SMART TM Follow-On Study

Final Presentation

Contract: 4000102900/10/NL/ AF

12/06/2012 - ESOC

Agenda

1. Introduction
2. Getting Knowledge about the Telemetry Data
3. Reducing the Amount of Data
4. Identifying the State of the Spacecraft
5. Conclusions & Discussion

1

Introduction

Objectives and Approach

Objectives

- This project is part of a group of studies in the frame of the development of the ITM concept.
  - Improve the operability of the spacecraft by reducing the amount of data and optimising the bandwidth used for transmitting telemetry data from spacecraft
- The main objective of this study has been to get knowledge from the spacecraft telemetry data
- Final objective has been to try to determine the spacecraft state in a OK/NOK status
  - Reducing the amount of the on-board data but without losing valuable information
  - To use this reduced set of data in order to take the decision about the spacecraft state
- Specific objectives:
  - To build an infrastructure for data management (more than 2000 parameters per mission/batch of data)
  - Initial statistical analyses for “seeing what we get .. ”
  - To explore different techniques or approaches analysing their effects on data reduction and possible application on spacecraft state determination
  - To define algorithms for detecting the spacecraft state based on the telemetry data. At least to differentiate between data with anomalies and without anomalies.

Approach

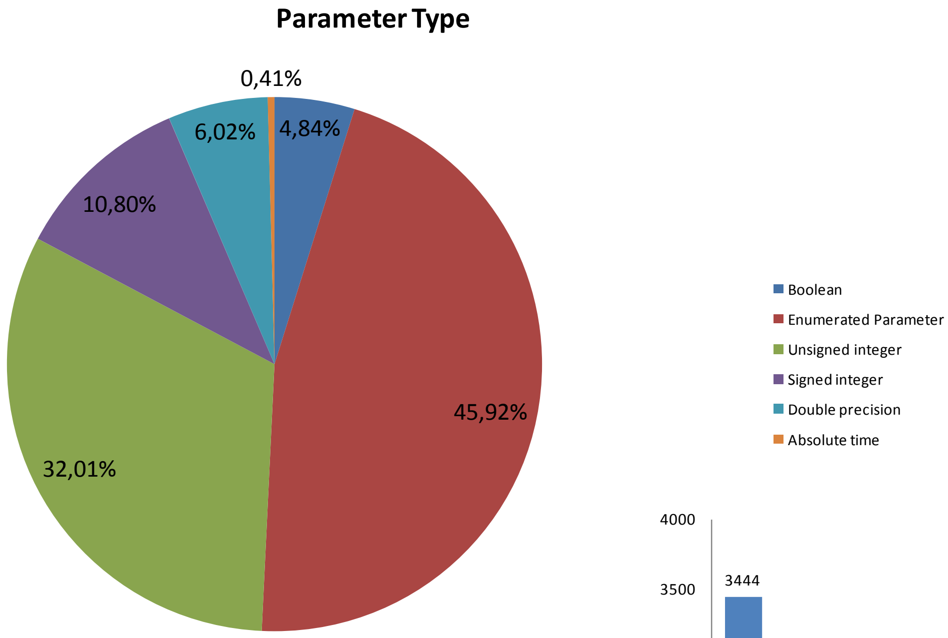
- Data has been analysed as a whole i.e. neither previous classification nor assumptions were previously done
- “Blind” analysis. The analyses did not look for any specific result
  - The approach has been modified “on the way..” depending on the results
- We are not experts on operations, therefore we find relations among data but we don’t know what they mean
  - The study did not look for “physical” explanations but for relationships or characteristics
- Initial runs did not take into account limitation on resources
- Selected missions for the study: Mars Express (MEX) and GOCE

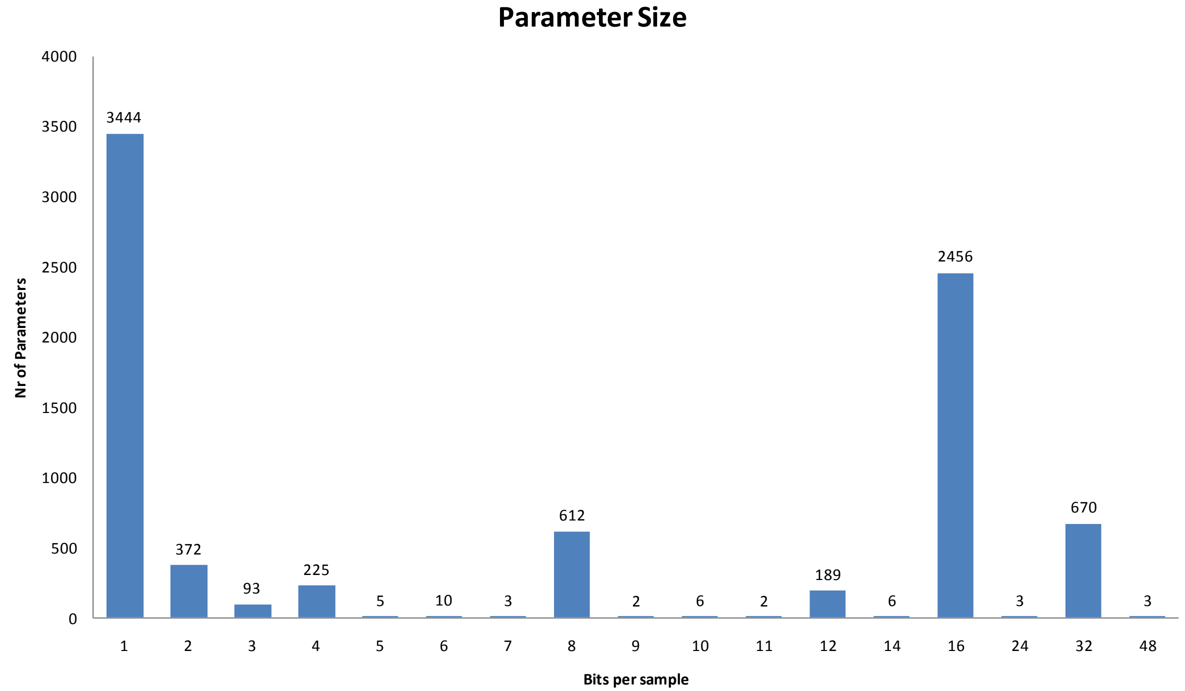
Getting Knowledge about the Telemetry Data

Initial Statistics

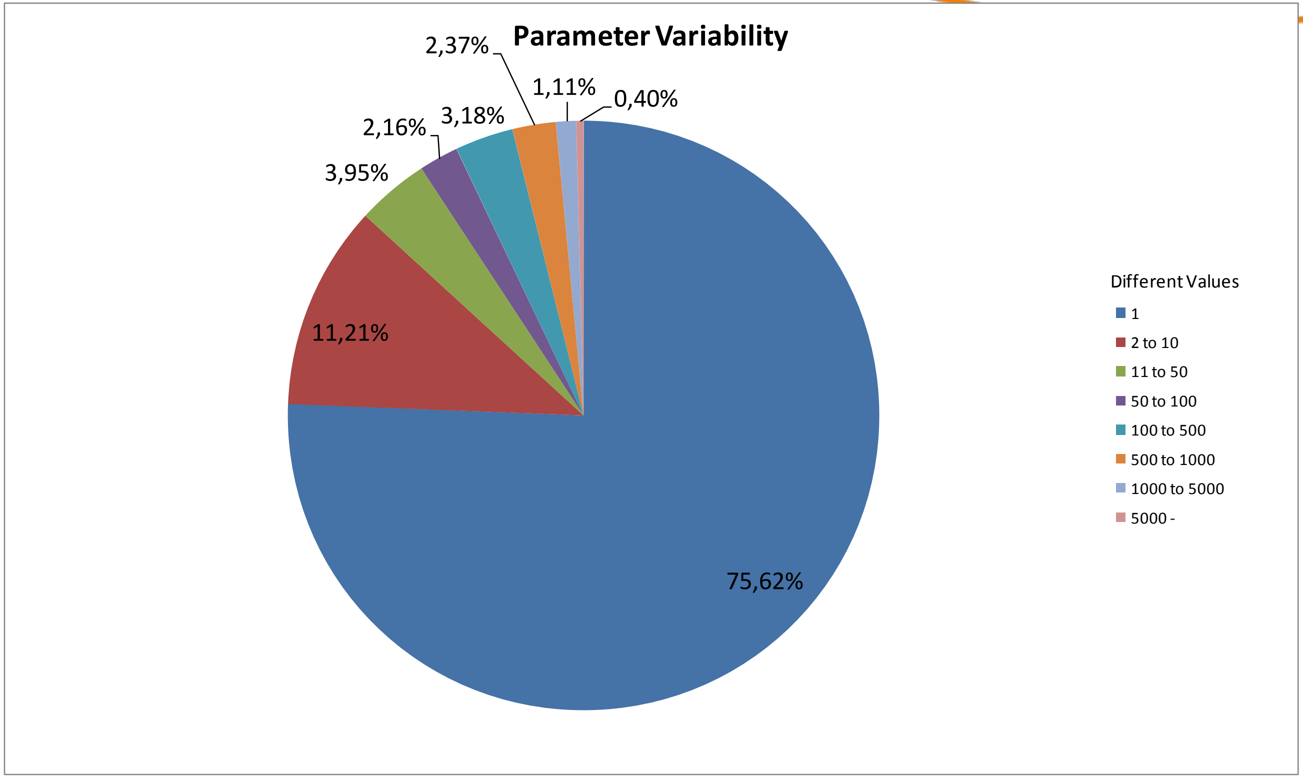
- Telemetry of MEX
  - First batch of 24 of hours of data
  - Second batch of 24 of hours of data
- Telemetry of GOCE
  - Batch of 24 hours
- Initial run was done on first batch of MEX and then, the results were verified with the second batch.
- Same analyses were done for GOCE data.

MEX - Parameters Type



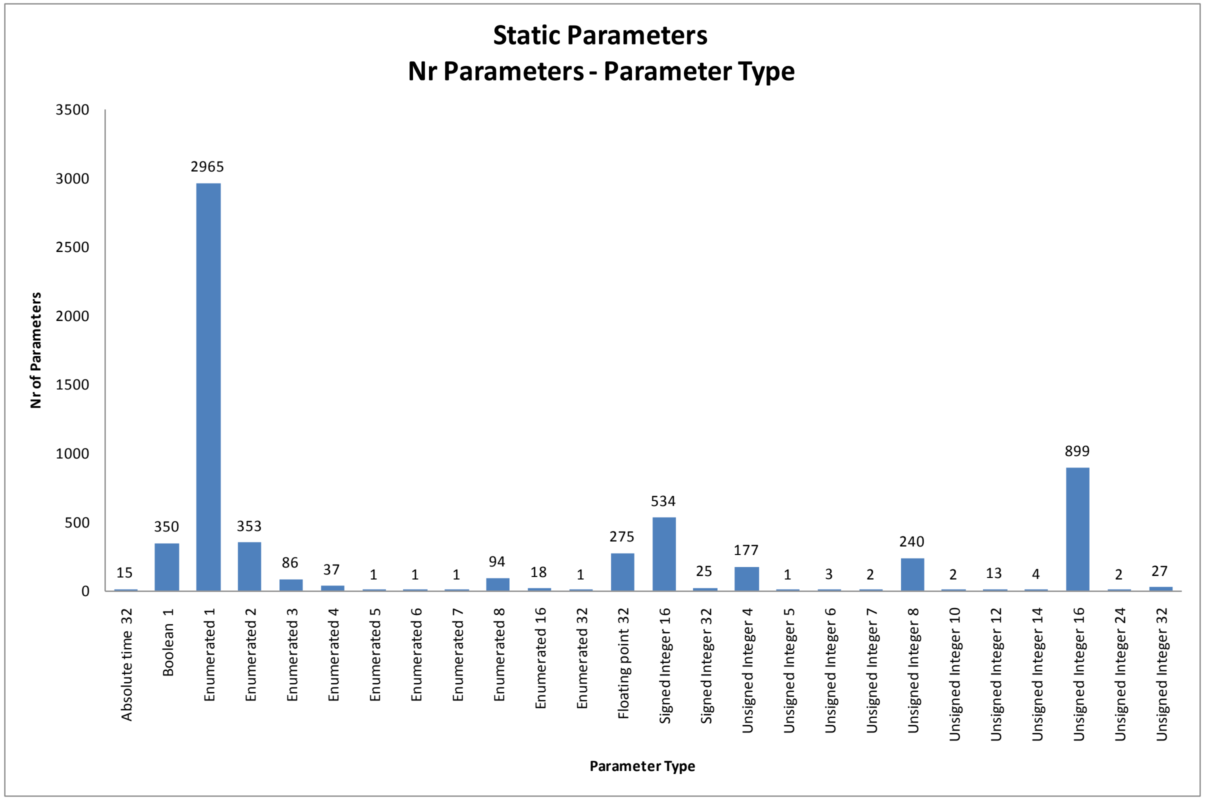


MEX - Variability



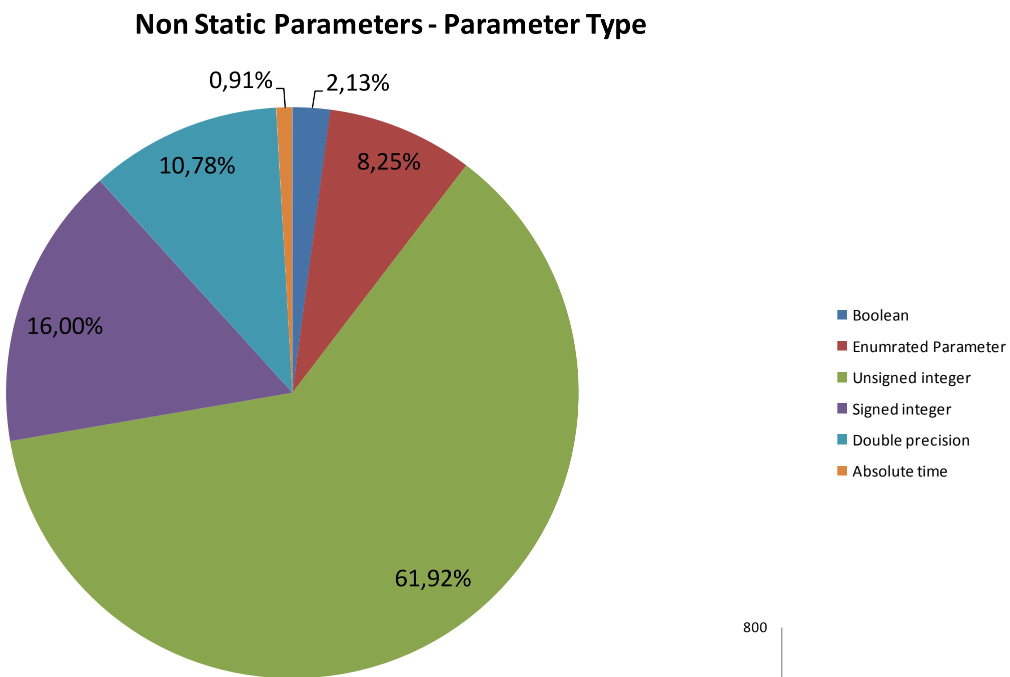
- **First main finding:** 75% of the parameters do not change during the sampled time
  - Parameters have been divided in Static and **Not-Static**

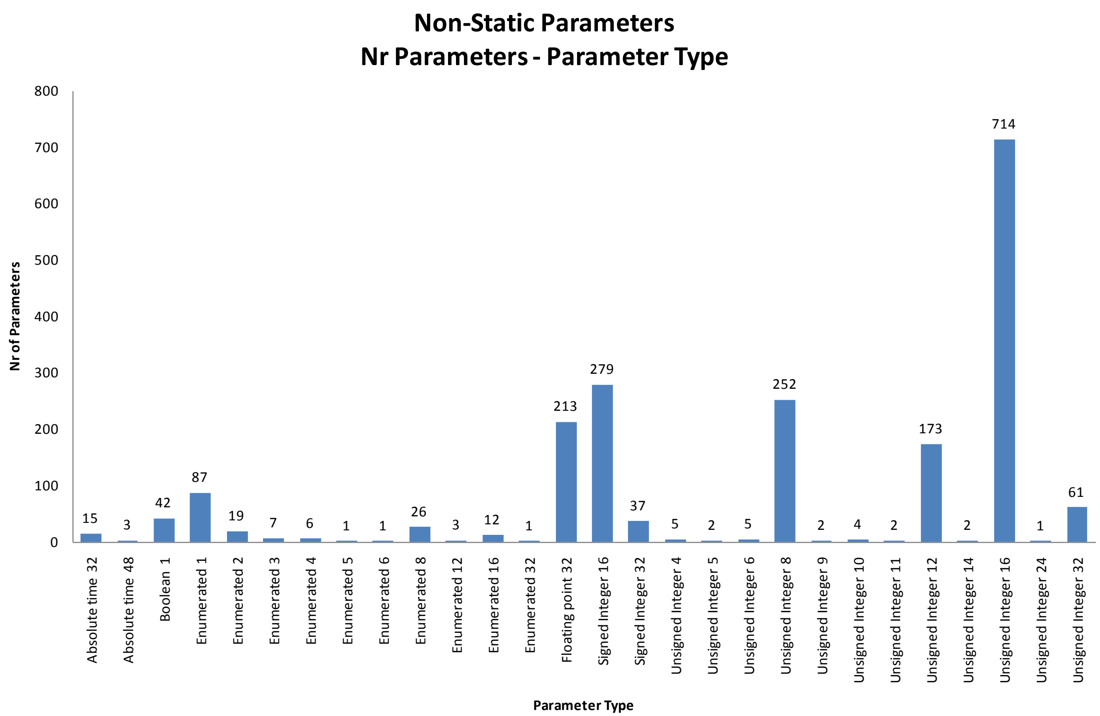
MEX - Static Parameters



- 58% of the static parameters are enumerated
  - Most of them encoded in 1 Bit
- Second amount of static are 16 Bit parameters (23% of static – 18% of the total)

MEX - Non Static Parameters

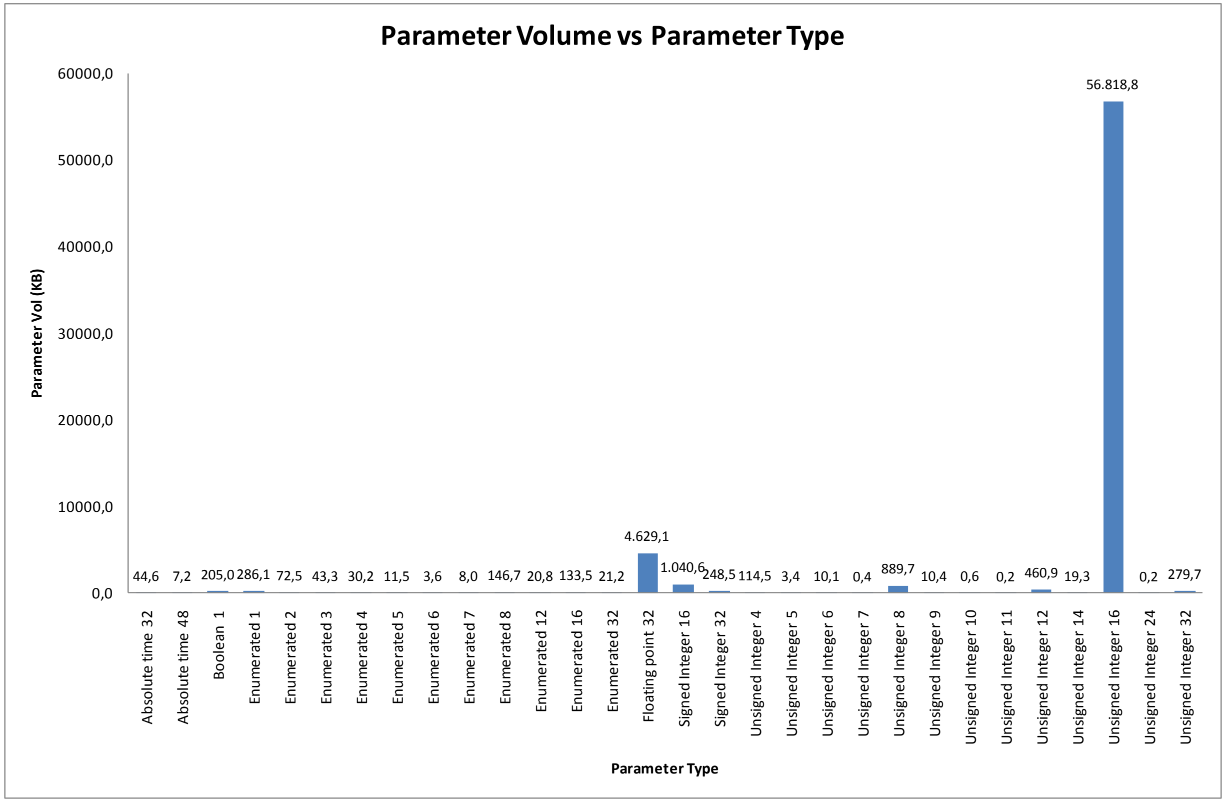




- 52% of non-static parameters are 16 Bit size
  - 71% are Unsigned Integers

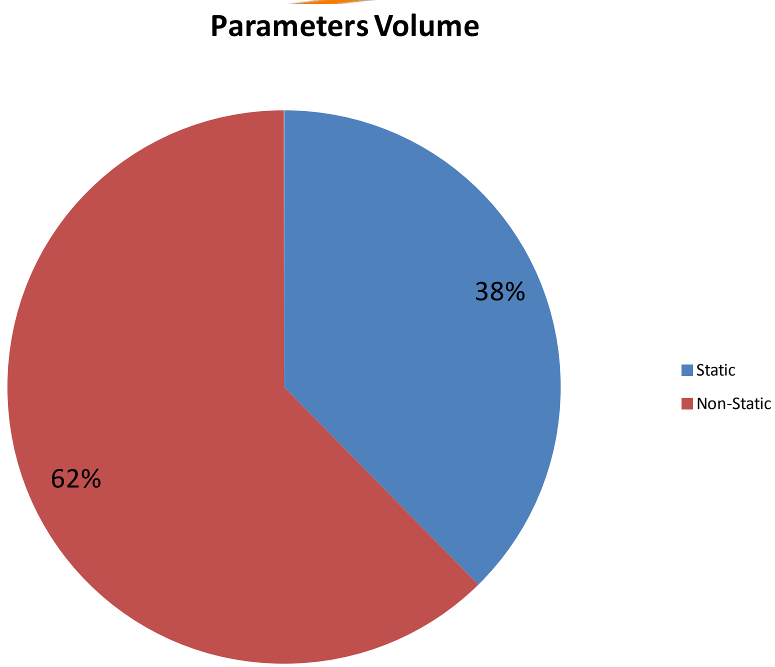
MEX - Volume Analysis

- To estimate the volume of the data produced
  - Assumption: The sample rate is estimated as number of samples per parameter during the analysed period
  - The volume per parameter is calculated as the number of observations by the bits used to encode it

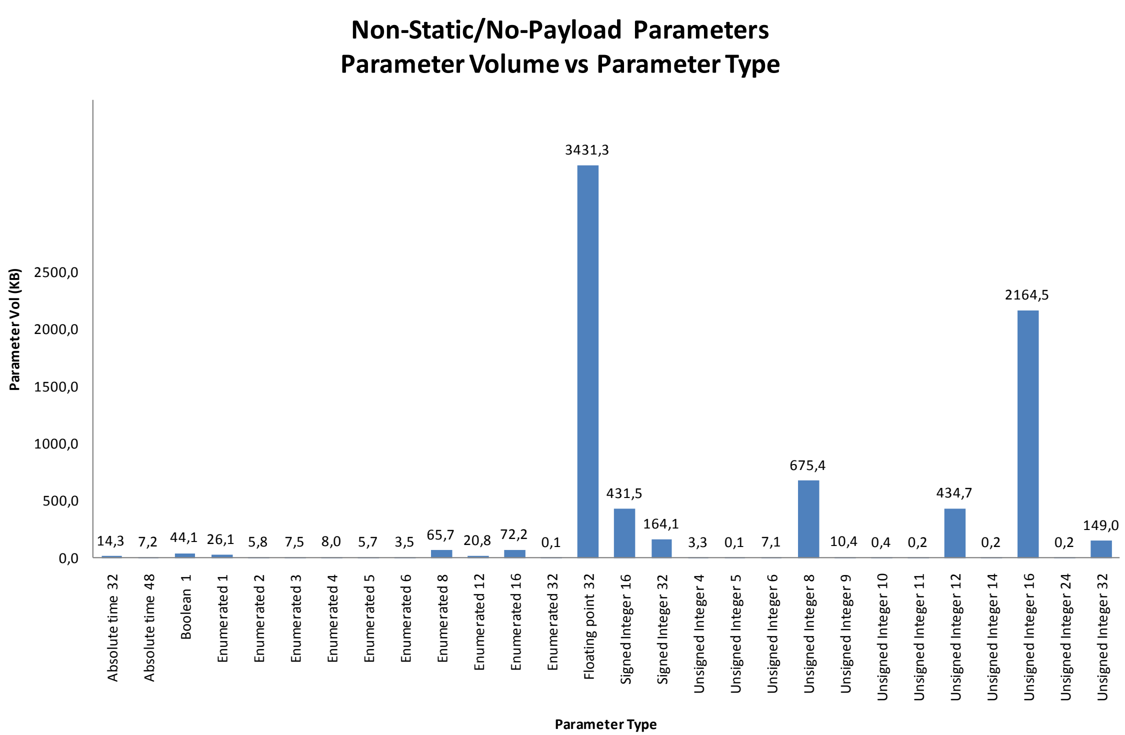


Total estimated volume: 64MB

MEX - Volume Analysis



Removing the payload parameters:
62 % used by non-static
38 % used by static

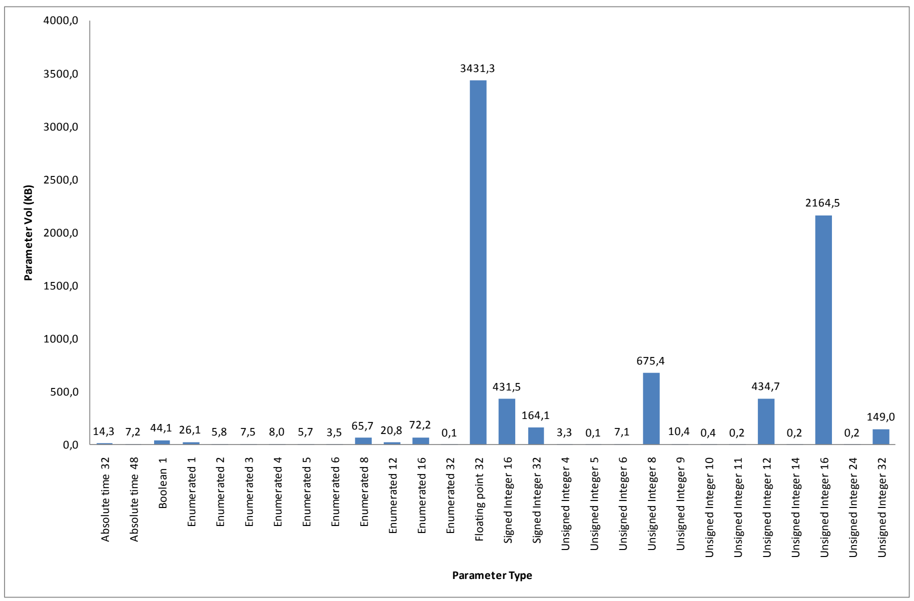


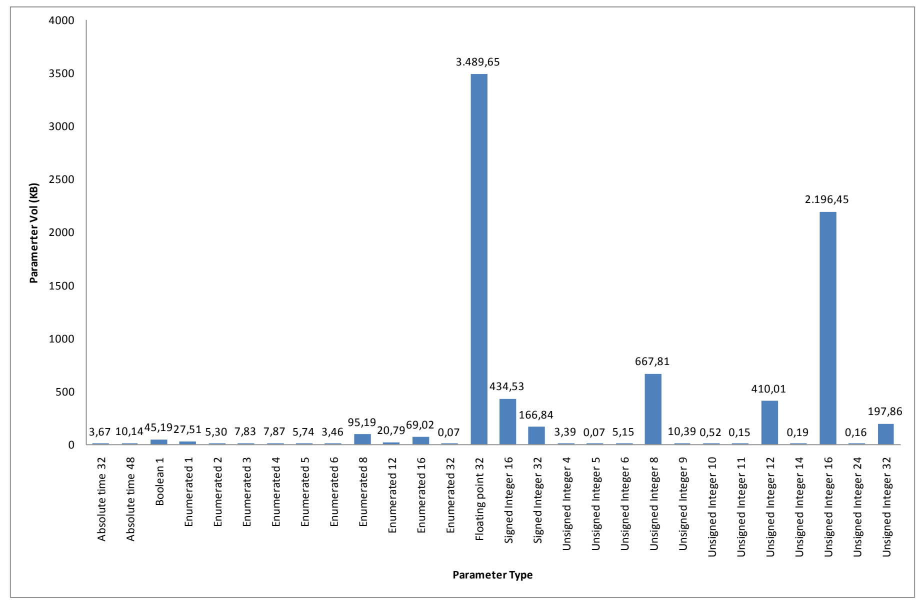
Non-static distribution:

- ❑ 32 Bits -> 44 % of volume
- ❑ 16 Bits -> 22 % of volume

MEX - Confirmation of the Results

- The same analyses were done for the second batch of telemetry data in order to check the results obtained
  - Results are confirmed, almost identical behaviour



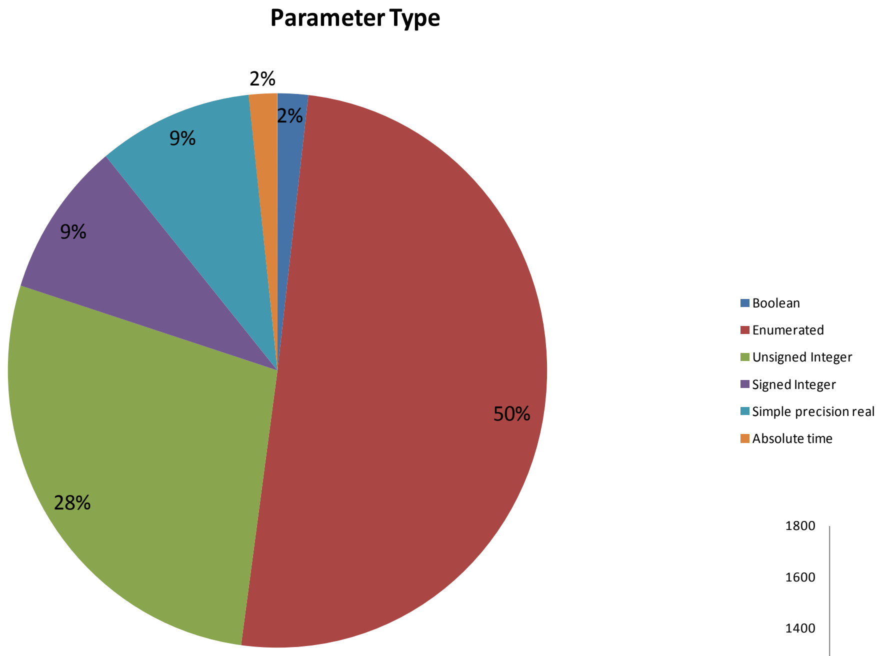


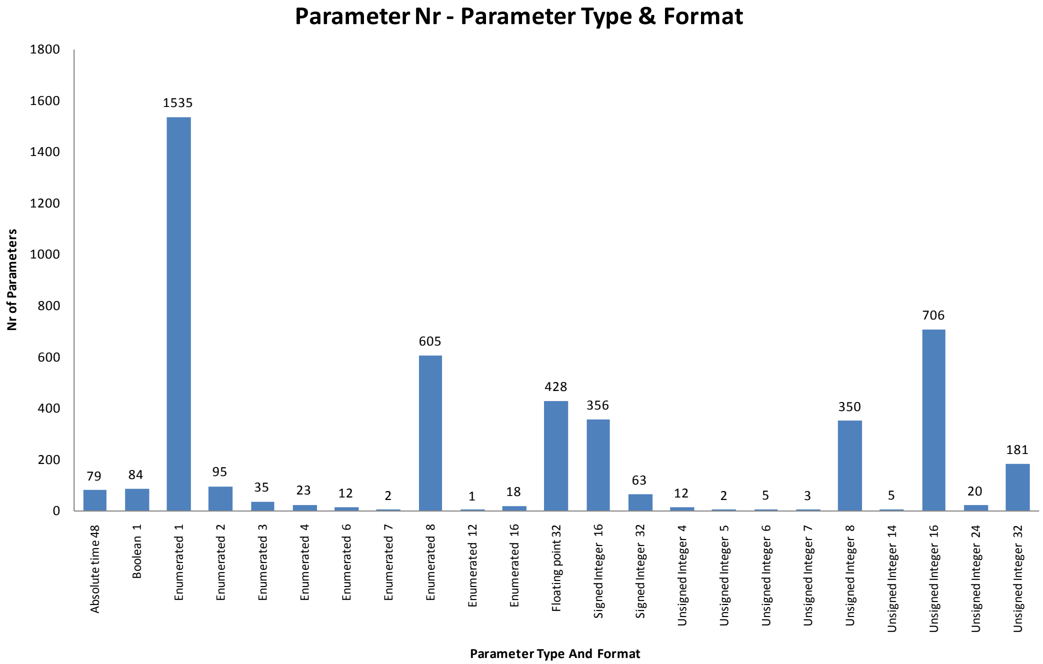
Batch 1

Parameter Type – Non Static

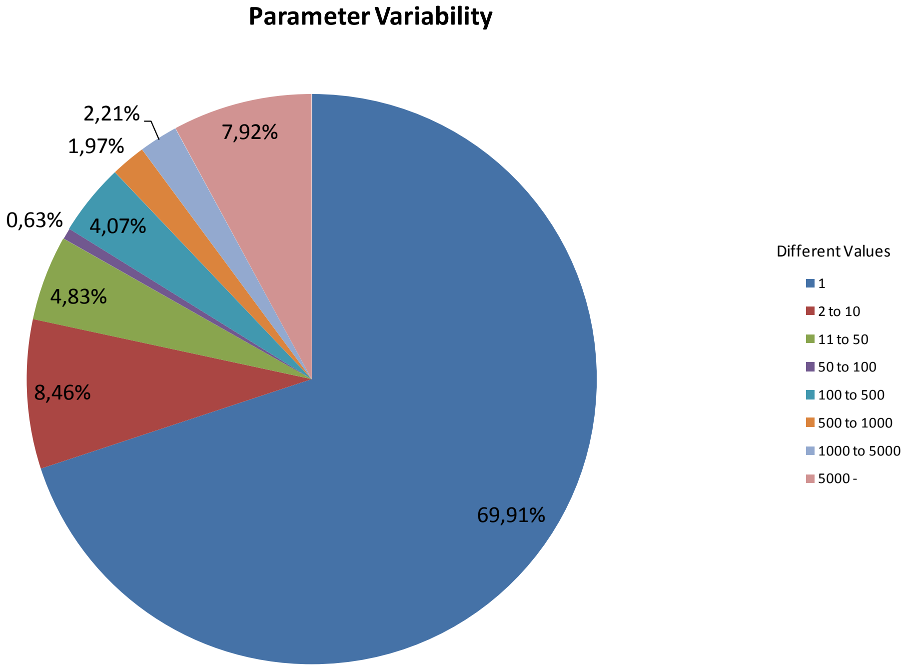
Batch 2

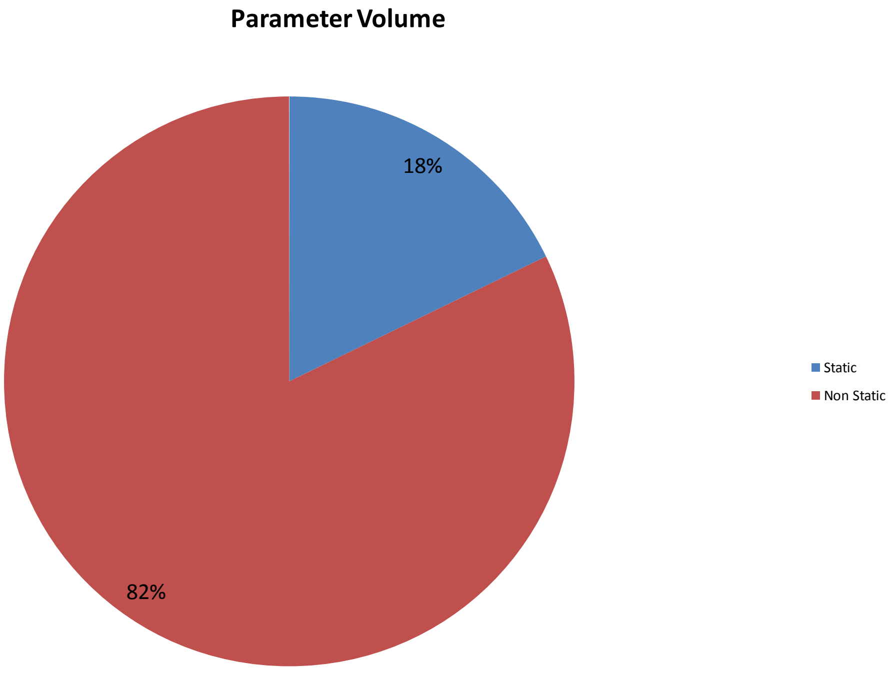
GOCE – Parameters Type





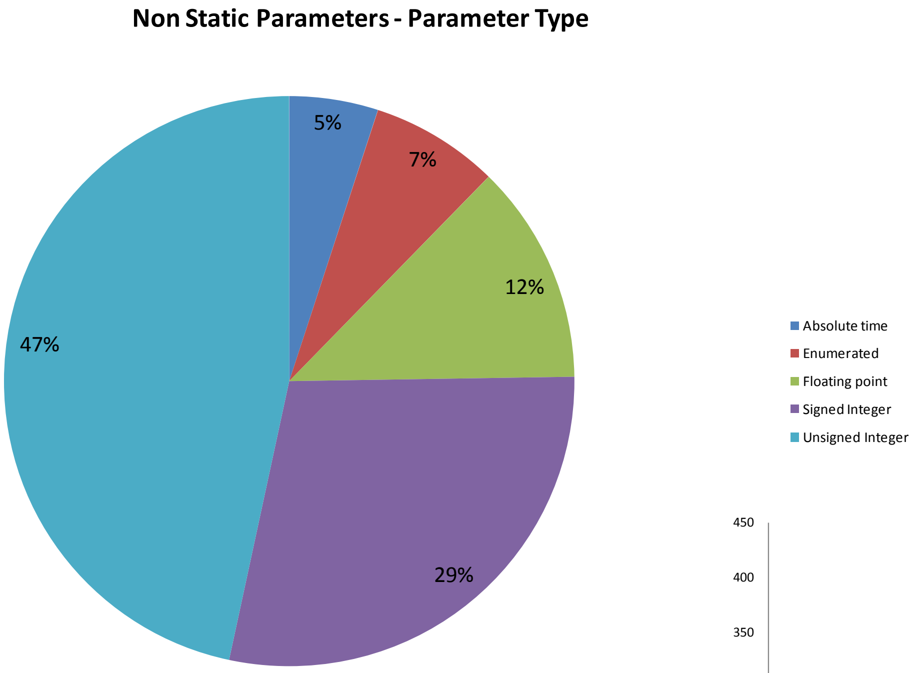
GOCE – Variability / Volume

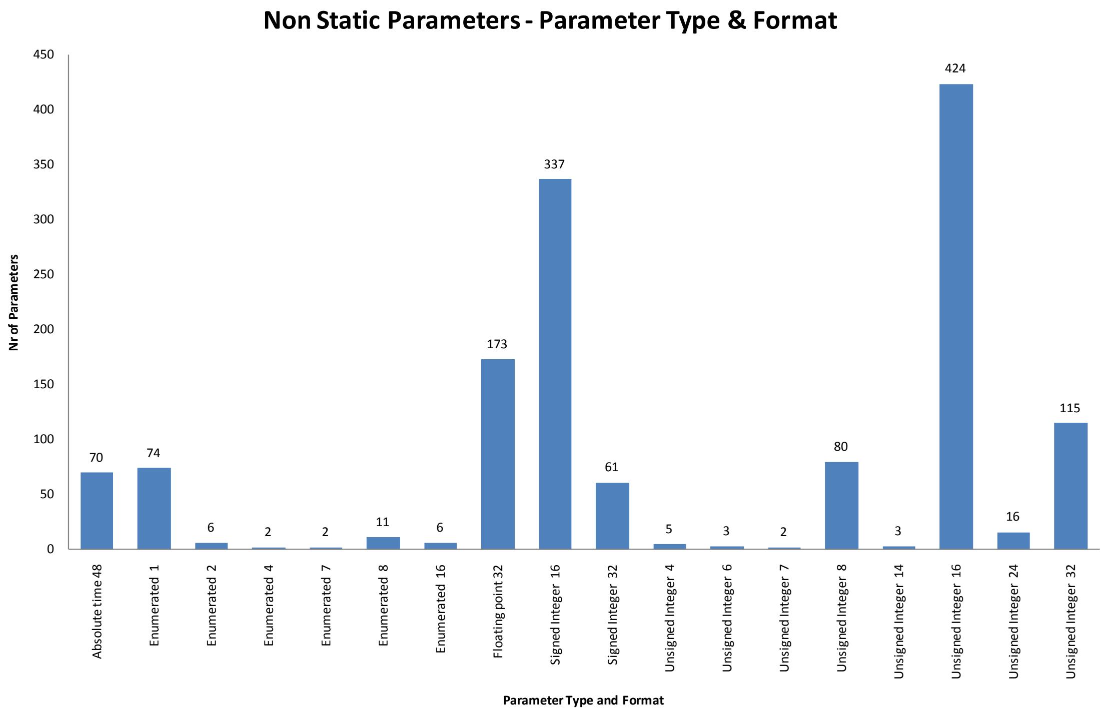




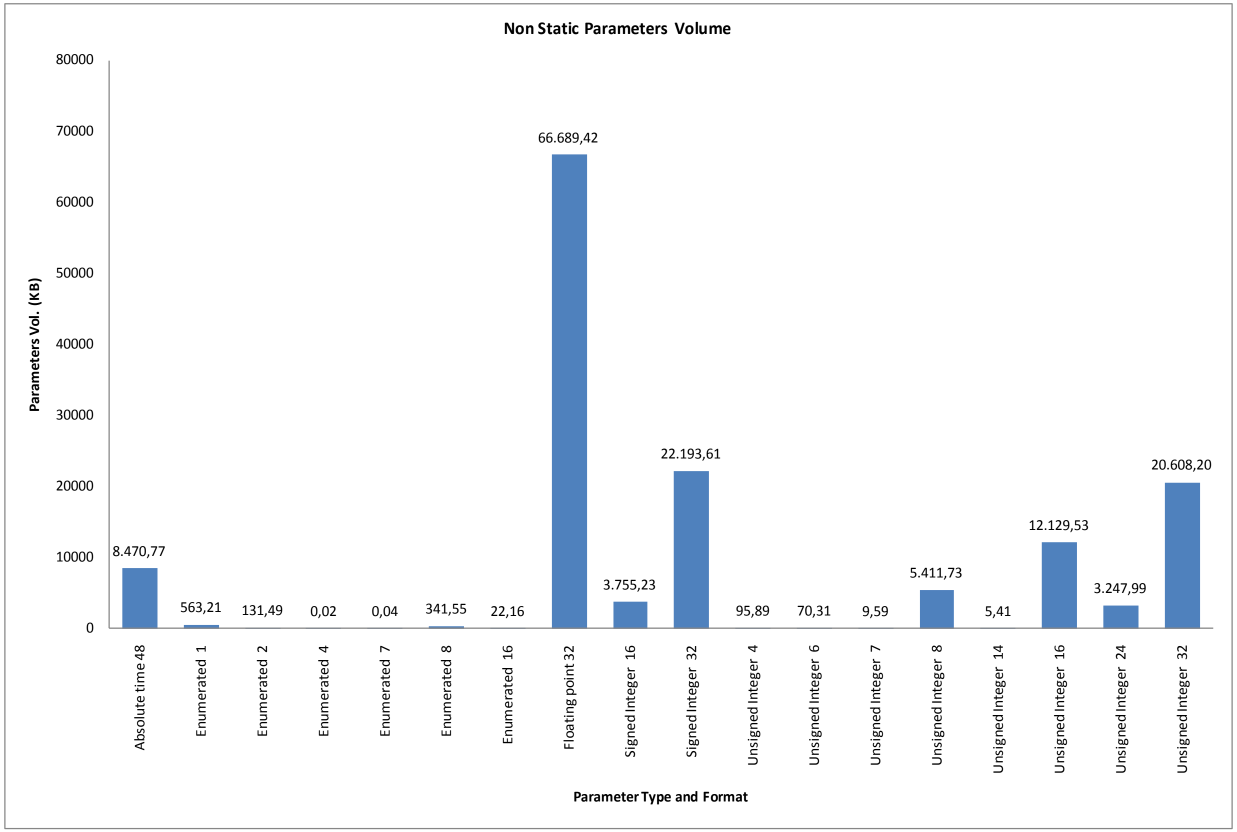
Total parameters Volume = 170.77 MB
Static parameters = 30.39 MB
Non-static parameters = 140.37 MB

GOCE – Non-Static Parameters





GOCE – Non-Static Parameters



Comparison between MEX - GOCE

- The distribution of parameters type is similar in both cases
  - Almost 50% of the parameters are enumerated, most of them are encoded by using only 1 bit
- In both cases there are a big number of parameters that do not change during the sampled time i.e. there are static and non-static parameters
  - In MEX around 75% are static. In GOCE around 70%
- In terms of volume non-static use much more volume
  - MEX: non-static parameters use 63.1% of volume
  - GOCE: non-static parameters use 82% of volume

Reducing the Amount of Data

- Clustering
- Optimal Re-Sampling
- Data Correlation
- Noise Evaluation

Approach

- Several approaches and algorithms were tested on non-static parameters
- Due to complexity of some of them and the large amount of data, a set of representative parameters was defined
  - 10 parameters were selected based on the statistics analysis
  - Criteria for selection: Type of parameter, number of changes and number of different values
- All algorithms were executed initially on these 10 parameters. Then the algorithm was applied to the complete set of data.
  - A initial run was done on the first batch of MEX data.
  - For those techniques that provided promising results the algorithms were re-run with additional data-sets in order to confirm the results (one additional data-set from MEX mission and one data-set from GOCE mission)
  - In some cases the data had to be trimmed due to resources limitation because of the proper nature of the algorithm.
- Criteria for evaluation/comparison used in the TN of Optimal Re-Sampling were used in the same manner
  - Performance (Time)
  - Average error
  - Compression ratio

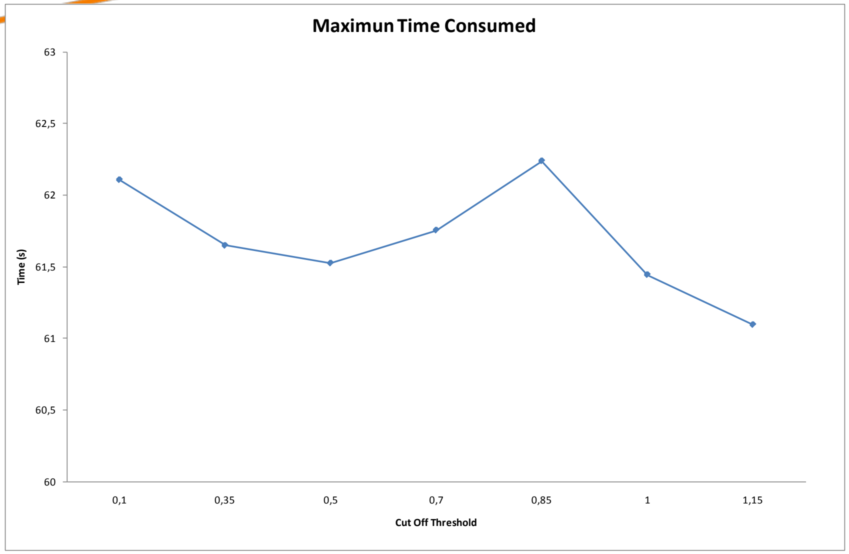
Clustering

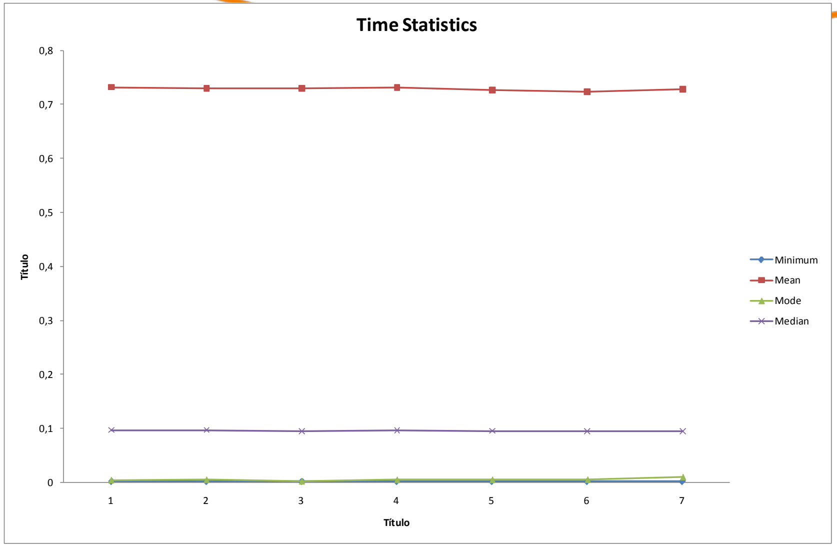
- Cluster Analysis consists of using different algorithms and methods for grouping objects of similar type into respective categories (clusters)
- The technique groups data objects considering only on the information that describes the objects and their relationships.
- The objects should be similar or related to other within the cluster and different from objects in other groups
- There are different types of clustering. Based on bibliography three types of clustering were selected due to their use in applications similar to the SMARTTM:
  - Hierarchical Clustering
  - K-Means Clustering
  - Gaussian Mixture Clustering

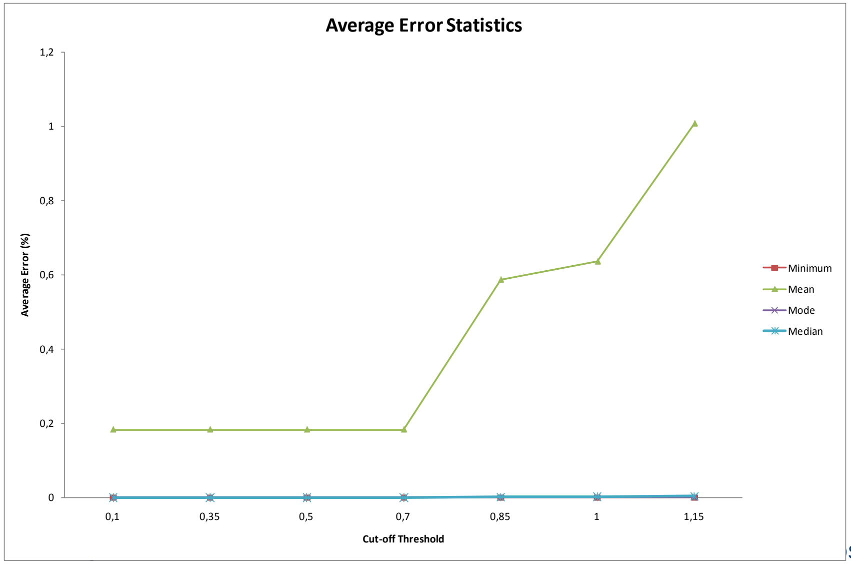
Hierarchical Clustering

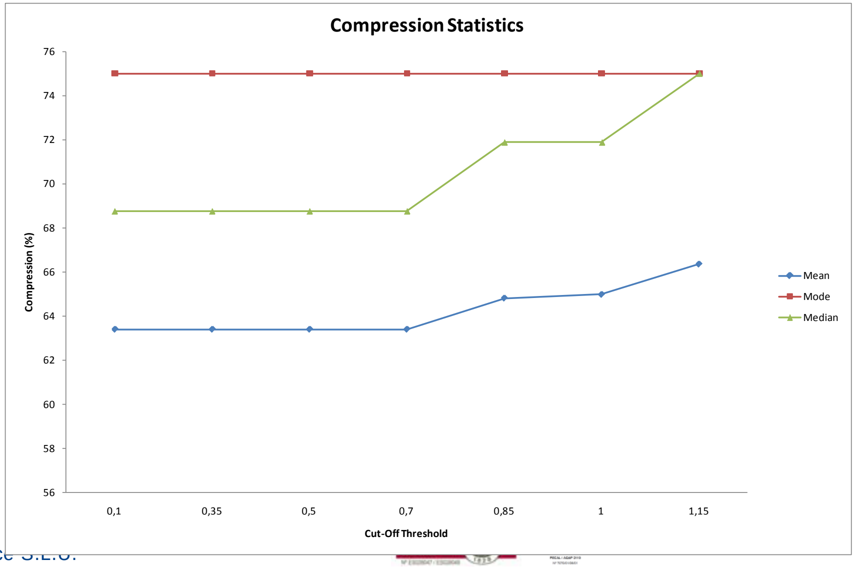
- The hierarchical clustering is based on a multilevel hierarchy; elements are grouped or divided on clusters depending on its distance (Euclidean distance or any other) to each other.
  - There is not necessary a predefined number of clusters
- Implementation using MATLAB
  - Drawback Only 25000 observations as maximum can be processed because it uses a matrix of M*(M-1)/2 where M is the number of samples.
- Statistics done for Cut-off threshold i.e. the threshold used by the algorithm to consider that two samples belongs to the same cluster

Hierarchical Clustering









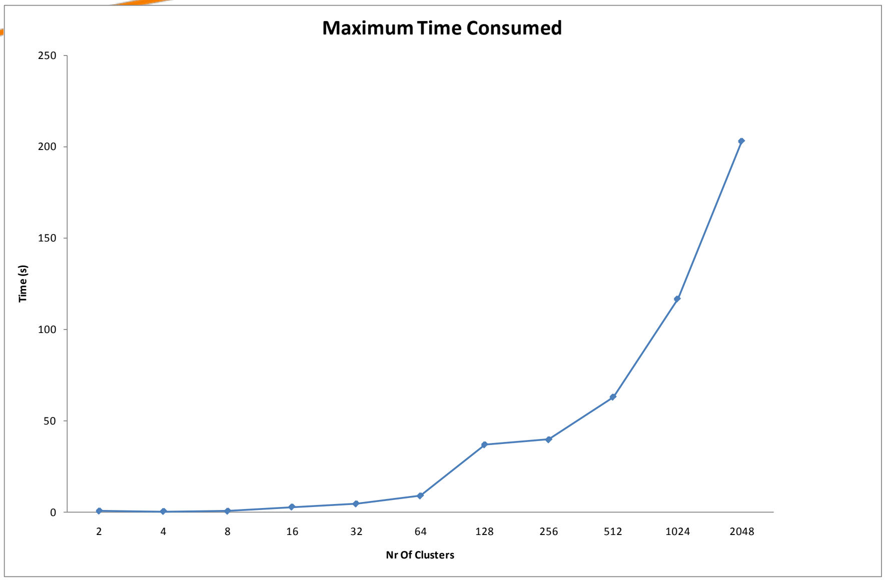
Hierarchical Clustering

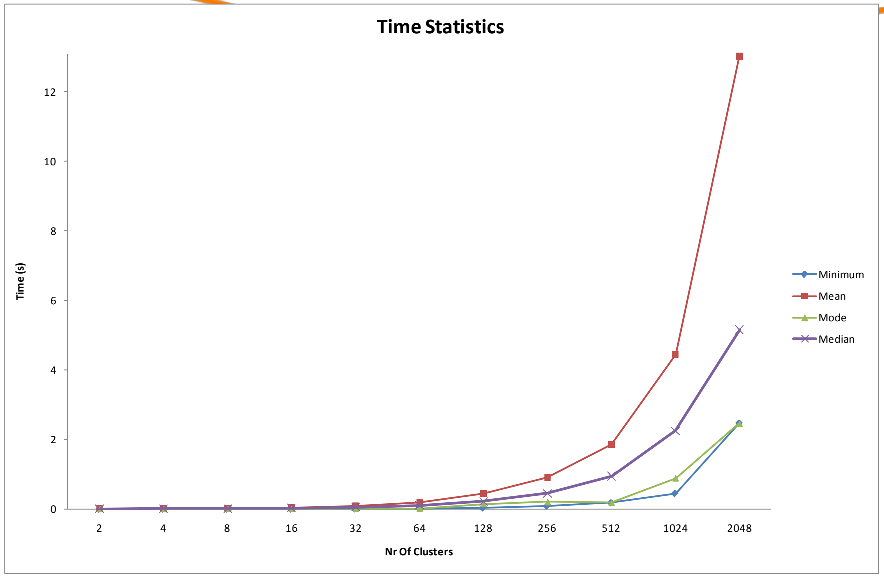
- The cut-off threshold does not affect too much the behaviour of the algorithm
- The output of algorithm is a good representation of the original time series: median, mode and minimum are very close to zero
- The compression statistics show that the algorithm has a good compression ratio considering the small error obtained
- Disadvantages of the hierarchical clustering:
  - Time and memory consumption. Strong related with the number of samples
  - The time increases by the square of the number of samples

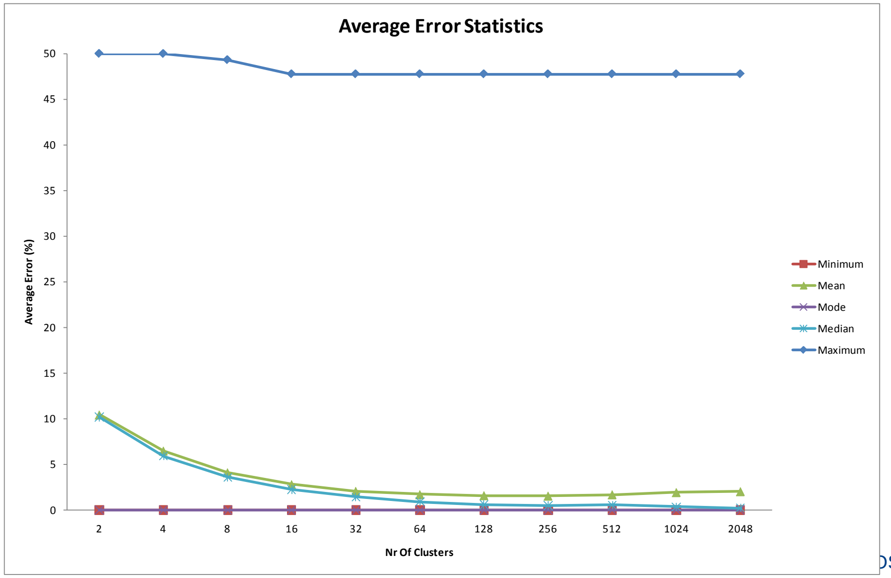
Gaussian Mixture Clustering

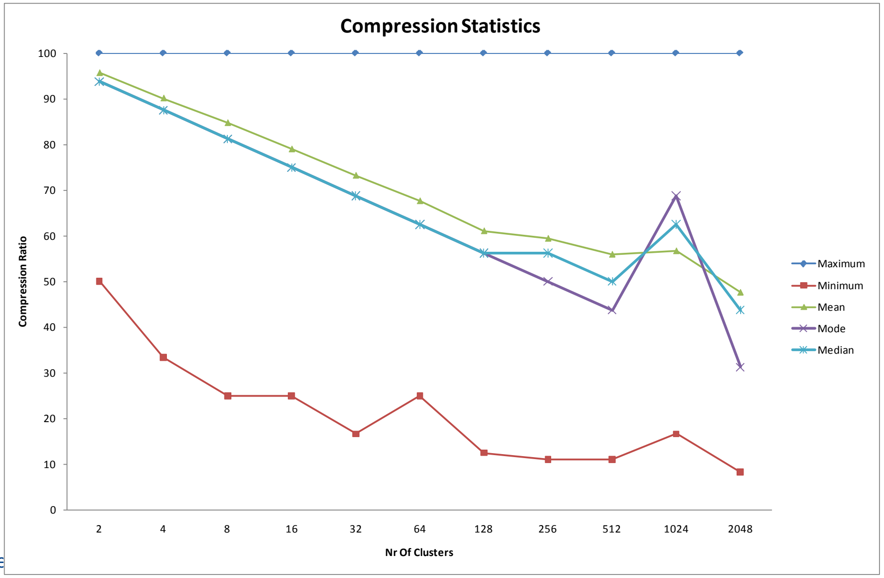
- A Gaussian Mixture Model is a probabilistic model for representing the presence of sub-populations within an overall population by using a weighted sum of Gaussian component densities.
- The models can be used to perform clustering by using an iterative algorithm which assigns posterior probabilities to each component density with respect to each observation.
- Algorithm receives as an input parameter the number of clusters.
  - This value was established in 64 clusters, which can be encoded by using 6 bits.
- Since the performance of this algorithm is strongly dependant on the number of clusters, results of these tests cannot be taken as the general behaviour, but as a first test using an arbitrary number of clusters.

Gaussian Mixture Clustering









Gaussian Mixture Clustering

- Even using a high number of bits to encode the parameters, this clustering algorithm produces a high average error, over 10% for about 20 parameters.
- Maximum time consumption increases exponentially with the number of clusters
- The maximum average error is almost constant, which means that even for a high number of clusters at least one parameter has a poor performance and all the information is lost
- The maximum compression is 100% regardless of the number of parameters. This value actually shows a failure of the algorithm that in some cases is not able to create the specified number of clusters
  - This algorithm is not suitable for telemetry parameters.

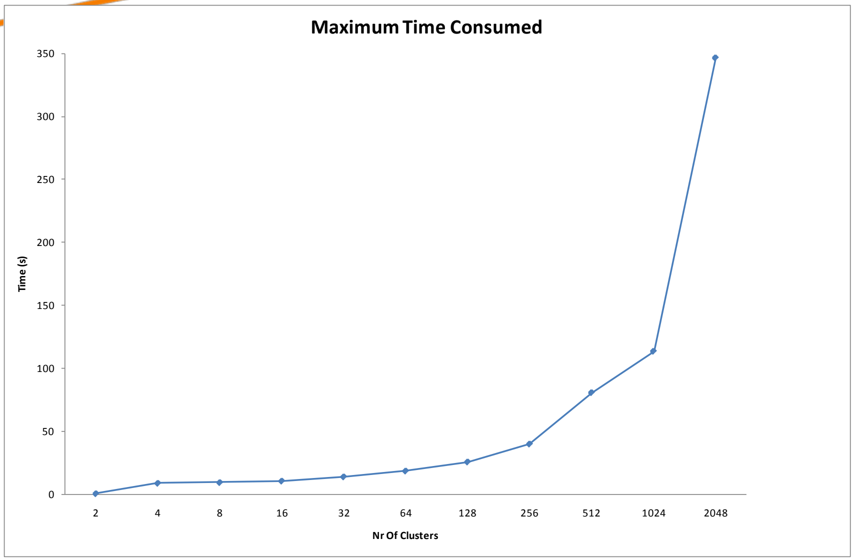
K-Means Clustering

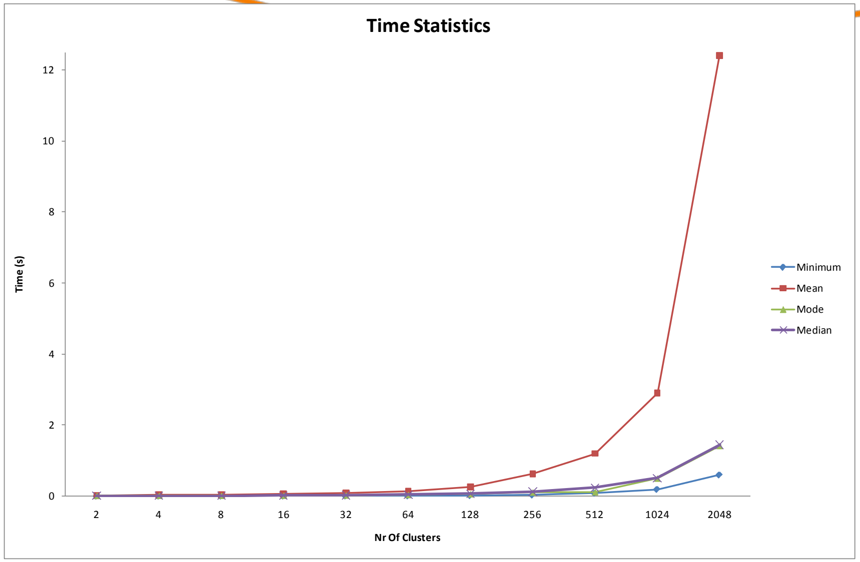
- K-Means is a clustering algorithm that partition n observations in K clusters. It starts with the K initial mean points and then uses an iterative process in which each observation is assigned to the cluster with the closer mean value, and by doing so it finds the natural center of the clusters

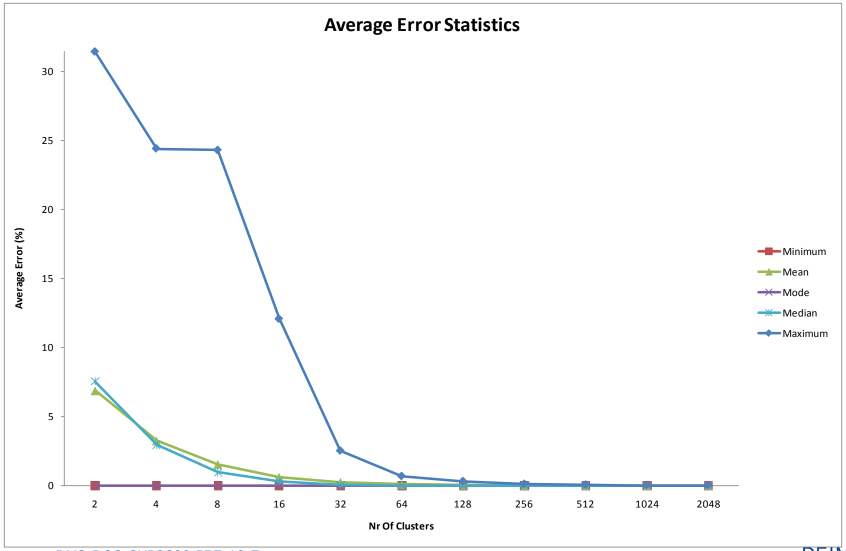
- The number of clusters again is a very important parameter and can produce difference between a good or poor performance. This algorithm works better on clusters with similar size.

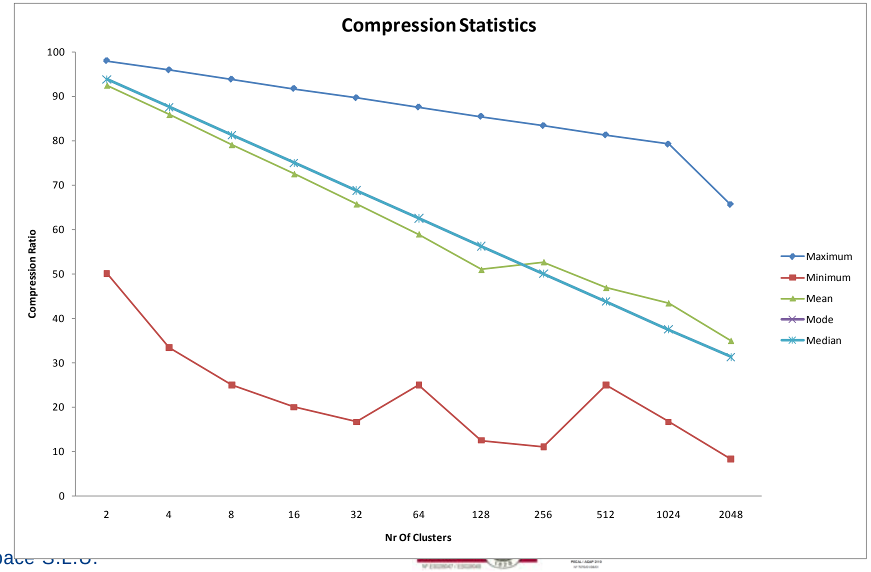
- Algorithm set for 64 clusters (6 Bit encoding). Only parameters originally encoded with 7 or more bits have been processed

K-Means Clustering





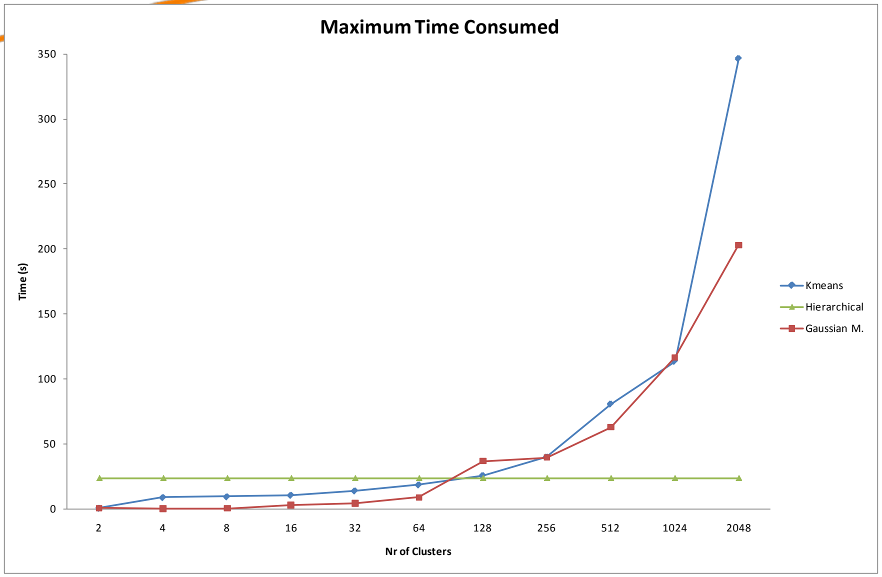


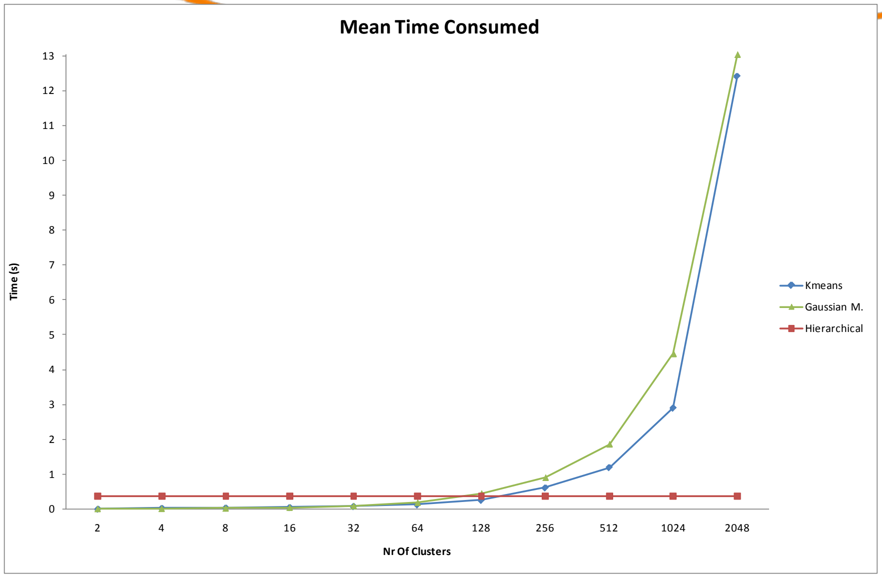


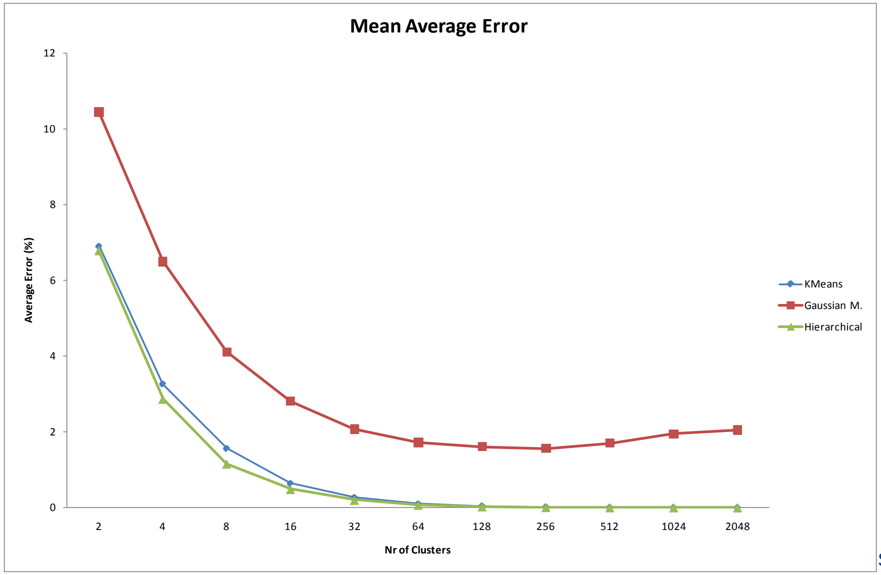
K-Means Clustering

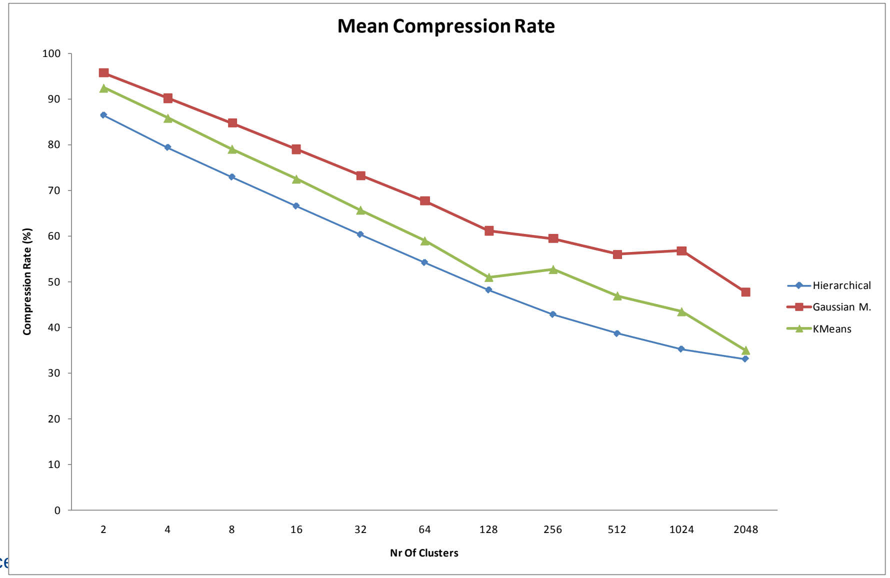
- The algorithm was able to work with a larger number of observations compared with hierarchical clustering
    - It takes less than 1 second to process each parameter
- For 32 or more clusters the average error for almost all the parameters is less than 1%, and when using 512 clusters the error is less than 0.1% for all parameters
    - This clustering algorithm has a very good behaviour in terms of representing the original time series.

Clustering Techniques Comparison





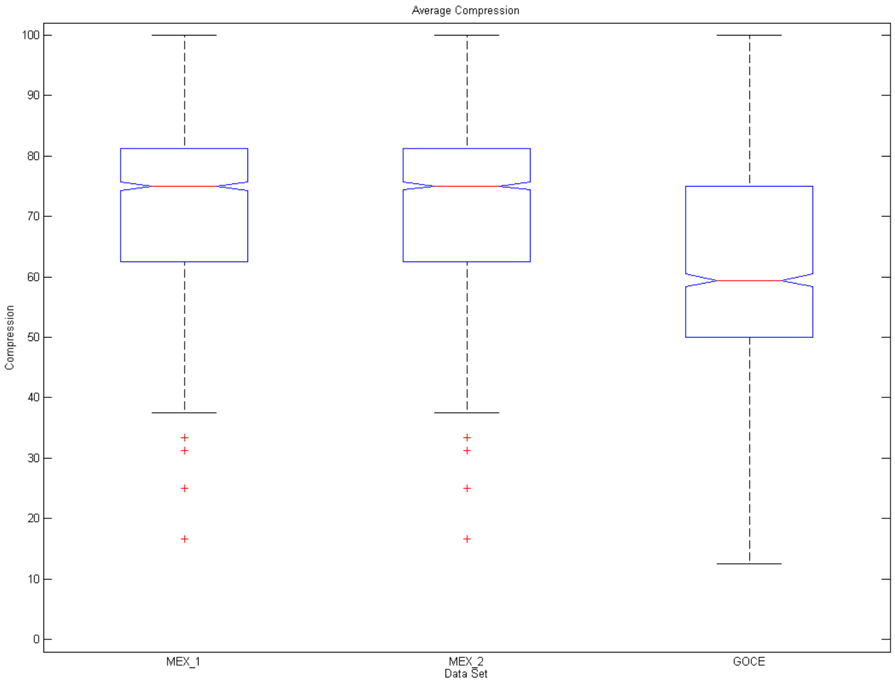


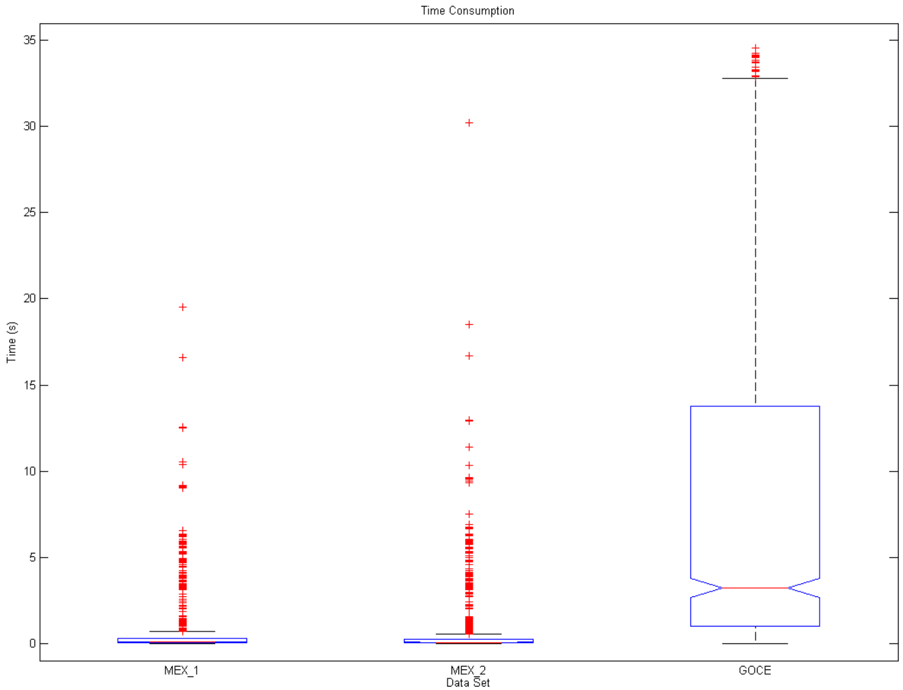


Clustering Techniques Comparison

- The comparison was made on 25000 samples due to limitation of the hierarchical clustering
- The K-means and Gaussian Mixture algorithms present an exponential behaviour when analysing the maximum time, the Gaussian is in general faster than the K-means
  - The hierarchical has a constant value and independent from the number of clusters
- It is confirmed that the Gaussian Mixture Clustering is not suitable for telemetry parameters, because it is oriented to obtain only a few clusters
  - In some cases the algorithm cannot produce the specified number of clusters
- Hierarchical clustering has the advantage of been able to automatically define the number of clusters to be used for each parameter
  - Not possible to work with large number of observations
- The K-Means clustering algorithm presents a good performance, and it is able to handle with larger amount of data than the hierarchical clustering, but it is strongly dependant on the number of clusters to be used
- According to the results above, the hierarchical clustering technique is selected for perform verification studies for checking the obtained results so far.

Hierarchical Clustering Verification





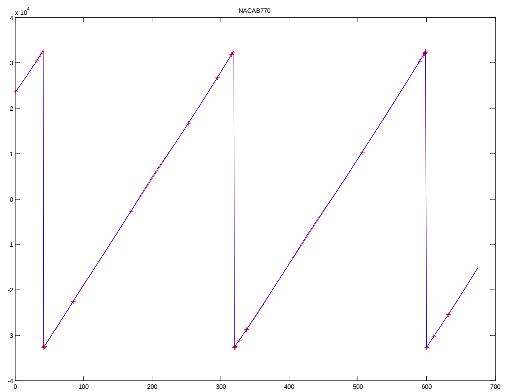
- Good results in terms of compresion and time comsuption although with a big variability

Optimal Re-Sampling

- The algorithm receives as input the time series samples ([time, value] pairs) and the maximum allowed error, and gives as output a new time series samples ([time, value] pairs) that has fewer or equal samples than the original series.
- By using linear interpolation between each consecutive pair of samples, the output series should resemble the original series guaranteeing the maximum error previously defined.

Optimal Re-Sampling

Some results on test parameters:



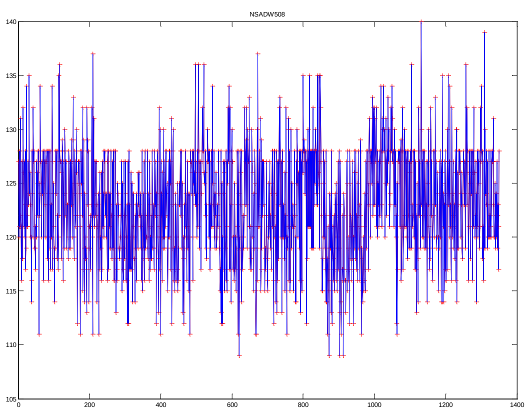
Parameter Name: NACAB770

Original Samples: 675

Fractal-Inspired samples: 29

Reduction Ratio: 95,7%

Time Consumed: 27.3 ms



Parameter Name: NSADW508

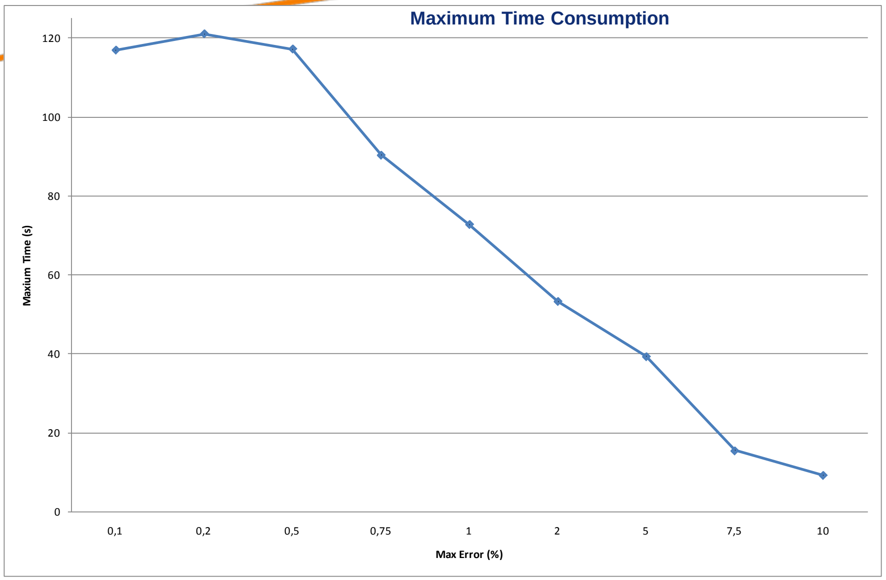
Original Samples: 1350

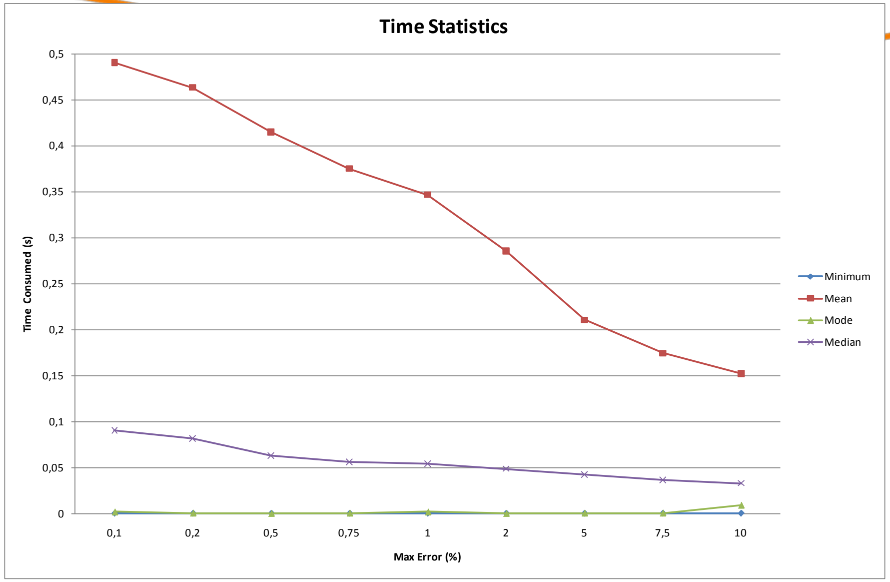
Fractal-Inspired samples: 1336

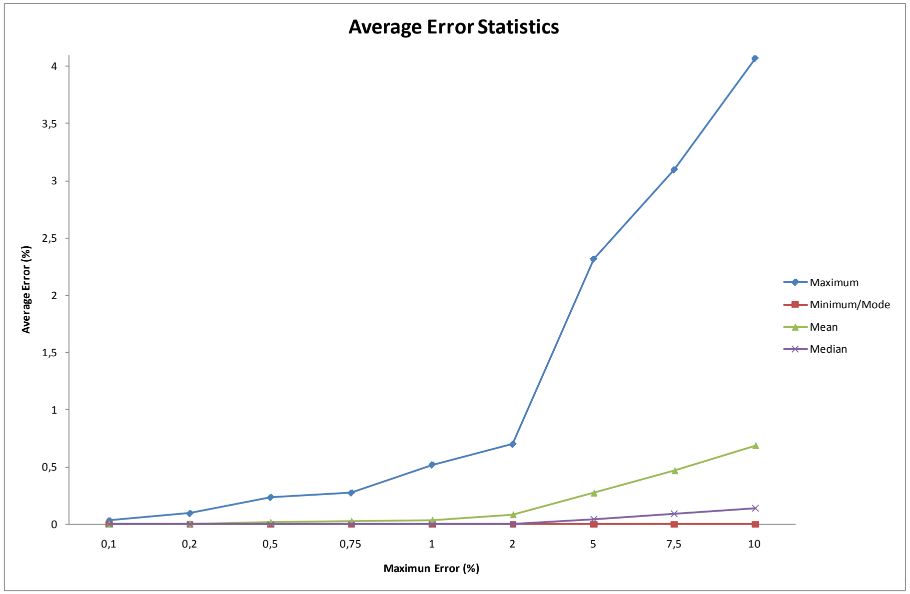
Reduction Ratio: 1,03 %

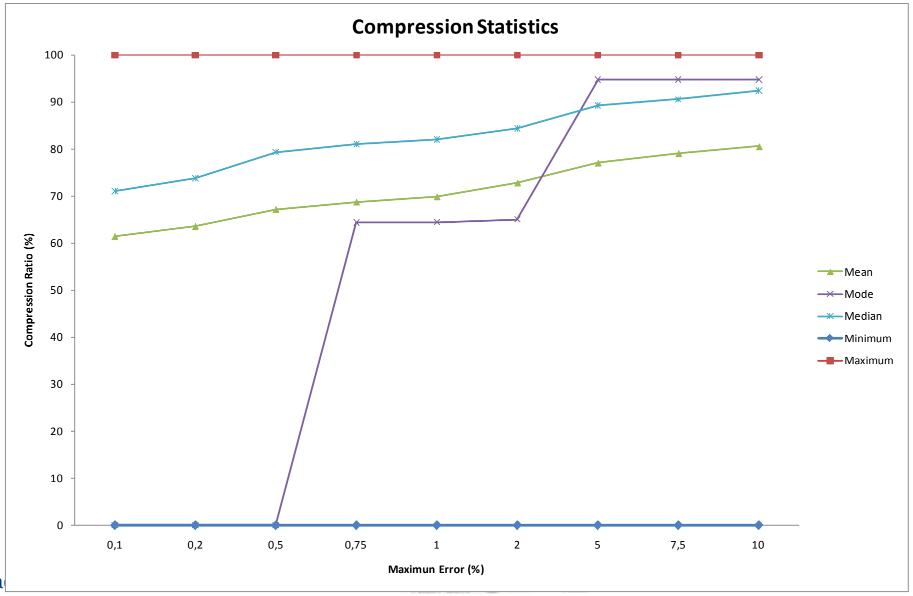
Time Consumed: 777.9 ms

Optimal Re-Sampling





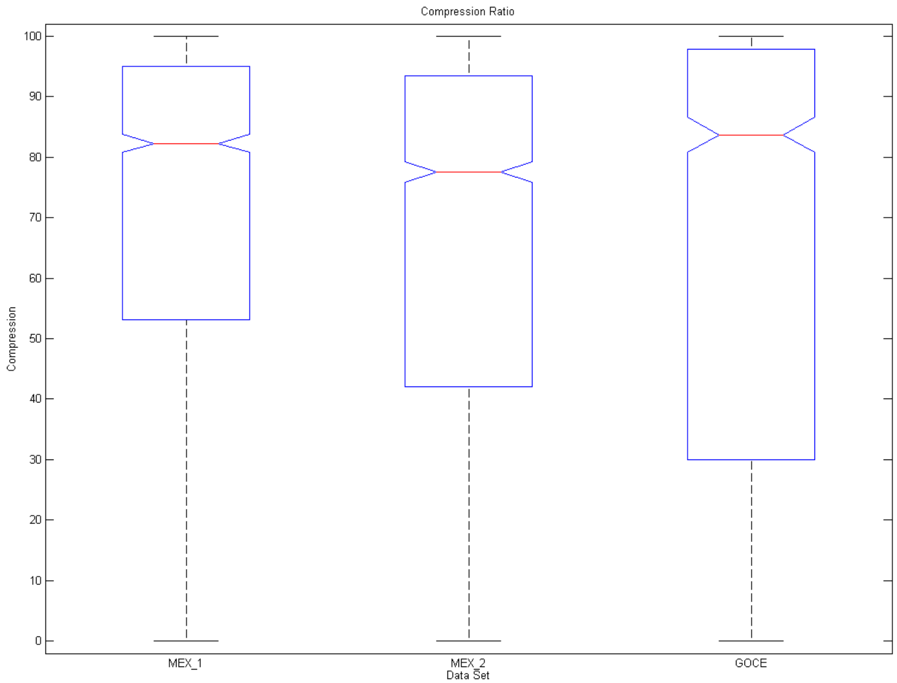


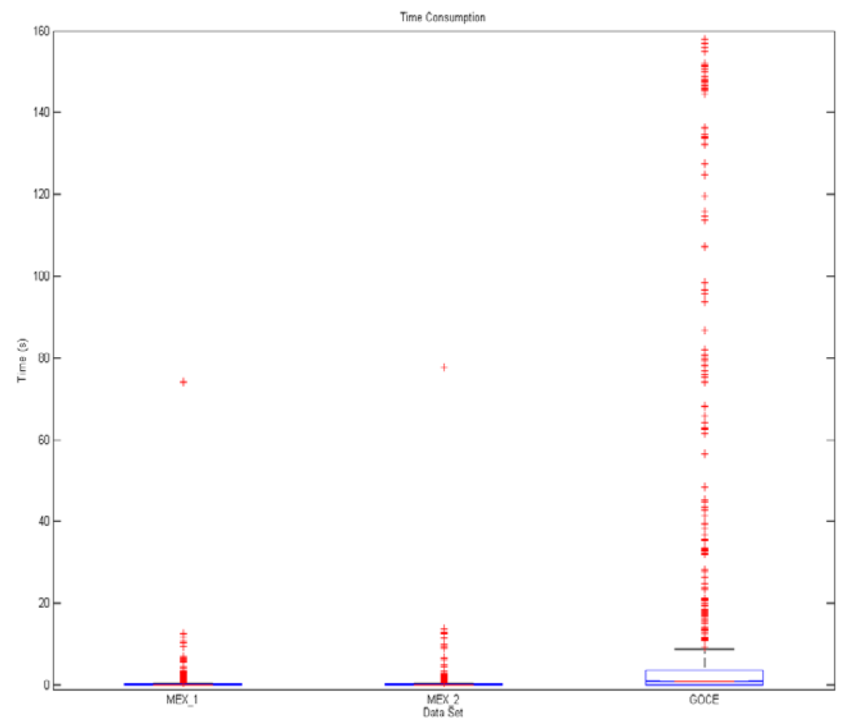


Optimal Re-Sampling

- Results show that the algorithm has a very good representation of the original time series
- When the time series is uniform or has linear behaviors (ramps, squared signals) the optimal re-sampling will be very efficient and the compression will be close to 100% (maximum value is 99.92% in all cases).
  - On the other hand when the time series has a lot of variations the compression could be 0% even if the maximum error allowed is very high.
- In terms of performance optimal re-sampling obtains very good results

Optimal Re-Sampling Verification





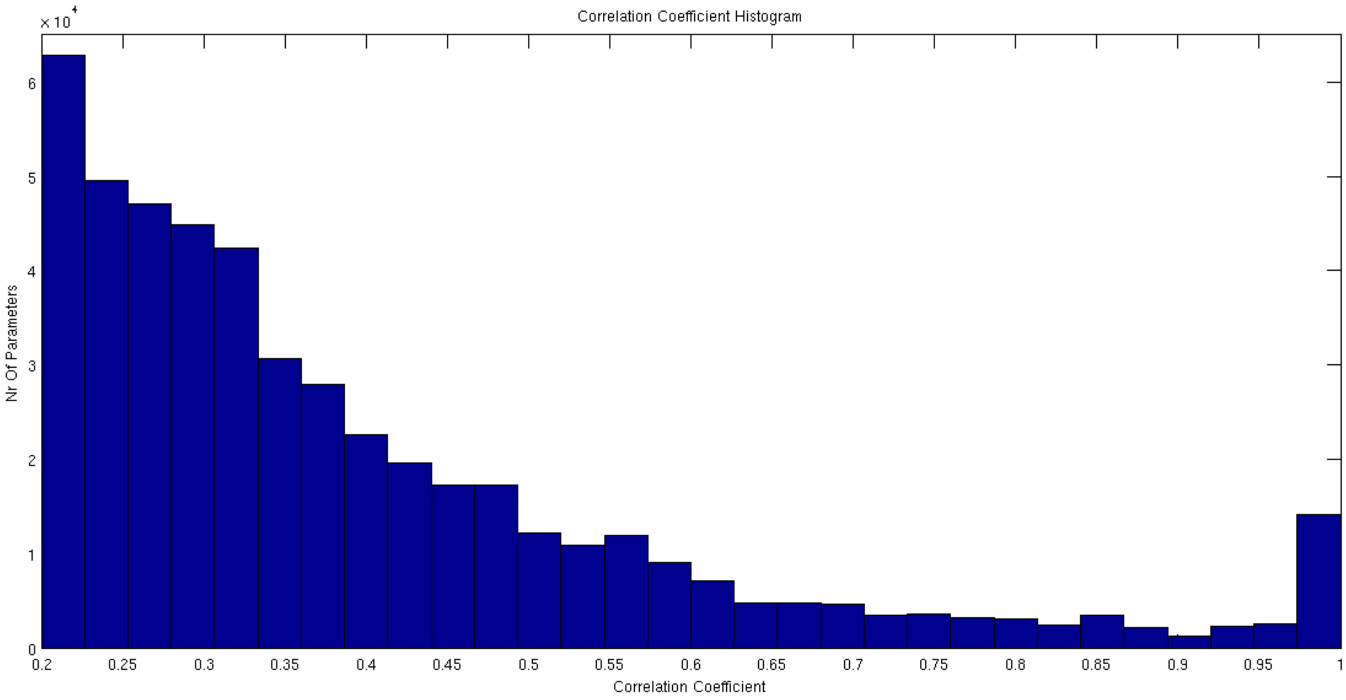
- ❑ Optimal Re-Sampling obtains very good results in terms of compression and performance in all data-sets

Data Correlation Analysis

- Objective: To identify significant relationship among parameters.
  - Difficulty: There is not previous knowledge about data therefore the analysis was made by a blind search
- Process:
  - Re-sampling all parameters to a pre-defined frequency and in the same timestamp
    - Minimum timestamp –> Time = 0
    - Definition of the increments based on the sampling period.
    - Linear interpolation was applied when the original value had not match the expected timestamp
  - Correlation analysis between all possible couples of parameters
- From each correlation analysis, two values P and R are obtained
  - P stands for the probability of finding the given correlation coefficient with a random time series
  - R is the actual correlation coefficient
  - If the value of P is greater than 0.05 or if the correlation coefficient is smaller than 0.2 the correlation is not meaningful and therefore discarded

Data Correlation Analysis

- A significant correlation coefficient (R > 0.2 or R < -0.2) was found for 25% (488840) of the possible combinations (1945378)
- When the value of the correlation coefficient increases, the number of correlations decreases

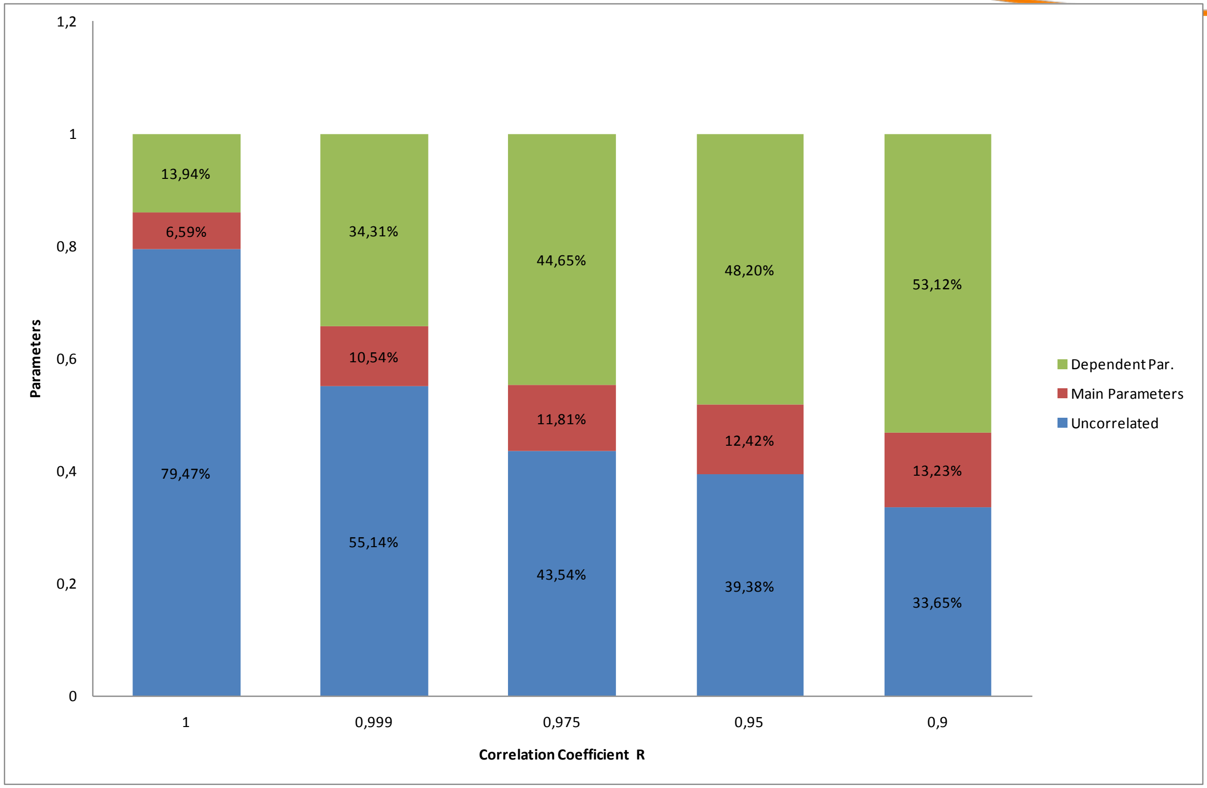


High number of correlations with R = 1, this means that those signals are identical or very strongly correlated

Optimising the Telemetry using Correlations

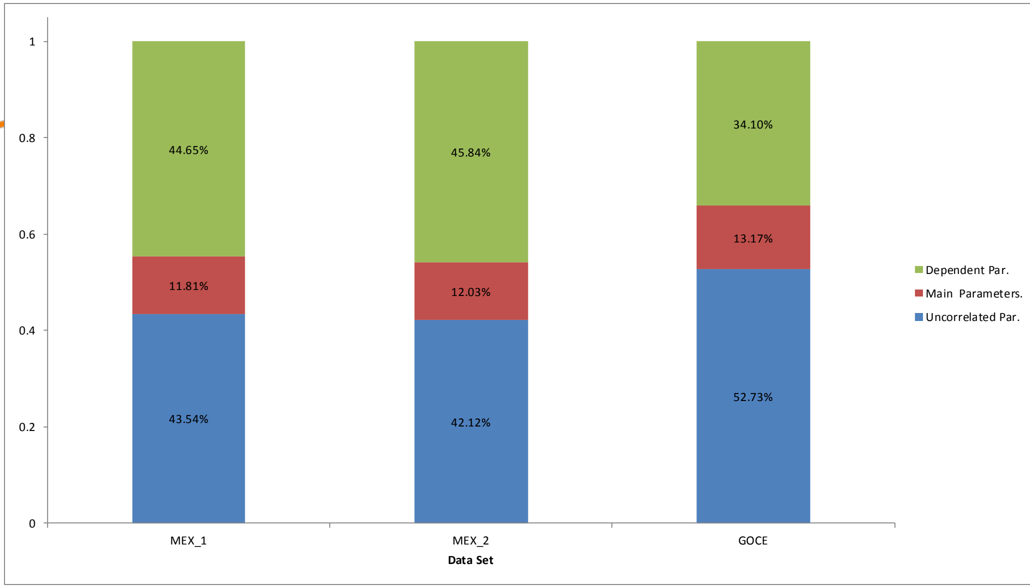
- Reduce the amount of parameters due to high number of correlations
  - Identify some “Master Parameters”: high level of correlation.
  - These Master parameters are suitable to be processed knowing that the correlated parameters have almost identical behaviour
- Work only with the parameters with R = 1 or R = -1
  - 79.47% (1568) of the parameters does not have correlation at this level i.e. they are independent (they should processed)
  - 6.59% (130) are the “master” parameters
  - 13.94% (275) are dependent of the “master” parameters (not needed to be processed)
- If the tolerance is reduced (correlation coefficient smaller than 1) the reduction in number of parameters will be higher
  - With correlation coefficient ($R > 0.999$ or $R < -0.999$) the reduction obtained is 34%.

Optimising the Telemetry using Correlations

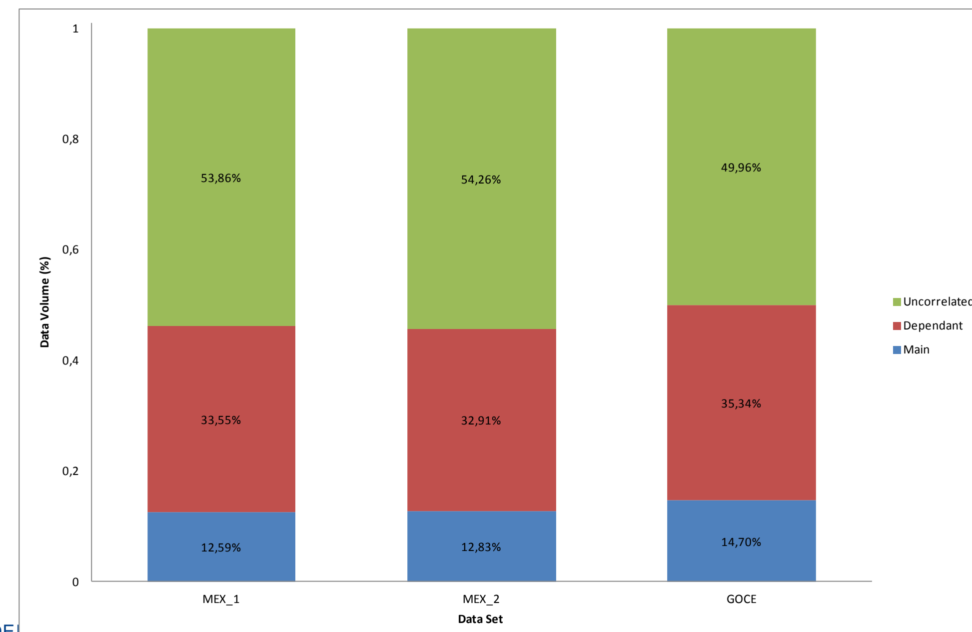


- With small reductions on the minimum correlation coefficient accepted, the number of dependent parameters increases
  - With R value over 0.975, almost one half of the parameters can be considered as dependent.

Data Correlation Verification

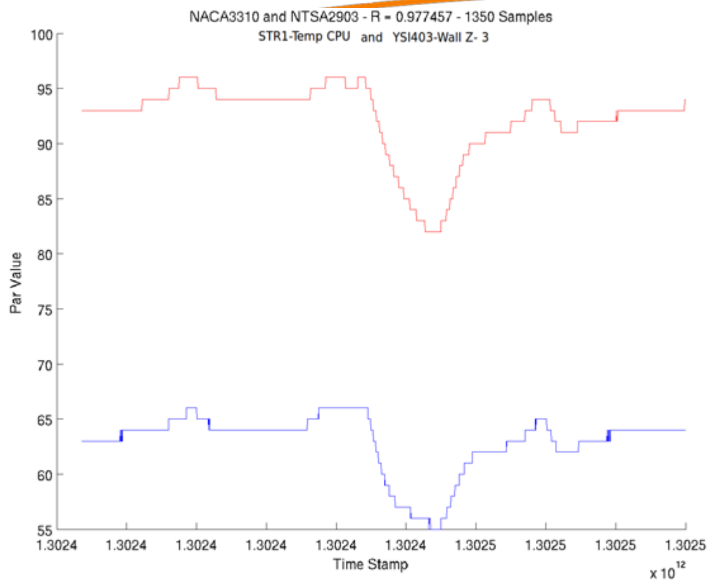


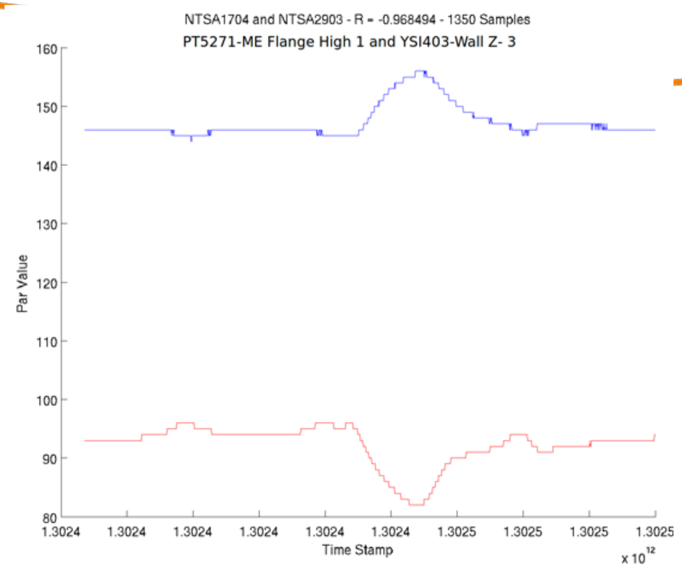
In terms of Volume

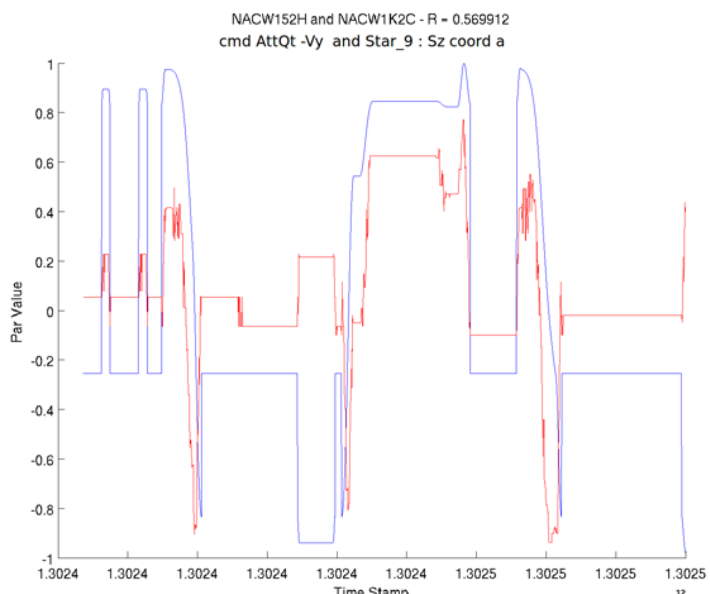


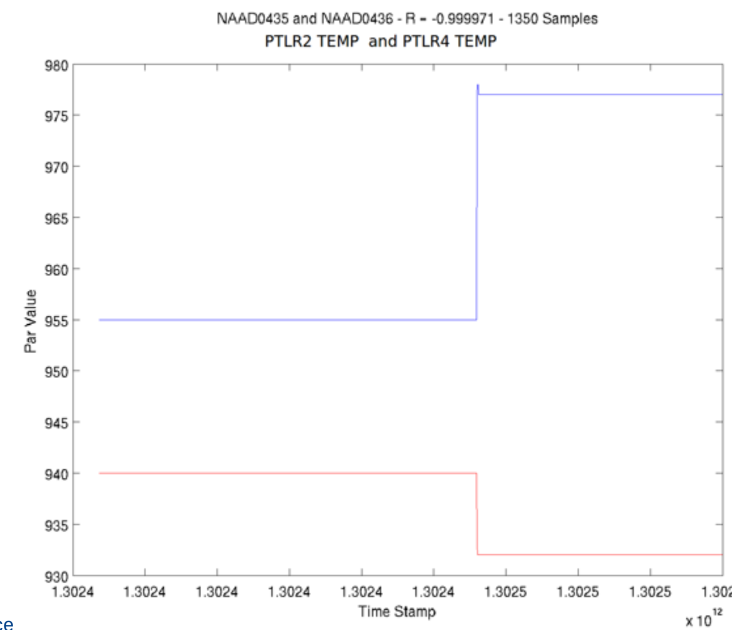
- The “master parameters” can be identified in all data-sets

Examples of Data-Correlations









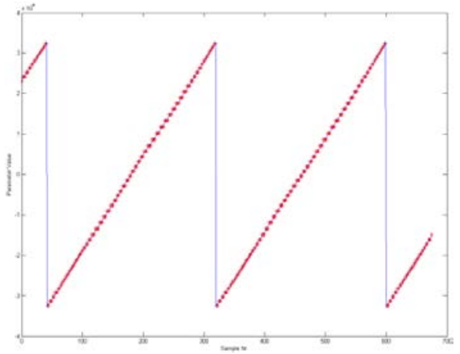
Assesment of Data Reduction

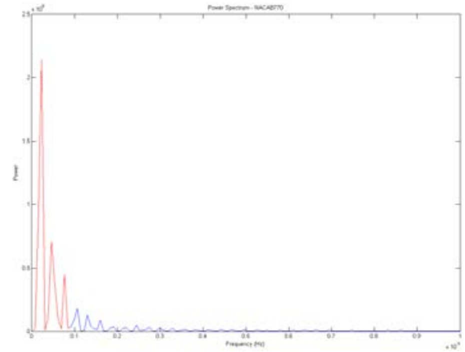
- It is possible to reduce the data volume by applying the proposed techniques
  - The results are consistent given that the techniques obtain similar results in all data sets
- Each technique evaluated reduces the data in a different form:
  - Optimal Re-Sampling reduces the number of samples
  - Clustering reduces the number of bits used for information coding
  - Data correlation reduces the amount of parameters to take into account
- In terms of time consumption, optimal re-sampling behaves better than the other two techniques.

Data Set	Mean Compression Optimal RS	Mean Compression H Clustering	Mean Compression Correlations
MEX 1	69,85	71,11	33,55
MEX 2	66,95	70,63	32,91
GOCE	65,28	63,68	35,33

Data compression for all techniques (in terms of volume)

Noise Evaluation

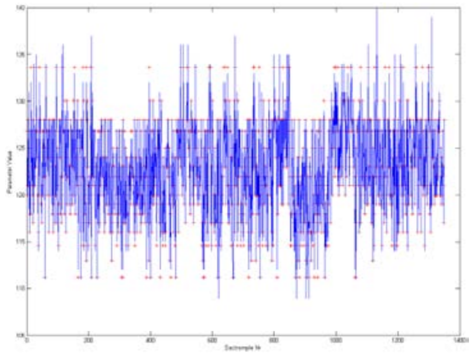


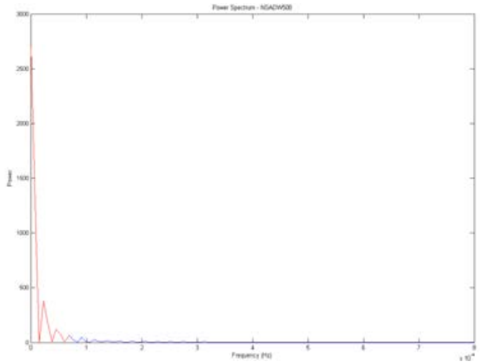


Parameter Name: NACAB770
Number of Samples: 675

Signal : 80.12%
Noise : 19.87%
S-to-N : 4.03

Main Frequency: 3.051 e-05 Hz





Parameter Name: NSADW508
Number of Samples: 1350

Signal : 92.96%
Noise : 7.03%
S-to-N : 13.21

Main Frequency: 7.62e-6 Hz

- The main problem with the noise is that it is very difficult to identify it. It could be done in a parameter by parameter basis but not in a general way.
  - It is not possible to define a general criterion for differentiate between noise and signal.

Identifying the State of the Spacecraft

- ❑ Data Correlation
- ❑ Clustering

Approach

- From the techniques evaluated for data reduction, two of them were selected for detecting the state of the spacecraft:
  - Data Correlation
  - Clustering

- Additional sets of data were used from MEX:
  - Four nominal data sets: MEX_1, MEX_2, MEX_3 & MEX_4
  - Two no-nominal data sets: MEX_minor & MEX_critical

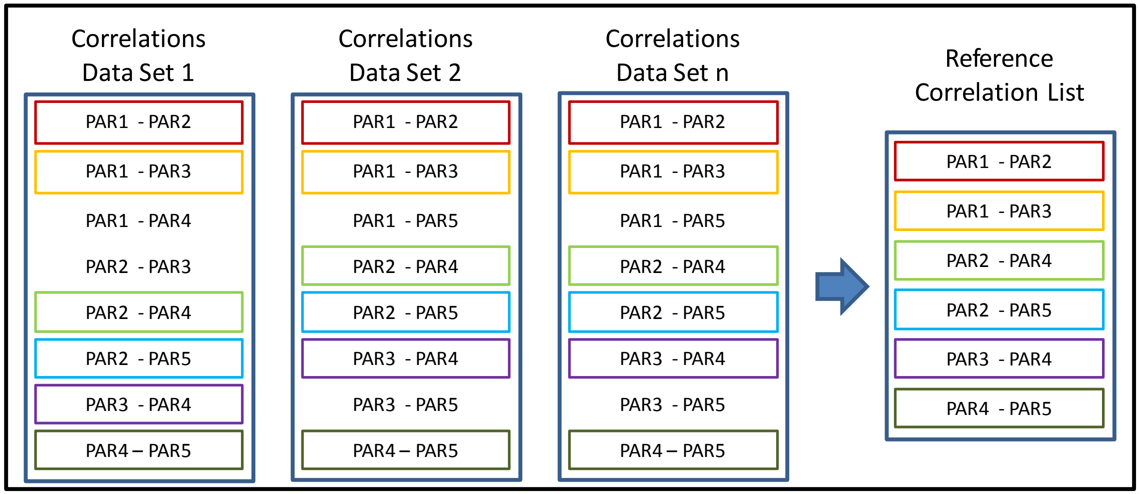
- Difficulty: All the data sets do not contain the same parameters

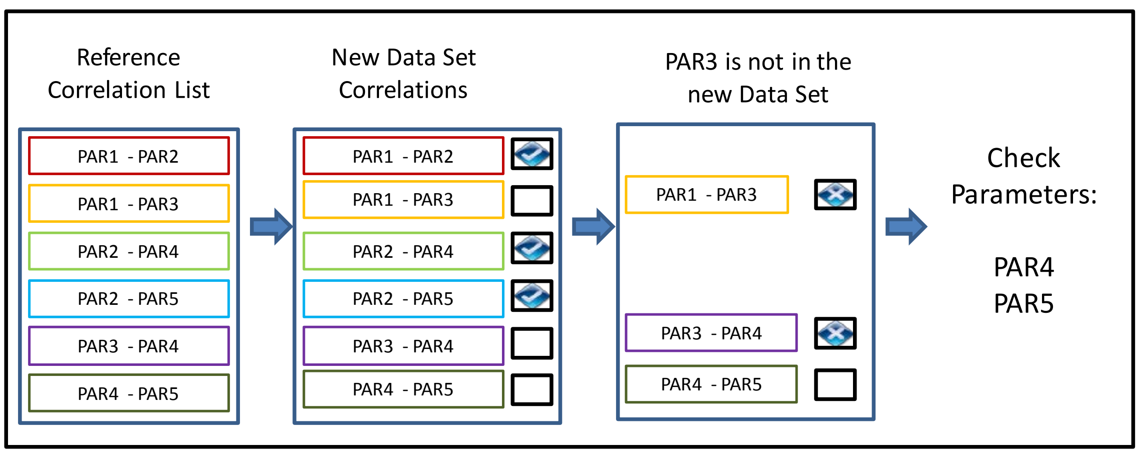
Decision Making based on Data Correlation

- Two approaches have been explored:
  - **Broken Correlations:** The broken correlations approach is based on detecting broken correlations as a result of anomalies in the spacecraft. The hypothesis is that if something goes wrong in the spacecraft some of the strong correlations could be broken due to the bad functioning.
  - **Master Parameters Variation:** This analysis is based on on checking the variation of the “master” parameters. The goal is to look for “strange” behaviour in the “master” parameters that could indicate a wrong behaviour in the spacecraft

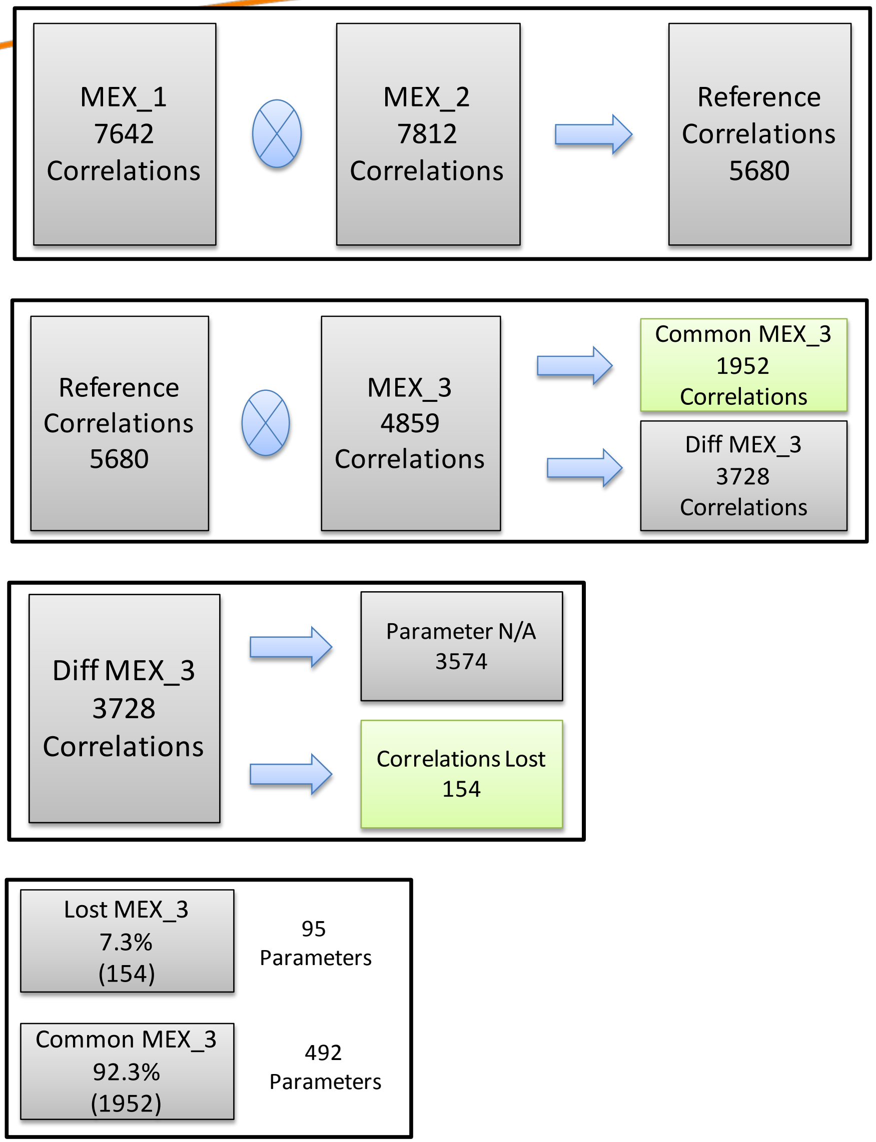
Broken Correlations

- Algorithm:





Broken Correlations



Broken Correlations

MEX_1 7642 Correlations ⊗ MEX_2 7812 Correlations → Reference Correlations 5680

Reference Correlations 5680 ⊗ MEX_3 4859 Correlations → Common MEX_3 1952 Correlations
→ Diff MEX_3 3728 Correlations

Diff MEX_3 3728 Correlations → Parameter N/A 3574
→ Correlations Lost 154

Lost MEX_3 7.3% (154) — 95 Parameters

Common MEX_3 92.3% (1952) — 492 Parameters

Data Set	Number of Corr.	Common Corr.	Different Corr.	Parameter N/A	Correlation Lost
MEX_3	4859	1952	3728	3574	154
MEX_4	4637	4071	1609	1461	148
MEX_minor	11596	3270	2410	1741	669
MEX_critical	6134	5137	543	102	441

Data Set	Correlation Lost	Common Corr.
MEX_3	7,30%	92,70%
MEX_4	3,50%	96,50%
MEX_minor	17,00%	83,00%
MEX_critical	7,90%	92,10%

Broken Correlations

MEX_3

Parameter Description	Lost Correlations
Earth to SC dist'	14
Fut segment start time'	14
Curr segment start time'	14
Sun to SC dist'	14
Fut segment end time'	13
Curr segment end time'	13
Star_5 : Sy coord a'	8
Star_6 : Sy coord a'	8
Star_9 : Sy coord a'	8
RM4-Synchro time 1 reg'	6
Last time in TM packet ('	6
Last time for User Synch'	6
Current time(COARSE)'	6
COARSE of CDMU(PM) acq'	6
COARSE of CDMU(RM1) acq'	6
COARSE of CDMU(RM2) acq'	6
COARSE of CDMU(RM3) acq'	6
COARSE of CDMU(RM4) acq'	6
Star_5 : Id Nb a'	6
Star_3 : Id Nb a'	6

MEX_4

Parameter Description	Lost Correlations
SAS1C4 - X(-X, +Zsc) R'	16
THR2 TEMP R'	13
Lock cnt : a_acm_swr_on'	13
Lock cnt : rcs_orb_on'	13
Fut segment start time'	12
Curr segment start time'	12
Fut segment end time'	11
Curr segment end time'	11
TM +15V R'	7
TM +15V R'	7
TM - 15V R'	6
TM-15R - SECOND VOLT TMR'	6
Reserved - 0433'	6
SAS1C3 - X(-Y, -Ysc) R'	5
THR1 TEMP R'	5
THR3 TEMP R'	5
PTLR2 TEMP'	5
Reserved - 0437'	5
Reserved - 0438'	5
RM4-Synchro time 1 reg'	5

MEX_minor

Parameter Description	Lost Correlations
TC -10 Received NUMBER'	42
IB Temp1 SetPoint'	35
SW Laser Temp SetPoint'	35
LW Laser Temp SetPoint'	35
SW Detector Temp SetPoin'	35
TC -15 Received NUMBER'	33
TC -16 Received NUMBER'	33
IB Temp5 SetPoint'	33
IB Temp6 SetPoint'	33
IB Temp7 SetPoint'	33
IB Temp8 SetPoint'	33
ADC SW Red Set'	21
ADC SW Nom Set'	21
Speed Reference Setup'	21
TC -5 Received NUMBER'	20
TC -8 Received NUMBER'	20
TC -12 Received NUMBER'	20
IB Temp2 SetPoint'	20
IB Temp3 SetPoint'	20
IB Temp4 SetPoint'	20

MEX_critical

Parameter Description	Lost Correlations
IB Temp2'	33
IB Temp1 wr'	33
TC -15 Received NUMBER'	20
TC -16 Received NUMBER'	20
IB Temp5 SetPoint'	20
IB Temp6 SetPoint'	20
IB Temp7 SetPoint'	20
IB Temp8 SetPoint'	20
TC -10 Received NUMBER'	19
SAS1C4 - X(-X, +Zsc) R'	15
TM +15V R'	14
TM +15V R'	14
Lock cnt : a_acm_swr_on'	13
Lock cnt : rcs_orb_on'	13
RM1-Synchro time 1 reg'	13
IB Temp7'	13
IB Temp2 wr'	13
LW Laser Temp'	12
LW Detector Temp'	12
LW TRW Current'	12

"Master" Parameters Variation

- Algorithm:
  - Identification of the "master" parameters
  - Process them in terms of mean and variance tests in order to detect significant differences between batches of data

Parameter ID	Parameter Description	MEX1	MEX2	MEX3	MEX4	MEX Minor	MEX Critical
'NDWD0300'	Start of ground activity	34	34	34	34	34	34
'NDMA5716'	RM1-Synchro time 2 reg	32	32	32	32	32	32
'NAAD0401'	THR1 TEMP R	26	20	24	10	16	14
'NDWDO100'	Start of OBCP MANAGER (C	25	1	12	25	0	25
'NAAD1905'	LCV 1A Open Status Nom	17	16	22	12	16	12
'NAWD0300'	Start of ground activity	17	17	21	17	17	17
'NACW0S00'	Earth to SC dist	15	16	1	12	12	15
'NPWA1470'	PCU-W47	11	1	1	1	3	1
'NPWD2521'	LCL1A curr EPC A	10	5	10	12	10	12
'NDWD0A07'	End SSMM_MGR wrk-fine	9	2	5	5	4	5
'NPWA1440'	PCU-W44	9	19	19	19	8	19
'NAWD0301'	Start of ground activity	7	4	8	4	9	4
'NAWD0A0U'	b_acm_swr_off	6	2	6	5	6	5
'NAAD0305'	ME TEMP N	5	1	1	1	1	1
NACW0M0J	THR_1 switchON Nb	5	5	7	5	5	5

Master parameters and the number of correlations in each data set

Mean comparison for Master Parameters

Parameter ID	MEX1	MEX2	MEX3	MEX4	MEX Minor	MEX Critical
'NDWD0300'	1566.047	2871.279	2678.105	-3538.176	1419.761	1322.238
'NDMA5716'	30401.095	28747.860	31944.552	34978.411	32601.380	30936.476
'NAAD0401'	1463.208	1438.646	1470.837	1443.828	1474.620	1480.918
'NDWDO100'	250516812	257601716	244641612	268574412	7819968	265982412
'NAAD1905'	0.9977	0.9963	0.9940	0.9962	0.9985	0.9962
'NAWD0300'	1295.919	2663.239	2341.378	-4112.396	895.508	1090.917
'NACW0S00'	350175224	332529239	355366653	231180574	149730288	265691635
'NPWA1470'	1508.008	1474.657	1501.840	1431.721	1432.775	1438.1711
'NPWD2521'	82.544	49.107	64.791	62.890	76.074	24.246
'NDWD0A07'	3342.678	3263.941	2876.517	2886.023	3176.916	3135.500
'NPWA1440'	644.762	600.647	640.115	591.409	600.0133	563.840
'NAWD0301'	1909.341	1932.842	1886.656	2344.091	1898.585	1938.804
'NAWD0A0U'	0.9979	0.9952	0.99604	0.9958	0.99868	0.99721
'NAAD0305'	1292.640	1345.329	1222.094	1352.848	1342.420	1357.606
'NACW0M0J'	11584.787	11715.028	11441.954	13184.438	6356.6803	12703.915

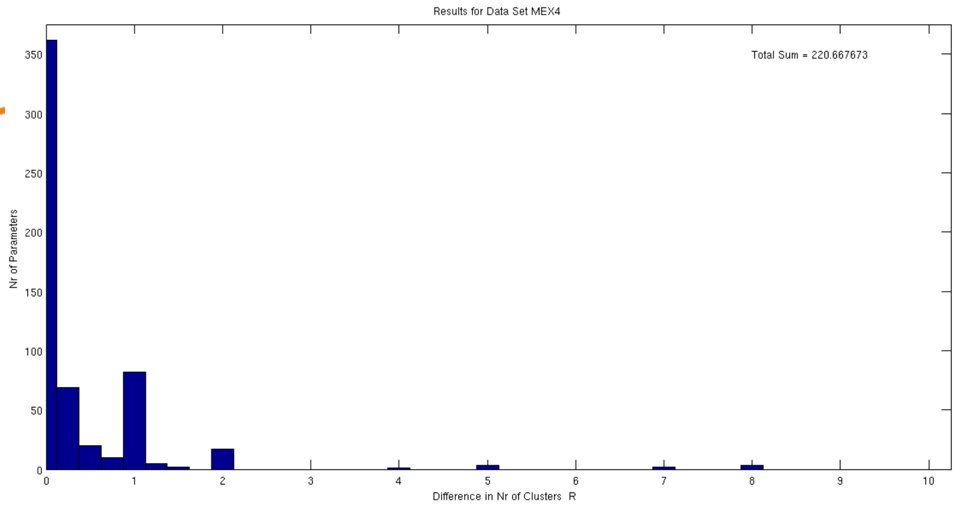
  - Nothing significant was found.
  - Variance analysis obtained similar results
- “Master” parameters variation does not provide conclusive results

Decision Making based on Clustering

- Two algorithms were developed:
  - Based on Hierarchical Clustering
  - Based on Hierarchical Clustering and Gaussian Clustering
- Algorithm based on Hierarchical Clustering:
  - Select the parameters that are common to all from all data sets.
  - Run the clustering algorithm for obtain the number of clusters, setting the cut-off value to 0.7
  - Divide the group of clusters in two groups: one with nominal data sets (MEX_1, MEX_2 & MEX_3) and the second one as a test group (MEX_4, MEX_minor & MEX_critical)
  - For each parameter, a vector with the number of clusters of all data sets in the reference group is created. The mean and variance of that vector is calculated, the objective is to obtain a mean-variance normal distribution of the nominal behaviour
  - When a new data set is analysed, the number of clusters obtained for the given parameter is compared with the distribution obtained and a response value R is given using a simple expression

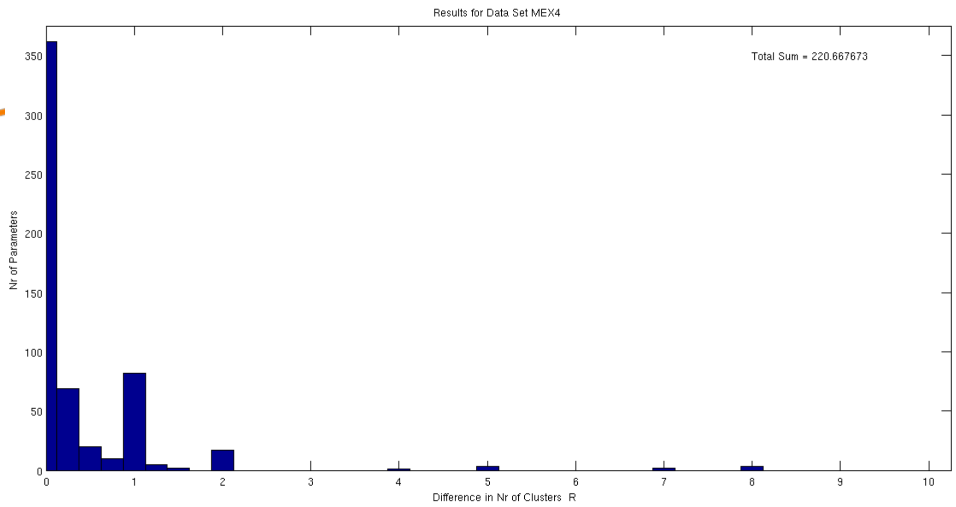
$$R = \frac{NClus_{new} - Mean}{Variance}$$

Decision Making based on H. Clustering



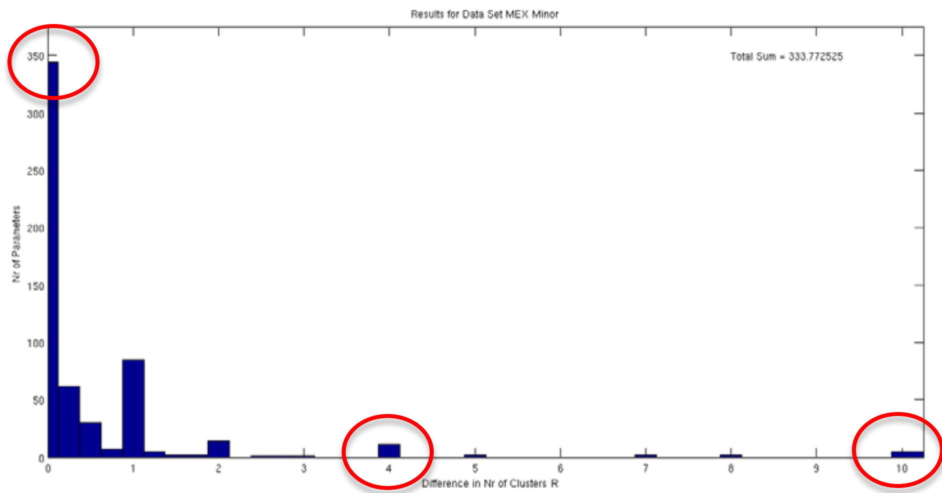
MEX_4

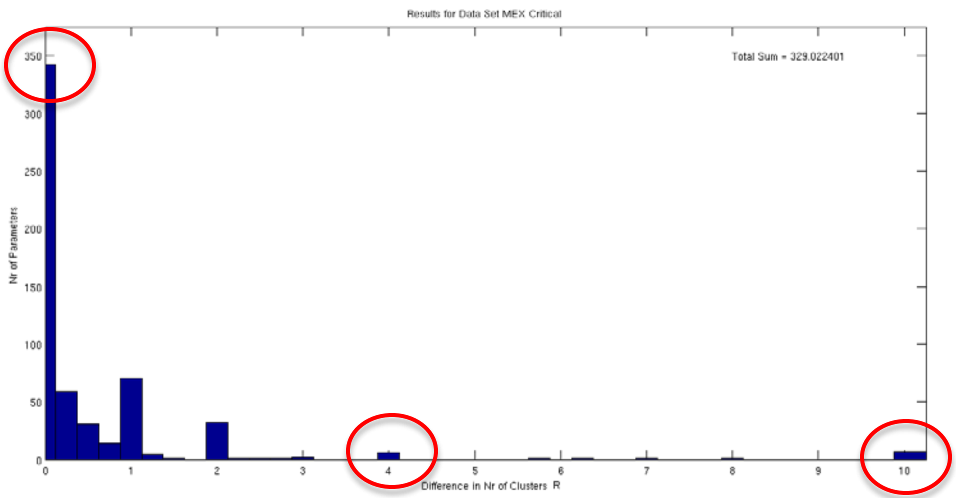
Decision Making based on H. Clustering



MEX_4

MEX_minor





MEX_critical

- The algorithm is able to differentiate between nominal and no-nominal but not between minor and critical anomalies elecnor deimos

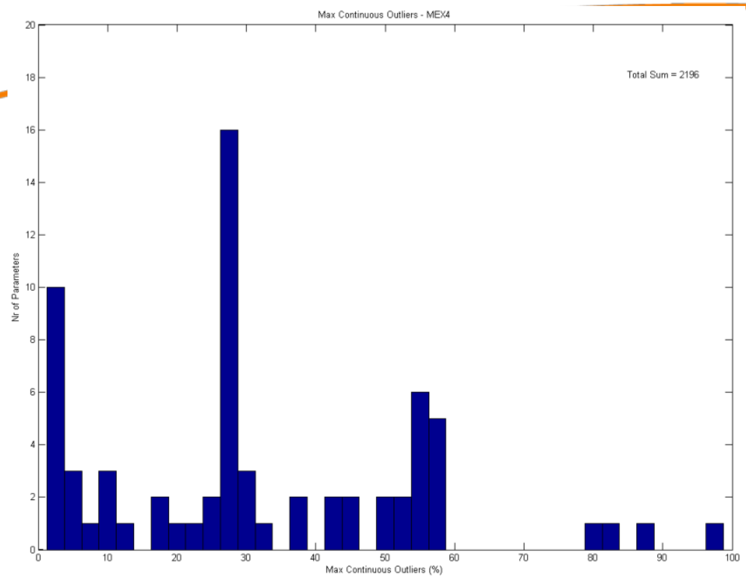
POLITÉCNICA

Algorithm based on Hierarchical + Gaussian Clustering

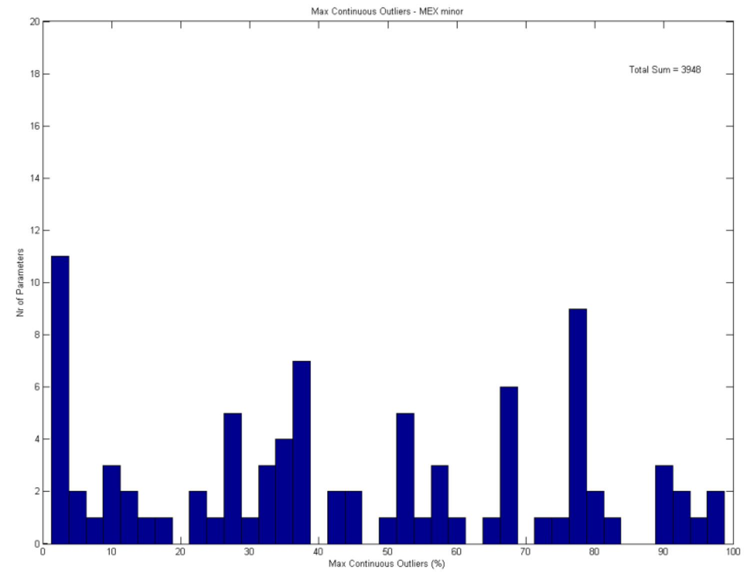
- The hierarchical clustering obtains the number and the center point of the clusters
- For each new sample, the Gaussian Clustering gives a vector with the same size as the number of clusters and in each position the probability of belonging to that cluster
- The algorithm developed takes the results (number of clusters and mean value of each one) from the hierarchical clustering applied to a reference data set, and creates a gaussian mixture distribution
- Once the Gaussian mixture distribution is created, it is used to categorize new data by assigning to each observation from a new data-set a vector with the probabilities of being in a given cluster
- If the maximum probability obtained is less than the threshold the observation is marked as an outlier.
- A counter of continuous outliers is defined, if the counter is small it means that the outliers are not representative but if it is a high number it may represent a new cluster (i.e. an anomaly or a new tendency).

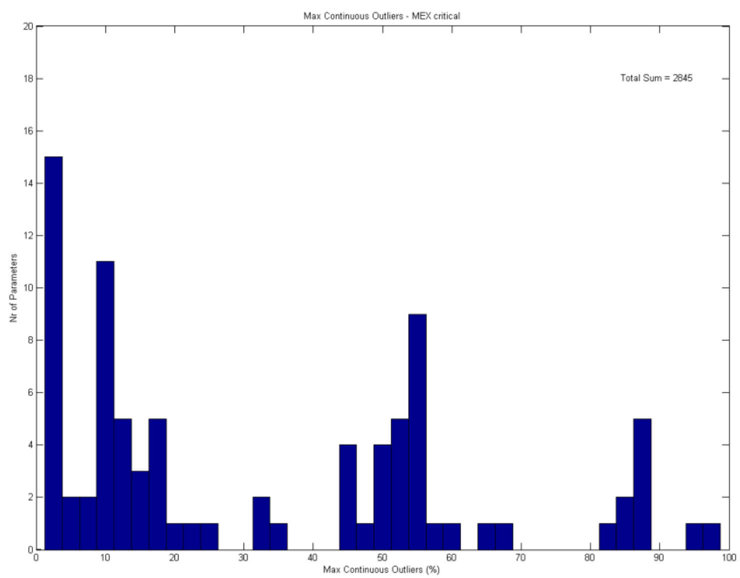
Decision Making based on Hierarchical and Gaussian Clustering

MEX_4



MEX_minor





MEX_critical

- ❑ This algorithm is not able to differentiate between nominal and no-nominal

Conclusions & Discussion

Conclusions & Discussion

❑ Overall Summary

- Logic of the project has been divided in three stages:
  - Getting basic knowledge of the telemetry data (Task 1)
  - Exploring different techniques for reducing the amount of data without losing information (Task 2 and Task 3)
  - Developing algorithms for detecting anomalies in the spacecraft (Task 4)
- The main summary of the study is that several thousands of parameters were analysed in a “blind” search finding that it could be possible to reduce the amount of data without losing information. Indeed a mean reduction of 60% of data volume was obtained (range between 33% - 70%)
- It was found that it could be possible to differentiate data sets with anomalies from data sets without anomalies. Although these results are not totally conclusive.

Data Set	Received Time Series Total Parameters	Processed Time Series Non-Static/No Payload	Processed Registers Non-Static/No Payload
MEX 1	8101	1973	4075692
MEX 2	8322	2180	4140339
MEX 3	7355	1479	657717
MEX 4	7467	1570	3054419
MEX Minor	8175	2221	4187515
MEX Critical	8024	1998	4436545
GOCE	4620	1391	50079040
Total	52064	12812	70631267

Number of parameters analysed

Conclusions & Discussion

- Getting knowledge of Telemetry Data
  - It was found that there are an important number of parameters that do not change during the sampled time (75% of the parameters) although in terms of volume is not too big (38% of volume).
- Exploration of techniques for reducing the amount data without losing information
  - From the techniques evaluated, three obtained good results for reducing the amount of data: Optimal Re-sampling, Clustering and Data Correlation.
  - The results obtained for the three techniques above have been verified using additional data sets of telemetry data: One additional set of MEX mission data and a set from GOCE mission. In all of them the results are similar.
  - Noise study does not provide any valuable result because it is not possible to isolate the noise from the real signal without any a-priori knowledge about the signal. Some proposals were implemented but the results were not conclusive.

Conclusions & Discussion

- Three data sets have been analysed corresponding to 5.538 time series (parameters) for a total of 36’332.141 registers or observations that have been processed.
- The main conclusion is that is possible to reduce the amount of data without losing valuable information. Depending on the technique used, the reduction can be up to 70%.

Data Set	Mean Compression Optimal RS	Mean Compression H Clustering	Mean Compression Correlations
MEX 1	69,85	71,11	33,55
MEX 2	66,95	70,63	32,91
GOCE	65,28	63,68	35,33

- It is important to note that each technique reduce the data in a different way:
  - Optimal Re-sampling reduce the number of samples of a given time series,
  - Clustering reduces the number of bits necessary for store the parameter value
  - Data correlation reduce the number of parameters used for represent the state of the spacecraft (removes the parameters that the information are already covered by other parameters).

Conclusions & Discussion

- Algorithms for detecting anomalies in the spacecraft
  - Two techniques were selected for determination of the spacecraft status: data correlation and clustering
  - The algorithm based on data correlation (searching for broken correlations) shows that it is possible to identify parameters with abnormal behaviour but adding additional analysis (i.e. doing mean tests and variance tests).
  - Using Hierarchical Clustering it is possible to detect when a data set has anomalies, although some tests for verifications should be done in order to have definitive conclusions.
  - Comparison of the performance of the different techniques

Technique	Representation of the Status	Time Consuming	Computing Resources
Correlation	LOW	HIGH	MID
Clustering Hier.	HIGH	MID	HIGH
Clustering Gauss + Hier.	MEDIUM	LOW/MID	LOW(*)

(*) Assuming that the clusters are previously created

Conclusions & Discussion

❑ Future Work

- To deepen on the real implementation of the techniques for reducing the amount of data. Techniques like Optimal Re-Sampling or Clustering are suitable to be implemented on-board
  - Combination of techniques: clustering for optimising the storage of the data on-board and optimal re-sampling for reducing the number of samples sent to the ground for re-build the signal
- Confirmation and fine tuning of clustering algorithms for differentiate between nominal and non-nominal behaviour
  - Algorithms should be robust against the difference in parameters between one data set and other in order to be able to compare

Thank you

¿Questions?