European Cooperation in the field of Scientific and Technical Research
- COST -

Brussels, 14 November 2014

COST 105/14

MEMORANDUM OF UNDERSTANDING

Subject :	Memorandum of Understanding for the implementation of a European Concerted Research Action designated as COST Action IC1408: Computationally-intensive methods for the robust analysis of non-standard data (CRoNoS)

Delegations will find attached the Memorandum of Understanding for COST Action IC1408 as approved by the COST Committee of Senior Officials (CSO) at its 191th meeting on 12-13 November 2014.

MEMORANDUM OF UNDERSTANDING
For the implementation of a European Concerted Research Action designated as

COST Action IC1408
COMPUTATIONALLY-INTENSIVE METHODS FOR THE ROBUST ANALYSIS OF NON-STANDARD DATA (CRoNoS)

The Parties to this Memorandum of Understanding, declaring their common intention to participate in the concerted Action referred to above and described in the technical Annex to the Memorandum, have reached the following understanding:

1. The Action will be carried out in accordance with the provisions of document COST 4114/13 "COST Action Management" and document 4112/13 "Rules for Participation in and Implementation of COST Activities", or in any new document amending or replacing them, the contents of which the Parties are fully aware of.

2. The main objective of the Action is to coordinate activities directed to the development of fast, robust, and efficient solutions to extract accurate knowledge from non-standard and imperfect data satisfying the requirements of the end-users.

3. The economic dimension of the activities carried out under the Action has been estimated, on the basis of information available during the planning of the Action, at EUR 76 million in 2014 prices.

4. The Memorandum of Understanding will take effect on being accepted by at least five Parties.

5. The Memorandum of Understanding will remain in force for a period of 4 years, calculated from the date of the first meeting of the Management Committee, unless the duration of the Action is modified according to the provisions of Section *2. Changes to a COST Action* in the document COST 4114/13.

A. ABSTRACT AND KEYWORDS

Real data sets from a wide variety of fields violate the idealized assumptions inherent in standard statistical theory. Robust data analysis methodology aims to mitigate the impact of such violations. Robust methods are usually developed to handle multivariate data. However, monitoring studies often contain information such as functional, set-valued, or different kinds of incomplete data. Robust methods for these complex data types are scarce and involve critical computational challenges. New models, methods and efficient, numerically stable, and well-conditioned robust strategies are essential to improve knowledge extraction from non-perfect and non-standard datasets. Applications include the analysis of climate data, medical monitoring and diagnosis, trading and financial forecasts. The aim is to create an interactive network spanning computing, statistics, machine learning, and mathematics with the necessary expertise required to develop such strategies in close collaboration with end-users. Software and guidelines will be developed. The Action will provide European scientists with cutting-edge data analysis tools which will be suitably disseminated by disparate means such as training schools, conferences and publications. Improved decision-making tools for preventing-mitigating policies will be derived. Thus, scientific, technological and social challenges will be tackled by the creation of a proper framework to coordinate and optimize research efforts.

Keywords: Robust methods, large non-perfect and non-standard datasets, numerical estimation, combinatorial optimization, parallel implementation, High-Performance Computing and software.

B. BACKGROUND

B.1 General background

Statistical methods are frequently highly sensitive to any deviation from the underlying ideal conditions under which they are optimal. Robust methods are used to mitigate the distortion that errors, outliers or, more generally, model misspecifications may produce in the data analysis. They are usually developed to handle standard multivariate data. However, the current means to collect information provide more and more complex and large databases containing mixed types of elements, such as functions, sets, georeferenced information or linguistic values. This is the case for electrocardiograms, spectrometric, rainfall or consumption curves, cell shapes, linguistic medical assessments or diagnostics, forest distributions and their evolution, or daily price distributions.

Monitoring studies usually produce contaminated datasets which are updated continuously and require fast analyses. Occasionally data are missing, for various reasons, leading to incomplete datasets. Robust strategies, including robust filtering methods, are even more important in this context, as big datasets are likely to contain atypical influential information whose effect is harmful when the aim focuses on the statistical behaviour of the majority.

While there is a vast literature concerning some descriptive problems for particular elements (e.g. image clustering or classification), the robust analysis of other kinds of elements, such as functional data, is sparse. The computational efficiency is an extremely important challenge even in the multivariate case. Mainstream problems including feature selection, clustering or classification/regression trees are NP-hard (Non-deterministic Polynomial-time hard) and cannot be optimally solved for large datasets. Different heuristic techniques have been proposed based on greedy strategies, (random) local search, genetic and other evolutionary algorithms, to name but a few. These algorithms can find adequate solutions, but they can also be arbitrarily inaccurate, with the corresponding misprediction cost. Thus, there is an imperative need for improvement.

The efficient computational treatment of well-founded general robust methods is a critical open problem. Existing tools do not meet the needs of the end-users either for the kind of data they can cope with or for precision. The capacity of analysis should be adapted to the real current problems, which include the applications across different domains that will be tackled, such as the accurate investigation of environmental, health or financial risks, or the detection of frauds. The high cost for mispredicting risks in all the mentioned areas ultimately falls on the society. Expertise in robustness, computational statistics, machine learning, numerical methods and optimization, data management, High-Performance Computing (HPC) and, in general, computing, mathematics and various applications is required to find realistic and useful solutions.

Important efforts are decisive to make Europe competitive in advanced, efficient, and robust data analysis, essential in virtually any decision-making process. The topic demands interdisciplinary expertise spanning Information and Communication Technologies (ICT) in a broad sense which has not been put together yet. COST is the unique framework providing the means to develop the indispensable synergies to face the considered problems. Opening a dialogue between the different communities, including end-users, will allow the identification of key problems, and the development of the most appropriate methods and the best numerical and computational strategies to solve them. The participants will have the opportunity to integrate their research in a larger context aiming at standardizing and avoiding replications, so that research costs are long-term reduced by fostering efficiency. The networking activities, such as working meetings and scientific missions, will stimulate cutting-edge research topics and proposals to be submitted to national and

EU research programmes. The training resources will contribute to the creation of a niche of young researchers, which will promote the European leadership in the topic beyond the duration of the Action. Open and users conferences will be used both as a meeting point for experts and practitioners and to disseminate the scientific results derived from the Action. The topic has a worldwide interest.

B.2 Current state of knowledge

The research will focus on the computationally efficient and robust analysis of non-standard and non-perfect datasets. It involves:
Computational Statistics: The need of deriving accurate numerical solutions for analytically complex statistical problems encouraged the widespread use of specific numerical and computational tools for data analysis such as factorization, regularization or branch-and-bound techniques. Resampling techniques such as bootstrap and cross-validation have also become essential in the development of more flexible and advanced statistical approaches. Nevertheless, the demand grows faster than the solutions. Bigger and more heterogeneous databases often make the computational problems intractable. Big data is currently one of the hot topics in computational statistics. HPC, and specifically parallel and GPU (graphics processing unit) computing has been sporadically used in data analysis problems. That supports its consideration in the context of this Action to alleviate the computational burden.
Robust Statistics: In almost all the real-life databases there is a number of errors or atypical values which highly affect the classical statistical procedures. Nowadays it is widely accepted that robust procedures are a suitable alternative. Nevertheless, their use is not as extensive as it should be, with the consequent economic and/or time cost associated with the need to properly cleaning the databases and the frequent lack of reliability of the statistical conclusions. Reasons are related to the complexity of analyzing heterogeneous and big datasets and obtaining in practice accurate robust estimators under flexible enough conditions.
Non-standard and non-perfect data: One of the most general statistical frameworks dealing with complex elements is the so-called “Object Oriented Data Analysis”. It covers Euclidean space-valued data such as functional data, and non-Euclidean space-valued data, such as some modern medical images. The theoretical analysis of these kinds of elements is quite recent and is being mainly developed in USA, although European researchers also contribute to the topic. From a computational point of view, the problems are huge even in the classical statistical setting. Other challenges arise, such as censoring, typical in medical studies, missing values or the consideration

of not precisely-defined information, such as subjective perceptions. Efforts to develop data analysis tools in these contexts have been made in the last years. Nevertheless, the computational issue is again still a challenge, especially when robustness is also pursued.
Data management: The technical research in the considered area needs to be complemented with the development of efficient storage, data processing and visualization tools, taking as a basis previous paradigms.

In order to cope with non-standard data in a robust framework, some methods based on distances can be extended. Some non-standard data can be reduced to multivariate data. Specific methods based on wavelets, support vector machine or neural networks have also been studied. However, there is a lack of efficient well-founded and general enough robust methods to cope with many types of data available nowadays. This is because the approaches with a potential higher statistical efficiency would involve either optimization problems of complex functions which do not admit a closed-form solution, or combinatorial problems which are computationally demanding and infeasible for large data-sets. This Action aims at filling this gap. Some case-studies in different domains will be considered which illustrate the current needs in the considered area. They include climate, medical, financial and trading data. Obviously, expertise in different areas of data analysis, computer sciences and mathematics is enforced to tackle these problems, which additionally require the collaboration with experts in data collection and storage, software and end-users to make the solutions significant for the society.
Summarizing, the major challenge that the development of robust methods for non-standard and non-perfect data faces nowadays is the computational burden of the problems, and this is due to: a) the need to employ inherently intensive methods even for small scale problems, such as cross-validation; b) the complexity of the involved optimization problems, and c) the practical limitations when large datasets are involved.
Comprehensive research breaking the barriers that must overcome a realistic, efficient and well-founded robust data analysis will be developed. The innovative approaches to set up the research are:
a) The unified formalization of non-standard data as points in Hilbert or metric spaces.
b) The exploration of alternatives to the usual ways of solving optimization and combinatorial problems in statistics to derive a bound the error when it is not computationally feasible to obtain the optimum through coresets, graph-based approaches and orthogonal decompositions.
c) The employment of weighted re-sampling schemes to re-utilize computations.
d) The consideration of advanced parallelization and implementations in different architectures,

including GPU and clusters of PCs for problems that cannot be tackled by sequential algorithms.
e) This Action is marginally related with IC0702, which finalized 2 years ago. Although the object of both Actions involves the area of data analysis, the approaches and outcomes are essentially different. IC0702 aimed to strengthen the dialogue between general statistics, as a mathematical branch relying on a rigorous derivation of conclusions, and soft computing, as an engineering science focused on obtaining working solutions quickly, accepting approximations and unconventional approaches with the risk of obtaining arbitrarily inaccurate results. In contrast, the current Action is focused on the development of sound computational methods, in the sense of efficiently storing and manipulating information, for unexplored problems in statistics.

B.3 Reasons for the Action

The research in the considered area requires the interaction of different expertise. Interdisciplinary work is needed due to the complexity of the data and the problems. However, even within each homogeneous community, the experts tend to be remarkably specialized in either a kind of data or framework or technique.
The Action arises from a real need to generate new scientific and technological advances which will ultimately have an effect on the society through the applications. The current models, methods and tools are not enough to generate the global solutions pursued by this Action. Non-perfect and non-standard datasets arise in almost any ambitious observational or experimental study, for instance, climate analysis (e.g., georeferenced temperature or rainfall curves), medical studies (e.g., images, electrocardiograms or expert assessments), trading inspections (e.g., georeferended price-quantity relationships) or financial research (e.g., daily stock price histograms). Practitioners of different fields demand more advanced statistical tools guaranteeing accuracy to cope with the non-perfect heterogeneous databases arising from their research. Important challenges such as improved risk assessments or the detection of frauds will be investigated with the approaches to be developed.
The Action will maximize its outcomes and pursue the effectiveness of the research by coordinating efforts through the different networking tools that COST provides. The networking activities, including Working Groups (WGs) meetings and the Short Term Scientific Missions (STSMs), will contribute to create the required synergies among experts and end-users and to promote the development of joint research in a standardized and unified way. The research costs will be long-term reduced by fostering efficiency. The conferences and training resources will be employed to disseminate the scientific and technological advances and to promote the use of the developed techniques among end-users. This will improve the quality of the knowledge extraction in the

research developed in Europe. The outcomes of the Action will also be disseminated by the publication of on-line material and the edition of books and special issues in renown journals in order to enlarge its impact. The training resources directed to young researchers will be capitalized to strengthen the research topic and to assure the leadership of Europe in an area whose applications are very relevant in Horizon 2020. The creation of a critical mass, competitive at a worldwide level beyond the duration of the Action, and able to support the science and the society with effective data analysis tools, will constitute an added value.

B.4 Complementarity with other research programmes

The vision of this Action is completely original, and there is no European research project with similar aims. However, it involves state-of-the-art topics, which will enable partial links. Namely,
COMPLEXDATA (ERC-SG-PE1): Research for the statistical treatment of certain non-standard data and its applications to biophysics is proposed. Namely, random integral transforms, random unlabeled shapes, random flows of functions, and random tensor fields are considered for standard (i.e. non-robust) statistical analyses. The aims of the project regard the methodological aspects rather than the computational ones.
MAZEST (ERC-SG-PE1): Semi-parametric inference for standard datasets is considered from a theoretical point of view. One of the objects, the M-estimators, are also of interest for robust theory.
HPC-GA (PEOPLE-2011-IRSES): The simulation of large-scale geophysics phenomenon using parallel programming and heterogeneous architecture is investigated. The results are of interest for this Action specially for the treatment of spatial data.
3D-MASSOMICS (FP7-HEALTH): Preprocessing, unsupervised, and supervised methods of statistical analysis of 3D MALDI-IMS data and its implementation using GPU architecture is proposed. This is specially related with the medical applications involving set-valued elements.
COST Action IC1305: A European research network focused on large scale computing and big data management, which can contribute in this specific task of the proposed research.
Other projects relate to disparate particular tasks, such as, DwB (INFRA-2010-1.1.3), LOD2 (ICT-2009-5), and OPENCUBE (ICT-2013.4.3), for the data collection and publication, TSSV (PEOPLE-2012-CIG), for the applications regarding tempo-spatial data, SIPA (ERC-SG-PE1), for matrix computations in learning problems and AMSTAT, for the monitoring applications involving signal processing. The collaboration of the Action with participants of those projects will benefit the state of the art, since complementary results may be derived.

C. OBJECTIVES AND BENEFITS

C.1 Aim

The aim is to coordinate activities directed to the development of fast, robust, and efficient solutions to extract accurate knowledge from non-standard and imperfect data satisfying the requirements of the end-users.

C.2 Objectives

A network spanning different communities, including computing, statistics, machine learning, and mathematics will be established to face one of the main challenges in statistics: to conciliate the trade-off between robustness and efficiency for generalized settings. New tools for the management of databases which are continuously being updated and monitored will be generated from the interaction of the experts. The results will be widely applicable in all sciences that use experimental or observational data. The suitable transfer of knowledge by using the tools provided by COST such as the organization of meetings and training schools will be one of the main targets. Stakeholders and end-users have been involved in the development of this Action with the aim of defining the specific objectives and outlining their expected benefits.

Within the general aim established in C.1, a number of secondary objectives will be pursued. In order to provide a quantitative way of measuring them, the corresponding types of deliverables from those listed below will be indicated in brackets:

Objective S1: To generate new models, methods and computational tools for robust location, classification and regression problems involving non-standard and non-perfect data. (Deliverables 5 and 7).

Objective S2: To develop robust inferences based on re-sampling. (Deliverables 5 and 7).

Objective S3: To investigate parallel strategies and their implementation on various computer platforms. (Deliverables 4 and 7).

Objective S4: To provide a set of benchmarks where the new techniques can be tested in order to show the degree of improvement, objectively quantified through statistical measures and inferences. (Deliverable 3).

Objective S5: To produce advanced and easy-to-use data analysis tools for generalized settings. (Deliverables 3 and 6).

The following specific organizational objectives are tackled:

Objective O1: To create a proper framework to foster the interaction among researchers whose

expertise is useful for the Action, which includes statistics and machine learning, mathematics, computer sciences and practitioners and data collection people. (Deliverables 1, 2, 6 and 10).
Objective O2: To contribute to the leadership of Europe in the efficient and robust data analysis beyond the Action, with the consequent benefits that other science will obtain from that. This objective is twofold. On the one hand, it comprises the investment in training (Deliverable 2) and, on the other hand, it refers to the production of tangible high-quality results (Deliverables 7 and 8).
Objective O3: To identify new research agendas (Deliverable 9).
Objective O4: To promote the effective transfer to the final users so that the scientific advances have a real impact on the society (Deliverables 1, 4, 6 and 10).
The primary deliverables can be summarized as follows:
Deliverables type 1. *Online resources:* A web page providing visibility to the Action will be set up at the beginning of the Action. This will be a space to disseminate and further the dialogue aimed at opening new research directions. It will also be used to encourage participation in the Action by interested researchers and organizations. Public and private on-line resources will be available. A forum and a tool to interchange scientific documents will be provided. Repositories derived from the COST Action activities (such as software, reports, benchmarks, publications, etc.) will be published and updated continuously.
Deliverables type 2. *Educational resources:* The material of the yearly courses organized by the Action will be published. A topic in the on-line forum will also be reserved for educational purposes.
Deliverables type 3. *Datasets and benchmarks:* The datasets that will be used as benchmarks will be published and frequently updated.
Deliverables type 4. *Algorithms and software*: The methods and their implementation in open-source packages will be periodically released so that cutting-edge data-analysis tools is accessible to the end-users.
Deliverables type 5. *Reports.* A number of reports with scientific or organizational content will be generated, namely, the results of at least 30 STSMs, 2 yearly Management Committee (MC) and WGs meetings, as well as the reports of the subcommittees and the Advisory Board.
Deliverables type 6. *Standards and usage guidelines.* Documents with detailed specifications and rules to help the practitioners to apply the results in an optimal way and to use the software will be developed.
Deliverables type 7. *Publications:* the new scientific advances will lead to a large number of journal publications, presentations at conferences, special issues of journals and at least 2 books.
Deliverable type 8. *A new specialized journal:* Once the topic has achieved the expectable strength,

a journal published by a prestigious editorial board and covering the methodological and computational aspects of the robust analysis in the considered context will be launched.

Deliverables type 9. *Submission of new research proposals.* Supported by the COST tools, at least 10 new (national and European) proposals integrating members of different groups of the network will be generated. Research proposals in collaboration with stakeholders will be especially encouraged.

Deliverables type 10.*Organization of events.* One yearly Action conference, forums with users, specialised workshops and sessions in related conferences will be organized.

All the deliverables refers to tangible outputs. The quantity and quality of such outputs will be internally assessed by the MC and externally by an Advisory Board with the help of end-users already involved in the Action, which cover the key applications mentioned in this Memorandum of Understanding, and those who will join later. They will take into account the results of the benchmarks, the publications in high-standard journals, the organization and participation in prestigious conference, the attendance to the course and satisfaction of the participants, the submissions to the new journal and the results of the evaluation of the proposals. Details about the delivery times are given in F.

C.3 How networking within the Action will yield the objectives?

The objectives will be achieved by combining efforts through the following networking activities in an iterative process:

Organizational meetings. The MC with the counsel of the WG leaders and the Advisory Board will design the strategies to create an interactive environment of scientists and professionals of academia, research centers and industry. Two meetings per year will be held complemented with frequent on-line communication. Details are given in E.1.

Scientific meetings. The WGs will exchange resources and collaborate in the development of the Objectives S1-S5 by combining the expertise of the participants of the different disciplines. The WGs will meet at least twice per year. They will also employ on-line tools for a more effective interchange. Details are given in E.2.

Face-to-face meetings. The close collaboration with end-users is essential for the success of the Action. Although stakeholders' representatives are participants of the Action, periodical face-to-face meetings will be organized to provide input to the WGs' activities.

Conferences and related events. User conferences, PhD workshops, forums, specialized sessions or tracks in related conferences (such as ERCIM, COMPSTAT, ICORS or PMAA) will be organized

in order to: 1- provide a wide and accessible networking environment; 2- foster the interaction with practitioners and get feedback; 3- disseminate the results of the Action and promote interest in the topic among the European researchers. At least 2 yearly events, one of each will be an Action international conference, will be coordinated. The Action will close with a specialized workshop gathering network participants and users.
Courses and tutorials. The training activities will be especially important in order to provide a common framework and to assure continuity beyond the Action lifetime. They will be organized either in summer or prior to the above-mentioned events in order to optimize resources. External experts that may contribute to the Action aims will occasionally be invited.
Mobility. The bilateral interchanges through STSMs will be proactively encouraged to allow Early-Stage Researchers (ESRs) to interact with senior experts, in order to generate innovative scientific knowledge, and to favour the submission of original research proposals.
Visibility: The publication resources, such as the website, on-line resources and scientific papers, will be exploited with the aim of attracting stakeholders, promoting an effective transfer of knowledge, and fostering long-term collaborations for new research projects.

C.4 Potential impact of the Action

The Action mainly aims at ICT-supported scientific and technological advances, but will also imply social benefits through its applications. Namely,
Scientific advances: A new paradigm in data analysis including new generalized models, methods and algorithms satisfying the demands of practitioners to analyze non-standard databases. Specifically, novel data analytic tools will be developed for estimating and testing concerning location, classification, clustering and data reduction problems and regression models involving complex random elements and different kinds of imperfections in a robust and efficient way. Strategies to address computationally-intensive combinatorial problems, which are common in general data analysis, will be provided. Significant progress on the interface of areas of computing, statistics, machine learning, and mathematics will be generated. This kind of knowledge is potentially useful for purposes not foreseen now, as it is common with any mathematical advance given its level of abstraction. There will be a direct impact on Big Data Science, which is permeating every field nowadays due to the presence of massive data sets. Specifically, Big Data management will benefit from the developed statistical techniques such as robust dimension reduction. Additionally, a critical mass with effective interdisciplinary background and bridging the gap between the theory and the applications, able to reinforce the European leadership in the area,

will be formed.
Technological advances: Repositories of open-source software and benchmarks which will favour a structured and standardized way to share information and will facilitate the application of the results. Easy-to-use on-line tools and best practice guides to employ the new methodologies will be provided.
Societal advances: The considered applications are eminently relevant for the society and affect important economic issues. Many environmental, medical and financial problems, among others, are essentially complex and large-scale. This is the case, for instance, of the global warming and its effects, or the sustainable management of natural resources. Their accurate analysis compels an efficient treatment of huge, and often incomplete, datasets involving different types of data. Data analytic tools such as clustering, decision trees, ANOVA, feature selection have been employed within these contexts and their robust and efficient versions are essential. Consequently, the results of this Action will have a direct impact in those settings. The improved decision-making tools will be directly useful for the design of preventing-mitigating strategies in different critical areas such as environment, health and economy. One of the social targets is to provide reliable and accurate tools to estimate financial risk, being it market, credit or operational risk. Current crises have pointed out that our society falls short of such methods and research should be a top-priority. Stakeholders aims a benefit from, e.g., a) the development of robust statistic methods to build early-warning detector indicators of different risks, including economic and financial crises ; b) the development of policies and regulations to better control for bank diversification, riskiness, interconnectedness, and liquidity and information contagion; c) the access to benchmark data and state-of-art models to test their internal models ; d) the access to network outputs and knowledge to improve statistical and data analysis literacy of their personnel. Moreover, since non-standard datasets may appear in virtually any observational or experimental study, the Action will benefit industrial applications. On the other hand, the training resources will pave the way to European early-stage researchers, making it easy to access to a variety of employment possibilities in the broad area of data analysis.

C.5 Target groups/end users

The results of this Action will be used at different horizontal levels.
a) Academics in areas of ICT, and more generally, computer science, statistics, machine learning, mathematics, big data science and finance, and especially young researchers, will take advantage of the scientific and technological advances and will build on them.
b) Health researchers and professionals, environmental and social scientists, quantitative analysts,

statistical consultants, professional of the industry and the banking sectors, forest managers and regulators are also final users through the applications, and are or will be involved as stakeholders. Experts from most of the mentioned fields have been involved in the preparation of this Action, namely, academics and practitioners in the areas of statistics, machine learning, mathematics, computer science, finance and professionals of medicine and engineering. Specifically, stakeholders in areas of medicine and finance have directly contributed to the Action, while others in areas such as environmental risks are involved through their data analysts.

D. SCIENTIFIC PROGRAMME

D.1 Scientific focus

The research process will be iterative, that is, each problem arising from the first research task will be subsequently analyzed at different stages. In the end, feedback will be provided by the practitioners so that the process starts again to refine the approach and gradually improve the tools until a satisfactory standard is attained. The main scientific tasks in the recursive process are as follows:

Task 1. Formalization of the problems. Firstly, the working framework will be established taking this memorandum as starting point. Further discussions and the input of new participants will be considered. The priority lines will be defined in light of the practical demands. A close interaction between the end-users and experts who will develop the methods will be fostered. In this way the critical shortcomings of the state-of-the-art methods can be properly analyzed and the essential problems can be formalized. Some of the researchers of the Action already work in collaboration with end-users and this will facilitate the initial contacts. However, forums and face-to-face meetings will be organized to enlarge the vision of the Action and, as a consequence, its impact.

Task 2. Data management. The aim is to develop methods tailored to specific real needs. Thus, the datasets to be analyzed have to be collected and be at disposal of the Action members in advance. To facilitate the storage and access to the data, open-source data platforms such as CKAN may be used. The datasets will be updated while the Action progresses.

Task 3. Models and methods: development of efficient solutions. The next step is to investigate different robust strategies to approach computationally intensive key-problems such as robust clustering, decision learning trees, multiple testing and feature selection. Fast and numerically stable recursive estimation strategies will be developed. Expertise in HPC will be exploited to tackle those problems that cannot be considered using conventional sequential strategies.

Task 4. Resampling-based inference: development of efficient solutions. Frequently the optimality of the traditional approaches relies on rigid distributional or model constraints that can be alleviated by using resampling-based methods. In these cases flexible, consistent and efficient resampling methods will be investigated. Given the computational burden of these kind of procedures, the strategies will be improved by using HPC.

Task 5. Implementation of the algorithms. In order to make the results accessible and easy-to-apply by the end-users, specialized software will be realised according to standard rules to be defined by the network. The algorithms will be implemented in open-source packages. The packages will include graphical tools and a detailed documentation assisting in the use of the packages will be delivered.

Task 6: Applications and benchmarking. Finally, the implemented procedures will be tested on the stored datasets. The practitioners will provide feedback so that the approaches are tuned if needed and best practice guides will be generated to assist end-users beyond the Action. Supporting results for data-based policies will be investigated, e.g. the determination of areas exposed to different environmental risks. As a final product, advanced and easy-to-use data-analysis tools ready to be exploited by stakeholders will be provided.

The participants interested in the Action will constitute the manpower needed to tackle the considered problems. These participants have their own equipment and research funding and will have the opportunity to integrate in the network thanks to the tools provided by COST to carry out the previously described networking activities.

D.2 Scientific work plan methods and means

The research tasks will be carried out by five WGs. High interaction among the researchers of all the WGs will be fostered. Each WG will have autonomy to develop their results, which will constitute a piece of the final outputs.

WG A- Data management and applications. This WG will mainly contribute to Tasks 1, 2 and 6. Climate, medical, trading and financial data will be used as initial benchmarks. Climate data are currently on-line available in multiple national agencies. Medical data will be provided by hospitals and research institutes which are represented in the Action. The participants of the Action have also access to large databases containing information about transactions from different countries and to financial data through benchmarks as *datastream*. The analysis of such datasets is important in the assessment of risks. Firstly, data will be collected and stored following protocols for standardization and respecting confidentiality matters. Distributed storage systems and massively parallel access to

data (e.g., Map-Reduce) will be exploited. Once the priority lines had been established, the problems will be analyzed by the others WGs. Then, the approaches that they develop will be tested and the results will be published as benchmarks. Any room for improvement will be highlighted. Further datasets arising from the interaction with the stakeholders are expected. The practical use of the approach will be the basis to generate the guidelines.

WG B - Models and methods. This WG will collaborate with WG A for the development of Task 1, and with WG C for the development of Task 3. A sound framework will be established to handle non-standard and non-perfect datasets, including elements such as functions, surfaces or sets, e.g. those provided by WG A. New robust population measures and models will be investigated, since the location-scale model is not always meaningful, as it relies on a Euclidean structure. Important problems which have not previously been considered will be clearly identified, formalized and tackled. Problems concerning location, including different regression frameworks, scale, clustering, and dimensionality reduction involving non-standard data, will be considered as starting points. Robust methods for estimation, such as those based on trimming, are intuitively generalizable. Alternatives using approaches such as M-type and minimum divergence estimators or distance-based methods will be tackled. Robust filtering techniques will be developed.

WG C – Computational tools. This WG will collaborate with WG B for the development of Task 3, and with WG C for the development of Task 4. Initially, the existing computational methods used for the most basic problems will be improved both in efficiency and accuracy. Subsequently, new algorithms to make the results of WG B applicable in practice will be developed. Cross-sectional research by providing alternatives to heuristic algorithms will be developed. Computationally intensive key-problems in robust clustering, decision learning trees, multiple testing and feature selection will be addressed. The setting includes problems traditionally solved by using stepwise algorithms which cannot be guaranteed to find either the optimal solution or any bound on the error. One of the innovative objectives is to develop new strategies able to provide such a bound when it is not computationally feasible to attain the optimum based on exact algorithms which are feasible for small size. The computational burden of these methods is high, but can be greatly reduced by parallelization. Thus, the parallelization of the derived strategies and their implementations in different architectures, such as GPU and clusters of PCs, will be explored. Algorithms based on the paradigms of computing coresets and/or sketches of data will be investigated. These algorithms can, by design, be easily distributed (for example, using Map-Reduce), parallelized or deal with streams of data.

WG D – Resampling-based inferences. This WG will collaborate with WG C for the development of Task 4. Cross-validation and bootstrapping are two of the most frequently employed resampling-

based approaches in statistics and data analysis. They are based on the idea that available samples can be used to obtain other versions of possible samples from the underlying population, and frequently provide more accurate results at a higher computational cost. Their robust versions are less extended and even more challenging. Computational strategies based on the re-utilization of the previous operations can significantly reduce the computational burden. However, the combinatorial nature of the problems will make the novel approaches computationally arduous and even infeasible for large-scale and high-dimensional problems. Consistency results providing the methods with a sound basis will be proved.

WG E- Software. This WG will develop Task 5. Open source software will be developed following standard rules in order to make the computational methods widely accessible for researchers in related topics and practitioners. The algorithms developed by WG C will be implemented in R, which is currently the leader free environment for data analysis, and other open-source languages. The capacity of these environments to cope with large datasets is still limited. Issues that relate to their ability to extract, transform, load and process big data needs to be addressed. Research in this direction will be developed, e.g., by incorporating fast matrix approximation techniques based on stochastic techniques. The elaboration of visualization tools to facilitate the inspection of large non-standard data sets will also be considered.

In a temporal point of view this working plan will be developed in combination with 5 stages:

Establishment of the network: Gather the knowledge of the experts, list the problems that will be addressed during a period of 6 months, identify the strategies to be investigated and establish the collaboration teams.

Development of the research: The WGs will interact to optimize the resources of the researchers to provide a unified outcome.

Diffusion: The new advances will lead to journal publications, presentations in conferences, special issues of journals and books.

Technology transfer: Workshops, special sessions and face-to-face meetings with stakeholders will be organized in order to assure that the results arrive to the final users.

Consolidation of the network and the research topic: At the end of the 4th year the WGs will determine the future directions for the established network.

These stages will make use of the COST tool as a mean to produce the targeted deliverables. The WGs will contribute to deliver some outcomes, such as

a) Reports of their activities and material to be published in the website.

b) The organization of special sessions for the conferences.

c) The cooperation in the organization of training activities.

d) The submission of new research proposals.

e) Publications through the different channels considered by the Action.

Some of the WGs will produce specialised outcomes, such as data-analysis tools (WG E) and datasets and guidelines (WG A).

E. ORGANISATION

E.1 Coordination and organisation

The Action will be guided by the Management Committee (MC), which will be formed according to standard rules for implementing COST Actions. The MC will meet twice per year to lead and monitor the Action so that the outcomes are timely delivered. The chair and vice-chair will be elected in the kick-off meeting. Five WGs will be created. A secretary will be appointed which will support the MC in administrative matters. The following subcommittees will assist on different issues:

Editorial board: It will be formed by experienced editors and will contribute to the review of the scientific results and advice on the dissemination policy.

Committee for STSMs : Representatives of all the WGs will be in charge of guiding the STSM policy by establishing a working programme and evaluating both the applications and the results. The committee will encourage the ESRs to participate in this programme and will award the most significant contributions, which will be invited for presentation in a special session of the Action conference. Visits of senior researchers to strengthen the connectivity of the different European groups will also be part of the programme.

Scientific Programme Committee (SPC): A balanced group of researchers of the different areas will assist in topics related to the organization of the conferences and courses. They will be in charge of creating a harmonized programme covering all the WGs topics. They will assist in the election of the topics for the tutorials and courses, specially focused to train ESRs. The suggestion of keynote speakers and the selection of external researchers who can be invited to collaborate will also be among their tasks.

On-line resources manager: One person will be responsible of all the aspects related to the web site, such as the design, content, forum, online tools, training, online communication or updates.

Committee for the gender-balance: Female researchers are still under-represented in the areas related to ICT. This Action aims at gradually improving the gender-balance. A committee

promoting this aim will be created. It will be in charge of developing special activities such as to organize special sessions and meetings, to develop a good practice manual or to keep an updated space for announcements.

ESRs Contact: The involvement of young scientists in the network is essential to guarantee its future. A person will be designated to promote the participation of the ESRs in the different activities designed for them, such as training courses, STSMs, educational forums, special sessions in conferences. They will also be encouraged to contribute to the rest of the activities of the network both at scientific and organizational level.

Project Contact: A person responsible of maintaining the network in contact with research projects related to the Action will be designated. This person will liaise with the group members in order to prepare and submit new research proposals within European frameworks.

Core Management Group (CMG): A small sub-group of the MC will serve as link between the subcommittees, the COST office, the stakeholders and the MC. Specifically, a member of the Core Management Group will be responsible for the contacts with stakeholders, the organization of the face-to-face meetings and the promotion of the forums and courses for end-users. They will prepare the documentation to be presented to the MC helping to make the MC meetings dynamic and well-focused. They will meet four times per year, some of them on-line.

The committees will be periodically renewed.

An active external Advisory Board (AB) with world-renowned experts of all the involved fields will be created. It will report on the activities and advise whenever necessary, especially if conflicts of interests arise.

Researchers from 19 countries spanning Europe have already expressed interest in the Action. They include top researchers, ESRs, practitioners and editors of important related journals (such as Journal of the American Statistical Society, Computational Statistics and Data Analysis, Statistics, Statistics & Computing, Journal of Machine Learning Research, Annals of Statistics, Numerical Linear Algebra with Applications or International Journal of Computer Mathematics). The research tasks will be carried out by the participants funded by their own countries and projects. The Action will coordinate the network in order to optimize the outcomes. A network of 60-80 participants is expected. Currently 48 partners have already signed up for the Action.

The milestones chronologically ordered will be detailed in F. The main types are:

Milestones type 1: MC meetings: Evaluation of the semester, approval of deliverables (reports, on-line resources, etc.) and planning of the next semester.

Milestones type 2: WGs meetings: Evaluation of the scientific activities, collection of deliverables and scientific planning of the next semester.

Milestones type 3: Website launch and annual revisions with on-line resources such as forums and repositories.
Milestones type 4: Finalization of data collection and updates.
Milestones type 5: Conferences and related events, including forums with end-users, Action conferences, specialized workshops and organization of sessions in related conferences.
Milestones type 6: Courses and tutorials for the establishment of common grounds, for ESRs and for end-users.
Milestones type 7: Annual reports both from the Core Management Group and the Advisory Board.
Milestones type 8: New journal launch.
Milestones type 9: Software releasing: New data analysis tools.
Milestones type 10: Usage guidelines releasing.

E.2 Working Groups

The five WGs listed in D.2 will be responsible for developing the scientific work plan. The following organizational rules will be considered:
a) Each WG will have a coordinator, approved by the MC, who will be responsible of guiding the WG according to the rules agreed by the MC. They will also be the link with the other WGs and all the committees of the Action. The coordinators will periodically report to the CMG.
b) The WGs will be connected through virtual tools and will physically meet twice per year in order to facilitate the dialogue, understanding and synergies among the participants.
c) Each participant will be able to participate in no more than 2 WGs to guarantee an effective networking, but avoiding over-dispersion and minimizing organizational problems.
d) The size of each WG will be at most 20-25 researchers. In case of exceeding this size, it is advisable to create sub-groups making the communication more efficient. The sub-groups will deliver their results in common with the WG, but will also have private time during the meetings.
e) The participants will communicate to the MC their preferences and they will be later distributed so that a proper balance is achieved.

E.3 Liaison and interaction with other research programmes

The European research programmes related to the topic from a wide perspective have been listed in B.4. There are currently other projects with a collateral relationship such as the ERC grants SMARTBAYES, MLCS, SMAC SCIENCEFORE, or NAMSEF, and the Marie Curie Action

MHDTURB. The Action will designate a Project Contact as described in E.1. This person will be in charge of maintaining contacts with the teams in related research projects. The list will be frequently updated. The Project Contact will appoint face-to-face meetings for the closer researchers to foster an effective interchange of knowledge. When appropriate, the experts involved in related projects will be invited to participate in the Action conferences, WGs meetings and other activities of the Action. Joint seminars will be organized if common goals are established.

E.4 Gender balance and involvement of early-stage researchers

This COST Action will respect an appropriate gender balance in all its activities and the Management Committee will place this as a standard item on all its MC agendas. The Action will also be committed to considerably involve early-stage researchers. This item will also be placed as a standard item on all MC agendas.

As already mentioned, there is lack of gender-balance in the involved research areas. The Action aims to achieve such a balance. The Action counts on a relative high proportion of females in comparison with the usual numbers in related publications and journals, which is often under 20%. They include both seniors and ESRs. A sub-committee will be created to foster the participation of females. Their duties are described in E.1 and cover the organization of events and the preparation of a good practice manual.

On the other hand, an ESRs contact will observe the interests of the ESRs. The Action will develop special activities aiming at the incorporation of ESRs and their motivation to develop high-quality research in the area, such as courses given by renown senior experts and the organization of special sessions to award the best results.

F. TIMETABLE

The duration of the Action will be 4 years. The timetable by semesters (from S1 to S8) will be as follows:

Semester	Milestones	Deliverables
S1	Kick-off MC meeting; WGs meetings; Website launching; Data	Initial website content and tools; Reports of the meetings and STSMs; Datasets

	collection	
S2	MC meeting; WGs meetings; Website updates; Courses to establish common ground; Action conference	Material of the courses; Reports of the meetings and STSMs; Book of Abstracts of the Action conference; Annual repository of presentations in conferences and publications of members of the Action; Annual evaluation of the CMG and the AB
S3	MC meeting; WGs meetings; Software releasing; Guidelines; Sessions in related conferences	Website updated with the new outcomes; Reports of the meetings and STSMs; Submission of research proposals; Software package and usage guide; Benchmarks; Special issue
S4	MC meeting; WGs meetings; Website updates; Course and forum with end-users; Action conference; Data collection	Material of the course; Reports of the meetings and STSMs; Book of Abstracts of the Action conference; Annual repository of presentations in conferences and publications of members of the Action; Annual evaluation of the CMG and the AB; Submission of research proposals; New datasets
S5	MC meeting; WGs meetings; Software releasing; Guidelines; Sessions in related conferences	Website updated with the new outcomes; Reports of the meetings and STSMs; Submission of research proposals; Software package and usage guide; Benchmarks updates; Special issue
S6	MC meeting; WGs meetings; Website updates; Courses for ESRs and end-users ; Action conference; Data collection	Material of the courses; Reports of the meetings and STSMs ; Book of Abstracts of the Action conference; Annual repository of presentations in conferences and publications of members of the Action; Annual evaluation of the CMG and the AB; Submission of research proposals; New datasets
S7	MC meeting; WGs meetings; New journal launching; Software releasing;	Website updated with the new outcomes; Reports of the meetings and STSMs; Submission of research proposals; Software package and usage guide; Benchmarks updates; Special issue

	Guidelines; Session in related conferences	
S8	MC meeting; WGs meetings; Website updates; Course; Closing conference; Final report; Software releasing; Guidelines releasing	Material of the course; Reports of the meetings and STSMs; Book of Abstracts of the Action conference; Annual repository of presentations in conferences and publications of members of the Action; Annual evaluation of the CMG and the AB; Submission of research proposals; Book edited; Guidelines

G. ECONOMIC DIMENSION

The following COST countries have actively participated in the preparation of the Action or otherwise indicated their interest: AT, BE, BG, CH, CY, CZ, DE, EL, ES, FI, FR, HU, IE, IT, NL, PT, RO, SE, UK. On the basis of national estimates, the economic dimension of the activities to be carried out under the Action has been estimated at 76 Million € for the total duration of the Action. This estimate is valid under the assumption that all the countries mentioned above but no other countries will participate in the Action. Any departure from this will change the total cost accordingly.

H. DISSEMINATION PLAN

H.1 Who?

The dissemination of the results will be carried out to arrive to the following target audiences:
a) Scientists using experimental or observational data, including those working for public institutions. Specifically, in environmental, medical and social science and in economy and finance.
b) Researchers in ICT, covering computer and big data science, statistics, machine learning, and some branches of mathematics, especially ESRs.
c) Statistics consultants and quantitative analysts.
d) Educational institutions which can use the outcomes in master programmes.
e) Professionals of the industry and banking sectors, including risk managers and financial

consultants.
f) Policy makers, especially in environmental areas.
g) Participants of collateral related European projects, as those listed in E.3.
h) Members of collateral related societies and working groups, as IASC, CFE or ERCIM WG on CMStatistics.
i) General public interested in the scientific and technological advances.

H.2 What?

The results of the Action will be disseminated using various channels, some of them being more adequate for specific audiences, which is indicated in brackets:
a) Website: announcements, official documentation, state of the art reviews, WGs and STSMs reports, slides of talks and seminars, call for papers, calls for potential STSM (easy to access for all audience).
b) On-line networking tools: forums, e-mail networks (mainly researchers close to the Action).
c) Use of e-news mailing lists of related areas (participants of collateral projects and societies).
d) Open seminars, tutorials, TS and courses (ESRs).
e) Publication of papers in peer-reviewed journals, books, special issues and a new specialized journal (researchers).
f) Action conferences: publication of Books of Abstracts and working papers freely available from the website of the Action (mainly researchers of the involved fields, but also scientists in other areas, professionals and policy-makers).
g) Participation in other related conferences (researchers, scientists in other areas, professionals and policy makers).
h) Face-to-face activities such as STSMs, meetings with members of the target audience (especially, consultants, professionals, policy-makers and representatives of related societies and educational institutions).
i) Software documentation, benchmarks and guidelines (mainly scientists in other areas, but also researchers close to the Action, consultants, professionals and policy-makers).
j) Non-technical publications in newspapers, magazines or TV (for general public).

H.3 How?

The dissemination plan will start as soon as the kick-off MC meeting establishes the guidelines for

the first year, according to this document. The Editorial Board will be responsible for the day-to-day control of the dissemination activities. It will serve for both supporting and encouraging the participants in their dissemination. The diffusion actions will be adapted when necessary to optimize the results.

The first dissemination method to be used is the website. In order to attract people to visit it, attention will be paid to design a clear, visual and easy-to-use structure. The contents will be organized combining the needs of the COST participants, other researchers in related areas, stakeholders and non-expert visitors. Original information and links to related pages, such as prestigious universities and research centers, societies and journals, will be updated permanently to stimulate frequent visits. Other virtual resources will be intensely exploited. The on-line resources manager will also be responsible for creating an international mailing list and collecting national mailing lists from the MC representatives of each country. They will be used to announce the website and the main novelties periodically.

The SPC will be in charge of designing a stimulating programme in order to attract researchers to attend the Action conferences. In addition to the outstanding invited keynote talks, several specialized invited sessions with leading researchers, and tutorials will be organized. The COST participants and other active experts will be encouraged to organize a number of sessions. The conferences will be organized in prestigious universities or cities and local information will be facilitated to create a positive environment for the meetings. Members of all the target audience will be invited to participate. The tutorials, TS and PhD courses will be carefully designed to cover challenging topics in order to foster initiatives among the young researchers. The Advisory Board will be actively involved in designing the scientific programme. The early-stage researchers will have the opportunity of disseminating their research in special sessions.

The conferences will be complemented by smaller meetings of professionals focused on disseminating the results of the Action. Additionally, the MC representatives of each country will contact relevant national organizations and serve as liaison with the COST Action members for diffusion purposes.

The scientific results will be submitted for publication to high-impact journals such as Journal of the American Statistical Association, Computational Statistics and Data Analysis, Statistics & Computing or Journal of Multivariate Statistics and the new journal to be created under the advise of the Editorial Board. The Editorial Board will also be responsible for the edition of special issues and specialized books.