Paper 065-2009

Effectiveness and Cost of SAS® Compression

Hitesh Sharma, GCE Solutions, San Francisco, CA

Abstract:

Base SAS usage has a broad spectrum. It varies from highly controlled and optimized production environments to non technical analysts trying to solve complex problems with Enterprise Guide drag and drop transformations. The data volumes vary and so does the structure of data. It is therefore helpful to have some guidelines on when to use compression. Thewhite paper identifies and analyzes data attributes which impact effectiveness of SAS compression. The goal is to be able to set some recommendations which would help analysts and developers to make initial judgment on using compression. More than 27000 SAS datasets were tested for compression and results were analyzed with data attributes. This helped in establishing guidelines for choosing compressions in SAS.

Introduction:

SAS compression works on a record (observation) by record basis. Therefore the attributes controlling the effectiveness of compression are primarily the record attributes. What contributes to the record attributes are - number of variables, total observation length, average variable length and most important the data values. Though the data values present an important factor, it is tough to quantify it. Other attributes are very easily quantifiable. Similarly, the effectiveness of SAS compression can be measured in two aspects – the cost and the benefit. The cost of compression is usually CPU bound and can be measured in terms of CPU cost data. The benefit of compression is usually measured in terms of I/O reduction and data size reduction. The time to create and access data may be higher or lower than in an uncompressed scenario so it can be a benefit or a cost in your scenario.

This analysis was done using SAS 9.1.3 SP4 on a Solaris platform. SAS datasets which were not in compressed state were compressed into a different library using a data step. The same action was taken with compression option set to no. The difference in CPU time, real time and size was measured to account for cost and effectiveness of creating compressed SAS datasets. Similarly, the compressed and then uncompressed version of the same dataset was read into _null_ data steps to compare the cost and benefit while reading compressed SAS datasets. The results were analyzed with the data attributes. The sample set was changed to make distribution of observation length and dataset size more uniform for some analysis. Compress = CHAR (RLE) option was used for compression. Compress=Binary (RDC) was not analyzed. The tests were not performed in a truly isolated environment so the interference from other factors cannot be ruled out. However, high sample size was taken to nullify unaccounted factors. Also data steps with compression on and compression off were processed just after each other. Efficiency and cost was measured by difference in the two processing. All cases where SAS chose not to compress were also identified and analyzed accordingly.

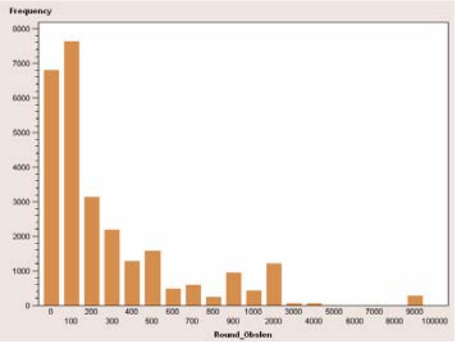
The following legend should be used to read the graph axis. Also one should take care of the varying scales when comparing graphs.

In_ObsLen	Observation Length for the sample SAS dataset
InpSize	Size of the input dataset in Bytes (Input dataset size)
Round_Obslen	Round(in_ObsLen, 50)
Round_Insize	case when (InpSize < 1000000) then Round(COMPARE_ALLS.InpSize,10000) when (Inpsize > 1000000 and inpsize < 1000000000) then Round(COMPARE_ALLS.InpSize,10000000) when (Inpsize > 1000000000) then Round(COMPARE_ALLS.InpSize,10000000000) end
SizeSaved_Per_obs	(Input dataset size – Output dataset size)/Number of Observations in dataset
Comp_ratio	(Input dataset size – Output dataset size)/ Input dataset size
DiffCPU_per_obs_1000	1000*(CPU time consumed by data step with compression - CPU time consumed by data step without compression) / Number of Observations in dataset
DiffReal_per_obs_1000	1000*(Real time consumed by data step with compression - Real time consumed by data step without compression) / Number of Observations in dataset
DiffReal_r_per_obs_1000	1000*(CPU time consumed by data step with compression - CPU time consumed by data step without compression) / Number of Observations in dataset
DiffCPU_r_per_obs_1000	1000*(Real time consumed by data step with compression - Real time consumed by data step without compression) / Number of Observations in dataset

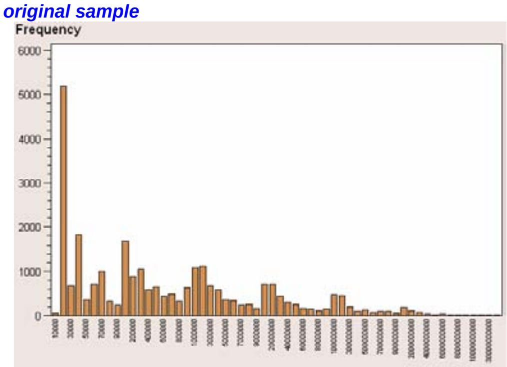
Understanding the sample sas datasets:

The graphs below show the kind of data used for analysis. Most of the original data had small size and short observation length. However, this only impacts the density of dots in a scatter plot. The distribution of the dots in not impacted. The number of datasets in the original sample set was 27000. The distribution however had impact on counts of datasets where benefitor loss was seen. To rule out the factor, a subset with more uniform distribution was created. This modified sample set had 5000 datasets. In the following text it would be mentioned when the modified sample was used.

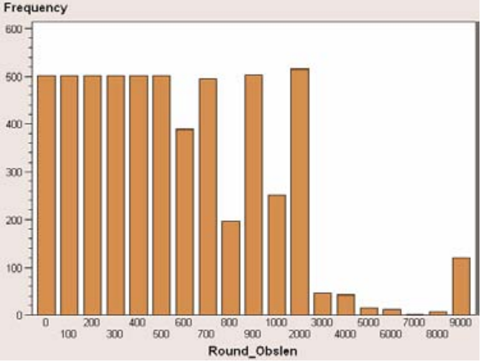
Distribution of input observation len gth in original sample



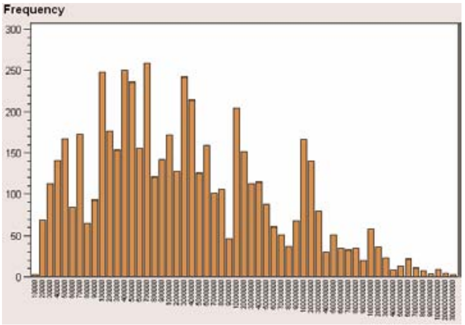
Distribution of input size in original sample



Distribution of input observation length in modified sample



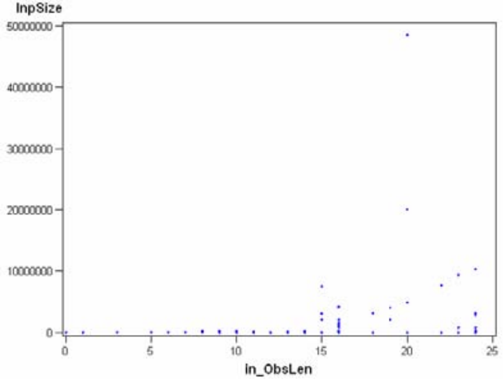
Distribution of input input in modified sample



SAS decided not to compress a SAS dataset:

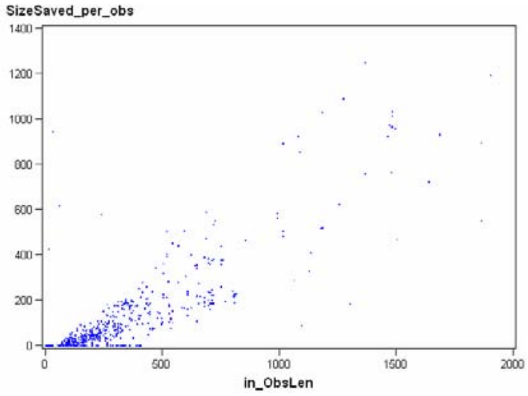
In 5% datasets out of a sample of more than 20000, datasets of observation length less than 25 bytes were not compressed with even the compression option on.

SAS datasets not compressed:

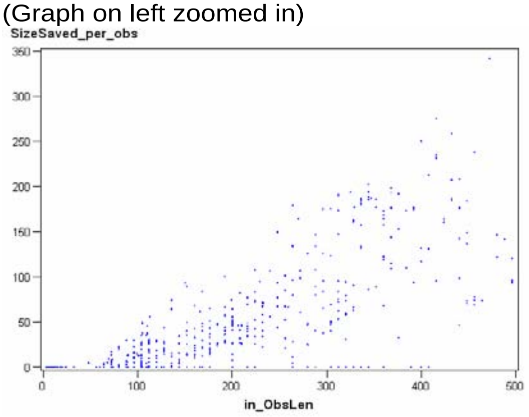


Size Saved per observation vs. Observation length: It has been observed that the size of data saved is very much correlated with the observation length. The SAS datasets that had observation length 100 bytes benefited well from compression. Good reduction in size was seen for most datasets above 200. The sas datasets having observation length smaller than 100 bytes faced expansion rather than reduction in size. Another representation can be seen in percent reduction vs. observation length.

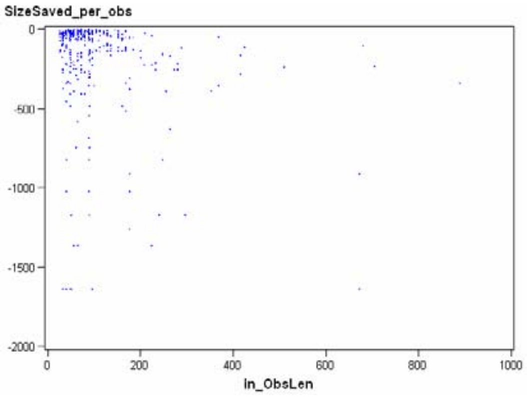
Compression Causing Reduction in Size:



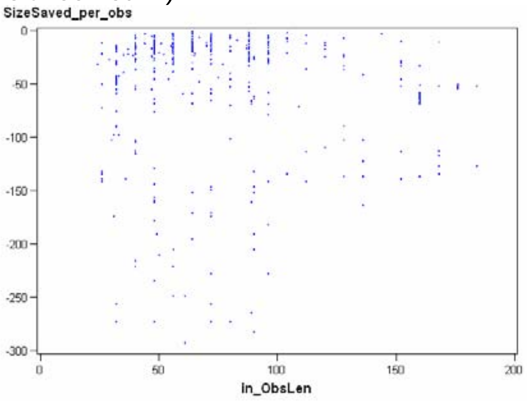
Compression Causing Reduction in Size: (Graph on left zoomed in)



Compression causing Expansion:

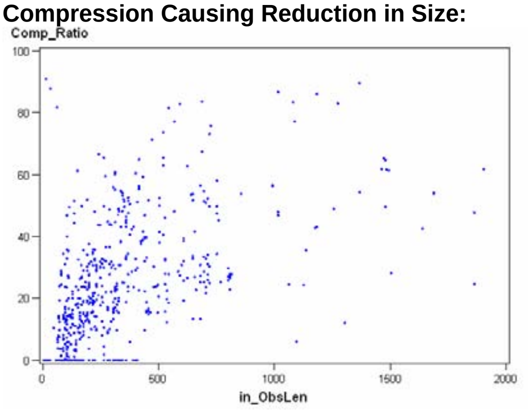


Compression causing Expansion: (Graph on left zoomed in)

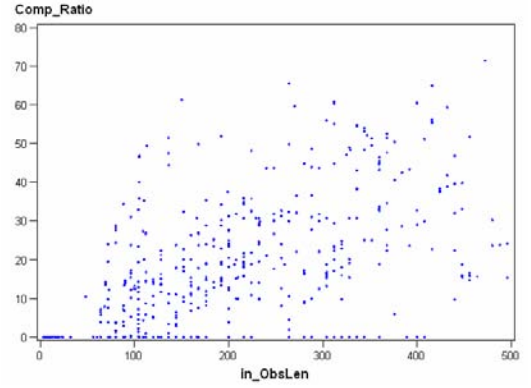


Percent Size Saved vs. Observation length: It has been observed that percent sized saved is also correlated with the observation length. The SAS datasets that had observation length100 bytes benefited well from compression. Good reduction in size was seen for most datasets. The sas datasets having observation length smaller than 100 bytes faced expansionrather than reduction in size. The pattern of size expansion is very consistent below 100 bytes.

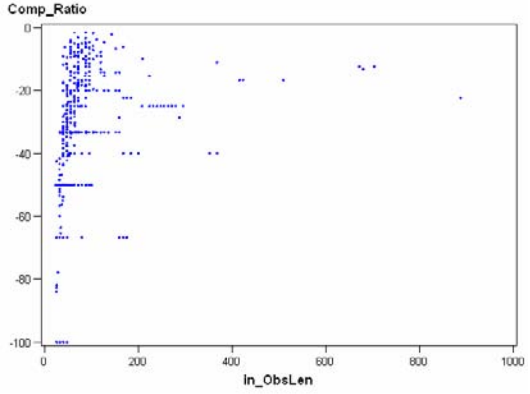
Compression Causing Reduction in Size:



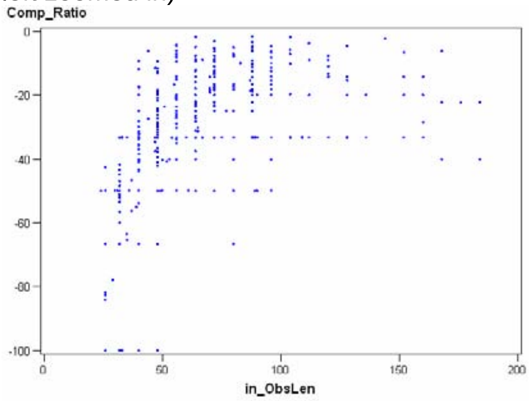
Compression Causing Reduction in Size: (Graph on left zoomed in)



Compression causing Expansion:

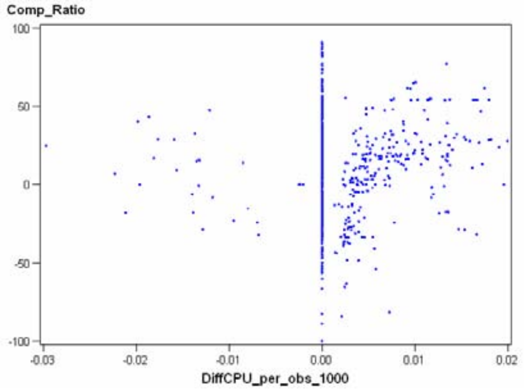


Compression causing Expansion: (Graph on left zoomed in)

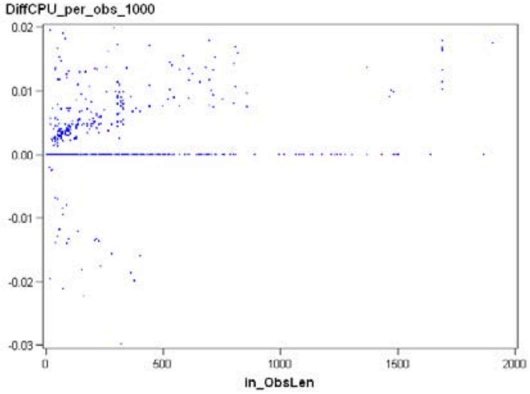


CPU Cost vs. Size Saved: CPU is seen as the biggest cost for compression. The pattern of compression ratio vs. additional CPU cost in creating a compressed dataset shows patterns similar to compression ratio vs. Observation length. This is supported by correlation between additional CPU cost vs. observation length. Zooming in on the graph shows a definitive increase in additional CPU cost with observation length. Few cases showing lesser CPU time with compression may have benefited because of extraneous factors.

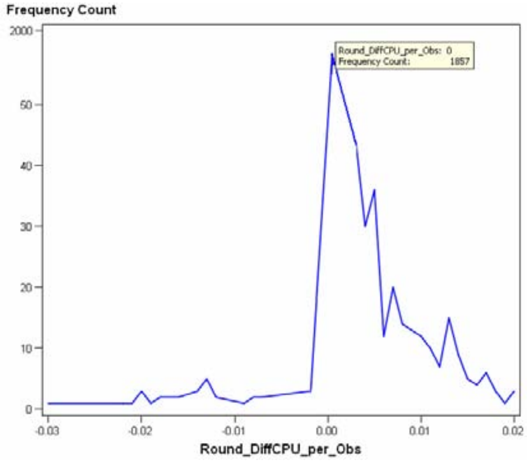
Both compression and expansion:



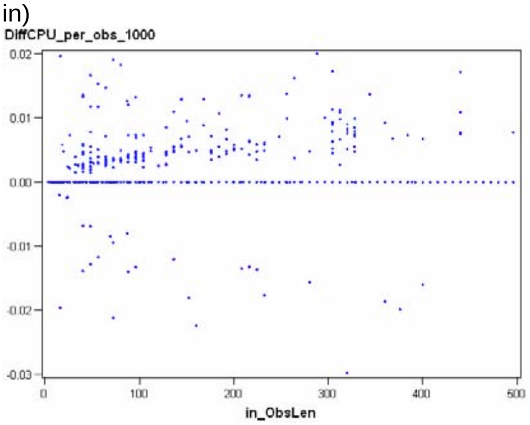
CPU difference per observation by Observation length:



Distribution of CPU difference per Observation:

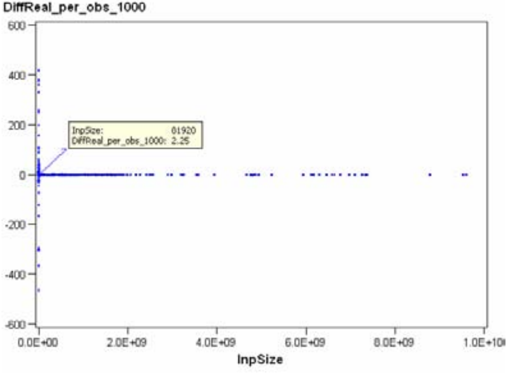


CPU difference per observation by Observation length: (Graph on left zoomed in)

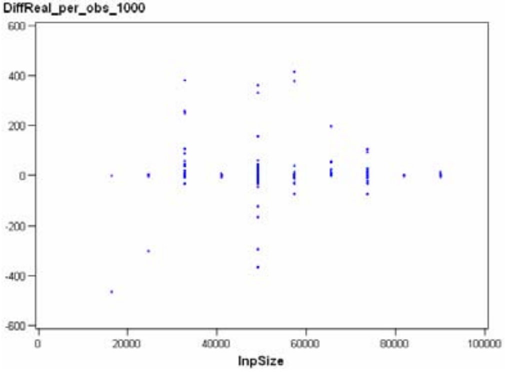


Difference in real Time per observation: Difference in real time per observation when creating dataset was taken as the efficacy parameter to eliminate the effect from number of observations. This analysis was done using the modified sample of 5000. The graphs between difference in real time per observation and observation length shows that some sas datasets with observation length 800 or below faced increase in real time and wider sas datasets did not face increase in real time. However, the graph between difference in real time per observation and compression ratio shows that the increase in real time was associated with negative compression ratio. Therefore it can be concluded that increase in real time happens only when compression fails. CPU processing is not the bottleneck in creation of successfully compressed sas datasets. Also, it should be noticed that the variation for difference in CPU time per observation is localized to datasets sized less than 100 KB.

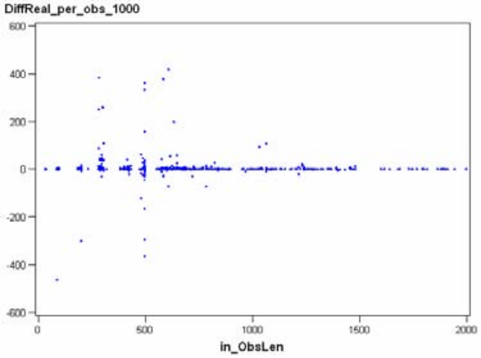
Difference in Real Time per observation vs. input size (zoomed in)**:**



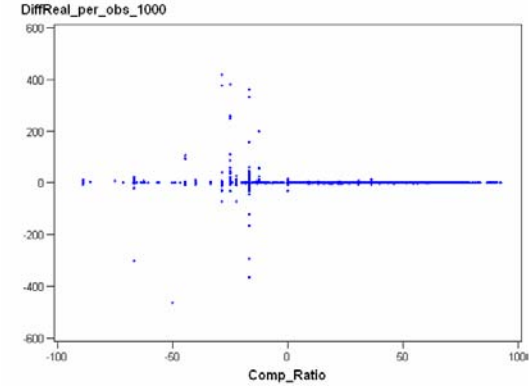
Difference in Real Time per observation vs. input size (zoomed in)**:**



Difference in Real Time per observation vs. Observation Length:

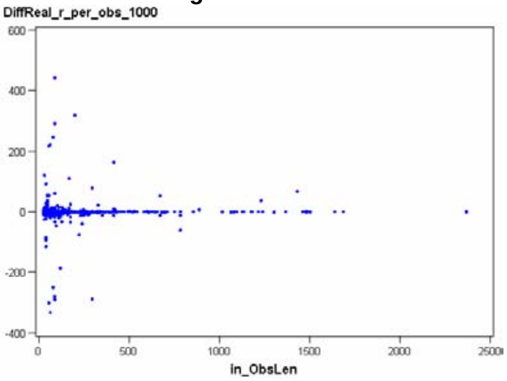


Real Time difference per observation by compression ratio:

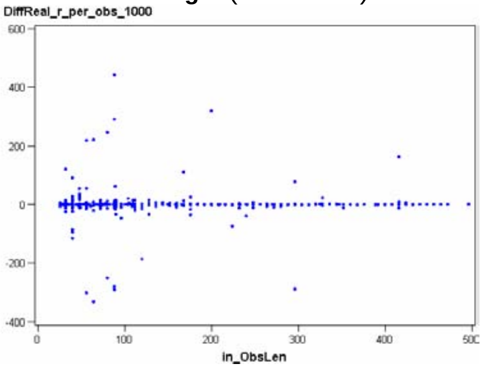


Behavior on reading compressed SAS data: Difference in real time and CPU time was also noted when reading from a compressed and uncompressed sas dataset into a _null_ data step The trends below show that there is a lot of variation in real time for smaller observation lengths. However as the observation length increases the difference goes small. This behavior can be explained by I/O benefit compensating for the CPU cost.

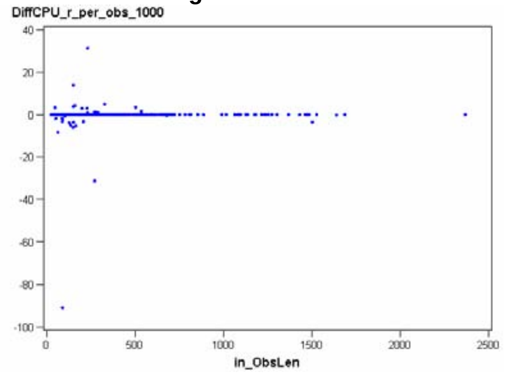
Difference in Real Time per observation vs. Observation Length:



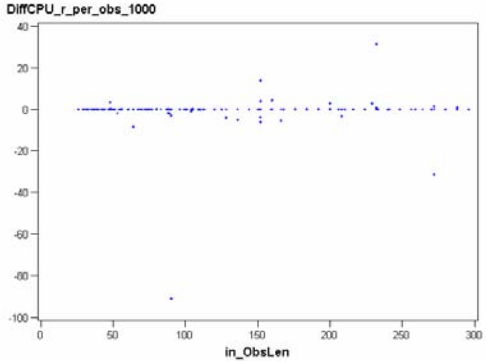
Difference in Real Time per observation vs. Observation length (zoomed in)**:**



Difference in CPU Time per observation vs. Observation Length:



Real Time difference per observation by compression ratio:



Conclusion:

The analysis shows that the benefit from compression is strongly seen for most datasets having observation lengths more than 150 bytes. Datasets between 75-150 bytes observation length should be handled carefully. The cost of compression is usually CPU and one must be careful if your environment is already CPU bound. However, the real time remained largely same for medium and large observation length. One must always review the notes in SAS log to see the impact of compression.

Contact Information:

Name: Hitesh Sharma
Phone: 203-685-8177
Email: Hitesh_sas@yahoo.com