FEDERAL OFFICE FOR SCIENTIFIC,
TECHNICAL AND CULTURAL AFFAIRS

rue de la Science 8 ▪ B-1000 BRUSSELS
Tel. 02 238 34 11 ▪ Fax 02 230 59 12
URL : www.belspo.be

OSTC

Interuniversity Attraction Poles (IAP)
Phase V

2002-2006

ANNEX I

TO CONTRACT P5/24

TECHNICAL SPECIFICATIONS

SECTION I

I. Description of the project and the network
II. Budget (global distribution per partner)
III. Composition of the follow-up committee

I. DESCRIPTION OF THE PROJECT AND THE NETWORK

I.1. TITLE OF THE PROJECT

TITLE (in English) : Statistical techniques and modeling for complex substantive questions with complex data

TITLE (in Dutch) : Statistische technieken en modellering voor complexe substantiële vragen op basis van complexe data

TITLE (in French) : Techniques statistiques et modélisation pour des questions substantives complexes et données complexes

I.2. NETWORK COMPOSITION

Network co-ordinator:

1. Name : Léopold Simar (UCL : administrative co-ordinator) and Irène Gijbels (UCL : scientific co-ordinator)
   Institution : Université catholique de Louvain, Institut de statistique
   Type : Principal partner

Partners

2. Name : Paul De Boeck (KUL-1)
   Institution : Katholieke Universiteit Leuven
   Type : Associate partner

3. Name : Emmanuel Lesaffre (KUL-2)
   Institution : Katholieke Universiteit Leuven
   Type : : Corresponding partner

4. Name : Noël Veraverbeke (LUC)
   Institution : Limburgs Universitair Centrum
   Type : : Associate partner

5. Name : Marc Hallin (ULB)
   Institution : Université Libre de Bruxelles
   Type : : Associate partner

6. Name : Hans-Hermann Bock (RWTH - Aachen)
   Institution : Aachen Technical University
   Type : : European partner

7. Name : Anestis Antoniadis (UJF - LMC - IMAG)
   Institution : Université Joseph Fourier
   Type : : European partner

8. Name :
   Institution :
   Type : : Principal partner

9. Name :
   Institution :
   Type : : Principal partner

10. Name :
    Institution :
    Type : : Principal partner

11. Name :
    Institution :
    Type : : Principal partner

12. Name :
    Institution :
    Type : : Principal partner

13. Name :
    Institution :
    Type : : Principal partner

14. Name :
    Institution :
    Type : : Principal partner

15. Name :
    Institution :
    Type : : Principal partner

16. Name :
    Institution :
    Type : : Principal partner

17. Name :
    Institution :
    Type : : Principal partner

18. Name :
    Institution :
    Type : : Principal partner

I.3. SUMMARY OF THE PROJECT

Provide a concise description of the project and indicate clearly and briefly the project's major objectives.

A. SUMMARY IN ENGLISH (max. 1 page)

A key task for statistics is to provide researchers with tools to frame their substantive questions within formal models so as to make them amenable to empirical research. Regarding the latter, an important related task in statistical analysis is to take into account the very nature of the data. In this respect, nowadays one may note an increasing demand in many fields to capture more adequately the complexity of the data that are collected to investigate substantive research questions. Moreover, the substantive research questions themselves also display an ever increasing complexity, especially since the last decade. Both types of complexity constitute a major challenge for contemporary statistics. Novel models and techniques are clearly needed to handle questions and data with a complicated underlying structure, using up-to-date methods in statistical modeling and inference and often involving adaptations/modifications of techniques available for simpler structures.

The point of departure of the proposed network activities is that of a broad range of complex substantive data sets and questions arising in various disciplines (including psychology, biomedical sciences, economics, and climatology). The overall aim of our project then is to develop appropriate statistical models and techniques to deal with these data and questions.

As such, the network activities will be organized into 6 work packages which have been further grouped in two major sections. Section I includes 4 work packages (WP1-WP4) that focus on 4 well-delineated *classes of models*. Section II includes 2 work packages (WP5-WP6) that can be considered to deal with statistical *meta-modeling aspects*; the latter can be studied in their own right but, in addition, can also be included within different classes of models as distinguished within Section I.

The key aims of the six work packages (WPs) can be summarized as follows:

- WP1 (*Functional estimation*): to expand classical functional estimation of one- and multidimensional curves in line with more realistic (but more complex) substantive theories (in particular: economic theories involved in frontier estimation) and to capture in appropriate ways change or break points;
- WP2 (*Time series*): to deal with two major sources of complexity in the analysis of multivariate time series: nonstationarity and high-dimensional data;
- WP3 (*Survival analysis*): the study of nonparametric regression models with a complex censoring mechanism or involving discontinuities, and of frailty models to capture heterogeneity;
- WP4 (*Mixed models*): to look for adequate random effects distributions;
- WP5 (*Classification and mixture models*): how to capture the heterogeneity in a population and what its exact nature is;
- WP6 (*Incompleteness and latent variables*): the development of (semi)parametric missingness models for incomplete and latent data, and the study of sensitivity to various assumptions implied by this modeling.

Integration of the network activities will be achieved on four different levels:

1) *substantive*: data sets will be shared by different work packages and as such will be analyzed in terms of different, complementary models;
2) *cross-links* will be established between pairs of work packages: e.g., survival models will be studied in WPs 3 and 4; latent variables will be addressed in WPs 5 and 6;
3) *interaction between Section I and Section II*: e.g., classification techniques and mixture models as studied in WP5 to capture heterogeneity will be applied within various WPs of Section I;
4) *common methodological ground*: (a) the vast majority of the work packages will make use of smoothing and bootstrap/Bayesian data analysis techniques as common methodological tools; (b) a number of methodological research topics will be addressed by assembling methodological findings from different work packages, which should finally lead to the drawing of generic conclusions.

The proposed research should result into novel types of statistical methods, models and model expansions that fit better to complex substantive theories as well as to complex data. As such they should provide researchers with more effective and useful analysis tools for answering important present-day questions.

B. SUMMARY IN DUTCH (max. 1 page)

Een belangrijke opdracht voor statistiekers bestaat in het voorzien van methoden die onderzoekers kunnen aanwenden voor het inkaderen van hun substantiële vragen in formele modellen, zodat deze handelbaar worden voor empirisch onderzoek. Wat dit laatste betreft, een belangrijke nauw verbonden opdracht in statistische analyse bestaat in het rekening houden met de natuur van de data. Tegenwoordig is er in vele domeinen een stijgende vraag naar het adequaat verwerken van de complexiteit van de data, die verzameld worden om substantiële vragen i.v.m. onderzoek te bestuderen. Bovendien vertonen de substantiële vragen i.v.m. onderzoek zelf een toenemende complexiteit, in het bijzonder vanaf het laatste decennium. Beide types van complexiteit betekenen een belangrijke uitdaging voor de hedendaagse statistiek. Nieuwe modellen en technieken zijn duidelijk noodzakelijk voor het verwerken van vragen en data met een complexe onderliggende structuur, gebruik makend van up-to-date methoden in statistische methodologie en besluitvorming en steunend op aanpassingen van bestaande technieken voor eenvoudigere structuren.

Het vertrekpunt van de voorgestelde netwerk activiteiten bestaat uit een breed gamma van complexe substantiële data sets en vragen die in verschillende disciplines voorkomen (waaronder psychologie, biomedische wetenschappen, economie en climatologie). Het hoofddoel van ons project is het ontwikkelen van geschikte statistische modellen en technieken om deze data en vragen te kunnen verwerken.

Als zodanig zullen de netwerk activiteiten georganiseerd worden in 6 werkmodulen die verder gegroepeerd worden in twee secties. Sectie I bestaat uit 4 werkmodulen (WP1-WP4) die zich toeleggen op 4 wel omschreven *klassen van modellen*. Sectie II omvat 2 werkmodulen (WP5-WP6) die statistische *aspecten van meta-modellering* behandelen; deze laatste kan op zich zelf bestudeerd worden, maar kan daarenboven ook deel uit maken van verschillende klassen van modellen, zoals onderscheiden in Sectie I. De belangrijkste doelstellingen van de zes werkmodulen (WPs) kan als volgt samengevat worden :

- WP1 (Schatten van functies) : het uitbreiden van klassieke methoden voor het schatten van één- of meer-dimensionele functies naar meer realistische (maar tegelijkertijd ook meer complexe) substantiële theoriëen (in het bijzonder : economische theoriëen die gebruik maken van het schatten van `frontiers') en het op een geschikte manier formaliseren en modelleren van heterogeniteit
- WP2 (Tijdreeksen) : het behandelen van twee belangrijke bronnen van complexiteit in de analyse van multivariate tijdreeksen : niet-stationariteit en hoog-dimensionale data
- WP3 (Overlevingsanalyse) : het bestuderen van niet-parametrische regressiemodellen met een complex censureringsmechanisme of die discontinuiteiten bevatten, en van frailty modellen voor het opvangen van heterogeniteit
- WP4 (`Mixed models') : het zoeken naar geschikte `random effects' verdelingen
- WP5 (Classificatie en `mixture' modellen) : het opvangen van heterogeniteit in een populatie en het bestuderen van zijn oorsprong
- WP6 (Onvolledigheid en latente variabelen) : het ontwikkelen van (semi)parametrische `missingness' modellen voor onvolledige en latente data, en het bestuderen van de sensitiviteit voor verschillende onderstellingen die gemaakt worden bij deze modellering

Integratie van de netwerk activiteiten zal op vier verschillende niveau's bereikt worden :

1) *substantieel* : data sets zullen gedeeld worden door verschillende werkmodulen en zullen als zodanig geanalyseerd worden m.b.v. verschillende, complementaire modellen;

2) *`cross-links'* zullen tot stand gebracht worden tussen paren van werkmodulen, bv. overlevingsmodellen zullen bestudeerd worden in WPs 3 en 4; latente variabelen zullen behandeld worden in WPs 5 en 6;

3*) interacties tussen Sectie I en Sectie II:* bv, classificatie technieken en `mixture' modellen zoals bestudeerd in WP5 voor het opvangen van heterogeniteit, zullen toegepast worden in verschillende WPs van Sectie I;

4) *Gemeenschappelijke methodologische basis :* (a) het grootste deel van de werkmodulen zal gebruik maken van smoothing en bootstrap/Bayesiaanse data analyse technieken als gemeenschappelijke methodologische hulpmiddelen; (b) een aantal methodologische onderzoeksprojecten zullen bestudeerd worden door gebruik te maken van methodologische resultaten van andere werkmodulen, hetgeen uiteindelijk moet leiden naar het trekken van algemene conclusies.

Het voorgestelde onderzoek moet resulteren in nieuwe types van statistische methoden, modellen en model uitbreidingen die beter geschikt zijn voor zowel complexe substantiële theoriëen als complexe data. Als zodanig zullen zij onderzoekers voorzien van efficientere en nuttige hulpmiddelen voor het beantwoorden van belangrijke hedendaagse vragen.

C. SUMMARY IN FRENCH (max. 1 page)

La tâche essentielle de la statistique consiste à fournir aux chercheurs des outils leur permettant de cadrer leurs questions substantives dans des modèles formels de manière à les rendre accessibles à la recherche empirique. En ce qui concerne cette dernière, une importante tâche connexe en analyse statistique est de prendre en compte la véritable nature des données. De ce point de vue, on peut constater aujourd'hui dans beaucoup de domaines une demande croissante d'appréhender plus adéquatement la complexité des données collectées pour pouvoir répondre à des questions substantives de recherche. De plus, les questions substantives de recherche elles-même font également apparaître une complexité toujours croissante, spécialement au cours de cette dernière décennie. Ces deux types de complexité constituent un défi majeur pour la statistique contemporaine. De nouveaux modèles et de nouvelles techniques sont clairement nécessaires pour traiter ces questions et ces données qui ont une structure sous-jacente compliquée. Cela se fera en utilisant des méthodes modernes de la modélisation et de l'inférence statistique et cela sous-entend des adaptations et/ou des modifications des techniques disponibles pour des modèles plus simples.

Le point de départ des activités du réseau proposé est une grande collection de données et de questions substantives issues de disciplines diverses (incluant la psychologie, les sciences biomédicales, les sciences économiques et la climatologie). Le but ultime de notre projet est alors de développer des modèles et des techniques appropriés pour traiter ces données et ces questions.

Les activités du réseau en tant que telles seront organisées en six modules de travail qui ont de plus été regroupés en deux sections majeures. La Section I comporte 4 modules (WP1-WP4) qui s'occuperont de 4 *classes de modèles* bien définis. La Section II comporte 2 modules (WP5-WP6) du type *méta-modélisation statistique*. Ils peuvent être étudiés en eux-mêmes, mais ils peuvent également être inclus dans les différentes classes de modèles de la Section I.

Les objectifs majeurs des six modules peuvent être résumés comme suit :

- WP1 (*Estimation fonctionnelle*) : étendre l'estimation fonctionnelle classique de courbes uni- ou multidimensionnelles en tenant compte de théories substantives plus réalistes (mais plus complexes) (en particulier : les théories économiques pour l'estimation de frontières) et déterminer de manière adéquate les points de rupture ou de changement ;
- WP2 (*Séries chronologiques*) traiter deux sources majeures de complexité dans l'analyse des séries multivariées : la non-stationnarité et les données de grande dimension ;
- WP3 (*Analyse de survie*) : étudier les modèles de régression non paramétriques avec un mécanisme de censure complexe ou impliquant des discontinuités et des modèles agrégés (« frailty ») pour tenir compte de l'hétérogénéité ;
- WP4 (*Modèles mixtes*) : rechercher des distributions adéquates pour les effets aléatoires ;
- WP5 (*Classification et modèles de mélange*) : capturer l'hétérogénéité dans une population et en comprendre sa nature exacte ;
- WP6 (*Données incomplètes et variables latentes*) : développer des modèles de non-réponses (semi)paramétriques pour des données incomplètes ou latentes et étudier la sensibilité aux diverses hypothèses intervenant dans la modélisation.

L'intégration des activités du réseau sera accomplie sur 4 niveaux :

(1) *substantif* : les données seront étudiées par différents modules et donc en termes de modèles distincts mais complémentaires ;

(2) des *liens croisés* seront établis entre paires de modules : ainsi les modèles de survie seront étudiés dans WP3 et WP4, les variables latentes dans WP5 et WP6 ;

(3) *interaction entre les Sections I et II* : par exemple, les techniques de classification et les modèles de mélange étudiés dans WP5 pour capturer l'hétérogénéité seront utilisés dans divers modules de la Section I ;

(4*) base méthodologique commune* : (a) la plupart des modules utiliseront comme outils méthodologiques les techniques de lissage et d'analyse des données par rééchantillonnage et Monte-Carlo ; (b) des sujets de recherche méthodologiques émergeront de la réunion des réalisations méthodologiques obtenues dans différents modules. Cela devrait permettre de tirer des conclusions génériques.

La recherche proposée devrait résulter en de nouveaux types de méthodes statistiques, de modèles et d'extensions de modèles qui épouseraient mieux les théories substantives complexes aussi bien que les données complexes. En tant que tels, ils devraient fournir aux chercheurs des outils utiles et plus efficaces pour répondre aux questions importantes actuelles.

I.4. OBJECTIVES, MOTIVATION AND STATE OF THE ART

(max. 5 pages)

Describe the project's objectives. Define the problems being addressed by positioning them in relation to the current state of knowledge. Justify the relevance of the proposed methods and approaches in accordance with the state of the art.

---

1. Motivation

A key task for statistics is to provide researchers with tools to frame their substantive questions within formal models so as to make them amenable to empirical research. Regarding the latter, an important related task in statistical analysis is to take into account the very nature of the data. In this respect one may note an increasing demand nowadays in many fields to capture more adequately the complexity of the data that are collected to investigate substantive research questions. Moreover, the substantive research questions themselves also display an ever increasing complexity, especially since the last decade. We illustrate both types of complexity.

(1) Regarding complexity of data, first *incomplete data* arise in many fields of applications. Examples include the occurrence of missing data in longitudinal biomedical research and in economics (where details of the financial situation of individuals or companies often are not fully available). Other examples include the occurrence of censoring; regarding the latter one may note the increasing importance of complex censoring schemes like interval censoring, for instance in AIDS research where the time until the development of AIDS for HIV patients is usually only known to lie in between two examination times. A second type of data complexity concerns *latent data*, that is, data that are intrinsically not directly observable such as personality traits in psychology or causal effects in medicine. A third illustration of data complexity concerns *high-dimensional data* which are the rule rather than the exception in stock exchange and macroeconomic studies as well as in studies based on clinical monitoring (where modern medical technology allows for an increasing number of simultaneous on-line measurements of a patient's status.)

(2) Regarding complexity of substantive questions, first one may wish to move to *more realistic accounts of phenomena*. Examples include the study of the economic productivity of companies, which can be translated into the problem of estimating the boundary curve of the support of a multivariate density function; one may wish to address such a frontier estimation problem with flexible nonparametric methods that do not rely on the unrealistic assumption that all companies produce under the theoretical optimal efficiency curve (which is implied by nonparametric frontier estimation methods that are restricted to the case of deterministic frontiers). Second, one may wish to refine universal accounts of phenomena by allowing for *heterogeneity* of experimental units under study. Examples include heterogeneity of individuals in the study of emotions in psychology and heterogeneity of medical centers taking part in a clinical trial. As a third illustration we point to questions regarding complex *structures in time and/or space*. Examples include the study of stock exchange and macroeconomic time series data that imply complex time structures such as nonstationarity and multiple breakpoints in trend or volatility (caused, for example, by external shocks on the stock market). Other examples include questions on the joint structure in time and space in longitudinal medical studies (e.g., of the caries status of various deciduous and permanent teeth).

Both types of complexity constitute a major challenge for contemporary statistics. Novel models and techniques are clearly needed to handle questions and data with a complicated underlying structure, using up-to-date methods in statistical modeling and inference and often involving adaptations/modifications of techniques available for simpler structures.

2. Aims and state of the art

2.1 Overall conceptual structure underlying the network activities

Figure 1 below presents an overview of the four major components of the proposed network activities. We will now briefly introduce each of them successively.

(1) The point of departure of the network activities is that of a broad range of complex substantive data sets and questions arising in various disciplines (including psychology, biomedical sciences, economics, and climatology).

(2) The overall aim of our project is to develop appropriate statistical models and techniques to deal with the data and questions referred to in (1). As such, the network activities will be organized into 6 work

packages which pertain to different modeling aspects. These work packages have been further grouped in two major sections:
(2a) Section I includes 4 work packages that focus on 4 well-delineated classes of models: functional estimation models, time series, survival analysis and mixed models.
(2b) Section II includes 2 work packages dealing with statistical modeling aspects that can be studied in their own right but that, in addition, can also be included within different classes of models as

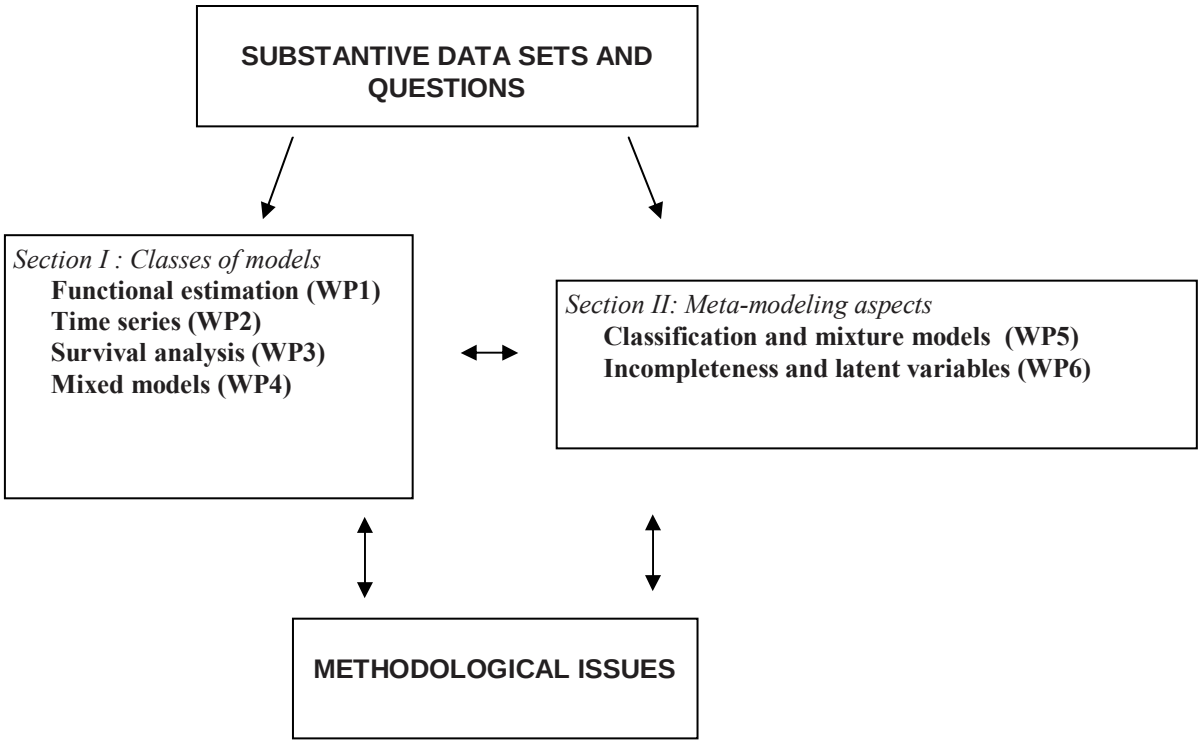

Figure 1: Overall conceptual structure of the proposed network activities

distinguished within Section I: classification & mixture models and incompleteness & latent variables. As such the work packages of this section can be located at a meta-level.
(3) Methodological issues will have to be addressed in all work packages and in all phases of the project. The lowermost box in Figure 1, however, refers to two types of common methodological aspects that play a key role through all of the network activities: (1) methodological tools that will be used in the vast majority of the work packages, and (2) methodological research topics that will be addressed by assembling methodological findings from different work packages and that should finally result in the drawing of generic conclusions.

2.2 Overview of work packages: Key aims and state of the art

We now proceed with a specification of the key aims per work package. Subsequently, we will briefly present per work package the state of the art with respect to these key aims, both in general scientific terms and in terms of previous accomplishments from the teams that will contribute to the work package in question.

Work package 1: Functional estimation

Aim

This workpackage is on non- and semiparametric estimation of one- and multidimensional curves, such as regression and autoregression, frontier curves, densities, quantile and hazard functions or (time-varying) spectral densities. Often it is assumed that the function to be estimated has a certain degree of smoothness all over its domain of definition. The key aim of this work package is to expand classical functional estimation of one- and multidimensional curves in line with more realistic (but more complex) substantive theories (in particular: economic theories involved in frontier estimation) and to capture in appropriate ways change or break points.

State of the art

Up to now the emphasis in frontier estimation research has been mainly on parametric and semi-parametric methods. The study of (more flexible) nonparametric methods has been restricted to the case

of deterministic frontiers, i.e. all companies produce under the theoretical optimal efficiency curve (see Gijbels, Mammen, Park and Simar, 1999; and Park, Simar and Weiner, 2000). The challenge here is to develop nonparametric methods for stochastic frontiers, which describe better the reality. Hall and Simar (2000) contains the first attempt to this challenging problem.

In the last decade, non- and semi-parametric methods for detecting change-points have been developed (see, for example, Müller and Stadtmüller, 1999; and Hall and Rau, 2000). But often the proposed methods assume that one needs to know the number of change-points; moreover, all methods depend on a (crucial) choice of smoothing parameters. Recently, automatic procedures for estimating the number of change points as well as selecting the smoothing parameters have been proposed in a regression context (see e.g. Gijbels and Goderniaux, 2000).

Nonparametric methods to deal with breakpoints in time series have been proposed in a variety of papers (as for breakpoints in trend, see, e.g., von Sachs and MacGibbon, 2000); wavelet-type estimators have been proposed recently also for nonparametric regression (Delouille, Franke and von Sachs, 2000), with possible generalizations to the autoregression context.

Workpackage 2: Time series

Aim

The aim of this workpackage is to deal with two major sources of complexity in the analysis of multivariate time series: nonstationarity and high-dimensional data (as opposed to long series lengths, which tend to facilitate the analysis). We will address these two important problems by means of (1) the development of applicable approaches to non-stationarity, and (2) adequate methods of dimension reduction. The common thread connecting (1) and (2) is a search for an appropriate decomposition of the data, based on spectral Principal Component Analysis techniques (Brillinger, 1981).

State of the art

Nonstationary time series have been discussed by Dahlhaus (1996, in press). Ombao et al (in press-c) developed a methodology to deal with nonstationarity using a specifically localized Fourier basis (SLEX: Smoothed Localized complex EXponentials); a number of important issues regarding this methodology remains untouched, in particular, to render the SLEX method fully multivariate (and not just bivariate, as is the state-of-the-art). Furthermore, Ombao et al. (in press-c) also devised a particular segmentation algorithm which minimizes a certain entropy cost based on the estimated spectral matrix in order to deliver the "best adapted" piecewise stationary approximation to the given signal.

An (econometric) introduction to multivariate time series van be found in Lütkepohl (1993).

In spectral Principal Component Analysis, the observations are decomposed into two mutually orthogonal components, a "common" and an "idiosyncratic" one, where the common component results from the action of a "low-dimensional" unobserved series of common shocks---the "factors". A consistent method of estimation of these common components, based on a dynamic factor-analytic technique, has been proposed by Forni et al. (2000, 2001, in press); this method has been successfully applied to large cross-sections of time series data, with dimension as high as 1,000.

Workpackage 3: Survival analysis

Aim

In this workpackage we focus on regression models in which the response variable is censored. Key aims are (1) the study of nonparametric regression models in which the censoring mechanism is complex or the regression function contains discontinuities and (2) the study of heterogeneity in terms of frailty models. The common main objective in both situations is to do inference on the object function (the regression function for (1) and the hazard function for (2)) which describes the relation between the response variable (the variable of final interest) and the explanatory variables.

State of the art

In regression with interval censored data, research has been mainly concentrated on studying (semi)parametric regression models (see e.g. Rabinowitz, Tsiatis and Aragon, 1995; Datta, Satten and Williamson, 2000). Estimation in models with discontinuities when the response is censored has been studied recently by Antoniadis, Gijbels and MacGibbon (2000) and Gijbels and Gürler (2001), using modern smoothing techniques such as wavelets and local approximation techniques.

In frailty modeling much effort has been put in understanding and implementation of the underlying likelihood theory. Key references include Nielsen, Gill, Andersen and Sörensen (1992) and Parner (1998). For an excellent applied approach, see Therneau and Grambsch (2000).

Workpackage 4: Mixed models

Aim

Various types of mixed models have been suggested in the literature (Verbeke and Molenberghs, 2000): linear mixed, generalized linear mixed, nonlinear mixed. All involve the combination of a fixed and a random effects part, both describing the influence of explanatory variables. Key aim of this work package is to look for adequate random effects distributions.

State of the art

Except for the nonparametric maximum likelihood method, the random effects distributions are usually taken from a specific parametric family. It seems that more emphasis should be put on the development of smooth densities for the random effect (Davidian and Gallant, 1993 and Shen and Louis, 1999). Considering a finite mixture of parametric distributions could be a start (see, e.g., Verbeke and Lesaffre, 1996), since it can be shown that they can approximate the correct underlying distribution arbitrarily close (Diaconis and Ylvisaker, 1979).

Lesaffre and Verbeke (1998) and Verbeke, Lesaffre and Brant (1998) have developed a variety of goodness-of-fit tests for the linear mixed model. Furthermore, the same authors have examined the properties of the normal mixed model with a mixture of normal distributions as random-effects distribution. Buyse and Molenberghs (1998) and Buyse et al. (2000) have developed a methodology to evaluate surrogate markers in a meta-analytic and longitudinal context by means of mixed models. Butter, De Boeck and Verhelst (1998) and Tuerlinckx and De Boeck (in press) studied IRT-type psychometric models (extensions of the Rasch and LLTM models), which involve a logistic link function and which are a particular case of non-linear mixed models.

Workpackage 5: Classification and mixture models

Aim

This work package deals with the problem of how to capture the heterogeneity in a population and what its exact nature is (e.g., are differences between objects primarily quantitative or qualitative in nature? If qualitative, how many components are involved?) Various mixture models will be focused on but other classification approaches will be considered as well.

State of the art

In the recent literature on mixture models a considerable number of novel types of mixture models have been developed, including mixed models with mixture random components (Verbeke and Lesaffre, 1996; Verbeke and Molenberghs, 2000; Lenk and DeSarbo, 2000), mixture structural equation models (Dolan and Maas, 1998; Yung, 1997; Zhu and Lee, 2001), mixed tree models of dissimilarities (Wedel and Bijmolt, 2000), mixture IRT models (Rost, 1997), and PMD models, which are a special type of contrained latent class models (Meulders, De Boeck, Van Mechelen, Gelman and Maris, in press; Meulders, De Boeck, and Van Mechelen, in press).

Inference on the number of mixture components has been dealt with by means of Bayes factors and posterior predictive checks (Berkhof, Van Mechelen and Hoijtink, 2000; Fraley and Raftery,1998), by modeling the number of components and making use of reversible jump MCMC methods (Richardson and Green, 1997, 1998); by modified likelihood ratio tests (Chen, Chen and Kalbfleisch, 2001), and by information approaches (Bozdogan 1994) (see also Bock, 1996).

For a state-of-the-art on classification models and algorithms more in general, see Bock (1996, 1999, 2001). For a novel broad decomposition family of classification models, see Leenen, Van Mechelen and De Boeck (1999).

Work package 6: Incompleteness and latent variables

Aim

The key aim of this work package is the development of parametric and semiparametric missingness models for incomplete longitudinal and latent data, and the study of sensitivity to various assumptions implied by this modeling.

State of the Art

For a general overview, see Little and Rubin (1986). Pattern mixture models have been developed by Hogan and Laird (1997a, 1997b), Little (1994), and Little and Wang (1996), models for incomplete data by, e.g., Diggle and Kenward (1994), and Little (1995), and semiparametric models by Robins et al. (1995a, 1995b, 1998). Sensitivity has been studied by Kenward (1998).

Publications of the teams contributing to this work package include Molenberghs, Kenward and Lesaffre (1997), Lipsitz et al. (1997) (on incomplete data modelling), and Michiels, Molenberghs and Lipsitz (1999), Kenward and Molenberghs (1998) and Verbeke et al. (2001) (on sensitivity analysis).

3. Coherence, complementarity and added value

3.1 Coherence

Coherence of the project will be achieved on four different levels:

1) substantive: applications in different substantive domains will be addressed, but in many places different substantive questions will imply important formal similarities (coming down to, e.g., heterogeneity, regression-type functional relations etc.); moreover, data sets will be shared by different work packages and as such will be analyzed in terms of different, complementary models.
2) cross-links between pairs of work packages: as an example, survival models will be studied in work packages 3 and 4; latent variables will be addressed in work packages 5 and 6.
3) interaction between Section I and Section II: for example, classification techniques and mixture models as studied in work package 5 will be applied within the time series analysis context of work package 2 and within the mixed models context of work package 4; sensitivity analysis, which will be primarily investigated within work package 6, will be subsequently applied within work packages 3 and 4.
4) there will be an important common methodological ground for all work packages in two respects: First, the vast majority of the work packages will make use of two common methodological tools -- smoothing techniques on the one hand and computer-intensive methods, such as bootstrap and Bayesian data analysis techniques, on the other hand--. Second, a number of well-defined methodological topics will be investigated in multiple work packages; in the final stage of the project, relevant findings on these topics will be assembled so as to draw more generic methodological conclusions; these methodological topics include: the applicability of bootstrap procedures for estimation and inference in nonstandard settings and the study of the comparative performance of smoothing and mixture modeling techniques.

3.2 Complementarity of the partners

The partners have complementary expertises on three different levels:

1) domains of substantive applications: e.g., UCL: economic and climatology problems, KUL-1: psychology, LUC/KUL-2: biomedical applications
2) types of models: e.g., ULB/UCL: time series analysis, KUL-2/LUC: mixed models,
3) methodology: e.g., UCL: smoothing techniques, LUC/UCL: bootstrap, KUL-1&2/LUC: Bayesian data analysis

3.3 Added value of the proposed research

From a content point of view, the proposed research should result into novel types of statistical methods, models and model expansions that fit better to complex substantive theories as well as to complex data. As such they should provide researchers with more effective and useful analysis tools for answering important contemporary questions.

From a formal, organizational, point of view, the proposed network should result in:

1) more intensive and effective exchanges within a considerable part of the Belgian statistical community
2) more and richer training opportunities for young statistical researchers
3) strengthening of possibilities to recruit suitable PhD students/researchers through joint advertising/operating on the international job market
4) strengthening of international position of the partners (benefit from complementarities in international contacts; increased visibility of their work.

I.5 DETAILED DESCRIPTION OF THE PROJECT

(min. 5 pages, max. 10 pages)

Submit a general description of the project as well as a precise description detailing each work package (coherent packages contributing to the pursuit of the project's interim objectives). At the same time, indicate the partners involved in each work package.

As explained earlier, the different applications we have in mind for the project share that the data are complex (e.g., high-dimensional, censored, unobserved) and that the research questions to be answered turn out to be complex as well. At a rather abstract level, the questions may be summarized as follows: First, in all cases the issue is to explain variables in terms of other variables (either observed variables as a function of other observed variables, or observed variables as a function of unobserved underlying or latent variables). Second, in all cases the functional form of the relation is an issue as well as the form of the distribution of the unobserved variables and of the effects of the explanatory variables.

In paragraph 1 on *work packages* it will be explained how these general issues take specific forms depending on the kind of application, the kind of data, and the kind of assumptions one wants to make. When formulated in this more specific context, the questions are quite complex, as will be clear from the description of the work packages. As explained in form D1, six work packages are distinguished in the project, and they are organized into two sections. Four of the work packages (WP1 to WP4) are related to classes of models we consider being of a basic kind: models for functional forms (WP1), time series models (WP2), survival analysis models (WP3), and mixed models (WP4). These four work packages are grouped into *section I* of the project. Two other work packages relate to models that are of interest not only in their own right but also for their potentialities to enrich the so-called basic classes of models: WP5 on classification and mixture models, and WP6 on models for incompleteness and latent variables. The latter two work packages are grouped into *section II* on meta-modeling aspects.

The work packages will each be described by indicating (a) the kind of substantive applications and problems one will concentrate on, (b) the models that will be used and the primary objectives that are set out for adapting or developing models and methods, to make them better suited for the complex issues one wants to tackle, (c) specific cross-links with other work packages, and (d) methodological problems. After the presentation of the work packages, it will be indicated which partners will contribute to each of them.

The following three paragraphs focus on the interaction between the two sections, common methodological issues in the project, and the time schedule for the activities:
-In paragraph 2 on *interaction between sections*, it will be explained how the concept of the project implies an intensive interaction between the two work packages from section II and the four work packages from section I, and what the content of that interaction will be.
-In paragraph 3, *methodological issues* with relevance for all of the network activities will be described. As indicated in form D1, we draw a distinction between common methodological tools and general methodological research topics. Regarding the latter, we will explain that the project will concentrate on: (1) the development and evaluation of smoothing techniques in a complex setting, (2) the evaluation of resampling/Monte Carlo methods, and (3) sensitivity and the effects of misspecification.
-In paragraph 4 *on the time schedule*, a timing will be given for the activities including those pertaining to dependencies between the two sections of the project, and those pertaining to methodology. The organizational aspects related to the time schedule are described in form H.

1. Work Packages

WP1: Functional estimation

Substantive applications/problems

Functional estimation and models for abrupt changes find their application in various contexts. In particular, functional estimation tools are important in economics and finances, for the analysis of productive efficiencies, for the assessment of performance scores of mutual funds in a stock market, and for finding indicators to compare new products in sustainable development perspectives. Abrupt changes from their part are relevant in various domains:

- It is well known that external shocks like a political crisis or global economical decisions are likely to have an impact on the trend and the volatility of stock market data. On the other hand, daily fluctuations

are not very interesting for the problem of estimating and predicting the development of returns. So for the statistical modeling and analysis of financial time series data we need to develop methods that are able to capture at the same time smooth parts and possibly abrupt transient behavior of the curve under study.
- It is well known that the growth of children occurs mainly in 'spurts' (abrupt changes in growth acceleration) and that the moments at which these spurts take place can differ across children.
- Detection of abrupt changes has also many applications in the bivariate case. Some examples of this are: detection of wind-fronts (sudden changes in wind direction) in the atmosphere, and detection of 'fault lines' in a surface undersea.

Models and primary objectives
As explained in form D1, the *aim* of this work package is to expand classical functional estimation to capture more accurately a complex reality, which includes capturing change-points in an appropriate way. Three *primary objectives* are formulated:
(1) *Nonparametric estimation of a frontier function in case of stochastic frontiers*. The objective is to improve the estimation of frontier functions in the presence of noise. For the purpose of estimation of the productivity of companies, a modelization by frontier curves, representing the maximum attainable output for a given input, has been proven very efficient. The data come into the form of a multivariate cloud of points, each point representing the productivity of one company. Hall and Simar (2000) show that the problem can be viewed as a problem of deconvolution, in which the interest is in estimating the boundary of the support of a density. Automatically selecting the smoothing parameter of a density estimator in the situation of contaminated data (the deconvolution context) has been studied recently in Delaigle and Gijbels (2001). These procedures need to be adapted to estimating the endpoints of the support of the density. In the frontier context, multivariate extensions of Hall and Simar (2000) need to be worked out. For both nonparametric stochastic frontier estimation and density deconvolution the use of bootstrap methods is rather non-standard.
(2) *Automatic detection of change-points in regression, hazard and density functions*. The objective is to develop fully data-driven methods for detecting change-points. The most important issues here are the automatic detection of the number of change-points, followed by fully data-driven estimation of their locations, their magnitudes and finally the whole curve (or surface). A possible approach is to develop bootstrap bandwidth selection procedures in this non-standard setup. These procedures can then also serve as ingredient for testing whether a nonparametric function is discontinuous or not. Also wavelet-based estimation methods seem to be quite promising (see Antoniadis, Gijbels and MacGibbon, 2000).
(3) *Modeling of heterogeneous regularities in time series and analysis of such time series*. Financial data often show a spatially varying degree of regularity and an approach is to model such data via (nonparametric, i.e. nonlinear) AR and (G)ARCH models. In these models for the conditional heteroscedastic variance of log-returns, a common feature is the occurrence of breakpoints in trend or volatility. Wavelet-based estimators that capture both these breakpoints and the smooth parts of the underlying function have been proposed. So-called second-generation wavelet schemes are needed here as the design of the data is no more equidistant and fixed. Construction of smoother wavelets than those of Delouille, Franke and von Sachs (2000) needs to be addressed. Again, in this time series context, bootstrap procedures are very useful, also for the more general problem of statistical inference (construction of confidence regions or hypothesis tests) (see Ombao, von Sachs and Guo, in press-b).

Cross-links with other work packages
- *with WP2:* Nonparametric estimation of functionals showing different degrees of smoothness is needed also in the context of factor analysis of panels of time series. Amato, Antoniadis and Grégoire (2001) have introduced a nonparametric method of Independent Component Analysis (ICA) for analyzing signals and images based on the search of independent components which are then estimated non-parametrically. Here we intend to extend the classical ANOVA to a functional setting for the analysis of variance of a set of curves. In the context of the dynamic factor approach of WP2 (Forni, Hallin, Lippi, and Reichlin, 2000) the parametric modeling of the 'idiosyncratic' components by linear combinations of the unobserved multi-dimensional noise is less satisfactory and hence needs to be replaced by nonparametric functional modeling, using in particular wavelet methods.
- *with WP3:* We will exchange common experience with WP3 regarding resampling techniques in the non-standard set-ups of censored/truncated data.

Methodological problems
The work package will concentrate mainly on two methodological problems:
- When dealing with complex structures such as those mentioned above, the application of resampling methods for inferential purpose can be rather non-standard, and the issue of consistency of a bootstrap procedure needs special attention.
- Adaptive estimation techniques will be used to attain optimal estimation directly from the data without making use of knowledge of the degree of smoothness of the unknown function (see, e.g., Lepskii and Spokoiny, 1997). The challenge here is to develop such adaptive techniques for our more complex situations, i.e. the situation of curves for which the degree of smoothness (unspecified and unknown) is different in different (unknown) regions of the domain. For the development of adaptive methods for time-varying spectral densities we aim to collaborate with workpackage 2. It will be necessary and fruitful to compare the adaptive approach with its complementary counterpart of using prior information in the Bayesian analysis of our partners from WP4, WP5 and WP6.

WP2: Time series
Substantive applications/problems
The issues we concentrate on are nonstationarity and high dimensionality in time series. Nonstationary time series data, in particular in higher dimensions, are found in a number of applied fields. In this project we will address the modeling and analysis of multivariate financial data (such as stock indices and individual returns), macroeconomic data, environmental data, multichannel EEG-curves in a large clinical study on epileptic patients, and other data from clinical monitoring.

Models and primary objectives
As explained in form D1, the *aim* of this work package is to deal with two major sources of complexity in the analysis of multivariate time series: (a) nonstationarity, and (b) high-dimensional data. Therefore we need (a) applicable approaches to nonstationarity, and (b) adequate methods of dimension reduction for complex time series data. In the past, these two issues have always been considered separately; combining them is highly desirable, though far from trivial. This work package has two *primary objectives*:
(1) *The development of approaches to nonstationarity in combination with dimension reduction.* We aim at extending SLEX (Smoothed Localized complex EXponentials) from the bivariate case to "moderately" higher dimensions, following the lines of Principal Component Analysis. Combined with a dimensionality reduction methodology, this hopefully will pave the way towards the treatment of much larger numbers of series. Furthermore, in order to improve upon our tree-structured segmentation algorithm (Ombao et al, in press-c), we aim to construct adaptive methods providing uniformly best segmentations into piecewise jointly stationary stretches; this will be done along the lines of the adaptive estimation techniques by Lepskii and Spokoiny (1997). An extension to other localized functional bases, such as wavelets (see also Nason et al, 2000), will be another main avenue of research.
(2) *The improvement of the dynamic factor model methodology*. One of the improvements is related to forecasting. The estimates in Forni et al (2000) indeed are based on bilateral filters, which are well adapted to the construction of (leading or coincident) indexes (see Forni et al., in press), but behave more poorly in forecasting problems. Other problems to be solved in this context are connected with seasonality problems, irregularly spaced data, unequal series lengths, and missing observations. Some of these will be solved by embedding the discrete time processes under study into a continuous time process framework. Finally, the basic models used so far are inherently linear. Also non-linear generalizations (e.g., using wavelet transforms), will be considered.

Cross-links with other work packages
- *with WP1:* Functional estimation is highly relevant for WP2. We will work on the elaboration of optimal local smoothing techniques of the SLEX periodograms, as the technique suggested in Ombao et al. (1999) for the development of a yet only global smoothing parameter has close connections to (generalized) cross-validation in general function estimation. Further we share with WP1 the implementation of adaptive segmentation methods; also wavelet extensions will heavily rely on collaboration with WP1.
- *with WP4*: Serial dependency of observations can be an important issue in longitudinal data, leading to mixed time series models.
- *with WP5*: The tree-structured segmentation algorithms to be used in connection with SLEX are similar to those being used in WP5 for the classification of entities. Different variants of tree-based models and more flexible models derived from these will be studied as well.

- *with WP6*: The forecasting issues in the dynamic factor approach are related with similar problems arising in WP6 (causal inference, missingness). The problem of missing observations is still a crucial unsolved problem for this approach.

Methodological problems
Methodological problems in this work package are connected with evaluating the performance of our estimators, either via asymptotic considerations or, as for inference, by bootstrap methods. To this aim, developments along the results by Ombao, Raz et al. (in press-a) on a powerful model for non-stationary time series offer the possibility of statistical inference, based on bootstrapping and other resampling techniques. The difficulties of nonstandard double-index asymptotics in a panel data context stem from the fact that the theory of multivariate time series is entirely built on a fixed-dimension basis, where asymptotics are taken as the length of the observation period tends to infinity. In our situation, both the cross-section dimension and the number of observations are tending to infinity, and this requires careful mathematical treatment. In the context of SLEX-like models, the handling of local stationarity, and the notions of consistency, efficiency, and asymptotic inference are associated with an increasingly finer partition of a fixed-length time interval, which also requires a more delicate treatment than standard asymptotics.

WP 3: Survival analysis
Substantive applications/problems
Interval-censored data occur, for example, in AIDS research when the time until the development of AIDS for HIV-patients is only known to lie in between two examination times. Discontinuities in the regression or hazard function when the response is subject to censoring occur, for example, when an operation or another medical treatment changes the risk function of the individual. Heterogeneity may be of relevance when different medical centers taking part in a clinical trial are a source of differences in the effects.

Models and primary objectives
As explained in form D1, the *aim* of this work package is to make inferences on an object function (a regression or hazard function), while taking into account complex censoring mechanisms, discontinuities, and heterogeneity. This work package has two *primary objectives*:
(1) *Nonparametric estimation with interval censored data and/or with abrupt changes*. (a) First, regarding the former, the domain is still in full development and in comparison with the area of right censored data, relatively few results have been obtained so far. Groeneboom and Wellner (1992) propose a nonparametric estimator of the distribution function under interval censoring and establish the weak convergence of this estimator. We would like to consider in a first stage the estimation of the conditional distribution of the response that is subject to interval censoring, given a random covariate. This situation is more complex, as it requires smoothing over the covariate space. In order to obtain asymptotic results for the estimator of the conditional distribution, we will make use of the aforementioned book of Groeneboom and Wellner (1992), as well as of Van Keilegom and Veraverbeke (1997), where the asymptotic properties of an estimator of the conditional distribution function are studied when the response is subject to right censoring. In a second stage, other functions of interest, like the conditional quantile, regression or hazard function will be considered. Bootstrap methods will be used to obtain alternative approximations for the normal limiting distributions of the estimators considered. (b) Second, the regression and hazard functions might show abrupt changes, and this creates difficult estimation problems, especially when the data are censored. Here the use of local approximation methods and wavelet-decomposition methods seem to be among the most promising ones.
(2) *Extending and improving inference for frailty models*. The domain of frailty models as models of heterogeneity has developed a lot during the last decade, but a number of interesting problems remain unexplored. In particular, we want to develop statistical inference for frailty models that include interaction effects. Another problem we would like to address in this context is the development of appropriate maximum likelihood theory for the variance component(s) of the frailty model (MLE boundary problems). We will also use sensitivity analysis ideas within the context of frailty models.

Cross-links with other work packages
- *with WP1*: See cross-links mentioned in the description of WP1.

- *with WP4*: There are close links with WP4 on mixed models. Both frailty models and mixed models capture heterogeneity, frailty models for incomplete (censored) data and mixed models for complete data.
- *with WP6:* Issues of causal inference and missing data often appear in the survival analysis context.

Methodological problems
Two methodological problems will be studied:
- We will use sensitivity analysis ideas within the context of frailty models.
- For frailty models as well as for mixed regression models we are interested in the use of resampling techniques for inference and model checking.

WP4: Mixed models
Substantive applications/problems
We will consider two important practical applications:
(1) The Signal Tandmobiel® study collects on a yearly basis (1996-2001) very detailed and calibrated dental information on tooth and surface level from about 4500 Flemish schoolchildren with a follow-up from 7 until 12 years of age. Also questionnaire data on diet and brushing behavior were collected. It is the purpose to examine the duration until caries development on surface and tooth level and to look for the determinant factors for caries. Also the interrelationship between the teeth is of major theoretical and practical importance, e.g., the relationship between caries on deciduous and permanent teeth.
(2) Individual differences and classification of subjects on the basis of a series of screening tests are becoming an essential part of preventive medicine. An example is the classification of subjects as cancer or non-cancer cases using longitudinal blood measurements of prostate specific antigen. The models involved here are roughly of the same type as models used in psychometrics: latent class models, IRT models (with estimated distributions of person effects), and combinations of both (e.g., mixture LLTM).

Models and primary objectives
The two examples imply generalized linear and non-linear mixed models. As explained in form D1, the *aim* of this work package is to model the random effects, while concentrating on the selection of adequate random effects distributions, including mixture distributions. Two *primary objectives* are formulated:
(1) *The implementation of multivariate random effects* with a highly structured and spatial covariance matrix embedded in a survival structure. Hence, random effects will be incorporated at various levels (tooth-, individual-, time- and school level). Classically, the implementation of these random effects involves parametric assumptions. Since the estimation of fixed effects depends on the distributional aspects of the random effects and can be distorted when they are violated, more relaxed assumptions are welcome (Davidian & Gallant, 1993; Shen & Louis, 1999). Indeed, one could assume that the random effects distribution is at least continuous (excluding the non-parametric maximum likelihood solution). There exist various procedures that deal with smoothing, like kernel density estimation, cubic- and B-splines. The latter is mathematically and numerically very tractable. First, univariate distributions are aimed at, later multivariate extensions will be explored. A combination of a longitudinal data model with a survival model will be necessary since we aim additionally at examining the degree of caries and to investigate repeated events.
(2) *The investigation of mixture models as an alternative for approaching the random effects distribution.* A finite mixture of classical distributions can approach arbitrarily close any distribution for the random effects. This approach is therefore a competitor to the above approach and could be more tractable numerically. Furthermore, formulating a mixture distribution for the random effects can be a basis for a classification of subjects (Verbeke & Lesaffre, 1996).

Cross-links with other work packages
- *with WP2:* See cross-links mentioned in the description of WP2.
- *with WP3:* See cross-links mentioned in the description of WP3 ; in addition WP4 will also rely on survival models as studied in WP3.
- *with WP5:* Given that some of the mixture models WP5 will concentrate on mixture extensions of nonlinear mixed models (from the IRT family), and given the binary or categorical type of part of the data in WP4, collaboration with WP5 seems quite natural. For example, the mixture LLTM allows for detecting groups with different patterns of factors affecting caries.
- *with WP6:* Random effects can be considered latent variables, so that the framework that is set up in WP6 is of direct relevance. Furthermore, we share the plan to set up our problems in a Bayesian

framework, so that the obtained results have mutual relevance, and so that sensitivity issues can be studied along the same lines.

Methodological problems
Three methodological problems can be seen for this work package:
- General problems concerning numerical convergence of the estimation procedures are of importance for several of the procedures we plan to use.
- Sensitivity of the estimations for assumptions on the random effects distribution.
- Detection of influential observations will be important for the assessment and interpretation of results.

WP 5: Classification and mixture models
Substantive applications/problems
In various substantive contexts the experimental units or objects under study stem from a heterogeneous population; questions then arise as to how to capture this heterogeneity and what its exact nature is (e.g., whether differences between objects are primarily quantitative or qualitative in nature). Applications we want to focus on in this work package include:
(1) the study of individual differences in psychology (in particular, the study of individual differences in personality-related behavioral profiles across situations, which immediately links up with the study of person x situation interactions),
(2) differences in the effect of educational and school-related variables (as in educational assessment), and the study of heterogeneity of patients in biomedical applications (in particular in case longitudinal data on these patients are available).
Also applications to image analysis, and heterogeneity in micro-arrays are possible if Poisson mixtures are used with spatial structures.

Models and primary objectives
As explained in form D1, the *aim* of this work package is to capture heterogeneity in a population and to find out what its exact nature is. Three *primary objectives* are formulated.
(1) *Studying specific types of mixture models*. First, we are interested in finite normal mixtures for random effects in linear and nonlinear mixed models. Linear mixed models apply both in psychological and biomedical applications, and nonlinear mixed models are very popular within psychometrics (IRT models, e.g., the Rasch model). Further, special attention will be given to:
- mixture Rasch models (Rasch scale within each mixture component) (Rost, 1997); related to these models we want to investigate to which extent discrete and continuous (multidimensional) representations are exchangeable;
- mixture models for items and situations (Janssen, Tuerlinckx, Meulders, & De Boeck, in press) in combination with mixtures for person effects;
- an extension of PMD models (mixture models with multiple binary latent variables and restrictions on the conditional and marginal probabilities) (Meulders, De Boeck, Van Mechelen, Gelman, & Maris, in press), so that person main effects and person by situation interactions can be fully captured.
(2) *Investigating methods to decide on the number and type of components*. Smoothing techniques, such as kernel estimation will be investigated, as well as approaching the number of components as a parameter within a Bayesian set up. Further, alternatives within the mixture context will be studied on their quality as approximations: models with only one component, models with point mass components (homogeneity within components). For example, the sensitivity of the inferences on the fixed effects will be tested for these alternatives.
(3) *Classification techniques other than mixtures will be studied* on their performance as an alternative for mixtures. Typical examples are k-means clustering, Kohonen networks, tree-based classification and extensions of these. For binary data, we will also investigate the potential of different variants of the latent binary decomposition models we have developed (Leenen, Van Mechelen, De Boeck, & Rosenberg, 1999) for the same purpose. Furthermore, the classification models that will be considered will be studied also for their extension to complex (e.g., symbolic) data, and, if appropriate given the kind of method, also to regression clustering (e.g., for time series) and to non-classical clustering criteria (involving convexity).

Cross-links with other work packages
- *with WP2:* Part of WP5 is the study of tree-based classification methods for capturing heterogeneity. A collaboration with WP2 will be set up for applications of tree-based methods to time series.
- *with WP4:* See cross-links mentioned in the description of WP4.
- *with WP6:* For estimation and model checking, the latent variables from IRT models, and mixture component membership can be considered unobserved data, so that results from WP6 are of direct relevance. Also the Bayesian framework to deal with these problems is shared with WP6.

Methodological problems
The following four problems will have to be dealt with:
-Problems related to the sample size and sampling design needed for detection of heterogeneity and model estimation will be studied.
- Identifiability problems need to be tackled (e.g., due to the possibility of label switching in mixtures with components belonging to the same parametric family). We plan to continue work on the identifiability of latent class models (San Martin & De Boeck, 2000).
-Tools for inference on the number of components will be studied and further developed.
-The choice of priors and sensitivity in Bayesian estimation deserves further study.

WP 6: Incompleteness and latent variables
Substantive applications/problems
The best known incomplete data problem occurs when a well designed study fails to record all outcomes it meant to analyze. This happens not only naturally in longitudinal studies where individuals drop out over time, but also in other settings where non-intentionally incomplete data complicate the analysis. Indeed, the lack of stochastic control over what is missing generates observed samples that no longer are representative of the original study population. Explicit statistical modeling of the missingness and its association with the variables of interest becomes necessary to avoid biased inference. Similar problems occur when the variables of interest are intrinsically not directly observable, like a personality trait (latent variable) in psychometrics or a causal effect in medicine, economics, or educational assessment. Also random effects are usefully cast within this framework. One has to address modeling of the unobserved features in relation to the measured variables before valid inferences can be drawn. The importance of these widespread data-analytic problems has become well recognized over the last decade.

Models and primary objectives
As explained in form D1, the *aim* of this work package is the development of parametric and semiparametric models for incomplete longitudinal and latent data, and the study of sensitivity to assumptions made in these models (Robins, Rotnitzky & Zhao, 1995; Robins, Rotnitzky & Sharfstein, 1998). Assessing the impact of missing data on subsequent statistical inference is an important but difficult question. Conditions can be formulated under which an analysis that proceeds as if the missing data are missing by design (i.e., ignoring the missing value process), can provide valid answers to study questions. The difficulty in practice is that such conditions can rarely be assumed to hold. A fundamental point is: when we undertake such analyses, assumptions will be required that cannot be assessed from the data under analysis. Hence, in this setting, there cannot be anything that could be termed a final analysis, and arguably the appropriate statistical framework is one of sensitivity analysis. When proposing a model to fit to the observed data, one shall ideally respect the fundamental structure of the data generating mechanism and incorporate what is known from the subject matter domain. This is a challenge in its own right. On the other hand, a balance must be achieved between a model realistic enough to reveal structural pathways and parsimonious enough to be identifiable with good precision from the observed data.

Against this background, we have the following *primary objectives*, each of them related to an open problem. We aim to further work on sensitivity analysis in several areas:
(1) We will work on sensitivity for assumptions about the *missingness mechanism* in parametric and semi-parametric models for incomplete longitudinal data, within and outside a Bayesian paradigm. The selection of models for measurements with nonrandom dropout can be made using a sensitivity approach (Kenward, 1998; Verbeke et al. 2001).
(2) Sensitivity analysis tools will be developed for *latent variable and mixed-effects* models in the psychometric and biomedical areas. The sensitivity analysis will relate to distributional assumptions and number of components, among other things.

(3) Sensitivity analysis questions *in causal inference* will be addressed. Recently, a new and general approach was developed for sensitivity analysis that distinguishes between ignorance and imprecision via overparameterized likelihoods. Another line of work has built further on properties of misspecified (log)likelihoods. Both will be developed further in the causal inference setting.

Cross-links with other work packages
- with WP4: See cross-links mentioned in the description of WP4
- with WP5: See cross-links mentioned in the description of WP5; especially objective 2 of WP6 is a subject of collaboration with WP5.

Methodological problems
The methodological problems are directly related to the focus of the work package on sensitivity. In the context of the proposal, sensitivity is considered a methodological issue that is of relevance for fully parametric and semi-parametric models. Furthermore, Bayesian sensitivity analysis tools for incomplete data will be studied and further developed.

Partners contributing to the work packages
For each work package two or three groups will collaborate on the realization of the objectives. One of the contributing groups takes the coordinating role. Table 1 shows the distribution of partner groups over the 6 work packages, with in each case the partner groups indicated in bold having the coordinating role.

Work package	Section	Contributing partners
WP1 Functional estimation	I	**UCL*** , ULB, UJF
WP2 Time series	I	**ULB**, UCL
WP3 Survival analysis	I	**LUC**, UCL
WP4 Mixed models	I	**KUL-2**, KUL-1, LUC
WP5 Classification and mixture models	II	**KUL-1**, KUL-2, RWTH
WP6 Incompleteness and latent variables	II	**LUC**, KUL-1, KUL-2

Table 1

Note: (*) The UCL group is the principal partner, all others are associated, corresponding or European partners.

Apart from the direct contributions as indicated in Table 1, other contributions will stem from groups with cross-links mentioned in each work package.

2. Interaction between sections
For all kinds of models it makes sense to think of two kinds of problems: (1) heterogeneity of the parameter values and/or the experimental units under study, and (2) how to deal with incomplete data. The work packages of section II will interact with those of section I on these problems:
(1) It is an explicit aim of WP5 to collaborate with WP3 and WP4 on mixture extensions of mixed models, and with WP1 to WP4 on classification in general. One specific example was given earlier and is related to tree-based methods for the segmentation of time series. More in general, the possibilities of a broader range of classification methods will be investigated for al WPs from section I.
(2) In a similar vein, it is the explicit aim of WP6 to collaborate with the other work packages on the issue of incomplete data. For example, longitudinal data may run over different lengths of time; data may be interval-censored, may be incomplete because of lack of compliance, or may be lacking for other reasons. Two very important problems are the modeling of missingness mechanisms and the testing of how sensitive conclusions are for the missingness mechanisms one assumes. Latent data (i.e., data that are by definition not observable) constitute a second type of incomplete data; as an example one may think of random person effects. For model estimation and testing, common approaches can be used to deal with both missing and latent data. Especially a Bayesian framework is very useful to formulate estimation and testing methods. Part of the work in WP6 will be spent on this.

3. Methodological issues
Some methodological issues will be studied in collaboration between work packages, while others will primarily be studied within a single work package. Issues that are most appropriate to do collaborative work on are: smoothing techniques, resampling/Monte Carlo techniques, and sensitivity. Smoothing and resampling/Monte Carlo techniques have a double role in the project: as tools one has to use, and as topics of research (about which we aim at generic results). In the description that follows, we will

concentrate on their role as topics of research (for the techniques as tools, see the plan for workshops in form H).
(1) *Smoothing* is an issue in several work packages. In general, smoothing techniques will be used in complex set-ups, which is a serious challenge for the project. The specific objectives aimed at include:
- Splines and kernel methods will be investigated on their performance, e.g. as alternatives for mixtures for the estimation of densities, and work is planned on extensions to the multivariate case.
- A useful nonlinear smoothing technique is the wavelet transform. Deleting and shrinking wavelet coefficients can be used for filtering and for approaching nonlinear relations. It will be investigated what the best options are in the context of work packages where this technique is of use, and how this approach compares with local polynomial approximation methods.
(2) *Resampling methods and Monte Carlo methods* will be used in several work packages, first, as a tool in curve estimation and in approaching discontinuity (e.g., bootstrap bandwidth selection), and second, as an inferential tool more in general. The objectives aimed at include:
- Specific evaluation of parameter estimates and model specifications will be taken care of using bootstrap methods and a Bayesian approach. Points of interest are the quality of the results in comparison with asymptotic results, and a frequentist evaluation of these tools.
- Both bootstrap methods and a Bayesian approach will be used in the project especially for non-standard estimation and testing, and an evident issue is what the qualities of these approaches are in comparison with one another.
(3) *Sensitivity* to model assumptions and misspecifications. This objective is at the heart of WP6, so that it is evident that work in this regard will be realized in that work package. Also in other work packages sensitivity is a topic of interest, for example, regarding the kind of distributions one wants to assume.

4. Time schedule
The network project has a large number of (to be sure, heavily intertwined) objectives. Note, however, that the partners of the network will invest also other means than those applied for with this proposal.

Regarding time, we want to be flexible (for example, spending more time on a problem if this is required to find a high-quality solution, or pursuing valuable unexpected insights). Nevertheless, for the main lines of the project we propose the following time schedule (graphically represented in Figure 2):

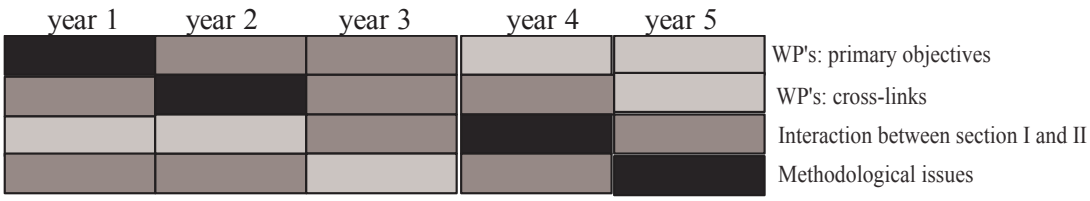

Figure 2: Bar chart representation of time schedule for work on the major lines of the proposal

Four degrees of activity are distinguished in terms of different levels of shading: white for no activity, light gray for regular activity, dark gray for intensified activity, and black for highly intensive activity. Note that white is never used, meaning that permanent attention will be given to all four kinds of activities. Two major periods are distinguished: the first three years and the last two years.
(1) During the first three years we will concentrate on the specifics of the different work packages, that is, the primary objectives as well as the cross-links. Besides, common methodological tools obviously also will need a lot of attention during this period, for example to fully explore their possibilities (e.g., workshops are planned about this, see form H).
(2) During the last two years the interaction between the two sections will be more important, in line with the concept of the project and its two levels. Furthermore, intensive work on methodological issues is planned for this period: Evaluative methodological research topics will be especially focused on (which will typically require information from earlier obtained results); methodological issues will finally also be of key importance for the kind of conclusions we will be able to draw in the distinct work packages.

References

Amato, U., Antoniadis, A. and Grégoire, G. (2001). Independent component nonparametric discriminant analysis. *Discussion Paper*. University of Grenoble.

Antoniadis, A., Gijbels, I. and MacGibbon, B. (2000). Nonparametric estimation for the location of a change-point in an otherwise smooth hazard function under random censoring. *Scandinavian Journal of Statistics,* 27*,* 501-519.

Berkhof, J., Van Mechelen, I. and Hoijtink, H. (2000). Posterior predictive checks: principles and discussion. *Computational Statistics,* 15, 337-354.

Bock, H. H. (1996). Probability models and hypotheses testing in partitioning cluster analysis. In P. Arabie et al. (Eds.). *Clustering and Classification* (pp. 377-453). River Edge, NJ: World Scientific.

Bock, H.H. (1999). Clustering and neural network approaches. In W. Gaul and H. Locarek-Junge (Eds.). *Classification in the Information Age* (pp. 42-57). Heidelberg: Springer.

Bock, H.H. (2001). *Convexity-based clustering criteria: a new approach.* Manuscript of a seminar given at the Academy of Economics, Krakow, Poland, October 2000.

Bozdogan, H. (1994). Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity. In H. Bozdogan et al. (Eds.). *Proceedings First US/Japan Conference on the Frontiers of Statistical Modeling: an Informational Approach* (pp. 69-113). Dordrecht: Kluwer.

Brillinger, D. (1981). *Time Series: Data Analysis and Theory*. New York: McGraw-Hill.

Butter, R., De Boeck, P. and Verhelst, N. (1998). An item response model with internal restrictions on item difficulties. *Psychometrika*, 63, 47-63.

Buyse, M. and Molenberghs, G. (1998). Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics*, 54, 186-201.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D. and Geys, H. (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*, 1, 1-19.

Chen, Hanfeng, Chen, Jiahua and Kalbfleisch, J.D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society B,* 63*,* 19-29.

Dahlhaus, R. (1996). Asymptotical statistical inference for nonstationary processes with evolutionary spectra. In P.M. Robinson and M. Rosenblatt (Eds.). *Athens Conference on Applied Probability and Time Series Analysis* 2. New York: Springer.

Dahlhaus, R. (2000). A likelihood approximation for locally stationary processes. *Annals of Statistics*, 28, 6, 1762 - 1794.

Datta, S., Satten, G.A. and Williamson, J.M. (2000). Consistency and asymptotic normality of estimators in a proportional hazards model with interval censoring and left truncation. *Annals of the Institute of Statistical Mathematics,* 52, 160-172.

Davidian, M. and Gallant, A.R. (1993). The nonlinear mixed effects model with a smooth random effects density. *Biometrika*, 80, 475-488.

Delaigle, A. and Gijbels, I. (2001). Bootstrap bandwidth selection in kernel density estimation from a contaminated sample. *Discussion Paper* 0116. Institut de Statistique, UCL, Louvain-la-Neuve.

Delouille, V., Franke J., and von Sachs, R. (2001). Nonparametric stochastic regression with design-adapted wavelets. *Sankhya,* 63, Ser. A, 328-366.

Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *Annals of Statistics*, 7, 269-281.

Diggle, P.J. and Kenward, M.G. (1994). Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics*, 43, 49-94.

Dolan, C.V., & van der Maas, H.L.J. (1998). Fitting multivariate normal finite mixtures subject to structural equation modeling. *Psychometrika,* 63*,* 227-253.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2000). The generalized dynamic factor model: identification and estimation. *Review of Economics and Statistics,* 82, 540-554.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2001). *The generalized dynamic factor model: consistency and rates*. Preprint, ISRO, ULB, Brussels.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (in press). Coincident and leading indicators for the Euro area. *The Economic Journal*.

Fraley, Chr. and Raftery, A.E. (1998). How many clusters? Which clustering method? - Answers via model-based cluster analysis. *Computer Journal,* 41, 578-588.

Gijbels, I. and Goderniaux, A.-C. (2000). Bandwidth selection for change point estimation in nonparametric regression. *Discussion Paper* 0024. Institut de Statistique, UCL, Louvain-la-Neuve.

Gijbels, I. and Gürler, Ü. (2001). Estimation in change point models for hazard function with censored data. *Discussion Paper* 0114. Institut de Statistique, UCL, Louvain-la-Neuve.

Gijbels, I., Mammen, E., Park, B. and Simar, L. (1999). On estimation of monotone and concave frontier functions. *Journal of the American Statistical Association*, 94, 220-228.

Groeneboom, P. and Wellner, J.A. (1992). *Information bounds and nonparametric maximum likelihood estimation*. Basel: Birkhäuser Verlag.

Hall, P. and Rau, C. (2000). Tracking a smooth fault line in a response surface. *Annals of Statistics*, 28, 713-733.

Hall, P. and Simar, L. (2000). Estimating a changepoint, boundary or frontier in the presence of observation error. *Discussion Paper* 0012. Institut de Statistique, UCL, Louvain-la-Neuve.

Hogan, J.W. and Laird, N.M. (1997a). Mixture models for joint distribution of repeated measures and event times. *Statistics in Medicine*, 16, 239-257.

Hogan, J.W. and Laird, N.M. (1997b). Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistics in Medicine*, 16, 259-272.

Janssen, R., Tuerlinckx, F., Meulders, M. and De Boeck, P. (2000). A hierarchical IRT model for standard setting. *Journal of Educational and Behavioral Statistics,* 25, 285-306.

Kenward, M.G. (1998). Selection models for repeated measurements with nonrandom dropout: an illustration of sensitivity. *Statistics in Medicine*, 17, 2723-2732.

Kenward, M.G. and Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science*, 13, 236-247.

Leenen, I., Van Mechelen, I. and De Boeck, P. (1999). A generic disjunctive/ conjunctive decomposition model for *n*-ary relations. *Journal of Mathematical Psychology*, 43, 102-122.

Leenen, I., Van Mechelen, I., De Boeck, P. and Rosenberg, S. (1999). INDCLAS: Individual differences hierarchical classes analysis. *Psychometrika*, 64, 9-24.

Lenk, P.J. and DeSarbo, W.S. (2000). Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika*, 65, 93-119.

Lepskii, O. and Spokoiny, V. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *Annals of Statistics,* 25, 2512-2546.

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

Lipsitz, S.R., Fitzmaurice, G.M., Molenberghs, G. and Zhao, L.P. (1997). Quantile regression methods for longitudinal data with drop-outs. *Applied Statistics*, 46, 463-476.

Little, R.J.A. (1994). A class of pattern-mixture models for multivariate incomplete data. *Biometrika,* 81, 471-483.

Little, R.J.A. (1995). Modeling the dropout mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90, 1112-1121.

Little, R.J.A. and Rubin, D.B. (1986). *Statistical Analysis with Missing Data.* Chichester: Wiley and Sons.

Little, R.J.A. and Wang Y. (1996). Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics*, 52, 98-111.

Luetkepohl, H. (1993). *Introduction to Multiple Time Series.* New York: Springer-Verlag.

Meulders, M., De Boeck, P. and Van Mechelen, I. (2001). Probability matrix decomposition and main-effects generalized linear models for the analysis of replicated binary associations. *Computational Statistics and Data Analysi,*38, 217-233..

Meulders, M., De Boeck, P., Van Mechelen, I., Gelman, A. and Maris, E. (2001). Bayesian inference with PMD models. *Journal of Educational and Behavioral Statistics,* 26, 153-179.

Michiels, B., Molenberghs, G. and Lipsitz, S.R. (1999). Selection models and pattern-mixture models for incomplete categorical data with covariates. *Biometrics*, 55, 978-983.

Molenberghs, G., Kenward, M. G. and Lesaffre, E. (1997). The analysis of longitudinal ordinal data with non-random dropout. *Biometrika*, 84, 33-44.

Müller, H.-G. and Stadtmüller, U. (1999). Discontinuous versus smooth regression. *Annals of Statistics*, 27, 299-337.

Nason, G., von Sachs, R. and Kroisandt, G. (2000). Wavelet processes and adaptive estimation of evolutionary wavelet spectra. *Journal of the Royal Statistical Society B,* 62, 271-292.

Nielsen, G., Gill, R., Andersen, P. and Sørensen, T. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics*, 19, 25-43.

Ombao, H., Raz, J., von Sachs, R. and Guo, W. (2002). The SLEX model of a non-stationary random process. *Annals of the Institute of Statistical Mathematic*, 54, 1, 171-200.

Ombao, H., Raz, J., Strawderman, R. and von Sachs, R. (2001). A simple GCV method of span selection for periodogram smoothing. *Biometrika*, 88, 1186-1192.

Ombao, H., von Sachs, R. and Guo, W. (2000). Estimation and inference for time-varying spectra of locally stationary SLEX processes. *Proceedings of the 2nd International Symposium on Frontiers of Time Series Modeling*. Nara, Japan, December 14-17, 2000.

Ombao, H., Raz, J., von Sachs , R. and Mallow, B. (2001). Automatical statistical analysis of bivariate nonstationary time series. *Journal of the American Statistical Association,* 96, 543-560.

Park, B., Simar, L. and Weiner, Ch. (2000). The FDH estimator for productivity efficiency scores: asymptotic properties. *Econometric Theory*, 16, 855-877.

Parner, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *Annals of Statistics*, 26, 183-214.

Rabinowitz, D., Tsiatis, A. and Aragon, J. (1995). Regression with interval-censored data. *Biometrika*, 82, 501-513.

Richardson, S. and Green, P.J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society B*, 59, 731-792.

Richardson, S. and Green, P.J. (1998). Corrigendum: On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society B*, 60, 661.

Robins, J.M., Rotnitzky A. and Zhao, L.P. (1995a). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90, 106-121.

Robins, J.M. and Rotnitzky, A. (1995b). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90, 122-129.

Robins, J.M., Rotnitzky, A. and Scharfstein, D.O. (1998). Semiparametric regression for repeated outcomes with non-ignorable non-response. *Journal of the American Statistical Association*, 93, 1321-1339.

Rost, J. (1997). Logistic mixture models. In W. van der Linden and R. Hambleton (Eds.). *Handbook of Modern Item Response Theory* (pp. 449-463). New York: Springer.

San Martin, E. and De Boeck, P. (2000). *Specification problems in latent class models.* Manuscript submitted for publication.

Shen, W. and Louis, T.A. (1999). Empirical Bayes estimation via the smoothing by roughening approach. *Journal of Computational and Graphical Statistics*, 8, 800-823.

Therneau, T.M. and Grambsch, P.M. (2000). *Modeling Survival Data. Extending the Cox Model*. New York: Springer-Verlag.

Tuerlinckx, F. and De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters. *Psychological Methods*, 6, 181-195.

Van Keilegom, I. and Veraverbeke, N. (1997). Estimation and bootstrap with censored data in fixed design nonparametric regression. *Annals of the Institute of Statistical Mathematics,* 49, 467-491.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91, 217-221.

Verbeke, G., Lesaffre, E. and Brant, L.J. (1998). The detection of residual serial correlation in linear mixed models. *Statistics in Medecine*, 17, 1391-1402.

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

Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E. and Kenward, M.G. (2001). Sensitivity analysis for non-random dropout: a local influence approach. *Biometrics*, 57, 7-14.

von Sachs, R. and MacGibbon, B. (2000). Non-parametric curve estimation by wavelet thresholding with locally stationary errors. *Journal of Scandinavian Statistics,* 27, 475-499.

Wedel, M. and Bijmolt, T.H.A. (2000). Mixed tree and spacial representations of dissimilarity judgments. *Journal of Classification,* 17*,* 243-271.

Yung, Yiu-Fai. (1997). Finite mixtures in confirmatory factor-analysis models. *Psychometrika,* 62*,* 297-330.

Zhu, Hong-Tu and Lee, Sik-Yum. (2001). A Bayesian analysis of finite mixtures in the LISREL model. *Psychometrika,* 66*,* 133-152.

I.6. SYNOPTIC LIST OF WORK PACKAGES

Record the title of each work package (as described in I.5) and the list of partners involved in its production.

---

1. Work package title: Functional estimation (WP1)
   Partners involved (promoter, institution) : L. Simar, UCL ; M. Hallin, ULB ; A. Antonadis, UJF-LMC-IMAG

2. Work package title: Time series (WP2)
   Partners involved (promoter, institution) : M. Hallin, ULB ; L. Simar, UCL

3. Work package title: Survival analysis (WP3)
   Partners involved (promoter, institution) : N. Veraverbeke, LUC ; L. Simar, UCL

4. Work package title: Mixed models (WP4)
   Partners involved (promoter, institution) : E. Lesaffre, KUL-2 ; N. Veraverbeke, LUC ; P. De Boeck, KUL-1

5. Work package title: Classification and mixture models (WP5)
   Partners involved (promoter, institution) : P. De Boeck, KUL-1 ; E. Lesaffre, KUL-2 ; H.-H. Bock, RWTH-Aachen

6. Work package title: Incompleteness and latent variables (WP6)
   Partners involved (promoter, institution) : N. Veraverbeke, LUC ; P. De Boeck, KUL-1 ; E. Lesaffre, KUL-2

7. Work package title:
   Partners involved (promoter, institution) :

8. Work package title:
   Partners involved (promoter, institution) :

9. Work package title:
   Partners involved (promoter, institution) :

10. Work package title:
    Partners involved (promoter, institution) :

11. Work package title:
    Partners involved (promoter, institution) :

12. Work package title:
    Partners involved (promoter, institution) :

13. Work package title:
    Partners involved (promoter, institution) :

14. Work package title:
    Partners involved (promoter, institution) :

15. Work package title:
    Partners involved (promoter, institution) :

I.7. MAIN SKILLS OF THE PARTNERS

Give the list of partners and record the main skill of each of them in relation to the project.

---

1. Partner (promoter, institution): L. Simar, UCL
   Main skill: Frontier estimation, non- and semiparametric methods for change-point detection, nonparametric density estimation based on contaminated data, resampling-techniques, time series analysis

2. Partner (promoter, institution): P. De Boeck, KUL-1
   Main skill: Bayesian data analysis, Boolean models for classification, latent class models, psychometric models (IRT, nonlinear mixed models)

3. Partner (promoter, institution): E. Lesaffre, KUL-2
   Main skill: Bayesian statistics, correlated data, longitudinal data, missing data, random effects distributions, repeated measures

4. Partner (promoter, institution): N. Veraverbeke, LUC
   Main skill: Causal inference, incomplete data, latent structures, longitudinal data, non-and semiparametric inference for incomplete data, resampling techniques

5. Partner (promoter, institution): M. Hallin, ULB
   Main skill: Asymptotic inference, rank-based and distribution-free methods, semiparametric inference

6. Partner (promoter, institution): H.-H. Bock, RWTH-Aachen
   Main skill: Classification and clustering, exploratory data analysis, Kohonen networks, mathematical finance and financial time series, probability theory and mathematical statistics

7. Partner (promoter, institution): A. Antoniadis, UJF-LMC-IMAG
   Main skill: Biostatistics, detection of discontinuities, multivariate analysis, non-and semiparametric inference, survival analysis models, wavelet-based estimation

8. Partner (promoter, institution): ,
   Main skill:

9. Partner (promoter, institution): ,
   Main skill:

10. Partner (promoter, institution): ,
    Main skill:

11. Partner (promoter, institution): ,
    Main skill:

12. Partner (promoter, institution): ,
    Main skill:

13. Partner (promoter, institution): ,
    Main skill:

14. Partner (promoter, institution): ,

Main skill:

15. Partner (promoter, institution): ,
    Main skill:

16. Partner (promoter, institution): ,
    Main skill:

17. Partner (promoter, institution): ,
    Main skill:

18. Partner (promoter, institution): ,
    Main skill:

I.8. NETWORK ORGANISATION AND MANAGEMENT

(max. 3 pages)

Describe the network's organisation as well as the practical terms governing collaboration and interaction between the partners (taking account of the fact that joint working is one of the IAP Programme's objectives)

The coordination of the project is taken care of by the principal partner group (UCL), in close collaboration with the other partner groups. In this section of the application, we will concentrate on means to further stimulate internal relations (communication, collaboration), training, and external relations.

First, a research network should really act as a network, meaning that we will to organize and manage collaborations between the various groups of the network. Second, we will encourage and train young researchers. Third, the network will not act as a closed network but as an open network, with connections to other statistical groups within Belgium as well as internationally.

We propose to organize eight types of research and research-related activities:
- for internal relations:
(1) research by doctoral students, (2) research by post-docs, (3) workshops, (4) seminars, (5) staff meetings, and a website and newsletter (see further)
- for training: (6) intensive courses
- for external relations: (7) a website and newsletter, (8) open network activities

1. Doctoral subprojects

We will encourage and train beginning researchers. The guess is that, given the budgets applied for, we will on average be able to work with 2 to 3 doctoral students for the main partner, 1 to 2 doctoral students for an associate partner, and 1 doctoral student for a European partner. The topics of the dissertations should be closely related to the primary objectives from the work packages. For example, "Detection of discontinuities and errors-in-variables" and " Modeling and analysis of multivariate non-stationary time series" could be topics for doctoral dissertations related to WP1 and WP2, respectively. The doctoral students would typically have a promotor from a first partner group and a co-promotor from a second partner group contributing to the same work package.

2. Postdoctoral subprojects

We consider it to be an optimal scheme that there will be two types of post-docs:
(1) The first type is associated to one work package (2 or 3 years of post-doc research), with the corresponding post-doc being mainly affiliated with the coordinating partner of the work package in question, but with incentives to work for some time also with partners coordinating a work package with specific cross-links to the former.
(2) The second type does research on common methodological tools and topics, and is affiliated mainly with one partner, but works part of the time as a visiting post-doc in the departments of several other partners. For example, 3 years of post-doc research could be invested in smoothing techniques in a complex set-up, with the personnel being affiliated with the UCL group, but visiting also other partners working on the same issue, for example the ULB group; as a second example, 2 years of post-doc research could be spent to resampling methods and Monte Carlo methods, with the personnel being affiliated with the KUL-1 group, but visiting also other partners, in particular the LUC group.

3. Workshops

Workshops are planned to play a key role in the organization and planning of the project. The scheduling over time links up with the phase concept of the project (see form D1) and with the time schedule (see form E).

Phase 0

Workshop 1: beginning of year 1

The aim of the workshop is to communicate to all partners the types of applications and substantive problems that will be focused on in the distinct work packages, so that one becomes maximally familiarized with the kind of applications and data of other partners/work packages. Available data sets related to the work packages will be explained, and a selection of them will be made, in order to be jointly studied in different work packages. Often, applications seem different because of the context they stem from, whereas when looked upon from a formal point of view, they are similar.

Phase 1

Workshop 2: beginning of year 2

Meeting on two categories of statistical tools that will be used in all or almost all work packages: (1) smoothing techniques, and (2) resampling/Monte Carlo techniques. This workshop has two purposes: (1) to give an overview of the techniques that are considered to be the most promising ones for partner groups who want to use them without concentrating on them as a topic of research, and (2) to report on the first results obtained from the project on these issues. For each category of statistical tools one or two international experts will be invited.

Phase 2

Workshop 3 (extended two-day meeting): half way of year 3

The workshop is planned in order

(1) exchange results obtained from the work packages regarding their primary objectives and cross-links and from research on methodological issues, and

(2) prepare intensive interactions between sections.

Workshop 4: half way of year 4

This will be a follow-up workshop on transfers started in workshop 3.

Phase 3

Workshop 5: beginning of year 5

The aim of the workshop is

(1) to report on further work package results and cross-link research, and to assemble methodological findings from the different work packages in all previous stages, and, most importantly,

(2) to prepare research on generic methodological conclusions**.**

For each of the workshops at least one external expert will be invited to comment on plans and results. Also the members of the follow-up committee will be invited for all workshops, so that they can follow the realization of the project and give their comments on previous achievements and suggestions for future work.

4. Research seminars

Most of the partners have a regular statistics seminar in their department. We do not foresee to organize extra sessions of seminars. Rather, we will ask each partner to focus their institutional seminar from time to time on topics of direct interest for the research network. For these network-focused seminars, the other partners of the network will be invited and, depending on the topic, also an expert from outside the network may be invited to give a presentation or to comment on the work that is discussed in the seminar.

5. Staff meetings

Two kinds of staff meetings are planned:

(1) General staff meetings with representatives of all partner groups will be organized at regular times, including at the occasion of the five workshops.

(2) Also staff meetings per work package are planned, with the staff that is involved in the work package and, depending on the agenda, possibly also with staff involved in linked work packages.

6. Intensive courses

On both categories of common statistical tools for the project – smoothing and resampling/Monte Carlo techniques – an intensive course will be organized during the first year or early in the second year (in association with workshop 2). In addition, we will ask the doctoral students to take two courses a year on a topic of the network. These will be courses that are either organized by one of the seven partners of the network or by another organization, for example the IOPS (Interuniversity Research School of Psychometrics and Sociometrics in the Netherlands). There are ample of such courses provided each year at the various institutes in the network. For example, the UCL proposes to organize a course on " Frontier estimation and detection of abrupt changes" to be taught by a visiting professor, and at the LUC there are each year visiting professors (in the Master of Biostatistics program) teaching courses on specific topics.

7. Newsletter and Website

In order to keep each other updated on the various special research activities related to the network we will establish a Network's Electronic Newsletter and a website. The newsletter will be sent out monthly and will refer to the Website for detailed and specific information. Of course, the website can be consulted on a continuous basis. The newsletters will be archived on the website. Among the information that we will exchange via these means, we mention:
(1) announcements of
- seminars related to the network-topics
- network workshops
- interesting conferences
- visitors related to network topics
- arrival of new post-docs and doctoral students
- interesting scientific results obtained

(2) reports
- short reports of research results, preliminary to complete manuscripts to be submitted
- manuscripts to be submitted, submitted, and published

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.

8. Open network activities

We distinguish between two types of openness: (1) to other Belgian statistical groups, including those who initially have participated in this preparation of this network, and (2) to international statistical groups.
(1) Members from other Belgian statistical groups will be invited at the workshops. A more intensive collaboration is possible via members of those other groups being co-promotor of a doctoral student associated with the network. Some of the objectives are quite appropriate for this purpose, as they have been initiated by these other groups during the preparation of the project. For example, part of the classification techniques objective from WP5 has been initiated by Jean-Paul Rasson (University of Namur), and the causal relations objective from WP6 has been initiated by Els Goetghebeur (University of Ghent).
(2) International experts will be invited for the workshops and some of the intensive courses will be taught by international experts (see above). Furthermore, all partner groups have extensive international contacts. In order to further stimulate international contacts, the network will look for high-quality statistical institutes and research groups abroad, in Europe (including Eastern Europe), but also in North America, in Asia and South-America. We hope to intensify these contacts and to extend them, if appropriate also in the form of research projects (e.g. bilateral projects thus far funded by the Flemish and French-speaking Communities, and EU projects). International networking is beneficient for the quality of research and for the recruitment of post-doctoral researchers and doctoral students.

I.9. RE-ORGANISATION OF THE PROJECT

To be completed only if the initial proposal has to be 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.
(max. 3 pages).

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 or IV) and the titles of projects in which the partners of the present network have participated.

I.11. VALORISATION OF SCIENTIFIC ACTIVITIES

(max. 3 pages)

List the different possibilities of valorisation of the network's activities, *i.e.* the dissemination, usage and transfer of acquired knowledge during the project in various ways such as expositions, publications, presentations, reports, teaching and other forms of transfer of knowledge. This should concern in particular the general public.

This list is indicative and is not subjected to the provisions of article 16.1 of the IAP-contract.

The valorisation of the network's activities can be done on various levels:

1. transfer of knowledge to (graduate) students and young researchers;
2. exposition and transfer of knowledge to the scientific community of statisticians;
3. exposition and transfer of knowledge to a larger community of statisticians, probabilists and mathematicians;
4. exposition and transfer of knowledge to the larger community consisting of people interested in statistics in general or using statistics in their discipline;
5. usage of the knowledge to statistics and other disciplines.

In order to ensure the valorisation of the network's activities at these various levels the following activities are planned. Each of these aims at the valorisation at one or more of the above levels.

(i). *Presentations at seminars and conferences*
A first way to exchange knowledge about results obtained via research in the network is by presenting these results during seminars or conferences. Members of the network should of course first of all communicate the results that they obtain to other members of the network. This can be done via, for example, reports posted on the website. See also items (ii) and (iv) below. Since almost all partners involved in the network organize on a regular basis one or more statistics seminar (most of them on a weekly basis) these seminars seem to be a first possible place to exchange the knowledge to other members of the network, as well as to other researchers outside the network.

Some partners of the network organize, separately, a so-called Doctoral students seminar', often organized every two or three weeks. Such a regular seminar is then the occasion for doctoral students to present their results obtained under the network.

When talks in these seminars are focused on scientific results obtained in the network, these seminars should be announced under a special heading (of the IAP-statistics network and the local organizer) and members of the network will be informed via the network of this event.

These actions aim at levels 1 and 5 of the valorisation.

All countries directly involved in the network (Belgium, France and Germany) have a quite strong society regrouping all statisticians/probabilists. As an example we mention the Belgian Statistical Society (BSS) which is a very active society counting among his members most of the Belgian statisticians working in academia and private or public sector. At this moment, the president of the society is the promotor of the LUC-partner of the network (prof. N. Veraverbeke), and the vice-president (future president-to-be) is Prof. A. Albert, who is a member of the follow-up-committee of the network. In addition, several members of the network are members of the board of the society, and almost all are regular members of the society. As most of the statistical societies, the BSS publishes a newsletter (B-Stat News), four times a year, and organizes annual scientific meetings of two days. The statistical societies in France and Germany organize similar activities. Talks during the annual meetings of the societies should expose the important results obtained in the IAP-statistics network. These actions aim at levels 2 and 5 of the valorisation.

With the above actions we will mainly reach researchers in Belgium, France and Germany. By participating in other international conferences, such as for example the European Meeting of Statisticians (organized each year), the meeting of the International Statistical Institute (organized each two years), the World Congress of the Bernoulli Society, the Annual Meeting of the Institute of Mathematical Statistics or the American Statistical Association, researchers of the network will have the possibility to expose their results to a more international audience of researchers. These actions aim at levels 3, 4 and 5 of the valorisation.

(ii). *IAP-statistics Technical Reports Series and Reprints Series*
The results that will be obtained by researchers working on research projects of the IAP-statistics network will be reported on in scientific papers that will be submitted for publication in international scientific journals. Two publication series of the IAP-network will be created to surround this activity. First we will create a IAP-statistics Technical Reports Series which will group all papers written under the IAP-statistics network. Each paper in this series should be submitted for publication in an international journal. Once a paper has been accepted for publication in an international journal and has been printed, we will list it into the IAP-statistics Reprints Series.

For the IAP-statistics Technical Report Series we will list (title and authors) all papers on our website and for each paper we will post a document (ps file or pdf file of the paper) that can be downloaded from the site. For the IAP-statistics Reprint Series we will provide on the website a list (title, authors, abstract) of all published papers. Both lists will be published in B-Stat News from time to time.
An additional possibility for both series could be to reproduce for each paper a number of hard copies that can then be used to distribute to interested readers (for example at conferences). The extra costs for this option have to be evaluated, as well as the need for it. This will be done shortly in consultation with the other network-members.

These actions will mainly help in achieving levels 1,2, 3 and 5 of the valorisation.

(iii). *Teaching of advanced courses and `Continued education' courses*
In the various educational programmes in statistics that are organized at the institutions, there are a number of statistics courses of second and third cycle that can benefit indirectly from the results obtained under the IAP-statistics network. A course such as ``Special research topics'' (see the programme of the Graduate School in Statistics of the UCL) certainly leaves an opportunity to summarize important findings or to present the state-of-the-art of research in a specific area. In short, transfer of knowledge to students at an advanced level can be done via advanced courses in the existing programmes. This can even be done in international courses in which members of the network act as Visiting Professors.

Apart from the above specialized courses, it occurs quite often that members of the network participate (as professors) in educational programmes aiming at a more general audience, interested in the use of statistics on a more general level. The audience often consists of people working at banks, in industries, for governmental organizations, etc. These educational programmes are either organized by the universities themselves (`Continued education' programmes) or are organized by private firms or other organizations for their employees (they then invite specialists from universities to give the courses). Members of the Institute of Statistics at the UCL for example teach each year several courses at this general level. Important findings from the research in the network, can be reported on (in general terms) in these courses.

These actions aim specifically at levels 4 and 5 of the valorisation.

(iv). *Short reports on new results in subprojects, informative publications in bulletins*
Short reports that explain briefly and in big lines important results obtained in the IAP-statistics network can be posted on our webpages. These reports can also be published in the bulletins of some statistical societies, for example in B-Stat News, the bulletin of the Belgian Statistical Society. With the latter publications we focus on levels 3, 4 and 5 of the valorisation.

These actitivities will be adjusted according to experiences, when needed.

II : Budget (global distribution per partner)

(not to be completed for the corresponding partner)
(in EURO, without decimals)

	Personnel	Operating costs	Equipment	Overheads	Subcontracting	Total
Partner : Léopold Simar (UCL partner)	600590	190452	69405	39553	0	900000
Partner : Paul De Boeck (KUL-1 partner)	472908	69350	55630	27112	0	625000
Partner : Noël Veraverbeke (LUC partner)	263188	85000	9402	17410	0	375000
Partner : Marc Hallin (ULB partner)	224992	97910	35936	16162	0	375000
Partner :	0	0	0	0	0	0
Partner :	0	0	0	0	0	0
Partner :	0	0	0	0	0	0
Partner :	0	0	0	0	0	0
Partner :	0	0	0	0	0	0
Partner :	0	0	0	0	0	0
Partner :	0	0	0	0	0	0
Partner :	0	0	0	0	0	0
Partner :	0	0	0	0	0	0
Partner :	0	0	0	0	0	0
Partner :	0	0	0	0	0	0
Partner :	0	0	0	0	0	0
European partnership 1: Hans-Hermann Bock (RWTH - Aachen partner)	50000		Not allowed	Not allowed	Not allowed	50000
European partnership 2 : Anestis Antoniadis (UJF-LMC-IMAG partner)	50000		Not allowed	Not allowed	Not allowed	50000
European partnership 3:	0		Not allowed	Not allowed	Not allowed	0
European partnership 4:	0		Not allowed	Not allowed	Not allowed	0
Total						**2375000**

Note: The budget for the European partnership in the table is the budget attributed by the IAP-programme only.

- Personnel: indexed gross remunerations; employer's social contributions; statutory insurance costs as well as any other compensation or allocation legally due in addition to the salary; the reimbursements for PhD grant holders (exempt from tax and benefiting from social security) . Personnel costs must account for minimum 60% of the total budget attributed to each partner of the network. The costs for the tax-free PhD grants may not account for more than 50% of the total personnel costs.

- Operating costs: basic supplies and products for laboratory, workshop or office; documentation; travel and accommodation; use of computing facilities; software; telecommunications; maintenance and operation of equipment and, more generally, consumables; hosting of visiting foreign researchers; reimbursements for non-EU grantees in accordance with the OSTC rules.

- Equipment: acquisition and installation of scientific and technical appliances and instruments, including IT equipment placed at the project's disposal. Equipment cannot be asked for during the last year of the programme.

- Overheads: general expenses of the institutions covering, on an inclusive basis, administrative, telephone, postal, maintenance, heating, electricity, rental, material depreciation and insurance costs. The total amount for this heading may not exceed 5% of total personnel and operating costs.

- Subcontracting: costs incurred by a third party in order to perform tasks or provide services necessitating specific scientific or technical skills outside the normal framework of the institution's activities Each request for subcontracting needs a approval from the programme administrator.

III. COMPOSITION OF THE FOLLOW-UP COMMITTEE

The follow-up committee is composed of the different partners of the network mentioned in Section I.2., a minimum of 3 experts from outside the network and belonging to the scientific community (Belgian or non-Belgian), and a representative of the OSTC.

Mention here the experts from outside the network who agreed to participate in the follow-up committee of the network.

---

1. Name of expert: Adelin Albert
   Speciality : Applied statistics in an advanced university medical centre
   Institution : University of Liège
   Research unit : Medical Informatics and Biostatistics
   Address :
   Road/Street : CHU Sart Tilman (B35)
   No. :
   Post code : 4000
   Town/City : Liège
   Country: Belgium
   Tel. : +32-4-366 2591
   Fax. : +32-4-366 2596
   Email : aalbert@ulg.ac.be

2. Name of expert: Delecroix Michel
   Speciality : Statistics - Applied Mathematics
   Institution : ENSAI (Ecole Nationale de la Statistique et de l'Analyse de l'Information)
   Research unit : Laboratory of statistics and modelling
   Address :
   Road/Street : Blaise Pascal, Campus de Ker Lann
   No. : BP 37203
   Post code : 35172 BRUZ CEDEX
   Town/City :
   Country: France
   Tel. : +33-2-99 05 32 42
   Fax. : +33-2-99 05 32 06
   Email : delecroi@ensai.fr

3. Name of expert: T.A.B. Snijders
   Speciality :
   Institution : University of Groningen
   Research unit : Department of Statistics & Measurement Theory
   Address :
   Road/Street : Grote Krulsstraat
   No. : 2/1
   Post code : 9712 TS
   Town/City : Groningen
   Country: The Netherlands
   Tel. : +31-50- 363 6188
   Fax. : +31-50 363 6304
   Email : t.a.b.snijders@ppsw.rug.nl

4. Name of expert:
   Speciality :
   Institution :
   Research unit :
   Address :
      Road/Street :
      No. :
      Post code :
      Town/City :
      Country:
   Tel. :
   Fax. :
   Email :

5. Name of expert:
   Speciality :
   Institution :
   Research unit :
   Address :
      Road/Street :
      No. :
      Post code :
      Town/City :
      Country:
   Tel. :
   Fax. :
   Email :

6. Name of expert:
   Speciality :
   Institution :
   Research unit :
   Address :
      Road/Street :
      No. :
      Post code :
      Town/City :
      Country:
   Tel. :
   Fax. :
   Email :