Frame Experiments on Dye Dataset

# Dye Dataset Description

The dye dataset contains measurements from dye track cruise W0908B, conducted by Oregon State University in the Pacific Ocean off the coast near Newport from August 26 to September 2, 2009. The columns in the dataset are: *Scan Number*, *Julian Days*, *Latitude*, *Longitude*, *Depth*, *Temperature*, *Conductivity*, *Salinity*, *Density*, *Fluorescence #1*, and *Fluorescence #2*. Dye is measured with a photometer at a sub-second frequency.

URL: <http://damp.coas.oregonstate.edu/latmix/w0908b/>

# Histograms

We first describe the types of histograms (data products) that we have computed and then discuss evaluation metrics than can be applied to compare them in terms of quality, computation effort and run-time performance.

## Types of Histograms

Each of the following histograms is computed based on a different technique to segment the incoming stream. To reduce noise, the average dye mass of each segment is calculated and multiplied by the depth of the segment. Finally, this value is summed to the corresponding density bin of the histogram.

### Oceanographer’s Histogram

The oceanographer’s histogram segments the dye dataset into predefined depth slices of a fixed size.

* **Advantage**: The oceanographer’s histogram yields an accurate representation of the dye distribution with respect to density bins.
* **Disadvantage**: Computing the oceanographer’s histogram is expensive as a large number of depth slices have to be processed. Depth slides are created regardless of the dye concentration in a given water region. (Can we go as far as claiming that it is difficult to compute online?)

### Windows-based Approximation

The windows-based approach segments the dye dataset based on its progressing attribute, i.e. the scan number into tuple-based windows of a fixed size.

* **Advantage**: The window-based approach can be computed efficiently.
* **Disadvantage**: Since windows are defined independent of the data, there are no guarantees as to the quality of the histogram in terms of how well it represents the actual dye distribution.

### Frames-based Approximation

We have conducted experiments with three types of delta-frames that each uses a different framing predicate. The first type of delta-frame starts a new frame whenever it has seen a shift in density (of the water). The second type uses the same approach to look for shifts in dye mass. Finally, the third type of delta-frame applies both predicates and starts a frame if either of them is true. In the last case, we have fixed the ratio between the two deltas to 1:2.

* **Advantage**: Frames segment the stream based physical properties such as water density (which relates to temperature and salinity) or the measured dye mass (is that derived from fluorescence? “delta dye concentration sum”?). Does not necessarily partition the stream, i.e. leaves out “uninteresting” regions.
* **Disadvantage**: While delta-frames are generally a good approximation that can be fine-tuned in a flexible manner, they are still an approximation. *Note that boundary-frames could be used to exactly compute the oceanographer’s histogram at the expense of processing lots of frames. There is no such option for windows.*

### Piece-wise Constant Optimal Approximation

Given a predefined number of partitions, we use dynamic programming to compute the optimal approximation of the oceanographer’s histogram in terms of partition boundaries.

* **Advantage**: Good approximation of the oceanographer’s histogram.
* **Disadvantage**: Partitions the stream by design. Cannot be computed online.

## Evaluation Metrics

To evaluate histogram data products, we use the following metrics.

### Earth Mover Distance

Earth mover distance measures how much data has been assigned to the wrong bin of the histogram in terms of how “far” it would have to be moved to be in the correct bin.

### Data Points in Computation

One goal of the approaches using frames or windows is to reduce the number of data points in the computation. This is evidenced by this “metric” that compares the number of data points in the original data sets, to the number of depth slides, and to the number of frames or windows.

### Runtime Performance

Our runtime performance analysis compares the frame approach to the window approach in terms of query execution times. We do not include the oceanographer’s histogram and the piece-wise constant optimal approximation as they cannot be computed in a stream data management system.

# Scatterplots

We first describe the types of scatterplots and, in particular, how summarized approximate plots are generated. We then discuss possible evaluation metric to quantify how well a summarized plot approximates the original unprocessed dataset.

## Types of Plots

The scatterplots over the day dataset compare the depth (x-axis) vs. fluorescence (y-axis) plot of the unprocessed dataset to average depth (x-axis) vs. average fluorescence (y-axis) plots of the summarized datasets.

### Original Dataset

The original dataset is represented as a straightforward xy-scatterplot of depth vs. fluorescence (#1).

### Frame-based Approximation

We use delta-frames on fluorescence (#1) to segment the data stream. For each frame, we report its average depth and its average fluorescence value.

### Window-based Approximation

We use tuple-based windows to segment the stream into a similar number of intervals as the frame-based approximation. For each window, we report its average depth and its average fluorescence (#1) value.

### Sampling-based Approximation

We sample the data stream at fixed intervals to generate a similar number of samplings as frames in the frame-based approximation. For each sampling, we report its depth and fluorescence (#1) value.

## Evaluation Metrics

To compare scatterplots, we use the following metrics.

### Average Euclidian Distance

We scale both the x and y-axis of the unprocessed and approximated dataset to [0..1). Then, we compute the Euclidian distance for each point of the original dataset to the closest data point in the approximated dataset. The error is given by the total sum or distances divided by the total number of points in the unprocessed dataset. The “problem” of this metric is that it is dominated by the high density of values where there is no dye. Since the frame-based approximation changes the density, while the windows-based and sampling-based approximations keep it the same, frames do not do particularly well based on this metric.

### Raster Bitmap Comparison

We scale the datasets as described above. Then, we overlay both datasets with an (*m*, *n*) raster, consisting of *m* columns and *n* rows. Every point in each dataset is mapped to a raster point. The error is then computed by comparing the two bitmaps and counting the points where they disagree, normalized by the size of the bitmap.

### Statistics

Mean? Standard deviation? Variance? Kristin?