D2.2
Methodology report for handling of data heterogeneity

WORK PACKAGE 2 – Metrology, quality and harmonization

LEADING BENEFICIARY: CNR

Authors:	Beneficiary/Institution
Laura Beranzoli (laura.beranzoli@ingv.it), Mariagrazia De Caro (Mariagrazia.decaro@ingv.it), Caterina Montuori (caterina.montuori@ingv.it)	INGV
Vito Vitale (v.vitale@isac.cnr.it), Mauro Mazzola (m.mazzola@isac.cnr.it), Boyan Petkov (b.petkov@isac.cnr.it	CNR
Herve Petetin (Herve.Petetin@aero.obs-mip.fr)	OBSERVATOIRE MIDI-PYRÉNÉES
Justin Buck (juck@bodc.ac.uk)	NOCS
Catherine Lund Myhre (Cathrine.Lund.Myhre@nilu.no)	NILU

Accepted by: Jean-Daniel Paris (WP 2 leader)

Deliverable type: REPORT

Dissemination level: PUBLIC

Deliverable due date: 28.4.2017/M24

Actual Date of Submission: 2.5.2017/M25

ABSTRACT

The deliverable is related to the work developed in the task T2.2- *Time-series heterogeneities: innovative user services* (Task leader: INGV[EMSO] , Participants: IFREMER[EURO-ARGO], UiT[ESONET-VI] , UvA[LIFEWATCH], NERC[FixO3]) of Wp2.

The deliverable, after setting a common vocabulary and terminology, analyses the most recurrent sources of heterogeneity affecting the time-series in spite of the usual standardisation internal to Research Infrastructures and focus on some of them, namely data gaps and breaking points.

For the sake of a feasibility study on the application of methods for heterogeneities detection, time-series provided by Research Infrastructures have been classified in 'very-long time-series', lasting from one years to several years (typical of parameters related to global changes), and 'short time-series', lasting few seconds to some months (typical of parameter related to abrupt phenomena).

The methodologies used in the feasibility study on heterogeneities detection on time-series from Research Infrastructures have been borrowed by Geophysics. The computation of the Probability Density Function of the Power Spectral Density of the time-series is used for data gap detection and the computation of the ratio between the Short-Time Average and the Long-Time Average is shown as an example for breaking point (start time of heterogeneities) detection. A brief description of the methods is given together to basic references. The heterogeneity treatment issues are considered very dependent from the features of the corresponding parameters, site of measurement acquisition, modeling scale, and a trans-disciplinary approach to the various ENVRIplus time-series deserve a more deepen analysis.

The promising results obtained across disciplines and across domains support the proposal for the implementation of services to help scientists and data managers during the selection process of the most suitable data for their original elaborations. In shared virtual environments (i.e., cloud computing), the service can provide basic processing tools for time-series in different domains based on the proposed methodologies for heterogeneities detection. The service can be a very helpful option assisting data managers in the regular Quality Assessment/Quality Check procedures and support scientists in accepting/discarding /correcting data before the final data selection in view of original analytical elaborations.

Project internal reviewer(s):

Project internal reviewer(s):	**Beneficiary/Institution**
Ari Asmi	University of Helsinki

Document history:

Date	**Version**
15.04.2017	Draft for comments
25.04.2017	Corrected version
30.04.2017	Accepted by J.-D. Paris, V. Vitale, A. Asmi

DOCUMENT AMENDMENT PROCEDURE

Amendments, comments and suggestions should be sent to the authors (Laura Beranzoli, laura.beranzoli@ingv.it; Mariagrazia De Caro, mariagrazia.decaro@ingv.it ; Caterina Montuori, caterina.montuori@ingv.it)

TERMINOLOGY

A complete project glossary is provided online here: https://envriplus.manageprojects.com/s/text-documents/LFCMXHHCwS5hh

PROJECT SUMMARY

ENVRIplus is a Horizon 2020 project bringing together Environmental and Earth System Research Infrastructures, projects and networks together with technical specialist partners to create a more coherent, interdisciplinary and interoperable cluster of Environmental Research Infrastructures across Europe. It is driven by three overarching goals: 1) promoting cross-fertilization between infrastructures, 2) implementing innovative concepts and devices across RIs, and 3) facilitating research and innovation in the field of environment for an increasing number of users outside the RIs.

ENVRIplus aligns its activities to a core strategic plan where sharing multi-disciplinary expertise will be most effective. The project aims to improve Earth observation monitoring systems and strategies, including actions to improve harmonization and innovation, and generate common solutions to many shared information technology and data related challenges. It also seeks to harmonize policies for access and provide strategies for knowledge transfer amongst RIs. ENVRIplus develops guidelines to enhance trans-disciplinary use of data and data-products supported by applied use-cases involving RIs from different domains. The project coordinates actions to improve communication and cooperation, addressing Environmental RIs at all levels, from management to end-users, implementing RI-staff exchange programs, generating material for RI personnel, and proposing common strategic developments and actions for enhancing services to users and evaluating the socio-economic impacts.

ENVRIplus is expected to facilitate structuration and improve quality of services offered both within single RIs and at the pan-RI level. It promotes efficient and multi-disciplinary research offering new opportunities to users, new tools to RI managers and new communication strategies for environmental RI communities. The resulting solutions, services and other project outcomes are made available to all environmental RI initiatives, thus contributing to the development of a coherent European RI ecosystem.

TABLE OF CONTENTS

ABSTRACT 2

DOCUMENT AMENDMENT PROCEDURE 3

TERMINOLOGY 3

PROJECT SUMMARY 3

TABLE OF CONTENTS 4

1. INTRODUCTION, TERMS OF REFERENCE AND DEFINITIONS 5

2. TIME SERIES ACROSS RESEARCH INFRASTRUCTURES 6

2.1 'VERY-LONG' TIME SERIES 6

2.2 SHORT TIME SERIES 7

3. METHOD FOR HETEROGENEITIES DETECTION IN SEISMOLOGICAL DATA 8

4. FEASIBILITY TEST FOR THE APPLICATION OF THE SEISMOLOGICAL METHOD FOR DETECTING DATA HETEROGENEITIES 10

5. RESULTS OF APPLICATION OF THE DETECTION HETEROGENEITY METHOD TO ENVRIplus RIS 11

5.1 SIOS 11

5.2 IAGOS 13

6. NOTE ON BIOGEOCHEMICAL OCEANOGRAPHY DATA FROM FIXO3 18

7. PRELIMINARY OUTLINE OF AN INNOVATIVE SERVICE FOR SCIENTIFIC USERS 19

8. CONCLUSIONS 21

REFERENCES 22

APPENDICES I 23

1. INTRODUCTION, TERMS OF REFERENCE AND DEFINITIONS

There is not a common accepted meaning for *heterogeneous data* across Earth Science sectors so far. Meteorology itself, which has a long and well established tradition of analysis of very long-time-series, and is used to deal with sharp changes over time of the data acquisition features, first introduced a terminology, that is *homogeneous data,* with opposite meaning.

Then, in ENVRIplus the adoption of a cross-domain and cross-disciplinary meaning of *heterogeneous data* passes through the acceptance of the meaning of its opposite term.

According to Meteorology, a long-term time-series is defined homogeneous when all the variability and change is due to the behavior of the phenomenon (Anguilar et al. (2003), WMO World Climate Data Monitoring Program). The time-series showing variability not due to the behavior of the phenomenon related to the data, can be defined heterogeneous; these variability can bias a time-series and lead to misinterpretations.

Heterogeneities in very-long time-series can be due to many factors influencing the data as for example

- change of instrument/acquisition system/calibration,
- change of site of measurements,
- movement or relocation of instrument/acquisition system,
- change of the environment around the site of measurements,
- adoption of new methodologies for data processing.

Common heterogeneities in time-series can be

- gaps, i.e. missed data,
- offset and offset level change,
- outlying values respect to an 'expected ' range,
- change of the range of values (mostly range reduction).

The most interesting and pressing scientific questions that scientists are called to face and hopefully reply, deal with global scale and medium- long-term trend of various basic physical parameters: temperature, sea level, green house gas cycle. Thus homogeneity of data can be really crucial in some analysis (e.g., climate change analysis) since the data have to be representative of actual variation occurring over time. It is important, therefore, to remove the heterogeneities or at least determine the possible error they may cause. Heterogeneity, can be reduced applying procedures to detect and identify the errors made in the process of recording, manipulating, formatting, transmitting and archiving data.

To identify the heterogeneities in a time-series, complete and updated metadata are necessary: they describe the conditions in which data have been recorded and can help the reconstruction of the history of the data acquisition. Basically, the documentation of the data is a pre-requirement for any attempt to restore data homogeneity. In particular, a fundamental recommendation is to keep the record homogeneous through appropriate management of the observations site and associated equipment. This recommendation is however very difficult to follow for time-series last from year to decades.

A general scheme to assess and treat the data homogeneity is recommended to follow the steps below:

i. Metadata Analysis and Data Quality Control and Assessment
ii. Creation of a reference time-series (not always feasible or meaningful)
iii. Heterogeneity detection and breakpoint (initial time/sample of the heterogeneity) identification
iv. Data adjustment, 'treatment' of heterogeneities

ENVRIplus, through the involvement of Research Infrastructures (RIs) addressing the four domains of Solid Earth, Oceans, Atmosphere and Biodiversity, is able to offer a very considerable range of parameter time-series with a large variety of features that one could hardly think to treat by a unique tool for detecting heterogeneities. Nevertheless, a long and continuous record not biased by major heterogeneities, such as data gaps and data offset, is the most relevant requirement common to all RIs for any procedure of production of reliable predictive models. A useful service for the time-series users can thus rely upon the development of an analytical tool able to identify these major heterogeneities and suitable for massive analysis. To this aim a preliminary analysis of the main features of the time-series produced by ENVRIplus RIs and the comprehension of the typical use of the time-series is fundamental for the development of the analytical tool for heterogeneities detection.

2. TIME SERIES ACROSS RESEARCH INFRASTRUCTURES

ENVRIplus RIs produce essentially two different categories of time-series:

- 'very-long 'time-series , meaning time-series lasting from years to some (1-2) decades,
- 'short time-series' lasting from seconds to some months.

In general, the first category is related to long-term trends possibly related to global changes, while the second category is related to transient or seasonal processes at regional/local scales. The lack of data (gaps) for failure/external causes/change of equipments/change of site/change of environmental site condition has been considered as the most recurrent heterogeneity.

2.1 'VERY-LONG' TIME SERIES

The parameters considered of interest for different RIs are: Temperature (T), CH_4 and CO_2.

T, CH_4 and CO_2 from ICOS can provide examples of greenhouse gases from 1984 and the treatment of their heterogeneities could help to corroborate/disprove some controversial trends-

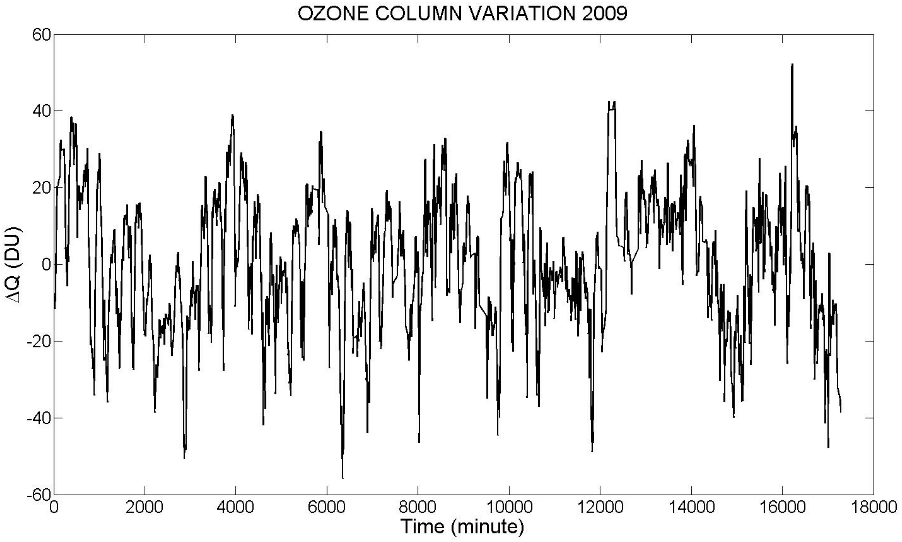

Fig.1: Example of 'very long' time-series: Ozone column variation acquired at Ny-Ålesund (78°56'N, 11°56'E, Elevation 10 msl) from SIOS infrastructure.

2.2 SHORT TIME SERIES

A typical example of short time-series is offered by geophysical data that are usually processed to detect fast event occurrence such as an earthquake. The occurrence of a fast event usually determines a sharp change in the characteristics of the geophysical time-series and the appearance of a heterogeneous signal (transient) in respect to the background signal level.

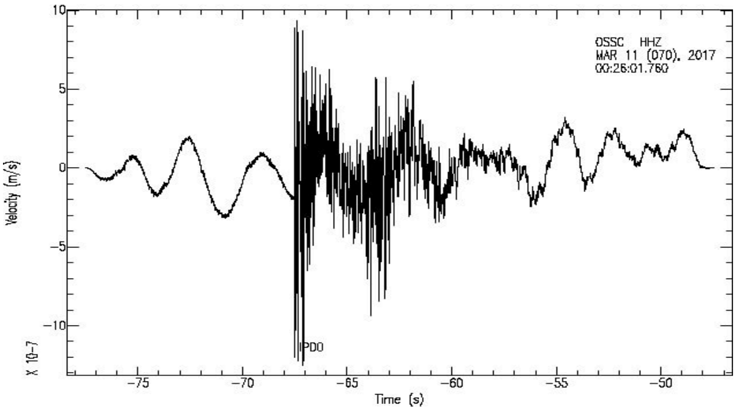

Fig.2. Example of seismogram (vertical ground velocity motion sampled at 100 Hz); the heterogeneity is a transient signal related to an earthquake of low magnitude recorded at short distance (15 km) from the epicenter and occurred in Tuscany (Italy) on 2017 March 11.

In Geophysics, namely in seismology, the detection of heterogeneities is thus the real first goal of the data processing and a well established analytical procedure has been developed and refined over time to make this procedure rapidly applicable also in real-time, that is simultaneously to the event occurrence. Here below we give a short description of this method.

3. METHOD FOR HETEROGENEITIES DETECTION IN SEISMOLOGICAL DATA

In the case of seismological data, a reference statistical model for the background signal, namely noise, is available thanks to Peterson (1993) who analysed the seismic noise spectra of a large number of worldwide seismometer land networks. Peterson's reference model is defined in the frequency domain and help to outline the main characteristics of the seismic noise worldwide (Figure. 3, left). The model is defined through the high-noise level curve and the low-noise level curve; thus the power spectral density of any seismological signal can be checked against this model in order to verify if it lays between the high and the low noise level curves. From the computation of the Probability Density Function (PDF) of the Power Spectral Density (PSD) of selected short time-series produced at the same recording station we can obtain the statistical significance of a any non-compliance with the reference model.

The algorithm for simultaneously computing PSD and PDF has been proposed by McNamara and Buland (2004) and McNamara et al. (2009). Following McNamara and Buland, each PSD is computed on one-hour time-series. Each hour-long time-series is divided into subsegments, ~4 min long and overlapping by 75% to reduce the variance of the final PSD estimates. Each subsegment is processed by

i. removing the mean and the long-period trend,
ii. applying a 10% cosine taper to reduce spectral leakage,
iii. computing the amplitude in the frequency domain using the fast Fourier transform (FFT) algorithm (Cooley and Tukey, 1965).

Subsegments are then averaged to provide a PSD for each one-hour long segment of the time-series. The PSD is converted into decibels (dB) with respect to acceleration so it can be compared with the high noise model of Peterson. In the last step, frequency distributions are constructed by gathering individual PSDs in the following steps: by binning periods in 1/8 octave intervals and by binning power in 1 dB intervals. Finally, each raw frequency distribution bin is normalized by the total number of PSDs to construct a PDF. A bin is defined as a given time interval including the original data values which, according to the described procedure, are replaced by a value representative of that interval, often the central value. The binning is a form of quantisation.

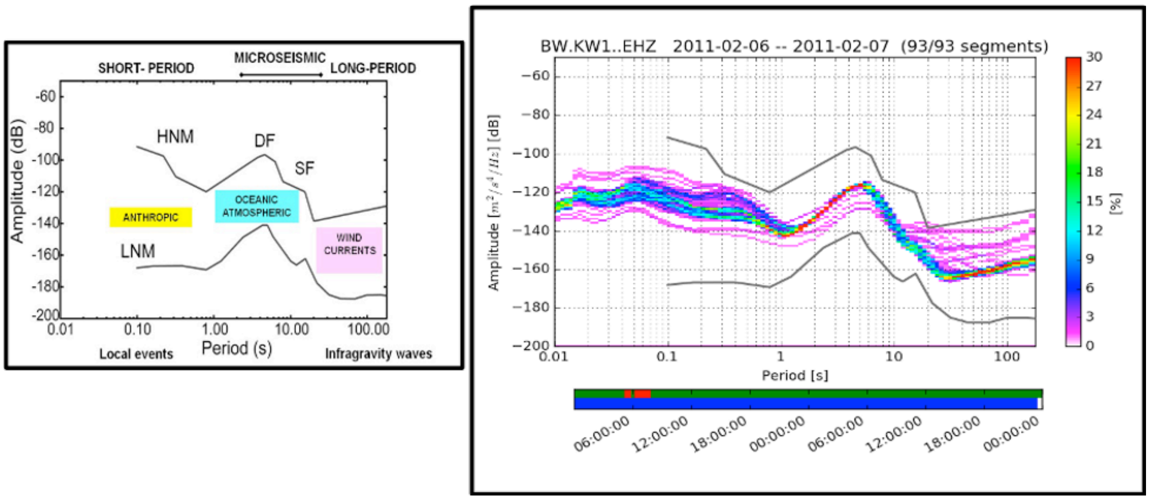

Fig.3. Example of geophysical (seismological) short time-series reference model: (left panel, black lines) the high-noise model (HNM) and low-noise model (LNM) of Peterson (1993) are reported delimiting the common values for the spectral amplitude of a seismological signal. The model give account of dominant sources of signal in the seismic spectrum as indicated in their corresponding frequency bands, spanning from infragravity waves >25 s (< 0:04 Hz) to local seismicity 0.05–0.25 s (4–20 Hz); (right panel) example of PDF constructed with 93 segments on vertical ground motion seismic component. The amplitude is given in units of decibels with respect to acceleration (m^2s^{-4}/Hz). The side color bar represents the probability of occurrence of each PSD normalized from 0% to 30%.
In the bottom right bar , dark green represents available data, red patches represent data gaps in the data streams.

Eventual artifacts (e.g outliers see Figure. 4) as well as breaking points can be accurately identified also in the time domain in order to clean up the signal. To this aim, algorithms have been developed using the ratio between the Short Time Average amplitude and the Long Time Average amplitude (STA/LTA) of a time-series. The short term average amplitude (STA) is sensitive to the transient/heterogeneity while the long term average (LTA) provides information about the temporal amplitude of the seismic noise at the site. When the STA/LTA ratio exceeds a pre-set value, the occurrence of a transient is 'declared'. This algorithm is used in seismology to detect the occurrence of transient signals connected to the occurrence of earthquakes: the time of the signal passing the pre-set threshold for STA/LTA is the earthquake arrival time at the recording station (Trnkoczy, 2002) .

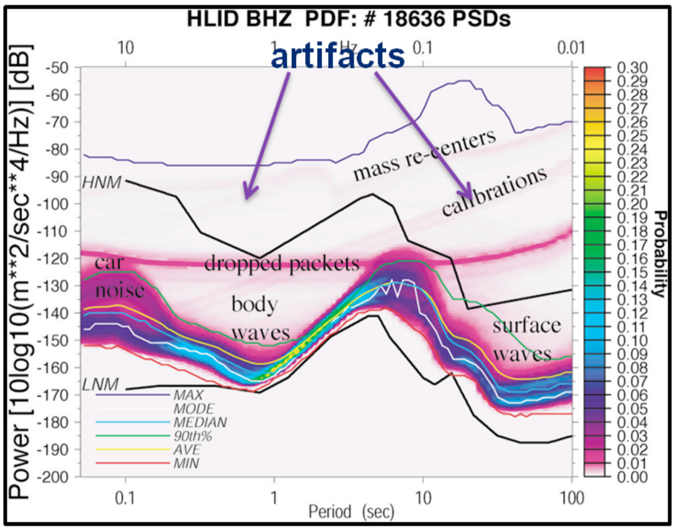

Fig.4: Example of PDF for a seismic land station (HLID, McNamara 2004). The PDF is computed using 18,636 PSDs of the vertical seismic ground motion during the period from September 2000 to September 2003. The Peterson model is also reported (black lines). Artifacts introduced by recording systems, considered as transients, are caused by data packets dropping and are recognizable in the PDF as low probability occurrence in respect to typical seismic noise, and do not show the typical trend of seismic noise.

4. FEASIBILITY TEST FOR THE APPLICATION OF THE SEISMOLOGICAL METHOD FOR DETECTING DATA HETEROGENEITIES

The applicability of PDF algorithm of the PSD has been proposed for very-long time-series provided by various ENVRIplus RIs.

The most important requirement for the applicability of the methodology is the definition of a reference *statistical model* (here after referred as to *model*) that is a typical behavior, for homogeneous time-series related to a selected physical parameter and a selected site. This requirement is not a trivial assumption and the possibility to identify a reference *model* has to be carefully explored.

In the case that a reference *model* is already available from the literature, it has to be studied in advance and eventual recurrent trends, seasonal variability, solar radiation dependent trends have to be known in advance in order to avoid misinterpretations of the data heterogeneities. Differently, if a local/regional reference *model* does not exist, the possibility to create a *model* 'on purpose' for the sake of heterogeneity detection has to be verified. A way to create a reference *model* is by considering representative values of noise during quiet and noisy periods from the time-series selected for the heterogeneity test. In fact, the correct approach for building the *model* is to skip data which are known to show particular natural events or other transient (natural or technical) phenomena and examine a sufficient number of time intervals from stations able to represent a consistent trend of noise on the desired scale (global, regional or local).

If a *model* can be computed for a given type of time-series, it can be used to check the compliance of any time-series of the same type. Any deviation from the *model* in term of probability can be considered a potential heterogeneity.

In order to start the feasibility test on the method, the collection of data from Ris was solicited during the 3rd ENVRIplus week in Prague (Nov. 2016).
The following RIs were available to provide data to test the applicability of the proposed method.

Infrastructure	Time series param.	Istitution	Key scientists
SIOS	Solar Radiation UV-B Ozone oscillations	CNR	Mauro Mazzola, Boyan Petkov
IAGOS	O_3 and CO data	CNRS	Susanne Rohs , Herve Petetin
ACTRIS	Aerosol absorption coeff.	NILU	Cathrine Lund Myhre
FixO3/EMSO	T	NOCS	Justin Buck

5. RESULTS OF APPLICATION OF THE DETECTION HETEROGENEITY METHOD TO ENVRIplus RIS

5.1 SIOS

Ozone oscillations

The daily Ozone oscillations time-series performed by UV-RAD at Ny-Ålesund (78°56'N, 11°56'E, Elevation 10 msl) of SIOS infrastructure are considered for the test. The possibility to generate a reference *model* has been explored. In Figure 4 the Power Spectral Densities (PSD) computed on daily Ozone oscillations time-series are showed. The time-series were acquired with a sample frequency of one sample per 5 minute in different period of 2009 (green line), 2010 (blue line), 2011 (pink line). The curves show to follow a common trend that could define a stable reference noise *model* for Ny-Ålesund site.

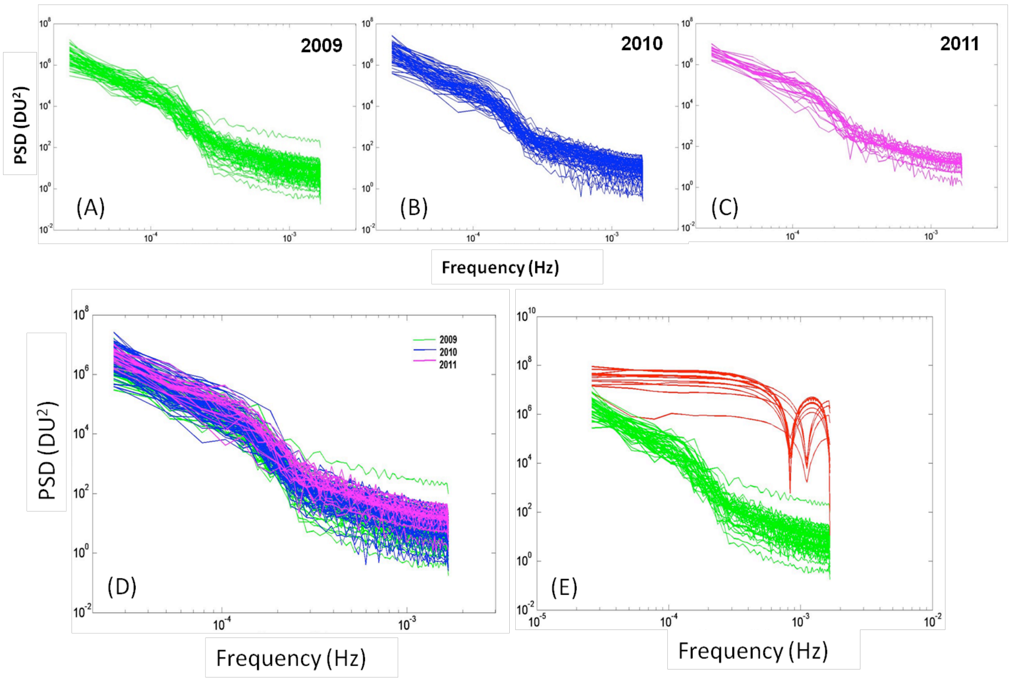

Fig.5: : Power Spectral Density (PSD) computed on daily Ozone Oscillations time-series performed by UV-RAD at Ny-Ålesund (78°56'N, 11°56'E, Elevation 10 msl) from SIOS infrastructure. The time-series were acquired with a sample frequency of one sample per 5 minute. (A) Green lines represent PSD computed on 2009 while blue (B) and pink lines (C) represent PSD computed on 2010 and 2011, respectively. (D) Shows all the top panels in the same frame. In (E) the red lines represent the PSD computed on 2009 daily time-series corrupted with outlier values.

The curves computed for the three periods (Figure5D) follow a common trend that could define a stable reference noise *model* for Ny-Ålesund infrastructure site.

The observed amplitude of the time-series are randomly replaced by 999, a number simulating an outlier value without physical meaning. The PSD calculated on the time-series windows with these outliers values appears with different features from the other ones. The PSDs have different trend and higher amplitude as shown in Figure 5E.

Solar Radiation UV-B

The same procedure has been applied for the doses solar UV-B time-series. The time-series have been acquired from 2005/04/27 to 2015/12/31 with a sample frequency of one sample per day. The sequence contains many missing value of several days.

The Figure 6 shows the Power Spectral Density (PSD) computed on two months of doses solar UV-B time-series. Black lines represent PSD estimated on UV-B time-series without gaps. The PSD define a reference noise *model* for UV-B. In the same figure, blue lines represent PSD computed on time-series with gaps and in which the missing values were filled with zero (zero padding is a common function of acquisition software) while the red lines show the result obtained by filling the missing values with an out of range value (in this case 9999999). The bottom panel in the

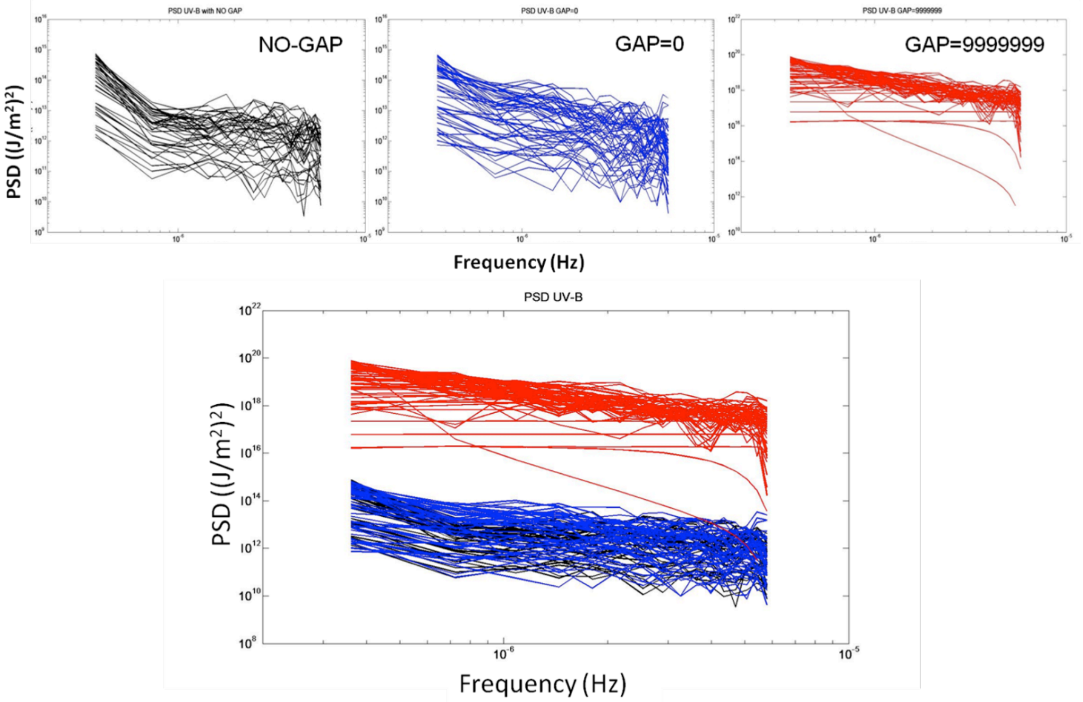

Fig.6: Power Spectral Density (PSD) computed on two months of doses solar UV-B time-series performed by UV-RAD at Bologna (44°31'N, 11°20'E, Elevation 50 msl) from SIOS infrastructure.

figure shows all the top panels in the same frame. The spectral analysis suggests that data gaps are better highlighted when the data samples in the gaps are replaced by an out of range values.

The time-series were acquired from 2005/04/27 to 2015/12/31 with a sample frequency of one sample per day. The sequence contains many missing value of several days. Black lines represent PSD estimated on UV-B time-series without gaps. Blue lines represent PSD computed on time-series in which the missing values are filled with zero while the red lines show the result obtained by filling the missing values with an out of range value (9999999). The figure at the bottom shows all the top panels in the same frame.

The spectral analysis suggests that the best way to highlight missing data is to fill the missing values with out of range values.

5.2 IAGOS

The possibility to get a reference *model* has been explored for ozone (O_3) and carbon monoxide (CO) time-series measured in three Troposphere layers (LT Lower- Ttroposphere, MT Mid-Troposphere, UT Upper Troposphere), from IAGOS infrastructure. Figures 7-8-9 show the Power Spectral Density results for O_3 time-series while the Figures 10-11-12 show the PSD results for CO time-series.
All the data have been acquired from 1st August 1994, with a sample frequency of one sample per day.

The spectra analysis suggest that a reference noise *model* could be identified for both O_3 and CO data in each troposphere layer. The spectral curves estimated on the time-series without missing values could be considered as an average trend for the possible noise *model* in which the spectra estimated on heterogeneous time-series can refer.

The main heterogeneity in the time-series from IAGOS is the lack of data that can either be non-existing values, or zero values or out of range value. Missing values are filled with an out of range value. This procedure is confirmed to be a good choice when it decide to use the PSD method for the heterogeneity detection.

O_3 measured in the Lower Troposphere

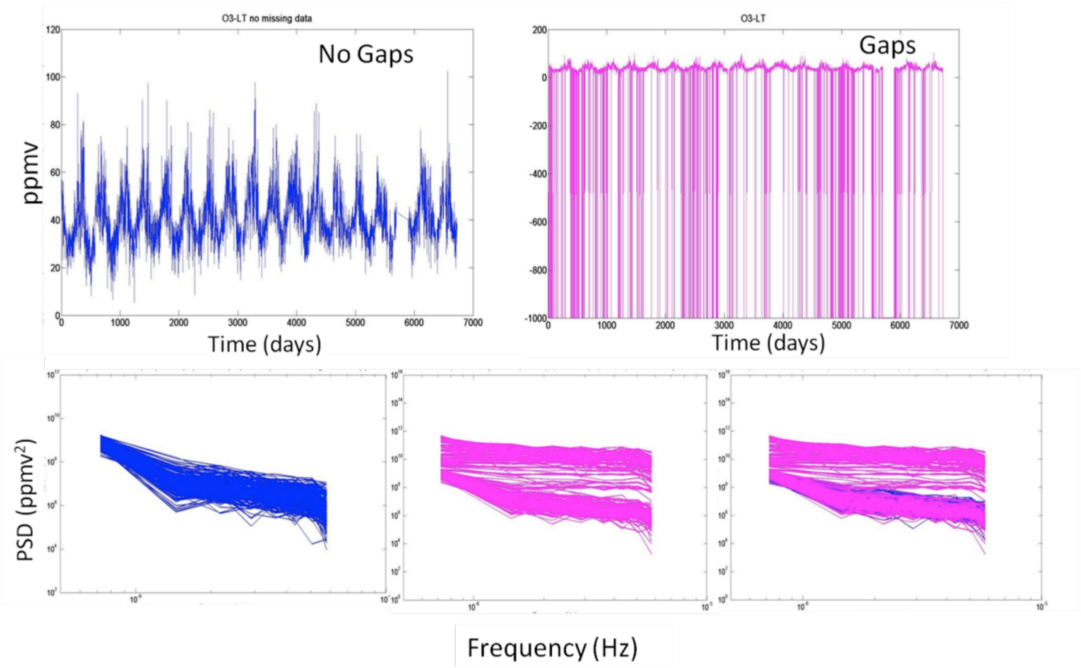

Fig.7: Power Spectral Density (PSD) computed on ozone (O_3) time-series measured in the Lower Troposphere (LT) from IAGOS infrastructure.

The time-series were acquired from 1st August 1994, with a sample frequency of one sample per day. Blue lines represent PSD estimated on O_3 time-series without gaps. Pink lines represent PSD computed on time-series in which the missing values are filled with -999.

O_3 measured in the Mid Troposphere

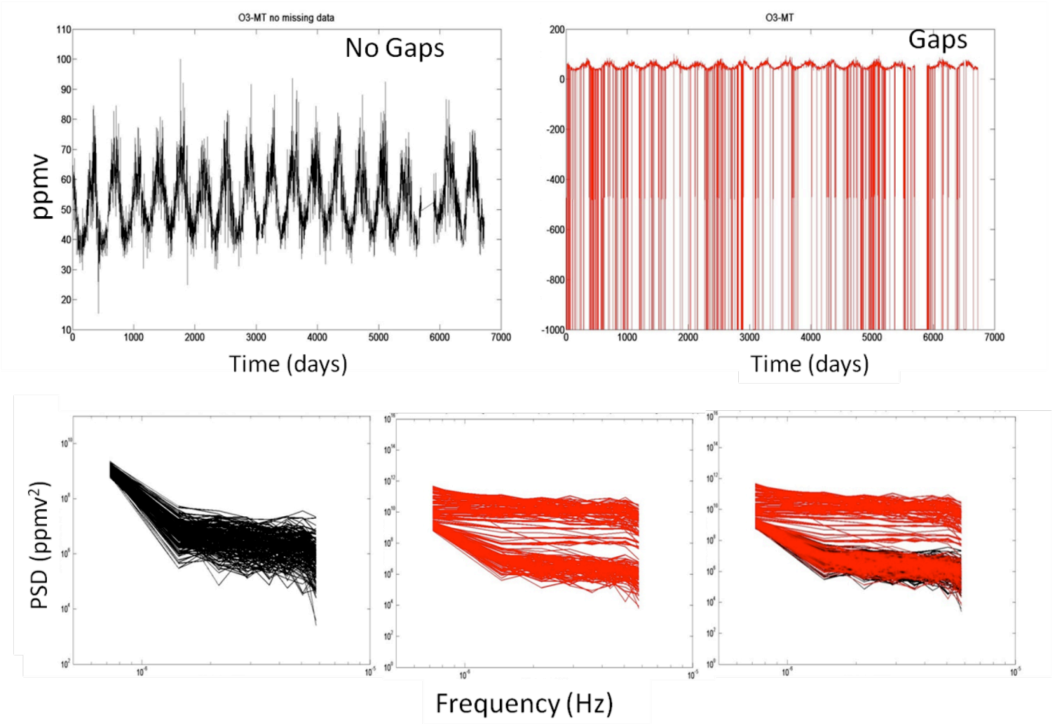

Fig. 8: Power Spectral Density (PSD) computed on ozone (O_3) time-series measured in the Mid Troposphere (MT) from IAGOS infrastructure. The time-series were acquired from 1st August 1994, with a sample frequency of one sample per day. Black lines represent PSD estimated on O_3 time-series without gaps. Red lines represent PSD computed on time-series in which the missing values are filled with -999.

O_3 measured in the Upper Troposphere

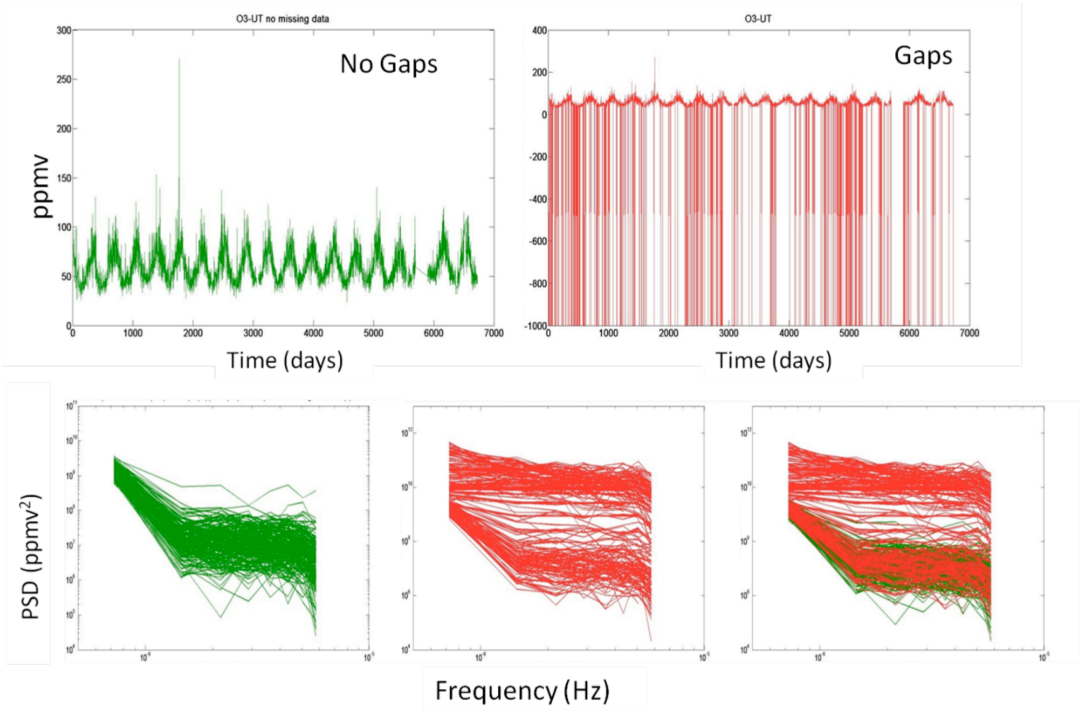

Fig.9: Power Spectral Density (PSD) computed on ozone (O_3) time-series measured in the Upper Troposphere (UT) from IAGOS infrastructure.

The time-series were acquired from 1st august 1994, with a sample frequency of one sample per day. Green lines represent PSD estimated on O_3 time-series without gaps. Red lines represent PSD computed on time-series in which the missing values are filled with -999.

CO measured in the Lower Troposphere

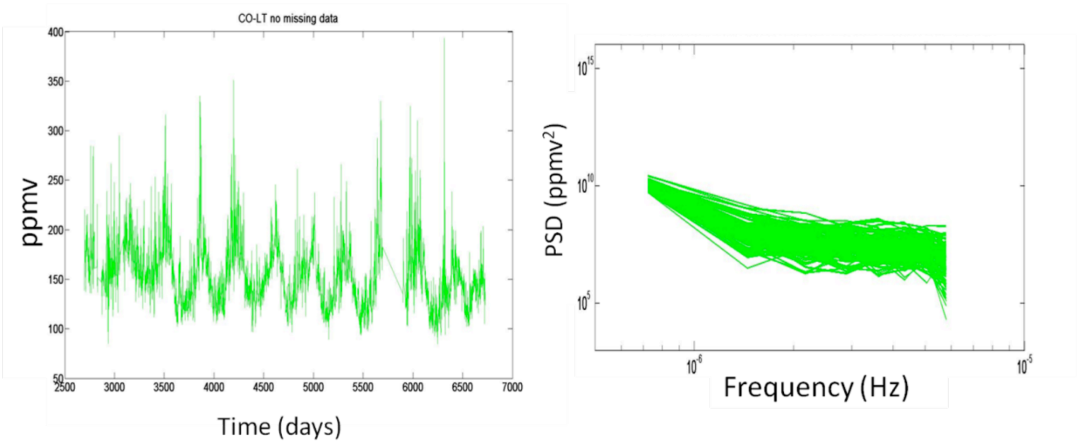

Fig. 10: Power Spectral Density (PSD) computed on carbon monoxide (CO) time-series measured in the Lower Troposphere (LT) belonging to IAGOS infrastructure. The time-series were acquired from 1st August 1994, with a sample frequency of one sample per day. The sequence contains no gaps.

CO measured in the Mid Troposphere

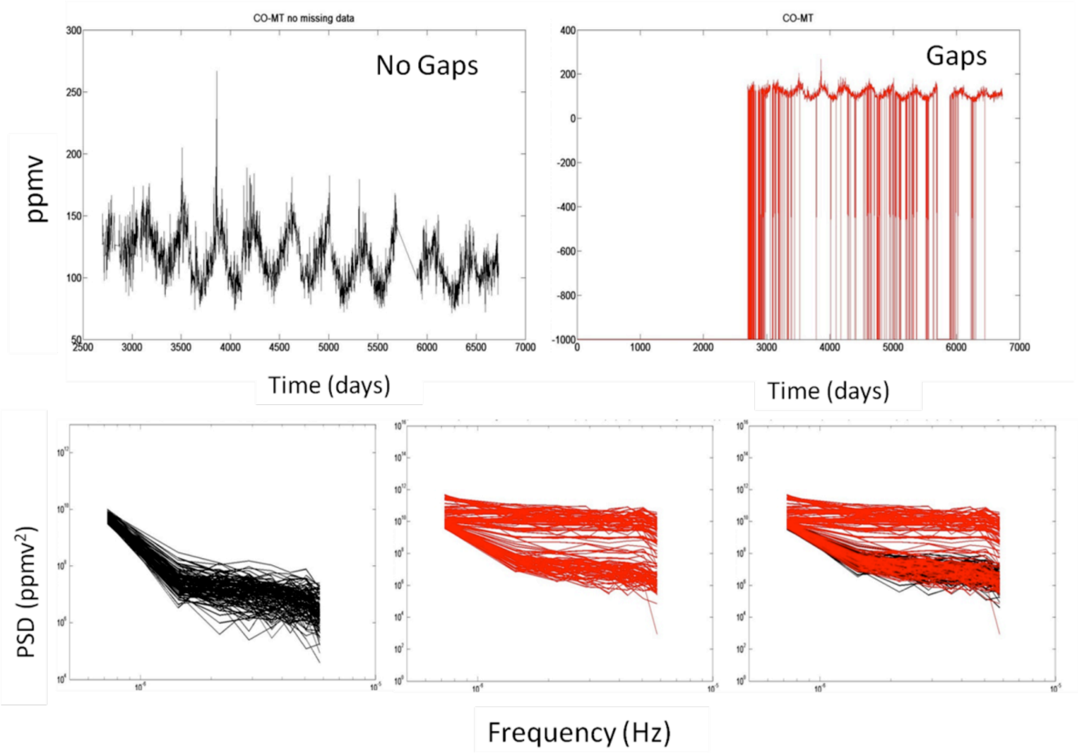

Fig. 11: Power Spectral Density (PSD) computed on carbon monoxide (CO) time-series measured in the Mid Troposphere (MT) belonging to IAGOS infrastructure.
The time-series were acquired from 1st August 1994, with a sample frequency of one sample per day. Black lines represent PSD estimated on CO time-series without gaps. Red lines represent PSD computed on time-series in which the missing values are filled with -999.

CO measured in the Upper Troposphere

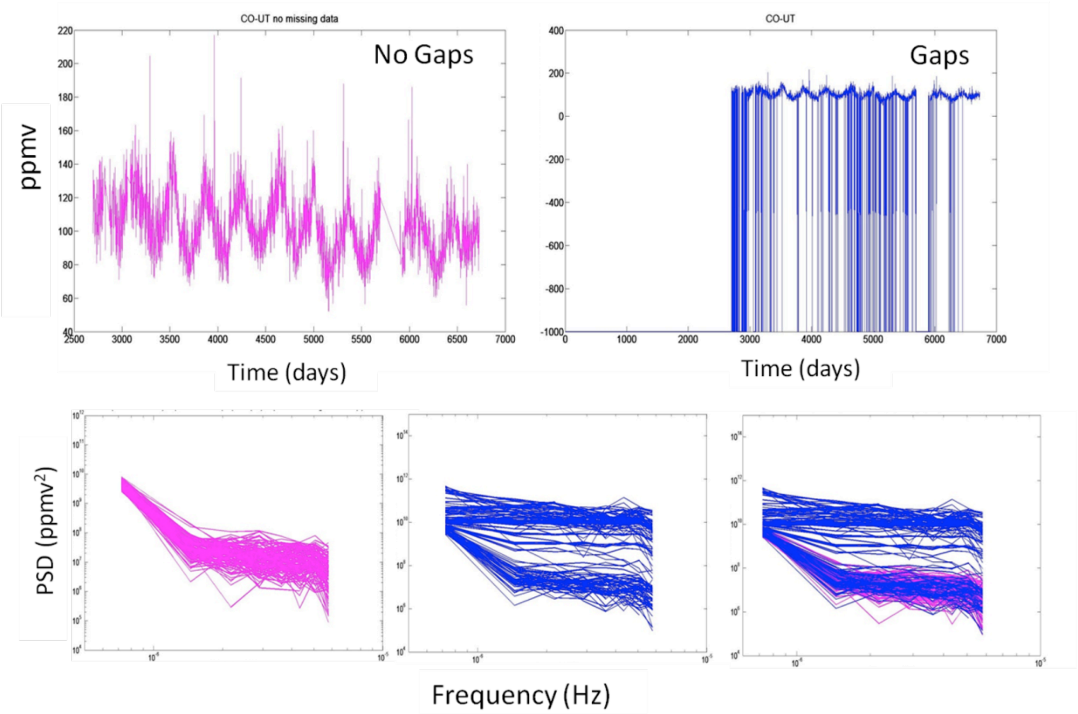

Fig. 12 Power Spectral Density (PSD) computed on carbon monoxide (CO) time-series measured in the Upper Troposphere (UT) belonging to IAGOS infrastructure.
The time-series were acquired from 1st August 1994, with a sample frequency of one sample per day. Pink lines represent PSD estimated on CO time-series without gaps. Blue lines represent PSD computed on time-series in which the missing values are filled with -999.

6. NOTE ON DATA FROM FIXO3/EMSO AND FROM ACRTIS

The application of the described approach to marine data is on-going at the time of submission of this report. Identification of temperature or CH_4 time series data that meet the criteria specified in Appendix I has proven challenging. Sensor models and types are rapidly developing meaning most long time series do not meet criteria "*v) the time-series has to be produced by a unique equipment*". An example of a long time series where the instrumentation has remained unchanged is the Continuous Plankton Recorder (CPR) run by SAPHOS in the UK. CPR sampling started in 1930 however this is deployed from moving platforms so criteria "*vi) the measurement site where the time-series has been acquired has to be known (Lat., Lon., height/depth)*" is not satisfied in the raw data. Gridded products are produced from raw CPR data but the goal of this task is to produce a tool that can be run in near-real time.

An example of long time-series that may be applicable, but does not include temperature or CH_4 data, is sea-level data from the tide gauge network. However, sea-level data falls outside the research infrastructures included in ENVRIplus.

In order to progress this analysis an attempt to test the method, the data collected at the Porcupine Abysal Plain (PAP) site between 2002 and 2007 will be conducted. This will be done in one year sections that are the duration of the deployment meaning that criteria v) and vi) are satisfied and inconsistencies because of different instrumentation or deployment locations are not introduced into the analysis. The results of this will be included in an updated version of deliverable 2.2 to be submitted prior to the end of the ENVRIplus project.

ACTRIS data of aerosol absorption deserve particular attention in order to set-up an actually controlled test and accurate analysis of the test results. The complexity of the data preparation will continue and the results will be presented in the updated version of the deliverable.

7. PRELIMINARY OUTLINE OF AN INNOVATIVE SERVICE FOR SCIENTIFIC USERS

The feasibility test on methodologies for detecting heterogeneities such as data gaps (non-existing values, zero values, out of range values) and breaking points has demonstrated that the selected methodologies, usually applied to some Solid Earth time-series, can also be suitable for application to Atmosphere time-series. This promising result encourage to extend the same methodologies to other type of time-series (i.e., Biodiversity) to verify eventual limitations and unsuitability.

At the same time the possibility to use a same methodology for the detection of data gaps and breaking points in time-series of different types suggest to develop basic shareable tools for the data Quality Assessment/Quality Control. These tools should be part of a larger suite of applications at disposal of data managers and scientists and should be adapted for the use in shared virtual environment such as cloud computing. In this respect the connection of the outcomes of this task to the ENVRIplus Theme 2 is evident and should be exploited to obtain complementary information and recommendation on the most suitable way to implement the service on a shared virtual environment.

In order to facilitate this interaction, a basic logical scheme of the methodology is presented without the presumption of providing a comprehensive and detailed design of the tool to be implemented for the service.

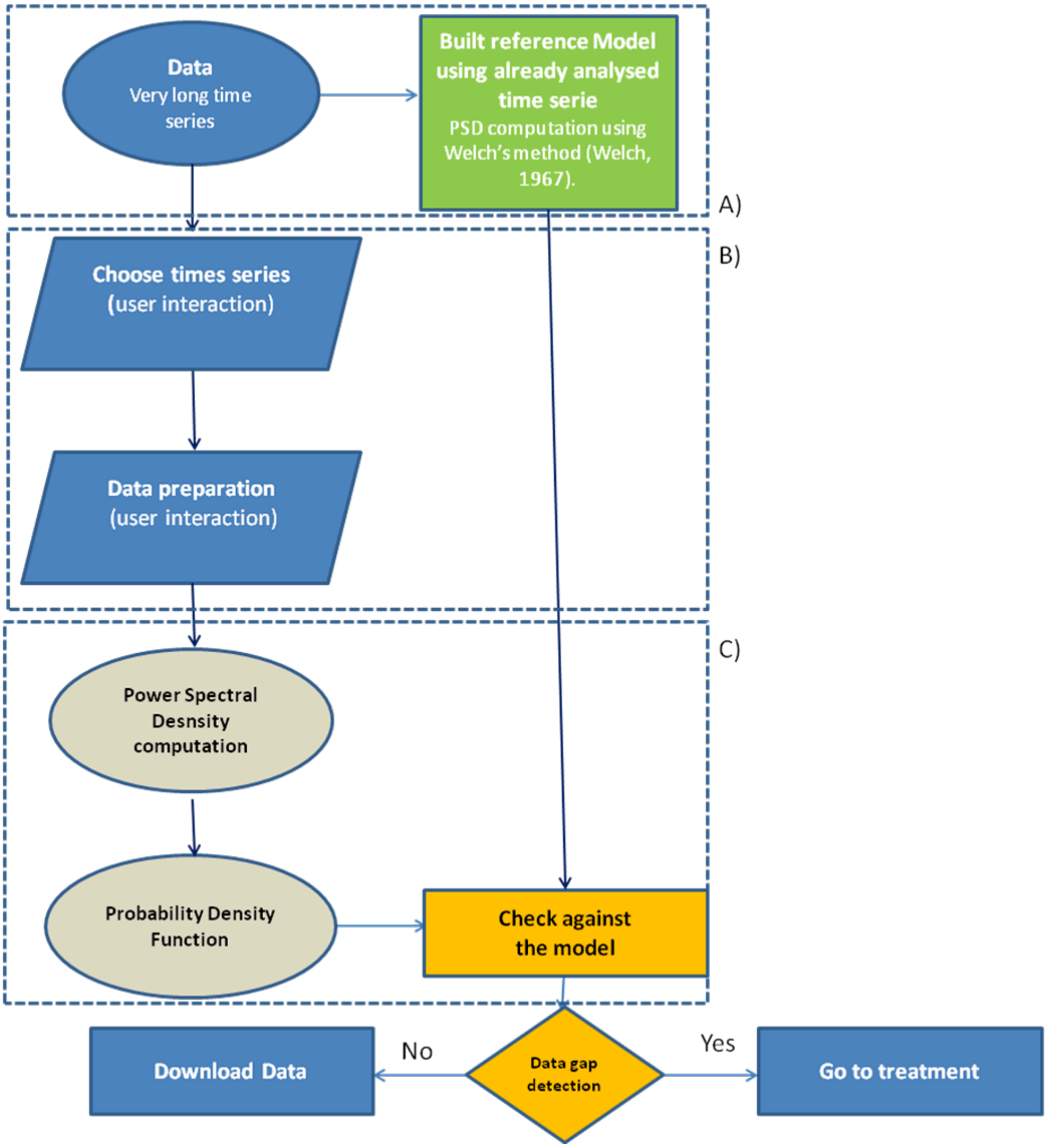
Data
Very long time series
Built reference Model using already analysed time serie
PSD computation using Welch's method (Welch, 1967).
A)
B)
Choose times series (user interaction)
Data preparation (user interaction)
C)
Power Spectral Desnsity computation
Probability Density Function
Check against the model
Download Data
No
Data gap detection
Yes
Go to treatment

Fig. 13 Simplified logical diagram describing the proposed tool for time-series heterogeneity detection: A) storage (permanent or temporary) of the time-series of interest ; B) interface for user input (criteria for choosing time series to be checked against gaps/outliers) ; C) computational phase for the verification of the existence of the heterogeneities and of their statistical weight. According to the outcome of the phase (C) the user can choose either to treat the heterogeneities (redirection to possible other service platforms) or to download the data.

8. CONCLUSIONS

A common definition for *data heterogeneity* across Research Infrastructures has been set up as a common accepted meaning was lacking.

Heterogeneity affecting the time-series in spite of the usual standardisation internal to Research Infrastructures were discussed and listed among the task participants. The most recurrent type of heterogeneities are the data gap (missing data or evident invalid data values) and outlier values.

A feasibility study was conducted on the application of methods derived from Seismology for heterogeneities detection to time-series provided by Research Infrastructures: the computation of the Probability Density Function of the Power Spectral Density of the time-series were applied for data gap and outlier detection, and the computation of the ratio between Short-Time Average and Long-Time Average was applied for breaking point (start time of heterogeneities) identification.

The feasibility study involved 'very-long time-series' (time scale from years to several years) typical of parameters related to global changes, and 'short time-series' (time scale from few seconds to some months) typical of parameters related to abrupt phenomena. For the two categories the following time series have been analysed:

- 'very-long time-series': Ozone oscillations; Solar Radiation UV-B; O_3 measured in the Lower, Mid and Upper Troposphere; CO measured in the Lower Troposphere;
- 'short time-series': ground velocity motion caused by an earthquake.

Marine very long-term time series from FixO3 with the required characteristics to check the applicability of the method (e.g., adoption a *model*) and perform the tests, were found complex to collect and prepare according to the requirements needed for the tests. The work of preparation of the data has been running however at the time of this report drafting and the results of this will be included in an updated version of deliverable 2.2 to be submitted prior to the end of the ENVRIplus project upon agreement to with the Coordinator of the project.

Methodologies considered typical for application in a specific sector of the Geophysics, that is Seismology, were found to be useful also to other disciplines and determined an increasing interest of task participants along the work development. The promising results support the proposal for the design and implementation of services to help scientists and data managers during the selection process of the most suitable data for their original elaborations.

The heterogeneity treatment issues are considered very dependent from the features of the corresponding parameters, site of measurement acquisition, modeling scale, and a trans-disciplinary approach to the various ENVRIplus time-series deserve a more deepen analysis.

A possible service to check for design can then include the development of basic processing tools for time-series in different domains based on the proposed methodologies for heterogeneities detection. The service should be developed in shared virtual environments (i.e., cloud), the can be a very helpful option assisting data managers in the regular Quality Assessment/Quality Check procedures and support scientists in accepting/discarding /correcting data before the final data selection in view of original analytical elaborations.

Impact on project: a common vocabulary on the meaning of *data heterogeneities* has been set up across different Research Infrastructures and disciplines. Records of the causes of heterogeneities have been analysed, discussed and shared. The need to have an effective but not sophisticated tools to quickly identify the data gaps and outliers is clearly spread across RIs and the task has given a reply to some time-series cases.

Impact on stakeholders: this task has highlighted one of the topic of converging interest of the ENVRIplus RIs community that is larger than the size of the partnership of the project: while languages and vocabularies are hard to be fully shared because of the differences of subjects and measurements and analysis techniques, Quality Control/Quality Assessment have evident common features. This task has put a seed for the future development and implementation of a practical tool that can be possibly used in a wider range of time-series than the one explored in the task T2.2. The implementation of a virtual environment shareable among managers and scientists of different infrastructures and disciplines is really feasible (see INDIGO - Data Cloud EC project, https://www.indigo-datacloud.eu/) and hopeful.

REFERENCES

Include here all references in a consistent form. You can also include references to earlier project deliverables – in this case make sure you are using direct links to the deliverable documents in the ENVRIplus website.

Beyreuther, M., R. Barsch, L. Krischer, T. Megies, Y. Behr, and J. Wassermann (2010). ObsPy: A Python toolbox for seismology, *Seismol. Res. Lett*. 81, no. 3, 530–533, doi: 10.1785/gssrl.81.3.530.

De Caro M., S. Monna, F. Frugoni, L. Beranzoli, P. Favali (2014). Seafloor Seismic Noise at Central Eastern Mediterranean Sites. *Seism. Res. Lett.,* 85, 5, 1019-1033, doi:10.1785/0220130203

McNamara, D. E., and R. P. Buland (2004). Ambient noise levels in the continental United States*,* *Bull. Seismol. Soc. Am*. 94, no. 4, 517–527.

McNamara, D. E., C. R. Hutt, L. S. Gee, H. M. Benz, and R. P. Buland (2009). A method to establish seismic noise baselines for automated station assessment*, Seismol. Res. Lett*. 80, no. 4, 628–637.

Peterson, J. (1993). Observation and Modeling of Seismic Background Noise, USGS Technical Report 93-322, 95pp.

Trnkoczy, A. (2002). Understanding and parameter setting of STA/LTA trigger algorithm: IASPEI *New Manual of Seismological Observatory Practice*, 2, 1–20; doi: 10.2312/GFZ.NMSOP-2 IS 8.1

Welch, P. D. (1967). The use of the fast Fourier transform for the estimation of power spectra: A method based on time averaging over short modified periodograms, *IEEE Trans. Audio* Electroacoustics AU-15, 70–73.

APPENDICES I

The requests addressed to the RI representatives in Task 2.2 in order to provide time-series for the feasibility test across ENVRIplus Domains (Solid Earth, Atmosphere, Ocean, Biodiversity) are reported below

1. Time-series: how you are requested to set up your data

(if a Power Spectral Density model is not available)

Step 1 - 'Individual' time-series selection and preparation: please check the following requirements

i. time-series must include measurements of **one** physical parameter only;

ii. time-series must have very-long duration (between 9 months to some years); the lower the sampling frequency the longer the time-series;

iii. time-series have to be already analysed and considered '*known-as-good*' with no anomalies or with very few and well known anomalies (type of anomaly, time of occurrence) such as data gaps, spikes, offset changes;

iv. the time-series has to be described (short text) in respect to types of phenomena (transient, periodical) usually observed and environmental influence (e.g., seasonal known variations, influence of meteorological conditions)

v. the time-series has to be produced by a unique equipment (no change of equipment within the time-series);

vi. the measurement site where the time-series has been acquired has to be known (Lat., Lon., height/depth)

vii. if Power Spectral Density (PSD) functions have been produced for the time-series, an example and a description of the main features of PSD has to be provided

viii. the following metadata have to be provided for the selected time-series:

Site name (also 'fake')
Network code (also 'fake')
Station code (also 'fake')
Location (Latitude and longitude, elevation/depth)
Channel code (in case of 'vectorial' time-series with x,y,z components)
Sampling rate
Start Time (year/month/day/hh/mm/ss.ss)
End Time (year/month/day)
Conversion factors from counts/volts (in case of digital/analogue equipment) to the international system of units (SI)

ix. data format description

Step 2 - Selection and preparation of additional time-series

If other time-series with the same features (similar PSD) as the one prepared in the Step 1 although with different metadata are available (e.g., your infrastructure provides CO_2 time-series in different measurement sites but in a region where the time-series is expected to be comparable) , then Step 1 can be developed for other individual time-series.

The repetition of Step 1 will provide a set of time-series (20-30 time-series would be a nice number)

The expected results from Step 1 and 2 is to obtain a set of time-series with similar behavior in the frequency domain and completely documented.

2. Power Spectral Density *model* computation

Step 3 - INGV compute and check the PSD of the selected time-series in order to find out a '*model*' a for your data in the frequency domain.

Step 4 - If a *'model'* can be identified, the partners will be asked to provide some *'known-as-bad'* time-series that is time-series with well known heterogeneities problems (e.g., data gaps, change of offset) and well documented (metadata as per our request). INGV will compute the PSD for this 'bad' time-series and compare to the *'model'*. The deviation from the *'model'* of the PSD of the 'bad' time-series is expected (the bad time-series results 'non homogeneous' by the comparison to the *'model'*).

On the basis of the detection of the deviations the test in declared successful.