Spacecraft State Representation through Pattern Recognition

SMART TM Follow-On Study

Executive Summary

Code : SMARTTM-DMS-PMD-PRR04
Issue : 1.0
Date : 25/05/2012

	Name	Function	Signature
Prepared by	Adrian Mora	SMARTTM Project Manager	
Reviewed by	José Luis Fernandez	SMARTTM PA Manager	
Approved by	Adrián Mora	SMARTTM Project Manager	
Signatures and approvals on original			

DEIMOS Space S.L.U.
Ronda de Poniente, 19, Edificio Fiteni VI, 2-2ª
28760 Tres Cantos (Madrid), SPAIN
Tel.: +34 91 806 34 50 / Fax: +34 91 806 34 51
E-mail: deimos@deimos-space.com

This page intentionally left blank

Document Information

Contract Data	
Contract Number:	4000102900/10/NL/ AF
Contract Issuer:	ESA

Internal Distribution		
Name	Unit	Copies
Mike Rennie	Real Time Systems	1
Internal Confidentiality Level (DMG-COV-POL05)		
Unclassified ☑	Restricted ☐	Confidential ☐

External Distribution		
Name	Organisation	Copies
David Evans	ESA-ESOC	1
Jose A. Martinez-Heras	ESA-ESOC	1

Archiving	
Word Processor:	MS Word 2007 (Word 97-2003 Compatibility Mode)
File Name:	SMARTTM-DMS-PMD-PRR04-10.doc

Document Status Log

Issue	Change Description	Date	Approved
1.0	Initial version of the document	25/05/2012	

Table of Contents

List of Figures

List of Tables

1. INTRODUCTION

Today's space missions greatly rely on the monitoring of simple housekeeping telemetry, which is often relatively limited due to bandwidth use limitations. Based on this data, flight control teams need to analyse the behaviour of the spacecrafts and investigate the causes and effects of any anomaly or unexpected behaviour detected. On this situation, studies are being performed in order to continue advancing in the autonomy of the spacecrafts and their observability on-ground, while at the same time reducing the effort necessary for repetitive tasks and automating some tasks in order to help the ground operation centres to focus on the most specialized activities.

The SMART TM Follow-On study has been investigated the possibility of analysing the on-board generated housekeeping data in order to reduce the amount of data without loss of information and evaluate the feasibility of implement some type of decision making in order to detect automatically anomalies in the spacecraft, at least OK/NOK state.

The present Executive Summary, pretends to show the main activities carried out during the study development as well as the main results obtained.

1.1. Purpose

This Executive Summary is part of the data-pack for the final project milestone: Final Presentation.

2. RELATED DOCUMENTS

2.1. Applicable Documents

The following table specifies the applicable documents that shall be complied with during project development.

Table 1: Applicable documents

Reference	Code	Title	Issue	Date
[SOW]	DOPS-GS-SOW-1003-OPS-HSA	STATEMENT OF WORK SMART TM Follow-On Study Spacecraft State Representation through Pattern Recognition	1.2	

2.2. Reference Documents

The following table specifies the reference documents that shall be taken into account during project development.

Table 2: Reference documents

Reference	Code	Title	Issue	Date
[RD 1]	SMARTTM-DMS-PMD-MOM0004	SMART TM – Way forward for new approach	1.0	
[RD 2]	SMARTTM-DMS-TEC-TN01	TN on Task1 & Task2 Activities – Historical Telemetry Analysis	3.0	
[RD.3]	SMARTTM-DMS-TEC-TN02	TN on Task 3, Task 4 and Task 5 Activities - Telemetry Representation Approach & Results (Final Report)	2.0	

3. OVERVIEW

3.1. Background

Today's space missions greatly rely on the monitoring of simple housekeeping telemetry, which is often relatively limited due to bandwidth use limitations. The housekeeping information is received in the form of packets consisting mainly of reports of various on-board applications and raw values measured across the spacecraft systems.

These data are then made available to the user on dedicated MMI applications provided by the mission control system on the ground, which usually provide facilities for displaying simple numerical-based data extracted from these packets, in the form of snapshot alphanumeric displays, graphs plotted over a period of time and simple mimics which graphically represent subsystems on-board.

Based on these facilities, flight control teams need to analyse the behaviour of the spacecrafts and investigate the causes and effects of any anomaly or unexpected behaviour detected. Understanding the effects may lead to actions to minimize them; understanding the cause allows in some cases to avoid it from happening again in the future. The analysis of the housekeeping data received helps to identify the status of the spacecraft and to detect or prepare any actions devoted to prevent, correct or counter-act any possible deviation or problem that could be occurring at present or appear in the future.

The current approach for telemetry analysis is based on the knowledge, experience and intuition from engineers, helped to some extent by some ground processing software that could help detecting simple anomalous conditions, as values out of expected ranges, etc. This task supposes a very time-consuming and complex effort that has to be undertaken over mostly raw data. Furthermore, ground analysis is constrained by the level of observability provided by the spacecraft, which maybe limited due to bandwidth limitations and may not fit the ideal observability level for the mission.

On the other hand, spacecrafts are becoming more autonomous, being able to respond automatically to some anomalies and relieving the operations centres from the critical management of these "short response" conditions during routine operational phases. Operations activities are moving to more advanced diagnostic engineering and management of the spacecraft configurations to prolong mission lifetime and to maximise data return.

On this situation, studies are being performed in order to continue advancing in the autonomy of the spacecrafts and their observability on-ground, while at the same time reducing the effort necessary for repetitive tasks and automating some tasks in order to help the ground operation centres to focus on the most specialized activities.

In the frame of telemetry downlink improvements, studies are focused in analysing different methods (sometimes complementary) to obtain the objective of reducing the bandwidth needs of the housekeeping telemetry while at the same time augmenting the spacecraft observability. The SMART TM study pertains to the group of studies looking for an Intelligent Telemetry (ITM) approach and will investigate the possibility of analysing the on-board generated housekeeping data.

3.2. Objective

The aim of this study is to understand the contents of the telemetry, use this understanding to reduce the amount of data in the telemetry while leaving enough information so that on-board decisions about spacecraft state can be determined by another algorithm.

This global objective is decomposed in three more specific objectives:

- Get knowledge about the telemetry data which are currently being generated by the spacecraft.
- To determine if is possible to reduce the amount of data but maintaining the same level of information.
- Identify algorithms for determine the state of the spacecraft based on the telemetry data

One of the main restrictions of the study is that all the analyses should be done in a "blind" way i.e. without any a-priory assumption or knowledge about the data. In addition, the data should be analysed as whole i.e. it should not be any pre-classification, as for instance, separate the data per sub-systems.

4. WORK DONE

The starting point for the study development is a collocation at ESOC where some meetings were held with FCTs of the two missions selected for this study: GOCE and Mars Express (MEX). Therefore all the telemetry data analysed during the study comes from real telemetry data from those ESA missions.

4.1. Getting Knowledge on Telemetry Data

The initial approach for getting knowledge of the telemetry data, consists in perform several statistical analyses on the data in order get some indicators that provide knowledge about the data. Following figures show some of the results obtained.

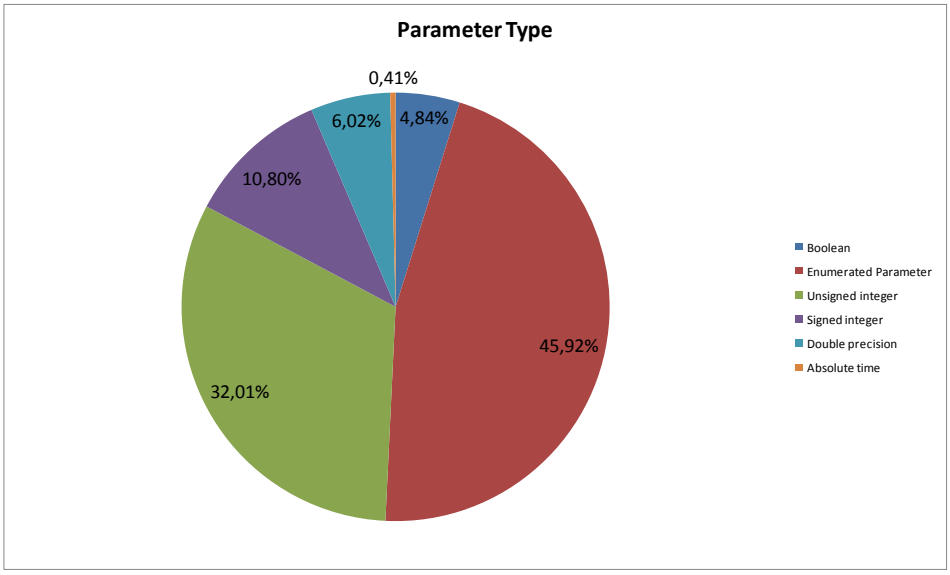

Figure 1: Distribution considering the type of parameter

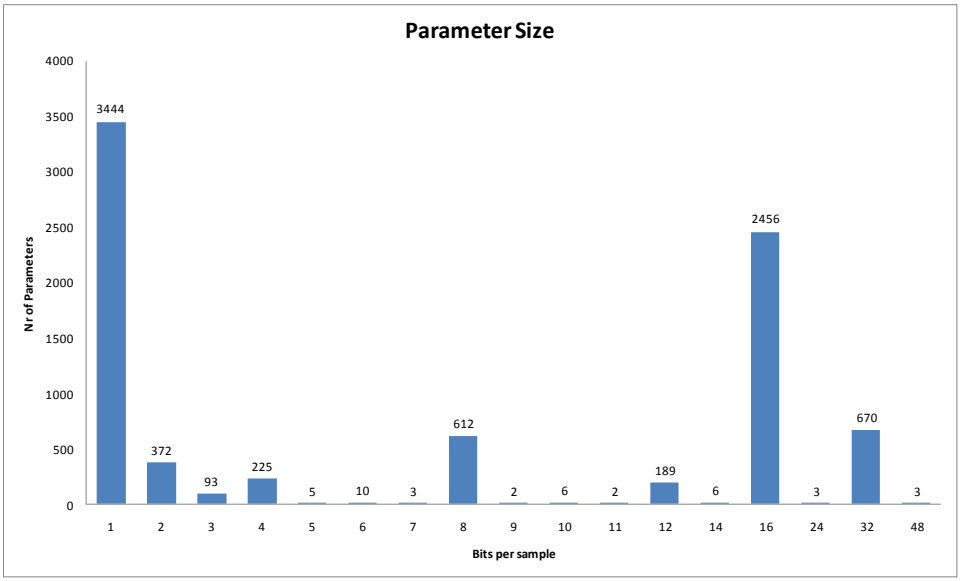

Figure 2: Distribution per Parameter size (in Bits)

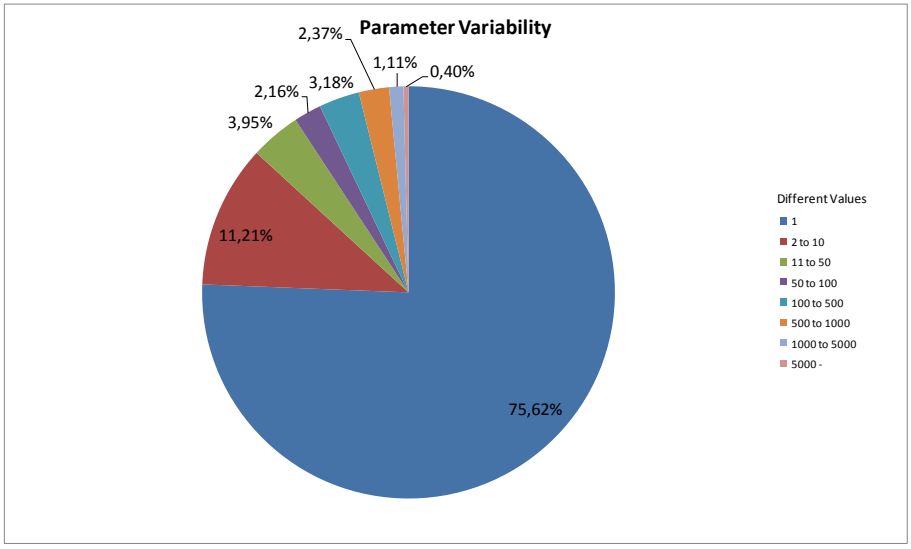

Figure 3: Parameters Variability

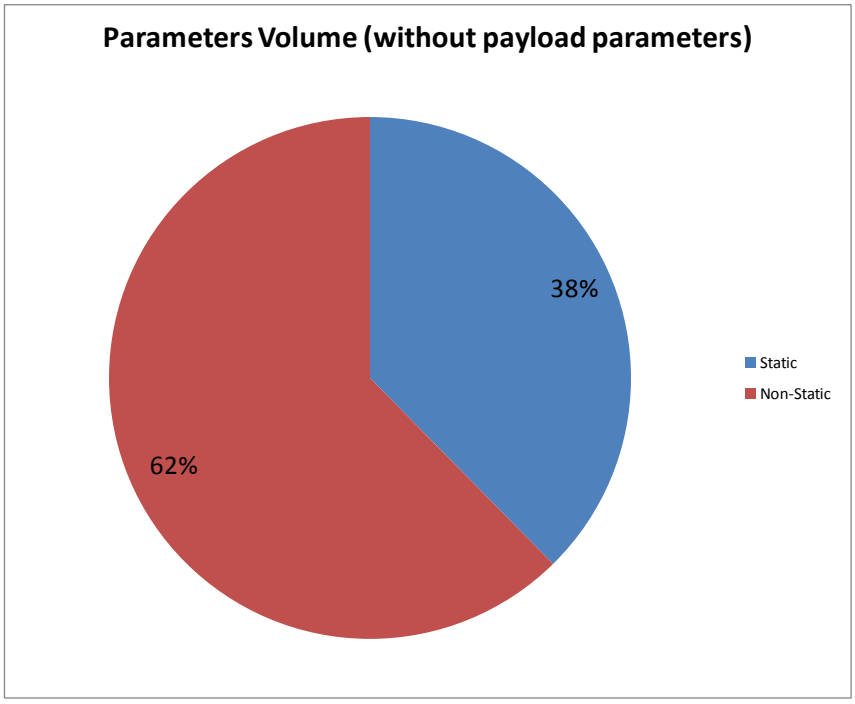

Figure 4: Distribution of volume between static and non-static parameters

From the figures above some observations could be done:

- ❑ Almost 50% of the data are Enumerated and only 5% are double precision (bigger size)
- ❑ Most of the data have only 1 bit size (42%), 16 bit data are about 30% and only 3 of the 8101 parameters have 48 bits length
- ❑ Most of the parameters of 1Bit size are Enumerated Type (88%)

The most important finding of this part of the analysis is that 75% of the parameters do not change at all during the sampled time. However, if this analysis is performed in terms of volume it can be observed that only 38% of the volume corresponds to those parameters that do not change (static parameters).

This means that the non-static parameters (those that change during the sample time) are sampled more frequently than the static parameters.

Those analyses were performed on two data sets from MEX mission with similar results in both cases, that confirm the results.

4.2. Reduction of Amount of Data

For reducing the amount of data several techniques were explored in order to evaluate their effectiveness for reducing the amount of data.

Several techniques were proposed as candidates to be used but finally the following techniques were evaluated:

- Optimal Re-Sampling of Time Series (Fractal Re-Sampling)
- Clustering
- Noise Evaluation
- Data Correlation
- Pattern Recognition

All the techniques above were explored on a set of data from MEX mission (MEX_1). In this first run the objective was to get results from all of them and after that, select those which obtain better results in terms of data reduction although other aspects like time consumption were also taken into account.

After this initial run three techniques were selected:

- Optimal Re-Sampling of Time Series (Fractal Re-Sampling)
- Hierarchical Clustering. For clustering techniques three different algorithms were evaluated initially: hierarchical, Gaussian Mixture and K-Means. From these three hierarchical clustering is selected to be re-evaluated given their positive results.
- Data Correlation

These selected techniques were then re-evaluated using an additional data set of MEX mission (MEX_2) and a data set from GOCE mission. Following main results of these techniques are presented.

4.2.1. Optimal Re-Sampling (Fractal Re-Sampling)

The algorithm for fractal re-sampling is a lossy compression method developed by ESA especially suitable for telemetry parameters. 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.

The algorithm is configured for allowing an error of 0.5%. The figure hereafter shows the results for compression ratio in all three data sets.

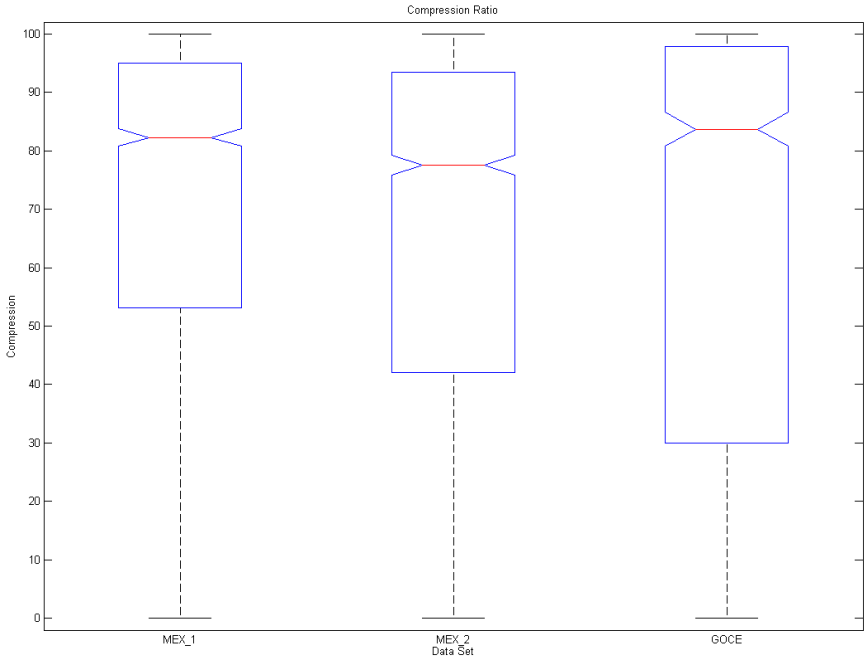

Figure 5: Compression Ratio Box and Whiskers (Optimal Re-Sampling)

The figure above shows that the MEX_1 and MEX_2 compression ratio has also similar behaviours, median values are 82.22% for MEX_1 and 77.47% for MEX_2.

The GOCE data set has a median of 83.69% which is in the same range than the other data sets, but the 25^{th} – 75^{th} percentile has a much wide range from 29% to 97% which shows that for this data set the compression ratio is more irregular. The maximum value in all cases in very close to 100% meaning that for some parameters in the three data sets, the algorithm gives a very good response, but on the other hand the minimum value is close to zero. That behaviour confirms the previous results where it was said that the compression obtained is highly dependent on the parameter itself.

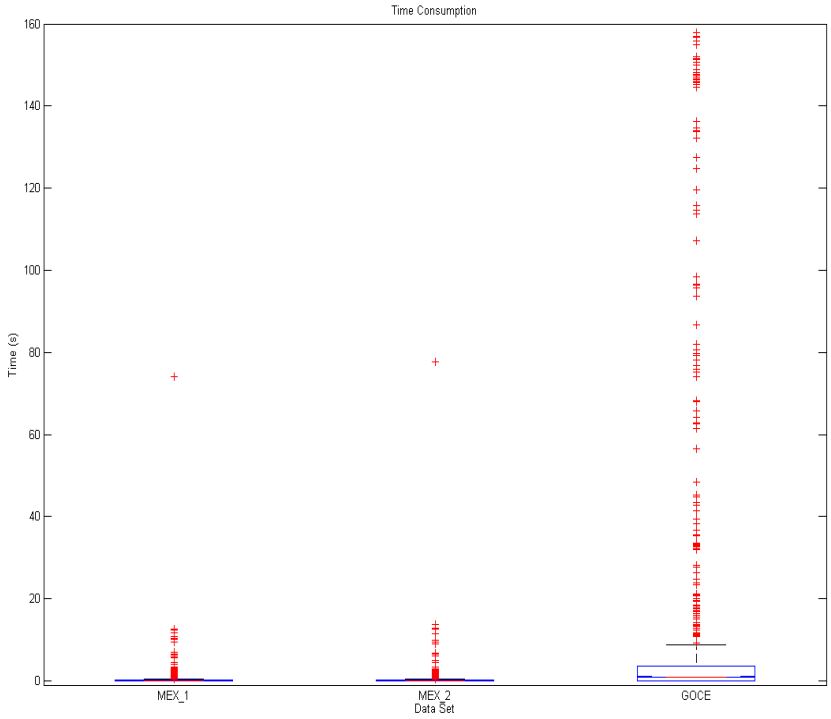

Figure 6: Time Consumption Box and Whiskers (Optimal Re-Sampling)

Analysing time consumption, it can be observed a big similarity between both MEX data sets, in both of them only one parameter has considerable major time consumption. The parameter is the same in both cases and is 'NACHKSDF' described as "SDF Place Holder for AOC".

In the GOCE data set, there are a larger number of parameters with high time consumptions, however the time consumed for most of the parameters is relative small. The median and mean values are similar for the three data sets: 0.049s for MEX_1, 0.049s for MEX_2 and 0.095s for GOCE.

The above results demonstrate that the Optimal Re-Sampling technique is very efficient in terms of data compression and time consumption.

4.2.2. 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. The groups can be obtained in top-bottom or bottom-top approach. In the first one, all the elements start in one single group (first level), then the most distant elements are separated in two groups, If the distance among the elements of a given cluster is greater than a maximum value, the elements of that second level cluster are separated again thus generating a new level. This process continues until all elements of the clusters have a maximum distance or until the maximum number of clusters is reached. In the bottom-top approach all the elements start as a cluster of one element, then they go up one level by joining the closer cluster, and again the process is repeated until the desired number of clusters is reached or until the distance between the elements of every cluster is within range.

An inconvenience of this technique is that it is not suitable for parameters of over 25.000 observations because the algorithm uses a matrix of M*(M-1)/2 items where M is the number of samples, for that reason and in order to be able to compare between data sets, all parameters were limited to 20000 samples.

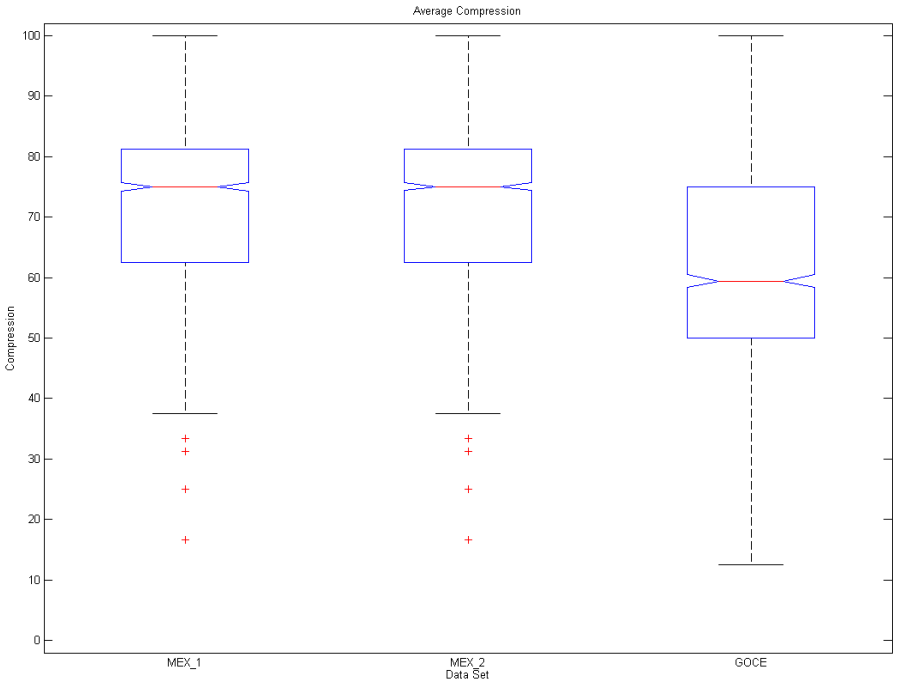

Figure 7: Compression ratio Box and Whiskers (Hierarchical Clustering)

The figure shows that for the two data sets of MEX the compression response is almost identical, the median value is 75%, also in both cases the blue boxes are in the interval of 62.5% to 81.25%, the maximum (99%) and minimum (16%) values are very similar as well.

For the GOCE data set the median value is 59% which means that the algorithm has a lower performance for this data set.

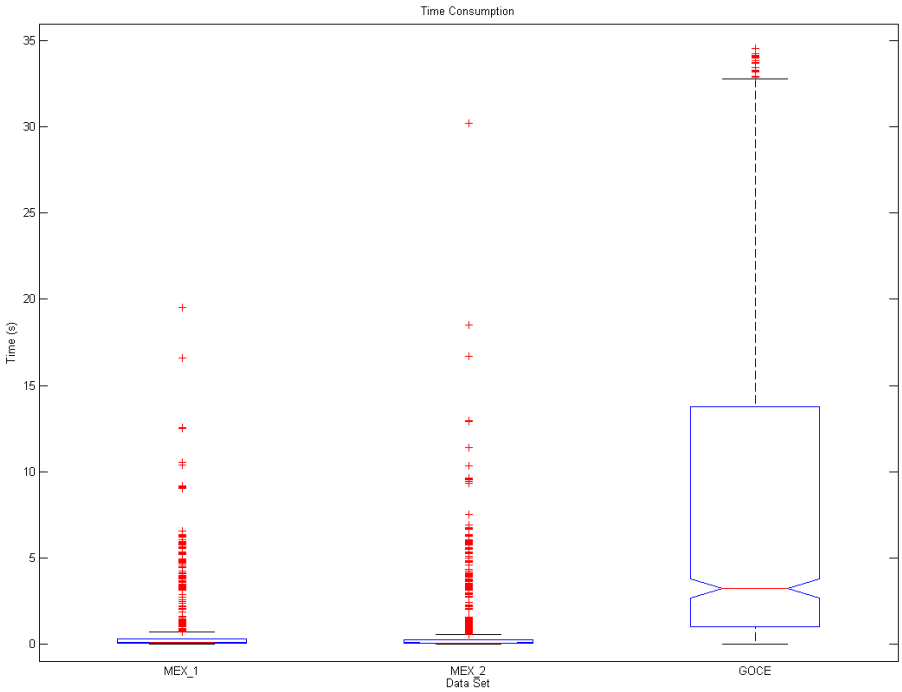

Figure 8: Time consumption Box and Whiskers (Hierarchical Clustering)

In terms of time consumption, the two data sets of the MEX mission have very similar figures while the statistics for GOCE shows longer execution time. The maximum time for MEX_2 is 30% bigger than MEX_2 (10 seconds approx.) and is the most significant difference, the median value is 0.04 seconds higher in MEX_1 (0.09s vs 0.05s), and the blue box that represents the range of 25% – 75% of the samples has a difference of 20% (0 - 0.287 s vs 0 – 0.232s).

4.2.3. Data Correlation

The correlation analysis was performed with the objective of finding significant relationships among parameters. This could lead to several interesting results, for instance it would allow identifying parameters with exactly the same behaviour and therefore reducing the volume of data transmitted to earth.

The analysis was made by a blind search, considering the correlation coefficient only for the non-static parameters obtaining a total of almost two million possible combinations.

The procedure for finding the correlations is divided in two steps, in the first part, all the parameters are re-sampled at a predefined frequency and in the same timestamps. In order to make this possible, the initial timestamp was obtained from the minimum time of all the parameters, denoted as time zero.

The second part of the procedure is the correlation analysis between all the possible couples of parameters, taking into account only the time intervals where both parameters are defined.

From each correlation analysis, two values P and R are obtained. The first one stands for the probability of finding the given correlation coefficient with a random time series. The second value 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.

A significant correlation coefficient (R > 0.2 or R < -0.2) was found for 25% (488840) of the possible combinations (1945378).

It is possible to take advantage of the high correlations found in this analysis in order to reduce the amount of parameters that have to be transmitted or stored, by identifying some "Master" parameters, with high number of correlations, and transmitting only those parameters knowing that the "dependent" or correlated parameters will have identical or almost identical behaviour.

A first approach for this case was developed, taking into account only the parameters with R = 1 or R = -1. For this case, the result was as following: 79.47% (1568) of the parameters have no correlation at this

level, so they have to be transmitted or stored; 6.59% (130) are the "Master" parameters which also have to be transmitted, and 13.94% (275) parameters are dependent from the master parameters, so a priory, there is no need to transmit those parameters.

If the tolerance is reduced, which means that the parameters with a correlation coefficient smaller than 1 are taken as valid, the reduction in number of parameters will be higher.

After this study, new tests were performed obtaining smaller values, for example with a very small decreasing of the correlation coefficient (R > 0.999 or R < -0.999) the reduction obtained is 34%.

The next figure shows the reduction in number of parameters when the correlation coefficient is changed for MEX_1 data set.

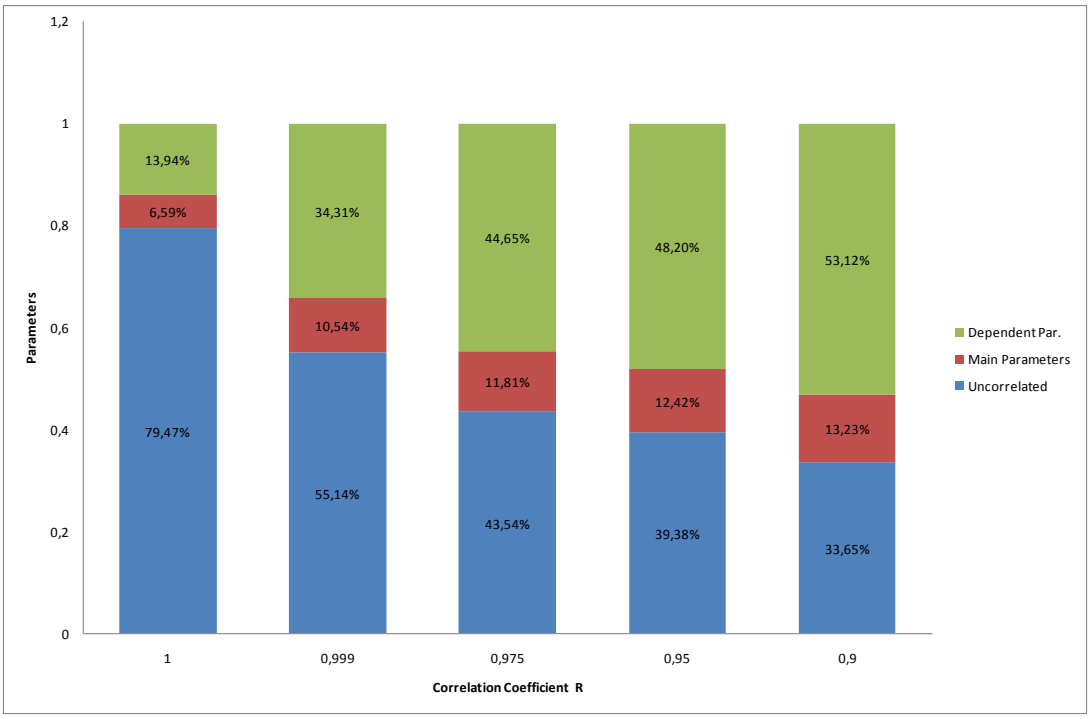

Figure 9: Correlation Coefficient Vs Number of Parameters

It can be observed that, with small reductions on the minimum correlation coefficient accepted, the number of dependent parameters increases, and with an R value over 0.975, almost one half of the parameters can be considered as dependent. It can also be observed that the amount of "Master" parameters doesn't change a lot. This means that there is an important group of parameters that can represent a big part of the status of the system.

The above described process was executed also for all three data sets: MEX_1, MEX_2 and GOCE with the following results. For this comparison the parameters with correlation coefficient greater than 0.975 are considered as identical, the next figure shows the reduction obtained for each set of parameters.

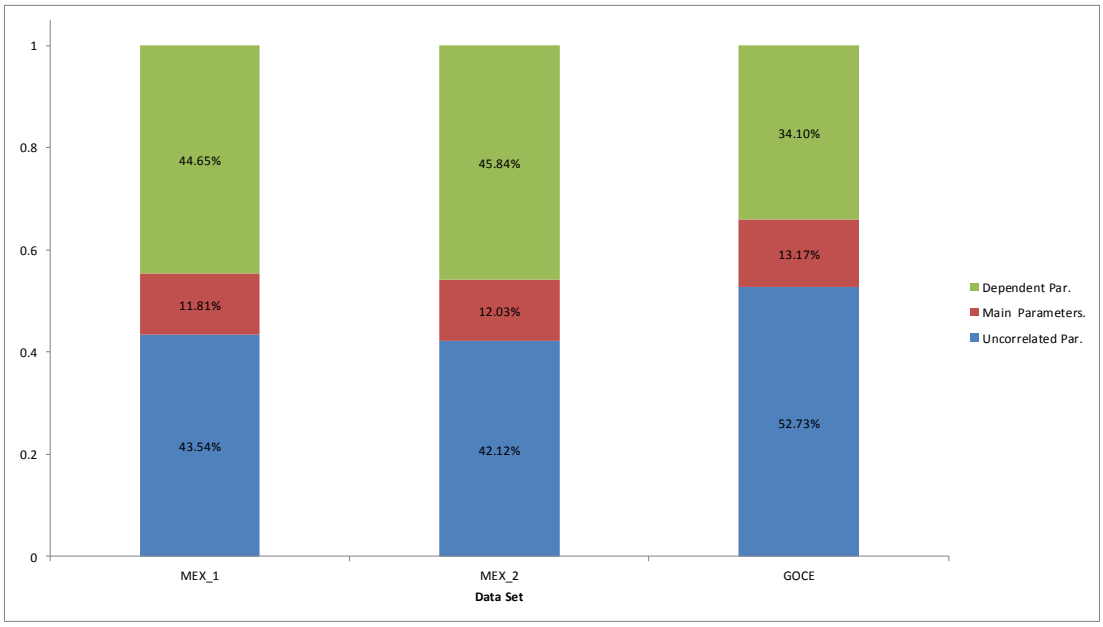

Figure 10: Comparison of the data reduction by correlation

It can be observed that the three percentages of main parameters are very similar between 11.8% and 13.1%. The number of uncorrelated parameters is more significant in the GOCE data (about a 10% more than in the MEX data sets), and that difference is reflected on the smaller quantity of dependent parameters.

The reduction shown in the Figure 10 is in terms of the number of parameters reduced. In order to compare it with the other two techinques that produce a reduction on the volume of data, it is neccesary to find the reduction for this technique also in terms of volume, to do so the volume of the main and uncorrelated parameters is calculated and compared with the total volume of the data set, the results of that conversion are shown in the next Figure.

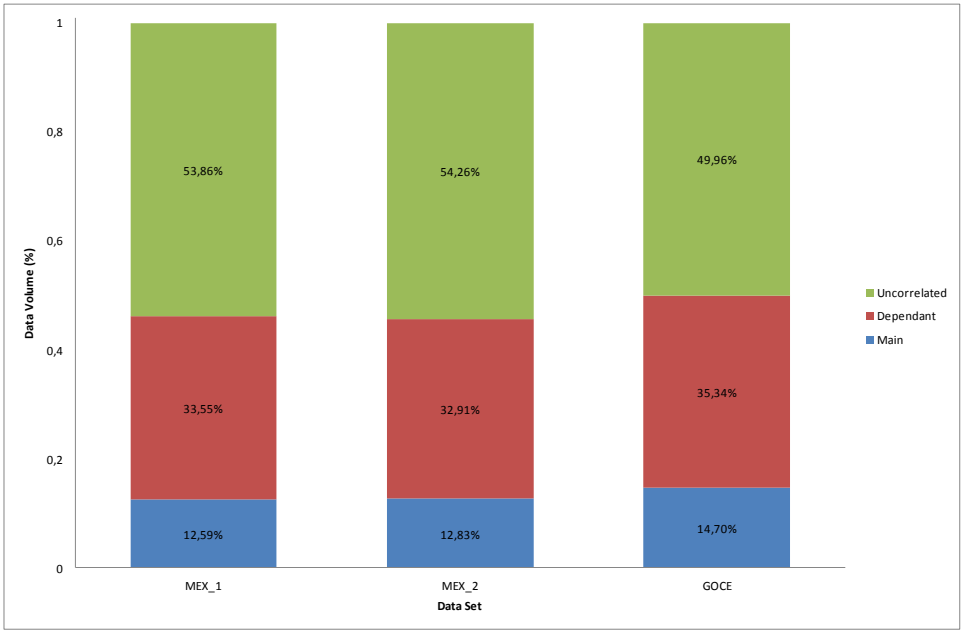

Figure 11: Comparison of the volume data reduction by correlation

From the figure above, it can be observed that the mean of volume reduction is 33.9%.

As a conclusion of this activity, it was found that it is possible to reduce the amount of data. Using the correlation for reducing the data volume, the volume reduction could be around 35% of the total data

volume. The compression ratio obtained with the other two techniques (Optimal Re-Sampling & Clustering) is better that with correlations. For instance, in GOCE data a reduction of 50% is obtained (setting the error to 0.07%), for MEX the reduction are above 76% with error of 0.05%.

4.3. Identifying the State of the Spacecraft

Based on the results from previous activities, the objective now is to explore different techniques for decision making that can determine the state of the spacecraft based on telemetry data.

After analysing the results it was decided that the most suitable techniques to be used for decision making are:

- Data correlation
- Clustering

Some algorithms are developed based in the above techniques to determine the state of the spacecraft.

For this Task, additional sets of telemetry data were used all of them of the same spacecraft:

- Four data sets of 24 hours each of nominal status (MEX_1, MEX_2, MEX_3, MEX_4)
- Two data sets (MEX_minor and MEX_critical) of 24 hours each of no-nominal status (i.e. including anomalies)

4.3.1. Decision making based on data correlation

Based on the fact that the correlation analysis showed that there are several parameters with high level of correlation with other parameters and that it is possible to identify some parameters as "master" parameters (i.e. parameters with a very strong correlation with the others). This means that the strong correlation between parameters could be used for detecting anomalies in the spacecraft.

Two possible uses of the data correlation analysis for decision making are tested:

- **Broken correlations.** The broken correlations approach consists in detecting broken correlations as a result of anomalies in the spacecraft. The hypothesis is that if something goes wrong in the spacecraft as there are so many strong correlations between parameters some of those correlations could be broken due to the bad functioning.
- **Master parameters variation**. This analysis consists on checking thee 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

A specific algorithm aimed to analyse and find differences between the nominal and no-nominal data sets were developed. The algorithm consists in:

1. First the correlation is calculated for one data set according with the previous analysis and the very high correlated combinations are extracted,
2. The procedure is repeated for other data sets available in which the status of the spacecraft is normal.
3. The next step is to find a group of correlations that are common to most of those data sets. Doing this, it is possible to obtain a set of correlations that can be used as reference set i.e. to have a "state vector" of the spacecraft. This "state vector" is used the next part of the study as a reference for looking for differences.

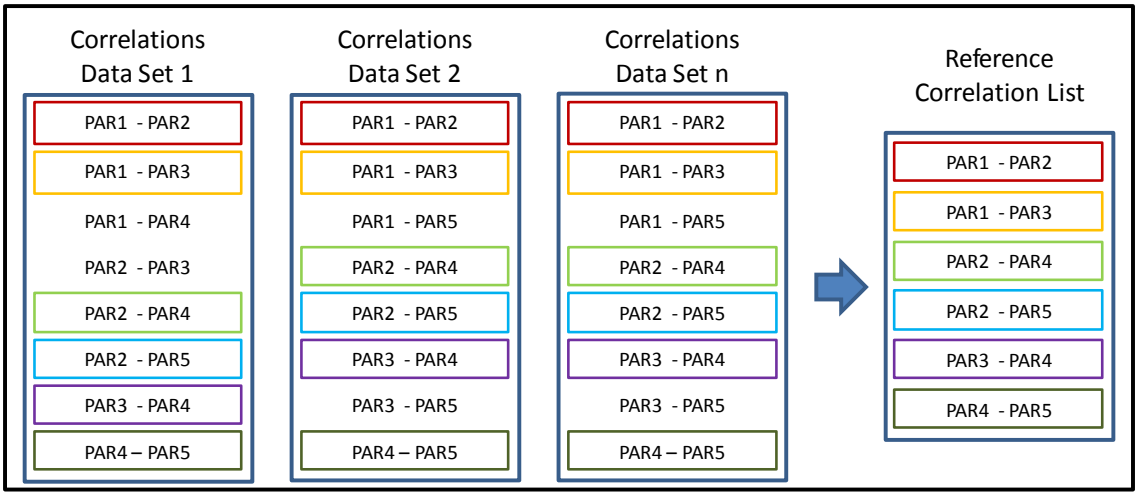

Figure 12: Graphical description of the method to find reference correlation list

For any new data set, it is analysed by checking if the correlations of the reference group are found also in the new data set, if the correlation is present in both of them that correlation is not taken into account. Therefore, only lost correlations are taken into account for the analysis.

Then the list of parameters involved in those lost correlations is extracted in order to check why the correlation is different in the new data set.

Finally, a list of the parameters involved in those lost correlations and the number of valid changes in the correlations for each one of the parameters is obtained. The parameters with higher number of changes are marked as relevant to the new status of the spacecraft. If the number of changes is not relevant then is possible to affirm that there is not major change in the status (See Figure 13).

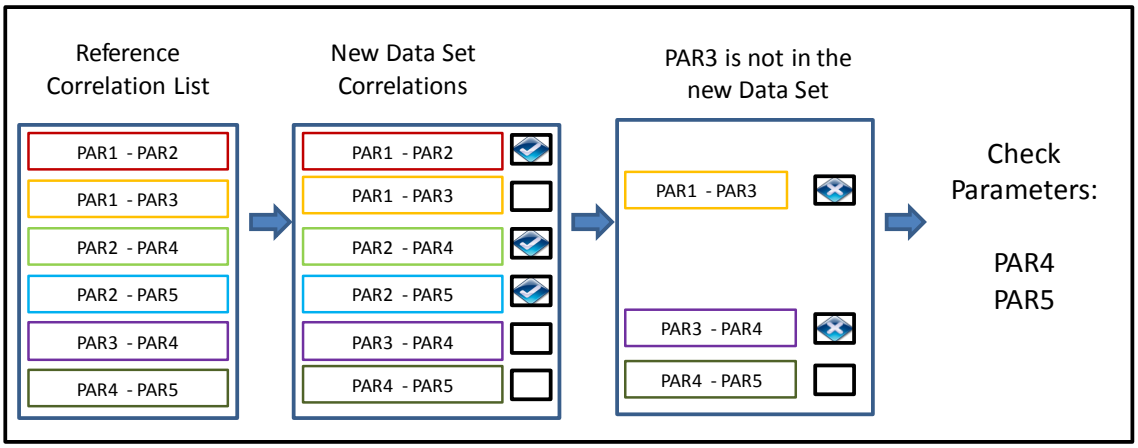

Figure 13: Graphical Description of the Correlation Analysis Algorithm

Decision Making based on "master" parameters variation

Another focus of the correlation study is oriented to find a list of "master" parameters and see if those parameters are able to represent the status of the spacecraft. To find those parameters simple criteria is used: the parameters with the highest number of high correlations are selected as "master" parameters, and the parameters correlated to that parameter are excluded.

The process used to obtain the list is described next:

- First, the list of very high correlations of each one of the six data sets obtained previously is reloaded
- From each group, the parameters involved in the correlations are listed together with the number of correlations in which a parameter appears. That list is sorted according to the number of correlations for each parameter

- The lists of parameters are compared in order to obtain a list of parameters that are common to all data sets.
- That list of parameters with high correlations in all of the data sets may contain parameters that are correlated to each other, so a final selection process has to be made to keep in the list only one parameter from a semi-group of parameters that are correlated to each other.

4.3.2. Decision making based on clustering

Two different algorithms for decision making based on clustering are tested:

- The first algorithm is based only on hierarchical clustering

  For this algorithm only the parameters that are common to all data sets will be taken into account, so the first step of the algorithm is to find the parameters that are common to all data sets, to do that an algorithm similar to the one used to find the master parameters in the correlation study was used, it starts with the parameters on the first data set and starts looking for it on the other data sets, if the parameter appears on all data sets it is included to the list of common parameters.

  After that the hierarchical clustering algorithm independently implemented on all the available data sets. The algorithm allows establishing either a fixed number of clusters or a fixed relative distance or "cut-off" value. In this case the cut-off value was the best selection because it allows to obtain a number of clusters depending on the characteristics of the parameter and it also gives clusters of the same size which will be useful in the second algorithm, the value for the cut-off was 0.7 because from the previous analysis it was obtained as a maximum value with good behaviour.

  For all parameters, the only outcome that will be used is the final number of clusters i.e. the centres or the size of the clusters are not be taken into account.

  The results of the clustering are divided in two groups of data sets, the first group is composed only of data sets on nominal conditions and will be used as reference, and the second group has both nominal and abnormal data sets and will be used for testing.

  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 (μ-σ) 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.

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

- Second algorithm is based on a combination of hierarchical and Gaussian clustering

  The second algorithm used to study the spacecraft status is based on a combination of hierarchical and Gaussian clustering. This combination allows to extract the best behaviour of each clustering algorithm.

  The hierarchical clustering gives the number and the center point of the clusters, but it works with a fixed and limited number of observations. The Gaussian clustering on the other hand is capable of working with new samples, for each new sample it gives a vector with the same size as the number of clusters and in each position the probability of belonging to that cluster, but it requires a pre-established clustering defined in terms of the mean value and variance of each 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, where the number of clusters and the mean value are given by the hierarchical clustering.

  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 value of that vector is equal or greater than an established threshold (0.9 in this case) it can be said that the observation belongs to that cluster and the a response value (R) of zero is assigned. If the maximum probability obtained is less than the threshold the observation is marked as

an outlier and a response value of 1/max(P) is assigned so the less the probability of being in a cluster the greater the response value. The sum of the response value is normalized by the number of observations so it can be compared between parameters or data sets with different number of samples in a given parameter. Also 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).

The results of applying the above described algorithms on the data available shows that that it is possible to identify parameters with abnormal behaviour by analysing the broken correlations although it is necessary to do additional statistical analyses.

One interesting result is that although the MEX_4 data set is considered as normal (without anomalies), the algorithm detects some type of abnormality. In this case a verification of the data should be done in order to check if really this data set belongs to a nominal situation or in contrast it is something that is causing that the algorithm is detecting it as abnormal (i.e. false positive)

The second approach, based on clustering techniques, found that Hierarchical Clustering could detect when a data set has anomalies. However, this result should be verified because it was not enough data to have definitive conclusions.

The combination of Heriarchical clustering and Gaussian clustering does not get expected results, although it is able to differentiate among nominal and no-nominal data-sets but it is necessary to perform some post-processing on the results.

5. CONCLUSIONS

5.1. Overall Summary

The logic work for the SMART TM study are clearly divided in three different stages:

- ❑ Get basic knowledge of the telemetry data (Task1)
- ❑ Explore different techniques for reducing the amount of data without losing information (Task 2 and Task 3)
- ❑ Develop algorithms for detecting anomalies in the spacecraft based on telemetry data (Task 4)

During the study development, the results obtained in each part has been used for define the following activity. All the studies were performed neither making any assumption about data nor getting information about the "physical" significance of each analysed parameter.

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%).

In addition, it was found that it could be possible to differentiate data sets with anomalies from data sets without anomalies. However, this result should be confirmed with additional studies.

5.2. Detailed Achievements

In each of the three above activities some key achievements have been produced:

- ❑ 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
  - Up to 5 different techniques have been explored for reducing the amount of data:
    - ➢ Optimal Re-Sampling
    - ➢ Clustering
    - ➢ Noise Evaluation
    - ➢ Data Correlation
    - ➢ Pattern recognition
  - From the techniques evaluated above , the three that obtained better results for reducing the amount of data were 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 new set from GOCE mission. The initial results were confirmed.
  - A total of 3 data sets have 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%. Table hereafter shows a summary of the results for compression.

The compression is calculated as the percentage of the data that will be not transmitted nor processed in relation with the total of data .

Table 3: Results for compression for each Technique / Data Set (in percentage)

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

- Algorithms for detecting anomalies in the spacecraft
  - For the techniques evaluated previously for reducing the amount of data, two of them 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.

5.3. Main Conclusions and Discussion

During the development of the study a big number of parameters have been processed using different techniques looking for possible data reduction (see Table 4). The results show that it is possible to obtain an important data reduction (around 70% in the best case).

Table 4: Summary of the parameters analysed during the study

	Received Time Series	Processed Time Series	Processed Registers
Data Set	Total Parameters	Non-Static/No Payload	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

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 and 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).

Taking into account that one of the main objectives of the study is to reduce the global amount of data managed during the spacecraft operations i.e. to reduce the amount of data managed during spacecraft operations in order to facilitate the diagnosis and recovery from anomalies, even more, to use this reduced amount of data for decision making on-board: any of the evaluated techniques could be used for this purpose.

For determining the state of the spacecraft, algorithms based on some of the techniques used previously have been developed for trying to discriminate between data sets with anomalies from data sets with nominal behaviour. Algorithms based on data correlation and clustering were developed.

Based on the result that a large number of parameters are highly correlated, this fact was used for try to identify anomalies by looking for broken correlations (it was supposed that when an anomaly occurs a high-level correlation could be broken).

The algorithms based on clustering (based on hierarchical clustering and Gaussian clustering plus hierarchical clustering), look for samples that could not be assigned to a predefined set of clusters that represents the "nominal state" of the spacecraft. If a sample is identified as that do not belongs to any of the defined clusters it is considered as a possible anomaly. The algorithms based on clustering have obtained good results in terms of differentiate between nominal data sets and the data sets with anomalies although it does not differentiate between the two data sets with anomalies.

Although the results seems to be good with clustering technique, they have to be reinforced by performing additional test using more data in order to have statistical foundation for the results.

The table hereafter shows a qualitative comparison of the performance of the different techniques for discriminating between data sets with anomalies from data sets with nominal behaviour. Some other evaluation/comparison criteria are added for checking the feasibility of the on-board implementation of each technique. The representation of the status is expressed as the ability of finding errors or variations between data sets using the given technique.

Table 5: Techniques performance and on-board feasibility

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

The table shows that the techniques based on clustering obtain better results for differentiate between nominal and no-nominal data sets of telemetry data. Correlation does not get conclusive results and in addition for on-board implementation the algorithm is high time consuming and requires important computing resources.

On the other hand, algorithms based on clustering are faster and depending on the implementation the computing resources are low if the clusters are previously created.

One of the main limitation of the work for decision making part is that it is necessary to rely on nominal data that represents the complete state of the spacecraft i.e. data that contains all the information relative for all the parameters representing the nominal behavior of the spacecraft (in addition it is necessary to have several sets of this type of data). Due to proper operations of the satellite and the mission planning it is common that some of the parameters are not sampled during some periods of time producing that different data sets contain different parameters although in all cases the satellite behavior is nominal. Those differences are difficult to manage from the point of view of the algorithmic for determining which is the reference point for identify abnormal behavior.

Even though the decision making part does not produce firm conclusions, the main results of the SMART TM study are considered successful taking into account the assumptions done for the study:

- It were not any a priory assumption about the data
- The search and algorithms were run in a "blind" way without introducing any knowledge about the functioning of the spacecraft but looking in the telemetry data as a whole.

Even with these "restrictions" it was found that it is possible compress the data using different techniques and that in some cases it could be possible to discriminate between data with and without anomalies.

The SMART TM study is part of the global NoC ITM study that develops the ITM concept for spacecraft telemetry data. The ITM concept has the main objective of optimize the observability of the spacecraft

behavior through reducing transmitted nominal housekeeping telemetry parameters and in certain cases to reduce it to an OK/NOK beacon concept. The SMART TM study provides some results that are useful for the global objective of the ITM concept:

- Using some of the techniques evaluated (e.g. optimal re-sampling) the data transmitted to ground could be reduced considerable. However, its implementation on-board is still pending to evaluate and prototype in order to check its feasibility.
- For determining the spacecraft state in a OK/NOK beacon, several algorithms were developed and tested, some of them with positive results although further tests should be done in order to confirm the results. It is also pending to check the feasibility for on-board implementation of these algorithms.

5.4. Future Work

Given the amount of data analysed, the number of techniques evaluated and the obtained results, it is clear that some additional work has to done for consolidate the main results obtained. It is considered that this study reinforces the ITM concept but some aspects have to be studied more in deep in order to consolidate the results.

Based on the results and the lessons learnt from this study, two main points should be the focus for future work:

- On-board implementation of algorithms/techniques for reducing the amount of telemetry data without losing information

  Regarding the on-board implementation of the techniques for reducing the amount of data, the work should be focused on optimal re-sampling and clustering techniques given that the correlation technique is high demanding in terms of computing resources. The other two techniques demonstrate that they are effective in terms of the level of data compression and time consumption, mainly the optimal re-sampling.

  Perhaps a combination of these techniques could obtain an optimal result: As described above, optimal re-sampling and clustering techniques achieve the compression acting over different aspects of the data i.e. optimal re-sampling reduces the amount of samples to represent the same time series signal and clustering reduces the number of bits necessary to represent the same information.

  Combining these techniques using the clustering for optimising the storage of the data on-board and using the optimal re-sampling for reducing the number of samples sent to the ground for re-build the signal, the sum of these techniques could reduce even more the data.

- Confirmation and fine tuning of clustering algorithms for differentiate between nominal and non-nominal behaviour

  One of the main aspects of the ITM concept is the possibility of automatically determine the spacecraft state at least in a OK/NOK form. Although the results obtained in the current study demonstrate that it is possible to differentiate between telemetry data with and without anomalies it is true also that it is not easy to do it with high confidence in the results.

  The spacecraft is such complex device that it is not easy to develop a deterministic algorithm that based on the telemetry data it could differentiate between a nominal and non-nominal situation. The future work should be centred in refine the proposed algorithms that obtained positive results (algorithms based on clustering technique).

End of Document