Supplementary Material for: “Designing Progressive and Interactive Analytics Processes for High-Dimensional Data Analysis”

Cagatay Turkay, Erdem Kaya, Selim Balcisoy, Helwig Hauser

1 Case Study

1.1 Dataset

The data analyzed during the case study comprises credit card transaction features, customer demographics and financial metrics, and normalized features.

1.1.1 Credit Card Transaction Features

- **cust_id** Customer ID
- **tx_date** Transaction Date
- **tx_time** Transaction Time
- **tx_amount** Transaction Amount
- **merch_type** Merchant Type (MMC Code)
- **merch_id** Anonymized Merchant ID
- **online_tx** Online Transaction Flag (True or False)
- **exp_type** Expenditure Type
- **currency** Currency
- **pos_x** Longitude of the Transaction Location
- **pos_y** Latitude of the Transaction Location

1.1.2 Customer Demographic and Financial Metrics

- **marital_stat** Marital status of the customer making the transaction
- **edu_stat** Educational status of the customer making the transaction
- **job_type** The category of the job of the customer making the transaction
- **income** Monthly income of the customer making the transaction
- **age** Age of the customer making the transaction
- **bank_age** The number years spent as a customer of the bank of the customer making the transaction
- **cc_mean_risk** A score showing how risky the customer is from the bank’s perspective
- **total_num_cc** Total number of the credit cards owned by the customer making the transaction
- **other_total_num_cc** Total number of credit cards issued by other banks and owned by the customer making the transaction
- **bank_cc_max_limit** The maximum credit card limit given to the customer by the bank during the data collection period
- **all_cc_max_limit** The maximum credit card limit given to the customer by all banks during the data collection period
- **total_transfer** The total within-bank wired transfer made by the customer making the transaction during the data collection period.
- **mean_transfer** The mean within-bank wired transfer amount made by the customer making the transaction during the data collection period
- **mean_eft** The total within-bank wired transfer made by the customer making the transaction during the data collection period
- **total_eft** The total between-banks EFT (Electronic Funds Transfer) amount made by the customer making the transaction during the data collection period
- **eft_entropy** A measure of the customer showing his/her diversity in issuing EFT to other banks. Higher value indicates that the customer distributed his/her EFT transfers evenly over the banks to which he/she performed an EFT transfer.
- **max_eft_bank** A categorical value showing the bank to which the customer performed EFT transaction most in terms of the money amount.

- **total_atm_withdrawal** The total amount of the ATM withdrawals made by the customer throughout the data collection period.
- **total_atm_deposit** The total amount of the ATM deposits made by the customer throughout the data collection period.
- **accept_percent** The rate of the total number of acceptance of the customers for various offerings made by the bank to the number of total offerings.
- **resp_score_mean** The average response score of the customers, calculated by the bank, indicating the responsiveness of the customers with respect to any interactions initiated by the bank.
- **resp_score_stddev** The standard deviation of the response scores of the customers calculated by the bank.
- **risk_score_mean**The average risk score of the customers, calculated by the bank.
- **mobile_total_transfer** The total within-bank wired transfer made by the customer over mobile application during the data collection period.
- **mobile_total_eft** The total between-banks EFT amount made by the customer over the mobile application during the data collection period.
- **mean_exp** The average credit card expenditures of the customer over the data collection period.
- **total_exp** The total credit card expenditures of the customer over the data collection period.

1.1.3 Normalized Features

- **tx_amount_n** The linear mapping of the tx_amount feature to the range [0,1].
- **income_n** The linear mapping of the income feature to the range [0,1].
- **age_n** The linear mapping of the age feature to the range [0,1].
- **bank_age_n** The linear mapping of the bank_age feature to the range [0,1].
- **total_num_cc_n** The linear mapping of the total_num_cc feature to the range [0,1].
- **bank_cc_max_limit_n** The linear mapping of the bank_cc_max_limit feature to the range [0,1].
- **all_cc_max_limit_n** The linear mapping of the all_cc_max_limit feature to the range [0,1].
- **total_transfer_n** The linear mapping of the total_transfer feature to the range [0,1].
- **mean_transfer_n** The linear mapping of the mean_transfer feature to the range [0,1].
- **mean_eft_n** The linear mapping of the mean_eft feature to the range [0,1].
- **total_eft_n** The linear mapping of the total_eft feature to the range [0,1].
- **total_atm_withdrawal_n** The linear mapping of the total_atm_withdrawal feature to the range [0,1].
- **total_atm_deposit_n** The linear mapping of the total_atm_deposit feature to the range [0,1].
- **mobile_total_transfer_n** The linear mapping of the mobile_total_transfer feature to the range [0,1].
- **mobile_total_eft_n** The linear mapping of the mobile_total_eft feature to the range [0,1].
- **mean_exp_n** The linear mapping of the mean_exp feature to the range [0,1].
- **total_exp_n** The linear mapping of the total_exp feature to the range [0,1].

1.1.4 Data Pruning

The data has been pruned in order to prevent various kinds of distortions in the visualizations such as stacking of data points to a relatively small area due to extremely high-valued points. The data has been pruned with respect to conditions listed below. For any condition, no more than 0.5% of the data has been pruned. As a result of pruning process, approximately 4% of the data have been dropped.

All the transaction tuples satisfying *all* of the following conditions formed the dataset used in the study.

- income <10000
- total_transfer <1000000
- mean_transfer <100000
- mean_eft <100000
- total_eft <1500000
- total_atm_withdrawal <80000
- total_atm_deposit <250000
- mobile_total_transfer <500000
- mobile_total_eft <800000
- mean_exp <5000

1.2 Tasks

The tasks and example workflow utilizing these tasks are discussed in detail in the paper. Below is the mapping between our high-level tasks and the task abstractions of Yi et al. [5].

Task #	High-level Task	Abstract Actions
Task-1	Automatic Feature-based Subsegment Generation	selection, filtering, clustering, comparison
Task-2	User-defined Subsegment Definition	selection, filtering, clustering, comparison
Task-3	Segment Composition	retrieve, reconfigure
Task-4	Segment Fine-tuning	retrieve, filtering, comparison, selection
Task-5	Composed Segment Description	retrieve, elaborate

Table 1. Mapping between high-level tasks and abstract actions based on Yi et al.'s [5] taxonomy.

1.3 Selectable Features

Selectable features in the difference view of the DimXplorer are listed in Table 1.3 and in "Normalized Features" subsection.

Feature Code	Feature Name	Feature Code	Feature Name
1	tx_amount	21	tx_amount_n
2	income	22	income_n
3	age	23	age_n
4	bank_age	24	bank_age_n
5	cc_mean_risk	25	total_num_cc_n
6	bank_cc_max_limit	26	bank_cc_max_limit_n
7	all_cc_max_limit	27	all_cc_max_limit_n
8	total_transfer	28	total_transfer_n
9	mean_transfer	29	mean_transfer_n
10	mean_eft	30	mean_eft_n
11	total_eft	31	total_eft_n
12	eft_entropy	32	total_atm_withdrawal_n
13	total_atm_withdrawal	33	total_atm_deposit_n
14	total_atm_deposit	34	mobile_total_transfer_n
15	accept_percent	35	mobile_total_eft_n
16	resp_score_mean	36	mean_exp_n
17	resp_score_stddev	37	total_exp_n
18	mobile_total_transfer		
19	mean_exp		
20	total_exp		

Table 2. Selectable features.

1.4 Feature Set Selection

The set of features selected by the analysts during the analysis sessions were recorded and represented in Table 3. Moment shows the time elapsed during the start of the corresponding sub-session until a feature set selection made by the analysts. Duration represents the time elapsed that the analysts spent on working on the selected feature set. The numbers in the feature set selection column are the index numbers of the features listed in Table 1.3. The progression below, recorded as multiples of 25, shows the maximum percentage of the whole data contributed to the calculations as of the moment the analysts renounced the selected feature set. Insights-Questions-Hypotheses column shows the total number of insights, questions, and hypotheses arose during the time period analysts worked with corresponding feature set.

Sub-session	Moment	Duration	Feature Set Selection	Progression Below	Insights-Questions-Hypotheses
2-1	1:23	2:10	1,3,7	75	0-2-0
2-2	0:05	7:11	1,4,8	75	1-0-0
2-2	9:54	2:52	1,2,8,18	25	1-1-0

2-2	17:30	7:10	1,9	75	2-0-0
2-3	00:00	4:45	1,9	100	1-0-1
2-3	5:58	4:19	3,6	50	1-0-1
2-3	10:23	9:32	3,12,17,18	100	2-2-0
2-4	5:14	4:55	1,3	50	0-2-0
2-4	10:09	1:35	1,3,6,12	25	0-0-0
2-4	11:59	5:26	1,3,9,10,15	25	0-0-0
2-4	20:32	9:50	3,5,7	75	2-1-0
3-1	2:02	4:03	22,23,26	25	1-0-0
3-1	27:00	6:05	22,23,27,34,35	50	0-2-0
3-1	33:05	4:15	23,27,34,35	25	1-0-0
3-1	37:20	17:10	23,27,32,33	75	2-0-0
3-2	0:00	4:03	23,27,29,33	25	0-0-0
3-2	4:03	4:37	12,15,16,24,25	50	0-0-1
3-2	12:40	4:29	12,15,16,36	25	0-0-1
3-2	17:09	3:22	12,15,16,37	50	0-0-1
4-1	2:00	6:06	25,31,33	NP	0-0-1
4-1	10:42	8:53	12,15,26,36	NP	0-3-0
4-1	24:34	25:34	22,23,25	NP	0-0-0
4-1	25:34	11:35	22,23,25,26	NP	1-0-0
4-2	19:03	11:54	25,32,33,36,37	NP	1-1-0

Table 3: Feature set selections performed by analysts during the course of the analysis sessions. NP stands for "non-progressive."

1.5 Insights-Questions-Hypotheses

Important take-away points were extracted from the analysis video records and are listed as follows.

Sub-session	Moment	Type of Inference	Content
Session 2-1	1:22	Question	(Q-1) Why do customers with high salary and credit card limits make transactions mainly around Istanbul? Why don't they spend their money in other regions? Why do the customers seeming to be traveling more and having less salary travel more compared to the ones with higher salary?
Session 2-1	7:01	Quote	(Qu-1) Analyst-1: "Let's try some other demographic features as this selection seems like not going to bring new patterns. We can generate so many new hypotheses in a very short time without waiting for the whole calculation to end." Researcher: "Do you think that making different feature set selections could be distracting?" Analyst-1: "No. Instead, visualization is quite engaging as we don't have to wait for even a moment to get some initial results."
Session 2-2	2:46	Insight	(I-1) Customers with higher salary and not using credit card frequently seems to make most of their transactions in Istanbul (might mean that they don't travel a lot).
Session 2-2	5:05	Insight	(I-2) Why do customers with high salary and credit card limits make transactions mainly around Istanbul? Why don't they spend their money in other regions? Why do the customers seeming to be traveling more and having less salary travel more compared to the ones with higher salary?
Session 2-2	5:26	Brief Interview	(Qu-2) Researcher: "How long would it take to reach to these insights if you were to use your own methods?" Analyst-2: "That would require me to form a database query selecting the transactions of the customers with high salary and compare the means of the transactions of those customers with mean of all transactions." Analyst-1: "And the total process could take up 4-5 queries. However, I'm not sure whether we could come up with that question in a very short time. Not to mention the time that would take for queries to run over the transaction database. [Progressive] Visualization seems to help a lot."
Session 2-2	13:35	Quote	(Qu-3) Analyst-1: "It seems like the clustering will not change. ... Almost all of the data has been calculated, let's switch to some other set."
Session 2-2	24:45	Insight	(I-3) Analyst-1: The customers working with other banks seem to be more profitable ones as their financial metrics draws a better picture (i.e. higher transfers, EFT, higher response score, etc.).
Session 2-2	26:37	Insight	(I-4) Customers with one credit card and high credit card limit seem to be have lower risk score.
Session 2-3	00:47	Testing	(Te-1) Hypothesis (insight) I-3 has been rejected. Customers with 2 or 3 credit cards and low credit card limits seem to represent low financial profile.
Session 2-3	03:41	Insight	(I-5) As the age of the customers increase the more likely their expenditures are grouped in the Istanbul area.
Session 2-3	14:12	Question	(Q-2) What is the relationship with the EFT entropy and the total credit number of a customer?
Session 2-3	14:33	Insight	(I-6) The EFT entropy seems to be independent from the total number of credit card.

Session 2-3	15:49	Insight	(I-7) Customers working with other banks seem to be managing their credit cards and their assets via those other banks, not via our bank. (Analyst-1)
Session 2-3	17:01	Question	(Q-3) Why do customers having more than one credit card have lower EFT entropy?
Session 2-4	5:17	Question	(Q-4) Can the transaction amount and the age of the customer be a good discriminator in terms of facilitating good segmentation? Can the transaction amount and the age of the customer be related to money deposit or withdrawal amount?
Session 2-4	11:24	Quote	(Qu-4) Analyst-4 accidentally deleted all the selections including the features, which immediately stopped all ongoing clustering and PCA calculations. Analyst-3: "Sigh, all the computations have gone." Analyst-1 "No worries. Make the selection again, please. They (clusters) should show up soon." However, after this dialog, they ended up starting with a slightly different feature set.
Session 2-4	12:06	Brief Interview	(Qu-5) Researcher: "To what extend do you think this tool (progressive visual analytic system) can effect your current analysis processes?" Analyst-1: "Recently we changed our policy which could be summarized as 'work on old but important hypotheses' to a stance encouraging our analyst teams for the production of new hypotheses. This approach, I believe, can take us one step forward as we will be trying out new alternatives. It seems like this tool is a good fit, at least conceptually, as our analysis is now more data-driven rather than goal-driven. ... (16:03) It is quite a new concept for my team to have almost-real-time response from a clustering calculation. ... (17:09) Typically, given the usual dataset size and analysis goals, we spare more than a day for clustering a model."
Session 2-4	21:18	Insight	(I-8) There seems to be a positive correlation with the age and income of the customer.
Session 2-4	25:47	Question	(Q-5) Do the transactions mostly done close to coastal line of the city form a cluster based on the feature set with age, mean risk and maximum credit card limit?
Session 2-4	29:25	Insight	(I-9) Online transactions seem to be clustered around the west half of the city.
Session 3-1	2:12	Quote	(Qu-6) Just after a new feature set has been selected, Analyst-3: "I think we can start as we have a good view of the clusters."
Session 3-1	3:10	Quote	(Qu-7) The team tried a new set of features and immediately observed a 'good' separation of data points. However, after only 15-20 seconds, the separation dramatically changed. After this, Analyst-2: "Well, I think waiting for a while might be a good thing."
Session 3-1	3:35	Insight	(I-10) As expected, there is a correlation between the credit card limit and income.
Session 3-1	27:29	Question	(Q-6) Are the features age and income together predictor of credit card limit or total transfer amount?
Session 3-1	30:02	Question	(Q-7) How do the clusters differ from each other in terms of total transfer amount?
Session 3-1	40:47	Insight	(I-11) The number of the credit card of the customer and the amount of expenditures made by him/her are not correlated.
Session 3-1	50:47	Observation	(Qu-8) Analyst-3 started to pose his idea and requested one of his fellow analyst to modify a particular visualization parameter. As soon as he noticed that visualization was changing on the screen, he said: "Well, let's wait for a moment to let it settle down." They waited for a couple of seconds so that DimXplorer could processed some more data.
Session 3-1	51:03	Design Informant	(Qu-9) During the course of the analysis, Analyst-3: "I've just seen a high response score for the selected cluster, but it has just gone away." As the clustering algorithm continued to its calculations, the relevant data points have moved to other clusters changing the pattern Analyst-3 previously discovered: "Wouldn't it be nice to have a button that simply pauses the progressive visualization?"
Session 3-2	08:11	Hypothesis	(H-1) The customers with low EFT entropy are also the ones with high response scores.
Session 3-2	08:40	Technical Problem	The analysis session has been paused. The session continued after approximately 4 minutes.
Session 3-2	14:59	Testing	(Te-2) H-1 has been verified. Customers with high response scores were also the ones with low EFT entropy. The reverse was also true, as it was clear on the comparison with PCA and cluster views.
Session 3-2	19:02	Hypothesis	(H-2) The customers with high response score are also the ones with high acceptance rate.

Session 3-2	20:31	Brief Interview	(Qu-10) Researcher: "I have been observing that you are making quick and brief inferences from the data." They were able to interactively try out many different filters and features for the clustering in a short time. Previously, they reported that these kind of changes would actually require a workday. "How do you think these brief findings can contribute to your actual analysis activities?" Analyst-3: "Well, we can further elaborate on the hypotheses that we derived from this tool. For example, we have seen that the customers with high EFT entropy tend to have low response scores. Well, we could verify this in more detail and, maybe, we can also investigate the outliers, the customers with high response scores and high EFT entropy, to find out why they discriminate from the majority." Analyst-2: "For example, if they are not credit card customers, I would try to 'get deep with them' with credit card offers." Analyst-3: "Moreover, if they make more EFT than they actually need, I would recommend them to apply for automatic payment services." Analyst-4: "It could also help us to locate a customer group that we have never been aware of. In that case, we could devise new actions for that new particular group."
Session 3-2	22:07	Brief Interview	(Qu-11) Researcher: "How do you usually go about inferring hypotheses during your daily analysis activities?" Analyst-4: "Experience and limited visualizations usually form a basis for our hypotheses extraction processes." Analyst-3: "Well, first of all, we couldn't get all those results that quickly. As we get some results from our data analytic models, we become aware of even more other options that we feel like we need to try. Well, we don't have that much time." Analyst-2: "And this situation may lead to indolence, particularly when your models take so much time to be calculated." Analyst-3: "And [when working on this tool] we are able to look at the model altogether which is way better than one person working on his own." Analyst-4: "Due to time limitations, we regularly eliminate the minor or insignificant cases, but we are able try them here in a short time." Research: "Aren't there situations that you work on the models as a team?" Analyst-2: "Of course there are. For example, one of us develops a model and we brainstorm on it. However, we usually get the feeling that the analyst working on that model must have spent much time on it and we cannot easily give up and try new models. However, during study I noticed that we can reject our own models as quickly as we generate them. It is easy to reject them on this tool." Analyst-4: "Analysis, especially exploratory one, is usually like a lottery, the more ticket you have, the more likely you win the prize. With this tool, we have many more tickets."
Session 4-1	2:30	Injection	DimXplorer has been configured to work in non-progressive mode. For any calculation, if the calculations are completed in less than 60 seconds, the DimXplorer will respond in at least 60 seconds.
Session 4-1	2:53	Observation	During the delay (60 seconds) that was intentionally inserted to simulate non-progressive analytics, Analyst-4 started to talk about what they would be expecting at the end of the clustering calculation.
Session 4-1	2:58	Hypothesis	(H-3) There is a positive correlation between EFT entropy and total number of credit cards per customer.
Session 4-1	12:17	Observation	At the end of the NP clustering calculation they were satisfied with the result of the clustering; however, they noticed that they forgot to change the feature on the axes of cluster small multiples. Instead of having the result to be recalculated, they decided to continue to the analysis with the old view.
Session 4-1	17:17	Question	(Q-8) How does the campaign acceptance rate of the customers change with high EFT entropy and high number of credit cards?
Session 4-1	17:36	Question	(Q-9) There are customers with high EFT entropy but also falling into one of the subgroups with high and low acceptance rate. What causes this discrimination?
Session 4-1	19:01	Question	(Q-10) Is there a positive correlation between the credit card number and the credit card limit that was granted by our bank?
Session 4-1	22:29	Quote	(Qu-12) Analyst-3: "I think it might also be beneficial to be able to set the step size of the progression. For example, it can be distracting to look at an ever-changing visualization. Can we follow a different approach, say, process a certain fraction of the data at each step of calculation. This way we can have some time to talk about intermediate results."
Session 4-1	32:24	Insight	(I-12) Partial answer has been found for the Q-10. The credit card limit granted by the bank varies depending on the total number of credit cards the customers possess. The 'age' of the customers seems to be a discriminating factor; however, this needs to be verified in a different context.

Session 4-2	0:20	Quote	(Qu-13) Analyst-4: "This new [non-progressive] version of the calculations are better from my point of view. In the progressive version, I was having hard time to mention about the patterns and express my ideas as the clusters were changing so quickly." Analyst-3: "Rather than having the results continuously updated, it might be better to have them with lags, step by step, with a considerable amount of time between the steps. By the way, I suspect that the sampling mechanism gets the data rows randomly. If it did, I would expect a little more uniform distribution across the clusters. However, even if we worked on the same set for a long time, we could barely see such a result." Analyst-4: "Even getting updates at every five minutes is also a great improvement for us as we get a clustering model developed in ten minutes or so with small datasets in commercial [monolithic, non-progressive] tools." Analyst-3: "I believe it would be great to define the amount of the data to be calculated in each step of the calculation. If I know the size of my data, I would divide it, say, into 10 chunks and set progression step accordingly."
Session 4-2	11:34	Injection	DimXplorer has been configured in a way that it will respond to calculations in at least 20 seconds.
Session 4-2	19:18	Question	(Q-11) Is there a correlation between the number of credit cards and expenditure, and money withdrawal from ATMs?
Session 4-2	21:51	Insight	(I-13) The customers using their credit cards more often seem to withdraw money from ATMs less often.

Table 4: Insights derived, questions formed, hypotheses generated by the analysts during the case study.

1.6 Self-reported Advantages of Progressive Approach on the Analysis Process

Subjective evaluations of progressive visualization and analytics were collected from the analysts verbally. Below is the summarized transcript relevant to the progressive aspect of the tool. *Brief Interviews* and notable *Quotes* listed on Table 4 are also additional sources used as subjective measures that were not listed below.

Analyst-3: "The progressive visualization provides instant feedback on the filtering of the data. We usually filter our data with database queries and if we can think better queries during the data retrieval, we are not able to apply it immediately. When the data is big, this becomes overwhelming." (Qu-13)

Analyst-2: "To add more on that, even the effort in finding the proper filter to the data can sometimes be time consuming. With this tool, I felt like I had more ideas, and maybe due to the visual matter we saw, we worked hands-on together and carried more fruitful brainstorming sessions." (Qu-14)

Researcher: "Do you do 'hands-on team work' together as a part of your daily analytic tasks?"

Analyst-3: "Well, it is usually quite impractical as we frequently have to wait for a certain period of time even for simple queries." (Qu-15)

Analyst-1: "What we look at during the analyses is usually the numbers supported with traditional visualizations. At the end of the day, if we happen to doubt our model and say, 'What if we tried this?', well, this instantly become a story for another day. And start all over again!" (Qu-16)

Analyst-2: "Well, from this perspective, we had many chances to try out different ideas today. And the visualization also facilitated emergence of many ideas." (Qu-17)

Analyst-4: "And, we try many clustering variants throughout an analysis period. The tool enabled us to try many options today. Well, I think I've become experienced with this data more quickly than I used to do in my other analysis activities." (Qu-18)

Analyst-3: "During the analysis, we feel obliged to look at to the portion of the data corresponding to the analysis goals and it is not quite often to run into different findings. With the visualization, we could notice also some interesting patterns that were not related to what we were supposed to analyze." (Qu-19)

Researcher: "Did you ever feel interrupted, distracted or disengaged during your analysis sessions today?"

Analyst-1: "No, we were so engaged that even the computer couldn't catch up with us [laughs]." The analyst implied the crashes of our visual analytics tool during the case study. Other analysts shortly answered that they did not feel interrupted. (Qu-20)

2 Numerical Evaluation of Online Computations – Further Details

In the following tests, we use the NPR metric suggested by van der Maaten and Hinton [4] which is computed by finding the set of k nearest neighbors of each point x_i in both of the projections ρ and ρ', which are denoted as G_i and $G_i^{'}$ respectively. We then compute the *NPR* between ρ and ρ' as: $1/n\cdot\sum_{i=1}^{n}(\left\|G_i\cap G_i^{'}\right\|/k)$.

We run 5 comparative tests (using $k=10$) with 5 different datasets which are either artificial or taken from the UCI repository [1]. The datasets are:

1. Artificial dataset $n=4050$, $p=35$ where the dimensions have distinct characteristics, e.g., normal, log-normal, uniform.
2. *US Census Dataset* with $n=2216$, $p=86$.
3. An artificial dataset with $n=1024$, $p=256$ where the dimensions all together encode 16 clusters.
4. *Low Resolution Spectrometer* dataset with $n=532$, $p=97$.
5. *Protein Homology Dataset* with $n=10498$, $p=77$.

Figure 1 displays the *NPR* scores for each of the dataset for 10% sample size increments. We observe that with the datasets that are taken from the UCI repository, even with 10% of the data, we obtained *NPR* scores close to 1, meaning that there is little difference with the projections computed by using only 10% of the data and those computed by an offline algorithm. However, for artificial datasets with structures, i.e., such as the 16 clusters in Test-3, or with dimensions that have very skewed distributions, i.e., such as log-normal distributions in Test-1, the *NPR* scores tend to be lower. This is due to the variability in sampling from these structured dimensions and more advanced sampling schemes may be utilized to overcome these problems [3]. These results show that even with very small portions of the data used, the online algorithm manages to provide approximate results that are reasonably reliable. It has been reported for PCA that in order to obtain reliable results, one has to use at least around 400 items or keep a 10:1 item to dimension ratio, i.e., at least 10 items per dimension [2]. When the amount of data that can be processed by our algorithm in 1 sec. is considered in the above tests (listed in Table 5), we observe that our algorithm manages to process sample sizes that are sufficient to achieve reliable results. The results also show that the number of data items consumed by the algorithm depends on the number of dimensions of the data. As a result, for the 256 dimensional dataset, our sampling method was not able to keep the 10:1 item to dimension ratio in 1 sec. but was able to maintain the 400 items consideration. This implies that for datasets with very large dimension counts, the results of the algorithm may be unstable due to the low number of samples that can be consumed within the temporal limitations.

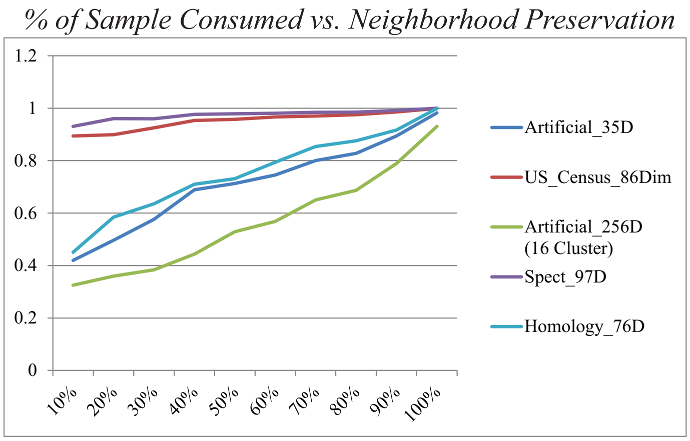

Fig. 1. *Neighborhood preservation ratio* values computed for 10 different sample sizes for 5 datasets.

Moreover, we cross the % of samples in Table 5 with the *NPR* scores in Figure 1. or three to four iterations (tests 3 and 5).

Test ID	# of dimensions	# of processed	% of data
1	35	1700	41%
2	86	1200	55%
3	256	210	20%
4	97	532	100%
5	77	1330	22%

Table 5. Performance evaluation for online PCA computations (the # and % of items processed in 1 second).

References

[1] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[2] J. Osborne and A. Costello. Sample size and subject to item ratio in principal components analysis. *Practical assessment, research & evaluation*, 9(11):8, 2004.

[3] M. Pechenizkiy, S. Puuronen, and A. Tsymbal. The impact of sample reduction on pca-based feature extraction for supervised learning. In *Proceedings of the 2006 ACM symposium on Applied computing*, pages 553–558. ACM, 2006.

[4] L. Van der Maaten and G. Hinton. Visualizing non-metric similarities in multiple maps. *Machine learning*, 87(1):33–55, 2012.

[5] J. Yi, Y. ah Kang, J. Stasko, and J. Jacko. Toward a deeper understanding of the role of interaction in information visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 13(6):1224–1231, 2007.