



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

---

**Secretariat**

-----

**Brussels, 30 November 2007**

**COST 257/07**

**MEMORANDUM OF UNDERSTANDING**

---

Subject : Memorandum of Understanding (MoU) for the implementation of a European Concerted Research Action designated as COST Action IC0702: Combining Soft Computing Techniques and Statistical Methods to Improve Data Analysis Solutions

---

Delegations will find attached the Memorandum of Understanding for COST Action IC0702 as approved by the COST Committee of Senior Officials (CSO) at its 169th meeting on 15 - 16 November 2007.

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

**COST Action IC0702**

**COMBINING SOFT COMPUTING TECHNIQUES AND STATISTICAL METHODS TO**  
**IMPROVE DATA ANALYSIS SOLUTIONS**

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 299/06 "Rules and Procedures for Implementing COST Actions" (or in any new document amending or replacing it), the contents of which the Parties are fully aware of.
2. The main objective of the Action is to strengthen the dialogue between the statistics and soft computing research communities in order to cross-pollinate both fields and generate mutual improvement activities.
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 14 million EUR in 2007 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 four 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 Chapter V of the document referred to in Point 1 above.

\_\_\_\_\_

## **A. ABSTRACT AND KEYWORDS**

The main objective of this COST Action is to strengthen the dialogue between the statistics and soft computing research communities in order to cross-pollinate both fields and generate mutual improvement activities. Soft computing, as an engineering science, and statistics, as a branch of mathematics, emphasize different aspects of data analysis. Soft computing focuses on obtaining working solutions quickly, accepting approximations and unconventional approaches. Its strength lies in its flexibility to create models that suit the needs arising in applications (context of discovery, model generation). In addition, it emphasizes the need for intuitive and interpretable models, which are tolerant to imprecision and uncertainty. Statistics is more rigorous and focuses on establishing objective conclusions based on experimental data by analysing the possible situations and their (relative) likelihood (context of justification, model validation). It emphasizes the need for mathematical methods and tools to assess solutions and guarantee performance. Bringing the two fields closer together will enhance the robustness and generalisability of data analysis methods, while preserving the flexibility to solve real-world problems efficiently and intuitively.

Keywords: data analysis, soft computing, statistics, model generation, model validation

## **B. BACKGROUND**

### **B.1 General background**

Soft computing is a discipline that deals with the design of hybrid intelligent systems which, in contrast to classical hard computing techniques, are tolerant to imprecision, uncertainty, partial truth, and approximation. Consequently tractable, intuitive, more robust, and lower cost solutions to real-world problems can be achieved. The main constituents of soft computing are fuzzy logic, neural networks, evolutionary computation, and probabilistic reasoning. Hence it is highly interdisciplinary, but not just a melange of different computational techniques; it is a partnership in which each technique contributes with a distinct and complementary methodology. Soft computing techniques have been successfully applied in many areas ranging from telecommunications, consumer electronics, and manufacturing systems to biology, agriculture, and social sciences.

As an engineering science, soft computing is very open to new and unconventional approaches that have theoretical characteristics which are not fully known. Its primary objective is to obtain working solutions for practical problems quickly, while proving theoretical properties is a secondary issue. Although this attitude produces the flexibility and swiftness needed in applications, it has also created the situation that many soft computing methods lack sound mathematical foundations. Its models are rarely checked rigorously with respect to their performance and robustness or monitored statistically. The underlying assumptions are rarely made explicit, thus impeding a reliable transfer to new applications.

Statistics, on the other hand, tries to establish objective conclusions based on experimental results. Performance is guaranteed by analysing the behaviour under all possible situations. Its methods have a sound mathematical basis. Robustness and generalisability are key issues. However, this creates the tendency to focus on models, whose mathematical properties are easy to analyse, thus constraining the eligible models and perhaps ruling out the most suitable or most promising one for a specific application. In addition, the strong mathematical orientation often leads to models that are difficult to understand and to apply for a non-mathematician, thus hindering plausibility checks by domain experts. Learning from the usually more intuitive and human-oriented way in which soft computing incorporates prior knowledge, treats uncertain or incomplete information, and handles changing environments can help to improve the use of complex statistical methods by practitioners.

## **B.2 Current state of knowledge**

The existing interaction between statisticians and soft computing researchers is very limited. In recent years it has mainly been focused on (simple) descriptive and exploratory statistics, which have even been used in parallel with soft computing techniques rather than in a true combination. Mathematical and inductive statistics have only scarcely been applied together with soft computing techniques, especially in applications.

A notable exception are support vector machines, which resulted from merging statistical learning theory with neural networks, thus giving rise to what has also become known as Vapnik-Chervonenkis theory. The general idea taken from neural networks is that the hidden layer(s) in a multilayer perceptron or radial basis function network can be seen as computing a transformation of the given data into a space, in which the addressed supervised learning problem (classification or regression) is easier to solve. In particular, the learning task can be restricted to linear classification (by separating hyperplanes) and to linear regression (provided the “right” transformation is carried out) in this space. The core ingredient taken from statistics is empirical risk minimisation, by which one tries to minimise the expected loss (with respect to a given loss function) that results from using a learned model to predict the outcome for new samples. In addition, the underlying assumptions (for instance, that the training set is an i.i.d. sample etc.) are made explicit. By combining these ideas, assessments of the expected performance (in probabilistic terms), estimates of the needed number of training samples, and new learning methods could be achieved. Today, support vector machines are among the most intensely researched areas in data analysis.

However, despite this success, which provides an enticing vision and a prototype for the possible emergence of new research areas, we are still far from exploiting the potential of combining statistical ideas with soft computing techniques. Even support vector machines cover only a small part of the wide range of neural network types, for most of which a proper statistical treatment is lacking or is in a fairly basic state. The probabilistic and statistical investigation of evolutionary algorithms is often restricted to highly simplified models (small populations, special fitness functions, very restricted genetic operations etc.) or provides only fairly general or somewhat vague insights into why they work. Combinations of other soft computing methods with statistical principles are even scarcer. At best cross-validation (to avoid overfitting and to assess the prediction performance) and simple statistical tests (to check whether a new method can be claimed to be better than an existing one) are used for model validation and evaluation. Deployed models are rarely monitored with statistical methods - if they are monitored at all - even though this offers powerful ways of improving the models.

Statistical techniques used in the induction process of soft computing models can, in most cases, safely be called trivial, and also focus mainly on classical machine learning approaches. Even though statistical model selection criteria in particular could provide a sound basis for many methods, little work in this direction can be found up to now.

Furthermore, several problems can be solved by means of either soft computing or statistics. For example, (binary) classification problems can be solved, for example, by inducing a fuzzy rule-based system, by training a neural network, or by logistic regression techniques. However, comparisons are usually restricted to methods originating from the same field (that is, soft computing methods are compared to other soft computing methods, statistical methods to other statistical methods). As a consequence, fairly little is known about the relative merits of methods originating from different fields. In addition, combined approaches (of which some, but not too many exist) may be able to maintain the strengths of both worlds and eliminate the weaknesses. Of particular interest is the statistical treatment of fuzzy data, which often occur in applications where objective measurement devices are not available or too costly to apply. However, the community working in this area is extremely small and publications are few.

Finally, a core concern of several soft computing approaches is interpretability. It is seen as highly important that the resulting model can be easily understood by a domain expert (who may not be a trained mathematician) and thus be checked for plausibility. One is even willing to give up prediction accuracy in order to obtain simple and interpretable models. This complements the statistical task of model selection, where the trade-off between a more complex model and a better fit to the data is carefully checked to avoid overfitting. However, soft computing and statistics emphasize different aspects. While statistics focuses on “statistical” complexity, which tries to capture the model complexity by counting the free parameters (as in the Akaike and Bayesian information criteria) or by evaluating a description of the model (as in minimum description length and minimum message length approaches), soft computing emphasizes the “psychological” simplicity of a model and its interpretability for human domain experts. Unfortunately, though, most existing approaches to measure this “psychological” simplicity are very heuristic in nature. For example, the complexity of a (fuzzy) rule-based system is often assessed by combining with some heuristic aggregation function the number of rules, antecedent terms, the parameters of fuzzy sets in these terms etc. There is also no proper comparison to “statistical” complexity that clarifies the differences and how the two notions should be weighted relative to each other (which could provide a way to handle the so-called accuracy-interpretability trade-off).

### **B.3 Reasons for the Action**

Since researchers in soft computing usually have a computer science background, while statisticians are educated as mathematicians, there tends to be a lack of common ground from the beginning. In addition, there is a lack of workshops and conferences which attract both statisticians and soft computing researchers, so that information exchange is scarce and accidental. As a consequence researchers in the two communities are usually not properly aware of the achievements and the state of the art in the other field. This has hindered joint research in the past, thus limiting possible results and leading to a waste of manpower. For example, it has happened several times that a method, which was well known in one research area, has been reinvented in the other, since a similar or even the same problem had been addressed. This can also lead to confusion, since the same methods often become known with different names.

It is clear, that researchers in both areas work on closely related problems and that both areas have powerful data analysis methods to offer. However, they emphasize different, though equally important aspects of data analysis. Therefore it is highly desirable to strengthen the dialogue between the statistics and soft computing research communities in order to cross-pollinate both fields. This will encourage mutual learning and will lead to a transfer of ideas and tools and a combination of known methods in order to merge the best of both fields. This intensified dialogue will lead to considerable scientific and technical progress, since the combination of soft computing techniques and statistics has a high potential and holds a large number of interesting scientific problems in stock. In the soft computing area, new intelligent systems will be developed with improved efficiency, robustness, generalisability, and an extended range of applications. In the statistical data analysis area, new methods will emerge from integrating fuzzy systems and expert knowledge in order to improve their interpretability and usability. In the best possible scenario whole new research fields could emerge, similar to how the field of support vector machines emerged from joining neural networks and statistical learning theory.

This COST Action will maximise its outcomes by a balanced combination of short-term missions, events to exchange ideas (i.e. focused seminars, semi-open workshops and large open conferences) and training courses. All of these activities will bring the two fields closer together and will permit to explore synergies between different computational and mathematical methods, to structure promising new ideas and research projects, to develop new research lines, to increase the multidisciplinary of European researchers, and to generate scientific and technical knowledge.

The COST scheme offers an excellent framework to carry out the conferences, workshops, training courses and scientific missions envisioned. The ESF also offers support to research conferences and networks and EU Framework Programme 7 funds research projects, coordination actions and scientific networks. However, the objectives of this Action require a scheme to foster the dialogue between two research communities across Europe and the COST funding permits to create a coordinated and connected series of activities. Individual conferences and coordination actions provide only limited means to open a fruitful exchange of ideas, transfer of tools and generation of new research lines, while the Action's objectives require a supervised series of activities such as the ones funded by COST. In contrast, networks are aimed at coordinating existing research groups and projects, while this Action's objective is to create new research ideas, projects and tools. The networking funding by COST could provide excellent opportunities to generate new research areas and capacities and is therefore the best funding scheme for the objectives aimed at by the proposers.

## **B.4 Complementarity with other research programmes**

This COST Action will provide new research lines in intelligent data analysis that could enable Europe to become a leader in the extraction of knowledge from the increasing amount of data collected, transferred and stored by modern information technologies. European researchers and companies could exploit the new data analysis tools leveraging the activities of this COST Action in new pan-European research projects funded by the Seventh EU Framework Programme (FP7). The new horizontal and wide-spreading research lines that will be generated by this COST Action will help to accomplish the seven challenges defined by the research theme Information and Communication Technologies of the Cooperation area of FP7. The stimulation of creative and multidisciplinary approaches provided by this COST Action will enable novel and imaginative research proposals that could be funded by the Ideas area of FP7. This COST Action will work in a more fundamental, inspiring and dialogue-based level than the more focused and applied research and networking activities funded by FP7. Finally, the coordinated and multidisciplinary approach of this COST Action will help Europe to move towards a world-leading knowledge-based economy.

This COST Action complements several previous COST Actions in the area of intelligent data analysis that address problems, which could be handled both by means of soft computing and statistics. For instance, in Action 274 (Theory and Application of Relational Structures as Knowledge Instruments), some soft computing and statistical techniques are used in order to solve the considered problems, although that Action has a much narrower focus. The running COST Action IC0602 (Algorithmic Decision Theory) is more closely related to this Action, because it combines researchers from different areas (such as decision theory, discrete mathematics and artificial intelligence) to develop new decision support tools. However, this COST Action takes a broader approach, because it does not focus on a particular problem (automatic decision), but aims at a wider variety of problems. An open and creative dialogue between soft computing and statistics communities will be initiated in order to generate several new research lines and projects combining soft computing and statistical data analysis strengths.

## **C. OBJECTIVES AND BENEFITS**

### **C.1 Main/primary objectives**

The main objective of the Action is to strengthen the dialogue between the statistics and soft computing research communities in order to cross-pollinate both fields, generate mutual improvement activities, spark new ideas, and trigger new research lines and projects.

The combination of soft computing and statistical approaches has the potential to transform the methodology used to generate solutions to real-world problems. Therefore the aim of this COST Action is to explore potential synergies between different computational and mathematical methods originating from these areas, to generate new scientific and technical knowledge by fusing the expertise of statisticians and soft computing researchers, and thus to develop new interdisciplinary research lines.

This wide and basic objective should start with a dialogue between researchers from soft computing and statistics in order to convey knowledge about the achievements and the state of the art in these two areas to each other and to make researchers aware of the shortcomings in their field that could be improved by methods and principles adopted from the other. In a second step, the most promising new research fields that could arise as a result of bringing together soft computing and statistical data analysis will be defined. Subsequently, new research projects and collaborations will be possible in order to develop more robust and universal data analysis methods with a sound statistical and mathematical basis, which preserve the flexibility, efficiency, and interpretability of soft computing techniques.

This Action covers a highly multidisciplinary field and targets a non-conventional and innovative research area in order to explore new research trends by nurturing new research communities. The COST Action will reinforce the multidisciplinary research at European centres by making it possible to deliver more robust and efficient data analysis tools to build knowledge-based applications for European economy and society. The successful implementation of the Action will permit to enhance European leading position in the field of intelligent data analysis. Derived beneficial effects will span basically all industrial, governmental, and academic sectors that use electronic data and require data and knowledge hidden in it to be used more effectively and efficiently. A successful COST Action will permit to generate at least two new techniques (comparable to support vector machines) combining the strengths of soft computing and statistical approaches. Furthermore, more than 10 research proposals in the area of soft computing and statistical methods are expected to emerge from this Action.

## **C.2 Secondary objectives**

- To improve the mutual awareness and knowledge of the achievements and the state of the art in statistics and soft computing among researchers in these areas. This includes crossing the language barrier separating the two fields, in which often the same term is used with a different meaning (like “classification”) and different terms are used for the same method or problem (for example, “discrimination” versus “classification”, “local models” versus “regression trees”).
- To initiate and sustain an open and intense dialogue between researchers in statistics and soft computing, thus providing a stimulating environment for sparking ideas that result from the synergies inherent in bringing these two fields closer together.
- To integrate soft computing and statistical data analysis approaches in a new research area that aims at combining their flexibility, interpretability, robustness and generalisability, since several weaknesses of existing methods in one field can be amended by techniques of the other.
- To develop specific new intelligent data analysis methods that integrate soft computing and statistical tools in order to generate novel, robust, and more efficient data analysis solutions, which have a sound mathematical basis and lead to interpretable, but still sufficiently accurate models.
- To enhance the capabilities of intelligent data analysis tools in order to leverage the increasing amounts of data collected, transmitted and stored by new information and communication technologies, converting them into useful knowledge that can be exploited to improve economic efficiency and social welfare.
- To coordinate researchers and teachers in soft computing and statistics fields in order to educate a new generation of multidisciplinary researchers and teachers in both areas in order to broaden the common ground, to ease information exchange, and to avoid waste of manpower.



- To disseminate the information and the results of the Action to the scientific community and to increase the awareness of technology developers and users of the new tools promoted.

### **C.3 How will the objectives be achieved?**

Soft computing is an interdisciplinary research area. This aspect will be strengthened by the integration of additional statistical methodologies in order to widen the scope of its applications and to ensure its robustness and efficiency. In addition, emphasising simplicity and interpretability will help to make statistical methods easier to use for practitioners. This COST Action intends to bridge the gap between soft computing (engineering science) and statistics (pure science) by creating a rich and open environment to develop new ideas, knowledge and applications. In the best possible scenario, whole new research fields could emerge, similar to how the research area of support vector machines emerged from combining neural networks and statistical learning theory.

In order to achieve the Action's goal, people in the soft computing community have to become better acquainted with state of the art statistical methodology. Statisticians, on the other hand, have to be made better aware of successful soft computing applications, in which models of unknown or insufficiently understood statistical characteristics have been used. These models need to be analysed in order to understand their success, to justify their applicability, and to specify conditions for a reliable transfer to other problems. It also needs to be understood better why these models are often more attractive to practitioners than standard statistical techniques. In particular, proper measures of "psychological" complexity need to be developed and related to known measures of "statistical" complexity.

A programme of regular bilateral visits, interspersed with seminars, workshops and conferences to disseminate individual achievements to other members of the communities is the best way to success in this direction.

### **C.4 Benefits of the Action**

The combination of soft computing and statistical data analysis techniques and the intensified dialogue between both research communities will encourage new scientific and technical progress by creating new research areas. This will produce new ideas, projects, computing tools and applications. The new intelligent systems that will arise from this COST Action will improve the efficiency and extend the range of applications of soft computing tools in order to strengthen the competitiveness of European industries and services.

From a scientific perspective, the main benefits expected from this COST Action are the following:

- Statistical evaluation of data, which was obtained by monitoring deployed soft computing solutions, can provide robustness and performance measures, which may be used to improve the models.
- Statistical analysis of soft computing models can provide guidelines about the appropriate soft computing techniques for different classes of problems and data.
- Identifying the assumptions underlying different soft computing techniques helps to predict their applicability and performance in different conditions. The results can be used for user support in semi-automatic software tools.
- Statistical investigations are directed towards practically relevant models, which may be less rigid from a theoretical point of view, but better complexity (in contrast to "statistical" complexity) will help proper model selection in applications.

- Transferring the more intuitive and human-oriented ways in which soft computing approaches incorporate prior knowledge, treat uncertain or incomplete information, and handle changing environments will help to improve the acceptance of (complicated) statistical methods by practitioners.
- New research fields, resulting from the combination of statistical approaches with soft computing techniques may result (in analogy to support vector machines).suited for humans (comprehensibility, intuitiveness, simplicity).
- New paradigms for the statistical evaluation of less rigid models will be developed.
- A better understanding of “psychological” factors.

### **C.5 Target groups/end users**

The Action will provide a series of coordinated activities to initiate and strengthen an interdisciplinary dialogue that permits to exchange knowledge and expertise between soft computing and statistical data analysis researchers more effectively and efficiently. In addition to the implied benefits of establishing such a stronger connection between the two research communities in order to generate improved multidisciplinary data analysis techniques, the COST Action will benefit engineers and technologists that will use the new and enhanced data analysis tools to improve current business processes and social services. The new data analysis tools will have a significant impact in applications in many different sectors such as personalised healthcare, environmental monitoring, security, web intelligence or cognitive robotics.

## **D. SCIENTIFIC PROGRAMME**

### **D.1 Scientific focus**

The scientific programme is focused on a pragmatic approach to develop new and specific data analysis tools combining soft computing and statistical approaches, instead of searching a top-down approach to fuse soft computing and statistical data analysis. Such a top-down approach, although it may appear more principled and promises a strong underlying theory, always bears the danger of getting lost in theoretical discussion and technical subtleties before producing practically relevant results, which are among the Action’s most prominent goals. This COST Action, since it is positioned at the border between two related fields that emphasize different aspects of data analysis, is expected to trigger inspiration and to spark several new ideas, as a natural result of seeing things from a different angle and shifting emphasis to new aspects. The pressure of having to solve all fundamental problems before progressing with practical applications will be avoided. This COST Action will thus provide an open and interdisciplinary framework that will help to generate a series of promising and practically relevant research projects with ideas from both fields. By bringing together top-level researchers and focusing on the particular strengths and contributions of soft computing and statistics, the development of new methods for data analysis will incorporate and combine several cutting-edge research areas.

The discussions and activities of the working groups of this COST Action could be based on the following six initial and particularly relevant research topics. This list is not completely exhaustive and this COST Action will keep the set of topics fairly open, in order to be able to incorporate any new ideas that result from the intense dialogue as well as information and knowledge exchange between the research areas of soft computing and statistics. These topics should rather be seen as seeds to start the discussion and as an outline of the potential achievements that the COST Action

will lead to. Nevertheless, the six topics described below are some of the most promising areas, where the joint activities of soft computing researchers and statisticians can be expected to produce high-quality results not only quickly, but also in such a way that practically relevant projects can be established.

#### Topic 1: “Statistical Validation and Monitoring of Soft Computing Models”

Currently only very simple statistical techniques are applied to validate and to assess soft computing models. However, a better, objective validation from the point of view of inferential statistics, which takes the state of the art into account, can lead to an integral methodology with interesting results and high application potential. For example, a good way to cope with the problem of overfitting (which means adapting a possibly a model with too many degrees of freedom to accidental and spurious properties of the training data set) is to consider a true underlying model (describable, of course, by a soft computing technique) from which the observed data has been sampled, but then this data has been disturbed by random noise. This view of the situation allows us to apply inferential statistical techniques to assess the performance of the model adequately and to handle the given data properly. Furthermore, re-sampling techniques like bootstrap and bagging provide means to measure the goodness of fit of the model, to objectively compare different learning methods, and to improve performance by combining multiple models, thus removing or at least mitigating the bias of the individual models. Statistical analysis can also be highly useful when monitoring the resulting system once it is deployed in an application. System failures, change points, loss of fit etc can be detected and maybe even predicted by analyzing the time series of the inputs and outputs and may lead to statistics-based model adaptation techniques.

#### Topic 2: “Model Selection and Validation for Neural Networks”

Even though support vector machines rely on statistical principles, only part of the power and flexibility of neural networks is used in combination with statistical techniques. For example, in support vector machines the kernel functions, by which the coordinate transformation is achieved, are not moved away from the data points and their parameters (like the “radius” or “window width” of the kernel functions) are not adapted during training. Neural networks, on the other hand, if trained with some gradient descent scheme like error back-propagation, allow for such adaptations: in principle, any parameter can be seen as an argument of the error function and thus be adapted. It is obviously desirable to have methods that can provide bounds on the expected performance in these cases. Statistical methods can also be useful in the design and training of neural networks. For instance, the number of neurons in the hidden layer(s) of both multilayer perceptrons and radial basis function networks and the structure of the network can be examined from a statistical point of view, providing rules for the best choice. Objective methods for comparing different network structures can be developed in terms of statistical hypothesis tests. On the other hand, since neural networks are more flexible than most of the usual predictive statistical methods, it is desirable to analyze possible combinations in order to balance predictive accuracy (which tends to be higher for more flexible methods) and interpretability as well as robustness (which prefers simple models).

#### Topic 3: “Evolutionary Algorithms as Estimators”

From a statistical point of view, evolutionary algorithms can be interpreted as iterative estimation procedures if one assumes an underlying population model and sees the fitness functions as a distribution the mode of which is to be estimated. The convergence of an evolutionary algorithm to the true or best solution can then be seen as the consistency of this estimator. By similar reasoning other statistical properties of estimators (like unbiasedness, efficiency, sufficiency etc.) can be transferred to the evolutionary algorithm framework, which may provide results about the expected number of generations to reach a solution (of given minimum quality) or the expected solution

quality after a given number of generations. Although there exists some seminal work along these lines, concrete results are often restricted to highly simplified models of genetic algorithms (small populations, special fitness functions, very restricted genetic operations etc.) or provide only fairly general or somewhat vague insights into why genetic algorithms work. It is definitely desirable to achieve a better and formally grounded understanding of the search as it is carried out by an evolutionary algorithm, for which statistical techniques are very well suited, since the evolutionary search is essentially a guided random process.

#### Topic 4: “Estimation of Distribution Algorithms”

Estimation of distribution algorithms are a newer addition to the range of evolutionary algorithms, which make stronger use of statistical principles, although they are still based on the classical paradigms. Instead of forming new individuals for the next generation by recombining the (genetic descriptions of) the individuals of the current generation (as it is the case in all “classical” evolutionary algorithms), estimation of distribution algorithms use the current population to estimate a distribution on the search space, which describes (in probabilistic terms) where good solutions may be located and what the quality of these solutions can be expected to be. The next generation of individuals is then obtained from this estimated distribution by random sampling, so that there is no direct connection between the (genetic description of the) individuals of two consecutive populations, thus enhancing flexibility and sometimes convergence speed, but also introducing new problems. Since distribution estimation and sampling are used, statistical techniques can be expected to solve these problems while maintaining the advantages. The investigation of the properties of estimation distribution algorithms (w.r.t., for example, expected number of populations, expected solution quality) is in a somewhat better state than for standard evolutionary algorithms, but still needs a lot of improvement.

### Topic 5: “Statistics with Fuzzy Data”

In many applications we meet data that appears to be crisp, but is actually derived from a human perception or estimation of some quantity. This is usually the case if an objective measurement device is not available (for example, because there is no simple underlying physical quantity or no appropriate physical measurement process is known) or carrying out a precise physical measurement would be too costly (at least w.r.t. the benefit that can be expected from the higher precision of the result), so that a human estimates the quantity or only provides a rating on a simple ordinal scale. Treating such data as if it were crisp is obviously inappropriate and it can be shown that loses information. However, classical statistics have few other possibilities if one does not want to invoke families of distribution functions, which may not even be applicable if there is no underlying measurable property. Therefore, specialized statistical techniques are needed, which take the fuzzy character of such data properly into account. The existing techniques, which transfer statistical notions like expected values and linear regression to the fuzzy setting, are far from sufficient and need expansion and improvement.

### Topic 6: “Psychological versus Statistical Complexity”

Soft computing usually emphasizes the “psychological” simplicity of a model, that is, how intuitive and easy to understand a model is for a human, who is not a data analysis expert. For this type of simplicity it is more important that the qualitative characteristics of, for instance, a functional relationship are captured than that the correct function class and the best estimates of the parameters are chosen. Two functions of considerably differing “statistical” complexity (in terms of free parameters or the description length of the model) can be equally simple “psychologically” if they exhibit the same qualitative behaviour. Furthermore, a large number of simple models, each describing a certain subspace, may be preferable to a complex overall model, even if the latter has fewer parameters. Finally, existing background knowledge, due to which only deviations from an expected behaviour need to be described instead of specifying a full model, lead to considerable differences compared to “statistical” complexity. Hence measures for the “statistical” complexity of a model are not suitable to measure “psychological” complexity. Since “psychological” complexity is what really counts in applications, but existing measures are essentially heuristic in nature, a better understanding of this type of complexity and its difference to “statistical” complexity is needed.

These six topics are built on top of current world-class research projects carried out by European researchers and will increase their success by the generation of new multidisciplinary research areas led by emerging young scientists. These topics will require an intense exchange of knowledge and a creative environment that will be possible by increasing the networking opportunities to extend this Action to a broader base of countries.

In order to avoid too much diversification, which could hinder the broader dialogue between researchers, the topics listed above will be initially classified into three Working Groups, each of them covering two topics. The first Working Group (WG) will be concerned with topics 1 and 2, thus focusing on validation and monitoring, in particular with respect to neural networks. The second WG will cover the two topics related to evolutionary computation (topics 3 and 4), while the third WG will target topics 5 and 6, both of which are oriented at human-friendly data processing and model generation.

## D.2 Scientific work plan – methods and means

During the first semester of the COST Action, the participation of new scientists and research groups will be encouraged. A common scientific programme based on the topics and objectives described in section D.1 will be agreed upon by all participants. The scientific work plan will be based on:

- Working Group Meetings to explore new opportunities to combine soft computing and statistical approaches
- Workshops and Seminars with a broad range of experts to address specific areas
- Tailored and structured scientific exchanges between research groups
- A large annual Conference
- Two training schools to educate a new generation of multidisciplinary researchers.

This Action will organise a yearly conference on “Soft Computing and Statistical Methods” in order to promote the dialogue between the research communities and to disseminate the new scientific paths envisaged. The first conference will be held in June 2008 in Oviedo, Spain, and serves the purpose to give the Action’s goals and objectives a high visibility among data analysis researchers, to make these researchers better aware of the potential of joining soft computing and statistics, to establish an overview of the state of the art in such interdisciplinary approaches, and to identify upcoming trends and research lines. These trends and research lines will then be chosen as the major topics of interest for the conferences that are to be held in the following years, revised, of course, according to the new developments that emerge over time. This COST Action envisages establishing these conferences as one of the main "exchange markets" for top-level research in hybrid approaches originating from soft computing and statistics, thus raising Europe’s standing in the practically highly important area of data analysis.

In addition to the yearly conference, a series of smaller discussion workshops and seminars will be held quarterly to promote more focused dialogues and develop specific collaboration opportunities. The research topics identified above may serve as a guideline for choosing the topics of the initial set of these workshops. These workshops will be coupled to workshops at the annual conference, which serve as means to make the outcome of the discussions and seminars known to the whole data analysis community.

An essential vehicle of this COST Action will be to proactively support the exchange of visitors and short term research stays in order to increase the detailed transfer of ideas and tools between the soft computing and statistical data analysis research communities. This Action will also seed new interdisciplinary project proposals to the Seventh Framework Programme by generating mutual understanding and trust between both research areas.

At the end of the COST Action, at least one workshop specially targeted at industry will be organised to disseminate the new data analysis tools arising from the Action to the private sector. This workshop will be strongly focused on applications of the developed methods. Furthermore, training schools will be supported to train interdisciplinary young scientist in both scientific areas.

Finally, a web page will be launched and maintained throughout this COST Action in order to provide an open forum to discuss and disseminate the emerging ideas and potential new tools arising from the enhanced dialogue between soft computing engineers and statisticians. This web page will be connected to a newsletter in order to inform interested people quickly about new

developments, upcoming workshops and conferences, successful industrial applications, open positions etc.

The main deliverables of this Action will be:

- More than 60 Short-Term Scientific Missions (at least one per signatory country per year).
- 4 large (>50 participants) and open conferences (one per year)
- 2 Training Schools
- 16 Workshops and Seminars
- Increased number of multi-disciplinary researchers that combine soft computing and statistical training in academia and industry.
- New collaborations between researchers in different research communities.
- More than 10 research proposals that will generate innovative soft computing techniques.
- More than 10 reports summarising the results of the Action (at least two per Working Group and one per year for the whole Action)
- Basic scientific knowledge and improved technical developments by opening new areas of exploration. This will lead to a number of publications and presentations in conferences.

The kick-off meeting of this COST Action will revise these assignments and also define a major area and an additional area for each member, as well as chose a coordinator for each Working Group. The initial objective of the Working Groups is to collect and distribute information about the particular expertise and research interests of the members in order to identify the most promising and best supported specific research topics. In a second step, a plan for small task forces (3 to 4 people), which are working on a specific subject, will be set up. These task forces should be in constant contact and meet at least four times a year in order to coordinate their efforts. The work groups serve as a coordinating instance for these task forces, making sure that the subjects of the different task forces are complementary. Each task force is expected to generate at least one concrete research project proposal.

## **E. ORGANISATION**

### **E.1 Coordination and organisation**

The partners of this COST Action will form a Management Committee (MC). The current participants envisaged for this COST Action cover more 9 countries and it will be open to new partners from other COST countries. The participation in the COST Action during the first stage will be promoted by actively advertising it in relevant conferences, journals, web pages and mailing lists, so as to permit to expand the geographical base of the Action and enhance the dialogue and the generation of new ideas.

The Management Committee will be the top-level supervisory body of the COST Action that will coordinate the COST Action activities following the “Rules and procedures for implementing COST Actions”. Working Group Leaders will be appointed to coordinate their activities. The MC will define the broad themes to be studied by the Working Groups and will supervise their activities in order stimulate their activities. The MC will prepare and present an annual scientific programme of the Action, annual progress reports and the final report after the completion of the Action. An Administrative Coordinator will be in charge of maintaining the communications with the National COST Coordinators, of expanding the network activities by involving new participants and of the interaction with the COST Office and other research programmes.

The Management Committee will meet at least every six months and its main responsibilities are to plan and design the activities of the Action, supervise their implementation, monitor progress, allocate resources and reconcile conflicting viewpoints. The Action aims to integrate the expertise from two different research areas to improve data analysis techniques. It will bring together scientists with a wide variety of expertise, encouraging new ideas and projects, providing support to emerging contacts and networks, facilitating the exchange of knowledge, and encouraging multidisciplinary training for young scientists. A Training School will be organised in second and fourth years of the Action.

The monitoring and evaluation of the implementation of this COST Action will be an essential duty of the Management Committee so that, in the end, the benefits from the new research projects and from the data analysis tools generated will be disseminated and exploited by European commercial and social organizations. The MC will have the mission to monitor the effective and correct development of the COST Action through its regular meetings. The progress in each scientific field and the progress of the different activities (especially of the short-term missions involving early-stage researchers) will be recorded in an annual Progress Report. The MC will also confirm and adjust the planning of the Action in order to increase the value of its achievements and to correct the deviations from the original plans. The MC will evaluate any adjustment or re-examination of the Action plan which may contribute to the improvement of its achievements and impact.

This COST Action will be coordinated as a dynamic organisation with numerous interactions among the different Working Groups to build an interdisciplinary approach to data analysis combining soft computing and statistical methods. The Management Committee and the Working Groups will also hold virtual meetings to deal with day-to-day management issues using web based tools every two months. Research in this COST Action will be primarily based in the Working Groups with a strong emphasis in horizontal contacts across the Action and coordination between Working Groups to exchange research directions, information and expertise.



The first MC Meeting will organise the Working Groups, plan the first annual conference and discuss proposals for the workshops and seminars. The MC will also nominate persons with responsibilities to contact and exchange of ideas with other COST Actions, EU projects and international programmes.

The establishment of a broader network and the attraction of new scientists to the research area promoted by this COST Action will be achieved through the planned annual conference, the regular update of the COST Action website and quarterly e-mail newsletters. The proactive support to the regular exchange of early-stage scientists between research groups will permit to enhance the links between organizations. Furthermore, the organisation of special workshops and seminars meetings that will focus on particular aspects of data analysis will permit to explore in depth specific topics and ideas. A series of reports summarizing the activities and results of this COST Action will be published and disseminated. The COST Action will promote scientific exchange of ideas and personnel and new information about research and opportunities through meetings, publications and websites.

## **E.2 Working Groups**

The work in this COST Action will be organised in three Working Groups: the first group (Group A) will focus on the first two topics, which are both related to statistical model selection and model validation, even though the second topic concentrates on neural networks as a special soft computing technique. The second group (Group B) will cover topics 3 and 4, both of which deal with evolutionary algorithms and their statistical analysis. Finally, the third group (Group C) will work on topics 5 and 6, which emphasize interpretability and human-oriented data analysis.

The three Working Groups will coordinate the activities in these research areas, but also new areas resulting from ideas sparked by the dialogue that will be created and sustained between the two research communities of soft computing and statistics. The new ideas emerging from the activities of the Action will be analysed collaboratively by the experts from soft computing and statistical data analysis fields. The specific outputs of the scientific programme will arise from the actual dialogue and analysis carried out by the working groups following a bottom-up approach.

## **E.3 Liaison and interaction with other research programmes**

Frequent contacts with other COST Actions and with coordinators of projects and networks under Seventh Framework Programme (e.g. first ICT theme calls and Marie Curie training networks) will ensure complementarity with the COST Action and will provide potential invitees for its activities.

## **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 participants in the Action will be encouraged to promote the involvement of female professionals in the different activities, and to adhere to gender equality when selecting and/or appointing new recruits. The gender balance will be overseen by the Management Committee.

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. The members of this COST Action will be encouraged to promote the participation of early stage researchers in the different activities of the

Action. When recruiting young scientists, the participants in this COST Action will follow the European Charter for Researchers and Code of Conduct for Recruitment.

## **F. TIMETABLE**

This COST Action is planned for four years. The Management Committee and the Working Groups will meet at least twice a year. In addition, there will be a kick-off meeting in Year 1 after the entry into force of the MoU to plan the initial organisation and initial activities of the Action.

The following timetable describes the basic activities to be carried out by the Action:

- Year 1:
  - Start-up Conference
  - At least three specific Workshops organised by the Working Groups
  - More than 10 Short-Term Scientific Missions
- Year 2:
  - 2nd Annual Conference
  - At least three specific Workshops organised by the Working Groups
  - Training School
  - More than 15 Short-Term Scientific Missions
  - More than 2 research proposals
- Year 3:
  - 3rd Annual Conference
  - At least three specific Workshops organised by the Working Groups
  - More than 15 Short-Term Scientific Missions
  - More than 3 research proposals
- Year 4:
  - Concluding Conference
  - At least three specific Workshops organised by the Working Groups
  - Training School
  - Workshop for Industry
  - More than 20 Short-Term Scientific Missions
  - More than 5 research proposals

## **G. ECONOMIC DIMENSION**

The following COST countries have actively participated in the preparation of the Action or otherwise indicated their interest: ES, DE, PL, CY, UK, FR, IT, AT, NL. On the basis of national estimates, the economic dimension of the activities to be carried out under the Action has been estimated at 14 Million € for the total duration of the Action.

This economic dimension takes into account the participation of 20 experienced scientists, 10 research assistants and 40 early stage researchers and 10 administration and support personnel.

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?**

All the members of the MC will play a pro-active role in the dissemination of information about the activities, events, reports and publications of this COST Action. The findings and recommendations of this COST Action will be disseminated to the following target audiences:

- Scientific communities involved in data analysis with emphasis in researchers in the fields of soft computing and statistics field.
- Relevant research funding authorities (e.g. Seventh Framework Programme, ESF, The Human Frontier Science Program)
- Current and potential users in private and public sectors of the new data analysis tools proposed.
- European level policy makers, governmental organizations and regional authorities promoting and supervising the development and the application of new information and communication technologies.
- Social groups and the general public to diffuse the importance, impact and implications of the new data analysis research lines proposed

### **H.2 What?**

Reports about all the meetings and events organised by this COST Action will be submitted to the MC for evaluation and to decide subsequent activities. There will be three different types of documents:

- Progress reports related to the COST Action activities
- Analytical reports with new interdisciplinary ideas and emerging research lines generated by the activities of this COST Action.
- Reviews of the scientific progress made in the fields of this COST Action

The preparation of these documents by experts is an essential component of the COST Action. The documents will provide up-to-date information and comments about new data analysis tools merging concepts from soft computing and statistics.

The main dissemination activities of the Action will be:

- Workshops, seminars and conferences organised by the MC
- Talks in other national and international conferences
- Papers in peer-reviewed scientific and technical journals
- Posting the documents described above on the public website of this COST Action
- Internet discussion forum and e-news mailing list for targeted scientific audiences

### **H.3 How?**

One of the main methods for disseminating the COST Action information will be via the Action website. This website will contain basic information about the Action and its progress including continuously updated and comprehensive project documentation. It will also contain links to relevant contextual material and to other relevant initiatives. A list of all publications, papers in conferences and documents of the COST Action will be produced yearly and will be disseminated in both paper and electronic forms. This list will include publications from members of the COST Action and publications arising from joint research projects generated by this COST Action.

An internal website will be managed by the Administrative Coordinator with information provided by the MC and by all the Working Group leaders. It will contain official documentation (e.g. MoU), the detailed work plan, the working documents and reports, and any useful information for the participants in the Action, such as an agenda with internal and relevant external events or information about potential Short-Term Scientific Missions and open positions for early-stage researchers in partner institutions.

It will be the responsibility of the MC members for each country to maintain and update a list of key scientists and research centres in that country and ensure that the list is available to other members of the MC. MC members will be responsible for ensuring that those interested in the COST Action are aware of activities, events and documents.

Two e-mail lists will be available, one for communication between Action participants and one for individuals or organisations that can be informed on specific progress of the Action (previously approved by the MC) without being directly involved in it. The second list will be open to anyone who wishes to join it via an electronic application form available on the project website as well as via expressions of interest gathered in attended events. An open and interactive web discussion forum will permit a flexible dialogue tool for interested scientists and stakeholders.

A general conference will be organized yearly where all the participants will meet with external scientists and discuss common problems and initiatives, and present the results of their collaborations. This COST Action will coordinate its activities with the Soft Methods in Probability and Statistics International Conferences and with the European Research Consortium for Informatics and Mathematics. The COST Action will invite to the Annual Congress representatives from:

- IEEE (in particular. the IEEE Conference on Data Mining – ICDM)
- EUSFLAT (European Society for Fuzzy Logic and Technology)
- ACM SIGKDD (Association for Computing Machinery: Special Interest Group Knowledge Discovery in Databases)

- BISC (Berkeley Initiative in Soft Computing)
- Fuzzy Logic Laboratorium Linz - Hagenberg
- SCI2S (Soft Computing and Intelligent Information Systems, University of Granada)

Specialized Workshops and Seminars focused in specific topics will be organised each year by the Working Groups, and early-stage scientists will be particularly encouraged to attend these events. Furthermore, the COST Action will aim to present ideas and results in the major international conferences and symposia about data analysis (e.g. Soft Methods in Probability and Statistics International Conference, IEEE Conference on Data Mining).

The MC will generate a detailed dissemination plan in the first semester of the Action. This plan will be evaluated and updated through the duration of the Action.

---