

## **MSc Projects 2008**

- 1. Title: Analysing network statistics  
Spectral analysis of networks**
  
- 2: Title: Analysing network statistics  
Fitting distributions for conditional uniform graph**

**Proposer: Gesine Reinert**

Brief Description: The statistical analysis of networks has a long tradition in social sciences, and over the last ten years or so ideas from statistical physics have complemented the toolbox. New network summaries are proposed on a regular basis.

The two projects would have as common core the idea of analysing networks using summary statistics. They would both have as data sets available

1. a yeast protein interaction network;
2. a food web network.

The first project would study a whole range of network summaries which are based on eigenvectors and eigenvalues of the adjacency matrix describing the network. The study could be a literature survey, or a simulation study, or a combination of both.

The second project would study the distribution of summaries for simulated networks, where the networks are drawn uniformly under the constraint that certain summaries, such as the number of edges in the network, are fixed. Such networks are called conditional uniform graphs. The distributions of other network summaries are in general not well understood. The proposal is to simulate networks and to try to fit a distribution to the empirical distributions.

Potential for it to be a joint project? Yes (they are actually two projects)

Any prerequisite courses MCMC and computer-intensive statistics

Computing required? Yes; R packages are available.

Are data available? Yes: food web and yeast.

- 3. Title: Combination of two diagnostic tests**

**Proposers: Sara Downs and Andy Mitchell (VLA)  
David Cox (Nuffield College)**

Two tests, the tuberculin skin test and the gamma-interferon blood test, are available for testing cattle for bovine tuberculosis. They may be used separately or in parallel. The tests have different sensitivities and specificities and may perform differently in different contexts. The gamma-interferon test is the more expensive. Some simple models are needed to guide choice between the tests. Data from the Veterinary Laboratories Agency (VLA) will be used to guide model formulation and to illustrate the results.

**4. Title: Monitoring of rare events**

**Proposers: Sara Downs and Andy Mitchell (VLA)  
David Cox (Nuffield College)**

A rather general formulation is that there is a population of individuals some of which have a rare condition that can be discovered only on test. It may be known that the incidence of the condition is very different in different sub-populations. How should testing effort be distributed between the different sub-populations?

Part of the work will involve reviewing and possibly extending a theoretical analysis of this problem but the main focus will be on studying the implications for testing for bovine TB in the UK where different areas in the country are assigned to one-year, two-year and four-year testing in the light of incidence of TB in the area in question.

**5. Title: Effects of erroneous and missing data in genetic studies**

**Proposers: YY Teo and Kerrin Small**

Brief Description:

Genetic studies aim to identify the hereditary causes of diseases. This is often achieved through the use of: (1) logistic regressions, if the study design is a case-control study with binary outcomes (for example, whether a subject has diabetes or does not have diabetes); (2) linear or robust regression, if the study design investigates a quantitative/numerical trait (for example, the measurement of blood pressure); (3) quantifying the extent of excessive transmission of a particular factor in a two-level categorical variable from parents to offsprings (for example, trying to find out whether it is more likely for an offspring to have malaria if the offspring inherits a particular gene from the parents). The process of identifying the genetic composition for each individual is not perfect and can lead to errors or high rates of missingness. The research project aims to identify the consequences of erroneous and missing data on the three different study designs above, specifically focusing on the impact on statistical power and the rate of false positives.

Potential for it to be a joint project?: No

Prerequisite Courses: Statistical Methods, Further Statistical Methods

Computing required? Yes

Data available? Yes, and via simulations

**6. Title: Resource Allocation in Pre-clinical Pharmaceutical Research**

**Proposer: J C Gittins**

Brief Description: Writing and testing code in C++ to augment the existing OPRRA (Optimising Pharmaceutical Research Resource Allocation) written by my research group and now being trialled commercially.

Potential for it to be a joint project?: There will be coworkers, but they will be DPhil rather than MSc students.

Prerequisite Courses: Previous knowledge of C++ would be helpful but not essential.

Computing required? Access to C++ Visual Interactive.

Data available? We have enough. Not much is needed.

7. **Title: Implementation and analysis of the Hall/Jewson hurricane landfall model in R**

**Proposers: Trevor Maynard (Lloyd's of London) and Gesine Reinert**

Brief Description: Tropical Cyclone (TC) frequency has increased in the North Atlantic in recent years, as has the number of very intensive TCs. It has been established that sea surface temperature (SST) is a major driver for TCs. With climate change, SSTs over the North Atlantic tend to increase. As TCs can cause major damage and distress, prediction of a TC landfall site is of great interest.

The Hall/Jewson model for SST and North American TC Landfall predicts the geographic distribution of hurricane landfall sites, given the geographic distribution with SST. The geographic landfall model is based on a TC track model, which has three components: genesis, propagation, and lysis (death). Genesis sites are simulated by sampling a kernel pdf built around historic sites. For propagation, mean latitude and longitude 6-hourly displacements and their variances are calculated. Standardized displacement anomalies are modelled as an AR(1) model, with latitude and longitude treated independently. Finally, TCs suffer lysis with a probability determined by optimal averaging of nearby historical lysis sites. Tracks are stochastic, so that two tracks originating from the same point can follow very different trajectories. To assess the effect of SST on TC landfall rates, the track model is constructed separately on each historical SST set.

There are a number of modelling issues which one could investigate. One issue is that TC landfall rates are estimated using a Poisson process model, which assumes that the number of TC landfalls in a year is Poisson distributed, with the Poisson parameter being based on the previous year. If there were very few, or no, landfalls in the previous year, then this Poisson model might clearly be disadvantageous.

In order to study the model, it would have to be implemented in R. The implementation of the model in R would indeed constitute a substantial part of the project. Once the implementation is in place, the assessment of the effects of different model assumptions is straightforward.

*Background reading:*

[http://pubs.giss.nasa.gov/docs/notyet/submitted\\_Hall\\_Jewson.pdf](http://pubs.giss.nasa.gov/docs/notyet/submitted_Hall_Jewson.pdf),  
[http://pubs.giss.nasa.gov/docs/notyet/inpress\\_Hall\\_Jewson\\_2.pdf](http://pubs.giss.nasa.gov/docs/notyet/inpress_Hall_Jewson_2.pdf)

Potential for it to be a joint project? None.

Any prerequisite courses: Time series and Applied Probability

Computing required? Yes, a large amount; independent programming would be necessary.

Are data available? Yes, from public web sites, HURDAT TCs and SST data from the UK Met Office Hadley Centre.

8. **Title: Devising a temporal and spatial model for predicting wildfires**

**Proposers: Trevor Maynard (Lloyd's of London) and Gesine Reinert**

Brief Description: Warmer temperatures appear to be increasing the duration and the intensity of the wildfire season in the Western United States. This observation is reminiscent of the increase in frequency and intensity of hurricanes, begging the question whether the methods developed for hurricane forecasts by Gray and Klotzbach, or by Saunders, can be extended for predict wildfires.

Time series (if available) which may be explanatory (though it is not expected that all would be studied in the time permitted) would be:

- Regional area average temperatures
- Snow melt timings or snow depth/cover indices
- Relative humidity
- The Santa Ana winds (if available)
- Sea surface temperatures (Pacific(mostly)/ Atlantic?)
- Atmospheric wind parameters (e.g. Meridional and Zonal wind vectors at various heights (such as 850 hPa, 200 hPa) )  
<http://www.cdc.noaa.gov/cdc/data.ncep.reanalysis.html>

It is believed that no forecasting method is currently widely published so such a project could be very useful to the insurance industry.

The project will not only involve time series analysis, but also some understanding of stochastic modelling. The model will have to be implemented in R; thus good coding skills are required.

Suggested reading:

Saunders and Lea, *Seasonal prediction of hurricane activity reaching the coast of the United States*, in *Nature* (2005) Vol. 434 pp.1005-1008;

The Gray and Klotzbach web site

<http://hurricane.atmos.colostate.edu/Forecasts/>

On the Santa Ana winds: [http://en.wikipedia.org/wiki/Santa\\_Ana\\_wind](http://en.wikipedia.org/wiki/Santa_Ana_wind)

The Science article by Westerling et al.,

*Warming and Earlier Spring Increase Western US Wildfire Activity*, *Science* (2006), vol.313. p.940-943

<http://www.sciencemag.org/cgi/content/abstract/313/5789/940>

and the article by Running in the same volume, p.927-928.

Potential for it to be a joint project? None.

Any prerequisite courses Time Series; and Applied Probability

Computing required? Yes, a considerable amount. .

Are data available? To be confirmed.

**9. Title: The effect of ambiguity on social network statistics**  
**Proposer: Dr Andrew Dalby**

Brief Description: Actors in a social network are rarely described unambiguously and so some contextual knowledge has to be built into automatic network generation. One example is the problem of alternative names for actors, so for instance I am Andrew Dalby but I may be called Andy Dalby or just Andy and people may know who they mean. Andy is ambiguous but it can be made unique using a contextual description such as the Andy who is a lecturer in bioinformatics. When we recreate networks the possibility of errors occurring because of these ambiguities need to be detected.

Ha: That the presence of ambiguity alters the network statistics.

Ho: Networks with and without ambiguity have the same statistical properties.

Potential for it to be a joint project?: None

Prerequisite Courses: None

Computing required? Windows based machine running Pajek  
Data available? Internal mail social network

**10. Title: A model of metabolic network evolution**  
**Proposer: Dr Andrew Dalby**

Brief Description: A simplified model of metabolic evolution will be created where new nodes are created by duplication and subsequent mutation. Nodes will be assumed to be essential so that duplication is a pre-requisite to mutation. Altered function can only occur through mutation. Once a function has occurred, then the descendents of that node can evolve both divergently and convergently - so the same function can evolve on multiple nodes which then allows the loss of one of these duplicates. The model can be built from either of two viewpoints; the functional landscape can be assumed to exist and the nodes are then discovered as the new nodes search this space or the mutations and duplications can create the functional landscape.

Prerequisite Courses: Stochastic Models in Mathematical Genetics

Computing required? Yes but this can be provided for on Garuda.  
Data available? Yes - databases of known metabolic networks are available.

**11. Title: Ancestral inference from DNA Data**  
**Proposer: Prof Bob Griffiths**

Brief Description: Data sets of DNA can be used to deduce unique gene trees representing the mutation history of the sequences. It is possible to deduce ages of mutations and time to the most recent common ancestor of the sequences using importance sampling on the ancestral stochastic process back in time. The project will involve editing DNA data, deducing gene trees, and running computer software GENETREE for inference. A brief review of how the importance sampling works is also needed.

Potential for it to be a joint project?: It works best as an individual project, however two students could undertake it with different data sets.

Prerequisite Courses: The Mathematical Genetics course is advisable.

Computing required; Familiarity with general computing tasks involving running command line programs in Linux or Windows, text editing, data handling and manipulation, and R use will be required. It is best to write the dissertation in LaTeX.  
Data available; This project will use either genbank Drosophila data, other organism data, or human data from the HapMap project.

**12. Title: Sex ratio variation in human populations**  
**Proposer: Prof Bob Griffiths**

Brief Description: Does the sex ratio vary in different countries or regions?

Use descriptive statistics to discuss variation among countries and regions and other multivariate analysis techniques to investigate variation with co-variates. Background reading on genetics, demography, and social factors affecting the sex ratio is expected.

Potential for it to be a joint project?: No

Prerequisite Courses: Courses on Descriptive Statistics, Multiple regression, and Multivariate analysis.

Computing required; Familiarity with general computing tasks, text editing, data handling and manipulation, and R use will be required. It is best to write the dissertation in LaTeX.

Data available; Central Intelligence Agency – The World Factbook

<https://www.cia.gov/library/publications/the-world-factbook/index.html>

Sex ratio data (at birth, persons <15, persons 15-64, persons >64, total population) for about 225 countries.

Data for these countries on numerous other variables such as infant mortality rate (male, female, total), total fertility rate, birth rate, death rate, life expectancy at birth (male, female, total), gdp per capita, population growth rate, age structure, median age, net migration rate, HIV/AIDS prevalence, rate of poverty, unemployment rate, inflation rate.

**13. Title: The population dynamics of plasmacytoid dendritic cells in the immune response to HIV infection: building and fitting nonlinear dynamics models to cell census data.**

**Proposer: Geoff Nicholls**

Potential for it to be a joint project?: No

Prerequisite Courses: Bayes, MCMC, some knowledge of numerical methods for non-linear systems of differential equations may be an advantage.

Computing required? Yes

Data available? Yes, though data acquisition is ongoing, and it is of interest to use the inference to help design future studies.

Brief Description: Plasmacytoid Dendritic cells (pDC) play a part in the immune response to viral infections, and in particular, HIV/AIDS and its counterpart in monkeys, SIV. The pDC cells are generated at a steady rate in bone marrow. Some time after infection the lymph nodes become inflamed. It seems that pDC cells are recruited via the bloodstream to the lymph nodes to fight the infection. There is some doubt as to the fate of pDC cells in the lymph nodes. They can become infected themselves, or they can be removed by other mechanisms. It is unclear which if any of these mechanisms dominate. The purpose of the study is to compare models representing the principal hypothetical dynamics. This research was proposed and is being undertaken by Prof. Simon Barratt-Boyes, of the Center for Vaccine Research, at the University of Pittsburgh.

Population density data for pDC cells is available. Also, it is possible to track cohorts of the cell population within the overall population by a cell-labeling technique. The aim of the study is to build nonlinear models of the population dynamics, fit them, and compare them to one another. The statistical inference could be conducted in a number of ways. It is natural to use Bayesian inference, and MCMC simulation to fit and compare models, though likelihood ratio tests and related AIC model comparison may also be practical. Problems of this kind (comparing non linear dynamical systems as models of data gathered from systems evolving in time) are of current interest from both a theoretical standpoint And in many areas of applied statistics.

References:

C.T.H. Baker et al., "Computational approaches to parameter estimation and model selection in immunology", *Journal of Computational and Applied Mathematics* 184, 50–76, (2005)

G. Bocharov et al., "Mathematical modeling of the antiviral type I interferon response", Proceedings of the FOSBE, Eds F. Allgower and M. Reuss. Fraunhofer IRB Verlag, 325-330, (2007)

F. Geissmann, "The origin of dendritic cells", Nature Immunology 8, 558-560 (2007)

K. Liu et al., "Origin of dendritic cells in peripheral lymphoid organs of mice", Nature Immunology 8, 578-583 (2007)

#### **14. Title: Multilevel analysis and model checking**

**Proposer: Tom Snijders**

The purpose of this study is to make available and document a variety of techniques for assumption checking and diagnosis of multilevel models, and apply them in a multilevel study of a data set still to be determined.

A large multilevel data set will be made available. One candidate is the publicly available PISA 2006 data set. PISA is a survey of students' skills and knowledge as they approach the end of compulsory education, carried out in a large number of countries on behalf of the Organisation for Economic Cooperation and Development (OECD). The data contains information about students, parents, and schools. Several multilevel models will be estimated using these data. Diagnostic methods will have to be applied to assess the fit of candidate models and to find models that have, as far as can be determined, a good fit.

The diagnostic techniques to be used include the techniques treated in Gelman & Hill (2006), Snijders & Berkhof (2008), and Snijders & Bosker (1999), but these reference are meant to be orienting, not limiting. In any case, the following techniques will be applied: within-group OLS residuals, with and without standardization; empirical Bayes level-two residuals, with and without standardization; measures for assessing influence of cases and groups; diagnostics for linearity. Some of these techniques are available in the R libraries nlme and lme4. In the course of this project, the student will have to see what is available in these libraries and study the program code to see what exactly is being calculated and plotted in the available functions. As far as the techniques are not, or incompletely, available, the student will have to write and document the R code to calculate them.

#### References

Gelman, A., & Hill, J. *Data Analysis using Regression and Multilevel/Hierarchical Models*. CUP, 2006.

Pinheiro, J.C., and Bates, D.M. *Mixed-Effects Models in S and S-PLUS*. Springer, 2000.

Snijders, T.A.B. & Berkhof, J. [Diagnostic checks for multilevel models](#). Chapter 3 of J. de Leeuw & E. Meijer (eds.), *Handbook of Multilevel Analysis*. Springer, 2008.

Preprint available at [http://stat.gamma.rug.nl/handbook\\_ml\\_ch3.pdf](http://stat.gamma.rug.nl/handbook_ml_ch3.pdf) .

Snijders, T.A.B. & Bosker, R.J. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Sage, 1999.

#### **15. Title: Handling of missing data in multilevel analysis**

**Proposer: Tom Snijders**

The purpose of this study is to elaborate and apply a variety of techniques for handling missing data in multilevel models.

A large multilevel data set will be made available. One candidate is the publicly available PISA 2006 data set. PISA is a survey of students' skills and knowledge as they approach the end of compulsory education, carried out in a large number of countries on behalf of the Organisation for Economic Cooperation and Development (OECD). The data contains information about students, parents, and schools. Several multilevel models will be estimated using these data. The PISA data set is a high quality data set with a limited amount of missing

data, but still there are enough missings in this data set to make it important to handle the missing data in the best possible way.

Steps to be made in this project can be tentatively indicated as follows.

1. Obtain an overview of what is contained in the data set and select some relevant dependent and explanatory variables. Determine the number of cases with missing data for the selected variables in all countries. Select 3 to 5 countries to be used in the following analyses, based on the amount of missing data.
2. Carry out and report multilevel analyses under the missing completely at random (MCAR) assumption – presumably by listwise deletion. This could be called the naïve approach of handling missing data (handling by ignoring).
3. Elaborate diverse ways of analysing the data under the missing at random (MAR) assumption. These will have to include the bootstrap applied to randomly imputed data, multiple random imputation, and maximum likelihood analysis under the ignorability assumption. Investigate R libraries that could carry out these analyses.
4. Carry out the analyses according to these three methods. (If for any of the methods it is not feasible to do this, argue why it is infeasible.) Discuss the suitability of the various methods, and the difference between results obtained under the MCAR and under the MAR assumption.
5. Discuss the plausibility of the MAR assumption in this case, and give some suggestions on further analyses that could be done to study the sensitivity of the results obtained to the MAR assumption.

#### References

- Gelman, A., & Hill, J. *Data Analysis using Regression and Multilevel/Hierarchical Models*. CUP, 2006.
- Roderick J.A. Little and Donald B. Rubin, *Statistical Analysis with Missing Data*, 2nd edition. Hoboken, NJ: Wiley, 2002.
- Pinheiro, J.C., and Bates, D.M. *Mixed-Effects Models in S and S-PLUS*. Springer, 2000.
- Snijders, T.A.B. & Bosker, R.J. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Sage, 1999.

## **16. Title: Network analysis of the Correlates of War data**

**Proposer: Tom Snijders**

The *Correlates of War (COW)* data sets available at <http://correlatesofwar.org/datasets.htm> include rich information about relations between countries including alliances, geographic contiguity, trade, etc. This source also contains data about wars and other militarized conflicts. In this project, the dynamics of wars will be modeled, using characteristics of countries, and of pairs of countries, as explanatory variables, and also using endogenous network dynamics in the model. Models used will be along the lines of Snijders (2001), as implemented in the *Siena* program (Snijders et al., 2007), but a special issue here is that the relation is antagonistic whereas the models in these publications are geared toward relations with a friendly interpretation; this will mean large differences in the endogenous network dynamics. The dynamics will use ideas of balance theory. This is a classical theory in social psychology and has been elaborated in stochastic models that are, however, not directly applicable for statistical data analysis (e.g., Antal, Krapivsky and Redner, 2006; Ludwig and Abell, 2007).

Work on this project will include not only statistical modeling but also data handling, development of new features in existing models, and possibly some programming.

#### References

- Antal T., Krapivsky P.L., and Redner S. Social balance on networks: The dynamics of friendship and enmity. *Physica D* 224, 130–136 (2006).

- Ludwig, M. and Abell, P. An evolutionary model of social networks. *The European Physical Journal B* 58, 97–105 (2007).
- Snijders, Tom A.B., [The statistical evaluation of social network dynamics](#). Pp. 361-395 in *Sociological Methodology - 2001*, edited by M.E. Sobel and M.P. Becker. Boston and London: Basil Blackwell.
- Snijders, Tom A.B., Steglich, Christian E.G., Michael Schweinberger and Mark Huisman. [Manual for SIENA version 3.1](#). University of Groningen: ICS / Department of Sociology; University of Oxford: Department of Statistics, (2007).

**17. Title: Modelling later life mortality**

**Industry supervisors:**

**As sponsors of this project Hymans Robertson will provide access to the following supervisors**

**Dr Ana Maria Madrigal, Senior Statistician**

**Dr Deven Patel, Statistician**

**Steven Baxter, Development Actuary**

**Supervisor from Department of Statistics - TBC**

**Brief description:**

The pension promises made by UK employers are estimated to have a value of almost £1 trillion. In order to assess this value a number of statistical assumptions need to be made, including the distribution of deaths by age.

It is estimated that someone aged 65 today will on average die in their late 80s. Until recently relatively few individuals lived to these older ages and so assumptions have had to be made as to how the probability of dying at a given age increases with age. This project will challenge these assumptions in light of modern data. It will develop ideas introduced by Thatcher et al<sup>1</sup> to assess the relative merits of various contending models for the way in which the probability of dying changes for the age range 85 and over.

**Statistical techniques involved:**

We anticipate that the project will call heavily upon the following techniques:

Survival analysis

Generalised Linear Modelling

This means the project will be best suited to students who have completed the courses in Statistical Methods and Survival Analysis. Study of the optional courses in Actuarial Sciences will be an advantage in forming a financial interpretation of the statistical analysis.

**Computing knowledge required:**

We would anticipate that this project will be carried out in R. The student should have a good working knowledge of R, but can expect support from the industry supervisors. A working knowledge of Microsoft Excel will be an advantage for some aspects of the project.

**Additional information:**

This project will require access to Hymans Robertson's database of mortality statistics. The proprietary and confidential nature of this data would mean that this project would need to be conducted from Hymans Robertson's London office. In light of the requirement to work away from Oxford a living allowance will be paid by Hymans Robertson. If you would like additional information on these arrangements please contact Dr Neil Laws or Jan Boylan.

<sup>1</sup>The force of mortality at ages 80 to 120, Thatcher, A. R., Kannisto, V, Vaupel, J.W. Odense University Press 1998

**18. Title: Power calculations for national study of environmental radiation and childhood leukaemia**

**Supervisor: John Bithell**

**Source: JFB/GMK/CCRG**

**Problem:** A Poisson regression of areal data (observed and expected numbers of cases of childhood leukaemia in county districts) on levels of background ionising radiation supplied by the Health Protection Agency reveal equivocal results. This is often taken by the radiation community to mean that such background radiation presents no risk of childhood leukaemia, though a detectable effect might be expected from current risk coefficient estimates in a large population. The detectability of an effect, however, depends crucially on the geographical variability of radiation levels. The object of the study would be to attempt an estimate of the power of a typical study either from theoretical considerations or by simulation (or both).

**Data:** Real data are immediately available from the Childhood Cancer Research Group.

**Methods:** Poisson regression, with methods for determining the power, which need to be studied from the literature. R programming is envisaged, with C as a possible alternative.

**Student:** The student would need to have above-average mathematical and computing ability.

**20. Title: Testing anti-ageing treatments**

**Proposer: David Steinsaltz**

Brief Description: I have a collection of data culled from multiple research papers which compare populations differing by drug, dietary, or genetic manipulation, comparing one paper has come out of this, fitting the Gompertz model by different methods, to compare the slope parameters and decide whether the manipulations have actually changed the rate of ageing.

The project would be to extend this to a model with mortality plateaux (slowing of the rate of ageing), and to use simulations to test the effectiveness of standard asymptotic confidence intervals.

Potential for it to be a joint project?: No

Prerequisite Courses: Only core courses

Computing required? Yes

Data available? Yes

**21. Title: Air Pollution in Mexico Metropolitan area**

**Supervisor: B. D. Ripley**

Daily air pollution data is available for various parts of Mexico city and the surroundings at [http://www.sma.df.gob.mx/simat/home\\_base.php](http://www.sma.df.gob.mx/simat/home_base.php). For more background see <http://www.sbg.ac.at/ipk/avstudio/pierofun/mexico/air.htm>

The data could be analysed either as time series of levels, or binary time series (whether the warning level was exceeded or not that day).

There will be a small spatial aspect (there are five main areas, and their values are related).

This will need some programming in R to fit models by e.g. maximum likelihood.

**22. Title: Spatial biodiversity of tropical forests**

**Supervisor: B. D. Ripley**

Condit et al (2002) Science 295, 666-9 measured biodiversity of tropical forests at three different sites. They used spatial sampling, but their analysis did not take this into account.

Objectives:

- find ways to visualize the datasets taking spatial aspects into account.
- find a more suitable analysis of the data.

The methods will be mainly from multivariate analysis, but it will be necessary to learn something about spatial autocorrelation.

Data and supplementary material are available at <http://www.sciencemag.org/cgi/content/full/295/5555/666/DC1> and (in part) in R package 'vegan'.

Computing: R and GGobi.

**23. Title: Who's who in 12th Century England: witness list ranking**

**Proposer: Geoff Nicholls**

Brief Description: In 12th and 13th centuries (at least), the court moved with the king. Royal charters were issued at court. These legal documents were witnessed by those present. A scribe recorded the names and offices of the witnesses. Historians believe that the individuals are listed in order by social rank. How reliably does list rank indicate social rank?

If list rank really is a reliable indicator of social rank, then historians will be able to infer social rank in cases where it is unknown. For example, there is a hypothesis that the Bishop of London had a high rank among Bishops. Is there evidence for this in the lists? Did the rank of this office change over time?

We need to discover the underlying partial order (if any) imposed by social rank. We can use Monte Carlo tests of significance to test for evidence of hypothesized rankings. We can fit observation models for witness lists, in which the parameter is the unknown partial order. Besides social rank, there are a number of possible covariates for list order, including the place and date of issue of the charter. Dr Nicholas Karn, Lecturer in medieval history at Oxford, has compiled a large electronic database of witness lists. We will select certain interesting subsets of these data for analysis.

In order to get started, we can make a simple test choosing 2 individuals, A and B say, for whom we know the social rank does not change in time. We can consider the lists in which both appear. If  $g(p)$  is a suitable link function, we can fit a logistic regression model  $g(p_i) = a + bt_i$  for the probability  $p_i$  that A appears before B in the  $i$ 'th list, with  $t_i$  the date (AD) the list was made, and test for  $b=0$  by analysis of deviance. If  $b=0$  then we can test for  $p_i = 1/2$ . We can also take a pair where we know the social rank did change. This time we expect to estimate  $b$  non-zero. A simple exploratory tool might be built up from just this analysis, without the need for any more sophisticated analysis.

Potential for it to be a joint project?: probably not  
Prerequisite Courses: Model fitting and validation for logistic regression (and/or) Monte Carlo methods, Monte Carlo tests for significance.  
Computing required? Yes, R.  
Data available? Yes

**24. Title: Analyzing livelihoods data for farmers in Uzbekistan**

Proposer: Alex Conliffe (DPhil student in the Geography Department via Francis Marriott)  
Supervisor - TBC

I am in the process of analyzing livelihoods data for farmers in Uzbekistan that I collected via 400 quantitative surveys. The surveys were done across 4 communities and with 2 different types of respondents: farmers and peasants. For each of the 4 communities, I have 70 surveys from peasants and 30 surveys from farmers (farmers actually represent only about 5% of the rural population but I increased the number I surveyed to be able to conduct statistical analysis using them as a group). My data is a combination of nominal and interval data. I am using SPSS for my analysis.

As a first step, I have been looking for significant differences between communities and between peasants and farmers in relation to different variables (i.e. types of crops they grow, number of animals they own, whether someone in their house migrates to earn an income, etc.) to compare across these groups.

I now want to be able to do the following:

1/ Simultaneously compare between the 4 communities AND between farmers and peasants for a given variable AND interpret the results properly given that with this configuration I will be comparing groups with very different sample sizes. Be able to do this for both nominal and interval data.

2/ Determine whether factors other than which community a household lives in and whether they are a peasant or farmer household impact certain variables, and which factors are most important. For example, through initial analysis, I know that, on average, farmer households own more livestock than peasant households and that they also grow more fodder on their land than do peasant households. Now I want to know if, for example, the number of livestock a household owns, regardless of whether they are peasants or farmers, is a better predictor of whether or not they will grow fodder on their land. What other factors might be more important? I'd like to be able to do this for nominal, interval and combinations of nominal and interval data, again knowing how to interpret results given comparisons between groups with different sample sizes.