**MPHIL IN STATISTICAL SCIENCES**
**2006-07**

## Applied Project Titles 2007

**(as summarised by their authors)**

| | |
|---|---|
| **Guillaume Bessi** | Models of default |
| **Christiana Charalambous** | Numerical solution of the broadcasting problem when peers have differing upload rates |
| **Xiaonon Che** | Markov-type models of the Real Time Gross Settlement payment system |
| **Chen Chen** | Ramp metering as a demand management strategy in road networks |
| **Georg Grull** | A large deviations approach to optimal business portfolios |
| **Ryan Li** | Calibrating a simple credit model |
| **Peter Man** | Calculating derivatives of parameters in population balances or `` What are detergents sensitive to?" |
| **Sanjeet Mangat** | Scaling in knock-out tournaments |
| **Donal Moore** | Pan-European return forecast modelling using a dynamic linear model |
| **Colin Prue** | Nonparametric smoothing issues in the oil industry |
| **Shijie Ren** | Higher order numerical schemes in finance |
| **Shiping Sun** | Developing techniques for forensic speaker identification |
| **Maria Vounou** | Regression trees in R, applied to health data |
| **Huizi Wang** | Modelling interest-rate data |
| **Georgios Zannoupas** | Re-evaluating survival data for cancer patients. Dichotomization and its consequences. |
| **Chao Zhang** | Statistical methods for non-linear dependence assets and sparse data |

**Guillaume Bessi**
**Models of default**

As the market for credit derivatives (also known as *contingent claims*, the two terms can be interchanged at will) has soared fiercely in recent years, so too has the demand from the banking industry for good models of credit risk. There are two main approaches for modelling the defaults of risky debt: the structural and reduced-form approaches. The present survey will include both a theoretical study about the valuation of defaultable claims within the framework of reduced-form models as well as an analysis of price data of a given type of derivatives, namely *Credit Default Swaps*, by implementing a statistical model.

**Christiana Charalambous**
**Numerical Solution of the Broadcasting Problem when Peers have Differing Upload Rates**

This project was supervised by Prof. R.R. Weber and Dr J. Mundinger.

BitTorrent, SplitStream and Avalanche, to name but a few, belong to the category of peer-to-peer (P2P) networks, which have proven to be a popular way of disseminating files such as videos, music or even software. Many P2P networks use the following very simple idea: a potentially large file is split by a single server $S$ into $M$ file parts and then distributed via the internet to a group of $N$ users, called the peers. As soon as a peer has downloaded a file part, he contributes by uploading it to other peers.

The following problem arises: How can we disseminate the file to a network of peers in the least possible time and subject to constraints on the upload capacities of the peers? This is the so-called Broadcasting problem and this project is motivated by the desire to find a solution by solving a series of mixed integer linear programs (MILPs). In practice, to solve an MILP, we first need to use the Primal and Dual simplex algorithm, to get a relaxation of the problem and then apply the Branch-and-Bound method for integer programming.

The main focus of the project is obtaining a numerical solution of the Broadcasting problem when peers have differing upload capacities. This is done by formulating the problem into an MILP and implementing it to an MILP solver. We have chosen ***LP_solve***, a free linear programming solver, mainly because it offers a user friendly Windows interface, namely ***LP_solve IDE***. It's very easy to use and more importantly doesn't require a computer programming background.

The solver uses the same methods as we would in practice but in this case, it suffices to input the objective function and constraints, press the solve button and wait for the result. The key idea is to try different combinations of $M$, $N$ and the capacities in order to investigate how the minimal make span depends on these values and also compare the computation time, which increases exponentially as these parameters get bigger.

We provide both an exact solution, as well as approximations and relaxations of the solution and compare the solving times. To begin with, we examined the exact solution and recorded the computational times. In addition, we obtained some formulae for the exact solution for certain small values of $M$ and $N$, which were verified via testing on the MILP solver.

In the case where a problem could not be solved within some reasonable time limit, an approximation method was applied. This gave us an upper and lower bound for the solution. In fact, in several cases we could even reach the exact solution via this method. We recorded many cases where an unsolved problem could be dealt with in just a few seconds when an approximation was used.

Furthermore, we solved relaxations of the MILPs, by removing the integer constraints of the problem and discovered that the relaxed problems could be tackled by the solver with strikingly great ease.

Taking everything into account, we realised that both the solution and the solving times are greatly dependent on the instance of the problem and particularly on the choice and combinations of the input parameters *M*, *N* and the capacities. Finally, we outlined the difficulties we came across while doing this research and offered some conjectures for further research.

**Xiaonon Che**
**Markov-type models of the Real Time Gross Settlement payment system**

This project is supervised by Dr. O. Lobunets and Prof. Y Suhov.

In recent years, the interest in modelling the large-valued inter-bank payment system has been growing rapidly. However, to our knowledge, there is no mathematical model which considers the inter-bank payment system as a stochastic process in existing literature. The aim of this project is to investigate the statistic properties of the inter-bank Real Time Gross Settlement (RTGS) payment system using a dynamic Monte Carlo algorithm.

In the response to the liquidity management with the RTGS payment system, more recent RTGS payment systems introduce the queuing facility. With queuing and collaterised borrowing facilities, the liquidity level of the system enhanced (chapter 2). Since stochastic process models have been wildly used in social networks, and some aspects of social network have similar statistic properties with the inter-bank payment system (chapter 3), therefore, we decided to fit the large-valued inter-bank payment as a stochastic process. Moreover, based on the empirical research on the inter-bank payment system (chapter 5), we have proposed a model with the parameters:

- N, the number of participating banks
- q, probability for a payment order is put in the queue
- p, probability for each participating bank has one unit cash at the beginning of a business day
- n, the number of groups among N banks
- $P_{ij}$, the probability for a payment order is transferred from bank in group i to a bank in group j
- $P_{ii}$, the probability for a payment moves between the banks in a single group i

Thus, the dynamic of the payment flow could be viewed as a finite-state space, discrete time Markov process determined by N, n, q and p, with transition probabilities $P_{ij}$ and $P_{ii}$.

Consider a random graph of N nodes where the nodes represent participating banks and the links indicate the payment flow. Each node is classified by three factors:

- cash position (positive, negative and zero)

- number of debts (outgoing payments)
- exposure to other participating banks (incoming banks)

We carry out a simulation with a dynamic Monte Carlo method (chapter 6), which provides us with the choices of the values of the parameters in the proposed model. The results are illustrated by two histograms. One graph shows the average number of end-of-the-day debts per node, while the other histogram is telling the averaged lifetime of a single debt after 480 Monte Carlo iterations.

The dynamic Monte Carlo simulation is carried out in MATLAB.

**Georg Grull**
**A large deviations approach to optimal business portfolios**

This project deals with the problem of optimal choice of new business. The effects of opening a new business on an insurer's infinite time ruin probability are studied in the Large Deviations (LD) regime.

Research on this topic had been conducted by applying methods of stochastic control to an extended Cramér-Lundberg model. Hipp and Taksar (2000) and Kelbert *et al*. (2005) showed that an insurer can reduce its infinite time ruin probability by opening a new business and selling it again at a later time. In this model the insurer decides to buy or sell a new business when the risk reserve reaches an optimal level and the new business does not even have to be profitable as premium income could exceed incoming claims for at least a short period.

Methods of LD were chosen for the verification of this result because LD is the field within probability theory that deals with the analysis of rare events such as ruin for example.

In the LD regime opening a new business reduces the infinite time ruin probability if and only if the new business satisfies a profitability condition that is stronger than the 'Net Profit Condition'. Thus in contrast to the stochastic control approach an insurer can only reduce its infinite time ruin probability by opening a profitable
business.

The results of the LD approach could be verified by carrying out simulations.

The models "Single business" and "New business" that are used in the LD approach are to be found in Chapter 2. Chapter 3 provides an introduction to the theory of Large Deviations. In Chapter 4 these techniques are applied to the problem of optimal choice of new business. After calculating the infinite time ruin probability in the "Single business" and in the "New business" model in Section 4.1 the equivalent profitability condition is derived in Section 4.2. Section 4.3 provides lower and upper bounds for the minimal profitability margin. In Section 4.4 numerical illustrations of the minimal profitability margin in the case of Gamma distributed claim sizes are presented. The C++ algorithm in the Appendix was used to produce these illustrations.

**Ryan Li**
**Calibrating a Simple Credit Model**

This project was supervised by Prof L.C.G. Rogers and Dr M.R. Tehranchi.

In this project we study a new portfolio credit risk model proposed by Di Graziano and Rogers (2005). The underlying dynamics of the model is driven by a continuous-time Markov chain, which can be regarded as the systematic risk factor or the ``state of the economy". Defaults occur independently conditional on the path of the chain, and the default time of each name is modelled as the first jump time of an inhomogeneous Poisson process, with intensity specified as a function of the chain. The default correlation is obtained endogenously from the model and dynamic, which means that the model is suitable to price products such as forward starting tranches and tranche options.

The model is tested through calibration on CDS term structures and standard tranche spreads observed in the market. In particular for the CDS data, we consider the daily term structure quotes available for a total of over 200 trading days (a time period of 10 months). This allows us to investigate in depth not only the calibration capabilities of the model, but also its robustness by analysing the parameter stability over time.

The paper is organized as follows. After a brief introduction in Chapter 1, Chapter 2 presents the model and discusses various results that will be used in later chapters. Chapter 3 illustrates the pricing of some standard credit derivative instruments based on Laplace transform. Chapter 4 develops a fast and accurate approximation to the Laplace transform of the loss distribution. Chapter 5 presents the numerical results and Chapter 6 concludes.

The implementation and calibration of the model were done in MathWorks MATLAB. Microsoft Excel was used for data manipulation and preliminary analysis.

**Peter Man**
**Calculating derivatives of parameters in population balances or ``What are detergents sensitive to?''**

In a coagulating particle system, we often would like to see how sensitive this system is to the model parameters. There are many practical applications for this - it may be that in the production of detergents, for example, the quality of the end product may be influenced strongly by the speed of the mixer. Also, in model selection we often wish to know whether certain parameters are important in the physical problem being considered, and in performing sensitivity analysis we can determine this.

In general, the coagulation of particles is modelled by the Smoluchowski Coagulation Equation, the solution to which we can approximate stochastically using Marcus Lushnikov processes. It can be shown that Marcus Lushnikov processes converge to the solution of the Smoluchowski Coagulation Equation in the limit as the number of particles tends to infinity. By using these processes combined with current Monte Carlo techniques, the solution can be approximated with reasonable speed and accuracy. In this project, only the Direct Simulation Monte Carlo method (DSMC) will be used to implement the Marcus Lushnikov process, in which each particle is given equal weighting.

It is possible then to perform sensitivity analysis by estimating the derivative of the solution with respect to the model parameters. However, current methods in general lack precision and this gives large variances for our derivative estimate. The main aim of this project is to implement in C two new algorithms (that Professor Norris has developed) which attempt to minimise this variance via coupling techniques. Once this has been achieved, the project will then analyse their performance using R. The performance will be measured by how quickly the algorithms run in order to achieve a specified confidence interval length in the parameter gradient estimate.

**Sanjeet Mangat**
**Scaling in Knock-out Tournaments**

We study knock-out tournaments, where at each round players compete in head-to-head matches. The winner progresses to the next round and the loser is eliminated, with the last remaining player declared the champion. Each player has a fixed intrinsic strength which may affect their chances of moving to the next round. This type of competitive process has relevance in many areas; we study the example of tennis using data from Grand Slams. All the data analysis and programming was carried out in R and S-Plus.

In the tennis setting it is natural to look at the world rankings as a strength measure. We define the strength of a player by the total number of ranking points they hold going into the competition. Intuitively we feel that stronger players tend to last longer in the tournament. But what can we say about how the strength distribution scales between rounds? And what is the initial strength distribution explicitly? We study 4 main issues in detail.

Firstly, prompted by a recent paper on tournament scaling, we investigate the hypothesis that the initial strength distribution has a \emph{power law tail}. We find the latter part of the distribution displays power-law behaviour, and we fit it via the methods of least squares and maximum likelihood estimation. A Kolmogorov-Smirnov goodness-of-fit test compares the two and confirms we have an adequate fit to the tail portion of the distribution by the MLE method.

Secondly, again prompted by the claim of a recent paper, we investigate the hypothesis that the strength distribution becomes asymptotically self-similar as the tournament progresses. Explicitly, this means that there exists a scaling function $g$ such that the distribution of the scaled player strengths is the same for all rounds $n$. This means the strength distribution has the nice property that we can scale it spatially as it varies through time, and collapse the later rounds back onto the first. In order to proceed, we conjecture a particular scaling function: dividing the data in round $n$ by the \emph{mean strength} of the round. This is seen qualitatively to have the desired effect on later rounds of data. We use a Kolmogorov-Smirnov two-sample test, applied with a bootstrap approach, to confirm our hypothesis that the data samples from different rounds scaled in this way are in fact drawn from the same distribution. This is clearly a significant result - players are eliminated via a stochastic process, there is no reason why each round should follow the same distribution when scaled by their respective means.

The third section answers the question of whether the upset probability (i.e. of a lower ranked player beating a higher ranked player) depends significantly upon the difference in

their strengths. This question arose during the analysis of the previous hypotheses. Papers in the past have assumed a constant $q$ for each match in a tournament - intuitively we realise this would not hold in the context of tennis. We answer this question positively, and use a method of logistic regression to estimate the exact dependence. We perform this analysis for both men's and women's tennis and compare the fitted upset probability curves for both sexes. We notice the curve for women lies above that for men, so we might conclude that women's tennis is the more competitive, where we define `more competitive' here to mean a higher probability of a surprise winner for a given strength difference.

The final section returns to the initial strength distribution and attempts to find a parametric model to improve on the power law tail fit. Empirical graphical techniques such as mean residual life and quantile-quantile plots lead us to consider heavy-tailed distributions such as the lognormal and Weibull. The Anderson-Darling goodness-of-fit test indicates the best fit for the full distribution is a lognormal. We further offer possible explanations for the shape of the distribution observed.

Throughout the project we study recent tournaments played on the same surface, and differentiate between the sexes in some areas to make comparisons. We find that the initial strength distributions are comparable between these tournaments, and so pleasingly our conclusions will not just be limited to the few tournaments we have studied. There are certainly areas open for further study, for example comparisons between different surfaces, or to examine how the characteristics of the strength distribution have changed over the past two decades say. Further, due to the ubiquity of the process studied, there are certainly many areas beyond the context of tennis where the methods and ideas of this project could be applied.

**Pan-European Returns Forecast Modelling using a Dynamic Linear Model}**
**Donal Moore**

In linear models structural instability is caused by parameter instability, where parameters change over time. Estimates derived from linear models where the relationships between the data and the parameters are incorrectly considered as stable are not meaningful. Inferences and forecasts obtained from these model can be very inaccurate. With this in mind, there are two questions which drive this project.

1. How do we detect parameter instability in linear models?
2. What methods are available to tacke the problem of parameter instability?

For the first question,we outline the Cumulative and Moving sums test of residuals, the Moving Estimates test and the Nyblom-Hansen test based the score statistics obtained from a linear model. These are designed to test the null hypothesis of parameter stability. With the exception of the Nyblom-Hansen test, the alternative hypothesis for these tests does not assume any particular structure, merely that the parameters are not stable. The Nyblom-Hasen test is designed to test the null hypothesis of parameter stability against the specific alternative that the parameters follow a martingale.

Armed with these tests, we can now test for parameter instability in linear models. However, if parameter instability is detected, how do we adjust our models to incorporate this problem? There are many different approaches but we concentrate on the Dynamic Linear

Model. The Dynamic Linear Model is a model where the parameters are allowed to evolve over time. They encompass a wide class of models. General Linear Models, Generalised Linear Models, ARIMA models and many more besides can be expressed within the Dynamic Linear Model framework. To estimate the parameters, we use the Kalman Filter. It is the recursive nature of the Kalman filter that makes the Dynamic Linear Model appropriate for modelling data which causes parameter instability in Linear Models. The Kalman Filter continuously projects forward estimates and update these estimates with each new piece of data that we observe. However, before we use the Kalman Filter, we must first estimate the unknown variance elements that are required to fully specify the Dynamic Linear Model. We outline two methods for estimating these variance matrices. The first method, is based on what is called Variance Discounting can be considered as a way around having to explicitly estimate the variance of the parameters. The second method is based on maximum likelihood.

The project arose from the desire to forecast returns in the Pan-European market. However, from the outset we believed that parameter instability in linear models may be a problem. We have a number of explanatory variables to use for modelling purposes, ranging from fundamental data, such as the Effective Exchange Rate and Dividends Yield, to technical data such as the Relative Strength Index. We use these variables to find an appropriate linear model to model the first half of the data. This model is then tested for parameter instability over the range of data that was used to fit the model. We then extend this linear model to a Dynamic Linear models where the parameters are allowed to evolve over time according to a random walk. Using the Maximum Likelihood method we find estimates of the variance matrices and fit a Dynamic Linear Model to the data. In the final sections, we assess the forecasting capabilities of each model using the second half of the data.

All the work is carried out in R. The packages, DLM used to fit Dynamic Linear Models and the package strucchange is used to carry out most of the tests of parameter stability.

**Nonparametric smoothing issues in the oil industry**
**Colin Prue,**

Supervised by Dr Richard Samworth (Statistical Laboratory), and Dr Mukund Unavane
and Dr Owen O'Loan (Spiral Software)

Crude oil, in its raw form, is of little practical use. It is a mixture of hydrocarbons with molecular weights ranging from just a few up to over 2000. Crude oil becomes useful when it is separated into fractions according to boiling point (which is determined predominantly by the molecular weights of the molecules).

The usual refining process, therefore, begins with fractional distillation of the crude oil, followed by further treatment processes to remove impurities. Tests are carried out in order to characterise the amount of hydrocarbon, as well as the amounts of impurities across the boiling range of the crude oil. However, due to limited time and resources, measurements are taken only for a selection of fractions. This measured data becomes much more useful when it can be used to predict the impurity content, say, for a fraction with any arbitrary boiling range together with an estimate of the uncertainty.

The first task at hand was to find the sulphur content of an arbitrary fraction given complete

information about the boiling point and a small number of measurements on the sulphur content. Estimation of this type cannot be performed with standard nonparametric techniques, as the sulphur measurements available are, rather than point-values, mean-values of the regression function.

Solving this problem required adaptation of the natural cubic smoothing spline methodology, and resulted in a generalisation of the standard theory able to accommodate minimisation of least squares on integrals of functions, obtaining many interesting results along the way.

The second component of the investigation involved an analysis of the uncertainty behind estimation using the above technique. It was concluded that the parametric bootstrap offered a good way of constructing a confidence interval on the regression function, the coverage of which matched that quoted by the construction of the technique.

This project had a large component of developing new theory, whilst also allowing scope to play with large sets of data and make interesting influence on a real life problem.

**Shijie Ren**
**Higher Order Numerical Schemes in Finance**

Mathematical finance helps investment decision-making and determination of fair prices for complex investment products. For example, Monte Carlo method can be applied to generate the price of a contingent claim. In Monte Carlo method, the biases of the estimate of a contingent claim payoff come from two parts: approximation of continuous time processes by time discretisation and averaging of different sample paths. This study aims to reduce the former bias, namely the time discretisation error, to provide a better numerical scheme of sample path pricing in financial modelling.

In Chapter 1, we first introduce the stochastic calculus and some important concepts in mathematical finance, such as replicating portfolio, discounted factor and the completeness of the market, which can allow us to apply Monte Carlo method. The discrete time model, the continuous time model and how the above mathematical financial concepts are used in such models will be shown in detail. Application of Monte Carlo method in financial engineering will also be introduced.

Chapter 2 presents four numerical schemes. Two weak order 1.0 schemes are the Euler scheme and the refined scheme. Two weak order 2.0 schemes are the second-order Taylor scheme and the extrapolation method. The higher orders of convergence are achieved using Itô-Taylor expansion and applying the method of extrapolation to two estimators at different levels of discretisation from weak order 1.0 scheme such as the Euler scheme.

In Chapter 3 and 4, we simulate the European call option under the Black-Scholes model and Heston's model, respectively. The codes of the simulations are written in C. Relative errors are compared for the four numerical schemes. Our conclusion is that for both models, higher order numerical schemes obtain smaller biases than weak order 1.0 schemes. Specially, the second-order Taylor scheme is the most accurate scheme. An alternative way is to use the extrapolation method, since it is much easier to implement than the second-order Taylor scheme.

**Shiping Sun**
**Developing Statistical Techniques for Forensic Speaker identification**

The aim of this project is to develop techniques for characterizing individual speakers by modeling the trajectories of formant frequency contours mathematically, using regression methods.

We have the data from the recordings of ten native standard English speakers taken at the Phonetics Laboratory at Cambridge University. Each speaker was asked to produce a number of utterances containing the sound /r/. We have the samples of all the five sequences, which are /sc_ri/, /sc_rae/, /sc_ra/, /sc_ro/ and /sc_ru/. For each target sequence(e.g/sc_ru/), and for each speaker we have four test sentences. Five repetitions were analyzed for each of these. First, second and third formant frequencies(F1, F2, F3) have been measured at regular intervals during each utterance of /r/ such that contours of time versus frequency can be plotted.

This project first investigates the results obtained using the LDA(Linear Discriminant Analysis) which enable us to make predictions about the identity of the speakers. Throughout the LDA, the sequence /sc_ru/ performs better than the other sequences, e.g /sc_rae/ and /sc_ro/. The application of LDA based on the original F3 data set shows that the sum of the first two linear discriminants of the sequence/sc_ru/ (84.66%) is larger than any other sequence. It is noted that by taking log on the full F3 data set, it only results in a tiny improvement in the proportion of the first two discriminants from the proportion of the trace for sequences /sc_ru/ and /sc_rae/, but a little decrease for the sequence /sc_ro/. We may conclude that by taking log on the data may give better result in some cases. The details were given in Chapter 4 of the project.

Secondly, we consider parameterizing the formant curves, that is, averaging over the data and doing the fit, so that we can add four cubic polynomial terms as a summary of these data. In this way, more information about the dynamics of different speakers has been obtained. We conclude that using the shapes of the curves for F3 does not significantly improve the previous results. This method was explained in Chapter 5.

Finally, we carried out the Quadratic Discriminant analysis and it suggests that using the curves of the F3 data could not improve the discrimination between speakers but instead might even make it worse in some cases.

ALL the programming was implemented in R 2.0.1

**Maria Vounou**
**Regression Trees in `R', Applied to Health Data**

Tree models were introduced to statistics in the 1980's and have become very popular since. Tree models form a statistical modelling technique for examining and predicting a response variable from several other explanatory variables. They constitute a non-parametric method in the sense that they do not make any assumptions about the functional parametric relationship between the response variable and the other explanatory variables of the dataset.

There are two types of tree models depending on the nature of the response variable. If the response is continuous a regression tree is constructed. On the other hand if it is categorical a

classification tree is constructed. These models are constructed by a recursive partitioning algorithm, acting on a mixture of continuous and categorical explanatory variables and can be represented graphically as binary trees.

This project attempts to give an insight into tree models and how these are used in practice to draw inferences and make predictions. The project concentrates on the application of regression tree models to health data.

The first section details the method of constructing trees in `R' using the recursive partitioning function "rpart" stored in the "rpart" library. The "rpart" function uses the following algorithm to construct the trees. Beginning from the root of the tree and using the whole dataset, "rpart" splits this dataset, according to a splitting criterion, into two mutually exclusive and exhaustive subsets defined to be the two new nodes of the tree. It then creates further binary splits on each of the resulting nodes in a similar fashion. This partitioning procedure continues until certain predefined criteria are met. These are termed as stopping criteria. If no further partition is achieved at a certain node the algorithm denotes the corresponding node as a terminal one. At each terminal node the recursive partitioning algorithm provides us with a predicted response.

Nevertheless, the best way for one to fully understand how tree models are constructed and utilised is to consider their application to a dataset. This is the topic of the second section where the dataset I used is taken from the Australian Health Survey 1977-1978. Using this dataset, I modelled the variation and possible dependence of the number of visits to the doctor on the other 18 explanatory variables of the dataset. In this section, I illustrated how one can construct the ``best" tree model for this particular dataset, taking into account the different nature of the explanatory variables (categorical/continuous), as well as the zero-inflated nature of the dataset (very large number of observations with zero response). I also illustrated how the ``optimal" size for our ``best" model can then be chosen using the so-called method of cross-validation.

Having illustrated and explained the use and application of tree models through my first dataset, I moved on to apply this theory to another dataset. This was the topic of the third section. I tried to determine whether my dataset suggested that maternal smoking relates to infant mortality. The dataset originated from the Child Health Development Studies that took place in California. However, the dataset consisted only of observations of infants that survived for at least 28 days after birth. Thus infant mortality could not be modelled directly. Hence I modelled the two principal causes of mortality, the length of gestation and the infant birth weight and tried to determine whether maternal smoking is related to these two causes. In this dataset, I also experienced the common phenomenon of missing values and I illustrated how tree models offer methods to handle with this problem.

At the end of the second and third sections I also considered how tree models compare with the well known parametric models, linear and generalised linear models. This was revisited in the final section of the project where I gave a detailed overview of the advantages and disadvantages of tree models. The most remarkable point to consider from examining the advantages and disadvantages of tree models is that the greatest disadvantage of tree models coincides with one of their important advantages, the fact that tree models are considered to be non-parametric.

Due to their relative novelty many ideas behind tree models are still not completely verified and several methods are not yet exploited to their full extent. Nevertheless, given their popularity, it is expected that tree models will become even more popular in the forthcoming years leading to the extension of the literature on trees and increasing the frequency of their use in statistical modelling.

**Modelling Interest-Rate Data**
**Huizi Wang**

The process of interest rate behaves like a process driven by a Brownian motion. The uncertainty of movement of interest rate opens up the possibility of interest rate derivatives whose payoffs are dependent on the level of interest rates, such as bonds, options, interest rate swaps and so on. They are widely traded in today's financial market. Estimating parameters is important as we need to use the fitted models to price interest rate derivatives. The aim of this project is to consider different methods of parameter estimation.

In this project, we mainly discussed about the calibration of Gaussian interest rate models. The 2 Gaussian-random-field models discussed in this project are the Vasicek model and the Kennedy model. Data used in this project are US Treasury Constant Maturity rates from the St Louis Federal Reserve Bank. There are daily, weekly and monthly data available form the web page
http://research.stlouisfed.org, we calibrated the model using the daily data in this project.

In the first part, we used spline techniques to plot the yield curves and the curve of short rates, from which we can see the movements of yield and short rate from 01/12/2005 to 01/12/2006. The Vasicek model expresses the evolution of the short rate in terms of a stochastic process with 3 parameters. We fit the Vasicek model to the short rates by 3 different ways to find the maximum likelihood estimates and iterative estimates for the parameters. In the second part, we discussed about the Kennedy model. Kennedy model is a Gaussian-random-field model that assumes the movement of forward rate is stationary and strictly Markov. Under these assumptions, the covariance structure of forward rates could be simplified to an explicit form that contains only 3 parameters. Again we used 2 different ways to do the parametrization. However, as the assumption of Kennedy model is stricter, some parts of our data do not satisfy the model. As forward rates with long maturities are more likely to be stationary, the Kennedy model can be fitted only to those with maturity greater than 2 years in this report.

The programs in this report were written in R and the typesetting was done in \LaTeX

**Project title: Re-evaluating survival data. Dichotomization and its consequences.**
**Georgios Zannoupas**

The investigation of the dependence of an outcome variable on several covariates forms the basic mechanism on which studies are conducted in many science fields. The simplest analysis that can then be carried out is linear regression, or multiple regression in the case where more than one independent variables are available. Many of these covariates are continuous and can therefore take values over a large range, yet it is a common practice to dichotomize them prior to statistical analysis. This is known as categorization which in simple terms means that the available data is split into two or more groups. Any subsequent

analyses usually investigate whether there is any significant difference in the mean of the dependent variable, between the groups of the dichotomized covariate. Despite the fact that there is extensive literature on the negative consequences of dichotomization, it seems that this method is still used regularly in various fields of science, ranging from psychology to all branches of medicine.

In this project the advantages and disadvantages of dichotomizing continuous covariates are discussed. Several methods of dichotomizing the data are presented, and a simulation study in survival analysis is conducted to compare these methods against the case where the covariate is kept continuous.

In order to illustrate the issues discussed, a detailed case study on breast and brain cancer patients is presented. This study also addresses recent advances in the Cox model, which include the use of martingale residuals to assess the functional form of continuous covariates, testing the model assumptions and the treatment of missing data.

All analysis was carried out using S-Plus and R.

**Chao Zhang**
**Statistical methods for non-linear dependence assets and sparse data.**

Asset Management is the branch of finance that involves making investment decision for clients who can either be institutional or retail in nature. The main areas of interest for a quantitative procedure in asset management are in performance analysis, risk management and portfolio construction. The focus of this project involves statistical techniques used in each of these main areas.

This project has been divided into two parts. The first part of the project is organized by using Stambaugh's Method and the expectation maximization (EM) algorithm to estimate the mean and covariance of the data set when I remove some data from my data set. Stambaugh's Method is presented first. This is mainly followed by Stambaugh (1997)'s paper: Analysing Investments Whose Histories Differ in Length *Journal of Financial Economics*, so I call Stambaugh's Method in my project. Furthermore, the EM algorithm is a general statistical method to deal with most missing data problems. I discuss the estimation of mean and variance for different cases, such as removing the top or bottom data in different percentages for each of the indices and removing data randomly from my data in different percentages for each of the indices. I use the error plot and histogram graph to compare the difference between Stambaugh's method and the EM algorithm. In conclusion, there is no strong conclusion enable us to decide whether Stambaugh's method or the EM algorithm is better in estimation, since the answer to these equations really depends on different cases and different indices.

The second part of the project investigates the non-linear dependence structure of the different asset classes. The concept of dependence plays an important role in finance and dependence across financial markets has been widely studied in the past decades. Traditionally, correlation is used to describe dependence between random variables. I give an example of the non-linear relationship showing that linear correlation does not capture the full dependency structure. I follow Embrechts (2001)'s paper: Modelling dependence with copula and Application to Risk Management. The various dependence concepts and two important copulas families: the elliptical copulas and the archimedean copulas are introduced. I develop practical algorithms for simulating from multivariate archimedean copulas, and apply the multivariate simulating copula algorithm to my data. These measures will then be used in the context of risk

management to evaluate the Value-at-Risk and the Conditional Value-at-Risk of the portfolio.

I used R and SPlus in programming and LaTeX to do the write-up.

This project was supervised by Dr. Meena Lakshmanan (Russell Investment Group) and Dr. Susan Pitts (Cambridge University).