that needed to be addressed to make our system generally practical. The solutions we derived for this purpose are described in the following subsections.

5.1 Distribution-Based Vector Magnitude Encoding

In Section 4.1.2 we argued for the use of opacity to encode the magnitude of a multivariate vector in the colored domain. Domain pixels with a larger magnitude will have a higher opacity and therefore a more pronounced visual appearance. Fig. 9a shows a colorization of the full 8-channel pollution dataset—its corresponding color map is shown in Fig. 7b. While we can see some areas with stronger colors, we also observe that overall the colors are somewhat washed out. This is because the simple uniform opacity mapping scheme cannot deal with the wide distribution of vector lengths.

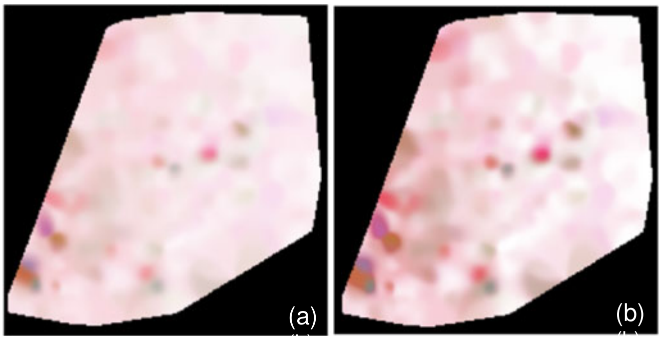

Fig. 9. Opacity encoding of vector magnitude using the 8-channel pollution dataset (a) linear encoding, (a) distribution encoding.

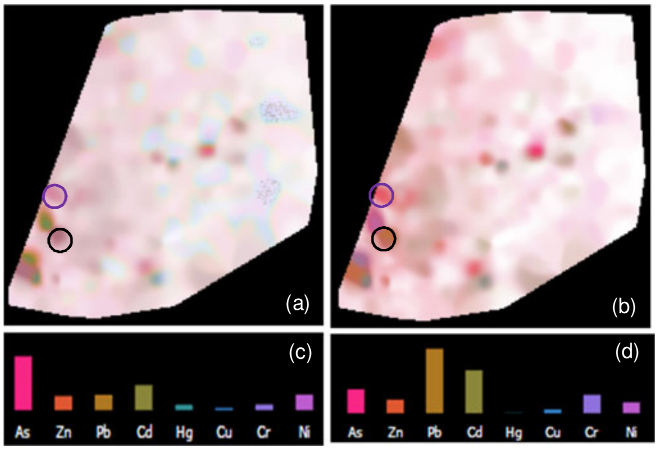

Fig. 10. Coloring irregularly sampled domains. (a) Color first, interpolate second; (b) weighted scheme (c) multivariate spectrum of the point circled in blue (d) spectrum of the point circled in black.

We devised a distribution-aware [11] mapping scheme to overcome this problem. We can reasonably approximate the distribution of vector lengths $[l_1, l_2, .., l_m]$ by a normal distribution, $G(\mu_l, \sigma_l)$. We then standardize and transform this distribution such that it has a more favorable dynamic range for mapping vector length to an opacity interval of [0,1]. A transformed vector length, l', is then given as:

$$l' = \left(\frac{l - \mu_l}{\sigma_l}\right)\sigma_g + \mu_g,$$

where l is the original vector length and $\sigma_g = 0.25$. For μ_g, the default value is 0.5, which can be changed in our interface to visually enhance certain detail. As such, 68% of the points will fall into the range $[\mu_g - \sigma_g, \mu_g + \sigma_g]$.

In experiments we found that it can be beneficial to taper off the tails of the distribution. This brings out smaller length variations more clearly and de-emphasizes noise and outliers. Suppose, for a given setting of μ_g the smallest value of $(l'_1, l'_2, .., l'_m)$ is l'_{min} and the largest is l'_{max}. We define an opacity encoding function, Φ, which takes a vector length value l', and converts it to an opacity $\Phi(l')$:

$$\Phi(l') = \begin{cases} (\mu_g - \sigma_g)\frac{(l' - l'_{min})}{(\mu_g - \sigma_g - l'_{min})} & l' < \mu_g - \sigma_g \\ l' & l' \in [\mu_g - \sigma_g, \mu_g + \sigma_g] \\ \frac{(l' - \mu_g - \sigma_g)}{(l'_{max} - \mu_g - \sigma_g)} + (\mu_g + \sigma_g) & l' > \mu_g + \sigma_g \end{cases}.$$

Since it is difficult to set a proper μ_g value for the opacity mapping in advance, we allow users to interactively change it within the range [0, 1]. This moves the unity-sloped midsection of the mapping function to the left (right) which decreases (increases) the overall opacity enhancement. Fig. 1 in the supplement material provides a visualization of this function.

Fig. 9b shows a colorization obtained with this method for $\mu_g = 0.3$. We see that it provides considerably more detail and contrast than the plain encoding of Fig. 9a. The video shows an animation across the range of μ_g.

5.2 AKDE Interpolation of Multivariate Colorizations

In Section 3 we discussed AKDE interpolation as a means to convert an irregularly sampled domain to a regular one. We demonstrated this method using a scalar field with a single color channel. AKDE interpolation of scalar domains has been well described in the literature [21]. In this section, we expand single-channel AKDE interpolation to multivariate colorized domains.

There are essentially three different strategies distinguished by where the color interpolation occurs—in the color space or in the domain image. All methods begin by embedding the multivariate irregularly spaced data samples into the HCL_{55} color map using the ICD widget.

Color First, Interpolate Second. In this scheme, each domain sample is mapped into the ICD to obtain its color. Then, AKDE-based interpolation is used to estimate the colors of the remaining pixels in the domain image. Fig. 10a shows a colorization obtained with this procedure. It has rather low quality—it looks quantized and has very little detail. A comparison with the true multivariate spectra confirms that the colors are not overly accurate. Compare, for example, the rather bland colors in the blue and black circle in Fig. 10a with the actual multivariate spectra of the corresponding data points shown in Figs. 10c and 10d, respectively.

Interpolate First, Color Second. Here a pixel color is obtained directly from (interpolated) multivariate values. In this procedure we would perform AKDE on the multivariate data and then look up the colors for each interpolated pixel. However, in order to convert a multivariate vector into color, it is necessary to compute its position in the ICD. This is not an easy undertaking since due to the non-linear optimization during the layout, the original position in the ICD to value has been lost. The only way to find the color would be to re-optimize the layout for both types of points—original and AKDE interpolated—an expensive operation.

Interpolate First, Indirect (Weighted) Color Second. This scheme is a compromise which is in some sense reminiscent to LLE [29]. It learns the interpolation weights in the image domain and applies them in the information domain (represented by the ICD). This is expressed in the following equation, which is based on Nadaraya–Watson kernel regression with kernel function $K_h()$. Here, P_i is one of the m original sample points, P_i^* is its corresponding location in the ICD, P is the pixel to be interpolated, and P^* is its corresponding location in the ICD, calculated using the weights learned

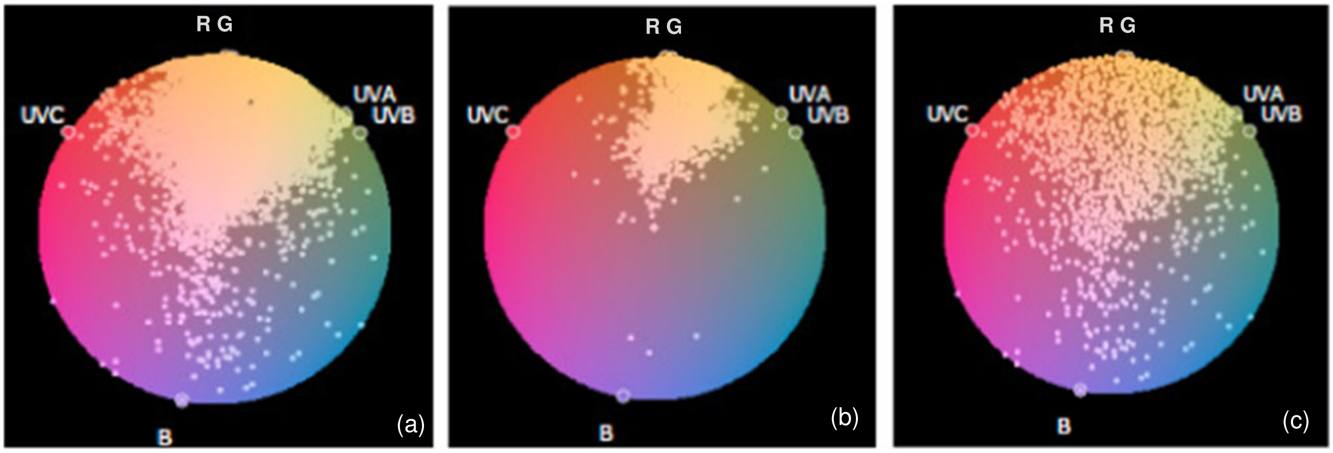

Fig. 11. Data sampling schemes: (a) original distribution, (b) down-sampled, (c) hashmap sampled.

from the AKDE in the image domain. Using this equation H and C of P are looked up in the ICD at location P^*:

$$P^* = \frac{\sum_{i=1}^{m} K_h(||P - P_i||)\Delta P_i^*}{\sum_{i=1}^{m} K_h(||P - P_i||)} \qquad HC = \mathrm{ICD}[P^*].$$

The computational cost is manageable since it does not require a re-optimization of the layout for each pixel.

Fig. 10b shows the result of this interpolation. We find that it preserves the original multivariate spectrum quite well (compare the blue and black circled points with the spectra on Figs. 10c and 10d, respectively).

5.3 Dealing with Large Data

Information displays such as our ICD suffer from overplotting when the number of data points gets large. In our case this leads to conditions where the colormap becomes difficult to read (see Fig. 11a). Such occasions arise when we use the ICD to colorize full-res multi-channel images, such as the multispectral images shown in Fig. 17. Likewise, a large number of attributes leads to an unrecognizable number of primary colors. In the following, we describe techniques that can deal with these problems.

5.3.1 Sparsification of Large Point Clouds

A first solution is to render the data points crowding the ICD semi-transparently. This can help somewhat in recognizing the colors in the colormap layer below, but the visualization is still too cluttered. We also experimented with traditional down-sampling methods which select samples based on density or randomly but none produced satisfactory results. However, all of these methods tend to neglect outliers and sparsely occupied areas. This is evident in Fig. 11b which shows the result we obtained by a density-based down-sampling of Fig. 11a.

Instead, we have opted for a stratified sampling approach based on a 2D hashmap. Our method imposes a 200×200 2D grid onto the color map and visits each point in turn. Initially, we create a global sample list to store the points after sampling. When a point maps into a so far unvisited grid cell, the point is added to the global sample list. At the same time, the point's four grid neighbors are frozen. This prevents any new point mapping to it from entering the sample list. The high density areas get more samples while low density areas do not, using the following mechanism. Every grid cell keeps a counter which increments whenever a point maps to it. If the count exceeds a set threshold, the neighbors of this grid cell are unfrozen, freeing them for the global sample list. Once finished, the global sample list is plotted onto the map.

Fig. 11c shows a result of the stratified sampling algorithm. We observe that the algorithm retains both the outlier points and the main distribution, but at the same time reveals the color map in the layer below.

5.4 Zooming and Contrast Enhancement

Oftentimes the color map is only partially filled by samples, with a few outliers in the remaining regions. While this is tolerable in conventional scatterplots with clusters, in our application it leads to an underuse of colors. The consequence is low color contrast in the image domain. See, for example, Fig. 12a. We observe that the points mostly use colors in the upper part of the HCL_{55} space. The resulting colorization (see Fig. 12e) is consequently somewhat flat with a few isolated hotspots. Compare this with Fig. 12g which uses the considerably more uniform point distribution of Fig. 12c for colorization. The resulting image is much more vivid and offers significantly more detail information. Some good examples are the areas enclosed in the small and large circles. The following subsections present several methods we designed.

5.4.1 Extracting the Main Cluster of Points

We use an approach akin to a *magnifying lens* to increase the spread of points on the color map. We chose an elliptical shape for this lens. We found that this makes the lens easy to manage and at the same time enables it to capture the typical shape of most point distributions.

In order to find this ellipse, we first use k-means clustering with $k = 1$. This yields the main cluster and its center M. Next, we use Principal Component Analysis (PCA) [19] to determine the distribution's extent as a set of two eigenvectors (black arrows in Fig. 12a), with two sorted eigenvalues λ_1 and λ_2. The ellipse is always drawn as a black outline (see Figs. 12a, 12b, , 12c, and 12d).

We consider the *interior points* falling into the elliptical lens the *core features*, and the *exterior points* the *peripheral features* and *outliers*. Users can increase (decrease) the extent of

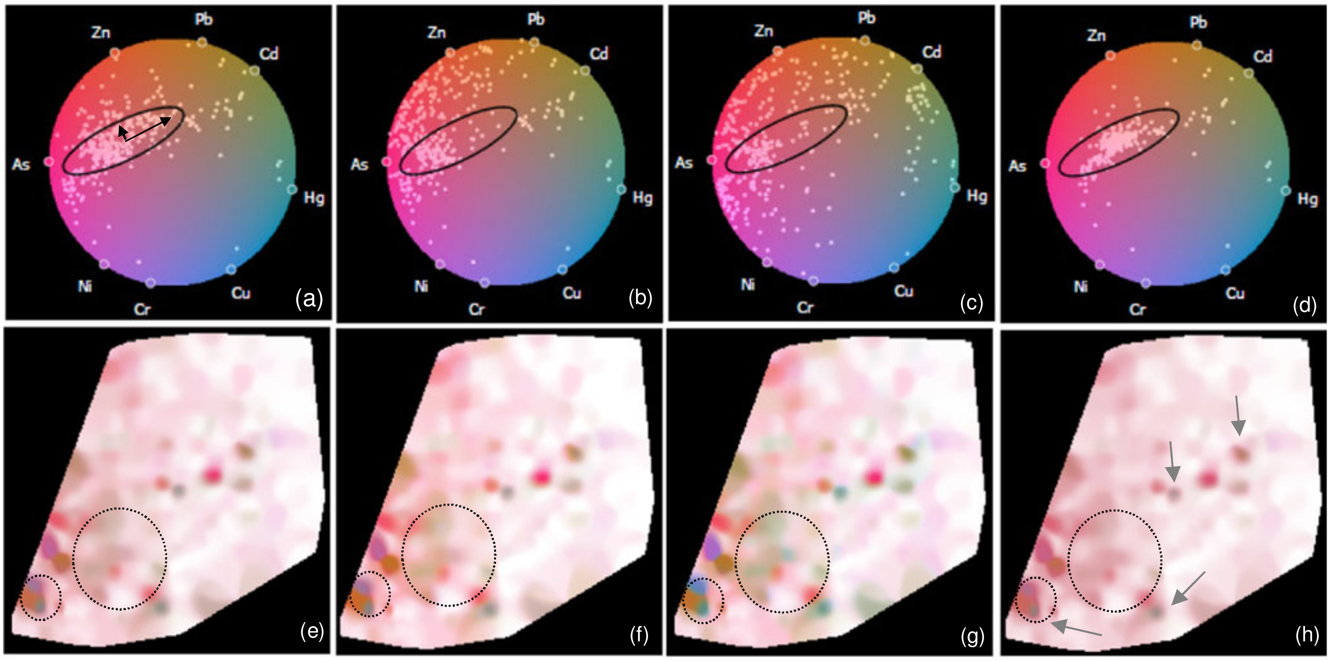

Fig. 12. Various color contrast enhancement schemes for the pollution dataset (top row: joint color map - point display, bottom row: colorized spatial domain. (a, e) original coloring, (b, f) color driven scheme, (c, g) data driven scheme, (d, h) outlier enhancement (CCC scheme).

the magnifying lens and so include (exclude) further interior points. This operation scales the eigenvectors and yields a larger (smaller) ellipse. In the limit the ellipse is the entire color space disk. This is technically done by increasing the lengths of the eigenvectors using an adjustment parameter β:

$$a = \frac{\lambda_1}{2}\beta \qquad b = \frac{\lambda_2}{2}\beta.$$

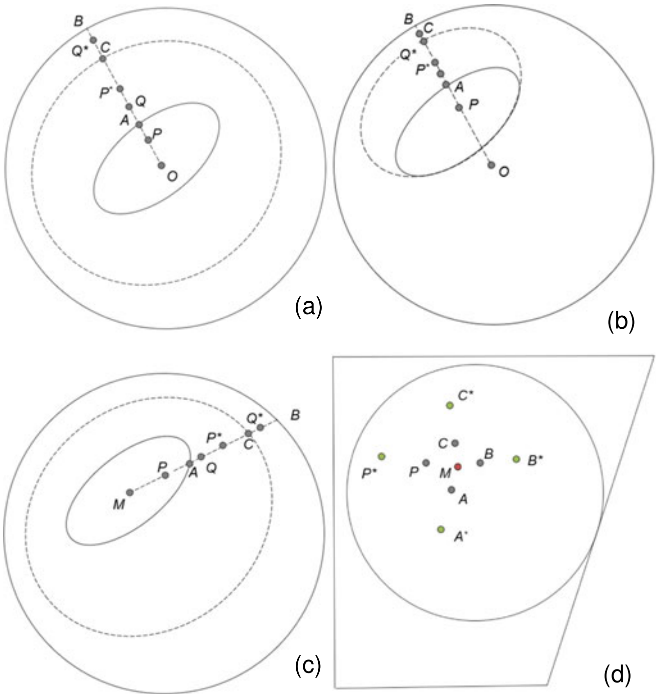

Fig. 13. Illustration of the color contrast enhancement schemes (a) color driven scheme with the color space center inside (b) or outside the ellipse (c) data driven scheme and (d) local color enhancement scheme (the polygon represents the UV color space).

As default, $\beta = 1$, and β can be adjusted via a slider.

With the interior and exterior points defined, expanding the ellipse during magnification will spread the interior points onto more color space and give them more contrast in the image. Exterior points on the other hand will compress and lose contrast. In that respect they behave like points that fall into a lens transition region.

There are some downsides of this general scheme. First, an increase in color contrast will diminish the visual effect of similarity. Second, points may change their hue in the expansion. This gives rise to two separate enhancement schemes. We will describe these two schemes in the following sections.

5.4.2 Color-Preserving Contrast Enhancement

This scheme seeks to preserve the hues of the points and only changes saturation. It observes the center of the color space, O, and pushes the interior points along lines emanating from O towards the border of the circular color space. Fig. 13a presents an illustration when the center of the color space is inside the lens. In this figure, the inner ellipse is the original shape of the lens while the outer ellipse is its coverage after magnification using the parameter θ:

$$\vec{\mathrm{OC}} = \vec{\mathrm{OA}} + \theta\left(\vec{\mathrm{OB}} - \vec{\mathrm{OA}}\right) \qquad \theta \in [0, 1], \tag{1}$$

When $\theta = 0$ then there is no magnification, while when $\theta = 1$ there is full magnification. In the latter case, the interior points are spread over the entire color map and the exterior points map to the map's boundary. For all other values of θ the interior points map to the larger ellipse and the exterior points map into the adjoining annulus region.

An original interior point P moves to a new location $P*$ per the following relationship:

$$\|OP^*\| = \frac{\|OC\|\|OP\|}{\|OA\|}. \tag{2}$$

An exterior point Q, on the other hand, moves to a new location $Q*$ computed as follows:

$$\|Q^*C\| = \frac{\|BC\|\|AQ\|}{\|AB\|} \tag{3}$$

When the color map's center is outside the elliptical lens (see Fig. 13b), the computations are unchanged. In this case, the enhancement is not that large but it preserves more similarity.

The result of this enhancement is shown in Fig. 12b for the color space, while the corresponding colorization is shown in Fig. 12f. Compared to the original layout in Fig. 12a, the points on the top left corner spread more towards the color map boundary. We find that the colors are more vivid than in the original colorization of Fig. 12e, but they are still comparable in hue (see for example the region circled in black). Overall, we find that color contrast is increased. On the other hand, the similarity relations are still well observable since this adjustment keeps the points in their original area of the color space.

5.4.3 Data-Driven Contrast Enhancement

The data-driven scheme focuses on the center of the data distribution, M. It starts from the center of the ellipse and pushes the interior points along lines emanating from M towards the border of the circular color space. This process is illustrated in Fig. 13c. Using again the parameter θ, the enlarged area can be obtained as:

$$\vec{MC} = \vec{MA} + \theta\left(\vec{MB} - \vec{MA}\right) \qquad \theta \in [0, 1]. \tag{4}$$

The new position of an interior point P is $P*$. It is computed as follows:

$$\|MP^*\| = \frac{\|MC\|\|MP\|}{\|MA\|}. \tag{5}$$

On the other hand, an exterior point Q will get compressed and its position $Q*$ can be obtained by:

$$\|Q^*C\| = \frac{\|BC\|\|AQ\|}{\|AB\|}. \tag{6}$$

The color mapping obtained with this scheme is shown in Fig. 12c and the corresponding colorization is shown in Fig. 12g. Compared to Fig. 12b, the points are now transferred across the color space center and use the color space more effectively than the color-preserving enhancement scheme. And indeed, the colorization in Fig. 12g better visualizes the disparity among the pollution chemicals by giving the levels more distinct colors. We can observe more detail, as can be seen, for example, in the region circled in black. However, this coloring loses some of the originally expressed similarity relations, compared to the color preserving enhancement coloring.

5.4.4 Outlier Enhancement

The color contrast enhancements presented so far emphasized the main distribution points. However, sometimes it can be important to specifically emphasize points outside the main distribution, while de-emphasizing the others. Such a scheme would show these former points in vivid colors according to their attribute affinities, while the latter points would visualize in a neutral uniform color.

For this purpose, we have developed what we call the *comparison compression coloring (CCC) scheme*. The CCC scheme works for both the color-preserving and the data-driven enhancement methods. It restricts the interior points into a smaller region such that they cannot take up many colors and distract the user. In this way, the color map will give more room to the exterior points. However, in this compression, we cannot simply set the parameter θ less than 0 (for shrinking the lens) and compute the layout via Equation (3) or (6). If so, any outliers should be pulled to the ellipse as well. Rather, we would like to preserve the isolated status of these outliers. For this reason, we build a weight function based on the distance from the center of the color space or the ellipse, respectively. The weight is defined as follows:

$$W_p = G(\mu, \sigma)(\|MP\|) \qquad (\mu = 0, \sigma = 0.5).$$

For the color driven scheme, Equation (1) becomes:

$$\vec{OC} = \vec{OA} + \theta\left(\vec{OB} - \vec{OA}\right)W_P \qquad \theta \in [-1, 0].$$

For the data driven scheme, equation (8) changes to:

$$\vec{MC} = \vec{MA} + \theta\left(\vec{MB} - \vec{MA}\right)W_P \qquad \theta \in [-1, 0].$$

The new location of point C can be obtained from the above equations. Based on the new location, we could then compute any point's new location via Equations (2), (3), (4), (5), and (6). The color map of this enhancement scheme is shown in Fig. 12d. We observe that the points inside the ellipse now occupy a smaller region, using only a few representative colors. The corresponding colorization is shown in Fig. 12h. We see that the most dominant main features are now visualized in a rather neutral and uniform color. They essentially form a contextual backdrop for the more color-enhanced outliers, where the color identifies the composition of the outlier. For example, in the circled regions we see outlier spots that were difficult to identify as such in the other colorings (for example in Fig. 12g) due to over-crowding, but they are now clearly visible. We also inserted arrows to point to some of the outliers.

5.4.5 Local Enhancement Using Colors Outside HC Disk

As mentioned in Section 4.1.1, some parts of the LUV color space are wasted since the HCL_{55} circle cannot cover the entire convex region of the UV space. To account for this, we provide a feature called *detail enhancement mode* that also makes use of colors outside the HC disk. In this mode, when we push the points toward the circular border, we allow them to cross the circle boundary and spill into the peripheral regions of the UV space. As shown in Fig. 5a, this gives the colorization access to stronger shades of purple, green, orange, and blue—the colors outside the HCL_{55} disk.

We distinguish between local and global color enhancement mode (see below). In local color enhancement mode, the user can specify an area of interest by drawing a

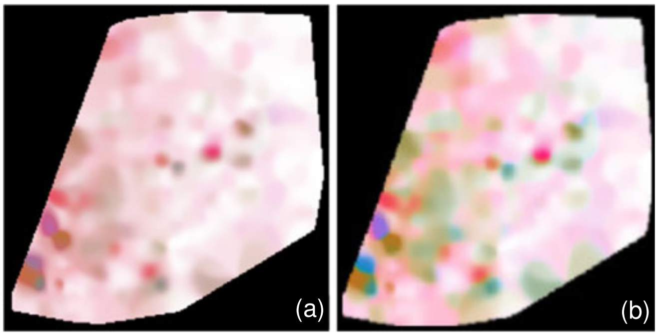

Fig. 14. Global color enhancement: (a) Original colorization using only UV colors within the HC disk. (b) Enhanced colorization also using UV colors outside the HC disk.

rectangle or polygon on the colorized image. The system then responds by providing a, possibly enlarged, detail image whose colorization only depends on the points that are part of the selected patch. Fig. 1c shows an example for this—the colorization of the image patch bounded with a square in Fig. 1b. It is easy to see the structural information coded by the variation in color in the detail patch, while it is not visible in the large image.

The algorithm works as follows. After a patch has been defined, the set S of all points falling into it is identified. Next, the center M of S is computed, and the points of S are either pushed away or dragged closer to M depending on the type of enhancement—exterior or interior. Fig. 13d shows an illustration of this process. Suppose S comprises points $\{A,\ B,\ C,\ P\}$ with center M. We perform a local enhancement using the displacement parameter θ. This moves S to $S*$ composed of $\{A*,\ B*,\ C*,\ P*\}$. $P*$ is computed from P as:

$$\vec{MP^*} = \theta\ \vec{MP} \qquad \theta \geq 0, \tag{7}$$

When $\theta<1$ this performs a compression, while when $\theta>1$, it performs an enhancement.

And indeed, we observe in Fig. 1c that these extra levels of pink have been used to fill in and expose the previously hidden structural variations.

5.4.6 *Global Enhancement Using Colors Outside HC Disk*

Global color enhancement mode expands the local area scheme to the entire image. We provide two options: (1) after users have enhanced the colors of a local area they can apply the local detail settings to the entire image, and (2) users can perform an enhancement to the entire image directly. The latter is equivalent to drawing the selection polygon to include the entire image.

A result of this procedure is shown in Fig. 14b using the pollution dataset. Compared to Fig. 14a, which is the original colorization only using colors within the HC disk, we obtain a significantly improved contrast and richer colors which allows more detail to be observed.

One might ask, why not always use these exterior UV regions. While the layout optimization schemes described in Sections 4.2 and 4.3, and [7] could easily support the convex shape, we would need to forego the ability to rotate the color space for user-defined color-attribute assignments. The two enhancement options we provide seemed to pose a good compromise.

We end by noting that whenever the user performs a rotation of the color space, or other operation, the points are pulled back into the HC disk and the image is reset.

6 Implementation and Use Cases

Our system is implemented as a client-server model. The client application uses the D3 JavaScript library [5] and can run on any modern web browser. The server application is written in C# and runs on an online compute server hosted in our laboratory.

Almost all aspects of our system were incrementally developed with domain scientists in the loop, giving us feedback and inspiring new features or modifications thereof on a routine basis. We worked with several groups of scientists, about 100 in total. They came from physics, material science, chemistry, computer science, environment science, and medical science. Proprietary restrictions preclude us from presenting some of the results we obtained in this paper. Yet, the following sections attempt to give an overview on the wide spectrum of applications in which our system has been deployed, tested, and evaluated.

6.1 Environmental Science—Pollution Data

We already used these data throughout the paper to demonstrate the various system features. Our collaborators are a group of environmental scientists who have been collecting a large amount of environmental monitoring data recording many toxic elements (see Section 3). The data originate from several major cities located in Shandong Province, China and hence they were not sampled on a regular grid. This inspired the development of the multivariate AKDE interpolation framework described in Section 5.2.

Due to the large number of variables, the scientists preferred the optimized attribute layout. This allowed them to capture the relations of the attributes directly in the display. They found this system feature rather convenient.

In the sessions we attended, the scientists applied both the color-preserving and the data-driven enhancement modes in their analyses. We also observed they used the outlier enhancement mode repeatedly. Moreover, they kept using the local detail contrast function, commenting that it enabled them to distinguish the color gamut by adjusting the opacity from different scale levels. The insight they gained using our system has been presented throughout the paper in figure captions and in the text.

6.2 Physics—Battery Data

Our scientific collaborators were a group of physicists and material scientists working at the National Synchrotron Light Source II (NSLS-II) at Brookhaven National Lab. They were looking for a tool that could help them understand a Fluorescence dataset of a battery material, scanned at the lab's hard X-ray nanoprobe beamline. The data are composed of an image stack of four different elements: “Ce”, “Co”, “Fe”, and “Gd”. This mixed ionic-electronic conductor denoted as CGO-CFO is widely used as battery in fuel cells. The key feature of this composition is the formation of a dual phase, thus, locating the new emerging phases is essential to

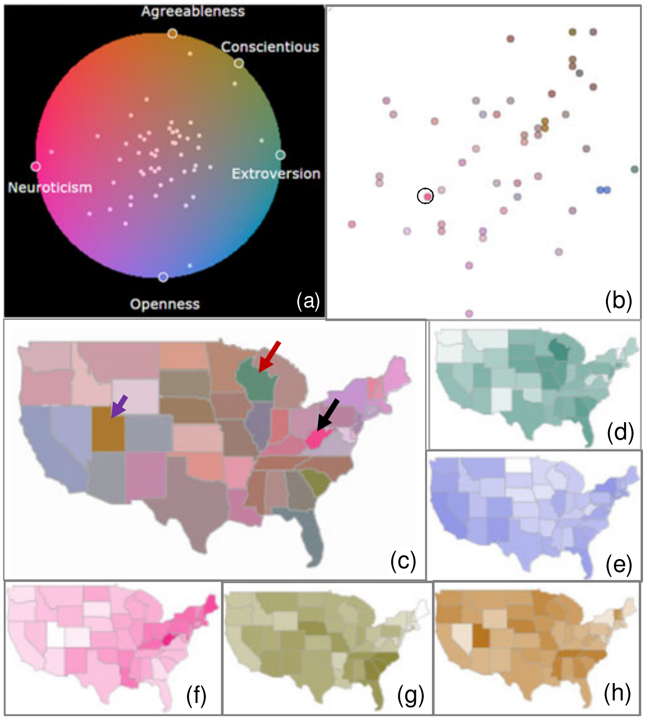

Fig. 15. A pseudo-coloring of the US states personality dataset (c) and its color legend (a).The arrows point at states with outlier behavior. (d)-(h) Individual choropleth maps of (d) extroversion, (e) openness, (f) neuroticism, (g) agreeableness, (h) conscientiousness. (b) MDS plot of all data and colorized with the ICD.

understand the conductivity and performance. Specifically, the scientists sought to (1) learn about possible interactions of the four elements, (2) see at which spatial locations these interactions occur, and (3) detect subtle component changes that might indicate the location of a new phase. They told us that their current tools were too tedious to use especially when the number of elements was beyond three when they could no longer fuse the data into RGB images.

Fig. 1 shows one of the dashboard visualizations the scientists created. The dashboard presents one of the key discoveries the scientists made when using our software. In the exploration that led to this dashboard they were looking for phase changes. It is difficult to see this type of incidence in an individual element map. Using our tool they could fuse the channels and soon they focused on the circular area pointed to by the arrow. They quickly identified the crescent area as being mostly composed of "Gd" since its color is purple. But in the upper portion of that area the color starts to be mixed with blue indicating the presence of more "Ce" than in the lower part. This apparently suggests the existence and potential location of a new phase.

Next the scientists focused on the small area delineated with the white box. By comparing the color with the color legend, the scientists learned that this area was mainly made of "Ce" and "Fe" since the color is a mixture of light green and pink. They wanted to see if this mixture had any structure in it, but the image could not reveal this. So they inspected this area in the local enhancement window on the right. They found that there indeed was a structural pattern composed of irregularly shaped zones of light green ("Co") and pink ("Fe"). By later checking the phase image, scientists confirmed this finding.

6.3 Choropleth Maps

Here we showcase the application of our system to multivariate choropleth maps. The dataset we have chosen is entitled "America's Mood Map". It contains data that seeks to characterize each state in the US by the personality and temperament of its population. The data was collected through an online survey [38] of more than 160,000 Americans. The dataset captures a set of psychological traits, specifically what psychologists call the Big Five: openness to experience, extroversion, agreeableness, conscientiousness, and neuroticism. We analyzed the dataset and found via correlation analysis that agreeableness is somewhat related to conscientiousness, but is only mildly correlated with extroversion. The final two traits, neuroticism and openness do not seem correlated with any other trait. All of these relations are visualized by arrangement in on the ICD color map boundary (see Fig. 15a).

We quickly spot a few outliers in the color map. The associated choropleth map (see Fig. 15c) we constructed using our framework just as quickly points out what states these outliers are: Utah (blue arrow) is predominantly conscientious, Wisconsin (red arrow) is predominantly extroverted, and surprisingly West Virginia (black arrow) is predominantly neurotic. There are also other states that have slight tendencies to certain traits but not as pronounced. Nevertheless, the combined choropleth map makes it easy to spot which states have similar (and dissimilar) personality profiles, which is much harder to do with the five individual maps of Figs. 15d, 15e, 15f, 15g, and 15h.

And so, one can quickly satisfy a strike of curiosity with regards to one's own state (or any other), and also look for similar states. For example, looking at Washington and Oregon, both have quite similar personalities but are rather different from the close neighbor California. The main difference is extroversion. On the other hand, Montana is a relatively "normal" and "peaceful" state—it has almost equal and low values in all of the attributes.

6.4 Colorizing MDS Plots and Other 2D Embeddings

Another useful aspect of the colorization is the added information it can provide in 2D data embeddings, such as MDS, t-SNE, etc. For example, Fig. 15b shows an MDS layout of the personality data, colorized using the ICD with the same setting than before. By colorizing the points, we can learn about their individual multivariate composition and possible biases in certain variables. These are semantic aspects that are lost in a conventional MDS optimization, but are returned in the colorization.

We also observe that the MDS and the colorization preserve similar associations. For the most part states with similar personalities have similar locations and are also colorized similarly. Likewise, outliers pop out with different colors, for example West Virginia (black circle).

Finally, our method could also be used in bivariate scatterplots, colorizing the points to reflect the other currently missing dimensions. This, however, can lead to confetti-like plots when the colorized variables have little correlation with those plotted. It works better with MDS since the embedding optimization provides the multivariate similarity structure needed for a coherent display.

Fig. 16. Conventional representations of multispectral images. (a) RGB image of the flower, (b) ultraviolet radiation image of the flower, (c) RGB image of the terrain, (d) thermal image of the terrain.

6.5 Multispectral Images

A popular type of image with more than three channels is the multispectral image. A multispectral image can have multiple bands taken from the visible and invisible (to humans) spectrum. Examples for the latter are the UV or the IR (thermal) bands. These bands can provide additional important information but are often viewed separately from the RBG image. Fig. 16a shows a flower's RGB image while Fig. 16b shows its UV radiation image [37]. Likewise, Fig. 16c shows a terrain RGB image and Fig. 16d shows a portion of the thermal image of the same terrain [39]. Fusing the visible and invisible channels into a single image can make the information more comprehensive. It essentially gives the human eye super vision, equipping it with the IV vision capabilities of fish, reptiles, etc. and the IR vision capabilities of snakes, etc. at the same time. We have studied our system with two examples of such imagery, presented next.

6.5.1 Flower Data Set

We utilized our tool to fuse the RGB and UV channels of the flower dataset (FigS. 16a and 16b). Fig. 17 shows the results we obtained. Comparing the colorization with the channel images as well as with the RGB and UV images, we can observe that the fused image has incorporated most if not all of the detail of these partial images. The local enhancement of the white box on top of the colorization exposes an interesting UVC irregularity in the top petal. It also shows a better rendition of the multispectral texture.

6.5.2 Terrain Dataset

Next, we colorized a multispectral terrain image comprised of three natural channels (RGB) and three thermal channels (IA, IB, IC). The result is shown in Fig. 18. We observe that the fused image depicts significantly more detail than the individual natural and thermal image channels. We can also quite easily pick out the individual channel images in the fused image based on their specific colors. For example, the ocean part has a higher "temperature" than the "mountain" part since its color is more "red". Finally, in the local enhancement image we can observe a few remarkable hot spots in the mountain area.

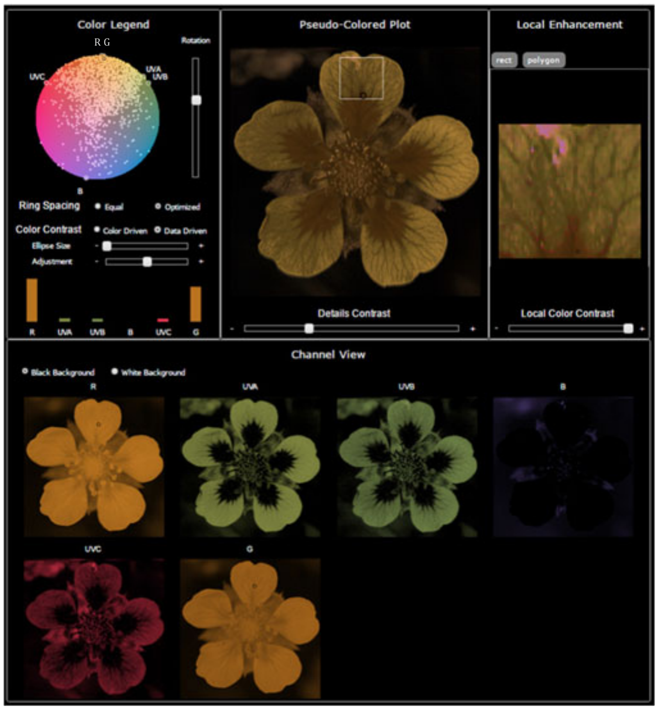

Fig. 17. Application to the 6-channel multispectral image of a flower. The bands are the natural RGB colors and the ultraviolet radiation UVA, UVB, UVC. The color map uses a more moderate level of stratified sampling to not over-emphasize the outliers.

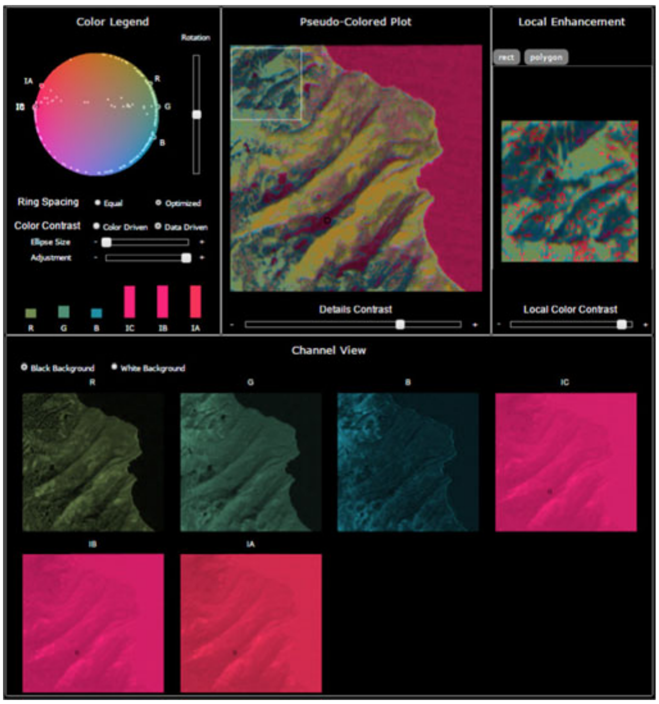

Fig. 18. Application to a 6-channel multispectral image of terrain, here an area around California. The bands are the natural RGB colors and the thermal with the channels IA, IB, IC.

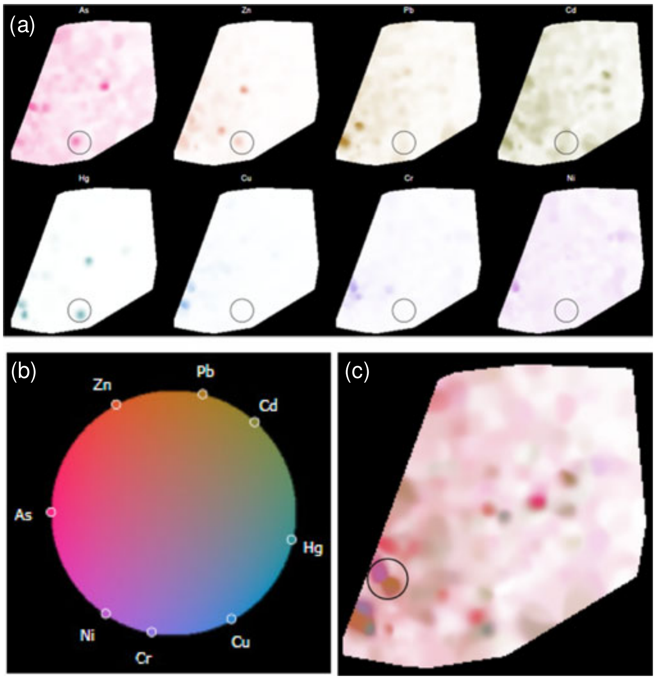

Fig. 19. User study setup (a) Segregated channel view display the circled region denotes the target. (b) ICD without scatterplot, just HCL color map; (c) Colorized domain with circled target region.

7 Assessing User Performance and Utility

We gathered insight on the effectiveness of our tool with respect to two aspects: (1) conciseness of the single-view ICD-based color encoding (in comparison to the segregated channel-based color encoding) and (2) utility of the overall interactive interface and system.

7.1 ICD-Based Encoding of Multivariate Data

To assess the strengths of the ICD-based encoding we conducted a somewhat informal (with respect to the statistical analysis) user study with 20 participants, recruited from our campus. These individuals came from various departments, such as computer science, physics, economics, and others. None of them was familiar with the types of tools that were subject of the study, namely, channel-based and ICD-based visualization of multivariate geo-referenced data. We started out with a training session to acquaint the participants with the two visual paradigms. The study was structured around the pollution dataset and the training session also educated the participants about the attributes and setting of this dataset (see Section 3). Questions were invited and a brief test was given.

In order to neutralize learning effects, the participants saw a random sequence of six cases with each being either a set of channel images (segregated view) or an ICD-based visualization. In each case we marked some area of interest by a circle and asked: "What are the heaviest pollutants in the circled area?"

Fig. 19a shows the segregated view while Fig. 19c shows the colorized image for a different target region. The channel images were of the same size than the colorized image in our study. Fig 19b shows the ICD for this dataset. We purposely left out the scatterplot to enable an unfettered view onto the color map. We did not provide the mouse-over interaction capabilities to locate the geo-points on the color map. The participants had to make their assessment using color similarity only.

At the end of each session we asked each participant which visualization paradigm he or she preferred. We asked "Do you prefer the Colormap-assist view or the segregated view?" We gave four options: colormap | segregated | both | none.

We found that both our tool and the channel images achieved similar accuracy (95%)—among the 20×6 questions, 114 questions are correct. There also was no significant difference in the time spent for coming up with an answer. The questionnaire, however, revealed that 90% of our users (18 out of 20) preferred the ICD over the set of channel images. We infer from this that looking just at one geo-image (and the ICD) is more convenient than scanning across the eight channel images. We feel that this is a good demonstration of the advantages of our approach with respect to channel scalability.

7.2 Overall Interactive Interface and System

Section 6.1 already reported some feedback we obtained from our collaborating scientists at BNL. All of them thought that our tool was very helpful since it reduced a large amount of tedious image comparison operations to just a few interactions. The linked interaction across the various panels helped them in color classification—they could easily pick the main features from the colorized image and connect them to the channel views. The bar charts helped them especially for areas with subtle color changes. They also thought the local enhancement with the selection interaction was very useful since they could go back and forth to explore more detailed features in a focused area. All in all, the NSLS-II scientists thought our tool was easy to use and very helpful in expediting scientific discovery.

8 Conclusions

We have presented an interactive framework, called Color-MapND which fuses principles from high-dimensional data visualization with principles from color science to address the longstanding problem of multi-field data visualization. A key element of our system is a multivariate scatterplot display that is overlaid onto a CIE HCL color map. Using this joint structure, a multivariate pseudo-coloring of the multi-field domain can be consistently obtained. We provide several extensions to this basic framework and apply it to regular and irregularly sampled multivariate domains, multivariate choropleth maps, and multispectral images.

We have already mentioned in Section 4.1 that standard color monitors are capable to display colors within the triangular sRGB space which exceeds our HCL disk in some CIE LUV space areas and leaves uncovered disk regions in others. The reader is referred to Fig. 2 in the supplement material for a visual depiction of this color space geometry. A possible solution for the former problem would be to provide visual cues, such as a shaded ring segment, to alert users to avoid these locations for the placement of important primaries. Alternatively, these colors can always be recovered on the fly by ways of our color contrast enhancement facility (within the extent of the sRGB color space).

ACKNOWLEDGMENTS

This research was partially supported by NSF grant IIS 1527200, by the MSIP, Korea, under the "ICT Consilience Creative Program" and by LDRD grant 16-041 from Brookhaven National Lab.

REFERENCES

[1] N. Ahmed, Z. Zheng, and K. Mueller, "Human computation in visualization: Using purpose driven games for robust evaluation of visualization algorithms," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 12, 2104–2113, Oct. 2012.
[2] L. Bergman, B. Rogowitz, and L. Treinish, "A rule-based tool for assisting colormap selection," in *Proc. IEEE Conf. Vis.*, 1995, pp. 118–125.
[3] B. Bollobás, A. Frieze, and T. Fenner, "An algorithm for finding Hamilton paths and cycles in random graphs," *Combinatorica*, vol. 7, no. 4, pp. 327–341, 1987.
[4] D. Borland and R. Taylor, "Rainbow color map (still) considered harmful," *IEEE Comput. Graph. Appl.*, vol. 27, no. 2, pp. 14–17, Mar. 2007.
[5] M. Bostock, V. Ogievetsky, and J. Heer, "D^3 data-driven documents," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 12, pp. 2301–2309, Dec. 2011.
[6] C. Brewer, "Color use guidelines for data representation," *Proc. Section Statistical Graph.*, pp. 55–60, 1999, [Online]. Available: http://www.colorbrewer.org
[7] S. Cheng and K. Mueller, "Improving the fidelity of contextual data layouts using a generalized barycentric coordinates framework," in *Proc. IEEE Pacific Vis. Symp.*, Apr. 2015, pp. 295–302.
[8] S. Cheng and K. Mueller, "The data context map: Fusing data and attributes into a unified display," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 1, pp. 121–130, Aug. 2016.
[9] J. Chuang, D. Weiskopf, and T. Möller, "Hue preserving color blending," *IEEE Trans. Vis. Comput. Graph.*, vol. 15, no. 6, pp. 1275–1282, Oct. 2009.
[10] C. Xie, W. Chen, X. Huang, Y. Hu, S. Barlowe, and J. Yang, "VAET: A visual analytics approach for e-transactions time-series," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 1743–1752, Nov. 2014.
[11] C. Xie, W. Zhong, and K. Mueller, "A visual analytics approach for categorical joint distribution reconstruction from marginal projections," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 51–60, Aug. 2016.
[12] D. Ebert, C. Morris, P. Rheingans, and T. Yoo, "Designing effective transfer functions for volume rendering from photographic volumes," *IEEE Trans. Vis. Comput. Graph.*, vol. 8, no. 2, pp. 183–197, Apr.-Jun 2002.
[13] H. Fang, S. Walton, E. Delahaye, J. Harris, D. Storchak, and M. Chen, "Categorical colormap optimization with visualization case studies," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 871–880, Jan. 2017.
[14] C. Gramazio, D. Laidlaw, and K. Schloss, "Colorgorical: Creating discriminable and preferable color palettes for information visualization," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 521–530, Jan. 2017.
[15] H. Hagh-Shenas, S. Kim, V. Interrante, and C. Healey, "Weaving versus blending: A quantitative assessment of the information carrying capacities of two alternative methods for conveying multivariate data with color," *IEEE Trans. Vis. Comput. Graph.*, vol. 13, no. 6, pp. 1270–1277, Nov./Dec. 2007.
[16] M. Harrower and C. Brewer, "ColorBrewer.org: An online tool for selecting colour schemes for maps," *The Cartographic J.*, vol. 40, no. 1, pp. 27–37, 2003.
[17] P. Hoffman, G. Grinstein, K. Marx, I. Grosse, and E. Stanley, "DNA visual and analytic data mining," in *Proc. IEEE Vis.*, pp. 437–441, Oct. 1997.
[18] I. Hotz and R. Peikert, "Definition of a multifield," *Scientific Visualization*. London, U.K.: Springer, pp. 105–109, 2014.
[19] I. Jolliffe, "Principal component analysis," *Springer Series in Statistics*, NY, New York, USA: Springer, 2002.
[20] E. Kandogan, "Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions," in *Proc. IEEE Inform. Vis. Late Breaking Topics*, pp. 9–12, 2000.
[21] P. Kerm, "Adaptive kernel density estimation," *Stata J.*, vol. 2, pp. 148–156, 2002.
[22] J. Kruskal and M. Wish, *Multidimensional Scaling*, Newbury Park, CA, USA: Sage Publications, 1977.
[23] J. Lawrence, S. Arietta, M. Kazhdan, D. Lepage, and C. Hagan, "A user-assisted approach to visualizing multidimensional images," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 10, pp. 1487–1498, Oct. 2011.
[24] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.
[25] M. Meyer, A. Barr, H. Lee, and M. Desbrun, "Generalized barycentric coordinates on irregular polygons," *J. Graph. Tools*, vol. 7, no. 1, pp. 13–22, 2002.
[26] W. Mokrzycki and M. Tatol, "Color difference ΔE: A survey," *Mach. Graph. Vis.*, vol. 20, no. 4, pp. 383–411, 2011.
[27] G. Paschos, "Perceptually uniform color spaces for color texture analysis: An empirical evaluation," *IEEE Trans. Image Process.* vol. 10, no. 6, pp. 932–937, Jun. 2001.
[28] P. Ready and P. Wintz, "Information extraction, SNR improvement, and data compression in multispectral imagery," *IEEE Trans. Commun.*, vol. 41, no. 3, pp. 1123–1131, Oct. 1973.
[29] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Sci.*, vol. 290, no. 5500, pp. 2323–2326, 2000.
[30] M. Safdar, G. Cui, Y. Kim, and M. Luo, "Perceptually uniform color space for image signals including high dynamic range and wide gamut," *Optics Express*, vol. 25, pp. 15131–15151, 2017.
[31] R. Stauffer, J. Georg, and D. Markus, "Somewhere over the rainbow: How to make effective use of colors in meteorological visualizations," *Bull. American Meteorological Society*, vol. 96, no. 2, pp. 203–216, 2016.
[32] J. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality Reduction," *Sci.*, vol. 290, pp. 2319–2323, 2000.
[33] L. Wang, J. Giesen, K. T. McDonnell, P. Zolliker, and K. Mueller, "Color design for illustrative visualization," *IEEE Trans. Vis. Comput. Graph.*, vol. 14, no. 6, pp. 1739–1754, Nov./Dec. 2008.
[34] J. Wang and K. Mueller, "The visual causality analyst: An interactive interface for causal reasoning," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 1, pp. 230–239, Aug. 2015.
[35] M. Wijffelaars, R. Vliegen, J. J. van Wijk, and E.-J. van der Linden, "Generating color palettes using intuitive parameters," *Comput. Graph. Forum*, vol. 27, no. 4, pp. 743–750, 2008.
[36] [Online]. Available: http://www.joshuastevens.net/cartography/make-a-bivariate-choropleth-map/, Accessed on: Mar. 1, 2017
[37] [Online]. Available: http://www.naturfotograf.com/, Accessed on: Jan. 15, 2017.
[38] [Online]. Available: http://time.com/7612/americas-mood-map-an-interactive-guide-to-the-united-states-of-attitude/, Accessed on: Jan. 15, 2017.
[39] Landsat, 2008. [Online]. Available: http://landsat.gsfc.nasa.gov, Accessed on: Jan. 15, 2017.
[40] [Online]. Available: https://en.wikipedia.org/wiki/HCL_color_space, Accessed on: Sep. 15, 2017.

Shenghui Cheng is working toward the PhD degree in the Visual Analytics and Imaging Lab, Computer Science Department, Stony Brook University. His research interests include visual analytics, information visualization and scientific visualization with a special focus on high-dimensional and multivariate data. For more information see http://www3.cs.stonybrook.edu/~shecheng/.

Wei Xu (M'2015) received the BS and MS degrees in computer science from Zhejiang University, in 2004 and 2006, and the PhD degree in computer science from Stony Brook University, in 2012. She is currently an assistant scientist in computational science initiative of Brookhaven National Laboratory and a visiting research assistant professor in Computer Science Department of Stony Brook University. Her current research interests include information visualization, visual analytics, and machine learning. She was/is PI and Co-PI of a few DOE grants and has been a reviewer or committee member for many top conferences and journals in the field of medical imaging and visualization. In 2014, she was selected as a women-@energy showcase representing female scientists in STEM fields. She is a member of the IEEE.

Klaus Mueller received the PhD degree in computer science from the Ohio State University. He is currently a professor of computer science with Stony Brook University and a senior adjunct scientist in Brookhaven National Lab. His current research interests include visualization, visual analytics, data science, and medical imaging. He received the US National Science Foundation (NSF) Early CAREER award in 2001, the SUNY Chancellor's Award for Excellence in Scholarship and Creative Activity in 2011, and the IEEE CS Meritorious Service Certificate in 2016. He has authored more than 170 papers which were cited more than 8,000 times. He currently is an associate editor-in-chief at the *IEEE Transactions on Visualization and Computer Graphics* and he is a senior member of the IEEE. For more information, see http://www.cs.sunysb.edu/~mueller.