BELGIAN SCIENCE POLICY

Wetenschapsstraat 8 rue de la Science
B-1000 BRUSSELS
Tel. +32 2 238 34 11 ▪ Fax +32 2 230 59 12
www.belspo.be

Interuniversity Attraction Poles (IAP)
Phase VI

2007 – 2011

ANNEX I

TO CONTRACT P6/03

TECHNICAL SPECIFICATIONS : SECTION I

Information on the network

Title of the project : **Statistical Analysis of Association and Dependence in Complex Data**
Name of the coordinator : Prof. Léopold Simar Institution : Université catholique de Louvain

I. 1. NETWORK COMPOSITION

BELGIAN PARTNERS *

Coordinator : Partner 1 (P1) Name : Simar, Léopold Institution : Université catholique de Louvain Institution's abbreviation : UCL	Partner 8 (P8) Name : Institution : Institution's abbreviation :
Partner 2 (P2) Name : Van Mechelen, Iven Institution : Katholieke Universiteit Leuven Institution's abbreviation : KUL-1	Partner 9 (P9) Name : Institution : Institution's abbreviation :
Partner 3 (P3) Name : Lesaffre, Emmanuel Institution : Katholieke Universiteit Leuven Institution's abbreviation : KUL-2	Partner 10 (P10) Name : Institution : Institution's abbreviation :
Partner 4 (P4) Name : Duchateau, Luc Institution : Universiteit Gent Institution's abbreviation : UG	Partner 11 (P11) Name : Institution : Institution's abbreviation :
Partner 5 (P5) Name : Veraverbeke, Noël Institution : Universiteit Hasselt Institution's abbreviation : UH	Partner 12 (P12) Name : Institution : Institution's abbreviation :
Partner 6 (P6) Name : Institution : Institution's abbreviation :	Partner 13 (P13) Name : Institution : Institution's abbreviation :
Partner 7 (P7) Name : Institution : Institution's abbreviation :	

* Mention only **one** name per partner. The person listed here should be the one in charge of the operational aspects of the project. Indicate the full name (family name + first name) of the partner.

EUROPEAN PARTNERS * (if applicable)

EU-Partner 1 (EU1) Name : Antoniadis, Anestis Institution : Université Joseph Fourier Institution's abbreviation : UJF Country : France	EU-Partner 3 (EU3) Name : González Manteiga, Wenceslao Institution : Universidad de Santiago de Compostela Institution's abbreviation : USC Country : Spain
EU-Partner 2 (EU2) Name : Eilers, Paul Institution : Universiteit Utrecht Institution's abbreviation : UU Country : The Netherlands	EU-Partner 4 (EU4) Name : Kenward, Mike Institution : London School of Hygiene and Tropical Medicine Institution's abbreviation : LSHTM Country : United Kingdom

* Mention only **one** name per partner. The person listed here should be the one in charge of the operational aspects of the project. Indicate the full name (family name + first name) of the partner.

I. 2. TITLE AND SUMMARY OF THE PROJECT

Indicate clearly and briefly the project's major objectives and provide a concise description of the project.

A. Title and summary in English (2 pages maximum)

Statistical analysis of association and dependence in complex data

One key aim of statistics is to analyse in an appropriate way the dependence and association present in a dataset. The data that are collected nowadays to analyse these dependence structures, are often of a complex nature and also the research questions are of an ever increasing complexity. This requires the construction of new models, or the adaptation of existing models, which is a challenging task. The development of new methods and intensive interaction between experts will also be required to cope with these complex data.

The global objective of the network is to develop new models and methodological tools to do inference and to analyse these complex data structures. To achieve this goal, the network will be structured in five interlocking workpackages, devoted to different types of complex data structures, that will be studied within the network.

1. *Workpackage 1: Multivariate data with qualitative constraints*
   In statistical analysis, the quantity one wants to estimate or test a hypothesis about, often satisfies certain natural qualitative constraints, which one has to take into account, if one wants to make fully use of the nature of the data. Examples of constraints include boundaries (e.g. in frontier analysis), monotonicity, convexity, ellipticity or independence of unobserved components, unimodality and sparsity. Qualitative constraints also arise when using dimension reduction techniques, analysing functional data or dealing with inverse problems. A wide spectrum of statistical techniques is required to analyse this type of complex data. Building further on the results obtained in the previous phase (Phase V) of the network, new challenging research questions will be studied in this area, like e.g. the estimation of stochastic boundaries, the use of dimension reduction techniques with incomplete data, etc.

2. *Workpackage 2: Temporally and spatially related data*
   Methods based on principal component analysis to forecast single variables on the basis of a large panel of time series are widely studied in the economic literature, and will be further explored and compared. Also, the work on dynamic factor models, already initiated during Phase V of the network, will be further pursued. Another topic the network is very much interested in, is the study of non-stationary time series. The achievements in the domain of locally stationary time series, which have been expensively studied during Phase V, will be further developed with a new emphasis on goodness-of-fit tests, adaptive inference and time-space modelling. A unified approach will be taken in order to jointly address several types of non-stationarity (such as time-varying coefficients models, unit roots models,…)

3. *Workpackage 3: Incomplete data*
   Several types of incompleteness in the data arise in practice: missing data, censored data (right censoring, interval censoring, ...), truncated data, misclassified data, coarse data, ... The main focus will be on censored and missing data. In particular, the work on nonparametric estimation with censored data and on frailty models studied already in detail during Phase V, will be further pursued, and it will be studied how repeated data and survival data can be jointly modelled. The focus in the analysis of missing data will be on sensitivity analysis and on the combination of latent structures and mixed and mixture modelling ideas.

Another research topic that the network is very much interested in, is the estimation of causal effects of observed exposures, measured with error, in randomised studies.

4. *Workpackage 4: Data with latent heterogeneity*
   Unobserved heterogeneity can be modelled in different ways. A natural and common way to model this heterogeneity is by means of mixed models, which have been extensively studied during Phase V. The gained expertise opens the door to study in particular mixed models with partially specified residual dependencies conditional on the values of the random effects, and generalised linear mixed models. For the latter model, flexible models for the random effects distribution will be investigated, like mixtures of normals to approximate a B-spline basis.

5. *Workpackage 5: Highdimensional and compound data*
   In many applications dealing with genomics, proteomics, metabolomics, etc., data to be dealt with typically include a very large number of variables. The information in such datasets often contains a lot of noise in the form of irrelevant information and masking variables. Additionally, the information can come from different types of sources. Major challenges for these datasets pertain to the detection of structure in highdimensional problems, the filtering of noise and irrelevant pieces of information, multiple testing in the presence of a high number of variables, and drawing much stronger inferences by means of suitable combinations of different data pieces at hand. To deal with these challenges, suitable non- and semiparametric techniques (including smoothing methods) will be developed, that can be used for noise reduction, appropriate dimension reduction and clustering techniques (including methods of mixture modelling) for highdimensional two- and multiway data, and data fusion methods in which several pieces of multiblock multiset data are jointly modelled.

Cross-links between the five workpackages will be established on at least three different levels:

I. *Interlocking complexities in the data*
   In practical situations, often data are encountered that imply interlocking complexities as studied in several workpackages. Methods will be developed for dealing in an appropriate way with such compound complexities. This will require more than a mere concatenation of results as obtained from different workpackages, because when considering e.g. missing data in a multivariate context new kind of complexities will have to be dealt with.

II. *Common modelling approaches*
   The study of dependence is a recurrent topic across all workpackages. In this regard different approaches will be taken to deal with dependence modelling, including the use of copula models, regression models that are based on e.g. various kinds of dimension reduction approaches, and random effects models, like e.g. generalised linear mixed models. Within the network, those different approaches will be compared, both on a theoretical level and on the level of analyses of several benchmarking datasets.

III. *Common methods and tools*
   The different workpackages will rely on a common set of tools, including techniques of kernel smoothing, semiparametric inference, Bayesian inference, optimisation, randomisation and bootstrap. Findings on these tools will be exchanged among workpackages, and generic results on these tools will be aimed at, allowing their use in a broad range of contexts.

B. Title and summary in Dutch (2 pages maximum)

Statistische analyse van associatie en afhankelijkheid in complexe data

Eén van de kerndoelen van statistiek is het op een geschikte manier analyseren van de afhankelijkheid en associatie aanwezig in een dataset. De data die tegenwoordig verzameld worden om deze afhankelijkheidsstructuren te analyseren, zijn vaak complex en ook de onderzoeksvragen worden steeds complexer van aard. Dit vereist het construeren van nieuwe modellen, of het aanpassen van bestaande modellen, wat een hele uitdaging is. Het ontwikkelen van nieuwe methoden en intensieve interactie tussen experten zal ook noodzakelijk zijn om met deze complexe data te kunnen werken.

Het globale objectief van het netwerk bestaat uit het ontwikkelen van nieuwe modellen en methodologische technieken om deze complexe datastructuren te analyseren. Om dit doel te bereiken, zal het netwerk gestructureerd worden in vijf aan elkaar gekoppelde werkpakketten, gewijd aan verschillende types van complexe data die bestudeerd zullen worden in het netwerk.

1. *Werkpakket 1: Multivariate data met kwalitatieve beperkingen*
   In statistische analyse, voldoet de grootheid die men wenst te schatten of waarvoor men een test wil uitvoeren, vaak aan bepaalde natuurlijke kwalitatieve beperkingen, die men in rekening moet nemen als men volledig wil gebruik maken van de aard van de data. Voorbeelden van zulke beperkingen kunnen gevonden worden in onder meer grenzen (bv. in 'frontier' analyse), monotoniciteit, convexiteit, ellipticiteit of onafhankelijkheid van niet-geobserveerde componenten, unimodaliteit en schaarsheid. Kwalitatieve beperkingen treden ook op bij het gebruik van dimensie-reductie technieken, bij het analyseren van funtionale data of bij het verwerken van inverse problemen. Een breed spectrum aan statistische technieken is vereist voor het analyseren van dit type complexe data. Door verder te bouwen op de resultaten die verkregen zijn tijdens de vorige fase (Fase V) van het netwerk, zullen nieuwe uitdagende onderzoeksvragen bestudeerd worden in dit gebied, zoals bv. het schatten van stochastische grenzen, het gebruik van dimensie-reductie technieken met onvolledige data, enz.

2. *Werkpakket 2: Tijd- en ruimte-gerelateerde data*
   Methoden gebaseerd op principaal component analyse voor het voorspellen van een variabele op basis van een grote groep tijdreeksen, zijn wel bestudeerd in de economie literatuur, en zullen verder worden onderzocht en vergeleken. Het werk i.v.m. dynamische factor modellen, reeds bestudeerd tijdens Fase V van het netwerk, zal ook voortgezet worden. Een ander onderzoeksonderwerp waarin het netwerk heel erg geïnteresseerd is, is de studie van niet-stationaire tijdreeksen. De bereikte resultaten in het domein van locaal stationaire tijdreeksen, welke uitgebreid bestudeerd zijn geweest tijdens Fase V, zullen verder ontwikkeld worden, met nadruk op goodness-of-fit testen, adaptieve inferentie en het modelleren van tijd-ruimte. Een verenigde benadering zal aangewend worden om verschillende types van niet-stationariteit gezamenlijk te bestuderen (zoals modellen met tijdsafhankelijke coefficienten, eenheidswortels modellen, ...).

3. *Werkpakket 3: Onvolledige data*
   Verschillende types van onvolledigheid in de data treden op in de praktijk : ontbrekende data, gecensureerde data (rechtse censurering, interval censurering, ...), getrunceerde data, verkeerd geclassifieerde data, verruwde ('coarse') data, ... De nadruk zal gelegd worden op gecensureerde en ontbrekende data. In het bijzonder, zal het onderzoek i.v.m. het niet-parametrisch schatten met gecensureerde data en i.v.m. 'frailty' modellen, reeds uitgebreid bestudeerd tijdens Fase V, verder onderzocht worden. Bovendien zal bestudeerd worden hoe gerepeteerde data en overlevenigsdata samen kunnen gemodelleerd worden. De nadruk bij het analyseren van ontbrekende data zal liggen op sensitiviteitsanalyse en op de combinatie van latente structuren en ideëen van gemengde modellen en mengverdelingen. Een ander onderzoeksonderwerp waarin het netwerk erg geïnteresseerd is, is het schatten van causale effecten van geobserveerde blootstellingen, gemeten met fout, in gerandomiseerde studies.

4. *Werkpakket 4: Data met latente heterogeneïteit*
   Niet-geobserveerde heterogeneïteit kan gemodelleerd worden op verschillende manieren. Een natuurlijke en vaak gebruikte manier om deze heterogeneïteit te modelleren, is het aanwenden van gemengde modellen, welke uitgebreid bestudeerd werden tijdens Fase V. De verkregen expertise opent de deur voor het bestuderen van gemengde modellen met partieel gespecifieerde afhankelijkheid van de residus, conditioneel op de waarden van de random effecten en voor het bestuderen van gegeneraliseerde lineaire gemengde modellen. Voor het laatste model zullen flexibele modellen voor de verdeling van de random effecten bestudeerd worden, zoals mengelingen van normale verdelingen om een B-'spline' basis te benaderen.

5. *Werkpakket 5: Hoog-dimensionele en samengestelde data*
   In vele toepassingen die te maken hebben met genomics, proteomics, metabolomics, enz., bevatten de data typisch een groot aantal variabelen. De informatie in zulke datasets bevat vaak veel ruis in de vorm van irrelevante informatie en gemaskeerde variabelen. Bovendien kan de informatie komen van verschillende bronnen. Belangrijke uitdagingen voor deze datasets hebben betrekking tot het detecteren van structuur in hoog-dimensionele problemen, het uitfilteren van ruis en van irrelevante informatie, het bestuderen van meervoudige testen in aanwezigheid van een groot aantal variabelen, en het trekken van sterkere conclusies door gebruik te maken van geschikte combinaties van de verschillende data onderdelen. Om deze uitdagingen aan te gaan, zullen geschikte niet- en semiparametrische technieken (zoals 'smoothing' methoden) ontwikkeld worden, die gebruikt kunnen worden voor ruis reductie, geschikte dimensie reductie en cluster technieken (zoals methoden voor modellering met mengverdelingen) voor hoog-dimensionele twee- en meer-weg data, en data fusie methoden waarbij verschillende onderdelen van 'multiblock multiset' data samen gemodelleerd worden.

Cross-links tussen de vijf werkpakketten zullen ontwikkeld worden op ten minste drie verschillende niveaus:

I. *Aan elkaar gekoppelde complexiteiten in de data*
   In de praktijk komen vaak data voor die aan elkaar gekoppelde complexiteiten bevatten, zoals bestudeerd in de verschillende werkpakketten. Methoden zullen ontwikkeld worden voor het op een geschikte manier omgaan met deze samengestelde complexiteiten. Dit zal meer vereisen dan het louter aan elkaar koppelen van de resultaten verkregen in de verschillende werkpakketten, vermits wanneer men bijvoorbeeld ontbrekende data in een multivariate context bestudeert, nieuwe types van complexiteiten zullen moeten bestudeerd worden.

II. *Gemeenschappelijke modelleringsbenaderingen*
   De studie van afhankelijkheid is een terugkomend onderwerp doorheen alle werkpakketten. In dit opzicht zullen verschillende benaderingen gevolgd worden om met het modelleren van afhankelijkheid om te gaan, zoals het gebruik van copula modellen, regressie modellen die gebaseerd zijn op verschillende soorten dimensie reductie benaderingen, en modellen voor random effecten, zoals bv. gegeneraliseerde lineaire modellen. In het netwerk zullen deze verschillende benaderingen met elkaar vergeleken worden, zowel op theoretisch niveau als op het niveau van het analyseren van verschillende referentie datasets.

III. *Gemeenschappelijke methoden*
   De verschillende werkpakketten zullen steunen op een gemeenschappelijk stel van methoden, zoals technieken voor kern 'smoothing', semiparametrische inferentie, Bayesiaanse inferentie, optimisatie, randomisatie en 'bootstrap'. Conclusies i.v.m. deze methoden zullen uitgewisseld worden over de werkpakketten, en generische resultaten i.v.m. deze methoden zullen nagestreefd worden, zodat ze aangewend kunnen worden in een rijk gamma contexten.

C. Title and summary in French (2 pages maximum)

Analyse statistique d'association et de dépendance pour des données complexes

Une tâche essentielle de la statistique est d'analyser de façon appropriée la dépendance et l'association se trouvant dans un ensemble de données. Les données collectées aujourd'hui afin d'analyser ces structures de dépendance sont très souvent d'une nature complexe et les questions de recherche sont elles-mêmes de plus en plus difficiles à résoudre. Ces questions nécessitent de construire de nouveaux modèles statistiques ou d'adapter des modèles existants. Pour faire face à la complexité des données, de nouvelles méthodes et une interaction soutenue entre différents experts sont également nécessaires.

L'objectif global du réseau est le développement de nouveaux modèles et outils méthodologiques en vue de développer de nouvelles techniques d'inférence et d'analyser ces structures complexes de données. Pour atteindre cet objectif, le réseau sera structuré en cinq modules de travail, selon les différents types de complexité rencontrés dans les données.

1. *Module 1: Données multivariées sous des contraintes qualitatives*
   Dans l'analyse statistique, la quantité que l'on souhaite estimer ou tester est très souvent soumise à certaines contraintes qualitatives naturelles dont on doit tenir compte dans l'inférence. Comme exemple de telles contraintes, on peut citer les bornes naturelles (par exemple en analyse de frontière), la monotonicité, la convexité, l'ellipticité ou l'indépendance de composantes inobservées, l'unimodalité ou les représentations creuses. Les contraintes qualitatives apparaissent également dans les techniques de réduction de dimension, dans l'analyse de données fonctionnelles, ou en travaillant sur des problèmes inverses. Un large spectre de techniques statistiques est nécessaire pour analyser ce type de données. En se basant sur les résultats obtenus lors de la précédente phase (Phase V) du réseau, de nouvelles questions de recherche seront étudiées dans ce domaine, comme par exemple l'estimation de frontières stochastiques, l'utilisation de techniques de réduction de la dimension pour des données incomplètes, etc.

2. *Module 2: Données temporelles et spatiales*
   Les méthodes basées sur l'analyse en composantes principales pour prédire une simple variable sur la base d'un panel de séries chronologiques, sont largement utilisées en économie, et seront étudiées et comparées dans ce module. Le travail sur les modèles à facteur dynamiques, déjà développés durant la Phase V du réseau, sera poursuivi. Un autre sujet de recherche qui sera développé dans le réseau est l'étude des séries chronologiques non-stationnaires. Les résultats dans le domaine des séries temporelles localement stationnaires, qui ont été intensément étudiées durant la Phase V, seront poursuivis avec une direction particulière dans les tests d'ajustement, de l'inférence adaptative et de la modélisation spatiale. Une approche unifiée sera introduite dans le but d'étudier conjointement plusieurs types de non-stationnarité (tels que les modèles à coefficients variables dans le temps, les modèles à racines unité,…)

3. *Module 3: Données incomplètes*
   Plusieurs types de données incomplètes sont rencontrés en pratique: les données manquantes, censurées (avec censure à droite, par intervalle,…), tronquées, mal classifiées, détériorées (« coarse »),… Ce module mettra surtout l'accent sur les données censurées et manquantes. En particulier, le travail réalisé au cours de la Phase V sur l'estimation nonparamétrique de données censurées et sur l'analyse de modèles frailty, sera poursuivi. Le module s'attachera également à modéliser conjointement les données répétées et les données de survie. L'analyse de données manquantes s'orientera vers l'analyse de la sensitivité, et sur une combinaison de structures latentes et des modèles mixtes et de mélange. Un autre sujet de recherche traité par le module concerne l'estimation d'effets induits par des expositions mesurées avec erreurs dans les études randomisées.

4. *Module 4: Données avec hétérogénéité latente*
   L'hétérogénéité inobservée peut être modélisée de différentes manières. Une modélisation naturelle utilise les modèles mixtes, qui ont été intensément étudiés durant la Phase V. L'expertise acquise ouvre la porte à l'étude des modèles mixtes dont la dépendance des résidus est partiellement spécifiée conditionnellement aux valeurs des effets aléatoires, ainsi qu'aux modèles linéaires mixtes généralisés. Pour ces derniers modèles, des modèles flexibles pour la distribution des effets aléatoires seront investiguées, comme par exemple des mélanges de Normales pour l'approximation de bases B-splines.

5. *Module 5: Données en grandes dimensions et données composées*
   Dans de nombreuses applications utilisant la génomique, la protéomique, la métabolomique, etc., les données à analyser possèdent typiquement un très grand nombre de variables. L'information présente dans ces bases de données contient souvent beaucoup de bruit sous la forme d'information non pertinente et des variables masquantes. De plus, l'information pertinente peut provenir de différents types de source. Les défis principaux dans ces bases de données concernent la détection de structures en grandes dimensions, le filtrage du bruit et des éléments d'information non pertinentes, les tests multiples en présence d'un grand nombre de variables, et le développement d'une inférence plus forte à l'aide de combinaisons adéquates des différentes parties de données disponibles. Afin de traiter ces problèmes, des techniques non- et semiparamétriques adéquates seront développées (incluant des méthodes de lissage), pouvant être utiles dans la réduction de bruit, une réduction appropriée de la dimension du problème, et des techniques de clustering (incluant les méthodes de modèles de mélange) pour les données à hautes dimensions multi-tableaux, et les méthodes de fusion de données dans lesquelles plusieurs éléments de données multiblock multiset sont modélisés conjointement.

Des liens croisés entre les cinq modules seront établis à au moins trois niveaux différents:

I. *Complexités combinées dans les données*
   Dans les situations pratiques, les données rencontrées combinent souvent plusieurs types de complexité étudiés dans différents modules. Des méthodes seront développées pour traiter de façon adéquate ces complexités croisées. Ces méthodes nécessiteront plus de travail que la simple concaténation des résultats obtenus dans les cinq modules, car en considérant, par exemple, les données manquantes dans un contexte multivarié, de nouveaux types de complexités apparaissent.

II. *Approches communes de modélisation*
   L'étude de la dépendance est un sujet récurrent dans l'ensemble des modules ci-dessus. Des modèles communs seront considérés pour analyser ces dépendances, incluant l'utilisation de modèles de copules, de modèles de régressions basés, par exemple, sur différentes techniques de réduction de la dimension et de modèles à effets aléatoires comme par exemple les modèles linéaires mixtes généralisés. Dans le réseau, ces différentes approches seront comparées au niveau théorique et sur base de l'analyse de bases de données de référence.

III. *Méthodes et outils communs*
   Les différents modules sont également reliés par une utilisation commune de certains outils, comme les techniques de lissage par noyau, l'inférence semiparamétrique, l'inférence Bayésienne, l'optimisation, la randomisation et le bootstrap. Les développements réalisés sur ces outils seront échangés parmi les modules, et des résultats génériques sur ces outils seront développés, permettant leur usage dans une grande variété de situations.

I. 3. OBJECTIVES, MOTIVATION AND STATE OF THE ART (5 pages maximum)

Describe the project's objectives and research goals.
Define the problems being addressed by positioning them in relation to the current state of knowledge.

1. Overall objectives and organisation of the project

In a very broad range of substantive research domains, key theories, hypotheses and questions pertain to associations and/or dependencies between variables. Moreover, recent evolutions in many areas of science imply a significant increase in the complexity of such research questions. For example, often interest shows up not only in associations and dependencies as such, but also in the dynamics of association and dependency structures, that is, in their evolution across time. Ample illustrations of this can be found in many areas, including macro- and micro-economics, climatology, geophysics, and epidemiology. As a second example, questions on associations and dependencies often become more complex because much higher numbers of variables get involved. As such, ubiquitous calls may be heard for comprehensive, holistic accounts of phenomena, which may include structural insights into associations and dependencies on different and interconnected levels. An obvious illustration of this can be found in the recently established research domain of systems biology, which aims at unravelling highly complex regulatory networks (comprising very large numbers of genes, transcriptional regulators and other proteins), preferably in terms of holistic accounts that simultaneously involve different levels of biological functioning.

Linking up with the considerable increase in complexity in substantive association- and dependence-related questions, an immediately related challenge is implied by dramatic increases in the complexity of data as collected by substantive researchers. One of the obvious reasons for this increase in data complexity is that suitable data are looked for in order to deal in an appropriate way with the increasing complexity of research questions as outlined above. Other reasons include the accumulation of knowledge across time, as well as the advent of new and much more refined measurement technologies. As to the latter, one may, for example, think of the broad availability of automated data recording facilities, and, more in general, of all new measurement possibilities implied by the so-called digital revolution; typical illustrations of this include fMRI data collection methods and high-throughput experiments. The form under which data complexity shows up further can be of quite different types. As a first example, one may simply think of the sheer amount of data. Secondly contemporary datasets not seldom include very large numbers of variables and relatively small numbers of experimental units, herewith reversing standard prescriptive ratios of data sizes. Thirdly, datasets often are subject to unusually large amounts of error (as illustrations one may think of highly blurred images as studied in the area of image processing, as well as of very noisy microarray gene expression datasets as extensively studied in bioinformatics). Fourthly, data not seldom are prone to various kinds of (structural as well as incidental) forms of missingness (including censoring) (due, e.g., to equipment failure, measurement limitations, as well as various other reasons).

In general, statisticians can rely on a broad variety of modelling possibilities to deal with association- and dependence-related issues. Examples include regression models, time series analysis, linear and nonlinear mixed models, and clustering models (mixture-based and other). In many cases, however, existing modelling possibilities are not tailored to the tremendous complexity of contemporary association- and dependence-related substantive research questions, let alone to the distinct complexities of the implied data. In this regard, various kinds of model expansion and model development seem to be urgently needed. Examples may include the development of new concepts of partial frontiers within the context of boundary modelling, the development of novel dynamic factor models of volatility, the expansion of mixed effects models with very flexible random effects distributions, as well as the formulation of simultaneous or comprehensive models for data fusion. Moreover, both existing and to be developed models are cursed with many open problems, including the correct mathematical characterisation of models, identifiability issues, and an in depth understanding of model interrelations. Furthermore, in view of the broad range of modelling tracks that can be taken to deal with association and dependences, often an almost complete lack of clues to select the most appropriate track is to be faced. The development of a sound comparison of different modelling possibilities, on a mathematical-theoretical as well as on an applied level, is a major research challenge.

A tailored model development and model expansion of course cannot go without the development of an associated set of tools to be used in the fitting of the models to empirical data. This may include the

development of consistent and efficient estimators, the development of feasible and converging computational procedures, the evaluation of algorithmic performance, and the study of inferential correctness. Examples are the development of suitable non- and semiparametric estimation procedures, sensitivity analyses with regard to modelling assumptions and outliers, and the development of data-driven methods for selecting optimal values of metaparameters. All of these developments are further especially challenging because of the manifold complexities of the substantive questions and the data, and the different possible modelling tracks as outlined above.

Taking all of the above into account, the major objectives of this project proposal pertain to the study and development of a broad range of suitable modelling procedures and efficient data-analytic tools to deal with typically highly complex substantive theories and research questions on association and dependence, while making use of data that are subject to a very broad range of complexities. As such, the project will require a very intense and optimised interaction between the vertices of the triangle as outlined below (Figure 1).

To achieve the major research objectives as outlined above, the project proposal as described in this document will be carried out. This implies a long series of research activities, which within this project have been grouped into five workpackages. The major organizing principle in constructing the grouping in question pertains to a division based on five different types of complexity that can be distinguished on the level of the data. The five types of complexity are the following ones:

(1) multivariate data with qualitative constraints
(2) temporally and spatially related data
(3) incomplete data
(4) data with latent heterogeneity
(5) highdimensional and compound data.

Within each workpackage ample research time will be devoted to the development of suitable novel models and model expansions, as well as of tools for the associated data analysis. All this will happen within an intensive interaction with the complex substantive research questions and data at hand.

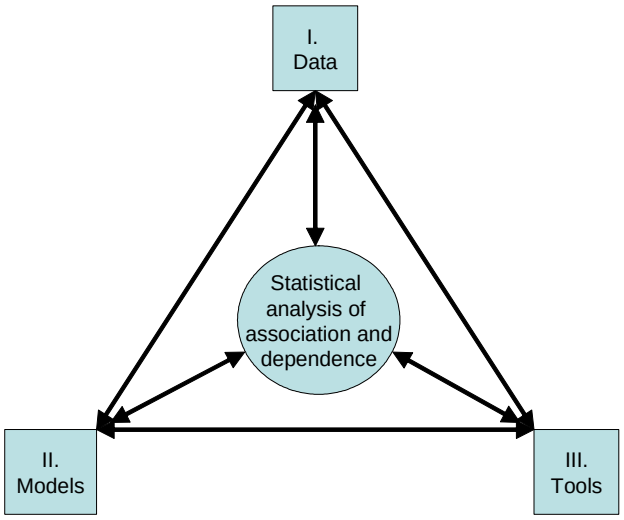

Figure 1: *Schematic representation of the three major cornerstones of the research proposal as well as their interactions.*

2. Objectives and state of the art per workpackage

Below the objectives for each of the five workpackages will be presented. The state of the art with respect to these objectives will also be specified, both in general scientific terms and in terms of previous accomplishments from the teams that will contribute to the workpackage in question.

Workpackage 1 : Multivariate data with qualitative constraints

Objectives

Often data and functions of interest satisfy certain qualitative constraints. These can be of very different nature: the function is a frontier or boundary; can show a certain degree of smoothness or non-smoothness; can only be assessed via indirect observations (leading to inverse and deconvolution problems), etc. Semiparametric techniques are a useful tool offering more flexibility than the very restrictive parametric methods. Data often show dependencies, of various levels. The main objectives here are: (1) to develop powerful nonparametric techniques for estimating boundaries or frontiers; for estimating smooth or non-smooth (multivariate) functions/images; for estimation in inverse/deconvolution problems; (2) to do inference in semiparametric regression models; (3) to study (robust) regression quantiles and multivariate extensions of these; and (4) to model dependencies and to test for independence; to study appropriate measures of dependencies, including analysis of robustness and efficiency.

State of the art

A recent survey on deterministic frontier estimation can be found in Simar and Wilson (2006). Recent robustified versions of estimators have been studied by Daouia and Simar (2006) and involve the concept of partial frontiers. The more realistic stochastic frontier models allow for noise.

Nonparametric data-driven techniques for estimating curves or surfaces that are possibly non-smooth at certain unknown locations have been studied during Phase V of the project. See e.g. Gijbels *et al.* (2006). For a survey of techniques see Qiu (2005).

An overview of techniques for inference for monotonic regression can be found in Gijbels (2005), whereas the use of wavelet techniques in combination with penalisation techniques for estimation of monotone regression functions was explored in Antoniadis *et al.* (2005).

During phase V of the project a lot of work has been devoted to the problem of density estimation based on data contaminated with errors. The performances of kernel-based methods were studied in detail and practical bandwidth selection methods were proposed in Delaigle and Gijbels (2004). An interesting deconvolution problem appears in the 'deblurring' of images. For wavelet techniques for deblurring images see Johnstone and Raimondo (2004).

Many models for human reaction time data are parametric, often proposing a convolution of two parametric distributions. Van Zandt (2002) discusses some of the most elementary nonparametric methods, not taking into account that reaction times are subject to a lower boundary and that censoring may occur at the higher end of the response time scale. Tuerlinckx and De Boeck (2005) study a specific parametric diffusion type model for reaction times. This could be used as a starting point in further nonparametric analysis.

Inference procedures for the parametric component of a general semiparametric regression model have been studied by e.g. Chen *et al.* (2003) and Van Keilegom and Carroll (2006). In single-index models, Geenens and Simar (2005) developed rank correlation based methods, and obtained estimators outperforming standard ones.

The power of quantile regression has been demonstrated in Koenker (2005). There is the need for a satisfactory concept of multivariate quantile.

Workpackage 2 : Temporally and spatially related data

Objectives

Serial or spatial correlation are two sources of complexity very often met in the applied sciences, for instance when data are collected in an automatic way. This workpackage will go beyond the work achieved in the previous Phase V on high dimensional time series and locally stationary processes. In particular, other types of nonstationarity will be studied, including structural breakpoints and unit root processes. Moreover, the analysis of spatial data will be considered.

State of the art

The statistical analysis of time series and spatial data has benefited from the huge progress in computation power. Topics such as multivariate modelling with a large number *n* of variables, or modelling non-stationary data are becoming easier. New challenges have appeared, though (see the next section). Among the achievements of Phase V, our work on dynamic factor models (see Forni *et al.,* 2004, 2005) and in locally stationary models, e.g. Van Bellegem and Dahlhaus (2006) and Ombao *et al.* (2005) are surely state of the art. The state of the art for three out of the five topics introduced in this workpackage will be discussed.

Development of local adaptive procedures in order to capture non-stationarity goes back to the concept of adaptivity of Lepski (1990), recently improved in Pohlehl and Spokoiny (2006). That local analysis necessities a deeper statistical analysis of likelihood ration tests of structural breakpoint, for which the paper of Andrews (1993) is still a major reference.

So far, modelling structural breaks like jumps was done via parametric models and assuming that breaks do not occur simultaneously in the mean and the variance. Nonparametric methods for detecting change points for independent data were studied within Phase V. In financial models inclusion of jumps is crucial in explaining stock prices and their volatility (Campolongo *et al.*, 2006).
Goodness-of-fit tests based on the periodogram and resampling are developed in the stationary situation by Paparoditis. Extension to spatial cases and/or non-stationary situations is original. Resampling of non-stationary signals is also addressed in Ombao *et al.* (2005).

Workpackage 3 : Incomplete data
Objectives
Incomplete data occur in time-to-events as censoring, and in longitudinal and multivariate studies as missing data. Correct (causal) inferences are needed. Further forms of incompleteness are misclassified data and measurement with error. Models for these types of incompleteness will be aimed at, and they will be corrected for varying intervention regimes. Models will be proposed for describing response shifts in test data. Such models typically need strong assumptions and thus appropriate sensitivity analysis tools will be proposed. For survival data, various types of censoring (left, right, interval) will be studied, with emphasis on informative and dependent censoring, and coupled to sensitivity analysis. Focus will be on nonparametric and parametric methods, in particular on frailty and copula-based models.

State of the art
The teams are internationally renowned in survival analysis, incomplete longitudinal data, test theory, and causal inference.
Interval censoring has been studied in a nonparametric context, but less so parametrically. Braekers and Veraverbeke (2005) proposed copula models for the dependence between lifetimes and censoring times; informativeness in regression models with censored data was studied by Braekers and Veraverbeke (2001). Frailty models for multivariate survival data are popular (Duchateau and Janssen, 2004) and there are many results regarding semiparametric regression with completely observed data. The theoretical properties of several of them (partial linear, additive, single-index, … models) are well studied, but much less is known about estimation in these models when the response variable is subject to random censoring.
Many models for incomplete longitudinal and multivariate data (Little and Rubin, 2002) exist (Verbeke and Molenberghs, 2000, Molenberghs and Verbeke, 2005). This produced a paradigm shift regarding standard analyses in, for example, regulated biopharmaceutical development. Also, sensitivity analysis to study the impact of modelling assumptions on inferences from incomplete data received attention (Cook, 1986, Lesaffre and Verbeke, 1998, Verbeke and Molenberghs, 2000, Molenberghs and Verbeke, 2005, Vansteelandt *et al.*, 2005). More flexible models and a consensus in the field are needed. The impact of misclassification is in need of further study. Little work has been done in the longitudinal setting. Further, a principled study of missingness and sensitivity in the area of test data is required, since item response theory models for dealing with strategy shifts and/or differential modelling received little attention (Yamamoto and Everson, 1997, Bolt *et al.*, 2002).
Current causal analysis of dynamic intervention regimes (Murphy, 2003, Robins, 2004) is restricted by the use of linear models, the exponential increase of the number of treatment sequences, and computational complexity. Hence, large datasets are required, available in marketing (Cho *et al.,* 2002), but not in all relevant areas. The theory as a whole needs further development (Rubin, 2003) and dose-response relationships are typically non-linear and recent development (Vansteelandt and Goetghebeur, 2003, van der Laan *et al.,* 2005) leaves open important questions such as how to infer these from causal models.

Workpackage 4 : Data with latent heterogeneity
Objectives
In WP4 traditional mixed models will be extended into various directions. More specifically, the following objectives have been fixed. First, models will be developed and compared with different flexible random effects distributions (e.g., based on P-splines, flexible normal mixtures, classical normal mixtures) for complex models possibly in the presence of coarsened data. Also, different computational procedures will be explored (e.g., the vertex-exchange-method) and mathematical and statistical properties will be explored of different variations of the basic approaches (e.g., L1 versus L2-based penalties for the P-spline approach). Further, novel methods (e.g., the copula method) will be developed to model dependencies in random effects models that are not accounted for by the random effects (hence, residual dependencies). Furthermore, mixed models will be studied for multivariate data measured repeatedly over time, possibly with time dependent covariates. On one hand, specification, hypothesis testing and computational aspects in the general context of multivariate mixed effects models will be examined. On the other hand, a model will be developed based on an Ornstein-Uhlenbeck process for the serially correlated error in which variances and covariances can vary randomly over

persons. Models with crossed random effects will also be studied, in which the units of a random variable are cross-classified with the units of another random variable. Algorithms for estimating such models will be compared, random effects models combined with an extreme value error distribution will be developed and the application of these models in actuarial and financial applications will be considered. Finally, the statistical concept of causality in complex models involving random effects (or latent variables) is intended to be explored.

State of the art

In general, mixed model methodology has spread widely over the last few decades. Finding the appropriate random effects distribution of the linear mixed model has received much interest recently. Ghidey *et al.* (2004) suggested a penalised Gaussian mixture for the random effects distribution where the normal components have been fixed on a grid and only the mixing weights are estimated using a penalised approach. This approach has also been applied in univariate and multivariate accelerated failure time models by Komarek and Lesaffre (2006), though limited research has been done in other complex mixed models. The choice of the penalty function for the P-spline approach is linked with the regularity assumptions that are imposed on the function to be estimated. This issue of regularisation has been discussed by Antoniadis and Fan (2001) in the context of wavelet techniques. Specific choices of penalties in penalised regression lead to techniques such as ridge type regression, LASSO, bridge regression, among others. Functional mixed-effects models and their applications to Mass-spectronomy data have been studied by Antoniadis and Sapatinas (2006).

The problem of residual dependencies has been studied extensively in psychometrics (for an overview, see Tuerlinckx and De Boeck, 2004). Many existing methods, however, suffer from interpretational drawbacks because modelling dependencies leads to univariate marginals (given the random effects) that do not belong to a nice functional class; an exception being the model proposed by Ip (2002). The application of copulas to discrete data has been studied by Tajar (2003).

The development of suitable models for multivariate longitudinal data is an active field of research, both in psychology (e.g., Blozis, 2004) and in biostatistics (e.g., Fieuws and Verbeke, 2004, 2006; Fieuws *et al*., 2006; O'Brien and Fitzmaurice, 2005). Models with serially correlated errors for bivariate longitudinal data have been described by, for instance, Sy *et al.* (1997) but they do not allow the variance and covariance parameters to differ over individuals.

The application of crossed random effects models in psychometrics has been proposed by e.g. Janssen *et al.* (2000) and Van den Noortgate *et al.* (2003). In the latter study an approximation method based on a linearization of the integrand has been used while the first is based on a Bayesian estimation algorithm. A related approach based on conditional linear mixed models has been proposed by Tibaldi *et al.* (2006). Approximate inferences for cross-classified data can be found in Schulz *et al.* (2005).

Russo *et al.* (2006) state that a sound statistical concept of causality should be based on the concepts of exogeneity and structural conditional models. It is also argued that although this approach is natural and satisfactory for simple models, i.e. two vectors of variables, serious difficulties arise when dealing with complex models involving latent variables. Further work is clearly required for a better understanding of causality in complex models. In particular, a connection with the literature, among others in econometrics, on "counterfactuals" should be developed with a view to obtain a better grasp of the concept of causality taking into account individual heterogeneity.

Workpackage 5 : Highdimensional and compound data

Objectives

Highdimensional and compound data require the development of specific analysis tools. To reveal and model association or dependence in real-valued object-by-variable data involving a high number of variables (e.g. micro-array data) novel models based on dimension-reduction techniques will be developed.

New multiple testing procedures are proposed to promote reproducibility in the presence of such high numbers of variables.

Additional information will be incorporated, in the form of one (or more) added dimensions to the object-by-variable data (thus creating multi-way data) or through extra highdimensional blocks of data from different sources pertaining to the same problem (compound data) and new statistical techniques will be developed for such multi-way and compound data.

State of the art

A structured overview of a broad range of biclustering methods that imply a simultaneous clustering of objects and variables has been presented by Van Mechelen *et al.* (2004). To deal with irrelevant or masking information, a broad range of tools has been proposed, including mixture modelling (George and McCulloch, 1993), subset selection methods (Brusco, 2004), and weighting techniques (Friedman and Meulman, 2004). To enhance reproducibility and overcome a plethora of false positive results, several statistical procedures

have been developed which protect the experimentwise proportion of false positives (Storey and Tibshirani, 2003; Efron, 2004). The growing concern about their decreased power to detect biologically important signals has led to new estimation/detection methods and to a rebalancing of type I and type II errors to optimise performance (Moerkerke *et al.*, 2006). Power can also be enhanced by bringing in new information from related sources. The analysis of compound data links up with techniques known in the psychometric literature as simultaneous components analysis (Kiers and ten Berge, 1989) and in the chemometric literature as common principal components analysis (Smilde *et al.*, 2003). In the bio-informatic literature it is termed genomic data fusion (Lanckriet *et al.,* 2004).

I. 4. DETAILED DESCRIPTION OF THE PROJECT (15 pages minimum, 25 pages maximum)

- Submit a general description of the project as well as a description detailing each workpackage and indicate the partners involved in each workpackage.
- Illustrate by means of a table or scheme the interaction between the partners within a workpackage and the interaction between the workpackages.
- Describe and justify the methods and proposed approaches in relation to the state of the art.
- Describe and justify how the contribution of the different partners will be integrated.

As outlined in Section I.3, the major challenge and objective of the network is to analyse the dependence and association present in a dataset. These dependence and association structures are often of a very complex nature. The complexities can be situated on three different levels, as well as on the connections between these levels : complex data structures (like functional data, highdimensional data, censored data, etc.), complexities in the model (for example, due to identifiability issues, or introduction of new elements in the model) and in the tools used to analyse data (including development of efficient estimation and testing procedures, sensitivity analysis, convergence of computational procedures, etc.). The network is structured in five workpackages, that are defined in terms of the different types of complex data that one encounters when analysing association and dependence structures.

This subsection is organised as follows. First, a detailed description of the research proposal is given for each of the five workpackages. Next, the interactions between the workpackages are explained in detail, for each of the three levels of complexity considered above (data, models and tools). The interactions between the different partners working on a same workpackage are also discussed. Finally, the added value gained through the network and the initiatives planned in order to integrate the contributions of the different partners, are outlined.

1. Workpackages

Workpackage 1 : Multivariate data with qualitative constraints

Substantive applications/problems
Frontier estimation deals with estimating the boundary of the support of a density of a multivariate random variable. Among the main applications is productivity analysis in which one seeks for optimal plans of production. The availability of panel data makes the problem somewhat easier, but semiparametric and nonparametric approaches are to be developed, in particular for data in which spatial and/or time dependencies are present. Further, explaining inefficiencies by external environmental factors is of practical importance.
Estimation of non-smooth surfaces is closely related to edge-detection (or boundary detection) and image analysis. For a variety of examples, see Qiu (2005). The analysis of a set of images is of interest in e.g. medical data - MRI or NMR, or for data from satellites.
Examples of situations where it is realistic to impose certain qualitative constraints (e.g. unimodality, monotonicity, convexity, concavity) are ample: in frontier estimation economic considerations lead to constraints of monotonicity and convexity; in medical applications monotonicity of a target function is often a realistic assumption.
The use of nonparametric and semiparametric deconvolution techniques will be investigated in certain application areas, such as astronomy. The aim is to relax the rather restrictive distributional assumptions. A more complex situation occurs when images need to be 'denoised' and 'deblurred'. Another application of inverse problems is tracking dynamic deformations in geostatistical images.
Human reaction time data are a common measure in psychological research. Reaction time data can be of different complexities: a large number of data for a single participant in an experiment; or a

shorter sequence of reaction times measured for different participants, both under various conditions. Often the reaction time consists of two components: a decision and a residual component. In this context, deconvolution is linked with the assumption of the existence of these two components. Semi- and nonparametric techniques for such data will be developed.

Models and primary objectives

1. Boundaries, frontiers, smooth versus non-smooth functions and images
   Partial frontiers have been introduced to obtain estimators of a frontier that are robust to extreme data points. These estimators however are quite different according to the way chosen for determining the frontier (input or output orientation). A first challenge is to define a new concept of partial frontier that would be independent of that orientation. A second challenge is considering stochastic frontier estimation and obtain estimators with reasonable rates of convergence. Among possible approaches to investigate further are: using local maximum likelihood approaches (see Kumbhakar *et al.,* 2006) or viewing the problem as a particular deconvolution problem. For panel data, dynamic modelling will be applied to deal with the issues of spatial and/or time dependencies. Also, statistical properties of conditional efficiency measures will be explored.
   Curve estimation problems can often be formulated in terms of a closed and convex parameter set embedded in a real Hilbert space. This is the case, for instance, if the curve of interest is a monotone or convex density or regression function. An estimator of the curve of interest can be obtained by projecting an arbitrary initial estimator onto this parameter set. Statistical properties of this estimator are studied in Fils-Villetard *et al.* (2005) and will be developed further.
   In non-smooth estimation of curves and surfaces and image analysis challenging research issues are: (1) use and implementation of local smoothing parameters in simple procedures to deal with homogeneous and heterogeneous regions in one estimation task; (2) application to three-dimensional functions and images; (3) application of the developed techniques to curves/surfaces with a bounded support, and a possible non-smooth behaviour at these boundaries. In a project A. Antoniadis, J. Bigot and R. von Sachs are working on a multiscale approach for statistical characterisation of temporally and spatially heterogeneous functional images. This work involves issues such as unsupervised clustering of the image, dimension reduction, wavelet thresholding and complexity-penalised likelihood.
   For nonparametric function estimation under qualitative constraints it seems quite natural to use the P-splines approach, putting as such constraints directly on coefficients in the representation.

2. Inverse problems and deconvolution problems
   In 'deblurring and denoising' images, it is a challenge to deal with the 'deblurring' part, that is in fact an inverse problem (or deconvolution problem). The available wavelet techniques are restricted to uniformly behaving measurement error. The use of simpler kernel-based methods as part of the basic ingredients will be explored. As mentioned already, the problem of stochastic frontier estimation is related to a deconvolution problem, and this connection will be further investigated and explored.
   Nonparametric estimation of the solution of inverse problems by means of regularisation techniques and by means of sparsity-enforcing constraints (for example via a weighted L1-norm) will also be considered. Closely related frameworks are penalised maximum likelihood estimation with L1-type penalties or Bayesian maximum a posteriori (MAP) estimation under Laplacian instead of Gaussian priors. More general penalties and data error statistics, such as Poisson-type noise (affecting e.g. astronomical images) will be considered. Most of the statistical studies on ill-posed inverse problems assume that the operator of the problem is known. The problem of an unknown operator is motivated by many practical situations, for instance blind deconvolution or nonparametric regression with instrumental variables. In inverse problems wavelet methods for continuous-time prediction by means of autoregressive Hilbert space processes will also be developed.

For studying reaction time data, the objectives are: (1) studying the deconvolution problem for an observed reaction time distribution without specifying the distribution of one of the components; (2) applying nonparametric density estimators to reaction time data that can handle lower bounded data (boundary problem) and possibly also right-censoring; (3) assessing the effects of conditions and/or covariates on the reaction time distribution, via for example the use of Cox' proportional hazards model (with the modification of a possible lower bound different from zero).

3. Semiparametric models and dimension reduction
   The work on general semiparametric regression models started during Phase V will be pursued. The focus will e.g. be on developing results on the use of empirical likelihood techniques for these models. The estimation of semiparametric transformation models, and the development of goodness-of-fit tests for semiparametric models will also be investigated. A single index model is a semiparametric extension of Generalised Linear Model (GLIM) with an unspecified link function. It will be investigated how the recently developed rank correlation-based methods could be used in testing issues.

4. Nonparametric inference and robust analysis
   The estimation of the nonparametric location-scale model $Y=m(X)+\sigma(X)\varepsilon$, with ε independent of X, has been well studied over the last years. The problem of developing testing procedures has been initiated by Van Keilegom *et al.* (2004) and Pardo Fernández *et al.* (2006). Several testing problems under this model are still unsolved and will be studied, among which testing for: (1) the form of the variance function; (2) independence between the error and the covariates; (3) changepoints in the model; (4) GARCH-type models (for which $m=\sigma$).
   A long-standing statistical challenge is the extension of regression quantiles to the multivariate setting (multiple output regression and/or multivariate autoregression). The development of a concept of multivariate regression quantiles will be proposed, in relation with regression contour ideas. The resulting regression quantile contours are expected to be widely applicable in areas such as economics, econometrics, biomedical applications and clinical monitoring. A second objective is to extend robust quantile regression techniques from linear models to the case of non linear regression models. Previous work trying to robustify quantile regression includes Rousseeuw and Hubert (1999), among others.

5. Modelling and measuring of dependencies and copula functions
   Finally, attention will be paid to the modelling of dependencies. A copula function directly links the marginal distributions of a random vector to their joint distribution function. Some well-known association measures such as for example Kendall's tau, are directly related to the copula function. Copulas have been used the last decade in a variety of areas (medical applications, actuarial sciences, ...). Analyses often rely on non-flexible parametric families of copulas. Nonparametric kernel-based estimators of copulas have been studied recently in Fermanian and Scaillet (2003). Powerful testing procedures will be developed, that test whether a copula belongs to a specific parametric family. Once such a test developed it will be also useful for testing for independence. A second challenge is to cope with data that are possibly right-censored, appearing for example in actuarial sciences.
   Related to the issue of robustification is also the search for robust measures of correlation. The aim is to study the robustness and efficiency properties of available nonparametric correlation measures. A measure of correlation with a good compromise between robustness and efficiencies properties will be aimed at. It will be a challenge to export these comparisons in a multivariate setting.

Cross-links with other workpackages
WP2: (1). Frontier estimation based on panel data resembles the type of data also encountered in WP2; (2). Estimation of non-smooth curves/surfaces is linked with the problem of estimation of trend or volatility in a time series with possible structural breaks.
WP3: (1). Reaction time data possibly involve right censoring, and hence resemble data having similar complexities as in WP3; (2). In WP3 semiparametric regression models and nonparametric location-scale models with censored data are among the research topics. There, modelling aspects of WP1 will show up; (3). Modelling of dependencies via copula functions is also an objective in WP3, where the focus is on survival data.
WP4: (1). Frontier estimation based on panel data using modelling techniques also appear in WP4; (2). Penalised likelihood approaches are studied and applied in WP4.
WP5: (1). Semiparametric regression models, such as for example single-index models, are able to deal with highdimensional covariate vectors, and hence there is a link with WP5 regarding complexity of data; (2). Robustification of methods is linked with similar issues of robust inference techniques in WP5.

Methodological problems
Several methodological problems need to be studied to deal with these types of complexity in the data:

1. In semi- and nonparametric frontier and boundary estimation as well as in smooth and non-smooth estimation and image analysis there is a need for data driven methods for selecting optimal smoothing parameters. Resampling methods such as bootstrap are among the used methodologies.

2. To compute regularised solutions in inverse problems, iterative algorithms based on the successive minimisation of surrogate functionals, will be needed.
   When analysing, for example, reaction time data, one is confronted with a deconvolution problem and a problem of a lower boundary. In addition, when assessing covariate effects for these data, one has to deal with an unknown starting point. Another problem is the determination of sample sizes required for obtaining stable non- or semiparametric estimators.

3. When using P-splines and other regularisation techniques, optimisation techniques over function spaces are part of the problem. Awareness and use of modern techniques of optimisation, including equivalence to dual problems, is an important issue.

4. When developing robust measures, e.g. of correlation, a methodological study of influence functions and maxbias curves is needed.

Partners involved: KUL-1, UU, UCL, UH, UJF, USC.

Workpackage 2 : Temporally and spatially related data

Substantive applications/problems
Nonstationary time series are encountered in several sciences (geophysics, neurosciences, economics, ...). A number of important problems in this area are still unsolved. Forecasting single variables on the basis of large panels is a well studied topic. With automated data collection, more data are under the form of time series, and often these time series are multivariate, like in financial applications. Most of the current techniques are not able to handle this type of data. Analysis of volatility is of prime importance for financial time series. Daily observations and other high frequency data raise other problems like superposition of several seasonalities. Moreover, extension to spatial data is important, e.g., for applications in geography of spatial economics.

Models and primary objectives

1. Forecasting using a large number of predictors
   Principal components in that context can be considered as a special case within a class of models that includes ridge regression, LASSO and others. The clarification of the relation between these different techniques in empirical macroeconomic and financial applications is planned. It will also be studied whether surveys forecast well industrial production, by performing out-of-sample forecasting experiments on the basis of a dynamic factor model. A new monthly dataset for the Euro area will help to investigate whether the improvement in forecasting accuracy using surveys is because they are more timely rather than for the information they contain about expectations on the state of the economy. A two-step estimator (combining OLS on principal components and a Kalman smoother) will be developed for large approximate dynamic factor models with series of length T and n variables, and study ($n; T$) consistency. The analysis provides theoretical backing for the estimator considered in previous papers. Also, the extension of dynamic factor models to locally stationary models is planned for multivariate time series which are driven by latent factors.

2. Beyond locally stationary models
   Beyond the locally stationary models that were studied in Phase V, the intension is to undertake the following research work: development of general tests of stationarity; for locally stationary processes, development of nonparametric goodness-of-fit tests based on the Kullback-Leibler distance in the frequency domain and estimating spectral densities (with subject-specific replications and random effects); combination of the two approaches that were investigated separately by Van Bellegem and Dahlhaus (2006) and Azrak and Mélard (2004), and their application to emotion data; nonparametric estimation of spectral density using asymmetric kernels; improved covariance and spectral estimation for multivariate time series using shrinkage-type estimators, under a double asymptotics with increasing dimensionality.
   On the other hand, in the area of volatility the following research is planned: implementation of a dynamic factor model for volatility, including application to highdimensional systems; efficient and/or adaptive estimation of semiparametric multivariate volatility models. We also intend develop spectral techniques for the goodness-of-fit of spatial models

3. Wavelets, signal processing and structural break analysis
   The collaboration across universities on wavelet techniques for time series for the analysis and forecasting of univariate and multivariate time series will be extended. Also, these techniques will be used for continuous-time prediction by means of autoregressive Hilbert space (ARH) processes. These are related to the theory of function estimation in linear ill-posed inverse problems.
   Structural breaks and changes of regime, either on the level or on the scatter, are worth more studies: adaptation of methods developed for independent observations within Phase V to time series in the case of breaks that can occur simultaneously in the mean and the variance, extending the existing limited parametric threshold structure to a more flexible nonparametric threshold setting; achievement of a very general, non-constrained and fully automatic method for break point detection and estimation of the spectral structure of time series, using hidden Markov switching regime models for the time-varying variance-covariance structure and the unbalanced Haar transform for automatic segmentation of time series.

Cross-links with other workpackages

WP1: Principal components in the context of forecasting using a large number of predictors can be understood as a regularisation method for inverse problems. Autoregressive Hilbert space processes are also related to estimation in ill-posed inverse problems.

WP4: In some configurations, panel data across time are related to mixed models. Estimation of spectral densities of locally stationary processes may have subject-specific replications and include random effects.
WP5: The dynamic factor model is aimed at highdimensional data.

Methodological problems
A certain number of methodological issues will be studied including the following ones:

1. For methods involving both time *T* and panel dimension *n* (topics 1 and 3 above), double asymptotics are to be treated adequately.

2. For topic 3, implementation will be needed of (1) equipment failure in on-line estimation; (2) fully automated model building procedures, assessed by comparing them through a benchmark; and (3) a model building procedure adapted to daily economic data taking care of several seasonalities, holidays, abnormal events.

3. For topic 2, development of local test of stationarity constructed using likelihood ratio. Our challenging task is to derive critical values using non-asymptotic arguments because the adaptive procedure applies the test to potentially small intervals.

Partners involved: KUL-1, UU, UCL, UG, UJF, USC.

Workpackage 3 : Incomplete data

Substantive applications/problems
Censored (correlated) survival data occur in the medical and epidemiological fields, e.g., when the validity of prognostic indices is studied, when the heterogeneity in outcome between participating hospitals in multi-center studies is of scientific interest, and when the presence of a random treatment by center interaction needs to be investigated. Incomplete longitudinal studies, combined with survival outcomes, are commonly encountered in the areas of dental epidemiology, HIV studies, psychiatric trials, quality of life in oncology, and mental health epidemiological studies, to name a few. Especially in a regulatory context, flexible sensitivity analysis tools are needed.
When a test is administered with a fixed time limit, some students may not have enough time to answer all questions, so-called test speededness, the effects of which are detrimental to the intended functioning of the test. Examinees affected by test speededness may hurry through, randomly guess or even fail to complete items, usually at the end of the test. The concept needs to be part of the model to avoid biased inferences.
Causal impact of observed exposure is of prime importance in therapeutic treatments and marketing interventions. Individualised treatments are gaining popularity, also in view of genetic breakthroughs (Prentice *et al.,* 2005). Inferences are hampered, though, by the very relationship between the treatment sequence and the outcome studied. Currently, large data streams are being recorded, forming a rich resource for evidence based marketing. Data mining techniques can help find association patterns (Cho *et al.,* 2002). Ideally, sequentially randomised experiments provide a basis for unbiased estimation of the causal impact of evolving interventions based on a treatment and response history.

Models and primary objectives

1. For censored survival data, copula modelling will be explored further, including model formulation, assessing goodness-of-fit, and bootstrap-based precision estimators. Various types of informative censoring will be combined into a single model, extending the copula-graphic estimator of Braekers and Veraverbeke (2005). The Koziol-Green model will be extended to handle dependent censoring and a mixture distribution for the censoring time will be considered. Regarding estimation in semiparametric regression with censored data, a first problem is that the mean function generally cannot be consistently estimated with censoring. Valid estimation procedures are in demand. These problems are currently under study for the single index model, but other semiparametric models need study too. Frailty models for right

censored data received a lot of attention, but questions on model diagnostics, hierarchical modelling, and methodological issues (e.g., likelihood ratio testing) are still open. Formulating frailty models for complex censoring are important but untouched. The link between copulas and frailties will be studied. Another research topic that will be studied is the analysis of recurrent events data by making use of transfer of tail information in order to improve the estimation of the so-called transition probabilities. This 'transfer of tail information' will be made possible by imposing a location-scale type regression model on the gap times between consecutive events.

2. Flexible models for incomplete longitudinal data, allowing for various types of sensitivity analysis, will be formulated within the selection model, pattern-mixture model, and shared-parameter model paradigms. Combined with mixture modelling ideas, hybrids between the various model families will be formulated, thereby allowing for misclassification. Sensitivity analysis tools, based on (semi-)parametric models, will be considered. Models for test speededness and/or shifts in test strategies, allowing for random guessing and omission of items, will be formulated. Subject-specific shifts and speeds will be allowed for. Sensitivity of inference will be assessed. The potential outcome formulation of causal parameters has connected causal inference and incomplete data methods. The focus will be on efficient semiparametric models for potential outcomes following different exposures, relying on instrumental variables or on the assumption of no unmeasured (time-varying) confounders. Focus will also be on non-linear effects, heterogeneity between subjects, and measurement error on exposure, extending linear model methods.

Cross-links with other workpackages
WP1: The non- and semiparametric models connect with WP1. For instance, the location-scale regression models, the semiparametric models and the copula models are encountered in both WPs. There are also cross-links on nonparametric methodology, like e.g. kernel smoothing and bootstrap methods.
WP4: Longitudinal, multivariate, and otherwise hierarchical data connect with WP4.

Methodological problems
The research proposed above entails the following methodological issues.

1. For censored data, the classical assumptions of non-informative and independent censoring will be relaxed, leading to new estimators, of which consistency and asymptotic normality will be shown. Methods for correlated survival times will be formulated. For recurrent events data the commonly imposed Markov assumption will be dropped leading to a new kind of estimation technique.

2. For incomplete longitudinal and multivariate data, possibly subject to misclassification, measurement, error, test speededness, and/or dynamic treatment regimes, complex models will be formulated, posing the following problems: (1) numerical convergence will be non-trivial, (2) model identification is not straightforward, and (3) sensitivity of conclusions regarding assumptions needs to be explored. The performance of models when assumptions do not hold and/or with outlying observations needs to be studied. Estimation and model properties are less than straightforward in the non-linear case. Early attempts (Robins *et al.,* 1994, van der Laan *et al.,* 2005, Vansteelandt and Goetghebeur, 2003) are in need of further development. Data sparseness coming from the large number of dynamic treatment regimes needs to be handled.

Partners involved: KUL-1, KUL-2, LSHTM, UCL, UG, UH, USC.

Workpackage 4 : Data with latent heterogeneity

Substantive applications/problems
Substantive areas of application that have motivated the research objectives of WP4 are biostatistics, psychometrics, actuarial and financial statistics.
The problem of modelling of residual dependencies arises in a biomedical study on birth defects (as a consequence of teratogens) because birth defects tend to cluster together on the same parts of the body, or due to some unknown genetic factor. In psychometrics, a similar problem occurs when items cluster together, for instance, because they follow the same reading passage in a test.
Models for multivariate longitudinal data and associated problems are inspired e.g. by the substantive problem of modelling repeated two-dimensional measurements of emotion and by a variety of medical longitudinal studies. However, the generality of the proposed stochastic model allows for applications in quite different substantive contexts as well.
Crossed random effects models have been applied in a psychometric context where both persons and items are considered random. The items are assumed to be sampled from a population or subpopulations of possible items.
Random effects models in combination with extreme value distributions are inspired by credibility models. Credibility is concerned with the prediction of future claims of a risk class, given past claims of that and related risk classes. Classical credibility models stem from simple models with normality assumption for both responses and random effects. It is useful to revisit the credibility models in the context of generalised linear models and to consider their specifications as a generalised linear mixed model. See for instance Antonio and Beirlant (2006a,b), the latter using B-spline smoothing with mixed models.
Apart from these substantive areas of application, the flexible distribution approach is inspired by the B-spline modelling of histograms and the fact that the current extensions of mixed models to non-normal random effects involve complex families of distributions implying considerable computations. The method can be used in almost any domain of application.

Models and primary objectives
The objectives are:

1. Developing and comparing different flexible random effects models suitable for complex mixed effects models possibly in the presence of time-varying covariates, coarsened data and both univariate as well as multivariate responses.

2. Modelling the univariate and multivariate residual dependence structure by different approaches and comparing their performance.

3. Application of these approaches to important biomedical, psychometric and financial applications.

4. Developing the notion of causality in random effects models.

Cross-links with other workpackages
WP1: The P-spline approach can be extended to smooth data with qualitative constraints. Modelling dependencies through copulas will be a topic of investigation in this WP.
WP2: Longitudinal data are examples of temporally-related data and therefore the approaches that will be explored in this WP are related to WP2. Further, the P-spline approach will be developed for particular classes of spatially related data.
WP3: There is an obvious connection between this WP and the workpackage on incomplete data. In fact, one of the reasons for the popularity of mixed effects models is their flexibility to deal with missing data. There has been a long standing collaboration in this sense between the KUL2- and UH research teams.
WP5: Specific techniques of penalised regression such as ridge type regression, bridge regression, LASSO regression will be exploited in the context of highdimensional data.

Methodological problems
The methodological challenges are mainly located on three levels: model formulation, estimation of the parameters, and model testing.

1. Model formulation: Defining flexible and smooth random effects distributions that preserve major characteristics such as discontinuities and fusing such flexible distributions with non-traditional data distributions (e.g., heavy-tailed distributions). Study random effects models in order to allow for causal inferences.

2. Model estimation: Develop and compare efficient traditional and Bayesian algorithms to estimate parameters from models for multivariate longitudinal data, residual dependency models, models with flexible random effects distributions and crossed random effects models

3. Model testing: The study of proper hypothesis testing methods for variance components and serial correlation structures, taking the non-standard nature into account; selection of the appropriate residual dependency copula function; selection of a suitable serial dependency process for multivariate longitudinal data.

4. Development of statistical concepts like causality in random effects models.

Partners involved: KUL-1, KUL-2, UU, UCL, UH, UJF, USC.

Workpackage 5 : Highdimensional and compound data

Substantive applications/problems
The digital revolution has created data streams leading to highdimensional and compound data in many scientific disciplines today with a resulting requirement of efficient information extraction techniques. Detecting genes, gene-gene interactions and gene-environment interactions that show strong associations with specific traits is a main research activity in the plants and animal sciences and in human medicine. Three-way data arise naturally in a broad range of domains, including genetics, proteomics (different biological settings (e.g., responders/non-responders) with multiple, independent biological samples, and multiple mass spectra for each of the samples), psychometrics (e.g., with the measurements of a set of responses for a set of persons in a set of situations), chemometrics (e.g., in fluorescence spectroscopy data, which measure the intensity of fluorescence at several combinations of excitation by emission wavelengths for different batches) and econometrics (e.g. measuring sales in response to incentives for different types of potential clients). Additional structure also imposes itself when causal questions are to be answered. Compound data are frequently encountered, for instance in bioinformatics where gene expression data from microarrays are linked to motif data and transcriptional regulator data pertaining to the same genes. Joint analysis of different data blocks is mandated when the noise level in each single block precludes reliable conclusions, and is encouraged when the different data blocks represent different views of the same underlying (biological) system so that, together, may yield a fundamental insight. As a second example of compound data, one may think of theories in psychology (and other disciplines) that rest on the assumption that outcomes (e.g., behaviours) come about through sequential processes (involving, e.g., perceptions, cognitions, affects), the links of which may be subject to important sources of variances (e.g., individual differences variance). Different data blocks may pertain to different parts of such processes, and a joint modelling of such blocks may yield a better understanding of the process under study as a whole.

Models and primary objectives
The primary objectives and model issues occurring in this WP can be divided in different categories that furthermore interrelate with each other:

1. Different dimension-reduction techniques are proposed and developed depending on the context and the model, based on standard techniques, such as (semi-)parametric regression techniques and partial least squares (PLS), different types of supervised and unsupervised clustering, (least-squares) support vector machines, generalised principal components techniques, penalised least squares and more general penalised maximum likelihood techniques. To deal with variable selection, mixture modelling, control of the FDR, maximising mutual information and genetic algorithms will be considered among others.
   Dimension reduction techniques can be applied to different types of information. In forecasting, for instance, the predictor space needs to be reduced, using techniques such as ridge regression and LASSO. Spectra-data, on the other hand, are of an entirely different nature. Such data often first need to be denoised by, for instance, nonparametric smoothing techniques, after which the denoised signal can be approximated using, for example, a spline approach. Another information type is a highly dimensional covariance matrix arising from a longitudinal study. The covariance matrix contains important information with respect to the relationship between the repeated measures and the aim is to model it as a function of covariates of multivariate normal distributions and of multivariate copula models.

2. In multiple testing, one can exploit the suspected correlation structures to increase the information content and hence power through a reinforcing combination of tests carrying similar signals. In the analysis of genotype-phenotype associations, one may for instance incorporate the structure of the genetic map and thus optimise performance under the combined perspective of the null and the alternative hypothesis. In some applications, the object dimension is small (e.g. micro-arrays) rendering asymptotic inference inappropriate. Resampling based multiple testing techniques can then be used as an alternative approach.

3. To join the information blocks in compound data global models including submodels need to be developed. The dimension reduction techniques also need to take into account the relationship between the different blocks of information. For instance, simultaneous clustering models will be developed with special emphasis on overlapping clustering models that imply a simultaneous dimension reduction of two or more blocks. Model justification is rarely feasible when faced with highdimensional data on relatively small samples. Nonparametric statistical models are a natural alternative. For parametric models sensitivity analysis and novel goodness-of-fit techniques are required. Furthermore, the robustness of the proposed models and techniques is also an issue.

Cross-links with other workpackages
WP1-WP2: The data structures handled in these workpackages are often highly dimensional, and the same dimension reduction techniques that will be used in WP5 are also highly relevant in these workpackages.
WP3: Incomplete data problems also occur in the highdimensional context of WP5, for instance in bio-informatics with time to event outcome, and also causal inference is relevant in WP5.
WP4: As data structures in WP5 are often hierarchically structured and some unobserved components can be modelled as random effects, the mixed model is also an essential tool in WP5.

Methodological problems
Methodological problems will be situated on different levels:

1. The study of the mathematical aspects of the models in question, including identifiability issues and model robustness. Also problems of model selection (with special emphasis on model complexity and goodness-of-fit) will be studied.
   More specifically, the high sparseness and the high peakness of spectra data can possibly be dealt with by P-splines with L-1 type penalties, by extending the standard quadratic loss function to the lasso type function or other more appropriate loss functions. Support vector machines, kernel based methods that are well suited for non-linear modelling of highdimensional data, will be studied with respect to their robustness properties towards

outlying observations. Modelling the highly dimensional covariance matrix is computationally problematic and Quasi Monte-Carlo techniques will be explored to try to tackle this problem. For all these problems, suitable estimation algorithms must be designed and implemented, along with a careful evaluation of their performance.

2. Multiple testing works under the assumption that the tests involved are uncorrelated which is often not true. The theory and implementations presently developed have not fully exploited information regarding the location of genetic markers on the genetic map. Especially the spatial information for gene-gene interactions promises to provide more powerful answers. Linked to this is the problem of correlated test statistics, especially when the correlation structure in the genetic material is acknowledged. To simplify matters, new methodology is often derived from the perspective of independent genetic markers which is often unrealistic. It will need to be studied where these techniques break down and how the extra information can be safely exploited.

3. For modelling multi-way data, extensions of techniques used in variable by object data are required. One such possible extension is the double structural equation model, which generalises the ordinary structural equation model. In the development of the new model, special attention will be given to the cases of normally distributed data and categorical data, and to the case in which one of the dimensions pertains to time, which will lead to a double structure latent growth curve model. The model interrelations should be studied in depth.

Partners involved: KUL-1, KUL-2, UU, UCL, UG, UH, UJF, USC.

2. Interactions between workpackages

As explained before, the aim is to develop interactions between the workpackages on at least three different levels: (1) Interlocking complexities in the data, (2) Common modelling approaches, and (3) Common methods and tools. Each of these cross-links will now be explained in detail.

I. Interlocking complexities in the data

In practical situations, often data are encountered that imply interlocking complexities as studied in several workpackages, like censored dependent data, highdimensional incomplete data, etc. To illustrate that a large number of complexities can be encountered in one and the same dataset, consider the following data, which have been studied extensively during phase V of the network : in a study on the presence of caries in childrens' teeth, the data are at the same time multivariate (many teeth per child), time dependent (several teeth exams over time), missing (missing dentist visits), with latent heterogeneity (unobserved variables), hierarchical (tooth surface on tooth within mouth within (sampled) school) and possibly with misclassified covariates and/or response (caries experience scored by dental examiner possibly with error).

Methods will be developed for dealing in an appropriate way with such compound complexities. This will require more than a mere concatenation of results as obtained from different workpackages, because when considering e.g. missing data in a multivariate context new kind of complexities will have to be dealt with.

II. Common modelling approaches

The study of dependence is a recurrent topic across all workpackages. In this regard different approaches will be taken to deal with dependence modelling. A non-exhaustive list of approaches that are studied by members of the network is given in the following table. Within the network, those different approaches will be compared, both on a theoretical level and on the level of analyses of several benchmarking datasets.

Common modelling approaches	WP1	WP2	WP3	WP4	WP5
Boundary and frontier models	x	x	x		
Copula models	x		x	x	x
Random effects models		x	x	x	x
Location-scale regression models	x		x		
Registration	x			x	x
Regression quantiles	x		x	x	x
Time series models	x	x			x
Structural equation models			x	x	x
Semiparametric models	x	x	x	x	x
Accelerated failure time models			x	x	
Longitudinal models			x	x	
Flexible models			x	x	
Dimension reduction models	x	x			x
Mixture models			x	x	x

In addition to the above description of the project per workpackage, the above table illustrates that a single model can be studied under several types of data complexity. Consider for instance the models for boundaries or frontiers. These models present a clear qualitative constraint, and are thus primarily linked to WP1. In some applications, one is interested in studying a time evolution of this frontier, for instance to define a dynamic model of firm efficiency. In that case, tools developed in WP2 will be useful. Frontier models can also be studied under WP3, if one considers a frontier as a truncated regression model.

III. Common methods and tools

The different workpackages will rely on a common set of tools, methods and algorithms. The following table gives some examples of tools that are used within the network across workpackages. Findings on these tools will be exchanged among workpackages, and generic results on these tools will be aimed at, allowing their use in a broad range of contexts.

Common methods and tools	WP1	WP2	WP3	WP4	WP5
Kernel smoothing techniques	X	x	x	x	x
Splines	X	x	x	x	x
Resampling methods	X	x	x	x	x
Regularisation	X	x			x
Empirical likelihood	X	x	x		
Optimisation		x		x	
EM algorithm		x	x		x
Bayesian methodology			x	x	

The organisation of the network will be accomplished taking into account the above interactions between the workpackages. For example, meetings, short courses, ... will be organised on particular themes that are of interest to researchers from several workpackages, working groups will be established that concentrate on well focused research problems that are attacked using different angles according to the chosen model or method, etc. See Section I.7 for more details on the organisation of the network.

3. Interaction between partners within a workpackage

Many of the contributors to a workpackage work actively together. Some explicit examples are summarised in the following table. Note that this list is not complete, but is merely meant to give an idea of the degree and kind of collaborations between partners within a workpackage.

Workpackage	Interaction	Description
WP1	UCL, UH	Copulas in location-scale regression models
	KUL-1, UCL, UJF	Inverse problems and deconvolution
	KUL-1, UJF	P-splines, regularisation and qualitative constraints
	UCL, USC	Empirical likelihood based goodness-of-fit testing
	KUL-1, UJF	Penalty methods for regression
WP2	KUL-1, UCL	Locally stationary models for emotion data
	KUL-1, UCL	Structural breaks in time series
WP3	UG, UH	Frailty models
	UCL, USC	Transfer of tail information for recurrent events data
	UH, LSHTM	Sensitivity analysis for incomplete data
WP4	KUL-2, UU	Flexible models in repeated measurements models
	KUL-2, UH	Flexible modelling for intermittent missing data
	KUL-1, KUL-2	Random effects models in psychometry
	KUL-2, UCL	Causality in random effects models
WP5	UG, UH	Micro-array data analysis
	KUL-1, UG	Robust multivariate statistics
	UCL, UJF	Image analysis and classification

In order to stimulate more interactions between the partners working in a same workpackage, and in order to get to know each others research better (especially of the partners who did not take part in the previous phase of the network), the main coordinator at UCL will organise a 'kick-off' meeting during the first six months. The objective of this meeting is that new collaborations between the partners will be established.

4. Added value gained through the network

The importance of the IAP network for its partners will be reflected in many different ways :

1. *Increased research activities and recognition of the partners of the network*
Thanks to the network, there will be more intensive and effective exchanges within a considerable part of the Belgian statistical community. It is also expected that the international position of the partners will be strengthened by taking advantage of the complementarities in the international contacts, the increased visibility of their work, etc.

2. *Increased interaction between different fields in statistics*
The researchers of the network are sometimes faced to similar problems but in different fields. An example is given by the analysis of heterogeneity, which is a recurrent topic in many fields in statistics and which has been the central theme of one of the workshops organised during phase V. In this respect, each partner clearly benefits from the expertise of the whole network.

3. *More training opportunities for PhD students*
During Phase V, many short courses were organised by the partners of the network. The announcements of these short courses were sent to all members of the network, and as a result the courses were attended by many of its PhD students. These activities will be continued. Moreover, the UCL partner takes part in a doctoral school in statistics and actuarial sciences,

recently created within the French speaking community of Belgium, which will enhance even more the training opportunities of young researchers.

4. *Facilitated application for additional research means*
The intensified contacts between the partners of the network under Phase V have brought them in an ideal position to obtain significant additional research means within the research domain of the network. It is expected that continued collaborations of the partners under Phase VI of the network will continue (if not increase) the chances of obtaining additional grants. For example, the partners of the network will investigate the possibility to take part in a network financed by the European Union (like e.g. a Marie Curie network).

5. Integration of the contribution of the partners

In order to stimulate discussions and collaborations on the three aforementioned levels of complexity (the three 'corners' in the triangle in Figure 1, namely data complexity, models and tools) and on the cross-links between them (the 'edges' of the triangle), a number of activities will be organised within the network, some of which overarch several workpackages:

- a `kick-off' meeting organised at the start of the project, which will make all partners familiar with the research interests and expertise of the other partners
- the annual workshops with the whole network, that will concentrate on the cornerstones of the project, namely the complexity of the data, the different modelling approaches and the different methods and tools
- meetings on specific research topics for smaller audience
- research seminars
- working groups of researchers collaborating on specific research problems
- intensive courses

More details on these activities can be found in Section I.7.

The European partners will participate to the kick-off meeting and the annual workshops, and will be informed of all other activities as well. This dissemination of information will be organised via an electronic newsletter and a continuously updated website. See Section I.7 for more details.

References

The references of papers and books mentioned in the above descriptions per workpackage, are given below. Note that more references, related to the work accomplished during phase V of the network, can be found on the web page of the network (http://www.stat.ucl.ac.be/IAP).

Antoniadis, A., Bigot, J. and I. Gijbels (2005), *Penalized wavelet monotone regression*, IAP-statistics Technical Report Series TR0532.

Antoniadis, A. and J. Fan (2001), Regularization of wavelets approximations (with discussion), *J. Amer. Statist. Assoc.*, 96, 939-967.

Antoniadis, A. and T. Sapatinas (2006), *Estimation and inference in functional mixed-effects models*, IAP-statistics Technical Report Series TR0621.

Antonio, K. and J. Beirlant (2006a), Actuarial statistics with generalized linear mixed models, *Insurance Math. Econom.*, to appear.

Antonio, K. and J. Beirlant (2006b), *Semiparametric regression models for claims reserving and credibility: the mixed model approach*, work in progress.

Azrak, R. and G. Mélard (2004), *Asymptotic properties of quasi-maximum likelihood estimators for ARMA models with time-dependent coefficients*, IAP-statistics Technical Report Series TR0442.

Blozis, S.A. (2004), Structured latent curve models for the study of change in multivariate repeated measures, *Psychol. Meth.*, 9, 334-353.

Bolt, D.M., Cohen, A.S. and J.A. Wollack (2002), Item parameter estimation under conditions of test speededness: application of a mixture Rasch model with ordinal constraints, *J. Educ. Meas.*, 39, 331-348.

Braekers, R. and N. Veraverbeke (2001), The partial Koziol - Green model with covariates, *J. Statist. Plann. Inference*, 92, 55-71.

Braekers, R. and N. Veraverbeke (2005), A copula-graphic estimator for the conditional survival function under dependent censoring, *Canad. J. Statist.*, 33, 429-447.

Brusco, M.J. (2004), Clustering binary data in the presence of masking variables, *Psychol. Meth.*, 9, 510-523.

Campolongo, F., Cariboni, J. and W. Schoutens (2006), The importance of jumps in pricing European options, *Reliab. Eng. Syst. Saf.*, to appear.

Chen, X., Linton, O. and I. Van Keilegom (2003), Estimation of semiparametric models when the criterion function is not smooth, *Econometrica*, 71, 1591-1608.

Cho, Y.H., Kim, J.K. and S.H. Kim (2002), A personalized recommender system based on web usage mining and decision tree induction, *Expert Syst. Appl.*, 23, 329-342.

Cook, R.D. (1986), Assessment of local influence, *J. Roy. Statist. Soc. Ser. B*, 48, 133-169.

Daouia, A. and L. Simar (2006), Nonparametric efficiency analysis: a multivariate conditional quantile approach, *J. Econometrics*, to appear.

Delaigle, A. and I. Gijbels (2004), Practical bandwidth selection in deconvolution kernel density estimation, *Comput. Statist. Data Anal.*, 45, 249-267.

Duchateau, L. and P. Janssen (2004), Penalized partial likelihood for frailties and smoothing splines in time to first insemination models for dairy cows, *Biometrics*, 60, 608-614.

Efron, B. (2004), Large-scale simultaneous hypothesis testing: The choice of a null hypothesis, *J. Amer. Statist. Assoc.*, 99, 96-104.

Fermanian, J.D. and O. Scaillet (2003), Nonparametric estimation of copulas for time series, *J. Risk*, 5, 25-54.

Fieuws, S. and G. Verbeke (2004), Joint modelling of multivariate longitudinal profiles: pitfalls of the random-effects approach, *Stat. Med.*, 23, 3093-3104.

Fieuws, S. and G. Verbeke (2006), Pairwise fitting of mixed models for the joint modelling of multivariate longitudinal profiles, *Biometrics*, 62, 2.

Fieuws, S., Verbeke, G., Boen, F. and C. Delecluse (2006), High dimensional multivariate mixed models for binary questionnaire data, *J. Roy. Statist. Soc. Ser. C*, to appear.

Fils-Villetard, A., Guillou, A. and J. Segers (2005), *Projection estimates of constrained functional parameters*, CentER Discussion Paper 2005-111.

Forni, M., Hallin, M., Lippi, M. and L. Reichlin (2004), The generalized dynamic factor model : consistency and rates, *J. Econometrics,* 119, 231-255.

Forni, M., Hallin, M., Lippi, M. and L. Reichlin (2005), The generalized dynamic factor model : one-sided estimation and forecasting, *J. Amer. Statist. Assoc.*, 100, 830-840.

Forni, M. and L. Reichlin (2001), Federal policies and local economies: Europe and the US, *European Economic Review,* 45, 109-134.

Friedman, J.H. and J.J. Meulman (2004), Clustering objects on subsets of attributes, *J. Roy. Statist. Soc. Ser. B*, 66, 815-849.

Geenens, G. and L. Simar (2005), *Index coefficients estimation in single-index models: the generalized maximum rank correlation estimator*, IAP-statistics Technical Report Series TR0560.

George, E.I. and R.E. McCulloch (1993), Variable selection via Gibbs sampling, *J. Amer. Statist. Assoc.*, 88, 881-889.

Ghidey, W., Lesaffre, E. and P. Eilers (2004), Smooth random effects distribution in a linear mixed model, *Biometrics*, 60, 945-953

Gijbels, I. (2005), Monotone regression. In Kotz, S., Johnson, N.L., Read, C. B., Balakrishnan, N. and B. Vidakovic (Eds), *Encyclopedia of Statistical Sciences*, New York: Wiley, pp 4951-4968.

Gijbels, I., Lambert, A. and P. Qiu (2006), Jump-preserving regression and smoothing using local linear fitting: a compromise, *Ann. Inst. Statist. Math.*, 58, to appear.

Ip, E. (2002), Locally dependent latent trait model and the Dutch identity revisited, *Psychometrika*, 67, 367-386.

Janssen, R., Tuerlinckx, F., Meulders, M. and P. De Boeck (2000), A hierarchical IRT model for criterion-referenced measurement, *J. Educ. Behav. Statist.,* 25, 285-306.

Jansson, M. (2005), *Semiparametric power envelopes for tests of the unit root hypothesis*, Preprint, Department of Economics, University of California, Berkeley.

Johnstone, I.M. and M. Raimondo (2004), Periodic boxcar deconvolution and diophantine approximation, *Ann. Statist.*, 32, 1781-1804.

Kiers, H.A.L. and J.M. ten Berge (1989), Alternating least squares algorithms for simultaneous components analysis with equal component weight matrices in two or more populations, *Psychometrika*, 54, 467-473.

Koenker, R. (2005), *Quantile Regression*, Cambridge University Press.

Koenker, R. and I. Mizera (2004), Penalized triograms: total variation regularization for bivariate smoothing, *J. Roy. Statist. Soc. Ser. B,* 66, 145-163.

Komarek, A. and E. Lesaffre (2006), Bayesian accelerated failure time model for correlated interval-censored data with a normal mixture as error distribution, *Statist. Sinica,* to appear.

Kumbhakar, S. C., Park, B. U., Simar, L. and E.G. Tsionas (2006), Nonparametric stochastic frontiers: a local likelihood approach, *J. Econometrics*, to appear.

Lanckriet, G.R.G., De Bie, T., Cristianini, N., Jordan, M.I. and W.S. Noble (2004), A statistical framework for genomic data fusion, *Bioinformatics*, 20, 2626-2635.

Lesaffre, E. and G. Verbeke (1998), Local influence in linear mixed models, *Biometrics*, 54, 570-582.

Little, R.J.A. and D.B. Rubin (2002), *Statistical Analysis With Missing Data (2nd ed.)*, Chichester: Wiley.

Moerkerke, B., Goetghebeur, E., De Riek, J. and I. Roldan-Ruiz (2006), Significance and impotence: towards a balanced view of the null and the alternative hypotheses in marker selection for plant breeding, *Roy. Statist. Soc. Ser. A*, 169, 61-79.

Molenberghs, G. and G. Verbeke (2005), *Models for Discrete Longitudinal Data,* New York: Springer.

Murphy, S. (2003), Optimal dynamic treatment regimes (with discussion), *J. Roy. Statist. Soc. Ser. B*, 65, 331-366.

O'Brien, L.M. and G.M. Fitzmaurice (2005), Regression models for the analysis of longitudinal Gaussian data from multiple sources, *Stat. Med.*, 24, 1725-1744.

Ombao, H., von Sachs, R. and W. Guo (2005), SLEX analysis of multivariate non-stationary time series, *J. Amer. Statist. Assoc.,* 100, 519-531.

Pardo Fernández, J.C., Van Keilegom, I. and W. González Manteiga (2006), Comparison of regression curves based on the estimation of the error distribution. *Statist. Sinica*, to appear.

Prentice, R.L., Pettinger, M., and G. L. Anderson (2005), Statistical issues arising in the Women's Health Initiative (with discussion), *Biometrics*, 61, 899-911.

Qiu, P. (2005), *Image Processing and Jump Regression Analysis,* New York: Wiley and Sons.

Robins, J.M. (2004), Optimal structural nested models for optimal sequential decisions. In Lin, D. Y. and P. Heagerty (Eds), *Proceedings of the Second Seattle Symposium on Biostatistics,* New York: Springer.

Robins, J.M., Rotnitzky, A. and L.P. Zhao (1994), Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.*, 89, 846-856.

Rousseeuw, P. and M. Hubert (1999), Regression Depth, *J. Amer. Statist. Assoc.*, 94, 388-433.

Rubin, D.B. (2003), Taking causality seriously: Propensity score methodology applied to estimate the effects of marketing interventions, *Lecture Notes in Artificial Intelligence*, 2837, 16-22.

Russo, F., Mouchart, M., Ghins, M. and G. Wunsch (2006), *Statistical modelling and causality in the social sciences*, IAP-statistics Technical Report Series TR0602.

Schulz, E.M., Lee, W.C. and K. Mullen (2005), A domain-level approach to describing growth in achievement, *J. Educ. Meas.*, 42, 1-26.

Simar, L. and P.W. Wilson (2006), Statistical inference in nonparametric frontier models: recent developments and perspectives. Forthcoming in Fried, H., Knox Lovell, C.A. and S. Schmidt, (Eds), *The Measurement of Productive Efficiency, (2nd Ed.)*, Oxford University Press.

Smilde, A.K., Westerhuis, J.A. and S. de Jong (2003), A framework for sequential multiblock component methods, *J. Chemometr.*, 17, 323-337.

Storey, J.D. and R. Tibshirani (2003), Statistical significance for genomewide studies, *Proceedings of the National Academy of Sciences of the USA*, 100, 9440-9445.

Sy, J.P., Taylor, J.M.G. and W.G. Cumberland (1997), A stochastic model for the analysis of bivariate longitudinal AIDS data, *Biometrics*, 53, 542-555.

Tajar, A. (2003), *Measuring and modelling dependence*, unpublished doctoral dissertation, UCL, Louvain-la-Neuve, Belgium.

Tibaldi, F., Molenberghs, G. and G. Verbeke (2006), Conditional mixed models with random effects, *British J. Math. Statist. Psych.*, to appear.

Tuerlinckx, F. and P. De Boeck (2004), Models for residual dependencies. In De Boeck, P. and M. Wilson (Eds), *Explanatory item response models: A generalized linear and nonlinear approach,* New York: Springer, pp. 289-316.

Tuerlinckx, F. and P. De Boeck (2005), Two interpretations of the discrimination parameter, *Psychometrika*, 70, 629-650.

Van Bellegem, S. and R. Dahlhaus (2006), Semiparametric estimation by model selection for locally stationary processes, *J. Roy. Statist. Soc. Ser. B,* to appear.

Van den Noortgate, W., De Boeck, P. and M. Meulders (2003), Cross-classification multilevel logistic models in psychometrics, *J. Educ. Behav. Statist.*, 28, 369-386.

van der Laan, M.J., Hubbard, A. and N. Jewell (2005), *Estimation of treatment effects in randomised trials with noncompliance and dichotomous outcome*, submitted.

Van Keilegom, I. and R.J. Carroll (2006), Backfitting versus profiling in general criterion functions, *Statist. Sinica,* to appear.

Van Keilegom, I., González Manteiga, W. and C. Sánchez Sellero (2004), *Goodness-of-fit tests in parametric regression based on the estimation of the error distribution*, IAP-statistics Technical Report Series TR0409.

Van Mechelen, I., Bock, H.-H. and P. De Boeck (2004), Two-mode clustering methods: a structural overview, *Stat. Methods Med. Res.*, 13, 363-394.

Vansteelandt, S. and E. Goetghebeur (2003), Causal inference with generalized structural mean models, *J. Roy. Statist. Soc. Ser. B*, 65, 817-835.

Vansteelandt, S., Goetghebeur, E., Kenward, M.G. and G. Molenberghs (2005), Ignorance and uncertainty regions as inferential tools in a sensitivity analysis, *Statist. Sinica*, to appear.

Van Zandt, T. (2002). Analysis of response time distributions. In Wixted, J. T. and H. Paschler (Eds), *Stevens' handbook of experimental psychology (3rd ed.), volume 4: Methodology in experimental psychology*, New York: Wiley, pp. 461-516.

Verbeke, G. and G. Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data*, New York: Springer.

Yamamoto, K. and H. Everson (1997), Modelling the effects of test length and test time on parameter estimation using the hybrid model. In Rost, J. and R. Langeheine (Eds), *Applications of latent trait and latent class models in the social sciences*, New York: Waxmann, pp. 89-99.

I. 5. PARTICIPATION OF THE PARTNERS IN THE DIFFERENT WORKPACKAGES

Tick off in the table the participation of the different partners in the different workpackages (delete not used rows and columns in the table). Mention for each partner his/her name and the institution's abbreviation.

	PARTNER	WP1	WP2	WP3	WP4	WP5
P1	Name : L. Simar Institution : UCL	X	X	X	X	X
P2	Name : I. Van Mechelen Institution : KUL-1	X	X	X	X	X
P3	Name : E. Lesaffre Institution : KUL-2			X	X	X
P4	Name : L. Duchateau Institution : UG		X	X		X
P5	Name : N. Veraverbeke Institution : UH	X		X	X	X
EU1	Name : A. Antoniadis Institution : UJF	X	X		X	X
EU2	Name : P. Eilers Institution : UU	X	X		X	X
EU3	Name : W. González Manteiga Institution : USC	X	X	X	X	X
EU4	Name : M. Kenward Institution : LSHTM			X		

I. 6. MAIN SKILLS OF THE PARTNERS

Describe the main skills of each of the partners in relation to the project (15 lines maximum per partner).
Delete not used lines.

P1 - Name : L. Simar
Institution : UCL
Main Skills : The UCL partner is a well recognised expert in the domains of non- and semiparametric regression, frontier estimation, survival analysis, extreme value theory, time series analysis, Bayesian statistics, repeated measurements, wavelets and copulas. In particular, it has made important contributions in the areas of stochastic frontier estimation, location-scale regression models, inference for semiparametric models with non-smooth criterion functions, empirical likelihood techniques and locally stationary time series. The main areas of application of the research of the UCL partner are: econometrics (e.g. instrumental regression, frontier estimation to obtain efficiency of firms), finance (e.g. multivariate volatility modelling), actuarial problems (e.g. risk modelling, modelling of extreme events), biostatistics (e.g. analysis of brain signals, censored and truncated data problems), chemometrics (e.g. drug discovery and development) and engineering (e.g. image analysis, form recognition).

P2 - Name : I. Van Mechelen
Institution : KUL-1
Main Skills : On the level of modelling and inference, the group has extensive expertise in: (a) the area of linear as well as nonlinear mixed models, in particular with regard to Item Response Theory (IRT) models; (b) boundary estimation, deconvolution problems, estimation and testing for smooth/non-smooth curves, inference for curves under qualitative constraints and modelling of extreme events; and (c) clustering models, including a broad range of simultaneous and multiway clustering approaches. Other major model families to which the group significantly contributed include: diffusion models, dimension reduction techniques (including multiway component analysis models, and multidimensional unfolding), and semi- and nonparametric flexible modelling (including partially linear models, single-index models, and additive modelling).
On the level of data-analytic tools, the group focuses on: (a) the development and efficient implementation of suitable estimation methods to deal with a broad range of continuous, discrete, and mixed discrete/continuous optimisation problems, (b) the evaluation of such methods, and (c) model selection and model checking procedures (including goodness-of-fit testing).
In all of the work, special emphasis is put on close links between the actual modelling and data analysis on the one hand, and substantive theories and research questions on the other hand. The most important fields of application are psychometrics (with ramifications into the psychology of individual differences and the psychology of emotion), bioinformatics, medical and engineering applications (including research on surface estimation and image analysis), actuarial and financial applications, econometrics, and chemometrics.

P3 - Name : E. Lesaffre
Institution : KUL-2
Main Skills : Firstly, KUL-2 is known for its statistical developments in mixed effects models. Initially, the (univariate) linear mixed model has been studied, but later (univariate) generalised and non-linear mixed effects models were examined. The technology of mixed effects models has also been transferred to survival models. For instance, the KUL-2 group has developed flexible random effects accelerated failure time models. The search for

tractable computational approaches to multivariate mixed effects models constitutes another class of recent developments. Secondly, the group has done research in a variety of missing data models. Nowadays, research is being done on intermittent missingness problems and joint survival and repeated measurements models. Thirdly, the KUL-2 group has done research in the application and development of new computational procedures for Bayesian computing. In all developments, the starting point was a practical problem which needed to be tackled. The first step was most often to apply a classical statistical approach. In a second step, new developments were explored in order to answer the research question to a more satisfactory degree. A major application area of the group is dental research.

P4 - Name : L. Duchateau
Institution : UG
Main Skills : The members of the UG-partner are attached to different faculties (Science, Agronomy, Veterinary Science, Economy and Sociology) and each member therefore contributes different skills. Quite a few team members are skilled in survival analysis, with extensions in causal inference and frailty modelling. Other main skills represented in the team are related to model checking, with goodness-of-fit and robustness of the model as main research topics. Finally, research in bio-informatics is the keen interest of most of the team members.

P5 - Name : N. Veraverbeke
Institution : UH
Main Skills : The UH has expertise in non- and semiparametric statistical inference. Special attention is given to censored data and frailty models in survival analysis. There is also an established expertise in mixed models and modelling of incomplete and longitudinal data. Smoothing, resampling and sensitivity analysis are some of the techniques used. Furthermore there are skills for dealing with statistical problems related to clinical trials, surrogate markers, infectious disease modelling, microarray data, bio-informatics.

EU1 - Name : A. Antoniadis
Institution : UJF
Main Skills : The research of this group focuses on the development of methodology and asymptotic inference for the analysis of signals, time series and image data containing structure at multiple scales. Such multi-scale phenomena arise in applications across a broad spectrum of the sciences. Currently, one of the primary ways in which to model multi-scale structure is through the use of wavelets, and the skills of the group are related to such approaches. In particular the UJF group studies and develops wavelet based methods for solving inverse problems, for smoothing noisy signals in regression and survival analysis and for developing supervised and unsupervised classification methods via dimension reduction for functional data.

EU2 - Name : P. Eilers
Institution : UU
Main Skills : This partner has extensive expertise and experience of direct relevance and value to the network. The central theme of his work is the efficient use of penalties to impose qualitative constraints on statistical models. This has resulted in many papers on smoothing of diverse types of data, shape-conserving function estimation, and quantile regression. The partner is an expert in statistical computation, which recently has led to ground-breaking work on regression and smoothing for highdimensional data. Inspired by chemometric problems, the group has contributed new approaches to inverse problems and signal processing. A large part of the consulting and research activities of the group considers high volume biological data from genomics and proteomics (SNP arrays, microarrays, mass spectra). The partner has an extensive network of contacts in the field and most of his papers have been written in cooperation with statisticians from many

universities (among which in Belgium, Hasselt (UH), Leuven (KUL), Louvain-la-Neuve (UCL)).

EU3 - Name : W. González Manteiga
Institution : USC
Main Skills : The USC partner is quite heterogeneous and comprises a group of researches with different skills. With a not so long but brilliant history, seminal contributions have been made on empirical processes, bootstrap inference and nonparametric techniques. Jointly with the deep theoretical developments, this group has succeeded in solving many practical problems in industrial, environmental and biomedical contexts. This network partner has strong connections with top researches in different topics, not only in Europe, but also in America. Indeed, being a part of this worthy project will strengthen the connections and will lead to collaborations in the fields of interest of the group : functional data, spatial statistics, financial models, set estimation, small area inference, genomics,… and, of course, within the cornerstone of the group, the goodness-of-fit topic.

EU4 - Name : M. Kenward
Institution : LSHTM
Main Skills : The group has a long track record of methodological research into missing data problems, in both observational and intervention trial settings. Work includes the development of selection, pattern mixture and latent variable (structural equation) models for incomplete data,
and methods for the joint modelling of event time and time varying outcomes. These have been developed in frequentist likelihood, Bayesian and hybrid (multiple imputation, stochastic EM) settings and has involved extensive collaborations with researchers at Hasselt (UH) and Leuven (KUL-2) Universities. The group has extensive experience in the application of such methods in clinical trial and in regulatory settings. Additional and related areas of expertise among the group includes methodological developments for dealing with measurement error, small sample inference in restricted maximum likelihood and the mixed model formulation of smoothing splines. The group has extensive experience of statistical application in a medical environment especially, but not exclusively, in epidemiology (including genetic), perinatal and cardiovascular research, and the neuro-sciences.

I. 7. NETWORK ORGANISATION AND MANAGEMENT (4 pages maximum)

Describe the network's organisation and the practical terms governing collaboration and interaction between the partners (meetings, newsletters, doctoral school, ...).

The organisation and management of the network will be taken care of by the principal partner (UCL), in close collaboration with the other partners. As was the case during Phase V of the network, a variety of activities and initiatives will take place in the network. This will allow the network to converge to a well established cluster of research groups, that acts as one block for matters like the organisation of workshops, meetings, intensive courses, mobility of researchers, etc. Moreover, a constantly updated website and an electronic newsletter will keep everyone informed of the activities going on in the network.

Below, each of the activities and initiatives to organise and manage the network will be explained in detail. Most of them are organised in a very similar way as during Phase V, given that experience during Phase V showed that this organisation functioned very well.

I. Network organisation

1. `*Kick-off*' *meeting*
In order to get to know each others research better (especially of the partners who did not take part in the previous phase of the network) and in order to stimulate more interactions between the partners working in a same workpackage, the main coordinator at UCL will organise a 'kick-off' meeting during the first six months of the project. During this meeting the research topics that will be focused on in the distinct workpackages will be communicated to all partners, so that one becomes maximally familiarised with the expertise of other partners/workpackages. The European partners will participate to this meeting as well.

2. *Workshops*
Five workshops in total are planned to be organised, one every year. All Belgian and European partners will take part in these workshops. A broad range of topics will be addressed during these workshops, related to the five workpackages of the network. In this way, the workshops will be accessible to everyone working in the network (this in contrast to the meetings, see point 3, that will focus on particular research topics). The target audience is not limited to statisticians working in the network. By inviting internationally reputed speakers the aim is to reach an international audience. The workshops will therefore be announced via international email lists and journals, by distribution of leaflets, etc., and the IAP network will be clearly mentioned as organiser of the workshop. The planning of the five workshops is as follows:

- Workshop 1 (during 2007)
  This workshop is a follow-up of the kick-off meeting organised by the UCL partner. During this first workshop, the aim is to strengthen the research connections between the partners of the network, and it will be tried to familiarise everyone with the three cornerstones of the project: the different types of data, models and tools that the partners are experts in. Workshops 2, 3 and 4 will then focus on one of these three cornerstones.

- Workshop 2 (during 2008)
  During this workshop, apart from the presentation of a broad range of research results obtained by members of the network, special attention will be given to the compilation of a set of (complex) benchmark datasets together with associated substantive questions on dependencies and associations. In particular, datasets with compound complexities will be focused on. The aim is to analyse these datasets by using several approaches and to present and compare the results from these approaches during the workshop.

- Workshop 3 (during 2009)
  This workshop will focus on the second cornerstone of the project, namely the comparison of different modelling approaches. Theoretical comparisons of different (and perhaps complementary) possible ways to capture dependencies and associations (including mathematical relations between these approaches) will be studied and presented during the workshop. This workshop will also contain `regular' sessions, during which researchers from in and outside the network will present results related to the themes of the network.

- Workshop 4 (during 2010)
  The third cornerstone, namely the evaluation of major data-analytic and algorithmic tools that show up in several workpackages (like resampling methods, likelihood methods, smoothing techniques, ...) will be the focus of the fourth workshop. In addition, recently obtained research results in a variety of domains related to the network will be presented.

- Workshop 5 (during 2011)
  The aim of this last workshop is:
  1. to report on further workpackage results and cross-link research, and to assemble methodological findings from the different workpackages in all previous stages,
  2. to prepare research on generic methodological conclusions,

3. *Meetings*
A set of meetings will be organised related to particular topics of the network. These meetings will be attended by researchers of the network, but will also be open to other researchers. They usually take one or two days, and are attended by a small group of researchers working in that particular research area. Among the meetings organised during Phase V are meetings on frailty models, missing data, mathematical statistics and interval censored data.

4. *Research seminars*
Each of the participating partners will organise on a regular basis statistics seminars at their universities. Announcements of these seminars will be sent out to most Belgian statisticians, including those participating in the network. These seminars will allow to facilitate the transmission of the research results within the network. In particular, each group will invite on a regular basis members of the partner teams to present their research. Moreover, apart from the regular statistics seminars, seminars will also be organised under the heading of the IAP-statistics network itself.

5. *Working groups*
The interaction between researchers of the network will be facilitated through the working groups. They allow an intensive discussion on specific topics between a limited number of researchers. The most significant working groups during Phase V of the network include groups of researchers working on frailty models, regularisation methods, discrete repeated measures and goodness-of-fit problems. Many of these groups will continue to collaborate during the next phase of the project.

6. *Training*
Several short (intensive) courses will be organised within the framework of the IAP-statistics network. These courses will be intended for all members of the network, and in particular (but not exclusively) for the PhD-students. The announcements will be sent out to all members and posted on the website. No (or reduced) registration fees will be required for IAP-members. Possible topics that could be covered are depth-based methods in multivariate analysis, spatial statistics, ... These courses will be taught by visiting professors or professors of the IAP network itself, depending on the topic of the short course.

Moreover, the UCL partner takes part in a doctoral school in statistics and actuarial sciences, recently created within the French speaking community of Belgium, which will enhance even more the training opportunities of young researchers.

7. *Mobility of researchers*
The exchange of researchers within the network will help to disseminate the gained expertise between different partners of the network. In particular, the mobility of doctoral students within the network will be encouraged. Senior members of the network will also be invited to take part in the jury of PhD defences at partner universities of the network.

II. Network management

1. *Logo*
A logo of the network has been designed for Phase V, and will continue to be used during the next phase. It will be put on all announcements of activities of the network mentioned above, in order to make the network more visible to the outside. The logo is given in Figure 2 and includes a number of important statistical concepts.

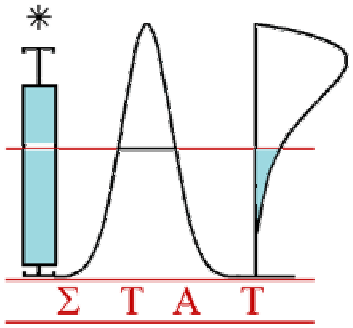

Figure 2: *Logo of the IAP network.*

2. *Website and electronic newsletter*
In order to keep each other updated on the various research activities related to the network a Network's Electronic Newsletter will be established, and the website will continue to be used, as during Phase V. The newsletter will be sent out monthly and will refer to the website for detailed and specific information. It will be sent not only to the members of the network, but also to all those who have asked to be kept informed of the activities of the network. Of course, the website can be consulted on a continuous basis. Its address is

http://www.stat.ucl.ac.be/IAP/PhaseVI/

Among the information that will be exchanged via these means, we mention:
- announcements of seminars related to the network-topics, workshops, meetings, short courses, ...
- visitors, arrival of new post-docs and doctoral students, ...
- calls for applications

Apart from these announcements, the website will also contain the following information:
- description of the project
- list of scientific personnel working under the IAP project
- downloadable member list
- downloadable technical reports, list of publications and list of books written my members of the network (see point 3 below)

- annual reports and reports of scientific meetings

A large part of the website information will be accessible to people from outside the network, so that also other interested researchers can follow the activities and results of the network project.

In a further development of the website, it is planned to use the website as a powerful interface to exchange information within the network. This includes exchange of datasets, programs, presentations, etc. It will also be used to exchange internal information (internal reports, confidential datasets, etc.) through special pages with privileged access.

3. *Technical Reports and Publications Series*
As during Phase V, two series available via the website, will report on scientific results obtained within the IAP-statistics network: the Technical Report Series and the Publications Series. The IAP-statistics Technical Reports Series groups all papers written under the IAP-statistics network. Each paper in this series has been submitted for publication in an international journal. Once a paper has been accepted for publication in an international journal and has been printed, it will be listed in the IAP-statistics Publications Series.
For the IAP-statistics Technical Reports Series all papers (title and authors) will be listed on the website of the network and for each paper a document (ps file or pdf file of the paper) will be posted that can be downloaded from the site. For the IAP-statistics Publications Series the complete reference of each publication will be provided on the website.
In addition, the references of all books written by members of the network will also be mentioned on the website.

4. *Staff meetings*
General staff meetings with representatives of all partner groups will be organised regularly**.** At these staff meetings the partners will discuss issues related to e.g. the organisation of workshops, the preparation of the annual reports, the organisation of short courses, initiatives to make the network more visible, the functioning of the web page, etc. During the annual workshops, a staff meeting will be organised with all Belgian and European partners.

I. 8. RE-ORGANISATION OF THE PROJECT (maximum 3 pages)

To be completed only if the initial proposal has to be adapted as a result of the selection outcome. If this implies changes in the composition of the network and/or the budget, it may be that it is not longer possible to pursue (achieve) the originally proposed objectives.

In this case, describe and clarify the re-organisation of the project compared to the initial proposal.

No re-organisation is necessary.

I. 9. BUDGET (global distribution per partner for the 5 years)

(in EURO, without decimals)

The detailed distribution per partner is given in Section II

	Name Partner	Institution	Budget
P1	L. Simar	UCL	900 000 euros
P2	I. Van Mechelen	KUL-1	650 000 euros
P3	E. Lesaffre	KUL-2	400 000 euros
P4	L. Duchateau	UG	400 000 euros
P5	N. Veraverbeke	UH	400 000 euros
EU1*	A. Antoniadis	UJF	25 000 euros
EU2*	P. Eilers	UU	25 000 euros
EU3*	W. González Manteiga	USC	25 000 euros
EU4*	M. Kenward	LSHTM	25 000 euros
TOTAL BUDGET			2 850 000 euros

* The budget for the EU-partner is the budget attributed by the IAP-programme only (without the 50% contribution of the EU-partner)

I. 10. PREVIOUS IAP-PHASES

To be completed only if the present network was funded during earlier phases of the IAP programme.

Mention the earlier phases of the IAP programme (I, II, III, IV, or V) and the titles of projects in which the partners of the present network has participated.

During Phase V the following partners have participated to the project entitled 'Statistical techniques and modelling for complex substantive questions with complex data' : the Belgian partners UCL, KUL-1, KUL-2, UH and ULB (Université Libre de Bruxelles), and the European partners UJF and RWTH-Aachen (Aachen Technical University). With the exception of the RWTH-Aachen partner and the ULB partner, all of these partners take part in the present project. More information on the accomplishments of the network during phase V (annual reports, overview report, ...) can be found on the web page of the network (http://www.stat.ucl.ac.be/IAP).